POLITECNICO DI MILANO
Doctoral School (Ph.D.) in Mathematical Engineering
XX cycle

ADAPTIVE AND REDUCED BASIS METHODS FOR OPTIMAL CONTROL PROBLEMS IN ENVIRONMENTAL APPLICATIONS

Presented to Dipartimento di Matematica "F. Brioschi"
POLITECNICO DI MILANO
by

Luca DEDE'

Graduated in Aerospace Engineering, Politecnico di Milano
ID number: *D*02085

Advisor

Prof. Alfio QUARTERONI

MOX, Modeling and Scientific Computing,
Dipartimento di Matematica "F.Brioschi",
Politecnico di Milano, Milano, Italy

Tutor: Prof. Luca FORMAGGIA
PhD Coordinator: Prof. Filippo GAZZOLA

Milano, March 2008

Acknowledgements

First of all I would like to acknowledge my advisor Prof. A. Quarteroni for having introduced me to the world of Numerical Analysis and Scientific Computing, for the interest, the research guidelines, the suggestions, the advices and his kindness.

I acknowledge Prof. A.T. Patera for having welcomed me in his group at M.I.T., for having introduced me to the world of Reduced Basis methods, for his availability, kindness and the support provided.

A special thank to Dr. G. Rozza for his support, suggestions, helpful discussions and friendship.

I acknowledge Prof. L. Formaggia, Dr. S. Micheletti, Dr. N.C. Nguyen, Dr. S. Perotto, Dr. M. Restelli and Prof. F. Saleri for the suggestions, help and useful discussions. Thanks also to Dr. L. Bonaventura, Prof. P. Gervasio, Dr. E. Miglio, Dr. F. Nobile, Prof. R. Sacco, Prof. A. Veneziani, Dr. Verani and Dr. P. Zunino.

Many thanks to all the members and persons met during these years at MOX and, in particular, to S. Minisini, Dr. A.B. Moura, R. Fronteddu and P. Ferrandi.

I acknowledge D. Blanchard for her kindness and friendly help. Many thanks also to the persons with whom I interacted at M.I.T., in particular, to Dr. G.S.H. Pau and Dr. S. Sen.

I acknowledge the financial support provided through the“Progetto Roberto Rocca”, M.I.T.–Politecnico di Milano collaboration. Many thanks to Dr. S. Sferza and Prof. F. Gazzola for the kind help.

Finally, special thanks to my Family for their support and patience.

Contents

Introduction

In this work we provide efficient numerical methods for the numerical solution of Partial Differential Equations (PDEs) and the computation of the associated outputs of interest, also in the frame of optimal control problems. In particular, we focus on environmental applications, specifically on atmospheric pollution problems, for which we are interested in evaluating the pollutant concentration in the computational domain and, moreover, in calculating the average concentration over certain areas, such as, e.g., a city. With this aim, a goal–oriented analysis [22] could be conveniently adopted in order to estimate the errors associated with the computation of such output functionals by means of Galerkin methods, remarkably the Finite Elements (FE) and Reduced Basis (RB) ones considered in this work.

In many engineering contexts, a major issue is represented by the prediction of the several quantities of interest for different physical and geometrical configurations. In these cases we handle with parametrized PDEs, which can be conveniently solved by means of the RB method [129, 136, 145], which allow to rapidly and accurately compute the associated output functionals. In this work we consider the RB method for the solution of parametrized advection–reaction PDEs which are formulated on the basis of a stabilized FE method. A "primal–dual" RB approach, allowing remarkable computational costs savings, is considered for the computation of the output functionals; both a priori and a posteriori RB error estimates are provided and discussed, thus outlining the properties of the RB method for this class of PDEs.

A possibly related issue is concerned with optimal control of parametrized PDEs. This is the case, e.g., of the regulation of the emissions from industrial chimneys in order to keep the pollutant concentration over critical areas below a prescribed threshold. In these cases, we are required to solve parametrized optimal control problems, for which the prediction of optimal control inputs is required every time that parameters change. The numerical solution of these problems could be computationally expensive, especially if described by unsteady PDEs, and emphasized in the many query context. In this work, we aim at extending the RB method to the case of parametrized optimal control problems, for which we expect the RB method to provide rapid and accurate solutions, as anticipated in [71, 149]. In particular, after having formulated the RB method for both unconstrained and constrained optimal control problems, we provide a posteriori RB error estimates based on a goal–oriented analysis. Moreover, we propose an "integrated" RB approach for the definition of the RB space, while taking into account the features of the optimal control problem.

Advection–dominated flows are often involved in atmospheric pollution phenomena, thus highlighting remarkable directional properties on the distribution of the pollutant concentration. In order to accurately capture these features while containing the computational costs, an efficient remedy is represented by the use of anisotropic mesh adaption procedures based on the FE method [7, 52, 54, 100]. These adaptive techniques, driven by a posteriori error estimates, allow us to smartly build meshes composed by suitably stretched and oriented elements, able to capture the directional characteristics of the solution. For this purpose, after having provided an anisotropic a posteriori error estimate in the energy norm for stabilized advection–diffusion–reaction PDEs, we aim at extending into the anisotropic setting the residual based a posteriori error estimate for output functionals.

The Thesis is organized as follows.
In Chapter 1 we deal with the environmental applications motivating this work and we handle with the aspects related to atmospheric pollution phenomena.
In Chapter 2 we discuss the optimal control theory for PDEs for both the unconstrained and constrained cases; then, after having recalled the optimization techniques, we discuss the numerical solution of an optimal control problem described by PDEs.
In Chapter 3 we consider a flow control problem, described by the incompressible Navier–Stokes equations, with the goal to minimize the drag acting on a blunt body. The Lagrangian functional framework together with a Lagrangian multipliers approach for the treatment of the Dirichlet boundary conditions are adopted.
In Chapter 4 we provide a general overview of the goal–oriented analysis based on Galerkin approximation methods and its extension to the case of optimal control problems described by PDEs.
In Chapter 5 we deal with the anisotropic setting for the goal–oriented analysis; we provide the a posteriori anisotropic error estimators for the energy norm and the computation of linear output functionals, together with the adaptive procedure for the anisotropic mesh generation.
In Chapter 6 we discuss the RB method for the approximation of parametrized steady advection–reaction PDEs. After having introduced the stabilized FE approximation, we provide and analyze the corresponding a priori and a posteriori RB error estimates; the "primal–dual" RB formulation is presented, together with an adaptive procedure for the choice of the RB spaces.
In Chapter 7 we formulate the RB method for the solution of parametrized optimal control problems described by parabolic PDEs. The a posteriori RB error estimates, based on the goal–oriented analysis, are proposed; our integrated RB approach for the choice of the RB space is then introduced.
Concluding remarks and an outlook to possible future developments follow.

Throughout this work, we use a standard notation to denote the Sobolev spaces of functions with Lebesgue measurable derivatives, their norms and, in general, the manifolds of Functional Analysis; see for instance [1, 27, 29, 109, 186]. Similarly, we use a standard notation for the Galerkin–Finite Element method, referring principally to [147, 153]; for further references, we refer the reader e.g. to [36, 49, 73, 96].

The simulations reported in this work have been carried out by means of *Matlab*®[192], *FreeFem++* [191] and *BAMG* [85] software. For the RB simulations reported in Chapter 7 the *rbMIT*©*MIT Software* [193] is used; in particular, for the numerical tests of Sec.7.2 an expanded module based on the *rbMIT*©*MIT Software* has been developed in collaboration with the Authors.

This work has been carried out at MOX–Modeling and Scientific Computing, Dipartimento di Matematica, *Politecnico di Milano*. Part of this work has been developed in collaboration with the Prof. A.T. Patera group, Fluids Laboratory, Mechanical Engineering Department, *Massachusetts Institute of Technology* and it has been partially supported by “Progetto Roberto Rocca”, M.I.T.–Politecnico di Milano collaboration.

Milano, 25 March 2008

Luca Dedè

Chapter 1

Environmental Applications: Atmospheric Pollution

During the last years themes generically indicated with “Environment” have made the object of a growing interest not only by the field specialists. In fact, environmental aspects are nowadays afforded and discussed from different points of view by a “multidisciplinary team” composed by scientists and politicians. This reflects a rising awareness that small perturbations on the delicate environmental equilibria and careless interactions between the human activities and the Environment could have significative social and economical consequences. Catastrophic events, such as e.g. earthquakes, tsunamis, flash floods, thunderstorms and hurricanes are extreme instances, however the human activities are dangerous as well since they produce air, water, soil and radioactive pollution. Hence, it is immediate to deduce that the forecast of environmental events and the control, regulation and planning of the human activities could enhance the life quality both in short and medium–long range periods; see e.g. [148].

In this work we address environmental aspects related to pollution phenomena as water and, above all, air pollution. In this context, it is immediate to notice that the study and the forecast of the dispersion of pollutants in the Atmosphere (above all in the low Troposphere) represents a crucial aspect in the monitoring of the air quality, the preservation of the Environment and the safeguard of the human being’s health. Among the sources of air pollution, those associated with human activities, such as industrial processes and transportation, are predominant. In particular, in this Chapter we focus on pollution phenomena due to the emissions by industrial chimneys (such as those associated with plants for energy production, industrial combustion or waste treatment). In this context, the evaluation of the impact of an industrial plant on the air quality represents a factor that must be taken into account in the design, the choice of the location and the daily management of such a plant. Similar considerations hold also in the case of water pollution phenomena.

Mathematical modeling and scientific computing are useful instruments to afford environmental problems, not only for the simulation and forecasts of environmental events, but also in view of the design and planning of human activities [148]. Customarily used in Meteorology for the weather forecasts since five decades already (see [26, 98, 171]), mathematical modeling and scientific computing are nowadays conveniently used in other nearby fields, such as e.g. hydrodynamic simulation, pollution phenomena, planning of the human activities and design of engineering works [44, 77, 149, 158]. This has been made possible, not only by the availability of faster and faster calculators, but also (and often above all) by the development and the use of efficient numerical methods [126]. In this context, we point out that some of the most recent techniques developed in Numerical Analysis, for instance mesh adaptivity [43, 52], Reduced Basis methods for parametrized problems [42, 136] and optimal control techniques [74, 107], represent useful and valid instruments for the analysis and simulation of environmental events.

For the environmental arguments treated in this Chapter we refer principally to [51]; for further deepening on atmospheric modeling, hydrodynamic applications, pollution phenomena and numerical methods for environmental applications, we refer the reader to, e.g., [3, 8, 25, 26, 46, 66, 77, 98, 123, 124, 137, 142, 156, 157, 158, 171, 175].
For the description of pollution phenomena, we make use of mathematical models based on advection–diffusion PDEs, while, for the numerical solution of these problems, the FE method [152, 153] is considered.

The Chapter is organized as follows. In Sec.1.1 we afford the aspects related to air pollution phenomena together with the corresponding mathematical models taking into account the influence of the Meteorology; finally, we provide some numerical examples. In Sec.1.2.1 we briefly recall the mathematical model applied to water pollution phenomena. In Sec.1.3 we discuss how mesh adaptivity, RB methods and optimal control techniques can be conveniently adopted for the solution of environmental problems.

1.1 Mathematical modeling for atmospheric pollution

In this Section we introduce the aspects related to air pollution phenomena, taking into account for the effects of meteorological conditions. We present a mathematical model and some numerical simulations referred to the case of pollutants emitted by industrial chimneys.

1.1.1 Atmospheric pollution

The Atmosphere is a stratified fluid composed essentially by noble gasses, Nitrogen in large part, Oxygen, Argon and traces of other gasses. We say that air is *contaminated* (polluted) if it contains composites (gasses or particles) able to produce, on the basis of their properties and concentrations, damages to living beings, vegetation and manufactures [51]. Even if the concentrations of the principal elements composing the air are substantially invariant, the amount of minor composites could largely vary, depending on both natural or human activities.
We observe that, once a pollutant is released in air, it is not immediate to forecast its evolution and distribution in time and space; in fact, diffusion, transport and reaction phenomena occur, often in combination with the evolution of other composites. In particular, each pollutant evolves with its own *temporal* and *spatial scales*, which could considerably vary from a composite to another. For this reason, it is important to evaluate the *life time*, i.e. the typical amount of time for which a given composite remains in Atmosphere before being removed by physical or chemical phenomena; on this basis it is possible to determine the typical temporal scale associated with a given pollutant. In the same manner, it is possible to classify the spatial scales as follows: *micro scale* (for phenomena with action limited to some meters), *urban scale* (till 10/50 km), *regional scale* (till 100/500 km), *synoptic scale* (till $1,000/5,000$ km)

and *global scale* (over $5,000\ km$). In particular, for pollutants emitted by industrial chimneys, the scales of major interest are the urban and the regional ones.

On the urban and regional scales the presence of pollutants is substantially due to the emissions of urban areas, road transport and industrial plants. In particular, typical pollutants for these scales are [51]:

- *Carbon Monoxide* (CO). It is due above all to traffic road (63%) and waste treatment 17%; industrial processes and other transportation systems contribute in minor part. Major effects are on human beings depending on concentration; in particular, behavior alteration, respiratory and cardiac problems and, in case of excessive concentrations, irreversible respiratory diseases.

- *Nitrogen Oxides* (NO and NO_2). Related to road transportation (49%), industrial combustion for manufacturing and energy production (32%) and other transportation systems (14%). Consequences on human beings are not completely clear; exposition to high concentrations could lead to significative respiratory and nervous problems.

- *Sulfur Dioxide* (SO_2). Natural emissions, such as volcanos, represent the major sources of this pollutant at a global level (61%); however, significative sources are due to the human activity, in particular to industrial combustion for manufacturing and energy production (31%) and road transportation (2%). Presence of this pollutant reveals on the vegetation, metals and construction materials, due to the generation of Sulfur Acid H_2SO_4 by interaction of Sulfur Oxides with water vapor. Consequences on the human being consist in irritations to the eyes and the respiratory system; high concentrations could lead to acute bronchial constriction and alteration of the nervous system. Moreover, SO_2 is able to generate a fine composite which penetrates deeply in the lungs.

- *Particulate* (PM_x). It indicates the solid particles dispersed in Atmosphere. The subscript x refers to the characteristic size of such particles in μm; PM_{10} is generated above all by natural erosion processes, while small particulate $PM_{2.5}$ is substantially due to combustion processes. The consequences of PM_{10} on human and living beings are limited, due to the presence of natural defences (mucous membranes and nostril hairs) which prevent the particulate to penetrate in the lungs. However, $PM_{2.5}$ is able to cross these defences, to penetrate deeply in the lungs and finally to reach the blood. This leads to respiratory diseases and increases the risk of cancer, especially if the particulate is chemically active.

The admitted concentration limits for pollutants are established by national or international regulations; in particular, Italian regulations (*D.M. 15/04/94*) define the following indicators:

- *Attention level*: a pollutant concentration which prefigures to reach the alarm level;

- *Alarm level*: a pollutant concentration which corresponds to an health and environmental risk.

These indicators are typically referred to short temporal scales; in Table 1.1 we report the attention and alarm levels for the pollutants previously considered.

Pollutant	CO	NO_2	SO_2	PM_{10}
Attention level $[\mu g/m^3]$	$15,000$	200	125	–
Alarm level $[\mu g/m^3]$	$30,000$	400	250	50

Table 1.1: Attention and alarm levels for the main pollutants (*D.M. 25/11/94*).

1.1.2 The general model

Let us consider a pollutant which on urban and regional scales does not significatively react with other composites. In this case, by considering an Eulerian approach, diffusion and transport phenomena of a pollutant could be described by means of the following parabolic advection–diffusion PDE:

$$\begin{cases} \dfrac{\partial C}{\partial t} - \nabla \cdot (\nu \nabla C - \mathbf{V} C) = f \qquad \text{in } \Omega \times (0, T_f) \\ + \, B.C.s \quad \text{and} \quad I.C., \end{cases} \tag{1.1}$$

endowed with appropriate initial (I.C.) and boundary conditions (B.C.s) for the pollutant concentration C; ν is the molecular diffusion of the pollutant in air, $\mathbf{V}$ is the advection field, while f is the pollutant source. We observe that, in general, ν, $\mathbf{V}$ and f depend space and time. With $\Omega \subset \mathbb{R}^3$, we indicate the computational domain in which we are interested in solving Eq.(1.1), while with T_f the final time step.

The numerical solution of Eq.(1.1) represents a fundamental issue in modeling and simulating atmospheric pollution phenomena. In fact, it is very important to use accurate and stable numerical methods able to preserve the total mass of the pollutant and the positivity of the variable C, especially for large spatial and temporal scales (see [156, 157]).

We observe that in Eq.(1.1) the diffusion of the pollutant in air is attributed to the molecular diffusivity ν; however, if the velocity of the fluid (air) is suitably described, dispersion due to turbulence phenomena is predominant. This is the case if compressible Navier–Stokes equations, or other suitable models as the Euler equations with turbulence models, are used to describe the air motions. However, the computational costs associated with the numerical solution of Eq.(1.1) could be large and unaffordable for problems of practical interest, which involve large temporal and spatial scales. To overcome this difficulty, the molecular diffusivity ν is replaced by a tensor $\mathbb{K}$ taking into account for the *turbulent diffusivity* in each direction; then Eq.(1.1) reads:

$$\begin{cases} \dfrac{\partial C}{\partial t} - \nabla \cdot (\mathbb{K} \nabla C - \mathbf{V} C) = f \qquad \text{in } \Omega \times (0, T), \\ + \, B.C.s \quad \text{and} \quad I.C.. \end{cases} \tag{1.2}$$

The components of the tensor $\mathbb{K}$ are in general unknown, even if they depend on the properties of the problem (Reynolds number), the geometry and the conditions of the air flux. Models for the study of pollutants in Atmosphere differ in the choice of the components of the tensor $\mathbb{K}$.

Experience shows that for an accurate numerical simulation of Eq.(1.2), a spatial resolution, greater than that associated with the advection field, is required.

We remark that Eq.(1.2) represents a general model only if chemical reactions are negligible. However, it is reasonable to assume that for the urban and regional scales, the pollutants

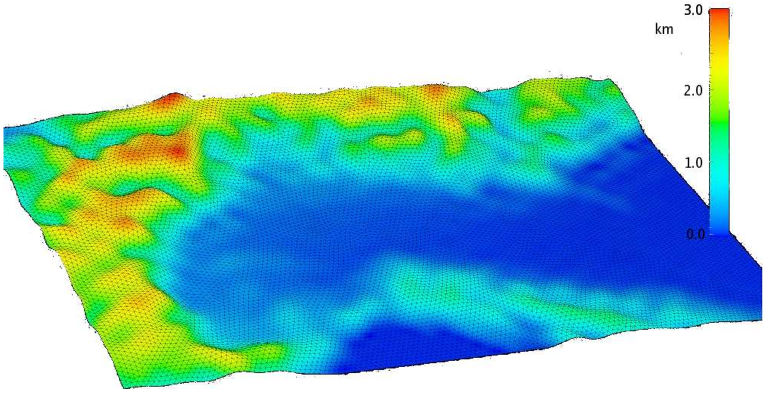

Figure 1.1: Soil orography of northern Italy; data provided by *ARPA–SIM, Emilia Romagna*, Italy.

considered in Sec.1.1.1 (CO, NO_x, SO_2 and PM_x) do not appreciably react with other components, for which model (1.2) reasonably holds.

1.1.3 The influence of Meteorology

As anticipated in Sec.1.1.2, a model describing the transport and diffusion of a pollutant in Atmosphere depends on the advection field $\mathbf{V}$ and the turbulent diffusion, described by $\mathbb{K}$. These depend in general on soil orography and Meteorology. But while for a given domain Ω the soil orography is known, the meteorological conditions could considerably vary. For this reason Meteorology represents a crucial aspect in the simulation of transport and diffusion of pollutants, hence the importance of the weather forecasts in foreseeing critical environmental conditions.

The use of mathematical modeling and scientific computing in meteorological applications has been widely developed since 1950s, above all for weather forecasts. In this context, several mathematical models and accurate and efficient numerical methods have been considered and are nowadays in continuous development; see e.g. [25, 26, 98, 137, 142, 171]. However, even if their contribution to Meteorology has been fundamental, it is important to remark that an intrinsic uncertainty factor remains in each forecast, due to the impossibility to correctly simulate phenomena involving small spatial and temporal scales and to consider the proper initial and boundary conditions. Different models and numerical methods are considered depending on the spatial and temporal scales on which weather forecasts refer; for example, on regional and synoptic scales, initial and boundary conditions are continuously corrected on the basis of experimental data and the results of global scales forecasts.

As an example, Fig.1.1 shows the soil orography of northern Italy; in Fig.1.2 we report the results of a synoptic scale weather forecast (including Italy) realized by *Agenzia Regionale Prevenzione Ambiente, Servizio Idro–Meteo (ARPA–SIM), Emilia Romagna*, Italy, in which the horizontal wind conditions at 1 km and 3 km over sea level are highlighted. These forecasts have been obtained by using the Euler equations with turbulence models and taking into account for the interaction with water vapor; for the numerical solution, the Finite Volume method [105] has been used. A structured mesh has been used with an horizontal spatial resolution of about 6.85 km and vertical resolution varying from 100 m near soil to

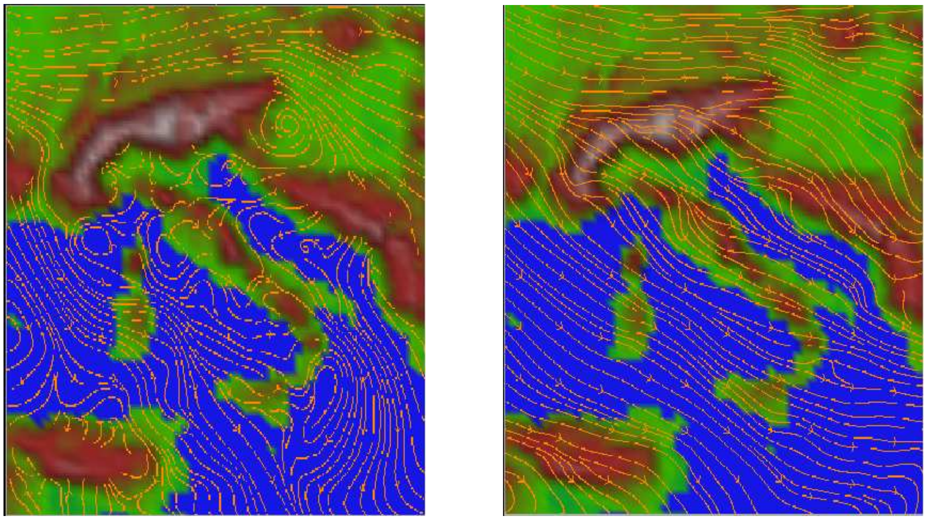

Figure 1.2: Weather forecast (20 July 2005, time 12:00) realized by *ARPA–SIM, Emilia Romagna*, Italy; wind field at 1 km (left) and 3 km (right) over seal level.

400 m at ceiling (9 km).

On the basis of the weather forecasts, which provide the wind field, it is immediate to deduce the advection field $\mathbf{V}$ in Eq.(1.2). However, the components of the tensor $\mathbb{K}$, which accounts for turbulence phenomena, must be determined on the basis of empirical laws based on further elaborations on the results of simulations. In fact, while mechanical turbulence depends on the orography, convective turbulence is generated by air motions related to thermal conditions; in particular, the temperature of the air provides a useful indication of the behavior of air masses and the dispersion properties of the Atmosphere. With this aim, let us introduce the *potential temperature* ϑ as [51]:

$$\vartheta := T \left(\frac{p}{10^5} \right)^{\frac{k-1}{k}}, \tag{1.3}$$

where p is the pressure expressed in Pa, T the temperature in K and $k = 1.4$ for air; ϑ corresponds to the temperature of a particle of air which is adiabatically led to the reference pressure of 10^5 Pa. Moreover, we define the *potential lapse rate* $\partial\vartheta/\partial z$ as the vertical gradient of the potential temperature, which is equal to:

$$\frac{\partial \vartheta}{\partial z} = \frac{\partial T}{\partial z} + \beta, \tag{1.4}$$

being $\partial T/\partial z$ the *lapse rate*, while $\beta = -9.8°C/km$ is the *dry adiabatic lapse rate.*
Depending on the values assumed by $\frac{\partial\vartheta}{\partial z}$, the stability of the Atmosphere could be classified as follows:

- if $\frac{\partial\vartheta}{\partial z} > 0$, air is *stable* (vertical air motions are inhibited and diffusion of pollutants is limited);
- if $\frac{\partial\vartheta}{\partial z} \simeq 0$, air is *neutral* (it is a situation of indifferent equilibrium, in which diffusion of pollutants is dominated by mechanical turbulence);

Class	Atmosphere	$\partial\vartheta/\partial z$ $[^oC/m]$
A	Strongly unstable	< -0.0092
B	Moderately unstable	$-0.0092 \div -0.0072$
C	Lightly unstable	$-0.0072 \div -0.0052$
D	Neutral	$-0.0052 \div 0.0048$
E	Stable	$0.0048 \div 0.0248$
F	Strongly stable	> 0.0248

Table 1.2: Atmosphere stability classes, Pasquill–Gifford classification.

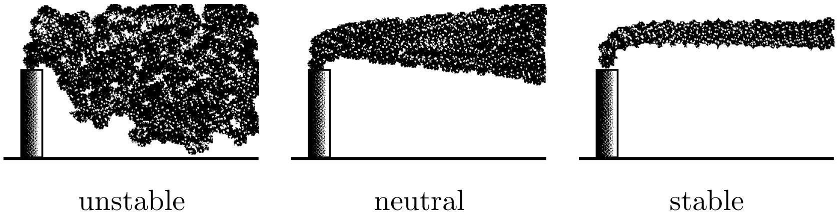

Figure 1.3: Illustrative example of the effects of air stability on a pollutants plume emitted by a chimney.

- if $\frac{\partial\vartheta}{\partial z} < 0$, air is *unstable* (convective motions of air masses occur, for which diffusion of pollutants is amplified).

Typically, stability classes are defined according with the Pasquill–Gifford classification (see Table 1.2 and [51]), which indicates with letters from A to F the atmospheric conditions, in increasing order, from strongly unstable to strongly stable.
We notice that atmospheric stability strongly depends on the luminous radiation at soil, varying considerably from day to night and from clear to cloudy conditions. Atmosphere in unstable conditions could represent a critical situation in the case of pollutants emitted by industrial chimneys; in fact, the pollutants tend to diffuse on a large area, but they could deposit in high concentrations at the soil near to the source [40]. Critical situations could occur in the case of inversion in quota of $\frac{\partial\vartheta}{\partial z}$, which is varying from negative or null values at soil to positive values in quota, thus confining the pollutants in the lower layer of the Atmosphere. In Fig.1.3 we report an illustrative example about the effects of the atmospheric stability on the pollutants plume emitted by an industrial chimney.

1.1.4 The Gaussian models

In this Section we propose a model for the transport and diffusion of pollutants emitted by industrial chimneys on urban and regional scales. With this aim, we make use of the *Gaussian models*, for which the concentration of the pollutants assumes the form of a Gaussian distribution in space [51]. This kind of models allows us to fit the coefficients of the tensor $\mathbb{K}$, which we assume to be in the following form:

$$\mathbb{K} = \begin{bmatrix} K_{\mathrm{xx}} & 0 & 0 \\ 0 & K_{\mathrm{yy}} & 0 \\ 0 & 0 & K_{\mathrm{zz}} \end{bmatrix}, \tag{1.5}$$

Class	σ_o $[m]$	σ_v $[m]$
A	$-2.50r^2+175r$	$-1.50r^2+195r$
B	$-2.08r^2+146r$	$-1.67r^2+167r$
C	$-1.67r^2+107r$	$-0.917r^2+134r$
D	$-1.07r^2+73.7r$	$-1.08r^2+65.8r$
E	$-0.750r^2+52.5r$	$-0.667r^2+41.7r$
D	$-0.750r^2+47.5r$	$-0.583r^2+35.8r$

Table 1.3: Dispersion coefficients for mixed urban and rural areas for different stability classes; r in km.

being z the vertical spatial coordinate, while x and y those associated with the horizontal plane. The Gaussian models allow to determine empirically the values of the coefficients K_{xx}, K_{yy} and K_{zz} on the basis of the orography, atmospheric stability conditions and spatial coordinates. However, the validity of these models is limited to some particular conditions; in fact, we require that the following hypotheses are satisfied:

- the advection–diffusion process (1.2) is stationary and the source term f is time independent;
- meteorological variables (pressure, temperature, wind velocity) are time independent;
- $K_{\mathrm{xx}} = K_{\mathrm{yy}} = K_h$, with $K_h(\mathrm{r})$ and $K_{\mathrm{zz}}(\mathrm{r})$ functions of the radial coordinate $\mathrm{r} = \sqrt{(\mathrm{x}-\mathrm{x}_0)^2+(\mathrm{z}-\mathrm{z}_0)^2+(\mathrm{z}-\mathrm{z}_0)^2}$, being $(\mathrm{x}_0, \mathrm{y}_0, \mathrm{z}_0)$ the coordinates of the source.

Coefficients K_h and K_{zz} are evaluated as:

$$K_h = \frac{\sigma_h^2 V}{2\mathrm{r}} \qquad \text{and} \qquad K_{\mathrm{zz}} = \frac{\sigma_v^2 V}{2\mathrm{r}}, \tag{1.6}$$

where $V = \|\mathbf{V}\|$ and σ_h and σ_v are the *dispersion coefficients* according with the Gaussian models. The dispersion coefficients are empirically related to the soil orography, stability class and distance from the pollutant source, as reported in Table 1.3 for mixed urban and rural areas (see [40, 51]). For example, we deduce that a pollutant particle located 1 km away from the emitting source in neutral air, moves in average of 73 m in horizontal direction, while 63 m in vertical direction w.r.t. the original wind trajectory.

By taking into account for the Gaussian model previously outlined, Eq.(1.2) finally reads:

$$\begin{cases} -\nabla\cdot(\mathbb{K}\nabla C - \mathbf{V}C) = f & \text{in } \Omega, \\ C = 0 & \text{on } \Gamma_D \\ \mathbb{K}\nabla C\cdot\hat{\mathbf{n}} = 0 & \text{on } \Gamma_N, \end{cases} \tag{1.7}$$

where Γ_D is the part of the boundary $\partial\Omega$ s.t. $\mathbf{V}\cdot\hat{\mathbf{n}} < 0$, being $\hat{\mathbf{n}}$ the outward directed unit vector normal to $\partial\Omega$, and $\Gamma_N := \partial\Omega\backslash\Gamma_D$. The source term f assumes the expression $f = Q/|E|\chi_E$, being χ_E the characteristic function of the localized emission subdomain $E \subset \Omega$ and Q the rate of emission in the subdomain E, whose volume is $|E|$. We observe that the subdomain E does not coincide with the chimney emission area; its location depends in fact on several factors, as the atmospheric stability class, wind field, chimney geometry, emission

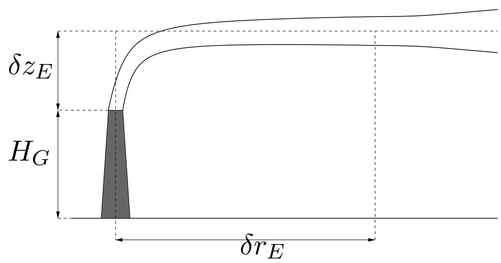

Figure 1.4: Displacement of the plume emitted from a chimney.

velocity of the plume, internal and external temperatures. In Fig.1.4 this effect is shown by means of an illustrative example. Experimental models, as the Briggs formulas, can be used to estimate the location of the subdomain E (see [51]); in particular, we assume that E is a block s.t. $|E| = 8\sigma_h^2\sigma_v$ for $r = 50$ m.

Remark 1.1. *Qualitative indications could be obtained by simpler models, as the quasi–3D model presented in [40, 44], for which a 2D advection–diffusion problem for the concentration is solved in the horizontal plane at a certain reference quota; concentration at soil is then obtained by projecting the solution in quota in analogy with Gaussian models.*

1.1.5 Numerical tests

In this Section we provide some numerical tests based on the model introduced in Sec.1.1.4. We consider the case of an hypothetic industrial plant for energy production, for which we are interested in monitoring the distribution of the concentration of SO_2 (typical of gasoline combustion) over a certain area. Two rates of emission for SO_2 are considered $Q_1 = 400$ kg/h and $Q_2 = 1800$ kg/h. The chimney, with geometrical height of 250 m, is located in a flat area.

The problem (1.6) is solved by means of the FE method with $\mathbb{P}^1$ basis functions with meshes composed by tetrahedral elements [147, 153]. The displacement of the subdomain E w.r.t. the original position of the chimney is taken into account while generating the computational domain Ω and the corresponding mesh. We observe that Ω assumes the shape of a tiny layer with vertical size of 4 km, which is by far lower than the other dimensions (about 40 km). For this reason, anisotropic meshes have been generated in order to increase the efficiency of the numerical solution; see Chapter 5.

Test 1

We firstly consider a test case with an imposed wind field $\mathbf{V} = V\hat{\mathbf{v}}$, with $V = 9$ km/h and $\hat{\mathbf{v}} = 1/\sqrt{2}\hat{\mathbf{x}} - 1/\sqrt{2}\hat{\mathbf{y}}$, for different classes of atmospheric stability; the emission rate Q_1 is considered. Vertical and horizontal displacements of E w.r.t. the chimney are assumed as 500 m and 1800 m along the wind direction, respectively. A mesh with $543,902$ tetrahedra and $92,788$ nodes has been considered.

In Fig.1.5(left), we report the iso–surfaces of the pollutant concentration corresponding to the values 3 $\mu g/m^3$, 10 $\mu g/m^3$ and 12.5 $\mu g/m^3$ for unstable atmosphere (class A). In Fig.1.5(right) we report the corresponding trace of pollutant at soil; the red dot indicates the location of the chimney, while the grey dots the location of hypothetic cities. Scales are in km, concentrations

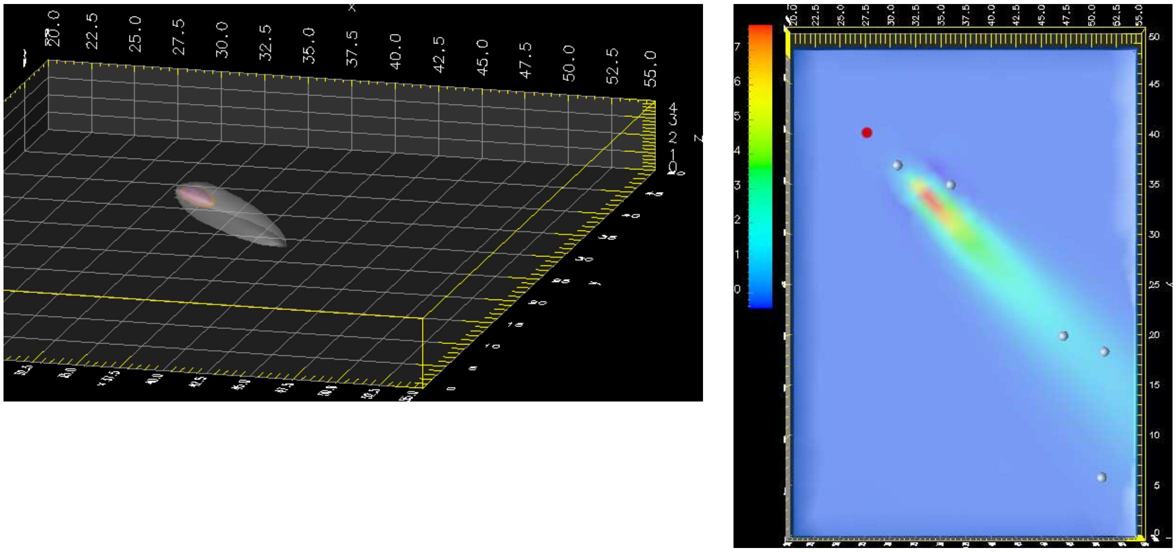

Figure 1.5: Test 1. Iso–surfaces of SO_2 concentration corresponding to the values 3 $\mu g/m^3$ (gray), 10 $\mu g/m^3$ (orange) and 12.5 $\mu g/m^3$ (violet) (left) and concentration map at soil (concentration in $\mu g/m^3$) (right); strongly unstable Atmosphere (class A).

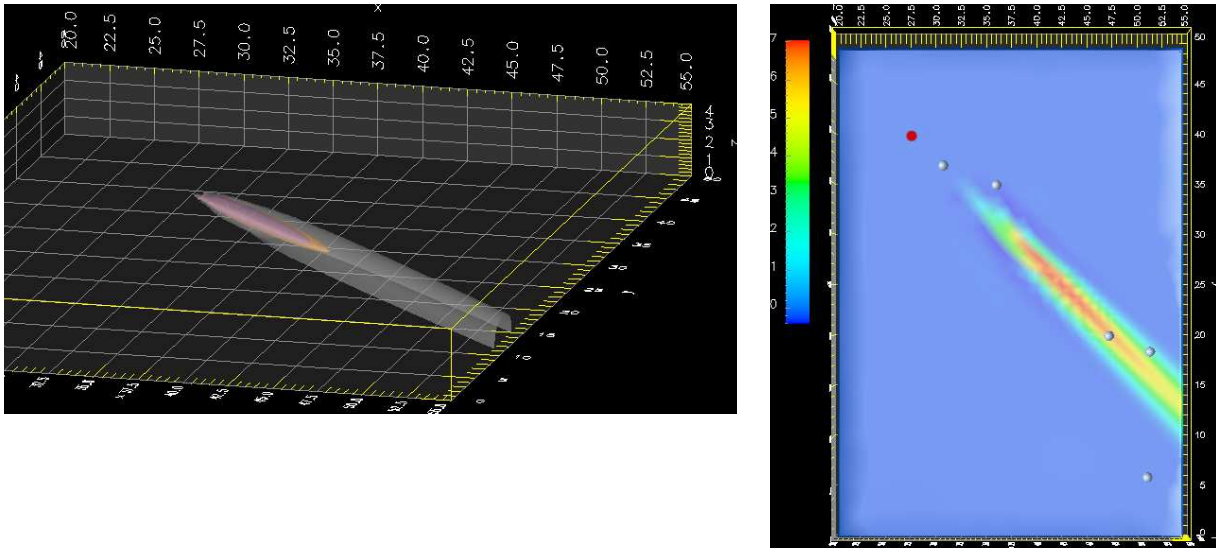

Figure 1.6: Test 1. Iso–surfaces of SO_2 concentration corresponding to the values 3 $\mu g/m^3$ (gray), 10 $\mu g/m^3$ (orange) and 12.5 $\mu g/m^3$ (violet) (left) and concentration map at soil (concentration in $\mu g/m^3$) (right); neutral Atmosphere (class D).

in $\mu g/m^3$. Similarly, in Fig.s 1.6 and 1.7, we report the cases of neutral (class D) and stable (class F) atmosphere, respectively. We notice that, as the atmospheric stability increases, then the peak of concentration at soil tends to move away from the chimney, even if affecting

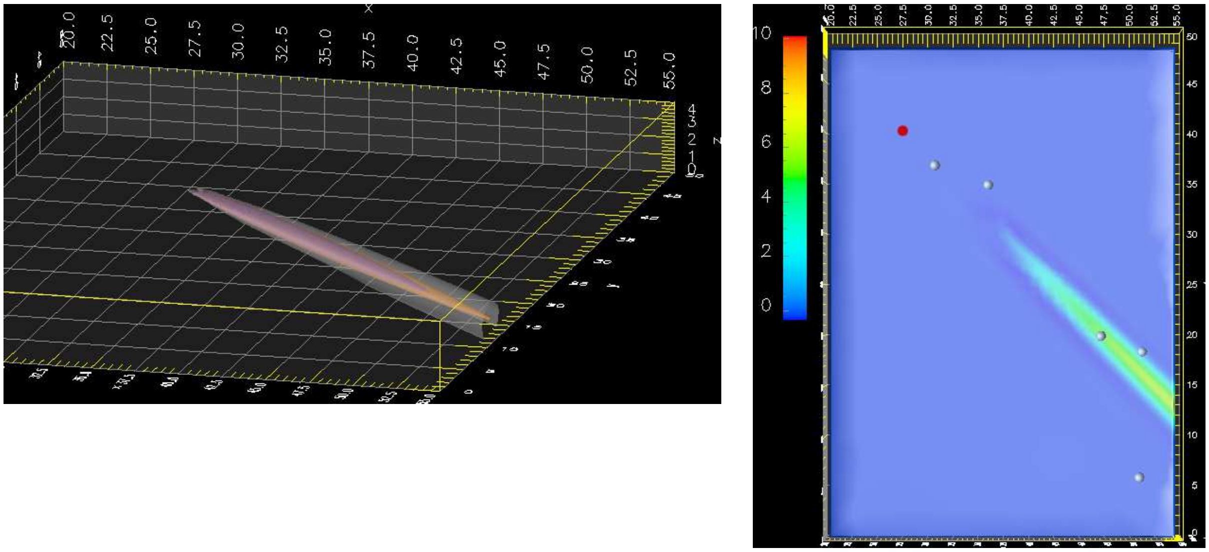

Figure 1.7: Test 1. Iso–surfaces of SO_2 concentration corresponding to the values 3 $\mu g/m^3$ (gray), 10 $\mu g/m^3$ (orange) and 12.5 $\mu g/m^3$ (violet) (left) and concentration map at soil (concentration in $\mu g/m^3$) (right); strongly stable Atmosphere (class F).

a smaller area.

Test 2

In this case we provide a numerical simulation based on the data obtained by a weather forecast, on which basis we deduce the wind field $\mathbf{V}$ and the stability class and hence, the tensor $\mathbb{K}$. In particular, we make use of the weather forecast provided by *ARPA–SIM, Emilia Romagna*, Italy, for the day 20 July 2005, time 12:00. The geographical coordinates of the hypothetic industrial chimney are $45.332°N$, $9.4350°E$. For this simulation, vertical and horizontal displacements of E w.r.t. the location of the chimney are assumed as 530 m and 1810 m, respectively. A mesh with $418,286$ tetrahedra and $70,450$ nodes has been considered. The computational cost (CPU time) associated with the numerical solution of this problem has been of $35,640$ s for the FE matrices assembling and $5,420$ s for the solution of the FE linear system; computations have been performed on an AMD Opteron^{TM} 1.8 GHz processor, with 1024 KB of memory cache and 8 GB of RAM.

In Fig.1.8 we report the wind field (left) and the potential slope rate $\partial\vartheta/\partial z$ (right) w.r.t. the quota z in the location of the chimney. In particular, we observe that the wind field changes direction and verse while increasing the quota. Similarly, atmospheric stability varies among the classes D and E (neutral and stable air). In Fig.1.9(left), we report the iso–surfaces of the pollutant concentration corresponding to the values 3 $\mu g/m^3$, 10 $\mu g/m^3$ and 12.5 $\mu g/m^3$ for the emission rate Q_1; atmospheric stability class and the tensor $\mathbb{K}$ are deduced by means of the profile of the potential slope rate. In Fig.1.9(right) we report the concentration trace of SO_2 at soil corresponding to the emission rate Q_2. As for Test 2, the red dot indicates the location of the chimney, while the grey ones the location of hypothetic cities; scales are in km, concentrations in $\mu g/m^3$.

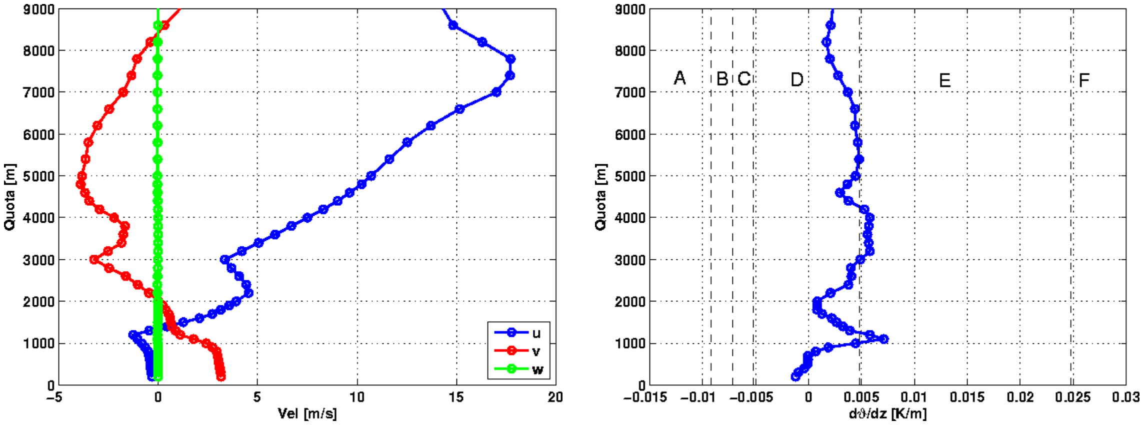

Figure 1.8: Test 2. Wind field $\mathbf{V} = [u, v, w]^T$ ($[m/s]$) (left) and potential slope rate ($[K/m]$) (right) w.r.t. the quota z ($[m]$) in the chimney location; atmospheric stability classes are shown. Elaborations of weather forecasts provided by *ARPA–SIM, Emilia Romagna*, Italy.

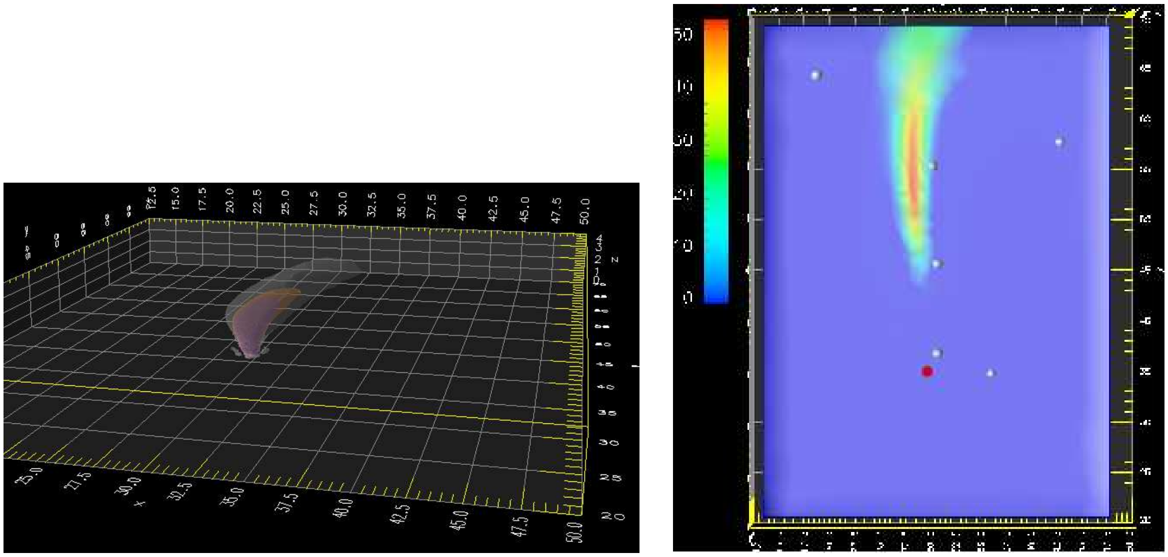

Figure 1.9: Test 2. Iso–surfaces of SO_2 concentration corresponding to the values 3 $\mu g/m^3$ (gray), 10 $\mu g/m^3$ (orange) and 12.5 $\mu g/m^3$ (violet) for Q_1 (left) and concentration map at soil for Q_2 (concentration in $\mu g/m^3$) (right).

1.2 Mathematical modeling for water pollution

In this Section we briefly describe mathematical models for pollution phenomena in water, in analogy with models for air pollution.

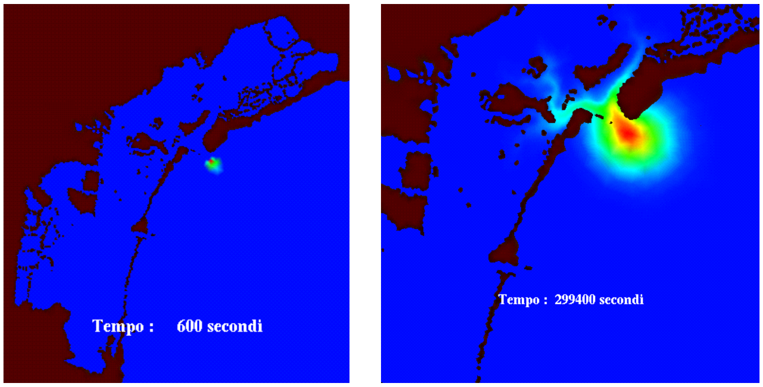

Figure 1.10: Concentration of a pollutant released in front of the Venice Lagoon at two different time steps. Courtesy of E. Miglio (MOX, Politecnico di Milano).

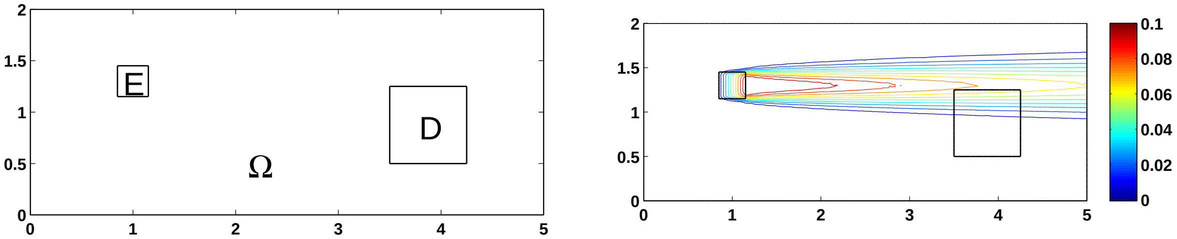

Figure 1.11: Computational domain and subdomains (left) and solution (right) of an academic $2D$ pollution problem.

1.2.1 The general model and Shallow water equations

As in the case of air pollution phenomena, advection–diffusion PDEs properly describe the distribution of a pollutant in water. In fact, under the same hypotheses of Sec.1.1.2, the parabolic Eq.(1.2) holds; even in this case, the (water) velocity field $\mathbf{V}$ and the tensor $\mathbb{K}$ must be provided. The velocity field $\mathbf{V}$ is usually obtained by means of the numerical solution of Shallow Water models which, given the bathometry of a river, lake or sea, allow to describe the motion of the water; see [3, 46, 123, 124]. Similarly to the atmospheric case, the diffusion tensor $\mathbb{K}$ should take into account for turbulence effects.

As example, in Fig.1.10 we report the results of a numerical simulation for which a pollutant, say an oil spot, is released in front of the Venice Lagoon, Italy, and it diffuses according with the tide motions, computed by means of a Shallow water model; see [123, 148].

1.3 Environmental applications and Numerical Analysis

In this Section we discuss how the numerical methods, as mesh adaption for the FE method, Reduced Basis methods and optimal control techniques, could be used for environmental applications.

With this aim, let us introduce an example concerning a $2D$ atmospheric pollution prob-

lem. In particular, referring to Eq.(1.7), we consider $\mathbb{K} = I$, $\mathbf{V} = [300, 0]^T$, $Q = 9$ and $\Omega = (0, 5) \times (0, 2)$. Moreover, we introduce an emission subdomain $E = (0.85, 1.15) \times (1.15, 1.45)$ and an *observation* subdomain $D = (3.5, 4.25) \times (0.5, 1.25)$; we recall that the pollutant source term is $f = Q/|E|$. In Fig.1.11(left) we report the computational domain Ω together with the subdomains E and D. In Fig.1.11(right) we report the concentration C, obtained by means of the FE method (with $\mathbb{P}^1$ basis functions and SUPG stabilization [122, 152, 153]) with a mesh composed by $9, 600$ equally distributed triangular elements and $4, 913$ nodes.
In particular, beyond the evaluation of the pollutant distribution in the computational domain Ω, we can be interested in calculating the average (or maximum) concentration of such a pollutant in a limited observation area (e.g. the subdomain D), which, in practise, could correspond to a city [40, 44]. With this aim, we define an *output functional*, say $s(C)$, as:

$$s(C) := \frac{1}{|D|} \int_D C \, dD, \tag{1.8}$$

which evaluates the average concentration of the pollutant in the hypothetic city located in D.

As anticipated, the numerical solution of Eq.(1.7) (and in general of Eq.(1.2)) and the evaluation of the associated output functional (1.8) could lead to considerable computational costs. In fact, considering the FE method, very fine meshes are required in order to avoid numerical instabilities and to obtain sufficiently accurate solutions [152, 153]. This represents an huge limit when numerical simulations are used to foresee risk situations associated with pollutant emissions; for example those occurring when weather conditions suddenly change or the rates of emissions go over the fixed limits. It is evident that in these cases the "response" of the numerical simulation should be sufficiently rapid and accurate at the same time. To overcome this difficulty, a possible strategy to increase the efficiency of the numerical simulation associated with the FE method consists in adopting well–suited meshes. These are typically generated by means of adaptive procedures based on a posteriori error estimates, see e.g. [16, 22, 61, 153, 178]; in particular, a posteriori error estimates could be defined for the error committed on the solution C of the problem or on the output functional $s(C)$ (goal–oriented estimates), if we are more interested in evaluating $s(C)$ rather than the complete solution. We observe that the meshes generated by means of this two types of estimates could sensibly differ, as we can deduce by observing Fig.1.11(right), where the output $s(C)$ is affected by a minor part of the pollutant plume [40, 44, 151]. More over, due to the anisotropic features of the solutions of advection–diffusion PDEs, as those associated with pollution problems, it is convenient to make use of meshes composed of anisotropic elements, see e.g. [43, 52, 53, 121, 122, 169]. For this reason, in Chapter 5, we propose and study an anisotropic a posteriori error estimate based on the goal–oriented analysis and an adaptive procedure for the mesh generation applied to advection–diffusion–reaction PDEs.

In Sec.1.1.3 we have observed that the distribution of a pollutant in a certain area strongly depends on the Meteorology and on the soil orography; by referring to Eq.(1.7), we recall that these factors influence the turbulence diffusivity tensor $\mathbb{K}$ and the velocity field $\mathbf{V}$. Hence, we deduce that each time the weather conditions vary, then we have to solve newly Eq.(1.7). However, as highlighted previously, this could be computationally expensive and prevent the possibility of "real–time" simulations of the pollutant concentration. Similarly, in a "many–query" context we could be interested in recursively solving Eq.(1.7) in order to study the effects of the location of a chimney for different weather conditions, e.g. as required by a preliminary study of the impact of an industrial plant. In this case, we can view Eq.(1.7) as parametrized PDE depending on a set of *parameters* $\boldsymbol{\mu}$ (the tensor $\mathbb{K}$, the velocity field $\mathbf{V}$ and the location of the chimney), for which the concentration C, solution of the parametrized Eq.(1.7), depends on $\boldsymbol{\mu}$ as $C = C(\boldsymbol{\mu})$; similarly, the output functional reads $s = s(\boldsymbol{\mu})$. For this kind of problems, a Reduced Basis (RB) method can be used, see e.g. [72, 113, 114, 129, 136, 145, 150, 162, 163, 165]; the RB method is a well–known technique which allows to evaluate rapidly and accurately an output $s(\boldsymbol{\mu})$ associated with PDEs with parametric dependence. In particular, in Chapter 6, we consider the RB method for the solution of parametrized advection–reaction PDEs, which describe transport dominated pollution phenomena, and the evaluation of the associated outputs [42].

An other interesting issue in environmental applications consists in regulating the pollutant emissions, i.e., referring to the case of an industrial chimney, in controlling the emission rate. In this context, optimization and optimal control techniques can be conveniently adopted to solve this kind of problems (see Chapter 2 and e.g. [2, 74, 107, 108]). By considering Eq.(1.2), the problem consists in regulating the rate of emission of a certain pollutant Q s.t. the pollutant concentration in a certain area and for a certain time interval is minimum. This problem can be stated as functional minimization problem, with a cost functional J in the following form:

$$J(C, Q) = \frac{1}{2} \int_0^T \int_D C^2 \; dDdt + \frac{1}{2}\gamma \int_0^T (Q - Q_d)^2 \; dt, \tag{1.9}$$

where Q_d represents the "ideal" emission rate of the industrial plant and γ is a positive number. Constraints on the rate of emissions Q are possible, as $Q_{min} \leq Q \leq Q_{max}$, related to the properties of the industrial plant, and $\int_0^T Q \; dt \leq T\overline{Q}$, expressing a limit on the total emissions, regardless the area interested by the pollutant. We observe that for an industrial plant for energy production, combustion of a mixture of gasoline and natural gasses could occur; in this case, a possible optimization problem consists in regulating the emission rate Q of pollutants associated with gasoline combustion, while preserving the total energy production by adding natural gasses.

Moreover, we notice that the optimal control problem could be set in a parameter dependent context, for which the parameters take into account for the weather conditions or the chimney location. In this case, we are interested in finding an optimal emission rate Q for any given wind field, atmospheric conditions and chimney location. The numerical solution of optimal control problems is in general computationally expensive, especially for problems dealing with environmental equations. For this reason, reduced basis techniques have been introduced for solving optimization problems, even if in a parameter independent context, see e.g. [78, 80, 91, 92, 93]; for parameter dependent problems a first use of the RB has been introduced in [149] in order to speed up the optimization process. In Chapter 7, we consider the RB method for the solution of optimal control problems described by parametrized PDEs, which we indicate as parametrized optimal control problems.

Chapter 2

Optimal Control for Partial Differential Equations: Theory and Numerical Methods

In this Chapter we report the basic concepts of the optimal control theory for PDEs together with the numerical methods used for the solution of optimal control problems. We consider both unconstrained and constrained optimal control problems, governed by steady and unsteady PDEs, specifically elliptic and parabolic PDEs.

The classical approach to optimal control for PDEs is based on the theory developed by *J.L. Lions* in [107, 108], which provides existence and uniqueness results for optimal control problems described by elliptic, parabolic, hyperbolic and mixed PDEs. However, by the theory of *J.L. Lions* a straightforward analysis for a broader class of optimal control problems is not always easy, e.g. for those described by more general equations, with non–linearities or boundary control (see e.g. [19, 39, 41, 44, 101] and Chapter 3). A complimentary approach which allows to handle this wider class of optimal control problems straightforwardly is based on the *Lagrangian* formalism, see [20, 74, 116, 118, 174], as well as, e.g. [75, 76, 119, 182]. This approach, which is based on the Lagrangian functional method for finite–dimensional optimization problems with equality constraints, is suitable also for problems described by ordinary differential equations or integral equations [74], as well as for shape optimization problems [94, 95, 125, 170]. Moreover, the Lagrangian setting allows to provide in a straightforward manner a posteriori error estimates for approximated optimal control problems as, e.g., in [19, 20, 40, 44, 75, 76, 119, 182] and as we discuss in Chapter 7. For these reasons we make use and discuss the optimal control theory based on the *Lagrangian* formalism. For a further deepening on the optimal control theory, we refer the reader to [2, 6, 10, 50, 107, 108, 144].

In this work, for the sake of simplicity, we indicate with *optimal control* a problem which is governed by both steady or unsteady PDEs. However, we notice that in literature a control problem described by a system of steady PDEs is usually indicated as an “optimization” problem, while, if described by unsteady PDEs, it is referred as an “optimal control” problem.

The optimal control problems are solved numerically by means of optimization techniques, such as the Steepest Descent, Conjugate Gradient, quasi–Newton and Sequential Quadratic Programming (SQP) methods [64, 130], which are applied to the approximated PDEs. Typically, the Galerkin–Finite Element method is used for the approximation of the PDEs (see e.g. [153]), even if other methods are possible, such as the Spectral (see e.g. [59]) or the Reduced

Basis method which we consider in Chapters 6 and 7 (see also e.g. [80]). We briefly describe the most widely used optimization techniques for both the unconstrained and the constrained cases; a numerical test case, referred to an unconstrained optimal control problem described by an elliptic PDE, is provided.

The Chapter is organized as follows. In Sec.2.1 we deal with the optimal control theory in the unconstrained case both for steady and unsteady PDEs; in similar way in Sec.2.2, we address the optimal control theory in the constrained case. Finally, in Sec.2.3 we briefly recall the optimization techniques together with an attempt of comparison for a specific test case. In Chapter 3 we report an example of an optimal flow control for a problem described by steady Navier–Stokes equations.

2.1 The Lagrangian formalism for unconstrained optimal control problems

In this Section we discuss the case of *unconstrained optimal control problems* by means of the Lagrangian formalism [20, 74, 75, 116, 118, 174]. Firstly, we introduce the optimal control problem in an abstract setting, then we distinguish the analysis in the cases of steady and unsteady PDEs, for which we refer to elliptic and parabolic PDEs, respectively.

Remark 2.1. *For the sake of simplicity, we present the Lagrangian formalism for scalar PDEs depending on scalar variables. However, the analysis can be straightforwardly extended to the vectorial PDEs with vectorial variables and to mixed problems, as we do in Chapter 3 and more specifically in Sec.3.1.*

2.1.1 The general case

The optimal control problem in the unconstrained case reads, in an abstract setting [74], as:

$$\text{find } \ u \in \mathcal{U}, \quad u = \ \text{argmin} \ J(v,u), \ \text{ where } v \in \mathcal{V} \ \text{ is solution of } \ r(v,u) = 0 \ \text{ in } \mathcal{W}, \tag{2.1}$$

being $J(v,u)$ the *cost functional*, while $r(v,u) = 0$ indicates a well–posed PDE (endowed with appropriate boundary and initial conditions), which is called the *primal equation* (often *state equation*); consequently, the variable $v \in \mathcal{V}$ is called the *primal variable*, while $u \in \mathcal{U}$ is referred as the *control variable*. The spaces $\mathcal{V}$, $\mathcal{U}$ and $\mathcal{W}$ are, in general, Banach spaces. The *Lagrangian functional* is defined on the basis of Eq.(2.1), as[1]:

$$\mathcal{L}(\mathbf{x}) = \mathcal{L}(v,z,u) := J(v,u) + \langle z, r(v,u)\rangle_{\mathcal{W}^*,\mathcal{W}}, \tag{2.2}$$

where $\mathcal{W}^*$ is the dual space of $\mathcal{W}$ and $\langle\cdot,\cdot\rangle_{\mathcal{W}^*,\mathcal{W}}$ indicates the corresponding crochet; the Lagrangian multiplier $z \in \mathcal{W}^*$ is called the *dual variable* (or adjoint variable). With $\mathbf{x} \in \mathcal{X}$, we indicate the variables (v,z,u), s.t. $\mathbf{x} := (v,z,u)$, where $\mathcal{X} := \mathcal{V} \times \mathcal{W}^* \times \mathcal{U}$. It follows that the Lagrangian functional is defined as $\mathcal{L} : \mathcal{X} \to \mathbb{R}$.
For the analysis of the optimal control problem, we recall the following definitions ([74]).

Definition 2.1. *By indicating with* $\mathbf{y} := (v,u) \in \mathcal{V} \times \mathcal{U}$ *and* $r(\mathbf{y}) = r(v,u)$, *we define the* feasible space *as* $\mathcal{Y} := \{\mathbf{y} \in \mathcal{V} \times \mathcal{U} \ : \ r(\mathbf{y}) = 0\}$.

[1] Sometimes the Lagrangian functional is defined as: $\mathcal{L}(v,z,u) := J(v,u) - \langle z, r(v,u)\rangle_{\mathcal{W}^*,\mathcal{W}}$.

Definition 2.2. *A feasible point* $\mathbf{y}^{**} := (v^{**}, u^{**}) \in \mathcal{Y}$ *is a* local optimal solution *if there exists* $\delta > 0$ *s.t.* $J(\mathbf{y}^{**}) \leq J(\mathbf{y})$ $\forall \mathbf{y} \in \mathcal{B}_\delta(\mathbf{y}^{**})$*, being* $\mathcal{B}_\delta(\mathbf{y}^{**}) := \{\mathbf{y} \in \mathcal{Y} \ : \ \|\mathbf{y} - \mathbf{y}^{**}\|_{\mathcal{Y}} < \delta\}$ *with* $\| \cdot \|_{\mathcal{Y}}$ *any norm of the product space* $\mathcal{Y}$*.*

Definition 2.3. *A local optimal solution* $\mathbf{y}^{**} \in \mathcal{Y}$ *is a* global optimal solution *if* $J(\mathbf{y}^{**}) \leq J(\mathbf{y})$ $\forall \mathbf{y} \in \mathcal{Y}$*.*

Definition 2.4. *Let* $J(\mathbf{y})$ *and* $r(\mathbf{y})$ *be continuously Fréchet differentiable*[2] *(see [99]) in* $\mathcal{B}_\delta(\mathbf{y}^\star)$*, being* $\mathbf{y}^\star := (v^\star, u^\star) \in \mathcal{Y}$ *a critical solution, i.e. satisfying the first order necessary conditions (2.4)–(2.6), and* $\mathcal{C}(\mathbf{y}^\star) := \mathcal{Y} \times \{\rho(u - u^\star) \ : \ u \in \mathcal{U}, \ \rho \geq 0\}$ *a convex cone*[3]*. We say that* $\mathbf{y}^\star \in \mathcal{Y}$ *is a* regular point *if* $\mathcal{W} \equiv \{r_{\mathbf{y}}(\mathbf{y}^\star)[\delta \mathbf{y}] \ : \ \delta \mathbf{y} \in \mathcal{C}(\mathbf{y}^\star)\}$*.*

For the first order necessary conditions for the local optimal solution of problem (2.1), the following Theorem holds; for the proof see [116, 174].

Theorem 2.1. *If* $\mathbf{y}^{**} = (v^{**}, u^{**}) \in \mathcal{Y}$ *is a regular point and a local optimal solution of problem (2.1), there exists a Lagrange multiplier* $z^{**} \in \mathcal{Z}^*$ *s.t.:*

$$\mathcal{L}_v(\mathbf{y}^{**}, z^{**}) = J_v(\mathbf{y}^{**}) + r_v^*(\mathbf{y}^{**})[z^{**}] = 0 \qquad \text{in } \mathcal{V}^*, \tag{2.4}$$

$$\mathcal{L}_u(\mathbf{y}^{**}, z^{**})[\psi] = \langle J_u(\mathbf{y}^{**}) + r_u^*(\mathbf{y}^{**})[z^{**}], \psi \rangle_{\mathcal{U}^*, \mathcal{U}} = 0 \qquad \forall \psi \in \mathcal{U}, \tag{2.5}$$

$$\mathcal{L}_z(\mathbf{y}^{**}, z^{**}) = r(\mathbf{y}^{**}) = 0 \qquad \text{in } \mathcal{W}, \tag{2.6}$$

where the differentiation of the Lagrangian functional (2.2) is in Fréchet sense[2]*. The operators* $r_v^*(\mathbf{y}^{**})$ *and* $r_u^*(\mathbf{y}^{**})$ *are the dual Fréchet derivatives of* $r(\mathbf{y})$ *in* $\mathbf{y}^{**}$ *w.r.t.* $v \in \mathcal{V}$ *and* $u \in \mathcal{U}$*, respectively; similarly,* $J_v(\mathbf{y}^{**})$ *and* $J_u(\mathbf{y}^{**})$ *are the Fréchet derivatives of* $J(\mathbf{y})$ *in* $\mathbf{y}^{**}$ *w.r.t.* $v \in \mathcal{V}$ *and* $u \in \mathcal{U}$*, respectively.*

The first order necessary conditions (2.4)–(2.6) are usually called the *Karush–Kuhn–Tucker conditions* (*KKT conditions*), while the solutions $\mathbf{x}^{**} \in \mathcal{X}$ satisfying the KKT conditions are indicated as *critical solutions*. Eq.(2.4) is referred as the *dual equation*, while Eq.(2.5) is the *optimality condition*; by defining the *sensitivity* $\delta u \in \mathcal{U}^*$ as:

$$\delta u = \delta u(\mathbf{y}, z) := J_u(\mathbf{y}) + r_u^*(\mathbf{y})[z], \tag{2.7}$$

the optimality condition (2.5) corresponds to $\langle \delta \mathbf{u}^{**}, \psi \rangle_{\mathcal{U}^*, \mathcal{U}} = 0$ $\forall \psi \in \mathcal{U}$. We observe that Eq.(2.6) is once again the primal equation defined in Eq.(2.1). Theorem 2.1 can be conveniently used in order to find a local optimal solution by means of the KKT conditions; however, a critical solution of the first order necessary conditions is not necessarily a local optimal solution. With this aim, we recall a second order sufficient optimality condition, whose proof is reported in [116, 118].

[2] Let be X and Y two spaces endowed with norm and $F(x) : x \in E \to Y$ an application defined on the open space $E \subset X$; we say that $F(x)$ is *Fréchet differentiable* in $\overline{x} \in E$ if there exists a linear and continuous operator $F_x(\overline{x}) \in \mathcal{L}(X, Y)$ s.t.:

$$\forall \varepsilon > 0, \ \exists \delta > 0 \quad : \quad \|F(\overline{x} + h) - F(\overline{x}) - F_x(\overline{x})[h]\|_Y \leq \varepsilon \|h\|_X \quad \forall h \in X \text{ s.t. } \|h\|_X < \delta. \tag{2.3}$$

The expression $F_x(\overline{x})[h]$, to which corresponds an element in Y for each $h \in X$, is said *Fréchet differential*, while the operator $F_x(\overline{x})$ is identified as *Fréchet derivative* of the application $F(x)$ in $\overline{x} \in E$. For a further deepening, we refer the reader to [99].

[3] A set $C \subset X$ is said convex if, $\forall x, y \in C$ and $\forall \alpha \in [0, 1] \subset \mathbb{R}$ s.t. $z := \alpha x + (1 - \alpha) y$, we have $z \in C$.

Theorem 2.2. *Let us suppose that $J(\mathbf{y})$ and $r(\mathbf{y})$ are twice Fréchet differentiable and let $\mathbf{x}^{**} = (\mathbf{y}^{**}, z^{**}) \in \mathcal{X}$ satisfy the first order necessary conditions (2.4)–(2.6). If there exists $\theta > 0$ s.t.*

$$\mathcal{L}_{\mathbf{yy}}(\mathbf{x}^{**})[\mathbf{y}, \mathbf{y}] \geq \theta \|\mathbf{y}\|^2_{\mathcal{Y}} \tag{2.8}$$

holds $\forall \mathbf{y} \in \mathcal{Y}$ which satisfies:

$$r_{\mathbf{y}}(\mathbf{y}^{**})[\mathbf{y}] = 0 \qquad \text{in } \mathcal{W}, \tag{2.9}$$

then $\mathbf{y}^{**} = (v^{**}, u^{**})$ is a (strict) local optimal solution. In Eq.(2.8), we indicate with $\mathcal{L}_{\mathbf{yy}}(\mathbf{x}^{**})$ the second Fréchet derivative of the Lagrangian functional w.r.t. $\mathbf{y}$, which corresponds to the Hessian of $\mathcal{L}(\mathbf{x}^{**})$ w.r.t. $\mathbf{y}$; similarly, in Eq.(2.9) $r_{\mathbf{y}}(\mathbf{y}^{**})$ corresponds to the Jacobian of the primal equation w.r.t. $\mathbf{y}$.

The use of Theorems 2.1 and 2.2 allows to find local optimal solutions of the problem (2.1); however, it is not always simple to establish if a local optimal solution is also a global one. This depends, in general, on the particular optimal control problem under consideration, for which the existence and uniqueness of a global optimal solution should be determined on the basis of its properties.

Remark 2.2. *Throughout this work we indicate with the double apex $**$ the optimal solution (local or global). In this Section and in Sec.2.2 the optimal solution $\mathbf{y}^{**} \in \mathcal{Y}$ is composed by the optimal primal solution $v^{**} \in \mathcal{V}$ and the optimal control $u^{**} \in \mathcal{U}$; however, in view of Chapters 4 and 7 it is useful to indicate the optimal solution as $\mathbf{x}^{**} := (v^{**}, z^{**}, u^{**}) \in \mathcal{X}$, taking into account also for the dual variable $z^{**} \in \mathcal{W}^*$.*

2.1.2 The case of steady PDEs

In this Section we apply the theory introduced in Sec.2.1.1 to the case of optimal control problems described by steady PDEs (optimization problems). In particular, we consider an elliptic PDE with quadratic cost functional. In Chapter 3 we discuss the case of an optimal flow control problem governed by steady Navier–Stokes equations.

By referring to the general optimal control problem (2.1), we introduce the following quadratic cost functional:

$$J(v, u) = \frac{1}{2} m_d(v - v_d, v - v_d) + \frac{1}{2}\gamma \; m_u(u - u_d, u - u_d), \tag{2.10}$$

where $\gamma > 0$, $m_d(\cdot, \cdot)$ and $m_u(\cdot, \cdot)$ are positive and symmetric bilinear forms, v_d is the desired solution, while u_d the desired control; the cost functional $J(v, u)$ is twice differentiable in $v \in \mathcal{V}$ and $u \in \mathcal{U}$. We notice that, for the sake of simplicity, we neglect to indicate the space dependence of the variables explicitly. Similarly, we introduce from Eq.(2.1) the primal equation in weak form:

$$\text{find } v \in \mathcal{V} \quad : \quad a(v, u)(\phi) = F(\phi) \qquad \forall \phi \in \mathcal{V}, \text{ with } u \in \mathcal{U}, \tag{2.11}$$

where $a(\cdot, \cdot)(\cdot)$ is a twice differentiable semilinear form (with linearity in the last argument) from $\mathcal{X} := \mathcal{V} \times \mathcal{V} \times \mathcal{U}$ and $F(\cdot)$ is a continuous and linear functional. The space $\mathcal{V}$ is an Hilbert space taking into account for any Dirichlet homogeneous boundary conditions, say e.g. $H^1(\Omega)$ or $H^1_0(\Omega)$, where Ω is the computational domain s.t. $\Omega \subset \mathbb{R}^n$, $n \geq 1$, with boundary $\partial\Omega$. The space of controls $\mathcal{U}$ is a Banach space, e.g. $\mathcal{U} = L^2(\omega)$, where ω is the

domain, subdomain or boundary where the controls are defined. Moreover, we assume that the primal equation (2.11) is well posed; we notice that in this case the space $\mathcal{W}^* \equiv \mathcal{V}$.
From Eq.(2.2) we define the Lagrangian functional for this particular case, which reads:

$$\begin{aligned} \mathcal{L}(\mathbf{x}) = \mathcal{L}(v,z,u) \quad = \quad & \frac{1}{2} m_d(v-v_d, v-v_d) + \frac{1}{2}\gamma\ m_u(u-u_d, u-u_d) \\ & +F(z) - a(v,u)(z). \end{aligned} \tag{2.12}$$

From the first order necessary conditions (2.4) and (2.5), we obtain respectively the dual equation:

$$\text{find } v \in \mathcal{V} \quad : \quad a_v(v,u)(z,\vartheta) = m_d(v-v_d,\vartheta) \qquad \forall \vartheta \in \mathcal{V} \tag{2.13}$$

and the optimality equation:

$$\gamma\ m_u(u^{**}-u_d,\psi) - a_u(v^{**},u^{**})(z^{**},\psi) = 0 \qquad \forall \psi \in \mathcal{U}, \tag{2.14}$$

with $\mathbf{x}^{**} = (v^{**}, z^{**}, u^{**})$ indicating the critical solution; we observe that the third condition (2.6) corresponds to the primal equation (2.11). Notice that $a_v(v,u)(z,\vartheta)$ represents the differential of the form $a(v,u)(z)$ w.r.t. v evaluated in $\vartheta \in \mathcal{V}$; similarly, $a_u(v,u)(z,\psi)$ represents the differential of the form $a(v,u)(z)$ w.r.t. u evaluated in $\psi \in \mathcal{U}$. In the same manner, it is possible to deduce a second order sufficient conditions from Eq.(2.8)[4].

Remark 2.3. *Let us suppose that $a(v,u)(z)$ can be written as $a(v,u)(z) = A(v,z) - B(u,z)$ with $A(\cdot,\cdot)$ and $B(\cdot,\cdot)$ continuous and bilinear forms, with $A(\cdot,\cdot)$ coercive. In this case an unique global optimal solution exists; see for instance [107]. Moreover, if $\gamma = 0$ in Eq.(2.10), the optimal control problem is not, in general, well posed, in the sense that the uniqueness of an optimal solution is not guaranteed [2].*

2.1.3 The case of unsteady PDEs

In this Section we consider the case of an optimal control problem for a time–dependent PDE. In particular, starting from the general case of Sec.2.1.1, we discuss the case of a parabolic PDE.

We start by introducing the time interval $[0,T]$; in this case the spaces $\mathcal{V}$ and $\mathcal{U}$ are, in general, Banach spaces over the space–time domain $\Omega \times (0,T)$. By adopting the notation introduced in Sec.2.1.2 for the steady case, we introduce the following cost functionals:

$$J(v,u) = \frac{1}{2}\int_0^T m_d(v-v_d, v-v_d)\ dt + \frac{1}{2}\gamma \int_0^T m_u(u-u_d,u-u_d)\ dt, \tag{2.15}$$

or:

$$J(v,u) = \frac{1}{2} m_d(v(T)-v_d, v(T)-v_d) + \frac{1}{2}\gamma \int_0^T m_u(u-u_d,u-u_d)\ dt, \tag{2.16}$$

where, for the sake of simplicity, the time and space dependence of the variables are understood. Starting from Eq.(2.1), we introduce the following primal equation:

$$\begin{aligned} & \text{find } v \in \mathcal{V} \quad : \quad m\left(\frac{\partial v}{\partial t}, \phi\right) + a(v,u)(\phi) = F(\phi) \qquad \forall \phi \in \mathcal{V},\ \text{with } u \in \mathcal{U},\ t \in (0,T), \\ & \text{with } v(0) = v_0, \end{aligned} \tag{2.17}$$

[4] For the computation of the Hessian (2.8) of the Lagrangian functional w.r.t. v and u is convenient to rewrite Eq.(2.12) as $\mathcal{L}(v,z,u) = J(v,u) + a(v,u)(z) - F(z)$.

where $m(\cdot,\cdot)$ is a time–independent symmetric and continuous bilinear form, while v_0 is the initial condition. As usual, essential boundary conditions are taken into account in the choice of the space $\mathcal{V}$. A typical instance is $\mathcal{V} = \{w \in L^2(0,T;H_0^1(\Omega)) \ : \ \frac{\partial w}{\partial t} \in L^2(0,T;H_0^1(\Omega)^*)\}$; similarly, the space of control variables could be $\mathcal{U} = L^2(0,T;L^2(\omega))$. Once again, we suppose that the parabolic PDE (2.17) is well posed and we observe that $\mathcal{W}^* \equiv \mathcal{V}$.
We observe that the optimal control $u \in \mathcal{U}$ is often indicated as "regulator" if the cost functional (2.15) is used, as "controller" if (2.16) is considered instead.

Let us start with the optimal control problem associated with the cost functional (2.15) (i.e. the "regulator" problem). In this case, the Lagrangian functional (2.2) reads from Eq.s (2.15) and (2.17):

$$\begin{aligned}\mathcal{L}(\mathbf{x}) = \mathcal{L}(v,z,u) \quad = \quad & \frac{1}{2}\int_0^T m_d(v - v_d, v - v_d)\ dt + \frac{1}{2}\gamma\int_0^T m_u(u - u_d, u - u_d)\ dt \\ & + \int_0^T F(z)\ dt - \int_0^T a(v,u)(z)\ dt - \int_0^T m\left(\frac{\partial v}{\partial t}, z\right)\ dt \\ & - m\left(v(0) - v_0, z(0)\right).\end{aligned} \tag{2.18}$$

In order to obtain the first order necessary conditions (2.4)–(2.6) for the local optimal solution, we differentiate $\mathcal{L}(\mathbf{x})$ w.r.t. v, z and u. In particular, by differentiating $\mathcal{L}(\mathbf{x})$ w.r.t. v, we obtain:

$$\begin{aligned}\mathcal{L}_v(\mathbf{x})[\vartheta] \quad = \quad & \int_0^T m_d(v - v_d, \vartheta)\ dt \\ & - \int_0^T a_v(v,u)(z,\vartheta)\ dt - \int_0^T m\left(\frac{\partial \vartheta}{\partial t}, z\right)\ dt \\ & - m\left(\vartheta(0), z(0)\right);\end{aligned} \tag{2.19}$$

by observing that $m\left(\frac{\partial v}{\partial t}, \vartheta\right) = m\left(\frac{\partial v}{\partial t}\vartheta, 1\right)$ and by integrating by parts, Eq.(2.19) reads:

$$\begin{aligned}\mathcal{L}_v(\mathbf{x})[\vartheta] \quad = \quad & \int_0^T m_d(v - v_d, \vartheta)\ dt \\ & - \int_0^T a_v(v,u)(z,\vartheta)\ dt + \int_0^T m\left(\vartheta, \frac{\partial z}{\partial t}\right)\ dt \\ & - m\left(\vartheta(T), z(T)\right).\end{aligned} \tag{2.20}$$

By using Eq.(2.20) in the first order necessary condition (2.4), we obtain the dual equation in weak form:

$$\begin{aligned}&\text{find } z \in \mathcal{V} \quad : \quad -m\left(\vartheta, \frac{\partial z}{\partial t}\right) + a_v(v,u)(z,\vartheta) = m_d(v - v_d, \vartheta) \qquad \forall \vartheta \in \mathcal{V},\ t \in (0,T), \\ &\text{with } z(T) = 0.\end{aligned} \tag{2.21}$$

In the same manner, by differentiating $\mathcal{L}(\mathbf{x})$ w.r.t. u, we obtain:

$$\begin{aligned}\mathcal{L}_u(\mathbf{x})[\psi] \quad = \quad & \gamma\int_0^T m_u(u - u_d, \psi)\ dt \\ & - \int_0^T a_u(v,u)(z,\psi)\ dt,\end{aligned} \tag{2.22}$$

from which we deduce the optimality condition (2.5):

$$\gamma\ m_u(u^{**} - u_d, \psi) - a_u(v^{**}, u^{**})(z^{**}, \psi) = 0 \qquad \forall \psi \in \mathcal{U},\ t \in (0,T). \tag{2.23}$$

Let us now consider the optimal control problem associated with the cost functional (2.16) (i.e. the "controller" problem), for which the Lagrangian functional (2.2) reads from Eq.s (2.16) and (2.17):

$$\begin{aligned} \mathcal{L}(\mathbf{x}) = \mathcal{L}(v,z,u) \ &= \ \frac{1}{2} m_d(v(T) - v_d, v(T) - v_d) + \frac{1}{2}\gamma \int_0^T m_u(u - u_d, u - u_d)\ dt \\ &\quad + \int_0^T F(z)\ dt - \int_0^T a(v,u)(z)\ dt - \int_0^T m\left(\frac{\partial v}{\partial t}, z\right)\ dt \\ &\quad - m\,(v(0) - v_0, z(0)). \end{aligned} \tag{2.24}$$

By differentiating $\mathcal{L}(\mathbf{x})$ w.r.t. v, we obtain:

$$\begin{aligned} \mathcal{L}_v(\mathbf{x})[\vartheta] \ &= \ m_d(v(T) - v_d, \vartheta) \\ &\quad - \int_0^T a_v(v,u)(z,\vartheta)\ dt - \int_0^T m\left(\frac{\partial \vartheta}{\partial t}, z\right)\ dt \\ &\quad - m\,(\vartheta(0), z(0)) \end{aligned} \tag{2.25}$$

and finally, by integrating by parts:

$$\begin{aligned} \mathcal{L}_v(\mathbf{x})[\vartheta] \ &= \ m_d(v(T) - v_d, \vartheta) \\ &\quad - \int_0^T a_v(v,u)(z,\vartheta)\ dt + \int_0^T m\left(\vartheta, \frac{\partial z}{\partial t}\right)\ dt \\ &\quad - m\,(\vartheta(T), z(T)). \end{aligned} \tag{2.26}$$

By using Eq.(2.26) in the first order necessary condition (2.4), we obtain the dual equation in weak form, which reads:

$$\begin{aligned} &\text{find } z \in \mathcal{V} \ \ : \ \ -m\left(\vartheta, \frac{\partial z}{\partial t}\right) + a_v(v,u)(z,\vartheta) = 0 \qquad \forall \vartheta \in \mathcal{V},\ t \in (0,T), \\ &\text{with } m\,(z(T), \vartheta(T)) = m_d\,(v(T) - v_d, \vartheta(T))\,, \end{aligned} \tag{2.27}$$

where the final condition is given in weak form. We observe that the optimality condition for this control problem is the same of Eq.(2.23).

The second order sufficient conditions of Theorem 2.2 are easily deducible for the optimal control problem endowed with both the cost functionals (2.15) and (2.16).

Remark 2.4. *The parabolic dual PDEs (2.21) and (2.27) evolve "backward" with the "initial" conditions given at the final time $t = T$. From a numerical point of view, this could lead to large computational costs. In fact, we need to solve the primal equation till the final time $t = T$ before solving the dual equation at each time step $t \in (0,T)$; this is due to the dependence of the dual problem on $v(t)$, $t \in (0,T)$ or $v(T)$. Decoupling techniques could be used in order to reduce the "complexity" of the optimal control problem in the case of linear parabolic PDEs.*

Remark 2.5. *Let us suppose, as in Remark 2.3, that $a(v,u)(z)$ could be written as $a(v,u)(z) = A(v,z) - B(u,z)$ with $A(\cdot,\cdot)$ and $B(\cdot,\cdot)$ continuous and bilinear forms and let $A(\cdot,\cdot)$ be coercive. Then, an unique global optimal solution exists [107]; similarly, if $\gamma = 0$ in Eq.(2.15) or Eq.(2.16), the optimal control problem is not, in general, well posed [2].*

2.2 The Lagrangian formalism for constrained optimal control problems

In this Section we discuss the case of *constrained optimal control problems* by using the Lagrangian functional approach introduced in Sec.2.1; see [74, 116, 118, 174]. First of all, we present the theory in an abstract setting by recalling the formalism of Sec.2.1; then, we consider the cases of steady and unsteady PDEs.

2.2.1 The general case

By adopting the notation of Sec.2.1, the optimal control problem in the control constrained case reads:

$$\text{find } \; u \in \mathcal{U}_{ad}, \;\; u = \; \text{argmin } J(v,u), \; \text{ where } v \in \mathcal{V} \; \text{ is solution of } \; r(v,u) = 0 \; \text{ in } \mathcal{W}; \tag{2.28}$$

the space $\mathcal{U}_{ad}$, which is indicated as the space of *admissible controls*, takes into account for the constraints on the control u. In particular, $\mathcal{U}_{ad}$ is a non–empty convex closed subset of $\mathcal{U}$, s.t. $\mathcal{U}_{ad} \subset \mathcal{U}$. We observe that the Definitions 2.1–2.4 hold also for the constrained case, where $\mathcal{U}$ is replaced by $\mathcal{U}_{ad}$. Moreover, the Lagrangian functional assumes the same expression as in Eq.(2.2), even if the functional is defined on the space $\mathcal{X}_{ad} := \mathcal{V} \times \mathcal{W}^* \times \mathcal{U}_{ad}$, i.e. $\mathcal{L} : \mathcal{X}_{ad} \to \mathbb{R}$. For the local optimal solutions of the problem (2.28), the following Theorem holds in analogy with Theorem 2.1 (for the proof see [116, 174]).

Theorem 2.3. *If $\mathbf{y}^{**} = (v^{**}, u^{**}) \in \mathcal{Y}_{ad}$, with $\mathcal{Y}_{ad} := \mathcal{V} \times \mathcal{U}_{ad}$ being the feasible space, is a regular point and a local optimal solution of problem (2.28), there exists a Lagrange multiplier $z^{**} \in \mathcal{Z}^*$ s.t.:*

$$\mathcal{L}_v(\mathbf{y}^{**}, z^{**}) = J_v(\mathbf{y}^{**}) + r_v^*(\mathbf{y}^{**})[z^{**}] = 0 \qquad \textit{in } \mathcal{V}^*, \tag{2.29}$$

$$\mathcal{L}_u(\mathbf{y}^{**}, z^{**})[\psi] = \langle J_u(\mathbf{y}^{**}) + r_u^*(\mathbf{y}^{**})[z^{**}], \psi - u^{**} \rangle_{\mathcal{U}^*,\mathcal{U}} \geq 0 \qquad \forall \psi \in \mathcal{U}_{ad}, \tag{2.30}$$

$$\mathcal{L}_z(\mathbf{y}^{**}, z^{**}) = r(\mathbf{y}^{**}) = 0 \qquad \textit{in } \mathcal{W}. \tag{2.31}$$

Conditions (2.29)–(2.31) are usually indicated as KKT conditions as in the unconstrained case. We observe that, in the constrained case, the optimality condition (2.30) is set in an inequality form and it replaces the equality (2.5) of the unconstrained case. However, the numerical treatment of the first order necessary conditions is not straightforward, due to the inequality expressing the optimality condition. In order to overcome this difficulty, a modified Lagrangian functional is introduced, say $\widetilde{\mathcal{L}}(\mathbf{x}, \mu_1, \ldots, \mu_n)$, taking into account for additional Lagrangian multipliers $\mu_1, \ldots, \mu_n$, $n \geq 1$, one for each inequality constraints delimitating the space $\mathcal{U}_{ad}$, which need to be specified . For the sake of simplicity, we prefer to distinguish between the steady and unsteady cases, which we afford in the Sec.s 2.2.2 and 2.2.3, instead of providing a general, but more involved, analysis.
We observe that for the second order sufficient conditions, Theorem 2.2 holds simply by replacing $\mathcal{U}$ with $\mathcal{U}_{ad}$.

2.2.2 The case of steady PDEs

In this Section we consider the particular case of a constrained optimal control problem described by steady PDEs.

First of all, as anticipated in Sec.2.2.1, we define the space of admissible controls $\mathcal{U}_{ad}$, for which we choose:

$$\mathcal{U}_{ad} = \{u \in \mathcal{U} \ : \ u_{min} \leq u \leq u_{max} \text{ a.e. in } \omega\}\,, \tag{2.32}$$

where ω is defined in Sec.2.1.2.
On this basis, for the general problem (2.28), we introduce the modified Lagrangian $\widetilde{\mathcal{L}}(\mathbf{x}, \mu_1, \mu_2)$ as:

$$\begin{aligned} \widetilde{\mathcal{L}}(\mathbf{x}, \mu_1, \mu_2) = \widetilde{\mathcal{L}}(y, z, u, \mu_1, \mu_2) \quad &:= \quad J(v, u) + \langle z, r(v, u)\rangle_{\mathcal{W}^*,\mathcal{W}} \\ &\qquad + \langle \mu_1, u_{min} - u\rangle_{\mathcal{U}^*,\mathcal{U}} + \langle \mu_2, u - u_{max}\rangle_{\mathcal{U}^*,\mathcal{U}}, \end{aligned} \tag{2.33}$$

where the Lagrangian multipliers $\mu_1, \mu_2 \in \mathcal{U}^*$. It follows that Theorem 2.3 can be reformulated as follows [74]; for the proof we refer the reader to [116, 174].

Theorem 2.4. *If $\mathbf{y}^{**} = (v^{**}, u^{**}) \in \mathcal{Y}_{ad}$ is a regular point and a local optimal solution of problem (2.28) with the modified Lagrangian functional (2.33), there exists a Lagrange multiplier $z^{**} \in \mathcal{Z}^*$ s.t.:*

$$\widetilde{\mathcal{L}}_v(\mathbf{y}^{**}, z^{**}, \mu_1, \mu_2) = J_v(\mathbf{y}^{**}) + r_v^*(\mathbf{y}^{**})[z^{**}] = 0 \qquad \text{in } \mathcal{V}^*, \tag{2.34}$$

$$\widetilde{\mathcal{L}}_u(\mathbf{y}^{**}, z^{**}, \mu_1, \mu_2)[\psi] = \langle J_u(\mathbf{y}^{**}) + r_u^*(\mathbf{y}^{**})[z^{**}] + \mu_2 - \mu_1, \psi\rangle_{\mathcal{U}^*,\mathcal{U}} = 0 \qquad \forall \psi \in \mathcal{U}, \tag{2.35}$$

$$\widetilde{\mathcal{L}}_z(\mathbf{y}^{**}, z^{**}) = r(\mathbf{y}^{**}) = 0 \qquad \text{in } \mathcal{W}, \tag{2.36}$$

together with the feasibility conditions*:*

$$u - u_{min} \geq 0 \qquad \text{and} \qquad u_{max} - u \geq 0 \qquad \text{a.e. in } \omega, \tag{2.37}$$

the multiplier sign conditions*:*

$$\langle \mu_1, \psi\rangle_{\mathcal{U}^*,\mathcal{U}} \geq 0 \qquad \text{and} \qquad \langle \mu_2, \psi\rangle_{\mathcal{U}^*,\mathcal{U}} \geq 0 \qquad \forall \psi \in \mathcal{U} \tag{2.38}$$

and the complimentary conditions*:*

$$\langle \mu_1, u - u_{min}\rangle_{\mathcal{U}^*,\mathcal{U}} = 0 \qquad \text{and} \qquad \langle \mu_2, u_{max} - u\rangle_{\mathcal{U}^*,\mathcal{U}} = 0. \tag{2.39}$$

If we consider the optimal control problem described by an elliptic PDE introduced in Sec.2.2.2, it is easy to deduce the modified Lagrangian functional (2.33) and the first order necessary conditions (2.34)–(2.36), together with the feasibility, sign and complimentary conditions (2.37)–(2.39).

Remark 2.6. *The considerations of Remark 2.3 hold also in the constrained case, for which it is sufficient to replace $\mathcal{U}$ with $\mathcal{U}_{ad}$.*

2.2.3 The case of unsteady PDEs

In this Section we discuss the case of a constrained optimal control problem descibed by unsteady PDEs.

As done for the steady case of Sec.2.2.2, we start by defining the space of admissible controls $\mathcal{U}_{ad}$. With this aim, we choose the particular case of a control u independent on the space ω, s.t. $u = u(t)$, and we assume $\mathcal{U} = L^2(0, T)$. We remark that in this case the control is a scalar function in time; however, it is straightforward to consider the more general case for which $u = u(x,t)$. We define $\mathcal{U}_{ad}$ as:

$$\mathcal{U}_{ad} = \left\{ u \in \mathcal{U} \ : \ u_{min} \leq u \leq u_{max} \text{ in } (0,T) \text{ and } \int_0^T u \ dt \leq T\overline{U} \right\}, \tag{2.40}$$

where an integral time inequality and a bound $\overline{U} \in \mathbb{R}$ have been introduced. Theorem 2.4 holds, even if a third Lagrangian multiplier, say $\mu_{\overline{U}}$ should be introduced. We observe that in this case $\mu_1, \mu_2 \in \mathcal{U}^*$, while $\mu_{\overline{U}} \in \mathbb{R}$. In fact, the modified Lagrangian functional (2.33) reads:

$$\begin{aligned} \widetilde{\mathcal{L}}(y,z,u,\mu_1,\mu_2,\mu_{\overline{U}}) \quad := \quad & J(v,u) + \langle z, r(v,u) \rangle_{\mathcal{W}^*,\mathcal{W}} \\ & + \int_0^T \mu_1(u_{min} - u) \ dt + \int_0^T \mu_2(u - u_{max}) \ dt \\ & + \mu_{\overline{U}} \left(\int_0^T u \ dt - T\overline{U} \right). \end{aligned} \tag{2.41}$$

It follows that the optimality condition (2.35) could be expressed as:

$$\begin{aligned} \widetilde{\mathcal{L}}_u(\mathbf{y}^{**}, z^{**}, \mu_1, \mu_2, \mu_{\overline{U}})[\psi] \quad = \quad & \langle J_u(\mathbf{y}^{**}) + r_u^*(\mathbf{y}^{**})[z^{**}], \psi \rangle_{\mathcal{U}^*,\mathcal{U}} \\ & + \int_0^T (\mu_2 - \mu_1)\psi \ dt + \mu_{\overline{U}} \int_0^T \psi \ dt = 0 \qquad \forall \psi \in \mathcal{U}, \end{aligned} \tag{2.42}$$

while the feasibility conditions (2.37) read:

$$u - u_{min} \geq 0, \qquad u_{max} - u \geq 0 \qquad \text{in } (0,T) \qquad \text{and} \qquad T\overline{U} - \int_0^T u \ dt \geq 0; \tag{2.43}$$

similarly, the multiplier sign conditions (2.38) follow:

$$\int_0^T \mu_1 \psi \ dt \geq 0, \qquad \int_0^T \mu_2 \psi \ dt \geq 0 \qquad \text{and} \qquad \mu_{\overline{U}} \int_0^T \psi \ dt \geq 0 \qquad \forall \psi \in \mathcal{U}, \tag{2.44}$$

together with the complementary conditions (2.39):

$$\int_0^T \mu_1(u - u_{min}) \ dt = 0, \qquad \int_0^T \mu_2(u_{max} - u) \ dt = 0 \qquad \text{and} \qquad \mu_{\overline{U}} \left(T\overline{U} - \int_0^T u \ dt \right) = 0. \tag{2.45}$$

Remark 2.7. *The considerations of Remark 2.5 are valid also in this case; the space $\mathcal{U}$ must simply be replaced by $\mathcal{U}_{ad}$.*

2.3 Optimization methods

In this Section we recall the most common optimization methods together with their application to optimal control problems described by PDEs. A numerical test case, referring to an unconstrained optimal control problem, is discussed.

2.3.1 A review of the optimization methods

In this Section we briefly discuss, in an abstract setting, the most common optimization techniques for both the unconstrained and constrained cases, like the Steepest–Descent, Conjugate Gradient, Newton, quasi–Newton and Sequential Quadratic Programming (SQP) methods. With this aim we refer principally to [64, 130].

Let us consider the following finite–dimensional optimization problem:

$$\text{find } \mathbf{u} \in U_{ad}, \quad \mathbf{u} = \text{argmin } J(\mathbf{u}), \qquad \text{with } U_{ad} \subseteq U := \mathbb{R}^n, \; n \geq 1; \tag{2.46}$$

we observe that, if $U_{ad} \equiv U$, optimization is unconstrained, while if $U_{ad} \subset U$, it is constrained. Moreover, let us suppose that $J(\mathbf{u})$ be twice differentiable in $\mathbf{u}$ and let us indicate with $\mathbf{g}(\mathbf{u}) \in \mathbb{R}^n$ and $H(\mathbf{u}) \in \mathbb{R}^{n\times n}$ the gradient and the Hessian, respectively, of the functional $J(\mathbf{u})$ and with $\mathbf{u}^{**}$ and $J(\mathbf{u}^{**})$ the optimal solution and the optimal cost functional.
Typically, iterative methods are used for the solution of optimization problems; in the unconstrained case, the stopping criterium is usually based on the norm of the gradient $\|\mathbf{g}(\mathbf{u})\| < tol$, where tol is a prescribed tolerance. In the constrained case, due to the presence of bounds delimitating the space U_{ad}, the previous stopping criterium is used together with another stopping criterium taking into account the satisfaction of the bounds on the control.

Steepest–Descent methods

Steepest–Descent (SD) methods are among the simplest methods which could be used for the solution of unconstrained optimization problems; see [152, 176].

Starting with an initial guess $\mathbf{u}^0 \in U$, the method consists in finding iteratively a sequence $\{\mathbf{u}^k\}$ s.t.:

$$\mathbf{u}^{k+1} = \mathbf{u}^k - \tau^k \mathbf{g}(\mathbf{u}^k) \qquad k = 0, 1, 2, \ldots, \tag{2.47}$$

where $\tau^k > 0$ is the relaxation parameter. Different choices for τ^k are available, such as:

1. $\tau^k = \tau \; \forall k = 0, 1, 2, \ldots$, with τ chosen s.t. $J(\mathbf{u}^{k+1}) < J(\mathbf{u}^k)$ (for some "simple" problems, an optimal relaxation parameter, say τ_{opt}, can be chosen a priori on the basis of the properties of the optimization problem [2]);

2. τ^k is chosen according with line search methods, such as e.g. the Armijo line search procedure [130, 152] (a simplified version consists in choosing $\tau^k = \overline{\tau}$, with $\overline{\tau}$ a prescribed relaxation parameter; then, $\tau^k = \overline{\tau}/2^j$ with $j = \min_{q\in\mathbb{N}}$ s.t. $J(\mathbf{u}^k - \overline{\tau}/2^q \; \mathbf{g}(\mathbf{u}^k)) < J(\mathbf{u}^k)$);

3. if $J(\mathbf{u}^{**}) > -\infty$ is known, τ^k is chosen as: $\tau^k = \left(J(\mathbf{u}^k) - J(\mathbf{u}^{**})\right)/\|\mathbf{g}(\mathbf{u}^k)\|^2$ (in this case the monotonicity of the sequence $\{J(\mathbf{u}^k)\}$, $k = 0, 1, 2, \ldots$, is not in general guaranteed).

We observe that the rate of convergence of SD methods is only linear. For this reason, even if the solution of an optimization problem by means of these methods requires a low amount of memory, the computational cost could be relevant.

Conjugate Gradient methods

Conjugate Gradient (CG) methods represent one of the most useful techniques for the solution of unconstrained optimization problems. Two variants of the method are available, which we indicate with the *linear* and *nonlinear* CG methods. For more on CG, we refer the reader to [64, 130, 152], but also to [172, 183, 184].

The CG methods consist in finding iteratively a sequence $\{\mathbf{u}^k\}$ s.t., given an initial guess $\mathbf{u}^0 \in U$:

$$\begin{aligned} \mathbf{u}^{k+1} &= \mathbf{u}^k + \delta^k \mathbf{d}^k \qquad k = 0, 1, 2, \dots, \\ \text{with } \mathbf{d}^k &= \begin{cases} -\mathbf{g}(\mathbf{u}^k) & \text{for } k = 0, \\ -\mathbf{g}(\mathbf{u}^k) + \beta^k \mathbf{d}^{k-1} & \text{for } k \geq 1, \end{cases} \end{aligned} \tag{2.48}$$

where $\delta^k > 0$ is a parameter obtained by a line search procedure, $\mathbf{d}^k$ is the search direction, while the parameter $\beta^k \in \mathbb{R}$ determines the type of CG method.

If the functional $J(\mathbf{u})$ in Eq.(2.46) is quadratic (with $H(\mathbf{u}^k) = H$), β^k is chosen s.t. $\{\mathbf{d}^0, \mathbf{d}^1, \dots, \mathbf{d}^{n-1}\}$ is a set of conjugate directions ($\mathbf{d}^{i\ T} H \mathbf{d}^j = 0\ \forall i \neq j$) and $\delta^k = -\left(\mathbf{g}(\mathbf{u}^k)^T \mathbf{d}^k\right) / \left(\mathbf{d}^{k\ T} H \mathbf{d}^k\right)$, the CG method is said *linear* (see e.g. [152]). The linear CG converges, in exact arithmetic, in at most n iterative steps; however, the optimal solution is often identified by the algorithm in less than n iterations. The efficiency of the method could be sensibly improved by using preconditioning techniques.

More in general the functional $J(\mathbf{u})$ is not quadratic. In these cases *nonlinear* CG methods, which corresponds to different choices of β^k in Eq.(2.48), are used. Possible choices of β^k are:

$$\begin{aligned} \beta^k_{PRP} &= \frac{\mathbf{g}(\mathbf{u}^k)^T \left(\mathbf{g}(\mathbf{u}^k) - \mathbf{g}(\mathbf{u}^{k-1})\right)}{\mathbf{g}(\mathbf{u}^{k-1})^T \mathbf{g}(\mathbf{u}^{k-1})} && CG_{PRP} \text{ (Polak–Ribiere–Poylak)}, \\ \beta^k_{PRP+} &= \max\{\beta^k_{PRP}, 0\} && CG_{PRP+} \text{ (mod. Polak–Ribiere–Poylak)}, \\ \beta^k_{CD} &= -\frac{\mathbf{g}(\mathbf{u}^k)^T \mathbf{g}(\mathbf{u}^k)}{\mathbf{d}^{k-1\ T} \mathbf{g}(\mathbf{u}^{k-1})} && CG_{CD} \text{ (Conjugate Descent)}; \end{aligned} \tag{2.49}$$

for other methods see e.g. [172, 183, 184]. The convergence behaviors of the formulas (2.49), together with line search conditions for δ^k (Armijo or Wolfe–Powell conditions [110, 184]), have been widely studied and tested for several test cases; see e.g. [130, 184]. However, it is not possible, in general, to determine which is the best method, but only to select the best one for particular classes of optimization problems.

Due to the low memory requirements, CG methods are often used for large scale unconstrained optimization problems (for which n is large).

Newton and quasi–Newton methods

Newton (N) and *quasi–Newton* (qN) methods represent, probably, the most used techniques for unconstrained optimization methods; we briefly recall here the formulations of the methods together with their properties. For further details, we refer the reader to [64, 110, 130] but also to [4, 5, 45].

Starting with an initial guess $\mathbf{u}^0 \in U$, a Newton type method consists in finding a sequence $\{\mathbf{u}^k\}$ s.t.:

$$\begin{aligned} \mathbf{u}^{k+1} &= \mathbf{u}^k + \delta^k \mathbf{d}^k \qquad k = 0, 1, 2, \dots, \\ &\text{with } H(\mathbf{u}^k)\mathbf{d}^k = -\mathbf{g}(\mathbf{u}^k), \end{aligned} \tag{2.50}$$

where, as for the CG methods, δ^k indicates a parameter which should be obtained by line search procedures.

In the case for which $\delta^k = 1$, the method (2.50) coincides with the Newton method [152], for which, if $\mathbf{u}^0$ is "sufficiently" near to $\mathbf{u}^{**}$, the convergence rate is quadratic; with this aim, SD methods are often used in combination with N and qN as initialization of the iterative procedure.

We observe that the N method requires the evaluation of the Hessian matrix $H(\mathbf{u}^k)$ and the solution of the linear system of Eq.(2.50) at each iterative step. This could lead to high computational costs especially for large optimization problems; we recall in fact that, in general, the Hessian matrix is full, for which a direct method is used for the solution of the linear system [152] (the computational cost is of order $O(n^3)$). To overcome this difficulties, an approximated inverse of the Hessian matrix, say B^k, is used to replace $H(\mathbf{u}^k)^{-1}$ in Eq.(2.50) s.t. $\mathbf{d}^k = -B^k\mathbf{g}(\mathbf{u}^k)$, where B^k must be symmetric positive definite. The so called quasi–Newton methods are based on the definition of such a matrix B^k; experience shows that, when line search methods are used, the most effective qN method is based on the BFGS (Broyden–Fletcher–Goldfarb–Shanno) update [110, 130] (qN$_{BFGS}$), s.t.:

$$B^{k+1} = \mathrm{bfgs}\left(B^k, \mathbf{w}^k, \mathbf{s}^k\right) \qquad k = 0, 1, 2, \dots, \tag{2.51}$$

with:

$$\mathrm{bfgs}\left(B, \mathbf{w}, \mathbf{s}\right) := \left(I - \frac{\mathbf{w}\mathbf{s}^T}{\mathbf{w}^T\mathbf{s}}\right) B \left(I - \frac{\mathbf{s}\mathbf{w}^T}{\mathbf{w}^T\mathbf{s}}\right) + \frac{\mathbf{w}\mathbf{w}^T}{\mathbf{w}^T\mathbf{s}}, \tag{2.52}$$

being $\mathbf{w}^k := \mathbf{u}^{k+1} - \mathbf{u}^k$ and $\mathbf{s}^k := \mathbf{g}(\mathbf{u}^{k+1}) - \mathbf{g}(\mathbf{u}^k)$; usually, $B^0 = I$. We notice that the convergence rate of a qN method is in general less than quadratic, due to the approximation of the Hessian matrix and the linear search procedure.

We observe that, even if the qN method based on the BFGS update does not require to solve at each iterative step a linear system, the memory storage of the full matrices B^k and B^{k+1} is required. This represents a limit on the effectiveness of the qN method when it is used for large scale optimization problems. A possibility to overcome this difficulty consists in generating a sparse matrix $\widetilde{B}^k$ approximating B^k, which is in general full; see [64]. In the literature, the most common approach is based on the *limited BFGS quasi–Newton* method (qN$_{L-BFGS}$), due to its simplicity and low memory storage requirements; see [110, 130] and [4, 5]. The qN$_{L-BFGS}$ method is based on the qN$_{BFGS}$ one, but the approximated matrix B^k is not explicitly generated neither stored in memory. A certain number of vectors, say $2m \ll n$, are used to store the information related to Hessian approximation; the amount of memory storage is controlled by the user by fixing m a priori according with the value of n and the features of the calculator. By referring to [4], the matrix B^{k+1} is defined implicitly by updating $\widehat{m}$ times, with $\widehat{m} := \min\{m, k\}$, a suitable matrix D^k in terms of the vectors $\mathbf{w}^i, \mathbf{s}^i$ $\forall i = k - \widehat{m} + 1, \dots, k$. At each iteration, the qN$_{L-BFGS}$ method requires the memory storage of the $2\widehat{m}$ vectors $\mathbf{w}^i, \mathbf{s}^i$; once that $k > m$, the oldest vectors are deleted and the newest updates stored in memory s.t. only at most $2m$ vectors are stored. For the qN$_{L-BFGS}$ method, Eq.(2.51) is replaced by:

$$B^{k+1} = \mathrm{lbfgs}\left(\widehat{m}, D^k, \mathbf{w}^k, \dots, \mathbf{w}^{k-\widehat{m}+1}, \mathbf{s}^k, \dots, \mathbf{s}^{k-\widehat{m}+1}\right) \qquad k = 0, 1, 2, \dots, \tag{2.53}$$

where:

$$\begin{gathered} \mathrm{lbfgs}\left(j, D, \mathbf{w}^i, \dots, \mathbf{w}^{i-j+1}, \mathbf{s}^i, \dots, \mathbf{s}^{i-j+1}\right) := \\ \mathrm{bfgs}\left(\mathrm{lbfgs}\left(j-1, D, \mathbf{w}^{i-1}, \dots, \mathbf{w}^{i-j+1}, \mathbf{s}^{i-1}, \dots, \mathbf{s}^{i-j+1}\right), \mathbf{w}^i, \mathbf{s}^i\right) \quad \forall i \geq j. \end{gathered} \tag{2.54}$$

The matrix D^k is chosen as $D^k = \nu^k I$, with $\nu^k = \max\left\{\rho^{k-\widehat{m}+1}, \rho^k\right\}$, being $\rho := (\mathbf{w}^T\mathbf{s})/(\mathbf{s}^T\mathbf{s})$.

Sequential Quadratic Programming methods

Sequential Quadratic Programming (SQP) methods represent the state of the art for the solution of constrained optimization problems. We recall here the basic concepts of the SQP methods; for further reading see [68, 130, 167] and also [192].

Let us suppose that the space of admissible controls U_{ad} is defined as $U_{ad} := \{\mathbf{u} \in U \ : \ c_i(\mathbf{u}) = 0 \quad i = 1, \ldots, m_e$ and $c_i(\mathbf{u}) \leq 0 \quad i = m_e + 1, \ldots, m\}$, with m the total number of constraints, while m_e the number of equality constraints; we define $\mathbf{c}(\mathbf{u}) \in \mathbb{R}^m$ as $\mathbf{c}(\mathbf{u}) := [c_i(\mathbf{u}), \ldots, c_{m_e}(\mathbf{u}), c_{m_e+1}(\mathbf{u}), \ldots, c_m(\mathbf{u})]^T$. Moreover, let us define the following Lagrangian functional:

$$\mathcal{L}(\mathbf{u}, \mathbf{z}) := J(\mathbf{u}) + \mathbf{z}^T \mathbf{c}(\mathbf{u}), \tag{2.55}$$

where $\mathbf{z} \in \mathbb{R}^m$ is the Lagrangian multiplier vector. SQP methods allow to closely mimic N and qN methods for constrained optimization in similar way as done for unconstrained optimization. In particular, at each iterative step, the Hessian of the Lagrangian functional (2.55) is approximated by means of qN updating techniques (typically by means of the BFGS technique). This is used to generate a Quadratic subproblem:

$$\begin{aligned} &\text{find } \mathbf{d}^k \in U, \ \mathbf{d}^k = \text{argmin } \left(\frac{1}{2} \mathbf{d}^{k\ T} H^k \mathbf{d} + \mathbf{g}(\mathbf{u}^k)^T \mathbf{d}^k \right), \\ &\text{with} \begin{cases} \nabla c_i(\mathbf{u}^k)^T \mathbf{d}^k + c_i(\mathbf{u}^k) = 0 & i = 1, \ldots, m_e, \\ \nabla c_i(\mathbf{u}^k)^T \mathbf{d}^k + c_i(\mathbf{u}^k) \leq 0 & i = m_e + 1, \ldots, m, \end{cases} \end{aligned} \tag{2.56}$$

whose solution $\mathbf{d}^k$ is used to compute $\mathbf{u}^{k+1} = \mathbf{u}^k + \delta^k \mathbf{d}^k$, being H^k and $\mathbf{g}^k$ the Hessian matrix and gradient vector associated to the Lagrangian functional (2.55) at the iterative step k, respectively. As usual, the parameter δ^k is determined on the basis of line search procedures. Problem (2.56) is solved by means of Quadratical Programming (QP) methods, for which we refer the reader to e.g. [64, 130]. SQP methods find the optimal solution by solving iteratively a sequence of QP subproblems, whose approximated Hessian matrix is associated with the Lagrangian functional (2.55) and not to the functional $J(\mathbf{u})$ as for N/qN methods. This means that, in general, N/qN and SQP methods are different. We notice that SQP methods coincide with N/qN methods if the functional $J(\mathbf{u})$ is quadratic and the constraints $\mathbf{c}(\mathbf{u})$ are linear in $\mathbf{u}$. In other words, SQP methods are equivalent to N/qN methods applied to the first–order necessary conditions (2.4)–(2.6); see [67]. On the basis of this assumption, it is possible to prove that SQP methods represent locally convergent algorithms; different techniques can be used to generate a global convergent algorithm.

2.3.2 Optimization methods for optimal control problems described by PDEs

In this Section we briefly discuss how to solve optimal control problems described by PDEs by means of the optimization methods reported in Sec.2.3.1; moreover, we present a "Direct method" for the solution of optimal control problems.

By referring to Sec.s 2.1.1 and 2.2.1, the optimal solution of the optimal control problems (2.2) or (2.28) is obtained by solving the first order conditions (2.4)–(2.6) or (2.29)–(2.31). The resulting system is in general fully coupled in the sense that each equation (or system of equations) depends on (v, z, u), with the exception of the primal equation which does not depend on the dual variable z.

In Sec.2.3.1 we have introduced several iterative methods, which can be conveniently used also for the solution of optimal control problems described by PDEs. With this aim, we need to use a numerical method for the approximation of the primal, dual and optimality equations which define the first order necessary conditions. As anticipated, several numerical methods are possible (such as Spectral or Reduced Basis methods). Specifically, we consider now the Finite Element (FE) method [147, 153] (in Chapter 7 we discuss the use of the Reduced Basis (RB) method for parametrized optimal control problems).

Let us indicate with N_h^ω the number of d.o.f. associated with the spatial approximation of a generic function defined in ω by means of the FE method, with N_h^Ω that associated with the spatial approximation of a function defined in Ω and with N_t the number of steps in which the time interval $[0, T]$ is divided due to the time discretization. It follows, by applying standard FE methods for elliptic and parabolic PDEs that the control function $u \in \mathcal{U}$ corresponds to a vector $\mathbf{u} \in \mathbb{R}^{N_h^\omega N_t}$; similarly, the approximated primal and dual variables read, respectively: $\mathbf{v} \in \mathbb{R}^{N_h^\Omega N_t}$ and $\mathbf{z} \in \mathbb{R}^{N_h^\Omega N_t}$. We observe that in the steady case we have $N_t = 1$. Let us suppose to use the notation introduced in Sec.2.1.2 and to consider the particular optimal control problems pointed out in Remarks 2.3 and 2.5, then the system corresponding to the approximated first order necessary conditions (2.4)–(2.6) can be generically written in matricial notation as:

$$A^T \mathbf{z}^{**} = M_d(\mathbf{v}^{**} - \mathbf{v}_d), \tag{2.57}$$

$$\gamma M_u(\mathbf{u}^{**} - \mathbf{u}_d) + B^T \mathbf{z}^{**} = 0, \tag{2.58}$$

$$A\mathbf{v}^{**} = \mathbf{F} + B\mathbf{u}^{**}, \tag{2.59}$$

where $A \in \mathbb{R}^{N_h^\Omega N_t \times N_h^\Omega N_t}$, $F \in \mathbb{R}^{N_h^\Omega N_t}$, $M_d \in \mathbb{R}^{N_h^\Omega N_t \times N_h^\Omega N_t}$, $B \in \mathbb{R}^{N_h^\Omega N_t \times N_h^\omega N_t}$, $M_u \in \mathbb{R}^{N_h^\omega N_t \times N_h^\omega N_t}$; the matrices and the vectors of Eq.s (2.57)–(2.59) take into account for appropriate boundary and initial conditions. In similar manner, it is possible to consider constrained optimal control problems and the nonlinear case. We remark that Eq.(2.58) represents the optimality condition (2.5), hence the sensitivity vector, say $\delta\mathbf{u} \in \mathbb{R}^{N_h^\omega N_t}$ (which corresponds to $\delta u \in \mathcal{U}^*$ in Eq.(2.7)), is $\delta\mathbf{u} = \gamma M_u(\mathbf{u} - \mathbf{u}_d) + B^T\mathbf{z}$. We observe that $\delta\mathbf{u}$ plays the same role of the gradient $\mathbf{g}$ of Sec.2.3.1.

On these basis, it is possible to solve the approximated optimal control problem by means of an optimization iterative method, such as SD, CG, N/qN methods or SQP methods for constrained problems. For a given approximated optimal control problem, the *iterative algorithm* reads (see also [40, 161]):

1. choose an initial guess $\mathbf{u}^0$;

2. solve the primal equation (2.59) given $\mathbf{u}$ in order to obtain $\mathbf{v}$ and compute the cost functional $J(\mathbf{v}, \mathbf{u})$;

3. solve the dual equation (2.57) given $\mathbf{v}$ in order to obtain $\mathbf{z}$;

4. compute the sensitivity vector $\delta\mathbf{u}$ given $\mathbf{z}$, $\mathbf{u}$ and, for nonlinear problems, $\mathbf{v}$;

5. if $\|\delta\mathbf{u}\| < Tol$, stop the algorithm (for a constrained problem the stopping criterium should take into account of the bounds delimitating the space of admissible controls);

6. update $\mathbf{u}$ according with one of the optimization methods (SD, CG, N/qN, or SQP) introduced in Sec.2.3.1 and return to Step 2.

Direct method

An other approach for the numerical solution of optimal control problems consists in using a "*Direct* method", which corresponds to solving Eq.s (2.57)–(2.59) by means of a direct method for linear systems (see [152]). In particular, we solve the following linear system:

$$\overline{A}\mathbf{x}^{**} = \overline{\mathbf{F}}, \tag{2.60}$$

where, $\mathbf{x}^{**} := [\mathbf{v}^{**}, \mathbf{z}^{**}, \mathbf{u}^{**}]^T$ and from Eq.s (2.57)–(2.59):

$$\overline{A} := \begin{bmatrix} A & 0 & -B \\ -M_d & A^T & 0 \\ 0 & B^T & \gamma M_u \end{bmatrix} \qquad \text{and} \qquad \overline{\mathbf{F}} := \begin{bmatrix} \mathbf{F} \\ -M_d\mathbf{v}_d \\ \gamma M_u \mathbf{u}_d \end{bmatrix}. \tag{2.61}$$

We remark that $\overline{A} \in \mathbb{R}^{N_t(2N_h^\Omega + N_h^\omega) \times N_t(2N_h^\Omega + N_h^\omega)}$ is a sparse matrix, hence, for the solution of system (2.60) it is convenient to use a direct method for sparse matrices ([152]). We observe that for this kind of methods the computational limit is represented by the amount of memory storage allowed during the numerical solution. Hence, for large scale optimization problems, the Direct method requires a large amount of memory and it could be not effective due to the size of matrix $\overline{A}$.
We observe that, by recalling Remarks 2.3 and 2.5, if $\gamma = 0$ system (2.60) is not well–posed, for which $\det(\overline{A}) = 0$; however, a solution can still be obtained by means of an iterative method, which evolves towards a solution (not unique) given a prescribed initial guess $\mathbf{u}^0$.
In similar way, for nonlinear optimal control problems, it is possible to obtain a nonlinear system corresponding to Eq.s (2.57)–(2.59). Standard techniques for the solution of nonlinear systems can be used, as the Newton method; see [152].

2.3.3 A comparison of the optimization methods

In this Section we provide a numerical test based on an unconstrained optimization problem described by an elliptic PDE. In particular, we compare, in terms of computational costs, the optimization methods of Sec.2.3.1 and the Direct method of Sec.2.3.2 for a large scale problem. For further examples of unconstrained optimal control problems described by PDEs, refer e.g. to [40, 161].

Example

By using the standard notation of Sec.2.1.2, we consider an optimal control problem for which $J(v,u) = \frac{1}{2}\int_\Omega (v - v_d)^2 \, d\Omega + \frac{1}{2}\gamma \int_\Omega (u - u_d)^2 \, d\Omega$ and the state equation is given by the elliptic PDE:

$$\begin{cases} -\Delta y = u & \text{in } \Omega, \\ y = 0 & \text{on } \partial\Omega, \end{cases} \tag{2.62}$$

being $\Omega = (0,1)^2 \subset \mathbb{R}^2$. In particular, we choose $v_d = 10\mathrm{x}_1(1-\mathrm{x}_1)\mathrm{x}_2(1-\mathrm{x}_2)$ and $u_d = 0$.
For the numerical approximation, we use the FE method with $\mathbb{P}^1$ continuous basis defined over uniform meshes, say $\mathcal{T}_h$, composed by triangular elements. For the solution of the unconstrained optimal control problem, we consider the Direct method of Sec.2.3.2 and the following iterative methods: SD (with $\tau = 1$), CG$_{PRP}$, CG$_{PRP+}$, CG$_{CD}$ and qN$_{L-BFGS}$ (see Sec.2.3.1), initialized with $\mathbf{u}^0 = \mathbf{0}$. An Armijo line search procedure [130] is used for

Mesh	$\mathcal{T}_{h1}$	$\mathcal{T}_{h2}$	$\mathcal{T}_{h3}$	$\mathcal{T}_{h4}$	$\mathcal{T}_{h5}$	$\mathcal{T}_{h6}$
♯ Triangles	$1,342$	$5,422$	$10,646$	$42,584$	$170,336$	$681,344$
♯ Nodes	712	$2,792$	$5,436$	$21,517$	$85,617$	$341,569$

Table 2.1: Meshes used for the numerical test; number of triangles and nodes.

$\mathcal{T}_{h1}$

	$\gamma = 10^{-3}$		$\gamma = 10^{-5}$		$\gamma = 0$	
Method	♯ it.	CPU $[s]$	♯ it.	CPU $[s]$	♯ it.	CPU $[s]$
SD	28	1.44	282	13.4	799	36.9
CG$_{PRP}$	35	1.94	189	13.5	509	35.1
CG$_{PRP+}$	29	1.59	401	33.5	1172	101
CG$_{CD}$	32	1.82	87	4.66	160	8.82
qN$_{L-BFGS}$	7	0.620	20	1.77	21	1.84
Direct	–	0.100	–	0.100	–	–†

$\mathcal{T}_{h2}$

	$\gamma = 10^{-3}$		$\gamma = 10^{-5}$		$\gamma = 0$	
Method	♯ it.	CPU $[s]$	♯ it.	CPU $[s]$	♯ it.	CPU $[s]$
SD	28	4.10	282	46.0	799	101
CG$_{PRP}$	35	5.42	169	39.2	507	99.2
CG$_{PRP+}$	29	4.50	400	107	1140	281
CG$_{CD}$	68	14.0	118	25.5	279	55.7
qN$_{L-BFGS}$	7	1.68	17	4.95	23	6.87
Direct	–	0.320	–	0.320	–	–†

† : $\texttt{det}(\overline{A}) = 0$

Table 2.2: Comparison of the computational costs and number of iterative steps for the meshes $\mathcal{T}_{h1}$ and $\mathcal{T}_{h2}$.

all the iterative methods, for which the stopping criterium is based on $\|\delta\mathbf{u}^k\|/\|\delta\mathbf{u}^0\| < Tol$, with $Tol = 10^{-4}$. We consider the cases for which $\gamma = 10^{-3}, 10^{-5}, 0$ and six meshes, say $\mathcal{T}_{h1}, \ldots, \mathcal{T}_{h6}$, whose properties are reported in Table 2.1. Computations are carried out by means of the Matlab® software [192] on an AMD Athlon 1.8 GHz processor, with 256 KB of memory cache and 1 GB of memory RAM.

In Table 2.2 we compare the optimization methods for the cases $\gamma = 10^{-3}, 10^{-5}, 0$ for the large scale problems associated with the meshes $\mathcal{T}_{h1}$ and $\mathcal{T}_{h2}$. We observe that, among the iterative methods, the qN$_{L-BFGS}$ is the most effective in terms of computational costs (CPU). The Direct method allows a very rapid solution of the optimal control problem; however, as anticipated, the case $\gamma = 0$ could not be solved by means of this approach. Moreover, we notice that, for all the iterative methods, the number of iterations and the computational costs increase as $\gamma \to 0$.

In Table 2.3 we compare the qN$_{L-BFGS}$ and Direct methods for the cases $\gamma = 10^{-3}, 10^{-5}, 0$ for "very" large scale problems (associated with the meshes $\mathcal{T}_{h3}, \ldots, \mathcal{T}_{h6}$). In particular, the Direct method results more efficient than the qN$_{L-BFGS}$ and hence than all the iterative methods considered (for example with $\mathcal{T}_{h3}$ all the CG methods converge to the optimal solution in more than $10,000$ iteration and with CPU times greater than 650 s). However, we observe that for $\mathcal{T}_{h6}$, the Direct method is not able to provide the optimal solution, due to memory

$\mathcal{T}_{h3}$

	$\gamma = 10^{-3}$		$\gamma = 10^{-5}$		$\gamma = 0$	
Method	$\sharp$ it.	CPU [s]	$\sharp$ it.	CPU [s]	$\sharp$ it.	CPU [s]
qN$_{L-BFGS}$	7	3.96	17	9.60	31	18.0
Direct	–	0.740	–	0.740	–	–†

$\mathcal{T}_{h4}$

	$\gamma = 10^{-3}$		$\gamma = 10^{-5}$		$\gamma = 0$	
Method	$\sharp$ it.	CPU [s]	$\sharp$ it.	CPU [s]	$\sharp$ it.	CPU [s]
qN$_{L-BFGS}$	7	20.2	14	41.6	31	95.9
Direct	–	4.64	–	4.67	–	–†

$\mathcal{T}_{h5}$

	$\gamma = 10^{-3}$		$\gamma = 10^{-5}$		$\gamma = 0$	
Method	$\sharp$ it.	CPU [s]	$\sharp$ it.	CPU [s]	$\sharp$ it.	CPU [s]
qN$_{L-BFGS}$	7	94.8	14	204	27	382
Direct	–	36.5	–	37.0	–	–†

$\mathcal{T}_{h6}$

	$\gamma = 10^{-3}$		$\gamma = 10^{-5}$		$\gamma = 0$	
Method	$\sharp$ it.	CPU [s]	$\sharp$ it.	CPU [s]	$\sharp$ it.	CPU [s]
qN$_{L-BFGS}$	7	452	15	1083	24	1678
Direct	–	–‡	–	–‡	–	–†

† : `det`$(\overline{A}) = 0$ ‡ : `out of memory`

Table 2.3: Comparison of the computational costs and number of iterative steps for the meshes $\mathcal{T}_{h3}, \ldots, \mathcal{T}_{h6}$.

problems: the memory requirements are in fact larger than those allowed by the computer in use. Conversely, the qN$_{L-BFGS}$ provides an optimal solution also for this "very" large scale problem.

We conclude that the Direct method is really effective for the solution of well–posed optimization problems compatibly with the memory performances of the computer in use. For "very" large scale optimal control problems and "ill–posed" optimal control problems, the qN$_{L-BFGS}$ represents the most effective method.

Chapter 3

Optimal Flow Control for Navier–Stokes Equations: Drag Minimization

In this Chapter we provide an applicative example of the optimal control theory discussed in Chapter 2 concerning an optimal flow control problem for Navier–Stokes equations.

A popular problem in Fluid Dynamics consists in minimizing the drag coefficient of a body in relative motion with a fluid [19, 57, 58, 63, 78, 80, 84, 94, 125]; in particular, one case commonly studied is that of the steady incompressible Navier–Stokes equations with constant density and viscosity. Some applications are, e.g., the design of airfoils in Aerodynamics [94] (at high Reynolds numbers) and the study of the geometry of blunt bodies in flows at low Reynolds numbers (see e.g. [166]). The drag minimization problem can be recast in the theory of optimal control for PDEs (see Chapter 2 and [2, 107]), or as a problem of shape optimization [63, 125]. We are interested in minimizing the drag coefficient of a blunt body by acting on the velocity at the boundaries of the body itself (see also [41]); this corresponds to regulate the aspiration or the blowing of the boundary layer in order to reduce the effects of the vortices coming off from the rear of the body. This problem can be formulated as an unconstrained optimal control problem, for which the control function is the Dirichlet boundary condition [18, 57, 58, 78, 80, 81, 84, 88]. With this aim we use the Lagrangian functional approach (see Sec.2.1 and [19, 20, 22, 44, 78]) instead of calculating directly the gradient of the functional subject to minimization [57, 58, 78, 81, 84]. In this context the main difficulty consists in the treatment of the Dirichlet boundary control function, which does not match with the variational setting provided by the Lagrangian functional approach, unless suitable lifting terms are introduced.

Different approaches have been considered in literature to overcome this difficulty in the field of Navier–Stokes equations. In [18] the Nitsche's method ([153]) is proposed in order to suppress the recirculation in a backward facing step in a channel by regulating the inflow velocity field. In [87, 88] both linear and non–linear penalized Neumann control approaches for the solution of optimal control problems with Dirichlet control functions are considered. Moreover in [88] a Lagrangian multiplier method ([15]) is proposed for the treatment of the Dirichlet control function (this strategy is also considered in [117] for optimal control problems governed by elliptic PDEs). The approaches presented in [87, 88] are applied for the minimization of the vorticity and the “distance” between the velocity field and a desired one; sequential quadratic programming (SQP) methods (see Sec.2.3.1 and, e.g., [64]) have been used for the solution of these optimal control problems.

We use, for our drag minimization problem, the Lagrangian multiplier method for the treatment of the Dirichlet velocity control function on the body, in analogy with the approaches outlined in [88] and in [117] (for elliptic PDEs). However for the evaluation of the drag, or more in general the force acting on the body, we exploit the variational form of the Navier–Stokes equations [60], instead of using the definition, the latter being directly related to the integral of the stress acting on the body boundaries. The two approaches, that are equivalent at the "continuous level", are in fact different at the "discrete level". The one that we follow yields more accurate computation of the drag coefficient [60] and, in the context of optimization, reduces the propagation of the approximation errors in the course of the iterative optimization procedure. Moreover, this approach allows us to identify the Lagrangian multiplier as the inward directed normal stress, to express the drag coefficient in terms of this multiplier and, consequently, to obtain a simple expression for the dual equations.

For the numerical solution of the optimal control problem we use the Steepest–Descent method (see Sec.2.3.1), while the PDE system is approximated by means of the Finite Element (FE) method. Finally, we report some numerical results concerning the minimization of the drag for blunt bodies (at low Reynolds numbers), which prove the effectiveness of the outlined procedures.

This Chapter is organized as follows. In Sec.3.1, we consider the drag coefficient minimization problem for the incompressible Navier–Stokes equations, introducing the evaluation of the drag coefficient by means of the variational form. Finally, we write the Lagrangian functional, using the Lagrangian multiplier method for the treatment of the Dirichlet boundary conditions, and we formulate the dual Navier–Stokes equations and the sensitivity equation. In Sec.3.2 we discuss the aspects related to the numerical resolution of the optimal control problem: in particular we recall the techniques considered for the numerical solution of the Stokes and Navier–Stokes equations. In Sec.3.3 we present some numerical results and compare them for different values of the Reynolds number. Finally, in Sec.3.4 concluding remarks follow.

3.1 Mathematical model: the optimal control problem

In this Section we treat a flow control problem governed by the Navier–Stokes equations for the minimization of the drag acting on a solid object.

3.1.1 Drag minimization for Navier–Stokes equations

We consider a body embedded in a $2D$ flow governed by the steady incompressible Navier–Stokes equations for a Newtonian fluid with constant density and viscosity.

The goal consists in minimizing the drag coefficient by regulating the flow $\mathbf{u}$ across the boundary Γ_{CTRL} of the body in relative steady motion with the fluid; see Fig.3.1. By defining $\mathbf{v}$ as the velocity field, p as the pressure, μ as dynamic viscosity coefficient, ρ as the density,

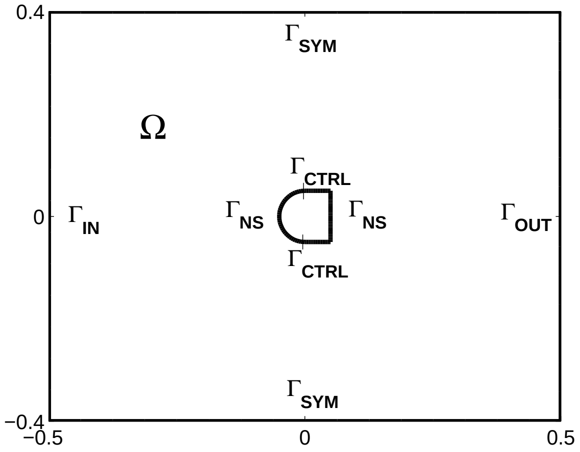

Figure 3.1: Domain for the control problem and boundary conditions.

the Navier–Stokes system reads [65, 153]:

$$\begin{cases} -\nabla\cdot\mathbb{T}(\mathbf{v},p)+\rho(\mathbf{v}\cdot\nabla)\mathbf{v}=\mathbf{0} & \text{in } \Omega, \\ \nabla\cdot\mathbf{v}=0 & \text{in } \Omega, \\ \mathbf{v}=\mathbf{v}_\infty & \text{on } \Gamma_{IN}, \\ \mathbf{v}\cdot\hat{\mathbf{n}}=0, \quad (\mathbb{T}(\mathbf{v},p)\hat{\mathbf{n}})\cdot\hat{\mathbf{t}}=0 & \text{on } \Gamma_{SYM}, \\ \mathbb{T}(\mathbf{v},p)\hat{\mathbf{n}}=\mathbf{0} & \text{on } \Gamma_{OUT}, \\ \mathbf{v}=\mathbf{0} & \text{on } \Gamma_{NS}, \\ \mathbf{v}=\mathbf{u} & \text{on } \Gamma_{CTRL}, \end{cases} \tag{3.1}$$

where $\hat{\mathbf{n}}$ and $\hat{\mathbf{t}}$ are respectively the outward directed normal and tangential unit vectors on the boundary Γ_i; the *stress tensor* $\mathbb{T}(\mathbf{v},p)$ reads:

$$\mathbb{T}(\mathbf{v},p)=\mu(\nabla\mathbf{v}+\nabla^T\mathbf{v})-p\mathbb{I}, \tag{3.2}$$

$\mathbb{I}$ being the identity tensor. We impose inflow boundary conditions on Γ_{IN}, symmetry conditions on Γ_{SYM}, no stress conditions on Γ_{OUT}, no slip conditions on Γ_{NS} and Dirichlet conditions ($\mathbf{u}$ is the control variable) on the control boundary Γ_{CTRL}, which corresponds to imposing the velocity on Γ_{CTRL}. Let us notice that $\cup_i\Gamma_i=\partial\Omega$ and $\Gamma_i\cap\Gamma_j=\emptyset$, $\forall i,j$.
The functional to minimize is the adimensional drag coefficient $c_D(\mathbf{v},p)$, for which the problem that we consider is:

$$\text{find } \mathbf{u}= \ \text{argmin } c_D(\mathbf{v},p) \ \text{ with } \ c_D(\mathbf{v},p):=-\frac{1}{q_\infty d}\oint_{\Gamma_{BODY}}(\mathbb{T}(\mathbf{v},p)\hat{\mathbf{n}})\cdot\hat{\mathbf{v}}_\infty\, d\Gamma, \tag{3.3}$$

and $(\mathbf{v},p)$ depends on $\mathbf{u}$ through the primal equations (3.1). The minus sign takes into account that, by convention, the force is positive if acting on the fluid. In Eq.(3.3) $\Gamma_{BODY}:=\Gamma_{NS}\cup\Gamma_{CTRL}$, $q_\infty:=\frac{1}{2}\rho V_\infty^2$, with $\mathbf{v}_\infty=V_\infty\hat{\mathbf{v}}_\infty$, $\hat{\mathbf{v}}_\infty$ is the unit vector directed as the incoming flow, V_∞ is taken constant and d is the characteristic dimension of the body. Let us notice that the drag $q_\infty d\, c_D$ has the dimension of a force per unit length.
We consider a parabolic profile for the control flow, written in the form:

$$\mathbf{u}=U\mathbf{g}(\mathrm{x}_1) \quad \text{with} \quad \mathbf{g}(\mathrm{x}_1):=-4\frac{(\mathrm{x}_1-\mathrm{x1}_{CTRL})(\mathrm{x2}_{CTRL}-\mathrm{x}_1)}{(\mathrm{x2}_{CTRL}-\mathrm{x1}_{CTRL})^2}\hat{\mathbf{n}}_{CTRL}, \tag{3.4}$$

where $\mathrm{x1}_{CTRL}$ and $\mathrm{x2}_{CTRL}$ are the abscissae of the endpoints of the boundary Γ_{CTRL}, while $\hat{\mathbf{n}}_{CTRL}$ is the outward directed unit vector normal to Γ_{CTRL}. Let us notice that the versus of the unit vector $\hat{\mathbf{n}}_{CTRL}$ is from the fluid towards the body, which is external to the computational domain; moreover, on Γ_{CTRL}, we have $\mathbf{g}(\mathrm{x}_1) \cdot \hat{\mathbf{n}}_{CTRL} \leq 0$. The effective control variable is the parameter $U \in \mathbb{R}$, which is the maximum value (in modulus) of the parabolic flow of Eq.(3.4). Let us observe that a positive value of U corresponds to a blowing of the fluid across the body walls Γ_{CTRL}; on the contrary, if U is negative, the fluid is aspirated (i.e. the flow is directed from the fluid towards the body).

Weak formulation

For the analysis of the optimal control problem we introduce, similarly to what was done in Sec.2.1.2 for the unconstrained steady case, the weak form of the Navier–Stokes equations. This can be done by introducing suitable functional spaces for $\mathbf{v}$ and p [65, 153], which account properly for Dirichlet boundary conditions; an alternative method consists in introducing the boundary conditions by means of a Lagrange multiplier approach [15, 88, 117]. This last approach, which we will follow, allows a straightforward treatment of the boundary conditions in the analysis of the optimal control problem, without introducing lifting terms.

Let us introduce the following forms:

$$a(\mathbf{v}, \mathbf{\Phi}) := \int_\Omega \mu(\nabla \mathbf{v} + \nabla^T \mathbf{v}) \cdot \nabla \mathbf{\Phi} \; d\Omega, \tag{3.5}$$

$$b(p, \mathbf{\Phi}) := -\int_\Omega p \nabla \cdot \mathbf{\Phi} \; d\Omega, \tag{3.6}$$

$$c(\mathbf{v}, \mathbf{v}, \mathbf{\Phi}) := \int_\Omega \rho(\mathbf{v} \cdot \nabla)\mathbf{v} \cdot \mathbf{\Phi} \; d\Omega. \tag{3.7}$$

The Navier–Stokes equations in weak form, with boundary conditions introduced through a Lagrange multiplier, read:

$$\begin{gathered} \text{find} \quad \mathbf{v} \in [H^1(\Omega)]^2, \quad p \in L^2(\Omega), \quad \mathbf{w} \in [H^{-\frac{1}{2}}(\Gamma_{DS})]^2 \quad : \\ a(\mathbf{v}, \mathbf{\Phi}) + b(p, \mathbf{\Phi}) + c(\mathbf{v}, \mathbf{v}, \mathbf{\Phi}) + \int_{\Gamma_D} \mathbf{w} \cdot \mathbf{\Phi} \; d\Gamma + \int_{\Gamma_{SYM}} \mathbf{w} \cdot \hat{\mathbf{n}} \; \mathbf{\Phi} \cdot \hat{\mathbf{n}} \; d\Gamma = 0, \\ b(\varphi, \mathbf{v}) = 0, \\ \int_{\Gamma_D} (\mathbf{v} - \mathbf{v}_D) \cdot \mathbf{\Lambda} \; d\Gamma + \int_{\Gamma_{SYM}} \mathbf{v} \cdot \hat{\mathbf{n}} \; \mathbf{\Lambda} \cdot \hat{\mathbf{n}} \; d\Gamma = 0 \\ \forall \mathbf{\Phi} \in [H^1(\Omega)]^2, \quad \forall \varphi \in L^2(\Omega), \quad \forall \mathbf{\Lambda} \in [H^{-\frac{1}{2}}(\Gamma_{DS})]^2, \end{gathered} \tag{3.8}$$

where $\Gamma_D := \Gamma_{IN} \cup \Gamma_{NS} \cup \Gamma_{CTRL}$, $\Gamma_{DS} := \Gamma_D \cup \Gamma_{SYM}$, $[H^1(\Omega)]^2$ and $[H^{-\frac{1}{2}}(\Gamma_{DS})]^2$ are the usual Sobolev spaces, while $\mathbf{v}_D$, which we consider in the Sobolev space $[H^{\frac{1}{2}}(\Gamma_D)]^2$, reads (Eq.(3.1)):

$$\mathbf{v}_D := \begin{cases} \mathbf{v}_\infty & \text{on } \Gamma_{IN}, \\ \mathbf{0} & \text{on } \Gamma_{NS}, \\ \mathbf{u} & \text{on } \Gamma_{CTRL}. \end{cases} \tag{3.9}$$

In particular, the control variable $\mathbf{u} \in [H^{\frac{1}{2}}(\Gamma_{CTRL})]^2$. The Lagrange multiplier $\mathbf{w} \in [H^{-\frac{1}{2}}(\Gamma_{DS})]^2$ is introduced to allow the variable $\mathbf{v} \in [H^1(\Omega)]^2$ to fit the Dirichlet boundary conditions, which includes the control variable $\mathbf{u}$. Let us observe that we suppose $\hat{\mathbf{n}}$ (i.e. Γ_{SYM}) to be "sufficiently regular" so that in Eq.(3.8) $\mathbf{w} \in [H^{-\frac{1}{2}}(\Gamma_{SYM})]^2$ implies $\mathbf{w} \cdot \hat{\mathbf{n}} \in H^{-\frac{1}{2}}(\Gamma_{SYM})$, $\mathbf{v} \in [H^{\frac{1}{2}}(\Gamma_{SYM})]^2$ implies $\mathbf{v} \cdot \hat{\mathbf{n}} \in H^{\frac{1}{2}}(\Gamma_{SYM})$, and so on. This is the case of the problem under investigation (see Fig.3.1), for which Γ_{SYM} is composed by two distinct segments. Moreover the boundary integrals of Eq.(3.8) can be seen as duality pairs: ${}_{H^{-\frac{1}{2}}(\Gamma_i)}\langle \cdot, \cdot \rangle_{H^{\frac{1}{2}}(\Gamma_i)}$ or ${}_{H^{\frac{1}{2}}(\Gamma_i)}\langle \cdot, \cdot \rangle_{H^{-\frac{1}{2}}(\Gamma_i)}$. Finally, in view of the finite element approximation, problem (3.8) will be reformulated in weak form [153]; the solution $\tilde{\mathbf{v}}$ is sought for in $[H^1_{|\Gamma_{DS}}(\Omega)]^2$, the test function $\mathbf{\Phi} \in [H^1_{|\Gamma_{DS}}(\Omega)]^2$, where $[H^1_{|\Gamma_{DS}}(\Omega)]^2 := \{\mathbf{s} \in [H^1(\Omega)]^2 : \ \mathbf{s}_{|\Gamma_D} = \mathbf{0} \text{ and } (\mathbf{s} \cdot \hat{\mathbf{n}})_{|\Gamma_{SYM}} = 0\}$, by modifying the source term in order to account for the non–homogeneous Dirichlet data. Thereby the direct evaluation of the variable $\mathbf{w} \in [H^{-\frac{1}{2}}(\Gamma_{DS})]^2$ is not necessary.

Drag evaluation

The drag coefficient c_D, defined in Eq.(3.3), can be evaluated once the Navier–Stokes equations are resolved and the velocity field $\mathbf{v}$ and the pressure p are known. However, the computation of c_D by means of the definition (3.3) can lead to inaccurate results even if the computational grid is quite fine, as we better specify later. Better results can be achieved by using an alternative expression for c_D, which we indicate with $\tilde{c}_D$, dependent on the variational forms used for the Navier–Stokes equations, as reported in [22, 60, 84].

In order to provide the expression of $\tilde{c}_D$, we express c_D (3.3) in the following form:

$$c_D(\mathbf{v}, p) = \frac{1}{q_\infty d} \oint_{\partial\Omega} (\mathbb{T}(\mathbf{v}, p)\hat{\mathbf{n}}) \cdot \mathbf{\Phi}_\infty \, d\Gamma, \tag{3.10}$$

where:

$$\mathbf{\Phi}_\infty \in [H^1(\Omega)]^2 \quad \text{with} \quad \mathbf{\Phi}_{\infty|\Gamma_{BODY}} = -\hat{\mathbf{v}}_\infty, \quad \mathbf{\Phi}_{\infty|\partial\Omega\setminus\Gamma_{BODY}} = \mathbf{0}. \tag{3.11}$$

By means of the Gauss theorem and the Green's identity ([60, 65]), Eq.(3.10) becomes:

$$\begin{aligned} c_D(\mathbf{v}, p) &= \frac{1}{q_\infty d} \int_\Omega \nabla \cdot (\mathbb{T}(\mathbf{v}, p)^T \mathbf{\Phi}_\infty) \, d\Omega \\ &= \frac{1}{q_\infty d} \int_\Omega (\nabla \cdot \mathbb{T}(\mathbf{v}, p) \cdot \mathbf{\Phi}_\infty + \mathbb{T}(\mathbf{v}, p) \cdot \nabla \mathbf{\Phi}_\infty) \, d\Omega. \end{aligned} \tag{3.12}$$

Owing to the first equation of the system (3.1) and Eq.(3.7), we have:

$$\int_\Omega \nabla \cdot \mathbb{T}(\mathbf{v}, p) \cdot \mathbf{\Phi}_\infty \, d\Omega = \int_\Omega \rho(\mathbf{v} \cdot \nabla)\mathbf{v} \cdot \mathbf{\Phi}_\infty \, d\Omega = c(\mathbf{v}, \mathbf{v}, \mathbf{\Phi}_\infty), \tag{3.13}$$

while, from Eq.(3.2), (3.5) and (3.6), we obtain:

$$\begin{aligned} \int_\Omega \mathbb{T}(\mathbf{v}, p) \cdot \nabla \mathbf{\Phi}_\infty \, d\Omega &= \int_\Omega \mu(\nabla \mathbf{v} + \nabla^T \mathbf{v}) \cdot \nabla \mathbf{\Phi}_\infty \, d\Omega - \int_\Omega p \nabla \cdot \mathbf{\Phi}_\infty \, d\Omega \\ &= a(\mathbf{v}, \mathbf{\Phi}_\infty) + b(p, \mathbf{\Phi}_\infty), \end{aligned} \tag{3.14}$$

for which, being $b(\varphi, \mathbf{v}) = 0 \ \forall \varphi \in L^2(\Omega)$, $\tilde{c}_D$ reads:

$$\tilde{c}_D(\mathbf{v}, p) = \frac{1}{q_\infty d} \mathcal{A}(\mathbf{v}, p, \mathbf{\Phi}_\infty, \varphi) \qquad \forall \varphi \in L^2(\Omega), \tag{3.15}$$

where:

$$\mathcal{A}(\mathbf{v}, p, \mathbf{\Phi}, \varphi) := a(\mathbf{v}, \mathbf{\Phi}) + b(p, \mathbf{\Phi}) + b(\varphi, \mathbf{v}) + c(\mathbf{v}, \mathbf{v}, \mathbf{\Phi}). \tag{3.16}$$

The form for the drag coefficient provided in Eq.(3.15) allows accurate computations and will be used in this work. Let us notice that c_D and $\tilde{c}_D$ are equivalent at the continuous level, however the representations (3.12) and (3.15) lead to different approximations at the discrete level. For instance, by using the FE method with continuous, piecewise polynomials of degree k for $\mathbf{v}$ and $k-1$ for p (Taylor–Hood elements, see [28, 153, 173]), the order of convergence of $\tilde{c}_{Dh}$ to the exact value c_D is $2k$, while for c_{Dh} is only k [60], being c_{Dh} and $\tilde{c}_{Dh}$ the computed quantities corresponding to c_D and $\tilde{c}_D$. This fact has great importance in the optimization context, where the dual variables depend on the sensitivity of the cost functional (in this case c_D) with respect to the primal variables. The choice of the representation, as verified by numerical tests, not only influences the computation of the dual variables, but it also affects the accuracy of the result obtained by the optimization procedure.

Let us observe that, according with Eq.(3.8), the variational form $\mathcal{A}(\mathbf{v}, p, \mathbf{\Phi}, \varphi)$ can be expressed as:

$$\mathcal{A}(\mathbf{v}, p, \mathbf{\Phi}, \varphi) = -\int_{\Gamma_D} \mathbf{w} \cdot \mathbf{\Phi} \, d\Gamma - \int_{\Gamma_{SYM}} \mathbf{w} \cdot \hat{\mathbf{n}} \; \mathbf{\Phi} \cdot \hat{\mathbf{n}} \, d\Gamma, \tag{3.17}$$

for which, upon the definition of $\mathbf{\Phi}_\infty$ given in Eq.(3.11), the drag coefficient $\tilde{c}_D$ admits the alternative expression:

$$\hat{c}_D(\mathbf{w}) = \frac{1}{q_\infty d} \int_{\Gamma_{BODY}} \mathbf{w} \cdot \hat{\mathbf{v}}_\infty \, d\Gamma. \tag{3.18}$$

Let us notice that Eq.(3.18) allows us to identify the Lagrange multiplier $\mathbf{w} \in [H^{-\frac{1}{2}}(\Gamma_{DS})]^2$ as the inward normal stress, more precisely:

$$\mathbf{w} = -\mathbb{T}(\mathbf{v}, p)\hat{\mathbf{n}} \qquad \text{on} \quad \Gamma_{DS}. \tag{3.19}$$

This last result allows us to derive in a straightforward manner the dual equations associated to the control problem, as we will show in the following Section.

3.1.2 The optimal control problem described by the Navier–Stokes equations

We apply the optimal control analysis based on the Lagrangian functional formalism given in Sec.s 2.1.1 and 2.1.2 for the unconstrained case to the drag minimization problem defined in Eq.(3.3).

According with Eq.s (3.8), (3.16) and (3.18), the associated Lagrangian functional (see Sec.2.1.2) is defined as:

$$\begin{aligned} \mathcal{L}(\mathbf{v}, p, \mathbf{w}, \mathbf{z}, q, \mathbf{r}, \mathbf{u}) \;:=\; & \hat{c}_D(\mathbf{w}) - \mathcal{A}(\mathbf{v}, p, \mathbf{z}, q) \\ & - \int_{\Gamma_D} \mathbf{w} \cdot \mathbf{z} \, d\Gamma - \int_{\Gamma_{SYM}} \mathbf{w} \cdot \hat{\mathbf{n}} \; \mathbf{z} \cdot \hat{\mathbf{n}} \, d\Gamma \\ & - \int_{\Gamma_D \backslash \Gamma_{CTRL}} (\mathbf{v} - \mathbf{v}_D) \cdot \mathbf{r} \, d\Gamma - \int_{\Gamma_{CTRL}} (\mathbf{v} - \mathbf{u}) \cdot \mathbf{r} \, d\Gamma \\ & - \int_{\Gamma_{SYM}} \mathbf{v} \cdot \hat{\mathbf{n}} \; \mathbf{r} \cdot \hat{\mathbf{n}} \, d\Gamma. \end{aligned} \tag{3.20}$$

By differentiating $\mathcal{L}(\cdot)$ w.r.t. $(\mathbf{z}, q, \mathbf{r})$ we obtain the primal equation in weak form reported in Eq.(3.8). Similarly, by differentiating $\mathcal{L}(\cdot)$ w.r.t. the primal variables $(\mathbf{v}, p, \mathbf{w})$, we obtain the dual equations. Their weak form reads:

$$\begin{gathered}
\text{find} \quad \mathbf{z} \in [H^1(\Omega)]^2, \quad q \in L^2(\Omega), \quad \mathbf{r} \in [H^{-\frac{1}{2}}(\Gamma_{DS})]^2 \quad : \\
a(\boldsymbol{\Theta}, \mathbf{z}) + b(q, \boldsymbol{\Theta}) + c'(\mathbf{v}, \boldsymbol{\Theta}, \mathbf{z}) + \int_{\Gamma_D} \boldsymbol{\Theta} \cdot \mathbf{r} \; d\Gamma + \int_{\Gamma_{SYM}} \boldsymbol{\Theta} \cdot \hat{\mathbf{n}} \; \mathbf{r} \cdot \hat{\mathbf{n}} \; d\Gamma = 0, \\
b(\vartheta, \mathbf{z}) = 0, \\
\int_{\Gamma_D} \boldsymbol{\Upsilon} \cdot \mathbf{z} \; d\Gamma + \int_{\Gamma_{SYM}} \boldsymbol{\Upsilon} \cdot \hat{\mathbf{n}} \; \mathbf{z} \cdot \hat{\mathbf{n}} \; d\Gamma = \hat{c}_D(\boldsymbol{\Upsilon}) \\
\forall \boldsymbol{\Theta} \in [H^1(\Omega)]^2, \quad \forall \vartheta \in L^2(\Omega), \quad \forall \boldsymbol{\Upsilon} \in [H^{-\frac{1}{2}}(\Gamma_{DS})]^2,
\end{gathered} \tag{3.21}$$

where, according with Eq.(3.7):

$$c'(\mathbf{v}, \boldsymbol{\Theta}, \mathbf{z}) := c(\boldsymbol{\Theta}, \mathbf{v}, \mathbf{z}) + c(\mathbf{v}, \boldsymbol{\Theta}, \mathbf{z}). \tag{3.22}$$

This dual problem corresponds to the linearized Navier–Stokes equations (Oseen equations) around $\mathbf{v}$ with the following boundary conditions:

$$\begin{cases}
\mathbf{z} = \mathbf{0} & \text{on } \Gamma_{IN}, \\
\mathbf{z} = \hat{\mathbf{v}}_\infty / (q_\infty d) & \text{on } \Gamma_{BODY}, \\
\mathbf{z} \cdot \hat{\mathbf{n}} = 0, \quad (\mathbb{T}(\mathbf{z}, q)\hat{\mathbf{n}}) \cdot \hat{\mathbf{t}} = 0 & \text{on } \Gamma_{SYM}, \\
\mathbb{T}(\mathbf{z}, q)\hat{\mathbf{n}} = \mathbf{0} & \text{on } \Gamma_{OUT}.
\end{cases} \tag{3.23}$$

Finally, by differentiating $\mathcal{L}(\cdot)$ with respect to the control variable $\mathbf{u}$, we have:

$$\int_{\Gamma_{CTRL}} \boldsymbol{\Psi} \cdot \mathbf{r} \; d\Gamma = 0 \qquad \forall \boldsymbol{\Psi} \in [H^{\frac{1}{2}}(\Gamma_{CTRL})]^2. \tag{3.24}$$

After considering the form (3.4) for $\mathbf{u}$, the previous equation reads:

$$\int_{\Gamma_{CTRL}} \psi \mathbf{g}(\mathrm{x}_\mathbf{1}) \cdot \mathbf{r} \; d\Gamma = 0 \qquad \forall \psi \in \mathbb{R}. \tag{3.25}$$

The sensitivity δU (see Eq.(2.7)) can be written in the following form:

$$\delta U = \int_{\Gamma_{CTRL}} \mathbf{g}(\mathrm{x}_\mathbf{1}) \cdot \mathbf{r} \; d\Gamma, \tag{3.26}$$

which is identically equal to zero at the optimum. Let us notice that the sensitivity δU depends on the dual stress $\mathbf{r}$; by similar arguments to those considered for the computation of the drag coefficient and by manipulating Eq.(3.21), Eq.(3.26) equivalently reads:

$$\delta U = -\mathcal{A}'(\mathbf{G}, \vartheta, \mathbf{z}, q, \mathbf{v}) \quad \forall \vartheta \in L^2(\Omega), \tag{3.27}$$

where:

$$\mathcal{A}'(\boldsymbol{\Theta}, \vartheta, \mathbf{z}, q, \mathbf{v}) := a(\boldsymbol{\Theta}, \mathbf{z}) + b(q, \boldsymbol{\Theta}) + b(\vartheta, \mathbf{z}) + c'(\boldsymbol{\Theta}, \mathbf{v}, \mathbf{z}), \tag{3.28}$$

being:

$$\mathbf{G} \in [H^1(\Omega)]^2 \quad \text{with} \quad \mathbf{G}_{|\Gamma_{CTRL}} = \mathbf{g}(\mathrm{x}_\mathbf{1}), \quad \mathbf{G}_{|\partial\Omega \backslash \Gamma_{CTRL}} = \mathbf{0}. \tag{3.29}$$

3.2 Numerical approximation

In order to solve the optimal control problem, we consider the Galerkin–Finite Element (FE) method for the numerical solution of the primal and dual equations (Navier–Stokes and Oseen type) and an iterative approach based on the Steepest–Descent method for the functional minimization.

3.2.1 The optimization iterative method

The optimization problem is solved by means of an iterative approach for optimal control problems described by PDEs, for which we refer the reader to the general algorithm discussed in Sec.2.3.2. In particular, we make use of the Steepest–Descent method introduced in Sec.2.3.1, with the relaxation parameter of Eq.(2.47) chosen as $\tau^k = \tau$.
Let us notice that for the optimal control problem under investigation, according with Eq.(3.27), the norm of sensitivity reads:

$$\|\delta u\|_{\mathcal{U}} = |\delta U| = |\mathcal{A}'(\mathbf{G}, \vartheta, \mathbf{z}, q, \mathbf{v})| \quad \forall \vartheta \in L^2(\Omega), \tag{3.30}$$

where $\mathbf{G}$ is chosen according with Eq.(3.29).

3.2.2 Numerical solution of primal and dual equations

Both the primal and dual equations enforce the Dirichlet boundary conditions in weak form, by means of Lagrangian multipliers, through the primal and dual stresses $\mathbf{w}$ and $\mathbf{r}$. However, for the numerical solution we can re–write these equations in a conventional manner introducing appropriate functional spaces and lifting terms. In this way, we omit to compute explicitly the primal and dual stresses $\mathbf{w}$ and $\mathbf{r}$.
The primal equations consist of the steady Navier–Stokes equations, while the dual ones correspond to a steady Oseen problem. In fact the dual equations are linear in both the dual variables (velocity and pressure), being the term $c'(\mathbf{v}, \mathbf{\Theta}, \mathbf{z})$ (3.22) linear w.r.t. the dual velocity $\mathbf{z}$. Let us notice that the dual equations depend on the primal variables uniquely through the term $c'(\mathbf{v}, \mathbf{\Theta}, \mathbf{z})$; for this reason the FE matrix corresponding to this term needs to be recomputed at each step of the optimization iterative procedure.
First, we recall briefly a method for the solution of the Stokes equations, which is employed to solve the dual equations and to initialize the primal Navier–Stokes equations. Then we consider the Newton method, with Uzawa preconditioning for the solution of the Navier–Stokes equations. Other and more efficient numerical methods for the solution of Stokes and Navier–Stokes equations, both steady and time dependent, can be used (see e.g. [65, 73, 79, 146, 153]).

Solution of steady Stokes equations

In order to satisfy the *inf–sup* (LBB) condition [153] we consider the $\mathbb{P}^2$–$\mathbb{P}^1$ continuous FE pair for the velocity and the pressure respectively (for both primal and dual ones). By considering a generic Stokes problem (see e.g. Eq.(3.1) without the non linear term) the corresponding approximated problem reads:

$$\begin{cases} A\mathbf{V} + B^T\mathbf{P} = \mathbf{F}, \\ B\mathbf{V} = \mathbf{0}, \end{cases} \tag{3.31}$$

where the matrices A and B corresponds to the variational forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ respectively (Eq.s (3.5) and (3.6)), $\mathbf{F}$ is the source term vector (including the lifting terms), while $\mathbf{V}$ and $\mathbf{P}$ are respectively the vectors of the unknown FE coefficients for the velocity and pressure. The problem is resolved by computing firstly the pressure vector $\mathbf{P}$ and then the velocity $\mathbf{V}$ vector:

$$BA^{-1}B^T\mathbf{P} = BA^{-1}\mathbf{F}, \tag{3.32}$$

$$A\mathbf{V} = \mathbf{F} - B^T\mathbf{P}. \tag{3.33}$$

Both systems admit unique solution, being the matrix A symmetric and positive definite and the rank of B maximum (due to the FE pair considered), that can be obtained by means of the Conjugate Gradient method (CG) [152], being the matrix $BA^{-1}B^T$ symmetric and positive definite.

Remark 3.1. *In the case of the Oseen problem for the dual equation, the matrix A arises from the variational forms $a(\cdot,\cdot)$ and $c'(\cdot,\cdot,\cdot)$ (Eq.s (3.5) and (3.22)), for which A is no longer symmetric; in this case for the solution of the linear systems (3.32) and (3.33) the GMRES method [152] has been considered.*

Solution of steady Navier–Stokes equations

The Navier–Stokes equations for an incompressible fluid with constant properties can be seen as a compact perturbation of the Stokes equations, with the addition of a non linear term in the velocity variable. The system can be solved by means of the Newton method, solving at each step of the iterative procedure an Oseen problem, which arises by linearizing the Navier–Stokes equations. These linearized equations at the generic step $k+1$, obtained from Eq.(3.1), read:

$$\begin{cases} -\nabla\cdot\mathbb{T}(\mathbf{v}^{(k+1)},p^{(k+1)}) + \mathbf{c}'(\mathbf{v}^{(k)},\mathbf{v}^{(k+1)}) + \mathbf{c}'(\mathbf{v}^{(k+1)},\mathbf{v}^{(k)}) = \mathbf{c}'(\mathbf{v}^{(k)},\mathbf{v}^{(k)}) & \text{in } \Omega, \\ \nabla\cdot\mathbf{v}^{(k+1)} = 0 & \text{in } \Omega, \\ B.C.s & \text{on } \partial\Omega, \end{cases} \tag{3.34}$$

where $\mathbf{v}^{(k)}$ is the velocity computed at the previous iterative step and $\mathbf{c}'(\cdot,\cdot)$ is defined as:

$$\mathbf{c}'(\mathbf{r},\mathbf{q}) := \rho(\mathbf{r}\cdot\nabla)\mathbf{q}. \tag{3.35}$$

The Newton algorithm reads:

1. solve the Stokes problem, computing the initial pressure and velocity field;

2. by considering the velocity computed at the previous step $\mathbf{v}^{(k)}$, solve the Oseen problem, corresponding to Eq.(3.34), obtaining $\mathbf{v}^{(k+1)}$ and $p^{(k+1)}$.

3. compute the appropriate norms of the incremental differences $\mathbf{v}^{(k+1)}-\mathbf{v}^{(k)}$ and $p^{(k+1)}-p^{(k)}$ and compare them with a prescribed tolerance;

4. if the stopping criterium is fulfilled, take $\mathbf{v}^{(k+1)}$ and $p^{(k+1)}$ as solutions of the Navier–Stokes equations; otherwise, set $\mathbf{v}^{(k)} = \mathbf{v}^{(k+1)}$, $p^{(k)} = p^{(k+1)}$ and return to point 2.

A preconditioning matrix can be used for the solution of the linear system (3.32) corresponding to the Stokes (or Oseen) equations, which are recursively solved in the Newton method. In particular, we consider the Uzawa preconditioning matrix D, which corresponds to the bilinear form $d(p, q) := \int_{\Omega} pq \, d\Omega$.

No stabilization terms for convection are used in the approximated Navier–Stokes equations, as we are interested in flows at low Reynolds numbers (inferior to 50) and we suppose to handle with sufficiently refined meshes. Moreover, as previously mentioned, a divergence–stable FE pair is considered, so that stabilization terms for the pressure do not need to be used.

Remark 3.2. *At each step of the optimization iterative procedure we solve the Navier–Stokes primal equations, for which an initialization by means of a Stokes problem occurs. In order to save computational costs, we can take as initial guess the primal velocity computed at the previous step of the optimization procedure, instead of computing it by means of a Stokes problem solver. The validity of this procedure is confirmed by numerical tests (see Sec.3.3). Other techniques for the reduction of the computational cost of the optimization procedure can be considered, such as, e.g., reduced order methods (see Chapter 7 and e.g. [78, 80]).*

3.3 Numerical results

We present some numerical results concerning the drag minimization control problem outlined in Sec.3.1, carried out by means of the FE library *FreeFEM++* [191].

In particular, referring to Eq.s (3.1) and (3.3) and Fig.3.2, we chose $\rho = 1$, $V_\infty = 1$ and $d = 0.1$ (body length), for which, defining the Reynolds number:

$$\mathbb{R}e := \frac{\rho V_\infty d}{\mu}, \tag{3.36}$$

the dynamic viscosity coefficient μ can be deduced. At the initial step of the optimization procedure, we initialize $\mathbf{u} = \mathbf{0}$, for which, according with Eq.(3.4), $U = 0$ and $U_{av} = \frac{2}{3}U = 0$, being U_{av} the average value of $\mathbf{u}$ on Γ_{CTRL}. We take the relaxation parameter of the Steepest–Descent method $\tau = 0.5$ and we consider a tolerance $Tol_{IT} = 1/1000|\delta U^0|$ for the stopping criterium of the optimization method, where the superscript 0 means at the initial step and δU is reported in Eq.(3.30). Moreover, we provide the computing times (CPU times) required for the solution of the optimal control problems; the computations have been carried out on an AMD Athlon 1.8 GHz processor, with 256 KB of memory cache and 1 GB of RAM.

Test 1

Referring to Fig.3.2 we consider the case for which $t = d = 0.1$. For the computations we use the mesh reported in Fig.3.3 with 2955 triangular elements (and 1565 vertices); in the sequel we provide also a comparison of the results for three different meshes.

Firstly we consider the case for which $\mathbb{R}e = 10$. In Fig.3.4 we report the velocity field and pressure, solution of the primal Navier–Stokes equations, obtained for $U_{av} = 0$; in Fig.3.5, the streamlines for the primal velocity field are presented. In Fig.3.6, we show the dual velocity and pressure, computed at the initial step of the optimization iterative procedure (corresponding to $U_{av} = 0$). At the end of the optimization procedure, we obtain the velocity field and the pressure reported in Fig.3.7; in Fig.3.8 the streamlines of the optimal velocity

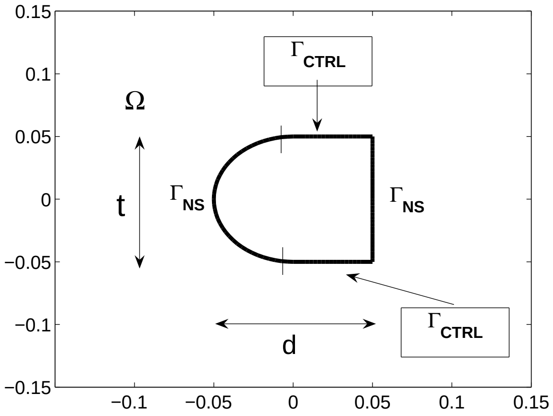

Figure 3.2: Zoom around the body, geometry and boundary conditions.

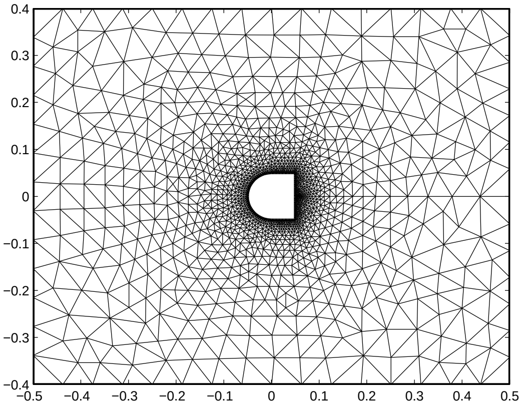

Figure 3.3: Test 1. Mesh with 2955 elements.

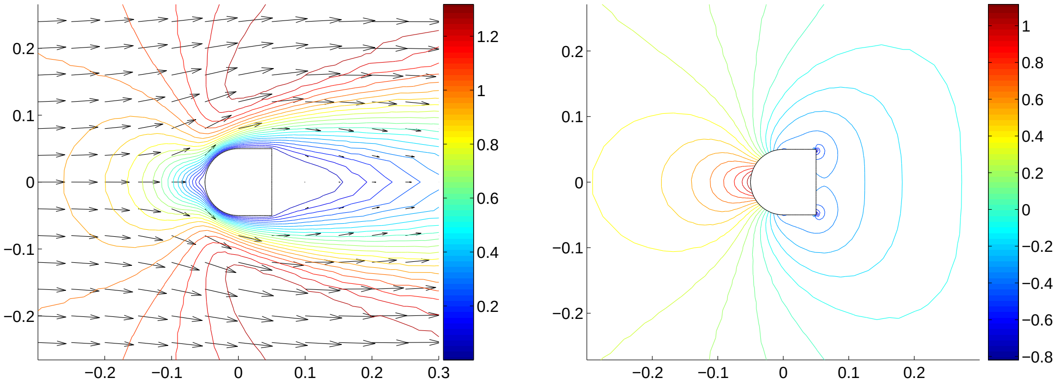

Figure 3.4: Test 1. Zoom of velocity field (left) and pressure (right) for the initial primal flow ($\mathbb{R}e = 10$).

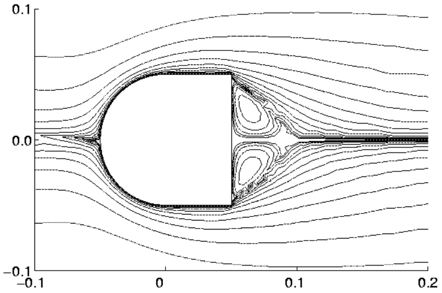

Figure 3.5: Test 1. Streamlines for the initial primal flow (zoom) ($\mathbb{R}e = 10$).

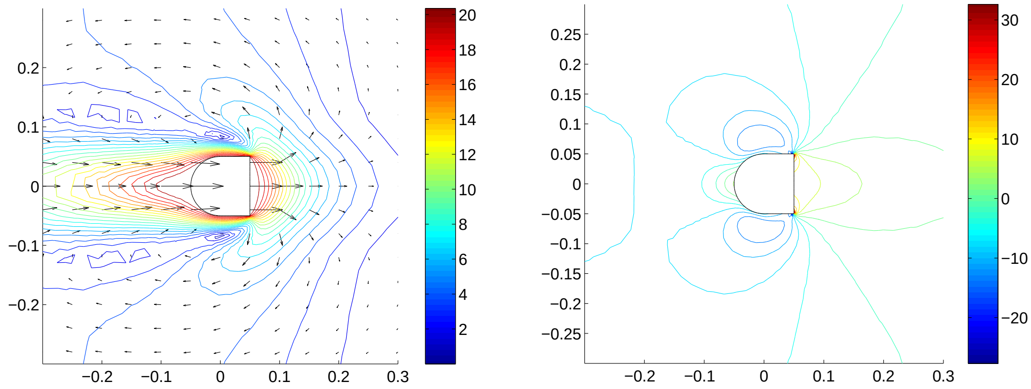

Figure 3.6: Test 1. Zoom of dual velocity field (left) and dual pressure (right) for the initial dual flow ($\mathbb{R}e = 10$).

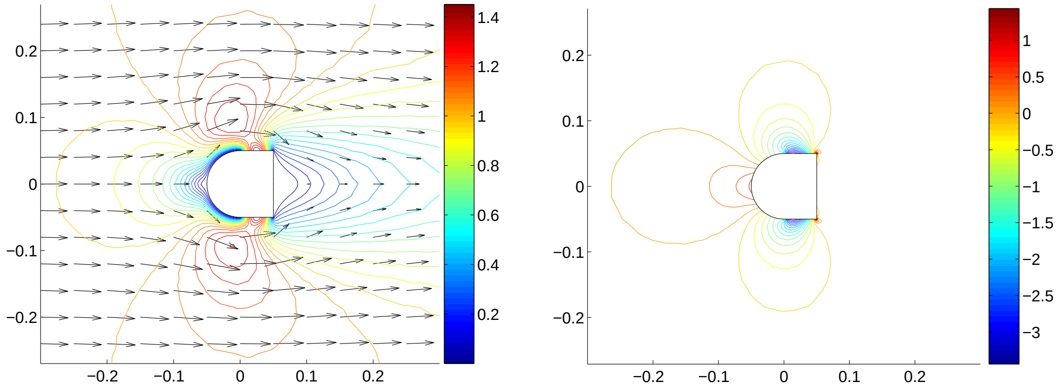

Figure 3.7: Test 1. Zoom of velocity field (left) and pressure (right) for the optimal primal flow ($\mathbb{R}e = 10$).

field show the effects of the optimization. In particular, the initial drag coefficient $c_D = 3.8332$, corresponding to $U_{av} = 0$, is reduced, by means of this procedure, to the optimal one, which is $c_D = 3.3066$, obtained for $U_{av} = -0.96676$ (this corresponds to an aspiration of the flow across the control boundaries of the body). In Fig.3.9 we report the behavior of the drag coefficient c_D and the sensitivity $|\delta U|$ (normalized w.r.t. $|\delta U^0|$) versus the number of iterations of the optimization procedure, for which the convergence occurs in 24 iterations. Let us observe

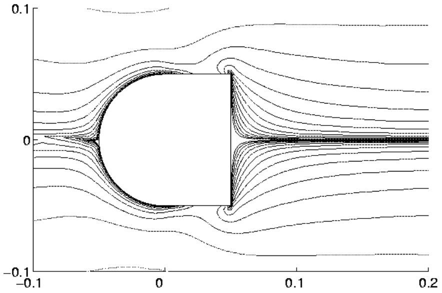

Figure 3.8: Test 1. Streamlines for the optimal primal flow ($\mathbb{R}e = 10$).

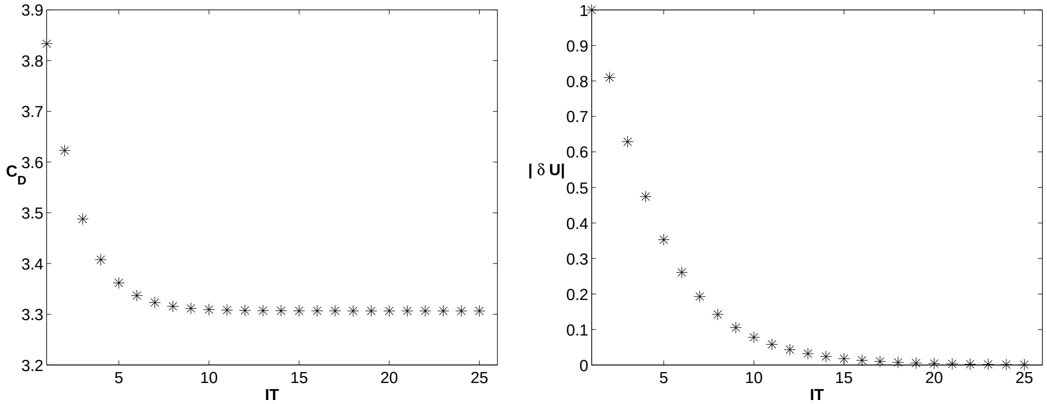

Figure 3.9: Test 1. Drag coefficient (left) and sensitivity (right), normalized with δU^0, vs. number of iterations of the optimization procedure ($\mathbb{R}e = 10$).

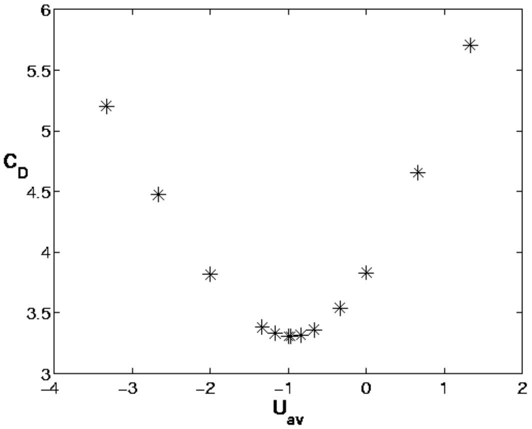

Figure 3.10: Test 1. Drag coefficient for different values of U_{av} (average value of control variable) ($\mathbb{R}e = 10$).

that the convergence of c_D to the optimal value is quick; e.g., after $k = 11$ optimization steps, we have that $|\delta U^k|/|\delta U^0| < 1/50$ and the relative percent errors on c_D and U_{av} (optimal), computed w.r.t. the optimal quantities, are respectively 0.03% and 3.8%. In Fig.3.10 we

Mesh	$\mathbb{R}e$	c_D ($\mathbf{u}=\mathbf{0}$)	c_D (opt.)	U_{av} (opt.)	$\sharp$ it.	CPU [s]
2955 el.	5	5.8036	4.7390	-1.7671	44	2295
	10	3.8332	3.3066	-0.96676	24	1761
	20	2.6710	2.3323	-0.57954	12	1316
	50	1.7737	1.5185	-0.32516	3	1050
4625 el.	5	5.8045	4.7401	-1.7679	44	4060
	10	3.8337	3.3074	-0.96700	24	2700
	20	2.6712	2.3328	-0.57950	12	2002
	50	1.7731	1.5186	-0.32489	3	1717
6467 el.	5	5.8046	4.7401	-1.7682	44	6405
	10	3.8338	3.3074	-0.96715	24	4446
	20	2.6712	2.3328	-0.57961	12	3406
	50	1.7732	1.5182	-0.32507	4	2935

Table 3.1: Test 1. Drag coefficients, minimum drag coefficients, average values (optimal) of control function, number of iterations of the optimization procedure and CPU times for different $\mathbb{R}e$ values; comparison of results for meshes with 2955 (see Fig.3.3), 4625 and 6467 elements.

show the values of c_D obtained by solving the Navier–Stokes primal equations for different values of U_{av}; this confirms that the local minimum value of c_D corresponds to that obtained by means of the optimization procedure, for which $U_{av} = -0.96676$.

In Table 3.1, we report the values of c_D (with $U_{av} = 0$) for different Reynolds numbers $\mathbb{R}e$, the corresponding optimal drag coefficients, the optimal control functions U_{av}, the number of iterations of the optimization procedure and the corresponding CPU time. In particular, we find that, when $\mathbb{R}e$ increases, the modulus of the optimal U_{av} decreases, as the number of iterations of the optimization procedure. However, let us observe that, as $\mathbb{R}e$ increases, the computational cost associated with the solution of the Navier–Stokes equations raises: for this reason the cost of the complete optimization procedure could be great, even if the number of iterations is limited. Moreover, in Table 3.1, we provide a comparison of the results for different triangular meshes with 2955 (see Fig.3.3), 4625 and 6467 elements. We show that, for the $\mathbb{R}e$ numbers considered, the sensitivity of the results w.r.t. the mesh is limited; this shows that the mesh with 2955 elements (Fig.3.3) is well suited for the resolution of this optimal control problem.

In Sec.3.2.2 we have anticipated that no stabilization terms for convection have been used for the FE approximation of the Navier–Stokes equations. Numerical tests confirm the validity of this choice, for the meshes and the $\mathbb{R}e$ numbers considered; for example, by using the mesh with 2955 elements and $\mathbb{R}e = 50$, the local Reynolds number $\mathbb{R}e_K := h/d\ \mathbb{R}e$, being h the diameter of the local mesh elements, is about $\mathbb{R}e_K \lesssim 1$ near the body.

In Remark 3.2 we have discussed a strategy for the reduction of the computational cost of the optimization procedure, for which we initialize the linearized Navier–Stokes primal equations with the velocity field computed at previous step of the optimization iterative procedure, instead of using a Stokes solver at each optimization step. Numerical tests, referring to the problem under investigation, outline that cost savings can be reached of about 37% and 25%, for $\mathbb{R}e = 5$ and $\mathbb{R}e = 10$, respectively (for the mesh with 2955 elements).

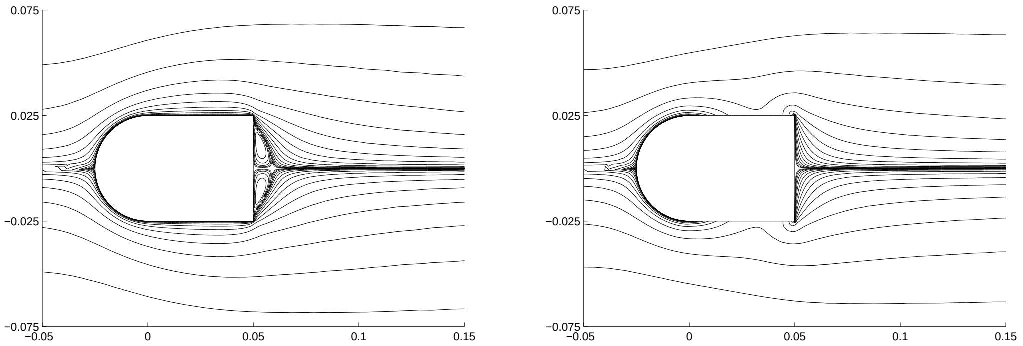

Figure 3.11: Test 2. Streamlines for the initial (left) and optimal (right) primal flows ($\mathbb{R}e = 10$).

$\mathbb{R}e$	c_D ($\mathbf{u} = \mathbf{0}$)	c_D (opt.)	U_{av} (opt.)	$\sharp$ it.	CPU $[s]$
5	3.9966	3.7802	-0.91485	56	3082
10	2.5624	2.5084	-0.37184	37	2466
20	1.7201	1.6991	-0.18668	21	2060
50	1.0759	1.0577	-0.11335	7	1517

Table 3.2: Test 2. Drag coefficients, minimum drag coefficients, average values (optimal) of control function, number of iterations of the optimization procedure and CPU times for different $\mathbb{R}e$ values.

Test 2

We consider the case for which $t = d/2 = 0.05$ (see Fig.3.2), i.e. the thickness of this body is the half of that considered for Test 1. For the computations we use a mesh similar to that reported in Fig.3.3 with 2987 triangular elements (and 1573 vertices).
In Fig.3.11 we show the streamlines of the initial (left) and optimal (right) flows for $\mathbb{R}e = 10$, which highlight the effect of the aspiration of the flow across the control boundaries. In Table 3.2 we report the results obtained for this test case, which confirm the considerations made for Test 1; more over, we notice that, for this body, the effects of the boundary layer aspiration on the drag coefficient is less evident than for the body of Test 1.

The Stokes case

If we define the drag minimization control problem for the Stokes equations, the drag coefficient does not admit a positive minimum, but tends to $-\infty$; this is due to the linearity in the velocity of the Stokes equations and the drag coefficient.
In Fig.3.12, we report, referring to Test 1, the drag coefficient (for a prescribed μ) computed for different values of U_{av}, for which the linear dependence of c_D on U_{av} is pointed out.

3.4 Concluding remarks

In this Chapter we have studied an optimal control problem for the drag reduction of a $2D$ body in relative motion with an incompressible fluid with constant properties. In particular,

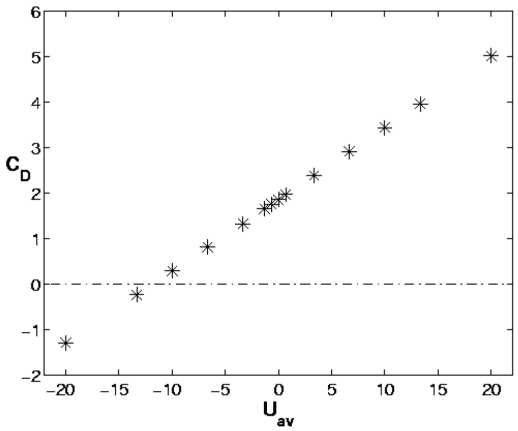

Figure 3.12: Stokes case. Drag coefficient vs. average value of control variable.

we have considered as control function the Dirichlet boundary conditions defined on a part of the boundary of the body itself. We have used the Lagrangian functional approach for the resolution of the optimal control problem, together with the Lagrangian multiplier method for the treatment of the Dirichlet boundary conditions. This has allowed a straightforward determination of the first order necessary conditions, without introducing lifting terms. We have proved the effectiveness of the procedure on numerical tests by computing the optimal flow for which the drag coefficient is minimum and comparing the results for different $\mathbb{R}e$ numbers and configurations.

Chapter 4

Goal–Oriented Analysis for Galerkin Methods

As anticipated in Chapter 1 for environmental problems, and more in general for a wide range of applications, one could be interested in evaluating an *output functional* which represents a meaningful quantity, such as an average concentration, stress, flux or displacement, rather than the entire solution of a problem described by a PDE. At the same time, the computation of this output functional should be sufficiently accurate in order to keep into account for prescribed tolerances on the error. In fact, it is in general appropriate to provide together with the values of the required output, an estimate of the error associated with its computation. This justifies the employment of a *goal–oriented* analysis, for which the final outcome is an *a posteriori* error estimate associated with any computed outputs; with this aim, an auxiliary problem, the so–called dual (or adjoint) problem is introduced. The use of a dual approach for a posteriori error estimates has been firstly proposed in [12, 13, 14] for the post–processing of output functionals for elliptic PDEs and then extended to more general problems [47, 48]; similar approaches have been considered in [112, 131, 132, 134, 138]. In [16, 21, 22, 154] this approach is further developed in a feedback method for the control of the output error and mesh adaptivity, which is commonly indicated as the *Dual Weighted Residual* (DWR) method; for an algebraic approach, see e.g. [61, 62, 63, 177]. The Lagrangian functional formalism is conveniently adopted in order to provide the goal–oriented analysis in a general context. As anticipated, the DWR method and other duality based approaches are often used not only to evaluate the errors associated with the computation of output functionals, but also in order to build suitable meshes by means of adaptive procedures led by a posteriori error estimates; see Chapter 5 and e.g. [16, 22, 43, 52, 54, 86].

The goal–oriented analysis can be conveniently extended to the case of optimal control problems described by PDEs (see Chapter 2 and e.g. [107]), for which the output functional is replaced by the cost functional; see [18, 19, 20, 22, 23, 24, 44, 119, 141, 151, 182]. Also in this case, a posteriori error estimates are provided and used to evaluate the approximation error associated with the numerical solution of the optimal control problem; mesh adaptivity procedure can be adopted on the basis of the error estimates.

Several numerical methods could be adopted in order to solve the problems described by PDEs and the corresponding output functionals.

In this Chapter we provide a general overview of the goal–oriented analysis for the approximation based on *Galerkin* methods such as the Finite Element method, the Spectral methods or the Reduced Basis method. Moreover, we briefly discuss the effects of the *"differentiate–then–discretize"* and *"discretize–then–differentiate"* approaches which underly the approximation strategy used for the goal–oriented analysis ([19, 23, 38, 151]). Similar considerations are discussed also for the case of optimal control problems.

The Chapter is organized as follows. In Sec.4.1 we discuss the goal–oriented analysis by introducing the Lagrangian functional approach, for which we put into evidence the aspects related to the "differentiate–then–discretize" and "discretize–then–differentiate" approaches; the general case and problems described by elliptic and parabolic PDEs are considered. Results for the linear case are also pointed out. In Sec.4.2 we extend the goal–oriented analysis to optimal control problems; with this aim, we provide the results for both the unconstrained and the constrained cases.

4.1 The goal–oriented analysis for output functionals

In this Section we recall the goal–oriented analysis and the basic ideas of the a posteriori error estimates based on the Lagrangian functional. After having introduced the general case, we briefly consider two examples regarding elliptic and parabolic PDEs and we discuss both the "differentiate–then–discretize" and "discretize–then–differentiate" approaches. The goal–oriented analysis is provided in the general case and then pointed out for linear problems.

4.1.1 The general case

By recalling the standard notation of Chapter 2, we introduce in an abstract setting the well–posed PDE $r(v) = 0$ in $\mathcal{W}$, with $v \in \mathcal{V}$, endowed with appropriate initial and boundary conditions. The goal consists in evaluating the output functional $s = L(v)$, being $L : \mathcal{V} \to \mathbb{R}$ a continuous functional. In analogy with the optimal control case of Sec.2.1.1, we define the Lagrangian functional as:

$$\mathcal{L}(\mathbf{x}) = \mathcal{L}(v, z) := L(v) + \langle z, r(v) \rangle_{\mathcal{W}^*,\mathcal{W}}, \tag{4.1}$$

where $z \in \mathcal{W}^*$ is the dual variable and $\mathbf{x} := (v, z) \in \mathcal{X}$, being $\mathcal{X} := \mathcal{V} \times \mathcal{W}^*$, s.t. $\mathcal{L} : \mathcal{X} \to \mathbb{R}$. The problem can be analyzed analogously to the optimal control problem described in Sec.2.1.1; for this reason, from Theorem 2.1, we obtain the following first order necessary conditions:

$$\mathcal{L}_v(v, z) = L_v(v) + r_v^*(v)[z] = 0 \qquad \text{in } \mathcal{V}^*, \tag{4.2}$$

$$\mathcal{L}_z(v, z) = r(v) = 0 \qquad \text{in } \mathcal{W}, \tag{4.3}$$

where differentiation is in Fréchet sense, $\mathbf{x} := (v, z) \in \mathcal{X}$ is the local minimum of the functional $\mathcal{L}(\cdot)$ and Eq.(4.3) is the dual equation. Let us notice that we do not use the notation $\mathbf{x}^{**}$ for the local minimum of $\mathcal{L}(\cdot)$, which is reserved to the minimum of optimal control problems.

4.1.2 The case of steady PDEs

In analogy and with the same notation of Sec.2.1.2, we consider the following elliptic PDE in weak form:

$$\text{find } v \in \mathcal{V} \quad : \quad a(v)(\phi) = F(\phi) \qquad \forall \phi \in \mathcal{V}, \tag{4.4}$$

where $\mathcal{V}$ is an Hilbert space and $\mathcal{W}^* \equiv \mathcal{V}$, s.t. $\mathcal{X} = \mathcal{V} \times \mathcal{V}$. We assume the semilinear form $a(\cdot)(\cdot)$ and the functional $L(\cdot)$ three times differentiable, while $F(\cdot)$ is a continuous, linear and infinitely differentiable functional. From Eq.s (4.1) and (4.4) it is immediate to define the Lagrangian functional:

$$\mathcal{L}(\mathbf{x}) = \mathcal{L}(v, z) := L(v) + F(z) - a(v)(z), \tag{4.5}$$

and to deduce the dual equation:

$$\text{find } z \in \mathcal{V} \quad : \quad a_v(v)(z, \vartheta) = L_v(v)(\vartheta) \qquad \forall \vartheta \in \mathcal{V}, \tag{4.6}$$

where $a_v(\cdot)(\cdot, \cdot)$ and $L_v(\cdot)(\cdot)$ are the differentials of $a(\cdot)(\cdot)$ and $L(\cdot)$ w.r.t. v, respectively. In view of the goal–oriented analysis which we provide in Sec.4.1.5, it is useful to define, from Eq.s (4.4) and (4.6), the following primal and dual *residuals* in weak form:

$$R^{pr}(\mathbf{x})(\phi) = R^{pr}(v)(\phi) := F(\phi) - a(v)(\phi) \qquad \forall \phi \in \mathcal{V}, \tag{4.7}$$

$$R^{du}(\mathbf{x})(\vartheta) = R^{du}(v, z)(\vartheta) := L_v(v)(\vartheta) - a_v(v)(z, \vartheta) \qquad \forall \vartheta \in \mathcal{V}, \tag{4.8}$$

s.t. $R^{pr} : \mathcal{X} \to \mathbb{R}$ and $R^{du} : \mathcal{X} \to \mathbb{R}$. We observe that Eq.s (4.4) and (4.6) can be conveniently rewritten in terms of the primal and dual residuals, respectively:

$$\text{find } v \in \mathcal{V} \quad : \quad R^{pr}(\mathbf{x})(\phi) = 0 \qquad \forall \phi \in \mathcal{V}, \tag{4.9}$$

$$\text{find } z \in \mathcal{V} \quad : \quad R^{du}(\mathbf{x})(\vartheta) = 0 \qquad \forall \vartheta \in \mathcal{V}. \tag{4.10}$$

Finally, we introduce the following notation:

$$\mathcal{L}'(\mathbf{x}_1)(\boldsymbol{\phi}) := R^{pr}(\mathbf{x}_1)(\phi) + R^{du}(\mathbf{x}_1)(\vartheta) \qquad \forall \mathbf{x}_1, \forall \boldsymbol{\phi} = (\phi, \vartheta) \in \mathcal{X}, \tag{4.11}$$

for which the first order necessary conditions (4.2)–(4.3) read:

$$\text{find } \mathbf{x} \in \mathcal{X} \quad : \quad \mathcal{L}'(\mathbf{x})(\boldsymbol{\phi}) = 0 \qquad \forall \boldsymbol{\phi} \in \mathcal{X}. \tag{4.12}$$

4.1.3 The case of unsteady PDEs

Similarly to the steady case, by recalling Sec.2.1.3, we provide the following parabolic PDE in weak form:

$$\begin{aligned} &\text{find } v \in \mathcal{V} \quad : \quad m\left(\frac{\partial v}{\partial t}, \phi\right) + a(v)(\phi) = F(\phi) \qquad \forall \phi \in \mathcal{V},\ t \in (0, T), \\ &\text{with } v(0) = v_0, \end{aligned} \tag{4.13}$$

under the same hypotheses of Sec.s 2.1.3 and 4.1.2 with $m(\cdot, \cdot)$ three times differentiable. The following two output functionals $L(v)$ could be considered:

$$L(v) = \int_0^T l(v)\ dt, \tag{4.14}$$

$$L(v) = l(v(T)), \tag{4.15}$$

being $l(\cdot)$ a continuous functional $l : \mathcal{V} \to \mathbb{R}$. Once again, from Eq.s (4.1) and (4.13) we define the Lagrangian functional:

$$\begin{aligned} \mathcal{L}(\mathbf{x}) = \mathcal{L}(v,z) \quad = \quad & L(v) + \int_0^T F(z)\ dt - \int_0^T a(v)(z)\ dt - \int_0^T m\left(\frac{\partial v}{\partial t}, z\right)\ dt \\ & -m\left(v(0) - v_0, z(0)\right). \end{aligned} \tag{4.16}$$

By differentiating $\mathcal{L}(\mathbf{x})$ w.r.t v, by integrating by parts in time and by recalling the output functionals (4.14) and (4.15), we obtain the corresponding dual equations in weak form, respectively:

$$\begin{aligned} & \text{find } z \in \mathcal{V} \quad : \quad -m\left(\vartheta, \frac{\partial z}{\partial t}\right) + a_v(v)(z,\vartheta) = l(v)(\vartheta) \qquad \forall\vartheta \in \mathcal{V},\ t \in (0,T), \\ & \text{with } z(T) = 0, \end{aligned} \tag{4.17}$$

$$\begin{aligned} & \text{find } z \in \mathcal{V} \quad : \quad -m\left(\vartheta, \frac{\partial z}{\partial t}\right) + a_v(v)(z,\vartheta) = 0 \qquad \forall\vartheta \in \mathcal{V},\ t \in (0,T), \\ & \text{with } m(z(T),\vartheta(T)) = l_v(v(T))(\vartheta(T)), \end{aligned} \tag{4.18}$$

where the final condition for the dual variable is given in weak form.
Similarly to the steady case, we define from Eq.(4.13) the primal residual $R^{pr} : \mathcal{X} \to \mathbb{R}$:

$$R^{pr}(\mathbf{x})(\phi) = R^{pr}(v)(\phi) := \int_0^T \left(F(\phi) - a(v)(\phi) - m\left(\frac{\partial v}{\partial t}, \phi\right)\right)\ dt \qquad \forall\phi \in \mathcal{V}, \tag{4.19}$$

and from Eq.s (4.17) and (4.18) the dual residuals $R^{du} : \mathcal{X} \to \mathbb{R}$, respectively:

$$R^{du}(\mathbf{x})(\vartheta) = R^{du}(v,z)(\vartheta) := \int_0^T \left(l_v(v)(\vartheta) - a_v(v)(z,\vartheta) + m\left(\vartheta, \frac{\partial z}{\partial t}\right)\right)\ dt \qquad \forall\vartheta \in \mathcal{V}, \tag{4.20}$$

$$R^{du}(\mathbf{x})(\vartheta) = R^{du}(v,z)(\vartheta) := \int_0^T \left(-a_v(v)(z,\vartheta) + m\left(\vartheta, \frac{\partial z}{\partial t}\right)\right)\ dt \qquad \forall\vartheta \in \mathcal{V}. \tag{4.21}$$

It follows that Eq.s (4.13), (4.17) and (4.18) read, respectively:

$$\begin{aligned} & \text{find } v \in \mathcal{V} \quad : \quad R^{pr}(\mathbf{x})(\phi) = 0 \qquad \forall\phi \in \mathcal{V}, \\ & \text{with } v(0) = v_0, \end{aligned} \tag{4.22}$$

$$\begin{aligned} & \text{find } z \in \mathcal{V} \quad : \quad R^{du}(\mathbf{x})(\vartheta) = 0 \qquad \forall\vartheta \in \mathcal{V}, \\ & \text{with } z(T) = 0, \end{aligned} \tag{4.23}$$

$$\begin{aligned} & \text{find } z \in \mathcal{V} \quad : \quad R^{du}(\mathbf{x})(\vartheta) = 0 \qquad \forall\vartheta \in \mathcal{V}, \\ & \text{with } m(z(T),\vartheta(T)) = l_v(v(T))(\vartheta(T)). \end{aligned} \tag{4.24}$$

Finally, in analogy with Eq.(4.11), we define, depending on the output functional (4.14) or (4.15), respectively:

$$\begin{aligned} \mathcal{L}'(\mathbf{x}_1)(\boldsymbol{\phi}) \quad := \quad & R^{pr}(\mathbf{x}_1)(\phi) + m(v_1(0) - v_0, \phi(0)) \\ & + R^{du}(\mathbf{x}_1)(\vartheta) + m(z_1(T), \vartheta(T)) \\ & \qquad \forall\mathbf{x}_1 = (v_1, z_1), \forall\boldsymbol{\phi} = (\phi, \vartheta) \in \mathcal{X}, \end{aligned} \tag{4.25}$$

$$\begin{aligned}\mathcal{L}'(\mathbf{x}_1)(\boldsymbol{\phi}) \quad := \quad & R^{pr}(\mathbf{x}_1)(\phi) + m(v_1(0) - v_0, \phi(0)) \\ & + R^{du}(\mathbf{x}_1)(\vartheta) + m(z_1(T), \vartheta(T)) - l_v(v_1(T))(\vartheta(T)) \qquad (4.26) \\ & \qquad \forall \mathbf{x}_1 = (v_1, z_1), \forall \boldsymbol{\phi} = (\phi, \vartheta) \in \mathcal{X}.\end{aligned}$$

In both the cases, first order necessary conditions (4.2)–(4.3) can be written as Eq.(4.12).

4.1.4 The “differentiate–then–discretize” and “discretize–then–differentiate” approaches

Let us now introduce the approximation based on a Galerkin method, such as the Finite Element, Spectral or Reduced Basis one. We recall that, for Galerkin methods, the solution of PDEs is searched in a subspace of the space $\mathcal{V}$, let say $\mathcal{V}_h$, s.t. $\mathcal{V}_h \subset \mathcal{V}$ [152, 153].

We have two possibilities for introducing the approximation in the problem and in the Lagrangian functional framework. The first approach, called *“differentiate–then–discretize”*, consists in defining the Lagrangian functional on the basis of the continuous primal equation and then to approximate the primal and dual equations in weak form, which corresponds to the continuous first order necessary conditions. The second possibility, called *“discretize–then–differentiate”*, introduces the approximation on the primal equation or directly on the Lagrangian functional, then the approximated dual equation is deduced on the basis of the differentiation of the approximated Lagrangian functional. These two approaches lead in general to different approximated dual problems; see [19, 38, 44, 151].

In order to better highlight these approaches, we consider, for example, the case of steady PDEs of Sec.4.1.2, for which the continuous primal and dual equations are reported in Eq.s (4.4) and (4.6), respectively.

Let us consider firstly the “differentiate–then–discretize” approach. On this basis, we introduce the approximation on the continuous primal and dual equations in weak form:

$$\text{find } v_h \in \mathcal{V}_h \ : \ a(v_h)(\phi_h) + d_h^{pr}(v_h)(\phi_h) = F(\phi_h) \qquad \forall \phi_h \in \mathcal{V}_h, \tag{4.27}$$

$$\text{find } z_h \in \mathcal{V}_h \ : \ a_v(v_h)(z_h, \vartheta_h) + d_h^{du}(z_h)(\vartheta_h) = L_v(v_h)(\vartheta_h) \qquad \forall \vartheta_h \in \mathcal{V}_h, \tag{4.28}$$

where $d_h^{pr}(\cdot)(\cdot)$ and $d_h^{du}(\cdot)(\cdot)$ are suitably semilinear forms taking into account for the effects of approximation on the original continuous forms $a(\cdot)(\cdot)$, $a_v(\cdot)(\cdot,\cdot)$ and functionals $F(\cdot)$, $L_v(\cdot)$. This is e.g. the case of the Galerkin–FE method with stabilization, for which the original weak form is modified by introducing suitable terms, in this case $d_h^{pr}(\cdot)(\cdot)$ and $d_h^{du}(\cdot)(\cdot)$. However, in general, the approximated dual equation does not represent a “true” duality with the approximated primal problem, being the approximation introduced independently of the continuous first order necessary conditions. For this reason, this approach is adopted together with “efficient” approximations, as strongly consistent stabilizations [152, 153]. However, when a posteriori error estimates are required, the goal–oriented analysis is not always straightforward, especially in presence of nonlinearities.

Conversely, in the case of the “discretize–then–differentiate” approach, we introduce first of all the approximated primal equation (4.27) and then the approximate Lagrangian functional, say $\mathcal{L}_h(\cdot)$:

$$\mathcal{L}_h(\mathbf{x}_h) = \mathcal{L}_h(v_h, z_h) := L(v_h) + F(z_h) - a(v_h)(z_h) - d_h^{pr}(v_h)(z_h). \tag{4.29}$$

We notice that $\mathcal{L}_h(\mathbf{x}_h) = \mathcal{L}(\mathbf{x}_h) - d_h^{pr}(v_h)(z_h)$, with $\mathcal{L}_h : \mathcal{X}_h \to \mathbb{R}$ and $\mathcal{X}_h := \mathcal{V}_h \times \mathcal{V}_h \subset \mathcal{X}$. By differentiating $\mathcal{L}_h$ w.r.t. v_h we obtain the approximated dual problem:

$$\text{find } z_h \in \mathcal{V}_h \;:\; a_v(v_h)(z_h, \vartheta_h) + d_{h,v}^{pr}(v_h)(z_h, \vartheta_h) = L_v(v_h)(\vartheta_h) \qquad \forall \vartheta_h \in \mathcal{V}_h, \tag{4.30}$$

which, due to $d_{h,v}^{pr}(v_h)(z_h, \cdot) \neq d_h^{du}(z_h)(\cdot)$, differs from the dual problem (4.28) obtained with the "differentiate–then–discretize" approach; with $d_{h,v}^{pr}(\cdot)(\cdot,\cdot)$ we indicate the differential of $d_h^{pr}(\cdot)(\cdot)$ w.r.t. v. With this approach the approximated dual equation represents a "true" duality with the approximated primal one, thus allowing a "coherent" and straightforward goal–oriented analysis, being from Eq.(4.29):

$$\mathcal{L}_h(\mathbf{x}_h) = \mathcal{L}(\mathbf{x}_h) - d_h^{pr}(v_h)(z_h). \tag{4.31}$$

For this reason, in this work we consider the "discretize–then–differentiate" approach.

Remark 4.1. *We remark that the Galerkin approximation is often introduced without modifying the weak forms of the PDEs and the Lagrangian functional, for which in the "differentiate–then–discretize" case $d_h^{pr}(\cdot)(\cdot) = 0$ and $d_h^{du}(\cdot)(\cdot) = 0$, while in the "discretize–then–differentiate" one $d_h^{pr}(\cdot)(\cdot) = 0$, and hence $d_{h,v}^{pr}(\cdot)(\cdot,\cdot) = 0$. In this case, both the approaches are in practise equivalent and the approximated dual equations coincide; in fact, first order necessary conditions (4.2)–(4.3) read, from Eq.(4.12):*

$$\textit{find } \mathbf{x}_h \in \mathcal{X}_h \quad : \quad \mathcal{L}'(\mathbf{x}_h)(\boldsymbol{\phi}_h) = 0 \qquad \forall \boldsymbol{\phi} \in \mathcal{X}_h. \tag{4.32}$$

4.1.5 The goal–oriented analysis

In this Section we provide the goal–oriented analysis for both the general and the linear cases and we discuss the basic ideas of a posteriori error estimates. Fur further deepening we refer the reader to [16, 19, 21, 22].

First of all, let us recall that we are interested in estimating the error due to approximation of the output functional $s = L(v)$, i.e. $s - s_h$, being $s_h := L(v_h)$. Moreover, we define the error on the primal and dual variables as $\mathbf{e}_h := \mathbf{x} - \mathbf{x}_h$, being $\mathbf{x} = (v, z) \in \mathcal{X}$ and $\mathbf{x}_h = (v_h, z_h) \in \mathcal{X}_h$, with $e_h^{pr} := v - v_h$ and $e_h^{du} := z - z_h$.

Proposition 4.1. *From Eq.s (4.1), (4.12) and (4.32) and under the hypotheses of Remark 4.1, if $\mathcal{L}(\cdot)$ is three times differentiable, the following result holds:*

$$s - s_h = \frac{1}{2}\mathcal{L}'(\mathbf{x}_h)(\mathbf{e}_h) + \Delta R, \tag{4.33}$$

with:

$$\Delta R := \frac{1}{2}\int_0^1 \mathcal{L}'''\left(\mathbf{x}_h + \varepsilon \mathbf{e}_h\right)(\mathbf{e}_h, \mathbf{e}_h, \mathbf{e}_h)\varepsilon(\varepsilon - 1)\; d\varepsilon. \tag{4.34}$$

From Eq.s (4.11), (4.25) and (4.26), Eq.(4.33) reads:

$$s - s_h = \frac{1}{2}\left(R^{pr}(\mathbf{x}_h)(e_h^{du}) + R^{du}(\mathbf{x}_h)(e_h^{pr})\right) + \Delta R, \tag{4.35}$$

which, if $\mathcal{L}(\cdot)$ is at most quadratic*, is:*

$$s - s_h = \frac{1}{2}\left(R^{pr}(\mathbf{x}_h)(e_h^{du}) + R^{du}(\mathbf{x}_h)(e_h^{pr})\right). \tag{4.36}$$

Finally, if $\mathcal{L}(\cdot)$ is linear*, we have:*

$$s - s_h = R^{pr}(\mathbf{x}_h)(e_h^{du}), \tag{4.37}$$

being in this case $R^{du}(\mathbf{x}_h)(e_h^{pr}) \equiv R^{pr}(\mathbf{x}_h)(e_h^{du})$.

For the proof we refer the reader to [22].

Remark 4.2. *In Sec.4.1.4 we have considered a modified Lagrangian functional $\mathcal{L}_h(\cdot)$ for the "discretize–then–differentiate" approach. The results of Proposition 4.1 hold also in this case, simply by recalling Eq.(4.31) and by evaluating $s - s_h + d_h^{pr}(v_h)(z_h) = \mathcal{L}(\mathbf{x}) - \mathcal{L}(\mathbf{x}_h)$ rather than $s - s_h$.*

Remark 4.3. *In the case of linear problems (with linear output functional and PDEs), the goal–oriented analysis is often provided without introducing explicitly the Lagrangian functional formalism, due to its simplicity.*

Remark 4.4. *We have considered the case for which both the approximated primal and dual variables are assumed in the same space $\mathcal{V}_h$. However, in general, $v_h \in \mathcal{V}_h^{pr}$ and $z_h \in \mathcal{V}_h^{du}$, with $\mathcal{V}_h^{pr}$, $\mathcal{V}_h^{du} \subset \mathcal{V}$. In this case a corrected output functional is often introduced, say $\widetilde{s}_h$, which is defined as $\widetilde{s}_h := \mathcal{L}(\mathbf{x}_h)$, with $R^{pr}(\mathbf{x}_h)(z_h) \neq 0$. The goal–oriented analysis could be provided also in this case, as we specify in Chapter 6 for the RB method in the linear case. We notice that, if $\mathcal{V}_h^{pr} \equiv \mathcal{V}_h^{du} = \mathcal{V}_h$, then $\widetilde{s}_h \equiv s_h$, being $R^{pr}(\mathbf{x}_h)(z_h) = 0$.2*

In order to evaluate the errors reported in Proposition 4.1, we need firstly to estimate the errors $\mathbf{e}_h$ or e_h^{du}, being the exact primal and dual solutions unknown. Hence, the necessity to provide a posteriori estimates for the output error $s - s_h$. From the results of Proposition 4.1 we can deduce the main features of the *Dual Weighted Residual* method (see [21, 22, 154]), for which, in the output error evaluation, the dual error e_h^{du} weights the primal residual $R^{pr}(\mathbf{x}_h)(\cdot)$ and, viceversa, the primal error e_h^{pr} weights the dual residual $R^{du}(\mathbf{x}_h)(\cdot)$. However, we notice that a posteriori error estimates strongly depend on the particular Galerkin method under consideration (FE, RB or Spectral methods); moreover, in the unsteady cases, the estimates depend also on the numerical scheme used for the approximation in time. With this aim, we refer the reader to Chapter 5 for the anisotropic goal–oriented analysis for the FE method in the steady case, while to Chapter 6 for a posteriori RB error estimates; for the Spectral method, we refer to [33].

The goal–oriented analysis and a posteriori error estimates could also be provided for algebraic systems, obtained by discretization of the first order necessary conditions; see e.g. [61, 62, 63, 177].

4.2 The goal–oriented analysis for optimal control problems

In this Section we extend the goal–oriented analysis provided in Sec.4.1 for the approximation of output functionals to the case of optimal control problems; see Chapter 2. With this aim, we briefly recall the particular optimal control problems introduced in Sec.s 2.1.2 and 2.1.3. Moreover, we discuss the "optimize–then–discretize" and "discretize–then–optimize" approximation approaches and the role of the "predictability" and "optimality" errors in the approximation of an optimal control problem. Finally, we provide the goal–oriented analysis for both unconstrained and constrained optimal control problems.

4.2.1 The general case

In Sec.s 2.1.1 and 2.2.1 we have discussed the Lagrangian functional formalism for unconstrained and constrained optimal control problems. In this Section, we make use of the standard notation introduced in Chapter 2 for the analysis of optimal control problems. We notice that, while in Sec.4.1 we have indicated with $\mathbf{x} = (v, z) \in \mathcal{X}$ the primal and dual solutions, for optimal control problems we set $\mathbf{x} = (v, z, u) \in \mathcal{X}$ (or $\mathcal{X}_{ad}$ in the constrained case), with $u \in \mathcal{U}$ (or $\mathcal{U}_{ad}$) the control function, and $\mathbf{x}^{**} = (v^{**}, z^{**}, u^{**})$ the optimal solution. Moreover, in this case, the output functional is replaced by the cost functional $J(v, u)$, on which the goal–oriented analysis is based. In fact, the error on the cost functional is assumed to be an appropriate indicator of the approximation error associated with the optimal control problem; see [19, 20, 23, 44, 119, 151, 182]. Other approaches to the error analysis are possible, even if less general; see e.g. [106].

4.2.2 The case of steady PDEs

We recall the optimal control problem described by elliptic PDEs introduced in Sec.s 2.1.2 and 2.2.2 for both the unconstrained and constrained cases.

On the basis of the first order necessary conditions (2.11), (2.13) and (2.14), we define the primal, dual and optimality residuals as:

$$R^{pr}(\mathbf{x})(\phi) = R^{pr}(v,u)(\phi) := F(\phi) - a(v,u)(\phi) \qquad \forall \phi \in \mathcal{V}, \tag{4.38}$$

$$R^{du}(\mathbf{x})(\vartheta) = R^{du}(v,z,u)(\vartheta) := m_d(v - v_d, \vartheta) - a_v(v,u)(z,\vartheta) \qquad \forall \vartheta \in \mathcal{V}, \tag{4.39}$$

$$R^{opt}(\mathbf{x})(\psi) = R^{opt}(v,z,u)(\psi) := \gamma m_u(u - u_d, \psi) - a_u(v,u)(z,\psi) \qquad \forall \psi \in \mathcal{U}_{ad}, \tag{4.40}$$

s.t. $R^{pr} : \mathcal{X}_{ad} \to \mathbb{R}$, $R^{du} : \mathcal{X}_{ad} \to \mathbb{R}$ and $R^{opt} : \mathcal{X}_{ad} \to \mathbb{R}$. We notice that in the unconstrained case $\mathcal{X}_{ad} \equiv \mathcal{X}$ and $\mathcal{U}_{ad} \equiv \mathcal{U}$. Once again, the first order necessary conditions (2.4)–(2.6) or (2.29)–(2.31) can be rewritten in terms of the residuals. In particular, by introducing the following notation:

$$\mathcal{L}'(\mathbf{x}_1)(\boldsymbol{\phi}) := R^{pr}(\mathbf{x}_1)(\phi) + R^{du}(\mathbf{x}_1)(\vartheta) + R^{opt}(\mathbf{x}_1)(\psi) \qquad \forall \mathbf{x}_1, \forall \boldsymbol{\phi} = (\phi, \vartheta, \psi) \in \mathcal{X}_{ad}, \tag{4.41}$$

the first order necessary conditions in the unconstrained case read:

$$\text{find } \mathbf{x}^{**} \in \mathcal{X} \quad : \quad \mathcal{L}'(\mathbf{x}^{**})(\boldsymbol{\phi}) = 0 \qquad \forall \boldsymbol{\phi} \in \mathcal{X}, \tag{4.42}$$

while in the constrained case:

$$\text{find } \mathbf{x}^{**} \in \mathcal{X}_{ad} \quad : \quad \mathcal{L}'(\mathbf{x}^{**})(\boldsymbol{\phi} - \mathbf{x}^{**}) \geq 0 \qquad \forall \boldsymbol{\phi} \in \mathcal{X}_{ad}, \tag{4.43}$$

with $R^{pr}(\mathbf{x}^{**})(\phi) = 0 \ \forall \phi \in \mathcal{V}$ and $R^{du}(\mathbf{x}^{**})(\vartheta) = 0 \ \forall \vartheta \in \mathcal{V}$. In analogous manner it is possible to analyze the constrained optimal control problem with Lagrangian multipliers (see Sec.s 2.2.2 and 2.2.3).

4.2.3 The case of unsteady PDEs

We briefly recall the optimal control problem described by parabolic PDEs introduced in Sec.s 2.1.3 and 2.2.3.

From Eq.s (2.17), (2.21), (2.27) and (2.23), we define, depending on the cost functionals (2.15) or (2.16), the following residuals, respectively:

$$R^{pr}(\mathbf{x})(\phi) = R^{pr}(v,u)(\phi) := \int_0^T \left(F(\phi) - a(v,u)(\phi) - m\left(\frac{\partial v}{\partial t}, \phi\right)\right)\, dt \qquad \forall \phi \in \mathcal{V}, \quad (4.44)$$

$$R^{du}(\mathbf{x})(\vartheta) = R^{du}(v,z,u)(\vartheta) := \int_0^T \left(m_d(v - v_d, \vartheta) - a_v(v,u)(z,\vartheta) + m\left(\vartheta, \frac{\partial z}{\partial t}\right)\right)\, dt \qquad \forall \vartheta \in \mathcal{V}, \quad (4.45)$$

$$R^{du}(\mathbf{x})(\vartheta) = R^{du}(v,z,u)(\vartheta) := \int_0^T \left(-a_v(v,u)(z,\vartheta) + m\left(\vartheta, \frac{\partial z}{\partial t}\right)\right)\, dt \qquad \forall \vartheta \in \mathcal{V}, \quad (4.46)$$

$$R^{opt}(\mathbf{x})(\psi) = R^{opt}(v,z,u)(\psi) := \int_0^T \left(\gamma m_u(u - u_d, \psi) - a_u(v,u)(z,\psi)\right)\, dt \qquad \forall \psi \in \mathcal{U}_{ad}. \quad (4.47)$$

Moreover, for the cost functionals (2.15) and (2.16), we introduce the following notations, respectively:

$$\begin{aligned} \mathcal{L}'(\mathbf{x}_1)(\boldsymbol{\phi}) \quad := \quad & R^{pr}(\mathbf{x}_1)(\phi) + m(v_1(0) - v_0, \phi(0)) \\ & + R^{du}(\mathbf{x}_1)(\vartheta) + m(z_1(T), \vartheta(T)) \\ & + R^{opt}(\mathbf{x}_1)(\psi) \\ & \qquad \forall \mathbf{x}_1 = (v_1, z_1, u_1), \forall \boldsymbol{\phi} = (\phi, \vartheta, \psi) \in \mathcal{X}_{ad}, \end{aligned} \quad (4.48)$$

$$\begin{aligned} \mathcal{L}'(\mathbf{x}_1)(\boldsymbol{\phi}) \quad := \quad & R^{pr}(\mathbf{x}_1)(\phi) + m(v_1(0) - v_0, \phi(0)) \\ & + R^{du}(\mathbf{x}_1)(\vartheta) + m(z_1(T), \vartheta(T)) - m_d(v_1(T) - v_d, \vartheta(T)) \\ & + R^{opt}(\mathbf{x}_1)(\psi) \\ & \qquad \forall \mathbf{x}_1 = (v_1, z_1, u_1), \forall \boldsymbol{\phi} = (\phi, \vartheta, \psi) \in \mathcal{X}_{ad}, \end{aligned} \quad (4.49)$$

for which the results (4.42) and (4.43) hold also in the unsteady case.

4.2.4 The "optimize–then–discretize" and "discretize–then–optimize" approaches

As done in Sec.4.1.4, we introduce the approximation of the optimal control problems by means of a Galerkin method. In particular, we search for the approximated optimal control solution, say $\mathbf{x}_h^*$, in the subspace $\mathcal{X}_h \subset \mathcal{X}$ (or $\mathcal{X}_{ad,h} \subset \mathcal{X}_{ad}$).

We observe that in Sec.2.3.2 we have discussed the optimization procedure for the numerical solution of optimal control problems described by PDEs; with this aim, we have introduced in an abstract setting the approximation of the first order necessary conditions. However, these approximated equations are obtained starting from the continuous optimal control problem according with the *"optimize–then–discretize"* or the *"discretize–then–optimize"* approach, which are the analogous of the strategies of Sec.4.1.4; see [38, 151]. We notice that we can extend all the considerations of Sec.4.1.4 to the context of optimal control problems. In particular, in this case, the use of an approach rather than the other could lead to evident differences, being the gradient of the Lagrangian functional used to detect the approximated

optimal solution by means of an optimization technique. For this reason, we adopt the "discretize–then–optimize" approach, being the duality arguments coherently handled at the approximated level.
By recalling Remark 4.1, approximation is often introduced without modifying the weak forms of the PDEs and the Lagrangian functional. In this case, the approximated optimal solution is obtained by solving the following first order necessary conditions in the unconstrained case:

$$\text{find } \mathbf{x}_h^* \in \mathcal{X}_h \quad : \quad \mathcal{L}'(\mathbf{x}_h^*)(\boldsymbol{\phi}_h) = 0 \qquad \forall \boldsymbol{\phi}_h \in \mathcal{X}_h, \tag{4.50}$$

and the following ones in the constrained case:

$$\text{find } \mathbf{x}_h^* \in \mathcal{X}_{ad,h} \quad : \quad \mathcal{L}'(\mathbf{x}_h^*)(\boldsymbol{\phi}_h - \mathbf{x}_h^*) \geq 0 \qquad \forall \boldsymbol{\phi}_h \in \mathcal{X}_{ad,h}, \tag{4.51}$$

with $R^{pr}(\mathbf{x}_h^*)(\phi_h) = 0 \ \forall \phi_h \in \mathcal{V}_h$ and $R^{du}(\mathbf{x}_h^*)(\vartheta_h) = 0 \ \forall \vartheta_h \in \mathcal{V}_h$.

4.2.5 Approximation: "optimality" and "predictability" errors

In the previous Section we have introduced the approximation of the optimal control problem by means of a Galerkin method, whose approximated solution is indicated with $\mathbf{x}_h^*$. We observe that such approximated optimal solution $\mathbf{x}_h^*$ is consistent, in the sense that, as the approximation improves (i.e. the space $\mathcal{X}_h$ tends to $\mathcal{X}$), then $\mathbf{x}_h^*$ tends to $\mathbf{x}^{**}$, the continuous optimal solution; see [143, 144] and also [40, 44]. As anticipated, the error on the cost functional is considered in order to evaluate the approximation error on the optimal control problem.

By indicating with $J^{**} := J(v^{**}, u^{**})$ the optimal cost functional associated with the continuous optimal control problem and with $J_h^* := J(v_h^*, u_h^*)$ that associated with the approximated problem, we define the *"optimality"* error on the cost functional as:

$$J^{**} - J_h^* := J(v^{**}, u^{**}) - J(v_h^*, u_h^*). \tag{4.52}$$

In particular, we notice that as $\mathbf{x}_h^*$ tends to $\mathbf{x}^{**}$, then the "optimality" error tends to zero, due to the consistency of the approximation. The "optimality" error expresses the capability of the approximation method to detect an optimal solution $\mathbf{x}_h^*$ which approximates the "truth" one $\mathbf{x}^{**}$. However, this error is affected by an other one, let say the *"predictability"* error, which deals with the capability of the numerical method to approximate the continuous solution for a prescribed control function. Let us indicate with v^* the primal solution obtained from the continuous problem by considering the optimal control function u_h^* of the approximated problem and with $J^* := J(v^*(u_h^*), u_h^*)$ the corresponding cost functional. Then, the "predictability" error reads:

$$J^* - J_h^* := J(v^*(u_h^*), u_h^*) - J(v_h^*, u_h^*). \tag{4.53}$$

In Fig.4.1 we report the idealized behavior of the cost functional w.r.t. the control function; in particular, we point out graphically the meaning of the cost functionals J^{**}, J^* and J_h^* (we remark that for optimal control problems we can write $J = J(v, u) = J(v(u), u) = J(u)$).
From Eq.s (4.52) and (4.53) we express the "optimality" error in terms of the "predictability" one as:

$$J^{**} - J_h^* = (J^{**} - J^*) + (J^* - J_h^*)\,; \tag{4.54}$$

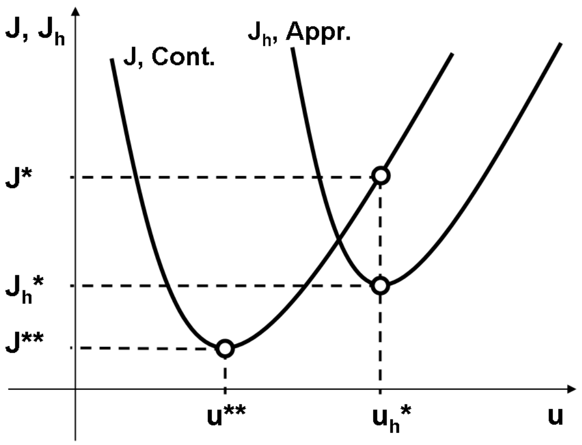

Figure 4.1: Idealized behavior of the continuous and approximated cost functionals, say J and J_h respectively, w.r.t. the control function u.

being $J^{**} \leq J^{*}$, we have:

$$J^{**} - J_h^{*} \leq J^{*} - J_h^{*}, \tag{4.55}$$

where the signum must be taken into account. Eq.(4.55) represents a first, rough estimate of the "optimality" error, which is based on the evaluation of the "predictability" error. We observe that the "predictability" error can be evaluated by means of the estimates provided in Sec.4.1.5 for the goal–oriented analysis for output functionals.
Similar considerations are reported in [24, 44].

4.2.6 The goal–oriented analysis: unconstrained and constrained cases

In this Section we discuss the goal–oriented analysis for optimal control problems, for both the unconstrained and constrained cases. For a further deepening, we refer the reader to [18, 19, 20, 22, 24, 44, 76, 119, 141, 182].

In view of the estimate of the "optimality" error $J^{**} - J_h^{*}$ (4.52), being $\mathbf{x}^{**} = (v^{**}, z^{**}, u^{**})$ and $\mathbf{x}_h^{*} = (v_h^{*}, z_h^{*}, u_h^{*})$, we define $\mathbf{e}_h^{*} := \mathbf{x}^{**} - \mathbf{x}_h^{*}$, $e_h^{pr,*} := v^{**} - v_h^{*}$, $e_h^{du,*} := z^{**} - z_h^{*}$ and $e_h^{opt,*} := u^{**} - u_h^{*}$.

Proposition 4.2. *By considering the first order necessary conditions (4.42) and their approximation (4.50) by means of a Galerkin method, if $\mathcal{L}(\cdot)$ is three times differentiable, then for an* unconstrained *optimal control problem the following equality holds:*

$$J^{**} - J_h^{*} = \frac{1}{2}\mathcal{L}'(\mathbf{x}_h^{*})(\mathbf{e}_h^{*}) + \Delta R^{*}, \tag{4.56}$$

with:

$$\Delta R^{*} := \frac{1}{2}\int_0^1 \mathcal{L}'''\left(\mathbf{x}_h^{*} + \varepsilon \mathbf{e}_h^{*}\right)(\mathbf{e}_h^{*}, \mathbf{e}_h^{*}, \mathbf{e}_h^{*})\varepsilon(\varepsilon - 1)\ d\varepsilon; \tag{4.57}$$

by introducing the residuals (4.38)–(4.40) or (4.44)–(4.47), Eq.(4.56) reads:

$$J^{**} - J_h^{*} = \frac{1}{2}\left(R^{pr}(\mathbf{x}_h^{*})(e_h^{du,*}) + R^{du}(\mathbf{x}_h^{*})(e_h^{pr,*})\right) + \Delta R^{*}. \tag{4.58}$$

For a constrained *optimal control problem, starting from Eq.s (4.43) and (4.51), we have:*

$$J^{**} - J_h^* \leq \frac{1}{2}\mathcal{L}'(\mathbf{x}_h^*)(\mathbf{e}_h^*) + \Delta R^*, \tag{4.59}$$

which, by introducing the residuals, reads:

$$J^{**} - J_h^* \leq \frac{1}{2}\left(R^{pr}(\mathbf{x}_h^*)(e_h^{du,*}) + R^{du}(\mathbf{x}_h^*)(e_h^{pr,*}) + R^{opt}(\mathbf{x}_h^*)(e_h^{opt,*})\right) + \Delta R^*. \tag{4.60}$$

If $\mathcal{L}(\cdot)$ is at most quadratic*, then $\Delta R^* = 0$ and the "optimality" error in the* unconstrained *case reads:*

$$J^{**} - J_h^* = \frac{1}{2}\left(R^{pr}(\mathbf{x}_h^*)(e_h^{du,*}) + R^{du}(\mathbf{x}_h^*)(e_h^{pr,*})\right), \tag{4.61}$$

while in the constrained *case the following estimate holds:*

$$J^{**} - J_h^* \leq \frac{1}{2}\left(R^{pr}(\mathbf{x}_h^*)(e_h^{du,*}) + R^{du}(\mathbf{x}_h^*)(e_h^{pr,*}) + R^{opt}(\mathbf{x}_h^*)(e_h^{opt,*})\right). \tag{4.62}$$

For the proof we refer the reader to [20, 182].
We observe that for unconstrained optimal control problems $R^{opt}(\mathbf{x}_h^*)(\psi_h^*) = 0 \ \forall \psi_h^* \in \mathcal{U}_h$ from Eq.(4.50), hence $R^{opt}(\mathbf{x}_h^*)(e_h^{opt,*}) = 0$, and the results (4.58) and (4.61) follow. In the constrained case we have in general $R^{opt}(\mathbf{x}_h^*)(e_h^{opt,*}) \neq 0$ (see Eq.(4.51)), from which the results (4.60) and (4.62) follow.
We notice that for an optimal control problem described by linear (primal) PDEs and with a quadratic cost functional, the estimates (4.61) and (4.62) hold.

As shown in Sec.4.1.5, the results of Proposition 4.2 depend on the primal, dual and optimality errors $\mathbf{e}_h^*$ ($e_h^{pr,*}$, $e_h^{du,*}$ and $e_h^{opt,*}$), which are unknown being dependent on the continuous optimal solution $\mathbf{x}^{**}$ (v^{**}, z^{**} and u^{**}). For this reason, in order to evaluate the results (4.56)–(4.62), suitable estimates of such errors must be provided depending on the particular Galerkin method under consideration and, for unsteady problems, on the numerical scheme used for the approximation in time. We observe that also in the case of optimal control problems, as for the goal–oriented analysis for output functionals, the main features of the DWR method are preserved. In fact, the primal residual $R^{pr}(\mathbf{x}_h^*)(\cdot)$ is weighted by the dual error $e_h^{du,*}$, the dual residual $R^{du}(\mathbf{x}_h^*)(\cdot)$ by the primal error $e_h^{pr,*}$ and, in the case of constrained optimal control problems, the optimality residual $R^{opt}(\mathbf{x}_h^*)(\cdot)$ is weighted by the optimality error $e_h^{opt,*}$.

Chapter 5

Goal–Oriented Anisotropic Mesh Adaptivity

In this Chapter, in view of environmental applications, we aim at devising an effective *mesh adaptivity* technique based on the goal–oriented analysis applied to the Finite Element method.

As anticipated in Chapter 1, we are interested in studying the distribution of some pollutants in atmosphere or in water, for which we are led to monitor or accurately compute quantities of interest, such as concentrations rather than fluxes in localized portions of the domain. Mathematically these quantities are identified by proper output functionals, typically represented by interior or boundary integrals. Hence, the necessity to deal with physically meaningful quantities justifies the employment of a *goal–oriented* analysis; see Chapter 4 and e.g. [16, 22, 61, 131]). The basic feature of this approach consists of estimating, within a user–defined tolerance, the exact (but unknown) functional, evaluating the functional itself on a suitable computable approximation. The final outcome of the goal–oriented analysis is an a posteriori estimator for the error on the target quantity.

Under reasonable assumptions (see Chapter 1 and e.g. [51]) the environmental phenomena at hand can be described by linear advection–diffusion–reaction PDEs. In particular, when considering the transport of a pollutant, strongly advection–dominated flows are often involved, thus leading to deal with situations characterized by evident directional features and steep gradients (as in the case of boundary or internal layers). To sharply capture these troublesome aspects without spoiling the overall computational cost, an efficient remedy is provided by mesh adaption techniques. With this respect, a further improvement in terms of saving of the computational cost can be obtained by means of an *anisotropic adaptivity* (see [7, 34, 43, 54, 56, 83, 100, 139, 169, 177]). The basic idea consists in smartly orienting, sizing and shaping the elements of the computational mesh in order to contain the number of the degrees of freedom's and to increase the numerical accuracy.

In this Chapter, after having generalized the anisotropic a posteriori error estimate reported in [140] for the energy error, we combine the goal–oriented analysis with the anisotropic setting; see [43]. The aim consists in extending the *Dual Weighted Residual* approach (see Chapter 4) to the case of anisotropic a posteriori error estimates. A first attempt in this direction has been made in [52]. We improve this analysis by carrying over to the goal–oriented framework the good property of the a posteriori error estimator to depend on the error itself, typical of the anisotropic residual based error analysis presented in [115, 120]. This dependence makes the estimator not immediately computable; however, after approximating this error via the Zienkiewicz–Zhu gradient recovery procedure [188, 189], the resulting estimator is expected to exhibit a higher convergence rate than the one in [52]. This allows us to monitor a quantity of interest on a properly adapted mesh with the least number of d.o.f.'s as possible, s.t. we can afford the environmental problems of interest in an effective way.

The Chapter is structured as follows. In Sec.5.1 we present the mathematical problem at hand and, in Sec.5.2, we introduce the anisotropic setting. An a posteriori error estimator for the energy norm is derived in Sec.5.3 in order to familiarize with the a posteriori framework in view of the goal–oriented setting. In Sec.5.4 we provide some numerical results related to this first a posteriori error estimator together with the actual adaptive algorithm. Sec.5.5 deals with the goal–oriented a posteriori analysis in the anisotropic setting. In particular, in Sec.5.6, two alternative anisotropic error estimators are addressed and compared numerically. Concluding remarks follow.

5.1 The problem at hand

In this Section we introduce the model problem, i.e., the advection–diffusion–reaction (ADR) PDE. In more detail, in view of advection dominated cases, we consider a SUPG type stabilized formulation [30]. The ADR equation reads:

$$\begin{cases} -\nabla\cdot(\nu\nabla v) + \mathbf{V}\cdot\nabla v + \gamma v = f & \text{in } \Omega, \\ v = 0 & \text{on } \Gamma_D, \\ \nu\nabla v\cdot\hat{\mathbf{n}} = g & \text{on } \Gamma_N, \end{cases} \tag{5.1}$$

where $\Omega \subset \mathbb{R}^2$ is a polygonal domain with boundary $\partial\Omega$, Γ_D and Γ_N are suitable measurable nonoverlapping partitions of $\partial\Omega$ with $\Gamma_D \neq \emptyset$ and such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$, and $\hat{\mathbf{n}}$ is the unit outward normal vector to $\partial\Omega$. Moreover we assume that the source $f \in L^2(\Omega)$, the diffusivity $\nu \in L^\infty(\Omega)$, with $\nu \geq \nu_0 > 0$, the reaction coefficient $\gamma \in L^\infty(\Omega)$, the advective field $\mathbf{V} \in [L^\infty(\Omega)]^2$, with $\nabla\cdot\mathbf{V} \in L^\infty(\Omega)$ and $-\frac{1}{2}\nabla\cdot\mathbf{V} + \gamma \geq 0$, a.e. in Ω, and the Neumann datum $g \in L^2(\Gamma_N)$ are assigned functions.
The weak form of (5.1) reads:

$$\text{find } v \in \mathcal{V} \ : \ a(v,\phi) = F(\phi) \qquad \forall\phi \in \mathcal{V}, \tag{5.2}$$

where $\mathcal{V} := H^1_{\Gamma_D}(\Omega) = \{w \in H^1(\Omega) \ : \ w|_{\Gamma_D} = 0\}$, while the bilinear form $a(\cdot,\cdot)$ and the linear functional $F(\cdot)$ are:

$$a(v,\phi) := \int_\Omega \left(\nu\nabla v\cdot\nabla\phi + \mathbf{V}\cdot\nabla v\ \phi + \gamma v\phi\right)\ d\Omega, \tag{5.3}$$

$$F(\phi) := \int_\Omega f\phi\ d\Omega + \int_{\Gamma_N} g\phi\ d\Gamma. \tag{5.4}$$

Existence and uniqueness of the solution of Eq.(5.2) follow immediately from the above hypotheses ([133]).
In order to approximate the problem (5.2), let $\{\mathcal{T}_h\}_h$ be a family of conforming decompositions [36] of $\overline{\Omega}$ into triangles K of diameter h_K, such that there is always a vertex of $\mathcal{T}_h$ on the interface between Γ_D and Γ_N. Let $\mathcal{V}_h = \{w_h \in C^0(\overline{\Omega}) \ : \ w_h|_K \in \mathbb{P}_1, \forall K \in \mathcal{T}_h,\ w_h|_{\Gamma_D} = 0\} \subset$

$\mathcal{V}$ denote the subspace of affine functions, $\mathbb{P}_1$ being the space of polynomials of (total) degree less than or equal to one.

As we are interested in applications where advection may strongly dominate over diffusion and reaction, a proper approximation scheme must be employed to limit the spurious oscillations of the numerical solution. For this reason, we consider the strongly consistent SUPG type method:

$$\text{find } v_h \in \mathcal{V}_h \ : \ a_h(v_h, \phi_h) = F_h(\phi_h) \qquad \forall \phi_h \in \mathcal{V}_h, \tag{5.5}$$

with:

$$a_h(v, \phi) := a(v, \phi) + \sum_{K \in \mathcal{T}_h} \tau_K \left(-\nabla \cdot (\nu \nabla v) + \mathbf{V} \cdot \nabla v + \gamma v, \mathbf{V} \cdot \nabla \phi\right)_K, \tag{5.6}$$

$$F_h(\phi) := F(\phi) + \sum_{K \in \mathcal{T}_h} \tau_K \left(f, \mathbf{V} \cdot \nabla \phi\right)_K, \tag{5.7}$$

where $(\cdot, \cdot)_K$ denotes the L^2–inner product on K, while τ_K are suitable stabilization parameters to be defined later (see Remark 5.3). By simply subtracting Eq.(5.5) from Eq.(5.2), with $\phi = \phi_h$, we get the "skew orthogonality" property, which reads:

$$a(e_h^{pr}, \phi_h) = \sum_{K \in \mathcal{T}_h} \tau_K \left(-\nabla \cdot (\nu \nabla v_h) + \mathbf{V} \cdot \nabla v_h + \gamma v_h - f, \mathbf{V} \cdot \nabla \phi_h\right)_K \qquad \forall \phi_h \in \mathcal{V}_h, \tag{5.8}$$

being $e_h^{pr} := v - v_h$ the approximation error. If further v enjoys a higher regularity, i.e. $v \in \{w \in H^1(\Omega) \ : \ \nabla \cdot (\nu \nabla w) \in L^2(K), \forall K \in \mathcal{T}_h\}$, the standard Galerkin orthogonality property w.r.t. the stabilized bilinear form $a_h(\cdot, \cdot)$ holds:

$$a_h(e_h^{pr}, \phi_h) = 0 \qquad \forall \phi_h \in \mathcal{V}_h. \tag{5.9}$$

5.2 The anisotropic framework

Throughout this Section we recall the anisotropic framework on which the a posteriori analysis is based; see [52, 54, 53].

Let us consider the standard invertible affine map $T_K : \widehat{K} \to K$ from a reference triangle $\widehat{K}$ to the general one $K \in \mathcal{T}_h$, s.t., for any $\mathbf{x} \in K$:

$$\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2)^T = T_K(\widehat{\mathbf{x}}) = M_K \widehat{\mathbf{x}} + \mathbf{t}_K \qquad \text{with} \quad \widehat{\mathbf{x}} \in \widehat{K}, \tag{5.10}$$

where $M_K \in \mathbb{R}^{2\times 2}$ and $\mathbf{t}_K \in \mathbb{R}^2$. The most suitable choice for $\widehat{K}$ coincides with the triangle with vertices $(-\sqrt{3}/2, -1/2)$, $(\sqrt{3}/2, -1/2)$, $(0, 1)$, that is with the equilateral triangle inscribed in the unit circle, with barycenter located at the origin.

The anisotropic information of each triangle K is derived moving from the spectral properties of the Jacobian M_K. Firstly, let us introduce the polar decomposition $M_K = B_K Z_K$ of M_K into a symmetric positive definite and an orthogonal matrix, B_K, $Z_K \in \mathbb{R}^{2\times 2}$, respectively. Then, let us further factorize the matrix B_K in terms of its eigenvectors $\mathbf{r}_{i,K}$ and eigenvalues $\lambda_{i,K}$, with $i = 1, 2$, as $B_K = R_K^T \Lambda_K R_K$ with $\Lambda_K = \mathrm{diag}\,(\lambda_{1,K}, \lambda_{2,K})$ and $R_K^T = [\mathbf{r}_{1,K}, \mathbf{r}_{2,K}]$. Thus the shape and the orientation of each element K are completely characterized by these quantities: the eigenvectors $\mathbf{r}_{i,K}$ provide us with the directions of the semi–axes of the ellipse circumscribing K, while the eigenvalues $\lambda_{i,K}$ measure the length of such semi–axes (see Fig.5.1).

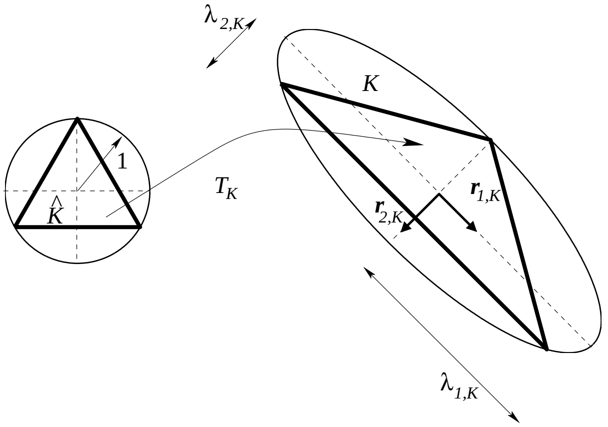

Figure 5.1: The affine map T_K from the reference triangle $\widehat{K}$ to the triangle K with the geometrical quantities $\lambda_{1,K}$, $\lambda_{2,K}$, $\mathbf{r}_{1,K}$ and $\mathbf{r}_{2,K}$ highlighted.

Remark 5.1. *In the analysis reported below Z_K and $\mathbf{t}_K$ do not play any role being associated with a rigid rotation and a shift, respectively.*

Without loosing in generality, we assume that $\lambda_{1,K} \geq \lambda_{2,K}$, so that the so–called stretching factor is:

$$s_K = \frac{\lambda_{1,K}}{\lambda_{2,K}} \geq 1, \tag{5.11}$$

being $s_{\widehat{K}}$ identically equal to 1.

In view of the a posteriori analysis in Sec.s 5.3 and 5.5, we introduce the Clément interpolant ([37]), defined, in the case of linear finite elements, as:

$$I_h^1 w(\mathbf{x}) = \sum_{N_j \in N_\Omega \cup N_{\Gamma_N}} P_j w(N_j) \varphi_j(\mathbf{x}) \qquad \forall w \in L^2(\Omega), \tag{5.12}$$

where φ_j is the Lagrangian basis function associated with the node N_j, while $P_j w$ denotes the plane associated with the patch Δ_j of the elements sharing node N_j and defined by the relations:

$$\int_{\Delta_j} (P_j w - w)\psi \; d\Omega = 0 \qquad \text{with } \psi = 1, \mathrm{x}_1, \mathrm{x}_2. \tag{5.13}$$

Notice that the sum in Eq.(5.12) runs only on the internal mesh vertices N_Ω and on those on the Neumann boundary N_{Γ_N}.

Due to the local feature of the anisotropic interpolation estimates, we refer to the restriction of $I_h^1 w$ to the element K as $I_K^1 w$.

Under the further following assumptions:

$$\mathrm{card}\,(\Delta_K) \leq N \quad \text{and} \quad \mathrm{diam}(T_K^{-1}(\Delta_K)) \leq C_\Delta \simeq O(1), \tag{5.14}$$

being $N \in \mathbb{N}$, with card$(\cdot)$ and diam$(\cdot)$ the cardinality and the diameter of a given set, the constant $C_\Delta \geq h_{\widehat{K}}$ and Δ_K denoting the patch of all the elements sharing at least a vertex with K, we can prove (see [54, 53, 121]) the following result.

Proposition 5.1. *Let $w \in H^1(\Omega)$. Then there exist constants $C_i = C_i(N, C_\Delta)$, with $i = 1, 2, 3$, s.t., for any $K \in \mathcal{T}_h$, it holds:*

$$\|w - I_K^1 w\|_{L^2(K)} \leq C_1 \left[\sum_{i=1}^{2} \lambda_{i,K}^2 \left(\mathbf{r}_{i,K}^T G_K(w) \mathbf{r}_{i,K} \right) \right]^{1/2}, \tag{5.15}$$

$$|w - I_K^1 w|_{H^1(K)} \leq C_2 \frac{1}{\lambda_{2,K}} \left[\sum_{i=1}^{2} \lambda_{i,K}^2 \left(\mathbf{r}_{i,K}^T G_K(w) \mathbf{r}_{i,K} \right) \right]^{1/2}, \tag{5.16}$$

$$\|w - I_K^1 w\|_{L^2(\partial K)} \leq C_3 h_K^{\frac{1}{2}} \left[s_K \left(\mathbf{r}_{1,K}^T G_K(w) \mathbf{r}_{1,K} \right) + \frac{1}{s_K} \left(\mathbf{r}_{2,K}^T G_K(w) \mathbf{r}_{2,K} \right) \right]^{1/2}, \tag{5.17}$$

with:

$$G_K(w) := \sum_{T \in \Delta_K} \begin{bmatrix} \displaystyle\int_T \left(\frac{\partial w}{\partial \mathrm{x}_1} \right)^2 d\Omega & \displaystyle\int_T \left(\frac{\partial w}{\partial \mathrm{x}_1} \right) \left(\frac{\partial w}{\partial \mathrm{x}_2} \right) d\Omega \\ \displaystyle\int_T \left(\frac{\partial w}{\partial \mathrm{x}_1} \right) \left(\frac{\partial w}{\partial \mathrm{x}_2} \right) d\Omega & \displaystyle\int_T \left(\frac{\partial w}{\partial \mathrm{x}_2} \right)^2 d\Omega \end{bmatrix} \tag{5.18}$$

a 2×2 symmetric positive semi–definite matrix.

Remark 5.2. *Conditions (5.14) essentially exclude too distorted patches in the reference framework. This simplifies the derivation of the interpolation estimates above as they are actually carried out in the reference setting and then mapped back to the general one. On the other hand, the same conditions do not limit the anisotropic features (stretching factor and orientation) of each K in Δ_K, but rather the corresponding variation in Δ_K (see [122] for examples of acceptable and not acceptable patches).*

Moreover the following result holds; see [121] for the proof.

Proposition 5.2. *For any function $w \in H^1(\Omega)$ and two constants $\alpha, \beta > 0$, it holds:*

$$\min\{\alpha, \beta\} \leq \frac{\alpha \left(\mathbf{r}_{1,K}^T G_K(w) \mathbf{r}_{1,K} \right) + \beta \left(\mathbf{r}_{2,K}^T G_K(w) \mathbf{r}_{2,K} \right)}{|v|_{H^1(\Delta_K)}^2} \leq \max\{\alpha, \beta\}, \tag{5.19}$$

being $G_K(\cdot)$ defined in Eq.(5.18).

Remark 5.3. *With reference to the SUPG type stabilized formulation (5.5)–(5.7), the following anisotropic stabilization parameter is used:*

$$\tau_K = \begin{cases} \dfrac{\lambda_{2,K}^2}{12 \nu_K} & \text{if } \ \mathbb{P}e_K < 1, \\ \dfrac{\lambda_{2,K}}{2\|\mathbf{V}\|_{L^\infty(K)}} & \text{if } \ \mathbb{P}e_K \geq 1, \end{cases} \tag{5.20}$$

where $\mathbb{P}e_K := \lambda_{2,K} \|\mathbf{V}\|_{L^\infty(K)} / (6\nu_K)$ is the local Péclet number, with $\nu_K = \min\limits_{\mathbf{x} \in K} \nu(\mathbf{x})$; see [122].

5.3 An a posteriori error estimator for the energy norm

In this Section we derive an anisotropic a posteriori error estimate for the energy norm $|||e_h^{pr}||| := \big(a(e_h^{pr}, e_h^{pr})\big)^{1/2}$ of the approximation error associated with the (primal) ADR equation (5.1). In more detail, we move from a standard residual–based approach [178] properly combined with the anisotropic analysis of Sec.5.2. The present analysis generalizes the result reported in [140] to the case of a non–constant convective term as well as to mixed boundary conditions.

Firstly, let us anticipate some notations used in the main result of this Section. For any $K \in \mathcal{T}_h$, let us define the interior and the boundary residuals associated with the FE approximation v_h, respectively:

$$r_K(v_h) := (f + \nabla \cdot (\nu \nabla v_h) - \mathbf{V} \cdot \nabla v_h - \gamma v_h)\,|_K, \tag{5.21}$$

and:

$$j_K(v_h)|_E := \begin{cases} 0 & \text{for any } E \in \partial K \cap \mathcal{E}_{h,D}, \\ 2\,(g - (\nu \nabla v_h)|_K \cdot \hat{\mathbf{n}}_K)\,|_E & \text{for any } E \in \partial K \cap \mathcal{E}_{h,N}, \\ -\,[\nu \nabla v_h \cdot \hat{\mathbf{n}}]_E & \text{for any } E \in \partial K \cap \mathcal{E}_{h,int}, \end{cases} \tag{5.22}$$

where $\mathcal{E}_{h,int}$ denotes the set of the internal edges of the skeleton $\mathcal{E}_h$ of the triangulation $\mathcal{T}_h$, while $\mathcal{E}_{h,D}$ and $\mathcal{E}_{h,N}$ stands for the Dirichlet and Neumann subset of $\mathcal{E}_h$, respectively. With the following notation:

$$[\nu \nabla v_h \cdot \hat{\mathbf{n}}]_E := (\nu \nabla v_h)|_K \cdot \hat{\mathbf{n}}_K + (\nu \nabla v_h)|_{K'} \cdot \hat{\mathbf{n}}_{K'}, \tag{5.23}$$

which is defined on the edge E separating elements K and K', we refer to the jump of the diffusive flux on the interface.

Remark 5.4. *In Eq.s (5.21) and (5.22) we have introduced the element–wise interior and boundary residuals associated with the primal ADR equation in a strong form, which is standard for the a posteriori FE error analysis; see e.g. [178]. However, the present analysis can be recast in terms of the residual (4.7) written in weak form introduced in Sec.4.1.2.*

We state now the desired anisotropic error estimate (see also [43, 121, 140]).

Proposition 5.3. *Let $v \in \mathcal{V}$ be the solution of the weak problem* (5.2) *and $v_h \in \mathcal{V}_h$ be the corresponding approximation via Eq.(5.5). Then, it holds:*

$$|||e_h^{pr}||| \leq C \left(\sum_{K \in \mathcal{T}_h} \alpha_K \rho_K(v_h) w_K(e_h^{pr}) \right)^{1/2}, \tag{5.24}$$

with $\alpha_K := (\lambda_{1,K}\lambda_{2,K})^{1/2}$,

$$\rho_K(v_h) := \left(1 + \tau_K \frac{\|\mathbf{V}\|_{L^\infty(K)}}{\lambda_{2,K}}\right) \|r_K(v_h)\|_{L^2(K)} + \frac{1}{2} \left(\frac{h_K}{\lambda_{1,K}\lambda_{2,K}}\right)^{1/2} \|j_K(v_h)\|_{L^2(\partial K)}, \tag{5.25}$$

$$w_K(e_h^{pr}) := \left[s_K \left(\mathbf{r}_{1,K}^T G_K(e_h^{pr}) \mathbf{r}_{1,K}\right) + \frac{1}{s_K} \left(\mathbf{r}_{2,K}^T G_K(e_h^{pr}) \mathbf{r}_{2,K}\right) \right]^{1/2}, \tag{5.26}$$

where $C = C(N, C_\Delta)$, τ_K is the stabilization parameter defined in Eq.(5.20), the matrix $G_K(\cdot)$ is defined in Eq.(5.18), while the residuals $r_K(v_h)$ and $j_K(v_h)$ are given by Eq.s (5.21) and (5.22), respectively.

Proof. From the definition (5.3) of the bilinear form $a(\cdot,\cdot)$ and thanks to the weak form (5.2), we have:

$$\begin{aligned} a(e_h^{pr},\phi) &= F(\phi) - a(v_h,\phi) = \sum_{K\in\mathcal{T}_h}\left\{\int_K f\phi\ d\Omega + \int_{\partial K\cap\Gamma_N} g\phi\ d\Gamma\right\} \\ &- \sum_{K\in\mathcal{T}_h}\int_K \left(\nu\nabla v_h\cdot\nabla\phi + \mathbf{V}\cdot\nabla v_h\ \phi + \gamma v_h\phi\right)\ d\Omega, \end{aligned} \tag{5.27}$$

for any $\phi\in\mathcal{V}$. Notice that we have split the integrals element–wise with the aim of localizing the a posteriori estimator. Moreover, we observe that in terms of the weak residual (4.7) introduced in Sec.4.1.2, we have $a(e_h^{pr},\phi) = R^{pr}(v_h)(\phi)$. After integrating by parts in the second sum at the r.h.s. of Eq.(5.27), we obtain:

$$\begin{aligned} a(e_h^{pr},\phi) &= \sum_{K\in\mathcal{T}_h}\left\{\int_K f\phi\ d\Omega + \int_{\partial K\cap\Gamma_N} g\phi\ d\Gamma\right\} \\ &- \sum_{K\in\mathcal{T}_h}\left\{\int_K \left(-\nabla\cdot(\nu\nabla v_h) + \mathbf{V}\cdot\nabla v_h + \gamma v_h\right)\phi\ d\Omega\right. \\ &+ \left.\int_{\partial K\cap\mathcal{E}_{h,N}} \nu\nabla v_h\cdot\hat{\mathbf{n}}_K\phi\ d\Gamma + \int_{\partial K\cap\mathcal{E}_{h,int}} \nu\nabla v_h\cdot\hat{\mathbf{n}}_K\phi\ d\Gamma\right\}. \end{aligned} \tag{5.28}$$

Let us observe that the integration by parts is possible consistently with the regularity of v_h since we are working on each element K. Thanks to the definitions (5.21) and (5.22), we get:

$$a(e_h^{pr},\phi) = \sum_{K\in\mathcal{T}_h}\left\{\int_K r_K(v_h)\phi\ d\Omega + \frac{1}{2}\int_{\partial K} j_K(v_h)\phi\ d\Gamma\right\}, \tag{5.29}$$

the factor 1/2 taking into account the fact that each internal edge E shares two elements of the triangulation. A suitable combination of the “skew orthogonality” property (5.8) together with the definition (5.21) of the internal residual $r_K(v_h)$ and with identity (5.29) (also used with $\phi = \phi_h$), yields:

$$\begin{aligned} a(e_h^{pr},\phi) &= a(e_h^{pr},\phi) - a(e_h^{pr},\phi_h) - \sum_{K\in\mathcal{T}_h}\tau_K\left(r_K(v_h), \mathbf{V}\cdot\nabla\phi_h\right)_{L^2(K)} \\ &= \sum_{K\in\mathcal{T}_h}\left\{\int_K r_K(v_h)(\phi-\phi_h)\ d\Omega + \frac{1}{2}\int_{\partial K} j_K(v_h)(\phi-\phi_h)\ d\Gamma\right. \\ &- \left.\tau_K\int_K r_K(v_h)\mathbf{V}\cdot\nabla\phi_h\ d\Omega\right\}. \end{aligned} \tag{5.30}$$

By adding and subtracting the quantity $\tau_K \displaystyle\int_K r_K(v_h)\mathbf{V}\cdot\nabla\phi\ d\Omega$ to Eq.(5.30) and by using the Cauchy–Schwarz inequality, we have:

$$
\begin{aligned}
|a(e_h^{pr},\phi)| \quad &\leq \quad \sum_{K\in\mathcal{T}_h} \Big\{ \|r_K(v_h)\|_{L^2(K)} \left[\|\phi-\phi_h\|_{L^2(K)} \right. \\
&+ \quad \tau_K \, \|\mathbf{V}\cdot\nabla(\phi-\phi_h)\|_{L^2(K)} + \tau_K \, \|\mathbf{V}\cdot\nabla\phi\|_{L^2(K)} \left. \right] \\
&+ \quad \frac{1}{2}\|j_K(v_h)\|_{L^2(\partial K)} \|\phi-\phi_h\|_{L^2(\partial K)} \Big\}.
\end{aligned}
\tag{5.31}
$$

The two terms involving the advective field $\mathbf{V}$ can be further bounded as:

$$
\begin{aligned}
\|\mathbf{V}\cdot\nabla(\phi-\phi_h)\|_{L^2(K)} \leq \|\mathbf{V}\|_{L^\infty(K)} \; |\phi-\phi_h|_{H^1(K)}, \\
\|\mathbf{V}\cdot\nabla\phi\|_{L^2(K)} \leq \|\mathbf{V}\|_{L^\infty(K)} \; |\phi|_{H^1(K)},
\end{aligned}
\tag{5.32}
$$

while, from Proposition 5.2 with $\alpha = s_K$ and $\beta = 1/s_K$, it holds:

$$
|\phi|_{H^1(K)} \leq |\phi|_{H^1(\Delta_K)} \leq \frac{1}{\lambda_{2,K}} \left[\sum_{i=1}^{2} \lambda_{i,K}^2 \left(\mathbf{r}_{i,K}^T G_K(\phi) \mathbf{r}_{i,K} \right) \right]^{1/2} . \tag{5.33}
$$

Coming back to Eq.(5.31) and choosing the arbitrary function ϕ_h as the Clément interpolant of ϕ, i.e. $\phi_h = I_h^1\phi \in \mathcal{V}_h$, we can exploit Proposition 5.1 along with relations (5.32) and (5.33) in order to get:

$$
\begin{aligned}
|a(e_h^{pr},\phi)| \leq \sum_{K\in\mathcal{T}_h} \Big\{ \Big[& (\lambda_{1,K}\lambda_{2,K})^{1/2} \Big(C_1 + (1+C_2)\,\tau_K \frac{\|\mathbf{V}\|_{L^\infty(K)}}{\lambda_{2,K}} \Big) \|r_K(v_h)\|_{L^2(K)} \\
+ C_3 \frac{1}{2} h_K^{1/2} \|j_K(v_h)\|_{L^2(\partial K)} \Big] & \Big[s_K \left(\mathbf{r}_{1,K}^T G_K(\phi) \mathbf{r}_{1,K} \right) + \frac{1}{s_K} \left(\mathbf{r}_{2,K}^T G_K(\phi) \mathbf{r}_{2,K} \right) \Big]^{1/2} \Big\}.
\end{aligned}
\tag{5.34}
$$

Result (5.24) follows after taking $\phi = e_h^{pr}$ and identifying C with $\max\{C_1, 1+C_2, C_3\}$. □

We notice that estimate (5.24) is not useful yet in view, for instance, of an adaptive procedure, the r.h.s. depending on the approximation error e_h^{pr} itself. On the other hand, due to the presence of such error, we expect the whole right-hand side of Eq.(5.24) to go to zero, as the mesh gets finer and finer. In order to exploit this nice feature we aim at approximating the weights $w_K(e_h^{pr})$ in Eq.(5.24) via suitable computable quantities. Due to their dependence on the first order derivatives of the error, a convenient strategy consists in resorting to the Zienkiewicz–Zhu (ZZ) recovery procedure [188, 189, 190]. Denoting the recovered gradient of v_h with $\nabla^{ZZ}v_h = \left((\nabla^{ZZ}v_h)_1, (\nabla^{ZZ}v_h)_2\right)^T$, we substitute the matrix $G_K(e_h^{pr})$ in the definition (5.26) of $w_K(e_h^{pr})$ with:

$$
[G_K(e_h^{pr,\star})]_{i,j} := \sum_{T\in\Delta_K} \int_T \Big((\nabla^{ZZ}v_h)_i - \frac{\partial v_h}{\partial \mathrm{x}_i} \Big) \Big((\nabla^{ZZ}v_h)_j - \frac{\partial v_h}{\partial \mathrm{x}_j} \Big) \, d\Omega, \tag{5.35}
$$

for $i,j = 1,2$. We can now define the fully computable a posteriori error estimator which will drive the anisotropic mesh adaption procedure used throughout the numerical validation in the Section below.

Definition 5.1. *Let $v_h \in \mathcal{V}_h$ be the solution of the discrete problem* (5.5)*. Then, the energy norm of the approximation error e_h^{pr} can be estimated by the quantity:*

$$\eta := \left(\sum_{K \in \mathcal{T}_h} \eta_K^2 \right)^{1/2}, \tag{5.36}$$

the local indicator η_K being given by:

$$\eta_K := \left(\alpha_K \rho_K(v_h) w_K(e_h^{pr,\star}) \right)^{1/2}, \tag{5.37}$$

where α_K and $\rho_K(v_h)$ are defined as in Proposition 5.3, while:

$$w_K(e_h^{pr,\star}) := \left[s_K \left(\mathbf{r}_{1,K}^T G_K(e_h^{pr,\star}) \mathbf{r}_{1,K} \right) + \frac{1}{s_K} \left(\mathbf{r}_{2,K}^T G_K(e_h^{pr,\star}) \mathbf{r}_{2,K} \right) \right]^{1/2}. \tag{5.38}$$

Remark 5.5. *The isotropic counterpart of the estimator (5.36) can be obtained by enforcing $\lambda_{1,K} = \lambda_{2,K} \simeq h_K$. We notice that with this choice the weight (5.26) collapses to $|e_h^{pr}|_{H^1(\Delta_K)}$. No gradient recovery procedure is needed, a simplification of such a term occurring in this case (see the derivation of standard residual based error estimators, for instance in [178]).*

Remark 5.6. *The general structure of the recovered gradient $\nabla^{ZZ} v_h$ is:*

$$\nabla^{ZZ} v_h(\mathbf{x}) = \sum_{N_j \in N_{\overline{\Omega}}} \nabla^{ZZ} v_h(N_j) \varphi_j(\mathbf{x}), \tag{5.39}$$

being $N_{\overline{\Omega}}$ the set of all the nodes in $\mathcal{T}_h$. Observe that $\nabla^{ZZ} v_h$ is piecewise linear, the hat functions φ_j coinciding with the hat basis functions of $\mathcal{V}_h$. Different recipes are available in the literature to compute the coefficients $\nabla^{ZZ} v_h(N_j)$ (see, for instance, [115, 159, 188, 189, 190]). One of the most popular choice, namely the continuous SPR procedure, can be related to the Clément interpolant I_h^1 defined previously. In more detail, it suffices picking $w = \partial v_h / \partial \mathrm{x}_i$, for $i = 1, 2$, in Eq.s (5.12) and (5.13). If further one approximates the mass matrix involved in Eq.(5.13) via the trapezoidal quadrature rule, then one obtains for the coefficients in Eq.(5.39) the following explicit expression:

$$\nabla^{ZZ} v_h(N_j) = \frac{1}{|\Delta_j|} \sum_{K \in \Delta_j} |K| \, (\nabla v_h)|_K, \tag{5.40}$$

$|\cdot|$ denoting the d–measure function, with $d = 1, 2$; see [159]. Recipes (5.39)–(5.40) will be employed in all the numerical test cases of this Chapter.

5.4 Monitoring of the energy norm: numerical assessment

The purpose of this Section is twofold: firstly, we provide an actual procedure to exploit the a posteriori error estimator (5.36), then some numerical results are discussed.

5.4.1 The adaptive procedure

We employ a metric–based adaptive procedure exploiting the estimator (5.36) in a predictive fashion. Two different approaches are typically pursued:

(a) given a constraint on the maximum number of elements, find the mesh providing the most accurate numerical solution;

(b) given a constraint on the accuracy of the numerical solution, find the mesh with the least number of elements.

We here detail the (b) approach, while providing some comments on the (a) one in Remark 5.7. Let us first recall that a metric is induced by a symmetric positive definite tensor field $\widetilde{M} : \Omega \to \mathbb{R}^2$. We aim at clarifying the interplay between metric and mesh. For any given mesh $\mathcal{T}_h$, we can define a piecewise constant metric $\widetilde{M}_{\mathcal{T}_h}$, such that, $\widetilde{M}_{\mathcal{T}_h}|_K = \widetilde{M}_K = B_K^{-2} = R_K^T \Lambda_K^{-2} R_K$, for any $K \in \mathcal{T}_h$, being the matrices defined in Sec.5.2. With respect to this metric, any triangle K is unit equilateral, i.e.:

$$\|e\|_{\widetilde{M}_{\mathcal{T}_h}} = \int_0^{|e|} \sqrt{\hat{\mathbf{t}}^T \widetilde{M}_{\mathcal{T}_h}(s) \hat{\mathbf{t}}}\; ds = 1, \tag{5.41}$$

with $\hat{\mathbf{t}}$ the unit tangent vector along the edge e.

Let us suppose now that a metric $\widetilde{M}$ is given. We show how an optimal mesh w.r.t. $\widetilde{M}$ can be defined in terms of a *matching condition*. With this respect, it is convenient to diagonalize the tensor field $\widetilde{M}$ as $\widetilde{M} = \widetilde{R}^T \widetilde{\Lambda}^{-2} \widetilde{R}$, with $\widetilde{\Lambda} = \text{diag}(\widetilde{\lambda}_1, \widetilde{\lambda}_2)$ and $\widetilde{R}^T = [\widetilde{\mathbf{r}}_1, \widetilde{\mathbf{r}}_2]$ positive diagonal and orthogonal matrices, respectively. For practical reasons, we approximate the quantities $\widetilde{\lambda}_1$, $\widetilde{\lambda}_2$, $\widetilde{\mathbf{r}}_1$ and $\widetilde{\mathbf{r}}_2$ defining $\widetilde{M}$ by piecewise constant functions over the triangulation $\mathcal{T}_h$, such that $\widetilde{\mathbf{r}}_i|_K = \widetilde{\mathbf{r}}_{i,K}$, $\widetilde{\lambda}_i|_K = \widetilde{\lambda}_{i,K}$, for any $K \in \mathcal{T}_h$ and with $i = 1, 2$. Then we introduce the following matching condition:

Definition 5.2. *A mesh $\mathcal{T}_h$ matches a given metric $\widetilde{M}$ if, for any $K \in \mathcal{T}_h$, we have:*

$$\widetilde{M}|_K = \widetilde{M}_{\mathcal{T}_h}|_K, \tag{5.42}$$

i.e. $\widetilde{\mathbf{r}}_{i,K} = \mathbf{r}_{i,K}$, $\widetilde{\lambda}_{i,K} = \lambda_{i,K}$, for $i = 1, 2$.

We stress the fact that in our case the tensor field $\widetilde{M}$ is not explicitly given. Rather it must be obtained by solving the optimization problem (b) reformulated w.r.t the optimal metric (rather than the optimal mesh) in view of Definition 5.2. Thus the optimal metric will be our actual unknown.

In more detail, the determination of $\widetilde{M}$ and of the corresponding matching triangulation is obtained via an iterative method. For clarity, we point out that, at each iteration j, we deal with three entities: the actual mesh $\mathcal{T}_h^{(j)}$, the new metric $\widetilde{M}^{(j+1)}$ computed on $\mathcal{T}_h^{(j)}$, and the updated mesh $\mathcal{T}_h^{(j+1)}$ matching $\widetilde{M}^{(j+1)}$. Problem (5.5) is first solved on $\mathcal{T}_h^{(j)}$. Then, its solution is used to set up suitable local optimization problems, with the aim of identifying the metric $\widetilde{M}^{(j+1)}$ approximating the optimal metric $\widetilde{M}$, solution of approach (b). By means of the matching condition (5.42), the new mesh $\mathcal{T}_h^{(j+1)}$ is then built. Let us detail the local optimization procedure. We rewrite the local estimator η_K in Eq.(5.37) as:

$$\eta_K^2 = \frac{|K|^{3/2}}{|\widehat{K}|^{1/2}}\, \widetilde{\rho}_K(v_h)\, \underbrace{\left[s_K \left(\mathbf{r}_{1,K}^T \widetilde{G}_K(e_h^{pr,\star}) \mathbf{r}_{1,K} \right) + \frac{1}{s_K} \left(\mathbf{r}_{2,K}^T \widetilde{G}_K(e_h^{pr,\star}) \mathbf{r}_{2,K} \right) \right]^{1/2}}_{(\diamondsuit)}, \tag{5.43}$$

being:

$$\widetilde{\rho}_K(v_h) := \frac{\rho_K(v_h)}{|K|^{1/2}} \quad \text{and} \quad \widetilde{G}_K(e_h^{pr,\star}) := \frac{G_K(e_h^{pr,\star})}{|K|}, \tag{5.44}$$

the scaled residual and recovered gradient matrix, respectively, and the relation $\alpha_K = \sqrt{|K|/|\widehat{K}|}$ advocated. This scaling is driven with the aim of making all terms in the r.h.s. of Eq.(5.43) approximately independent on the measure of triangle K, at least asymptotically (i.e., when the mesh is sufficiently fine), thus lumping this information only in a multiplicative constant. In view of approach (b) we firstly observe that minimizing the number of elements is equivalent to maximizing the area $|K|$ of each element. As we demand also that the local error indicator η_K be equal to a desired constant (the local tolerance τ) according with an equidistribution criterion, the only way to satisfy (b) is to minimize the term ($\diamondsuit$) in Eq.(5.43). This corresponds to solve the following local constrained minimization problem:

$$\begin{gathered}\text{find } s_K,\ \mathbf{r}_{1,K} \quad \text{s.t.} \\ I(s_K, \mathbf{r}_{1,K}) = s_K \left(\mathbf{r}_{1,K}^T \widetilde{G}_K(e_h^{pr,\star})\mathbf{r}_{1,K}\right) + \frac{1}{s_K}\left(\mathbf{r}_{2,K}^T \widetilde{G}_K(e_h^{pr,\star})\mathbf{r}_{2,K}\right) \qquad \text{is minimum,}\end{gathered} \tag{5.45}$$

with $s_K \geq 1$, $\|\mathbf{r}_{1,K}\|_2 = \|\mathbf{r}_{2,K}\|_2 = 1$ and $\mathbf{r}_{1,K} \cdot \mathbf{r}_{2,K} = 0$, $\|\cdot\|_2$ denoting the standard Euclidean norm. The solution of this problem is given in the following Proposition.

Proposition 5.4. *The solution $(\widetilde{s}_K, \widetilde{\mathbf{r}}_{1,K})$ of problem (5.45) is given by:*

$$\widetilde{s}_K = \sqrt{\frac{\sigma_{1,K}}{\sigma_{2,K}}}, \qquad \widetilde{\mathbf{r}}_{1,K} = \mathbf{p}_{2,K}, \tag{5.46}$$

being $\sigma_{1,K}$ and $\sigma_{2,K}$ the maximum and minimum eigenvalues of the matrix $G_K(e_h^{pr,\star})$, while $\mathbf{p}_{1,K}$ and $\mathbf{p}_{2,K}$ are the associated eigenvectors.

For the corresponding proof see [121] where is provided also a practical recipe to bypass the rare occurrence for which $\sigma_{2,K} = 0$. In order to fully compute $\widetilde{M}^{(j+1)}$, after computing $\widetilde{s}_K$ and $\widetilde{\mathbf{r}}_{1,K}$, we need only to compute the two eigenvalues $\widetilde{\lambda}_{1,K}$ and $\widetilde{\lambda}_{2,K}$ separately. This is achieved by resorting to the equidistribution principle ($\eta_K = \tau$, $\forall K$) cited previously, thus yielding:

$$\widetilde{\lambda}_{1,K} = \sqrt{\widetilde{s}_K q}, \qquad \widetilde{\lambda}_{2,K} = \sqrt{\frac{q}{\widetilde{s}_K}}, \tag{5.47}$$

with:

$$q := \left[\frac{\tau^4}{|\widehat{K}|^2(\widetilde{\rho}_K(v_h))^2\left(\widetilde{s}_K\sigma_{2,K} + \sigma_{1,K}/\widetilde{s}_K\right)}\right]^{1/3}. \tag{5.48}$$

The adaptive algorithm used in practice reads:

1. set $j = 0$ and build the background mesh $\mathcal{T}_h^{(j)}$;
2. solve the problem (5.5);
3. solve the local minimization problem (5.45) for $\widetilde{s}_K$ and $\widetilde{\mathbf{r}}_{1,K}$;
4. by means of the equidistribution principle, compute $\widetilde{\lambda}_{1,K}$ and $\widetilde{\lambda}_{2,K}$;

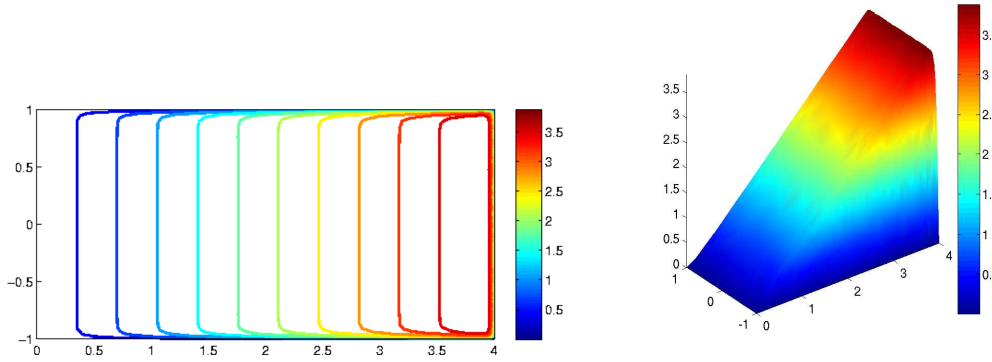

Figure 5.2: Test $E1$. Contourlines (left) and surface plot (right) of the exact solution.

5. build up the new metric $\widetilde{M}^{(j+1)}$;
6. construct the new mesh $\mathcal{T}_h^{(j+1)}$ matching the metric $\widetilde{M}^{(j+1)}$;
7. if a suitable stopping criterion is met, exit; else go back to step (2).

Remark 5.7. *If one is interested in the approach (a), the previous adaptive procedure can be adopted except for the choice of the tolerance τ which is not any more user–defined, but it depends on the desired number of elements.*

5.4.2 Numerical tests

We show now the reliability of the anisotropic a posteriori error estimator (5.36) by means of numerical tests. In particular, we compare the performance of such estimator with the isotropic corresponding one (see Remark 5.5), still driven by a metric based approach. The mesh generator employed in all test cases below is BAMG [85].

Test $E1$: the "ramp" case.

This is an academic test case with available exact solution aiming at providing a quantitative analysis of the proposed estimator. In more detail, referring to the ADR equation (5.1), we assume $\nu = 10^{-3}$, $\mathbf{V} = (1,0)^T$, $\gamma = 0$, $\Omega = (0,4) \times (-1,1)$, $\Gamma_D = \partial\Omega$, and the source term f is chosen s.t. the exact solution of Eq.(5.1) is:

$$v(\mathrm{x}_1,\mathrm{x}_2) = \mathrm{x}_1 \left(1 - e^{-50(4-\mathrm{x}_1)}\right) \left(1 - e^{-50(\mathrm{x}_2+1)} - e^{-50(1-\mathrm{x}_2)}\right), \tag{5.49}$$

which shows three boundary layers along the horizontal and outflow boundaries (see Fig.5.2). We notice that we are in the presence of a highly advection dominated problem as the global Peclét number is $\mathbb{P}e := \|\mathbf{V}\|_{L^\infty(\Omega)} L/(2\nu) = 2 \cdot 10^3$, with $L = 4$.

In the spirit of approaches (a) and (b) of the above adaptive procedure, we carry out two comparisons, with the same accuracy and with the same number of elements. In Fig.5.3 we compare the meshes obtained by employing the anisotropic (left) and isotropic (right) estimator with a similar accuracy on the exact error equal to 0.33. Notice that in the anisotropic

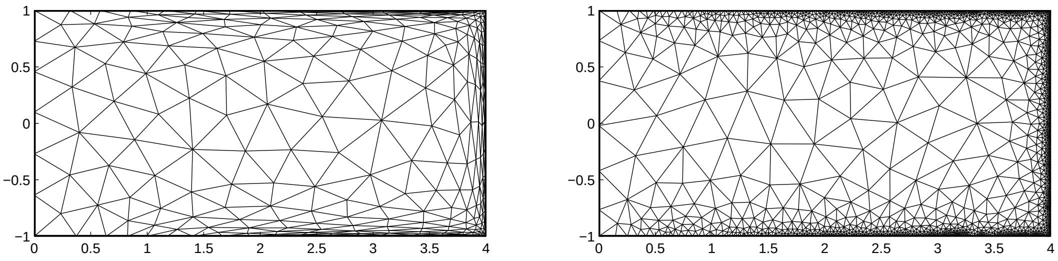

Figure 5.3: Test $E1$. Anisotropic (left) versus isotropic (right) adapted meshes with a similar accuracy (0.33).

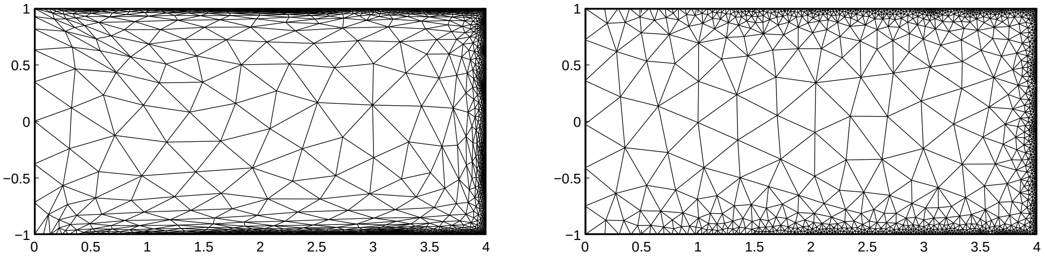

Figure 5.4: Test $E1$. Anisotropic (left) versus isotropic (right) adapted mesh with the same number of elements $(2,500)$.

case only 500 elements are required versus $3,200$ demanded in the isotropic case. Moreover, most of the triangles in the anisotropic case are stretched along the three boundary layers.

Fig.5.4 shows the meshes (anisotropic on the left and isotropic on the right) obtained by fixing the number of mesh elements, chosen equal to $2,500$. The energy norm of the error in the anisotropic case is $1.38 \cdot 10^{-1}$ compared with the value $3.92 \cdot 10^{-1}$ for the isotropic case, yielding a gain of $1/3$ in accuracy. Let us observe that the additional $2,000$ triangles of the anisotropic mesh w.r.t. Fig.5.3 are essentially "squeezed" in the boundary layers, the central area remaining almost unchanged. In Fig.5.5 we highlight a detail of the meshes in Fig.5.3(left) and Fig.5.4(right) in correspondence with a portion of the outflow boundary. In both the cases the anisotropic meshes are clearly stretched along the boundary layer, thus allowing to sharply capture the steep solution. Analogous comments hold for the horizontal boundary layers except that, as these are parabolic layers, their thickness is $O(\sqrt{\nu})$ while the outflow boundary is $O(\nu)$ large.

We check now the convergence properties of the proposed adaptive procedure. In particular, in Fig.5.6 we compare the convergence histories characterizing both the anisotropic and isotropic error estimators along with that associated with a uniform refinement. The trend for the first two estimators exhibits the same slope even if the line corresponding to the anisotropic estimator is shifted below the isotropic one. The mesh adaption driven by uniform refinement seems to lead to a slower convergence. The conclusions drawn from Figs 5.3 and 5.4 are further

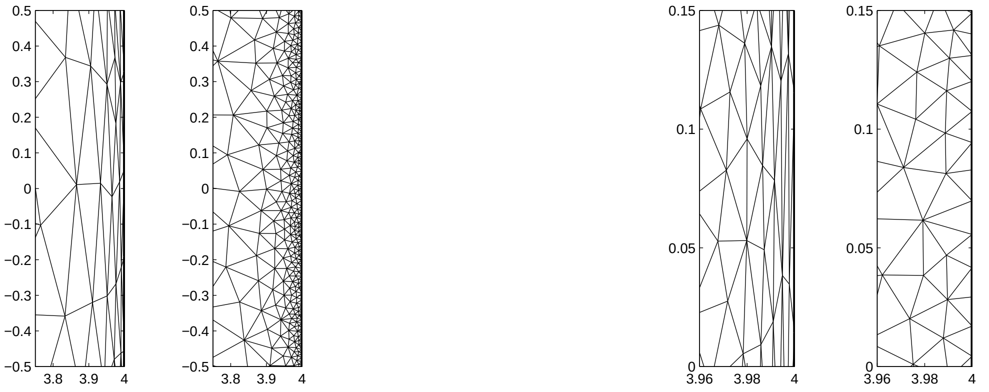

Figure 5.5: Test $E1$. Zooms of the adapted meshes in Fig.5.3(left) and 5.4(right).

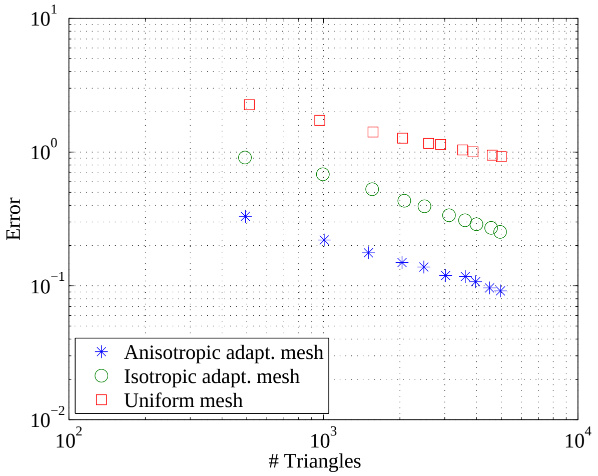

Figure 5.6: Test $E1$. Convergence histories associated with the uniform refinement (□), anisotropic adapted mesh ($*$) and the isotropic one (O).

	Anisotropic	Isotropic
$\|\|\|e_h\|\|\| \simeq 0.33$	7.86	108
$\sharp\mathcal{T}_h = 2500$	83.4	68.8

Table 5.1: Test $E1$. CPU time (in seconds) for the anisotropic and isotropic mesh adaption procedure with the same accuracy (first row) and number of elements (second row).

justified by Fig.5.6: the error reduction is about 1/3 for the anisotropic versus the isotropic case with the same number of elements, while the saving in the number of elements is about 6–7 times as much.

In Table 5.1 we deal with the computational cost issue[1]. From the first row the shorter

[1] The computations are carried out on an AMD Athlon 1.33 GHz processor, with 256 KB of Memory Cache and 256 MB of RAM.

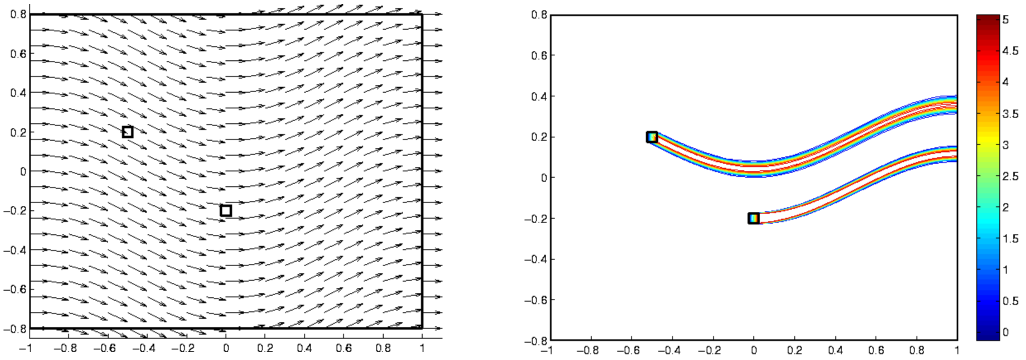

Figure 5.7: Test $E2$. Advective field (left) and contourlines of the reference solution (right).

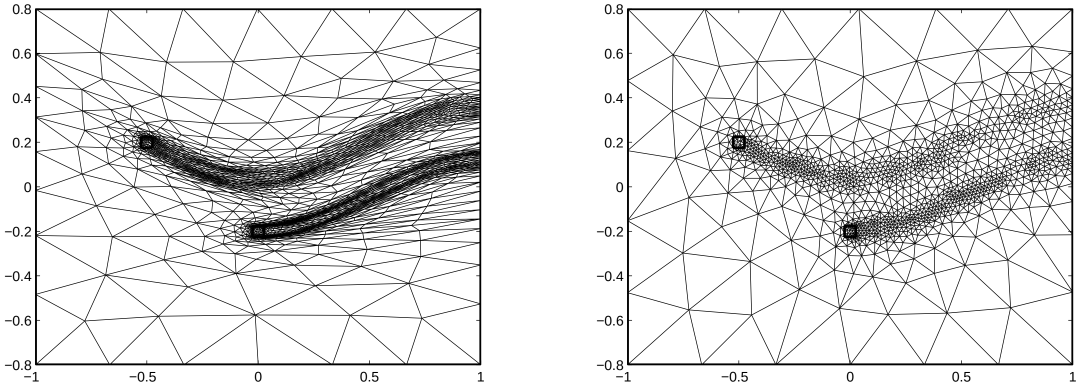

Figure 5.8: Test 2. Anisotropic (left) versus isotropic (right) adapted mesh with the same number of elements (2, 300).

CPU time of the anisotropic adaptive procedure is clearly due to the much lower number of elements at hand. On the other hand with the same number of elements the anisotropic procedure is slightly more expensive: this is no surprise, since the anisotropic error estimator involves the approximation error e_h^{pr} in its definition, this requiring an overhead due to the ZZ recovery procedure.

Test $E2$: the "two chimney" case.

We deal now with a problem motivated by a real environmental issue. In particular, we want to study the diffusion and the transport of a pollutant emitted in air by industrial chimneys in the presence of strong wind; see Chapter 1.

For this test case we choose in (5.1) $\nu = 10^{-3}$, $\mathbf{V} = (1, 0.5\sin(\pi \mathrm{x}_1))^T$, $\gamma = 0$, $\Omega = (-1, 1) \times (-0.8, 0.8)$, $\Gamma_D = \{\mathrm{x}_1 = -1\}$, $g = 0$ and $f = 100\chi_{E_1 \cup E_2}$, where χ_{E_i} is the characteristic function associated with the emission area E_i, with $i = 1, 2$, E_i being the squared subdomains centered at $(-0.5, 0.2)$ and $(0, -0.2)$, respectively, with side equal to 0.05. Fig.5.7 displays the

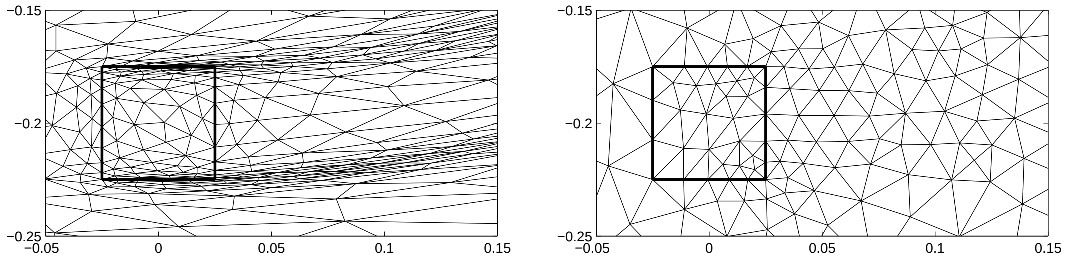

Figure 5.9: Test 2. Detail of the anisotropic (left) versus isotropic (right) adapted meshes around the lower chimney.

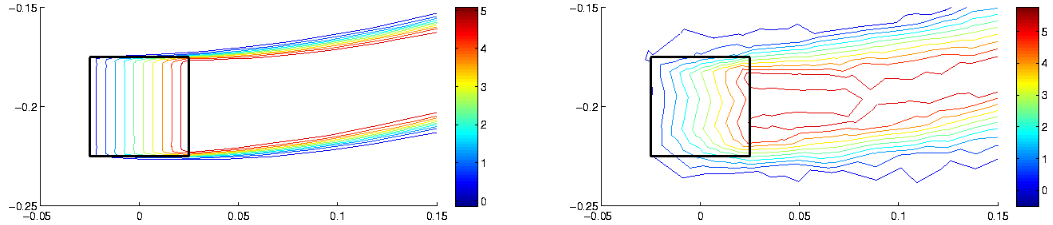

Figure 5.10: Test 2. Contourlines around the lower chimney on the anisotropic (left) and isotropic (right) meshes.

advective flow field (left) along with the reference solution (right) computed on a sufficiently fine uniform mesh.

In Fig.5.8 we contrast the meshes yielded by the anisotropic (left) and isotropic (right) adaptive procedure with the same number of triangles, i.e. $2,300$. By comparing Fig.5.8 with Fig.5.7 it is easily seen that both the adapted meshes detect the pollutant wakes. However, it is also evident that the anisotropic mesh fits more accurately all the internal layers characterizing the reference solution. In fact, zooming on the lower chimney (see Fig.5.9), it can be appreciated the “coarsening” in the middle of the wake in the anisotropic case which is completely absent in the isotropic mesh. Note also that the misleading uniform distribution of the elements in the isotropic case is actually due to the small dimension of the zoom box. The better accuracy of the anisotropic approximation is further corroborated by the contourlines in Fig.5.10, which are compact in the anisotropic case while scattered on the isotropic mesh. Also remarkable is the accuracy of the anisotropic solution inside the emission area.

5.5 The goal–oriented anisotropic analysis

In view of environmental applications we are interested in accurately approximating physically relevant quantities such as concentrations around critical areas of the domain, or fluxes across sections of interest. As anticipated in Chapter 1, these goal quantities can be mathematically represented by suitable linear (or nonlinear) output functionals $L(\cdot)$ of the solution ($s = L(v)$). The goal–oriented framework fully fits this need ([22, 61]) as highlighted and described in Chapter 4. The merging of this approach with the anisotropic framework has already been carried out for the ADR equation and also the Stokes problem in [52, 55]. In this Section after deriving an error estimate equivalent, up to a constant, to the one in [52], we provide an alternative approach for the case of a linear output functional $L(\cdot)$ preserving the nice feature of the estimate (5.24) to depend on the error, i.e. to likely converge at a faster rate. We observe that with such an estimate we aim at extending the Dual Weighted Residual (DWR) method introduced in Sec.4.1.5 into the anisotropic framework.

The main ingredient of the goal–oriented analysis is the dual problem related to the functional $L(\cdot)$ at hand, as we have discussed in Sec.4.1.1. Let $L : \mathcal{V} \to \mathbb{R}$ be the linear goal functional. By considering the "discretize–then–differentiate" approach discussed in Sec.4.1.4, the dual problem associated with the ADR (primal) problem (5.2) reads:

$$\text{find } z \in \mathcal{V} \ : \ a_h(\vartheta, z) = L(\vartheta) \qquad \forall \vartheta \in \mathcal{V}, \tag{5.50}$$

the stabilized bilinear form $a_h(\cdot,\cdot)$ being defined in Eq.(5.6). The corresponding approximated problem is:

$$\text{find } z_h \in \mathcal{V}_h \ : \ a_h(\vartheta_h, z_h) = L(\vartheta_h) \qquad \forall \vartheta_h \in \mathcal{V}_h. \tag{5.51}$$

We observe that the dual problem is defined without introducing explicitly the Lagrange functional formalism. This is due to the simplicity of the analysis for linear problems with linear output functionals; see Remark 4.3.
A first anisotropic bound on the functional $s - s_h = L(e_h^{pr})$ of the approximation error can now be stated, being $s_h := L(v_h)$.

Proposition 5.5. *Let v, $z \in \mathcal{V}$ be the solutions of the primal and dual problems (5.2) and (5.50), respectively, and v_h, $z_h \in \mathcal{V}_h$ the corresponding approximations satisfying Eq.s (5.5) and (5.51), respectively. Moreover, let us assume that v is smooth enough s.t. the standard Galerkin orthogonality (5.9) holds. Then, the following estimate can be proved:*

$$|L(e_h^{pr})| \leq C \sum_{K \in \mathcal{T}_h} \alpha_K \rho_K(v_h) w_K(z), \tag{5.52}$$

with $C = C(N, C_\Delta)$, α_K, $\rho_K(v_h)$ defined as in Proposition 5.3, while:

$$w_K(z) := \left[s_K \left(\mathbf{r}_{1,K}^T G_K(z) \mathbf{r}_{1,K} \right) + \frac{1}{s_K} \left(\mathbf{r}_{2,K}^T G_K(z) \mathbf{r}_{2,K} \right) \right]^{1/2}.$$

Proof. By suitably combining the dual formulation (5.50) together with the Galerkin orthogonality (5.9), the definitions of $a_h(\cdot,\cdot)$ in Eq.(5.6) and of the strong form of the ADR equation in Eq.(5.1), we get:

$$\begin{aligned} L(e_h^{pr}) &= a_h(e_h^{pr}, z) = a_h(e_h^{pr}, z - \phi_h) \\ &= a(e_h^{pr}, z - \phi_h) + \sum_{K \in \mathcal{T}_h} \tau_K \left(r_K(v_h), \mathbf{V} \cdot \nabla (z - \phi_h) \right) \qquad \forall \phi_h \in \mathcal{V}_h. \end{aligned} \tag{5.53}$$

From Eq.(5.29) with $\phi = z - \phi_h \in \mathcal{V}$, we have:

$$\begin{aligned} L(e_h^{pr}) &= \sum_{K \in \mathcal{T}_h} \left\{ \int_K r_K(v_h)(z - \phi_h) \, d\Omega + \frac{1}{2} \int_{\partial K} j_K(v_h)(z - \phi_h) \, d\Gamma \right. \\ & \left. + \tau_K \int_K r_K(v_h) \mathbf{V} \cdot \nabla (z - \phi_h) \, d\Omega \right\}. \end{aligned} \tag{5.54}$$

By using the Cauchy–Schwarz inequality, the functional of the error can be bounded as:

$$
\begin{aligned}
|L(e_h^{pr})| \quad \leq \quad \sum_{K\in\mathcal{T}_h} \Bigg\{ \quad & \|r_K(v_h)\|_{L^2(K)} \Big[\quad \|z-\phi_h\|_{L^2(K)} + \tau_K \|\mathbf{V}\|_{L^\infty(K)} |z-\phi_h|_{H^1(K)} \Big] \\
+ \quad & \frac{1}{2} \|j_K(v_h)\|_{L^2(\partial K)} \|z-\phi_h\|_{L^2(\partial K)} \quad \Bigg\} \qquad \forall \phi_h \in \mathcal{V}_h.
\end{aligned}
\tag{5.55}
$$

The choice $\phi_h = I_h^1 z \in \mathcal{V}_h$ together with Proposition 5.1 leads us to the final result (5.52), C coinciding now with the maximum of the interpolation constants C_1, C_2, C_3 in Proposition 5.1.

□

Remark 5.8. *The use of the stabilized dual weak form (5.50), at variance with [52], is justified in view of Proposition 5.6. The same kind of estimate is obtained in both cases, the only change consisting of a different value for the constant C in Eq.(5.52), with $C = \max(C_1, 1+C_2, C_3)$ in [52].*

As in the case of Proposition 5.3 the r.h.s. of Eq.(5.52) cannot be used yet as an a posteriori error estimator, the exact dual solution z being in general unknown. With this aim, let us introduce the following Definition.

Definition 5.3. *Let v_h, $z_h \in \mathcal{V}_h$ be the solutions of the approximated problems (5.5) and (5.51), respectively. Then, the error on the functional $L(\cdot)$ can be estimated by the quantity:*

$$
\eta^{L_1} := \sum_{K\in\mathcal{T}_h} \eta_K^{L_1}, \tag{5.56}
$$

being $\eta_K^{L_1}$ the local error indicator given by:

$$
\eta_K^{L_1} := \alpha_K \rho_K(v_h) w_K(z^\star) \tag{5.57}
$$

where α_K and $\rho_K(v_h)$ are still defined as in Proposition 5.3, while:

$$
w_K(z^\star) := \left[s_K \left(\mathbf{r}_{1,K}^T G_K(z^\star) \mathbf{r}_{1,K} \right) + \frac{1}{s_K} \left(\mathbf{r}_{2,K}^T G_K(z^\star) \mathbf{r}_{2,K} \right) \right]^{1/2}, \tag{5.58}
$$

$G_K(z^\star)$ being computed relying on the ZZ recovery procedure applied to z_h.

We have now all the tools necessary to introduce the variant of the estimator η^{L_1}, hopefully enjoying better approximation properties.

Proposition 5.6. *An estimate alternative to that reported in Eq.(5.52) for the functional of the primal approximation error is provided by the following relation:*

$$
|L(e_h^{pr})| \leq C \sum_{K\in\mathcal{T}_h} \alpha_K \rho_K(v_h) w_K(e_h^{du}), \tag{5.59}
$$

being $e_h^{du} := z - z_h$ the dual approximation error and with:

$$
w_K(e_h^{du}) := \left[s_K \left(\mathbf{r}_{1,K}^T G_K(e_h^{du}) \mathbf{r}_{1,K} \right) + \frac{1}{s_K} \left(\mathbf{r}_{2,K}^T G_K(e_h^{du}) \mathbf{r}_{2,K} \right) \right]^{1/2}. \tag{5.60}
$$

Proof. Result (5.59) follows simply by mimicking the proof of Proposition 5.5 choosing in the end $\phi_h = z_h + I_h^1(z - z_h)$. □

Analogously to estimate (5.24) the weight of the r.h.s. of Eq.(5.59) depends on the unknown error, in this case the dual one. We follow the same approach as in Sec.5.3 (see Eq.(5.35)) where the matrix $G_K(e_h^{du})$ in the definition of $w_K(e_h^{du})$ is replaced by the matrix of the recovered gradients $G_K(e_h^{du,\star})$ given by:

$$\left[G_K(e_h^{du,\star})\right]_{i,j} := \sum_{T\in\Delta_K} \int_T \left((\nabla^{ZZ} z_h)_i - \frac{\partial z_h}{\partial \mathrm{x}_i}\right)\left((\nabla^{ZZ} z_h)_j - \frac{\partial z_h}{\partial \mathrm{x}_j}\right) d\Omega, \tag{5.61}$$

for $i, j = 1, 2$. The fully computable estimator for $L(e_h^{pr})$ is provided in the following Definition.

Definition 5.4. *By adopting the same notation of Definition 5.3, the global anisotropic a posteriori error estimator for the functional of the approximation error e_h^{pr} is:*

$$\eta^{L_2} := \sum_{K\in\mathcal{T}_h} \eta_K^{L_2}, \tag{5.62}$$

being $\eta_K^{L_2}$ the local error indicator given by:

$$\eta_K^{L_2} := \alpha_K \rho_K(v_h) w_K(e_h^{du,\star}), \tag{5.63}$$

being the weight $w_K(e_h^{du,\star})$ defined as:

$$w_K(e_h^{du,\star}) = \left[s_K\left(\mathbf{r}_{1,K}^T G_K(e_h^{du,\star})\mathbf{r}_{1,K}\right) + \frac{1}{s_K}\left(\mathbf{r}_{2,K}^T G_K(e_h^{du,\star})\mathbf{r}_{2,K}\right)\right]^{1/2}. \tag{5.64}$$

5.6 Goal–oriented analysis: numerical assessment

This Section replies Sec.5.4 in the goal–oriented framework and with reference to the estimators (5.56) and (5.62). In particular, due to the similar structure of the estimator (5.36) with the new ones η^{L_1} and η^{L_2}, the description of the adaptive procedure is here kept to a minimum.

5.6.1 The adaptive procedure

The procedure employed to derive the constrained minimization problem (5.45) can be mimicked also for the goal–oriented estimators (5.56) and (5.62). The substantial differences are: the new definition of the objective function $I(s_K, \mathbf{r}_{1,K})$ involves the scaled matrices $\widetilde{G}_K(z^\star)$ and $\widetilde{G}_K(e_h^{du,\star})$ instead of the matrix $\widetilde{G}_K(e_h^{pr,\star})$; then, we notice that the local estimators (5.57) and (5.63) do not involve any square root in contrast to the local estimator (5.37). The analogue of Proposition 5.4 can thus be stated.

Proposition 5.7. *The optimal metric $\widetilde{M}$ is identified by the following choices:*

$$\widetilde{\lambda}_{1,K} = \sqrt{\widetilde{s}_K q}, \qquad \widetilde{\lambda}_{2,K} = \sqrt{\frac{q}{\widetilde{s}_K}}, \qquad \widetilde{\mathbf{r}}_{1,K} = \mathbf{p}_{2,K}, \tag{5.65}$$

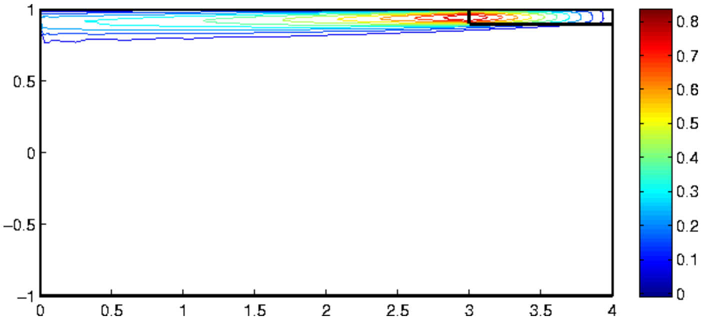

Figure 5.11: Test $G1$. Contourlines of the dual solution.

where:

$$\widetilde{s}_K := \sqrt{\frac{\sigma_{1,K}}{\sigma_{2,K}}}, \qquad q := \left[\frac{\tau^2}{|\widehat{K}|^2(\widetilde{\rho}_K(v_h))^2\,(\widetilde{s}_K\sigma_{2,K}+\sigma_{1,K}/\widetilde{s}_K)}\right]^{1/3}, \tag{5.66}$$

being $\sigma_{1,K}$ and $\sigma_{2,K}$ the maximum and the minimum eigenvalues of the matrix $\widetilde{G}_K(z^\star)$ or $\widetilde{G}_K(e_h^{du,\star})$ according with the choice η^{L_1} or η^{L_2}, respectively and with $\mathbf{p}_{i,K}$ the corresponding eigenvectors.

The adaptive algorithm detailed in Sec.5.4.1 will be exploited as is in the numerical validation below.

5.6.2 Numerical tests

We compare now the performances of the estimators η^{L_1} and η^{L_2} on three test cases, the first one purely academic, the others being instead aimed at environmental applications. In particular, as we are dealing with a goal–oriented approach, the linear functional $L(\cdot)$ in Eq.(5.50) will be chosen on the basis of environmental motivations.

Test $G1$: the "ramp" case.

We come back to the test case $E1$ in Sec.5.4.2, now reviewed in a goal–oriented setting. The target functional is $L(v) = \int_D v\, d\Omega = 0.2741$, D being the rectangle with coordinates $(3, 0.9)$, $(4, 0.9)$, $(4, 1)$ and $(3, 1)$ (see the boxed area in the top-right corner of Fig.5.11). The corresponding dual solution is displayed in Fig.5.11: as the dual source term involves a localized quantity and the problem is strongly advective dominated, the dual solution is confined to the upper horizontal slab of the domain, i.e. the region feeding the information into the region D.

In Fig.5.12 we show the anisotropic meshes driven by the estimators η^{L_1} (left) and η^{L_2} (right), sharing the same number of triangles (about $1,600$). A qualitative comparison among these meshes and the anisotropic ones in Fig.s 5.3 and 5.4 highlight the significant role played by the functional $L(\cdot)$ in the goal–oriented case: the three boundary layers exhibited by the primal solution are no longer detected, while only the region carrying information towards the rectangle D is refined. A suitable zoom of the meshes in Fig.5.12 in correspondence with the top side of the domain is provided in Fig.5.13. The two details emphasize the less anisotropic

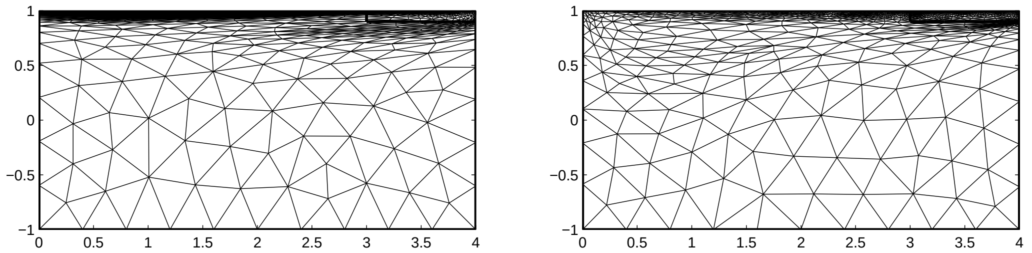

Figure 5.12: Test $G1$. Anisotropic adapted mesh driven by η^{L_1} (left) versus η^{L_2} (right) with the same number of elements $(1,600)$.

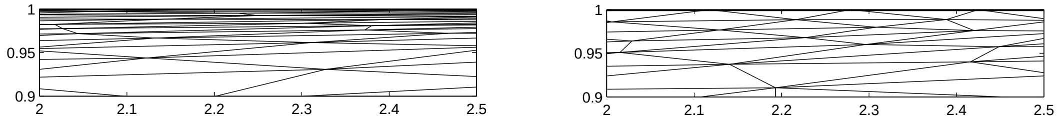

Figure 5.13: Test $G1$. Zooms of the meshes in Fig.5.12 associated with η^{L_1} (left) and η^{L_2} (right).

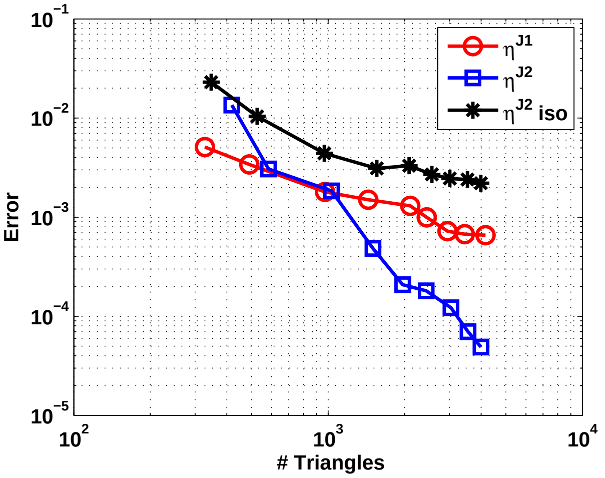

Figure 5.14: Test $G1$. Convergence histories associated with the anisotropic estimators η^{L_1} (O), η^{L_2} (□) and the isotropic estimator $\eta^{L_2}_{iso}$ $(*)$.

exasperated nature of the mesh identified by the estimator (5.62). This is to be expected due to the dependence of the matrix G_K on the dual error rather than on the dual solution.

Finally, we compare the convergence histories associated with η^{L_1}, η^{L_2} and the isotropic counterpart of η^{L_2}, i.e. $\eta^{L_2}_{iso}$ (see [22] for an instance of the corresponding recipe). As Fig.5.14 demonstrates, the anisotropic estimator (5.62) exhibits a faster convergence compared with both η^{L_1} and $\eta^{L_2}_{iso}$. This implies a lower number of triangles in order to guarantee a given

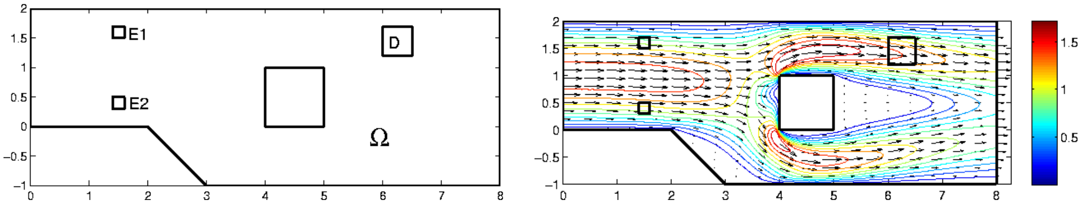

Figure 5.15: Test $G2$. Domain (left) and advective field $\mathbf{V}$ (right).

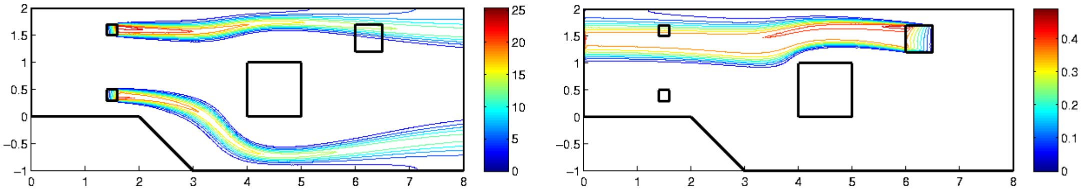

Figure 5.16: Test $G2$. Contourlines for the primal (left) and dual (right) solution.

accuracy or, likewise, a higher accuracy for a fixed number of d.o.f.'s.

Test $G2$: the "channel" case.

This test case deals with an environmental problem modeling transport of pollution in water. In particular, two pollutant sources E_1 and E_2 are placed in a river characterized by a dry area (an island, for example) localized 8in the center of the domain. Our aim is the evaluation of the average value of the pollutant concentration in a zone D of interest (e.g., a fish or beach area). In Fig.5.15(left) is represented this setting. With reference to the ADR equation (5.1), the advective field $\mathbf{V}$ (see Fig.5.15(right)) is computed by solving the incompressible Navier–Stokes equations with the following data: the Reynolds number is chosen equal to 100, a parabolic inflow profile with average value 1 is enforced at the inflow boundary $\{\mathrm{x}_1 = 0\}$, while a no slip condition holds on the land borders and a homogeneous Neumann condition is assigned at the outflow $\{\mathrm{x}_1 = 8\}$. As far as the other data of equation (5.1) is concerned, we take $\nu = 10^{-3}$, $\gamma = 0$, $\Gamma_D = \{\mathrm{x}_1 = 0\}$ and the source term $f = 100\,\chi_{E_1 \cup E_2}$. Finally, the output functional is $L(v) = \int_D v \; d\Omega$.

As reference solution we choose the approximation computed on a uniform mesh with $51,008$ triangles, thus the goal value being equal to 1.2355. The contourlines of the primal and dual solutions are displayed in Fig.5.16(left) and (right), respectively. We can appreciate that only the emission area E_1 influences the zone of interest D as a consequence of the strong horizontal advective field (notice also the perturbation on the dual solution due to the dry area). This is further confirmed by the adaptive meshes yielded by η^{L_1} and η^{L_2}, for a fixed number of elements (about 1,500), as shown in Fig.5.17(left) and (right), respectively). A zoom of the adapted mesh around the emission source E_1 is provided in Fig.5.18: the mesh associated with the estimator η^{L_1} is clearly the most anisotropic one.

Finally, concerning the converge history, similar conclusions as in Fig.5.14 can be drawn (see Fig.5.19).

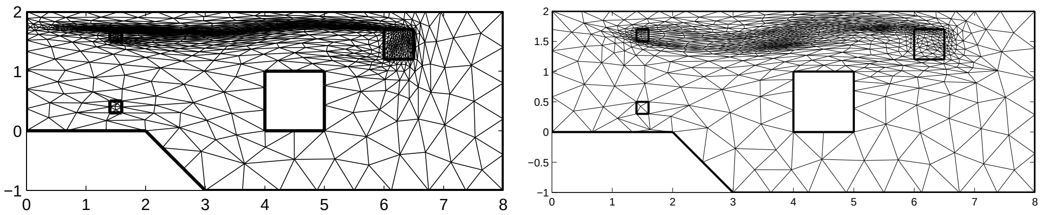

Figure 5.17: Test $G2$. Anisotropic adapted meshes driven by η^{L_1} (left) versus η^{L_2} (right) with the same number of elements $(1,500)$.

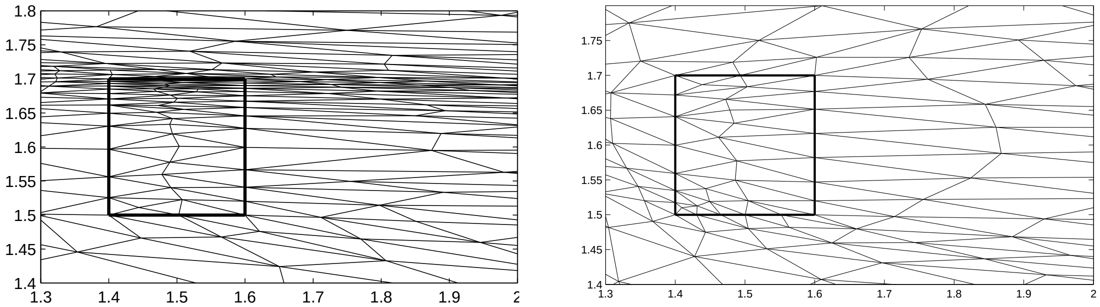

Figure 5.18: Test $G2$. Zooms of the meshes in Fig.5.17 associated with η^{L_1} (left) and η^{L_2} (right).

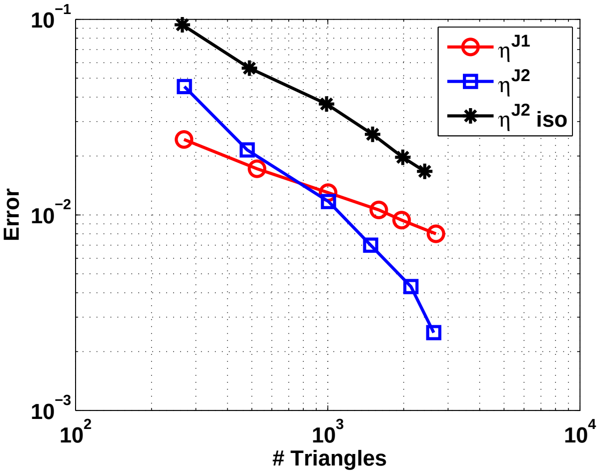

Figure 5.19: Test $G2$. Convergence histories associated with the anisotropic estimators η^{L_1} (O), η^{L_2} (□) and the isotropic estimator $\eta^{L_2}_{iso}$ $(*)$.

Test $G3$: the "PIT tag detection" case.

This test case can be regarded as an environmental problem concerning Radio Frequency IDentification (RFID) of animals. In particular, we deal with a PIT (Passive Integrated

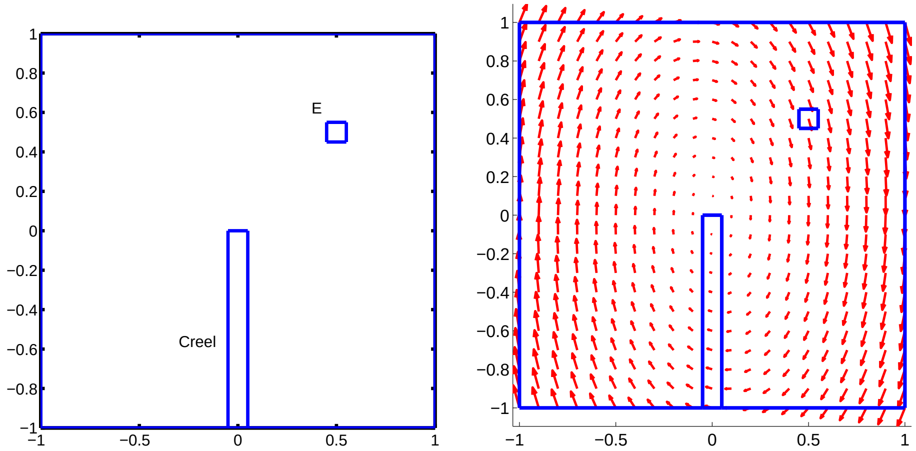

Figure 5.20: Test $G3$. Domain (left) and advective flow field (right).

Transponder) technology providing a variety of identification and monitoring solutions for fish and wildlife research (see e.g., [160]). PIT tags have been used for over twenty years to permanently identify individual animals. The small size of PIT tags, also known as "microchips", virtually eliminates negative impact on animals with little or no influence on growth-rate, behavior or health, and makes recapture unnecessary, thus reducing handling time and stress to the animal. The principle of RFID is to use a signal transmitted between an electronic device, such as a "tag", "transponder" or "microchip" and a reading device, such as a "scanner", "reader" or "transceiver". The RFID or EID (Electronic IDentification) devices most widely used in animals are passive. Passive integrated transponders have no battery so the microchip remains inactive until read with a scanner. The scanner sends a low frequency signal to the microchip within the tag providing the power needed to send its unique code back to the scanner and positively identify the animal.

In our test case we simulate a typical monitoring situation in the case of fishes, under the $2D$ approximation that the vertical motion of the fishes is negligible. Suppose that a school of fishes are PIT tagged and then continuously released in a small area off sea. We are then interested in measuring the fish flux across a rectangular creel located downwind a strong eddy. We also assume that the phenomenon takes place in an area whose size is large compared with the dimension of the fishes so that the fish random motion by "diffusion" is dominated by the convective effects. We model the fish evolution by the steady ADR equation (5.1). The domain $\Omega = (-1,1)^2$ is reported in Fig.5.20(left), along with the dump area E and the creel $(-0.05, 0.05) \times (-1, 0)$. The advective field $\mathbf{V}$ is approximated by the following elliptic contracting spiral:

$$\mathbf{V} = (V_1, V_2)^T = (\mathrm{x}_2 - 0.1\mathrm{x}_1, 3(-\mathrm{x}_1 - 0.1\mathrm{x}_2))^T\,, \tag{5.67}$$

with $\nabla \cdot \mathbf{V} = -0.4$, and the corresponding flow field is shown in Fig.5.20(right). As for the other data of the ADR equation (5.1), we take $\nu = 10^{-3}$, $\gamma = 0$, $f = 100\chi_E$, E being the squared release area of side 0.1 centered at (0.5,0.5), and $\Gamma_N = \emptyset$. The goal functional is given by $L(v) = -\int_{\mathrm{Creel}} V_1 v \, d\Omega \simeq 7.885 \cdot 10^{-2}$, as approximated on a uniform fine mesh consisting of 86,144 elements. The color plots of the reference primal and dual solutions are displayed in Fig.5.21(left) and (right), respectively.

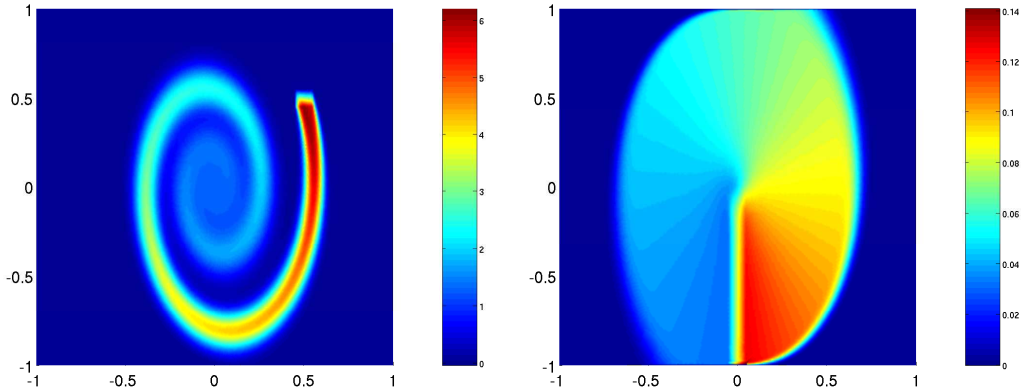

Figure 5.21: Test $G3$. Color plot of the primal (left) and dual (right) reference solutions.

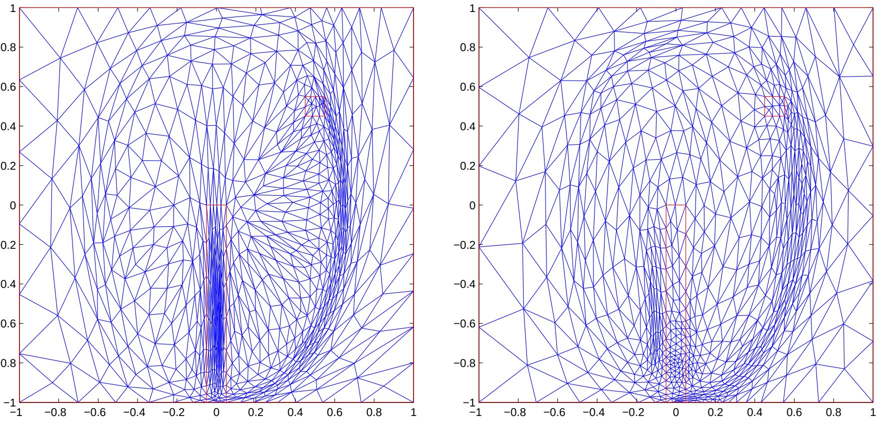

Figure 5.22: Test $G3$. Anisotropic adapted meshes driven by η^{L_1} (left) and η^{L_2} (right) with the same number of elements $(1,250)$.

In Fig.5.22 we show the anisotropic adapted meshes, with about $1,250$ triangles, obtained by means of the estimators η^{L_1} (left) and η^{L_2} (right). Both the meshes highlight the regions mostly affecting the computation of the output functional $L(\cdot)$ according with the interplay between the primal and the dual solutions. In particular, as shown in Fig.5.22, the dual solution provides us with a qualitative information by selecting the area involved in the adaptivity. On the other hand, the details of the mesh inside this area are due to both the strong gradients of the dual solution and to the intensity of the primal solution which gets lower and lower as we move towards the center of the domain (see Fig.5.21). Both the estimators detect the same critical area by placing anisotropic triangles along the primal streamline, though the layers associated with η^{L_2} are larger than the ones corresponding to η^{L_1}. Notice also that the orientation of the triangles around the center of the domain is quite different. In Fig.5.23 we provide the zooms around the creel of the adapted meshes of Fig.5.22. Observe the crowding and the strong anisotropic features of the triangles yielded by η^{L_1} (left) due to the large gradient of the dual solution in that area. On the other hand the estimator η^{L_2}, depending on the gradient of the error of the dual solution, exhibits less

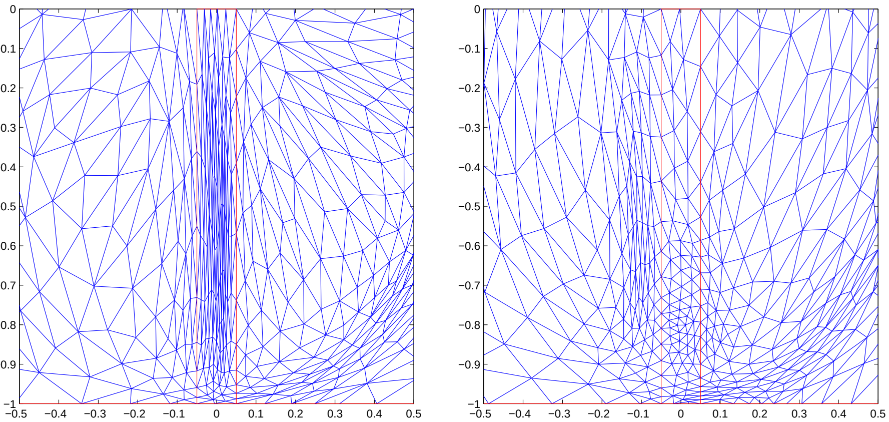

Figure 5.23: Test $G3$. Zooms around the creel of the meshes in Fig.5.22 associated with η^{L_1} (left) and η^{L_2} (right).

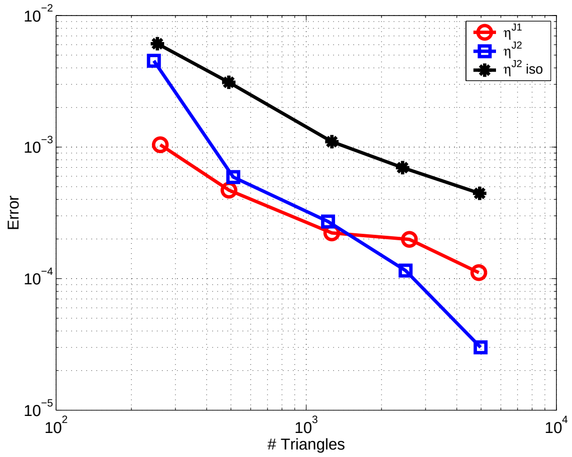

Figure 5.24: Test $G3$. Convergence histories associated with the anisotropic estimators η^{L_1} (O), η^{L_2} (□) and the isotropic version of estimator η^{L_2}, $\eta^{L_2}_{iso}$, $(*)$.

apparent anisotropic characteristics.

In Fig.5.24 we gather the convergence histories associated with the estimators η^{L_1}, η^{L_2} and $\eta^{L_2}_{iso}$. The same type of conclusions as in the previous test cases can be inferred: the estimator η^{L_2} allows for a better convergence rate w.r.t. η^{L_1}. The isotropic version of the estimator η^{L_2}, i.e., $\eta^{L_2}_{iso}$, leads to errors larger than those yielded by η^{L_1} and η^{L_2}, for a fixed number of elements, or equivalently, to a saving of d.o.f's when the same tolerance is considered.

We now consider a variant of the previous test case in which we include the effect of a strong (and possibly fraudulent) fishing activity taking place downwind the monitoring region. We model this phenomenon by adding a reaction term γ of value 100 in the region $F = (-1, -0.25) \times (-0.05, 0.05)$. The color plots of the reference primal and dual solutions, computed on the same fine mesh as above, are shown in Fig.5.25(left) and (right), respectively. Note the "barrier" effect due to the presence of the fishing area. The goal functional is still

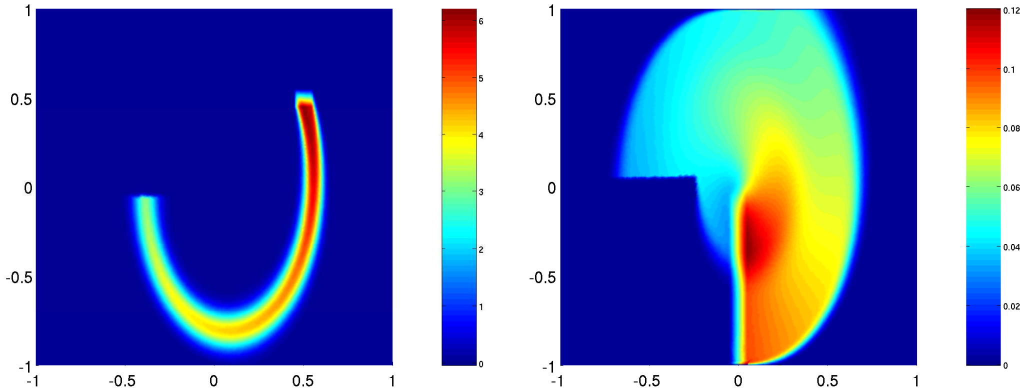

Figure 5.25: Test $G3$ with fishing. Color plot of the primal (left) and dual (right) reference solutions.

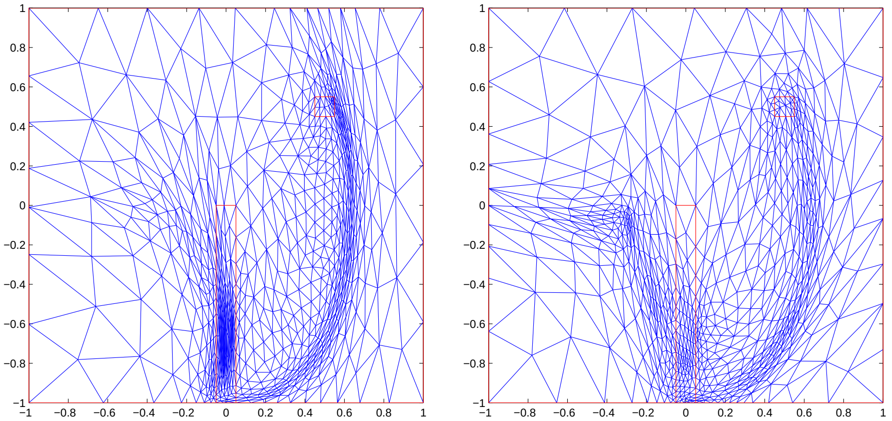

Figure 5.26: Test $G3$ with fishing. Anisotropic adapted mesh driven by η^{L_1} (left) and η^{L_2} (right) with the same number of elements $(1,250)$.

$L(v) = -\int_{\text{Creel}} V_1 v \, d\Omega \simeq 6.057 \cdot 10^{-2}$, which is slightly lower than the previous case due the capture of fishes.

In Fig.5.26 we provide the anisotropic adapted meshes, with about $1,250$ triangles, associated with the estimators η^{L_1} (left) and η^{L_2} (right). On comparing Fig.5.22 with Fig.5.26, the most striking evidence is that in the left and top parts of the domain the mesh is not refined as the fishing activity is actually interrupting the flow. Moreover, the refinement localized along the region F in Fig.5.26(right) is maybe due to a poor approximation of the dual solution z. Fig.5.27 details the adapted meshes in Fig.5.26 around the creel. As in Fig.5.23, it is apparent the more anisotropic nature of the mesh associated with η^{L_1}.

In Fig.5.28 we plot the convergence histories corresponding to the estimators η^{L_1}, η^{L_2} and $\eta^{L_2}_{iso}$. Similar conclusions as in the previous test cases hold.

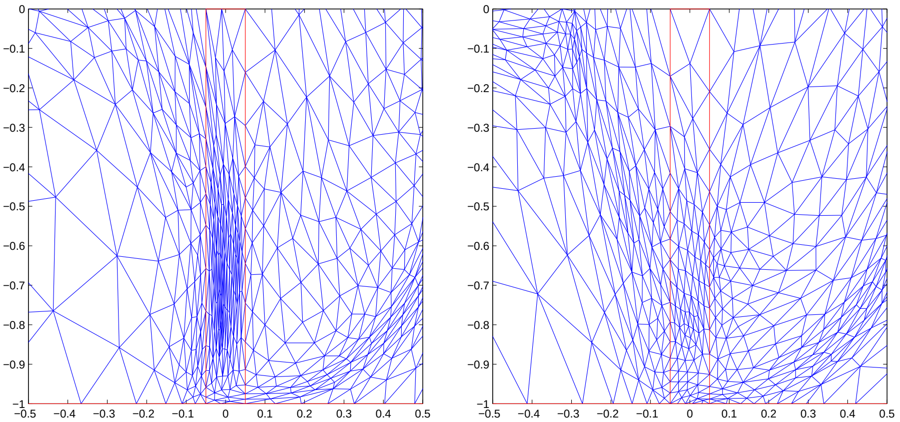

Figure 5.27: Test $G3$ with fishing. Zooms of the meshes in Fig.5.26 associated with η^{L_1} (left) and η^{L_2} (right).

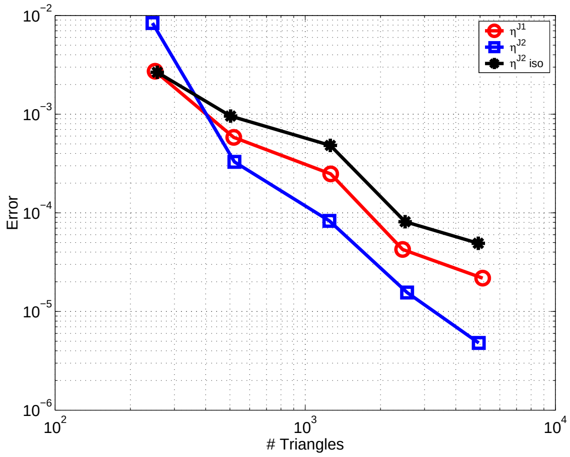

Figure 5.28: Test $G3$ with fishing. Convergence histories associated with the anisotropic estimators η^{L_1} (O), η^{L_2} (□) and the isotropic version of estimator η^{L_2}, $\eta^{L_2}_{iso}$, ($*$).

5.7 Concluding remarks

We have shown via some basic test cases that the mesh adaption procedure driven by the anisotropic a posteriori error estimate can be effective in tackling environmental applications modeled by the ADR equation (5.1), especially in the presence of strong advective fields. In more detail, for a desired accuracy the CPU time reduces considerably due to the lower number of elements required by the anisotropic procedure. In addition, the combined use of anisotropic adaption together with a goal–oriented analysis turns out to be resolvent in view of the monitoring of quantities of physical interest in environmental applications.

Chapter 6

Reduced Basis Method for Parametrized Advection–Reaction Problems

The Reduced Basis (RB) method is a computational approach which allows rapid and reliable predictions of functional outputs of elliptic and parabolic Partial Differential Equations (PDEs) with parametric dependence [72, 129, 136, 145]. In recent works, other kinds of problems have been considered too, particularly also the case of parametrized hyperbolic PDEs has been afforded [135]. Indeed, the RB method has a wide range of relevant applications in the characterization of engineering components or systems which require the prediction of certain "quantities of interest" or performance metrics: e.g., deflections, maximum stresses, maximum temperatures, heat transfer rates, flow rates, aerodynamical forces or momentum. This method has already been applied to the cases of Stokes and Navier–Stokes equations [111, 150, 164, 165, 180], as well as to linear elasticity problems [129, 179] and many other physical applications (see e.g. [32, 70, 145]). As anticipated in Chapter 1, environmental problems represent a promising field of application for the RB method. Preliminary investigations have been considered in [149, 151] for pollution problems in air, for which the RB method has been adopted to evaluate the concentration of pollutants emitted by industrial sites in certain zones of observation, such as cities [44]. Parametrized steady advection–diffusion PDEs have been adopted in order to describe such phenomena. In particular, both geometrical and physical parameters have been considered, such as the location of industrial plants, the intensity or direction of the wind field, or the diffusion coefficient.

In this Chapter we deal with the investigation of the RB method for the evaluation of outputs, which depend on the solution of parametrized advection–reaction PDEs; this in view of environmental applications, for which the diffusion phenomena are negligible w.r.t. the transport and reaction ones. The solution of such PDEs stands for the concentration of a pollutant in a medium, such as air or water, in a $2D$ domain.

The RB method is based on the decoupling of the generation and the projection stages of the approximation procedures, which leads to a decoupled offline–online computational approach. The complexity of the offline step, in which the basis are generated, depends on the dimension of the "truth" space, say N_t, to which belongs the "truth" solution corresponding to a given value of the parameters vector. The complexity of the online stage depends on the dimension of the RB space, say N, with $N \ll N_t$, and the complexity of the parametric dependence.

For the definition of the "truth" space, we make use of the Finite Element (FE) method [153]. In order to get rid of the instabilities caused by the transport term of the hyperbolic advection–reaction PDE, we use the Streamline Diffusion Finite Element (SDFE) stabilized method [153, 187]. This leads to the transformation of the original hyperbolic PDE into a new one, with elliptic nature. We define the RB method for this parametrized stabilized advection–reaction problem, for which the affine decomposition property holds. Moreover, we consider the *"primal–dual"* RB approach [129, 136, 168], which requires the definition of a dual problem. This approach is well–suited both for the approximation and the error evaluation of the output and, as we highlight in this work, also for the reduction of the computational costs associated with the RB online stage w.r.t. those of the "only primal" RB approach (without the dual problem). We provide a priori RB error estimates for both the solution and the output, thus evidencing the role of the FE approximation and stabilization in the RB method. In fact, the total error on the solution or the output is composed by two parts corresponding to the FE and RB approximations, respectively. In particular, we show that the "complexity" of the RB increases, as the FE approximation improves by reducing the mesh size (i.e. the parametrized stabilized problem tends to assume an hyperbolic nature). We also provide the a posteriori RB error estimate for the output according with [136, 168]. We remark that the idea of using stabilized FE for the definition of the "truth" space has been already introduced in [149], even if a priori and a posteriori RB estimates, as well as an error analysis for the FE and RB approximations, have not been provided. Then, we propose a general adaptive algorithm for the choice of the sample sets (used for the definition of the RB basis), for which the a posteriori RB estimate is used. We base our adaptive algorithm on a criterium of minimization of the computational costs associated with the online step. Moreover, a general sensitivity analysis of the output w.r.t. the parameters is provided directly in the RB context, in order to inherit the advantage of the offline–online decomposition.
Two numerical tests, inspired by environmental problems, are provided dealing with both physical and geometrical parametrizations. The tests highlight the effectiveness of the proposed adaptive algorithm and the savings of online computational costs allowed by the "primal–dual" RB approach. Moreover, we experimentally show that the RB approximation is stable, if the FE one is stable.

The Chapter is organized as follows. In Sec.6.1 we introduce the parametrized advection–reaction PDEs in an abstract setting. We consider two specific problems (regarded as Problem 1 and Problem 2) with physical and geometrical parameters. A particular case of Problem 1, say Problem 1(bis), is also introduced. In Sec.6.2 we provide the FE approximation of the parametrized problem, after having introduced the stabilization by means of the SDFE method. An a priori error analysis for the Problem 1(bis) is provided in order to evaluate the effects of the stabilization in the FE approximation. Some estimates and other mathematical tools, which we use in the following Sections, are introduced as well. Sec.6.3 deals with the RB method, which is described for the particular problem in consideration; the "primal–dual" RB approach is considered for the solution of the RB problem. Both a priori and a posteriori RB estimates are provided. Moreover, the adaptive algorithm for the choice of the samples sets is outlined. Finally, a sensitivity analysis on the output is considered. In Sec.6.4 we report some considerations about the numerical solution of the parametrized advection–reaction PDEs, by using jointly the FE and the RB methods. We provide in Sec.6.5 two numerical tests referring to Problems 1 and 2 introduced in Sec.6.1, which prove the effectiveness of the proposed procedures. Concluding remarks follow.

6.1 Parametrized advection–reaction equations

In this Section we introduce the parametrized advection–reaction PDEs, which describe the problem of interest in an abstract setting and, in view of the numerical tests of Sec.6.4, we specify two particular problems with physical and geometrical parameters.

6.1.1 An abstract parametrized problem

Let us indicate with $\boldsymbol{\mu}$ the vector of parameters s.t. $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_P) \in \mathcal{D}$, being $\mathcal{D} \subset \mathbb{R}^P$, with $P \in \mathbb{N}$ the domain of parameters. We consider the following advection–reaction PDE:

$$\begin{cases} \mathbf{V}(\boldsymbol{\mu}) \cdot \nabla v + \gamma(\boldsymbol{\mu}) v = f(\boldsymbol{\mu}) & \text{in } \Omega, \\ v = 0 & \text{on } \Gamma_D, \end{cases} \tag{6.1}$$

where $\Omega \subset \mathbb{R}^2$ is a bi–dimensional domain with boundary $\partial\Omega$. The parametrized advection field $\mathbf{V}(\boldsymbol{\mu}) \in [L^\infty(\Omega)]^2$ $\forall \boldsymbol{\mu} \in \mathcal{D}$ is chosen s.t. $\nabla \cdot \mathbf{V}(\boldsymbol{\mu}) = 0$ $\forall \boldsymbol{\mu} \in \mathcal{D}$; in the same manner we define the parametrized reaction term $\gamma(\boldsymbol{\mu}) \in L^\infty(\Omega)$ $\forall \boldsymbol{\mu} \in \mathcal{D}$, s.t. $\gamma(\boldsymbol{\mu}) > 0$ $\forall \boldsymbol{\mu} \in \mathcal{D}$, and the parametrized source term $f(\boldsymbol{\mu}) \in L^2(\Omega)$ $\forall \boldsymbol{\mu} \in \mathcal{D}$. Let us observe that, for the sake of simplicity, we have omitted to explicitly express the dependence of $\mathbf{V}(\boldsymbol{\mu})$, $\gamma(\boldsymbol{\mu})$ and $f(\boldsymbol{\mu})$ on the spatial coordinate $\mathbf{x} \in \mathbb{R}^2$, which should be read as $\mathbf{V}(\boldsymbol{\mu}, \mathbf{x})$, $\gamma(\boldsymbol{\mu}, \mathbf{x})$ and $f(\boldsymbol{\mu}, \mathbf{x})$ respectively. We also observe that not all the functions $\mathbf{V}$, γ and f must necessarily depend on the set of parameters.

Moreover, we suppose that the parametrized data admit the *affine decomposition* property, e.g.: $\mathbf{V}(\boldsymbol{\mu}) = \mathbf{V}(\boldsymbol{\mu}, \mathbf{x}) = \left(\sum_{i=1}^{M_{V1}} \Theta_i^{V1}(\boldsymbol{\mu}) g_i^{V1}(\mathbf{x}), \sum_{j=1}^{M_{V2}} \Theta_j^{V2}(\boldsymbol{\mu}) g_j^{V2}(\mathbf{x})\right)$, with $\Theta_i^{V1}(\boldsymbol{\mu})$, $\Theta_j^{V2}(\boldsymbol{\mu}) \in C^1(\mathcal{D})$, $g_i^{V1}(\mathbf{x})$, $g_j^{V2}(\mathbf{x}) \in L^\infty(\Omega)$, $i = 1, \ldots, M_{V1}$, $j = 1, \ldots, M_{V2}$, for some M_{V1}, $M_{V2} \in \mathbb{N}$. In the same manner the reaction term reads: $\gamma(\boldsymbol{\mu}) = \gamma(\boldsymbol{\mu}, \mathbf{x}) = \sum_{i=1}^{M_\gamma} \Theta_i^\gamma(\boldsymbol{\mu}) g_i^\gamma(\mathbf{x})$, with $\Theta_j^\gamma(\boldsymbol{\mu}) \in C^1(\mathcal{D})$ and $g_i^\gamma(\mathbf{x}) \in L^\infty(\Omega)$, $i = 1, \ldots, M_\gamma$, for some $M_\gamma \in \mathbb{N}$; finally, the source term is: $f(\boldsymbol{\mu}) = f(\boldsymbol{\mu}, \mathbf{x}) = \sum_{i=1}^{M_f} \Theta_i^f(\boldsymbol{\mu}) g_i^f(\mathbf{x})$, with $\Theta_j^f(\boldsymbol{\mu}) \in C^1(\mathcal{D})$ and $g_i^f(\mathbf{x}) \in L^2(\Omega)$, $i = 1, \ldots, M_f$, for some $M_f \in \mathbb{N}$.

We define the part of the boundary $\partial\Omega$ indicated with Γ_D as $\Gamma_D(\boldsymbol{\mu}) := \{\mathbf{x} \in \partial\Omega \ : \ \mathbf{V}(\boldsymbol{\mu}) \cdot \hat{\mathbf{n}} < 0 \ \forall \boldsymbol{\mu} \in \mathcal{D}\}$ which corresponds to the inflow boundary, being $\hat{\mathbf{n}}$ the outward directed unit vector normal to $\partial\Omega$. Moreover, we assume that the boundary $\Gamma_D(\boldsymbol{\mu})$ is "fixed", in the sense that $\Gamma_D(\boldsymbol{\mu}) = \Gamma_D$ $\forall \boldsymbol{\mu} \in \mathcal{D}$; finally, we define Γ_N as $\Gamma_N := \partial\Omega \backslash \Gamma_D$, i.e. the outflow boundary. The weak form of problem (6.1) reads:

$$\text{find } \ v(\boldsymbol{\mu}) \in \mathcal{V} \ \ : \ \ a(v(\boldsymbol{\mu}), \phi; \boldsymbol{\mu}) = F(\phi; \boldsymbol{\mu}) \qquad \forall \phi \in \mathcal{V}, \ \ \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.2}$$

where $\mathcal{V} := H^1_{\Gamma_D}(\Omega)$, being $H^1_{\Gamma_D}(\Omega)$ the usual Hilbert space of functions with null trace on Γ_D (see e.g. [109]), and:

$$\begin{aligned} a(w, \phi; \boldsymbol{\mu}) &:= \int_\Omega \left(\mathbf{V}(\boldsymbol{\mu}) \cdot \nabla w \, \phi + \gamma(\boldsymbol{\mu}) \, w \, \phi\right) \, d\Omega, \\ F(\phi; \boldsymbol{\mu}) &:= \int_\Omega f(\boldsymbol{\mu}) \, \phi \, d\Omega. \end{aligned} \tag{6.3}$$

Let us observe that, due to the affine decomposition assumptions made for the data $\mathbf{V}(\boldsymbol{\mu})$,

$\gamma(\boldsymbol{\mu})$ and $f(\boldsymbol{\mu})$, the bilinear form $a(\cdot,\cdot;\boldsymbol{\mu})$ and the functional $F(\cdot;\boldsymbol{\mu})$ can be rewritten as:

$$\begin{aligned} a(w,\phi;\boldsymbol{\mu}) &= \sum_{q=1}^{Q} \vartheta_q(\boldsymbol{\mu}) a_q(w,\phi), \\ F(\phi;\boldsymbol{\mu}) &= \sum_{q=1}^{Q^F} \vartheta_q^F(\boldsymbol{\mu}) F_q(\phi). \end{aligned} \tag{6.4}$$

for some $Q \in \mathbb{N}$ and $Q^F \in \mathbb{N}$ and the bilinear forms $a_q(\cdot,\cdot)$ and the linear functionals $F_q(\cdot)$ not depending on the parameter vector $\boldsymbol{\mu}$.

Our goal consists in calculating an output $s(\boldsymbol{\mu})$ depending on the parameter $\boldsymbol{\mu}$:

$$s(\boldsymbol{\mu}) = L\left(v(\boldsymbol{\mu});\boldsymbol{\mu}\right) \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.5}$$

where $L(\cdot;\boldsymbol{\mu})$ is a linear and continuous functional acting from $\mathcal{V}$ to $\mathbb{R}$, s.t.:

$$L(\phi;\boldsymbol{\mu}) := \int_\Omega \sigma(\boldsymbol{\mu})\, \phi \; d\Omega, \tag{6.6}$$

being $\sigma(\boldsymbol{\mu})$ a parameter function for which the same hypothesis of affine decomposition made for $f(\boldsymbol{\mu})$ holds, s.t.:

$$s(\boldsymbol{\mu}) = \sum_{q=1}^{Q^L} \vartheta_q^L(\boldsymbol{\mu}) L_q\left(v(\boldsymbol{\mu})\right), \tag{6.7}$$

for some $Q^L \in \mathbb{N}$ and the linear functionals $L_q(\cdot)$ independent on $\boldsymbol{\mu}$.

6.1.2 Problem 1: physical parametrization

In this Section we consider a particular case of the general advection–reaction problem described in Sec.6.1.1, with a physical parameter; in the following Sections, we will refer to this problem as "Problem 1".

We set $\boldsymbol{\mu} = \mu$, with μ a physical parameter which can be regarded as the magnitude of an advection field, whose direction is given by the vector $\mathbf{b} \in \mathbb{R}^2$. By referring to the abstract problem (6.1) we chose $\mathbf{V}(\mu) = \mu\mathbf{b}$, $\gamma(\mu) = 1$ and $f(\mu) = g$ (independent on μ); moreover, we assume $\mathcal{D} = [\mu_{min}, \mu_{max}] \subset \mathbb{R}$ and $\sigma(\mu) = \delta$ (independent on μ). The corresponding advection–reaction problem reads:

$$\begin{cases} \mu\mathbf{b}\cdot\nabla v + v = g & \text{in } \Omega, \\ v = 0 & \text{on } \Gamma_D, \end{cases} \tag{6.8}$$

for which the generic weak form (6.2) holds with the following bilinear form $a(\cdot,\cdot;\mu)$ and functional $F(\cdot;\mu)$:

$$\begin{aligned} a(w,\phi;\mu) &= \int_\Omega \left(\mu\mathbf{b}\cdot\nabla w\,\phi + w\,\phi\right)\; d\Omega, \\ F(\phi;\mu) &= \int_\Omega g\,\phi\; d\Omega, \end{aligned} \tag{6.9}$$

while the output functional reads:

$$L(\phi;\mu) = \int_\Omega \delta\,\phi\; d\Omega. \tag{6.10}$$

Let us observe that in this case we have $Q = 2$, $Q^F = 1$ and $Q^L = 1$, $\vartheta_1(\mu) = \mu$, $\vartheta_2(\mu) = 1$, $\vartheta_1^F(\mu) = 1$, $\vartheta_1^L(\mu) = 1$, while the decomposed forms and functionals read: $A_1(w,\phi) = \int_\Omega \mathbf{b}\cdot\nabla w\,\phi\; d\Omega$, $A_2(w,\phi) = \int_\Omega w\,\phi\; d\Omega$, $F_1(\phi) \equiv F(\phi;\mu)$ and $L_1(\phi) \equiv L(\phi;\mu)$, both independent on μ.

6.1.3 Problem 2: physical and geometrical parametrization

We consider now a parametrized problem with geometrical and physical parameters; this allows us to vary the shape of domain by acting on the values of the geometrical parameter (see [150, 162, 163, 164, 165]). We will refer to this problem as "Problem 2".

Let us introduce a geometrical parameter μ_g, s.t. $\boldsymbol{\mu} = (\mu_p, \mu_g) \in \mathcal{D} \subset \mathbb{R}^2$, where μ_p is the physical parameter introduced in Sec.6.1.2. We remark that the domain depends now on μ_g; from a computational point of view this could lead to a relevant computational effort in terms of time–work. In order to overcome this difficulty, we map the real domain into a reference domain, which is "fixed" as the geometrical parameter vary, and we transform the original problem into a new one set on the reference domain.
By indicating with the subscript 0 the quantities defined on the real domain $\Omega_0 = \Omega_0(\mu_g)$, the parametrized advection–reaction PDE reads:

$$\begin{cases} \mu_p \mathbf{b}_0 \cdot \nabla_0 v_0 + v_0 = g_0 & \text{in } \Omega_0(\mu_g), \\ v_0 = 0 & \text{on } \Gamma_{0D}(\mu_g), \end{cases} \tag{6.11}$$

where we have adopted the same notation of Sec.6.1.2. The problem in weak form reads:

$$\text{find } \; v_0(\boldsymbol{\mu}) \in \mathcal{V}_0 \quad : \quad a_0(\phi_0(\boldsymbol{\mu}), \phi; \boldsymbol{\mu}) = F_0(\phi; \boldsymbol{\mu}) \qquad \forall \phi \in \mathcal{V}_0, \;\; \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.12}$$

where $\mathcal{V}_0 := H^1_{\Gamma_{0D}}(\Omega_0(\mu_g))$ and:

$$\begin{aligned} a_0(w,\phi;\boldsymbol{\mu}) &:= \int_{\Omega_0(\mu_g)} \left(\mu_p \mathbf{b}_0 \cdot \nabla_0 w\,\phi + w\,\phi\right)\; d\Omega_0(\mu_g), \\ F_0(\phi;\boldsymbol{\mu}) &:= \int_{\Omega_0(\mu_g)} g_0\,\phi\; d\Omega_0(\mu_g). \end{aligned} \tag{6.13}$$

In the same manner the output functional $s(\boldsymbol{\mu})$ reads:

$$s(\boldsymbol{\mu}) = L_0(v_0(\boldsymbol{\mu}); \mu_g) \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.14}$$

with:

$$L_0(\phi;\mu_g) := \int_{\Omega_0(\mu_g)} \delta_0\,\phi\; d\Omega_0(\mu_g). \tag{6.15}$$

Let us suppose that an *affine map*, acting from a reference domain Ω to the real domain $\Omega_0(\mu_g)$, could be provided and expressed in the following form:

$$\mathbf{x}_0 = T(\mu_g)\mathbf{x} + \mathbf{t}(\mu_g), \tag{6.16}$$

being the tensor $T(\mu_g) \in \mathbb{R}^{2\times 2}$ and the vector $\mathbf{t}(\mu_g) \in \mathbb{R}^2$. In the case that the domains Ω_0 and Ω are partitioned into subdomains Ω_{0i}, Ω_i, s.t. $\cup_i \Omega_{0i} = \Omega_0$ and $\cup_i \Omega_i = \Omega$, it is necessary to define an affine map for each subdomain; however, for the sake of simplicity, we consider only the case of non partitioned domains, even if it is a straightforward matter to generalize to the case with subdomains.

By using the affine map, the weak problem (6.8) can be rewritten in weak form as in Eq.(6.2), where $\mathcal{V} := H^1_{\Gamma_D}(\Omega)$ and $a(\cdot,\cdot;\boldsymbol{\mu})$, $F(\cdot;\boldsymbol{\mu})$ are defined in Eq.(6.3) and set on the reference domain Ω. By referring to the general problem of Sec.6.1.1 and by inspection of the weak form (6.2), it is possible to deduce the data:

$$\begin{aligned}
\mathbf{V}(\boldsymbol{\mu}) &= \mu_p \det(T(\mu_g))\ T(\mu_g)^{-T}\ \mathbf{b}(\mu_g),\\
\gamma(\boldsymbol{\mu}) &= \det(T(\mu_g)),\\
f(\boldsymbol{\mu}) &= \det(T(\mu_g))\ g(\mu_g),\\
\sigma(\boldsymbol{\mu}) &= \det(T(\mu_g))\ \delta(\mu_g),
\end{aligned} \tag{6.17}$$

where $\mathbf{b}$, g and δ correspond to the data $\mathbf{b}_0$, g_0 and δ_0 given on the real domain, respectively.

6.2 Finite Element approximation: stabilization

In this Section we consider the Finite Element (FE) method for the numerical approximation of the hyperbolic advection–reaction PDE introduced in Sec.6.1. With this aim, we add suitable stabilization terms to the weak form of the problem according with the Streamline Diffusion Finite Element (SDFE) method (see e.g. [96, 97, 187]). Moreover, we provide an a priori FE error estimate and some estimates which we adopt in Sec.6.3.2 for the a priori reduced basis error estimate.

6.2.1 Stabilization: the SDFE method

We introduce now the Streamline Diffusion Finite Element (SDFE) method for the numerical approximation of the abstract problem introduced in Sec.6.1.1, as proposed in [96] and discussed in [187].

Let us indicate with $\{K\}$ the triangular elements of a quasi–uniform unstructured mesh $\mathcal{T}_h$ of the domain Ω, s.t. $\cup_{K\in\mathcal{T}_h} = \overline{\Omega}$, and with h the largest diameter of the mesh triangles, i.e. $h := \max_{K\in\mathcal{T}_h} \mathrm{diam}(K)$. For the FE approximation we use piecewise linear basis functions on $K \in \mathcal{T}_h$ and we define the space $X_h := \left\{ w \in C^0(\overline{\Omega})\ :\ w_{|K} \in \mathbb{P}^1(K)\ \ \forall K \in \mathcal{T}_h \right\}$.

The stabilized approximated weak form of the problem (6.1) reads:

$$\text{find } \ v_h(\boldsymbol{\mu}) \in \mathcal{V}_h \ \ :\ \ a_h(v_h(\boldsymbol{\mu}), \phi_h; \boldsymbol{\mu}) = F_h(\phi_h; \boldsymbol{\mu}) \qquad \forall \phi_h \in \mathcal{V}_h,\ \ \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.18}$$

where $\mathcal{V}_h \subset \mathcal{V}$ is the FE space, being $\mathcal{V}_h := \{w_h \in X_h\ :\ w_h(\mathbf{x}) = 0\ \ \forall \mathbf{x} \in \Gamma_D\}$, and, from

Eq.(6.3):

$$\begin{aligned} a_h(w,\phi;\boldsymbol{\mu}) \quad := \quad & \varepsilon_h(h,\boldsymbol{\mu}) \int_\Omega \nabla w \cdot \nabla\phi \; d\Omega \\ & + \delta_h(h,\boldsymbol{\mu}) \int_\Omega (\mathbf{V}(\boldsymbol{\mu}) \cdot \nabla w) \;\; (\mathbf{V}(\boldsymbol{\mu}) \cdot \nabla\phi) \;\; d\Omega \\ & + \delta_h(h,\boldsymbol{\mu}) \int_\Omega w \; \mathbf{V}(\boldsymbol{\mu}) \cdot \nabla\phi \; d\Omega \\ & + a(w,\phi;\boldsymbol{\mu}), \\ F_h(\phi;\boldsymbol{\mu}) \quad := \quad & \delta_h(h,\boldsymbol{\mu}) \int_\Omega f(\boldsymbol{\mu}) \; \mathbf{V}(\boldsymbol{\mu}) \cdot \nabla\phi \; d\Omega \\ & + F(\phi;\boldsymbol{\mu}). \end{aligned} \tag{6.19}$$

The coefficients $\varepsilon_h(h,\boldsymbol{\mu})$ and $\delta_h(h,\boldsymbol{\mu})$ are chosen as in [96, 187]:

$$\varepsilon_h(h,\boldsymbol{\mu}) := C_\varepsilon(\boldsymbol{\mu})h^{3/2}, \qquad \delta_h(h,\boldsymbol{\mu}) := C_\delta(\boldsymbol{\mu})h, \tag{6.20}$$

and are considered "small". By indicating with V and L the measure units of the advection field $\mathbf{V}(\boldsymbol{\mu})$ and the length respectively, we observe that, for dimensional reasons applied to the weak form (6.18), $[C_\varepsilon(\boldsymbol{\mu})] = V/L^{1/2}$ and $[C_\delta(\boldsymbol{\mu})] = 1/V$. For example, for the Problems 1 and 2 outlined in Sec.s 6.1.2 and 6.1.3, we choose $C_\varepsilon(\boldsymbol{\mu}) = c_\varepsilon \mu_p$ and $C_\delta(\boldsymbol{\mu}) = c_\delta/\mu_p$.
Let us observe that the continuous stabilized version of the problem (6.18) would read:

$$\text{find} \;\; v_c(\boldsymbol{\mu}) \in \mathcal{V} \;\; : \;\; a_h(v_c(\boldsymbol{\mu}), \phi_c; \boldsymbol{\mu}) = F_h(\phi_c;\boldsymbol{\mu}) \qquad \forall \phi_c \in \mathcal{V}, \;\; \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.21}$$

where the continuous solution $v_c(\boldsymbol{\mu}) \in \mathcal{V}$ of the stabilized problem differs from the continuous solution $v(\boldsymbol{\mu}) \in \mathcal{V}$ of the original problem in weak form (6.2).
Let us notice that the stabilization acts only on the weak form (6.2), but not directly on the functional output (6.5). The output corresponding to the FE solution is:

$$s_h(\boldsymbol{\mu}) := L(v_h(\boldsymbol{\mu});\boldsymbol{\mu}). \tag{6.22}$$

Remark 6.1. *The stabilized advection–reaction problem (6.18) corresponds to an elliptic PDE, for which boundary conditions are necessary on the whole boundary $\partial\Omega$ and not only on Γ_D. The choice made for the FE space $\mathcal{V}_h$ consists in assuming implicitly homogeneous Neumann condition on Γ_N.*

Let us observe that, due to the affine decomposition assumptions made in Sec.6.1.1 on the problem data $\mathbf{V}(\boldsymbol{\mu})$, $\gamma(\boldsymbol{\mu})$ and $f(\boldsymbol{\mu})$, even for the stabilized problem, as the "original" one (see Eq.(6.4)), the affine decomposition property holds, i.e.:

$$\begin{aligned} a_h(w,\phi;\boldsymbol{\mu}) &= \sum_{q=1}^{Q_h} \vartheta_{hq}(\boldsymbol{\mu}) a_{hq}(w,\phi), \\ F_h(\phi;\boldsymbol{\mu}) &:= \sum_{q=1}^{Q_h^F} \vartheta_{hq}^F(\boldsymbol{\mu}) F_{hq}(\phi). \end{aligned} \tag{6.23}$$

for some $Q_h, Q_h^F \in \mathbb{N}$.
By writing the FE solution as:

$$v_h(\boldsymbol{\mu}) = \sum_{j=1}^{N_h} v_{hj}(\boldsymbol{\mu})\varphi_j, \tag{6.24}$$

being φ_j the FE lagrangian basis function associated with the FE space $\mathcal{V}_h$ and N_h the number of the nodes of the triangulation $\mathcal{T}_h$; it follows that the FE problem corresponds to the following linear system:

$$\text{find } \boldsymbol{v}_h(\boldsymbol{\mu}) \in \mathbb{R}^{N_h} \ : \ A_h(\boldsymbol{\mu})\boldsymbol{v}_h(\boldsymbol{\mu}) = \mathbf{F}_h(\boldsymbol{\mu}), \tag{6.25}$$

where $(\boldsymbol{v}_h(\boldsymbol{\mu}))_i = v_{hi}(\boldsymbol{\mu})$; the matrix $A_h(\boldsymbol{\mu}) \in \mathbb{R}^{N_h \times N_h}$ and the vector $\mathbf{F}_h(\boldsymbol{\mu}) \in \mathbb{R}^{N_h}$ are defined as:

$$\begin{aligned} A_h(\boldsymbol{\mu}) &:= \sum_{q=1}^{Q_h} \vartheta_{hq}(\boldsymbol{\mu}) A_{hq} \qquad \text{with } (A_{hq})_{i,j} := a_{hq}(\varphi_i, \varphi_j), \\ \mathbf{F}_h(\boldsymbol{\mu}) &:= \sum_{q=1}^{Q_h^F} \vartheta_q^F(\boldsymbol{\mu}) \mathbf{F}_{hq} \qquad \text{with } (\mathbf{F}_{hq})_i := F_{hq}(\varphi_i). \end{aligned} \tag{6.26}$$

By recalling Eq.s (6.5), (6.7) and (6.22) and by defining the vector $\mathbf{L}_h(\boldsymbol{\mu}) \in \mathbb{R}^{N_h}$ as:

$$\mathbf{L}_h(\boldsymbol{\mu}) := \sum_{q=1}^{Q^L} \vartheta_q^L(\boldsymbol{\mu}) \mathbf{L}_{hq} \qquad \text{with } (\mathbf{L}_{hq})_i := L_q(\varphi_i) \tag{6.27}$$

the output can finally be computed as:

$$s_h(\boldsymbol{\mu}) = \boldsymbol{v}_h(\boldsymbol{\mu}) \cdot \mathbf{L}_h(\boldsymbol{\mu}). \tag{6.28}$$

Let us now introduce the stabilized *dual problem* associated with the weak form (6.18) and the output (6.5):

$$\text{find } \ z(\boldsymbol{\mu}) \in \mathcal{V} \ \ : \ \ a_h(\phi, z(\boldsymbol{\mu}); \boldsymbol{\mu}) = -L(\phi; \boldsymbol{\mu}) \qquad \forall \phi \in \mathcal{V}, \ \ \forall \boldsymbol{\mu} \in \mathcal{D} \tag{6.29}$$

and the corresponding approximated problem:

$$\text{find } \ z_h(\boldsymbol{\mu}) \in \mathcal{V}_h \ \ : \ \ a_h(\phi_h, z_h(\boldsymbol{\mu}); \boldsymbol{\mu}) = -L(\phi_h; \boldsymbol{\mu}) \qquad \forall \phi_h \in \mathcal{V}_h, \ \ \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{6.30}$$

Let us observe that, for the sake of simplicity, we have avoided the subscript 'c' for $z(\boldsymbol{\mu}) \in \mathcal{V}$ in Eq.(6.29). In fact, we stress the fact that Eq.(6.29) does not represent the continuous dual problem, but only the continuous version of the stabilized one. We observe that the minus signum on the r.h.s. of Eq.s (6.29) and (6.30) is introduced to take into account for the standard notation of the RB method.

Remark 6.2. *In Eq.s (6.29) and (6.30) we have defined the dual problems by implicitly using the Lagrangian functional formalism according with the "discretize–then–differentiate" approach discussed in Sec.4.1.4.*

In analogy with the primal FE problem, the FE dual solution corresponds to the solution of the following linear system:

$$\text{find } \boldsymbol{z}_h(\boldsymbol{\mu}) \in \mathbb{R}^{N_h} \ : \ A_h(\boldsymbol{\mu})^T \boldsymbol{z}_h(\boldsymbol{\mu}) = -\mathbf{L}_h(\boldsymbol{\mu}), \tag{6.31}$$

where $(\boldsymbol{z}_h(\boldsymbol{\mu}))_i = z_{hi}(\boldsymbol{\mu})$, being $z_h(\boldsymbol{\mu}) = \sum_{j=1}^{N_h} z_{hj}(\boldsymbol{\mu})\varphi_j$ similarly to Eq.(6.24). In the same manner as the stabilized primal problem (6.18), also the dual one admits an affine decomposition, due to the assumption made for $\sigma(\boldsymbol{\mu})$ in Sec.6.1.1.

Remark 6.3. *Let us now consider an advection–diffusion–reaction problem with diffusion coefficient ε "small", s.t. the solution of this elliptic problem assumes a hyperbolic behavior. Moreover, let us assume an homogeneous Dirichlet condition on the whole boundary $\partial\Omega$. The FE problem should be stabilized in order to avoid numerical instabilities. If we consider the SDFE method, the stabilized problem assumes the form (6.18), where we have neglected the diffusion term ε, being $\varepsilon \ll \varepsilon_h$ due to the hypothesis ε "small". Let us observe that in this case we have $\mathcal{V}_h := \{w_h \in X_h \ : \ w_h(\mathbf{x}) = 0 \ \ \forall \mathbf{x} \in \partial\Omega\}$. We refer to this problem in the Sec.6.2.2 as Problem 1(bis) for the definition of a priori error estimate.*

6.2.2 A priori FE error estimate

In this Section we consider the a priori FE error estimate associated with the solution of the stabilized problem (6.18) according with the SDFE method and the output $s(\boldsymbol{\mu})$. First of all, we provide the a priori FE error estimate for the Problem 1(bis) described in Remark 6.3; then, we provide some estimates for the general problem endowed with stabilization described in Sec.6.2.1.

Problem 1(bis)

By considering the Problem 1(bis) described in Remark 6.3, we recall here the a priori error estimate of the solution of the problem (6.18) and the corresponding output (for more details, see [187]).

Let us define the following norm, which depends on the parameter vector $\boldsymbol{\mu}$:

$$|||w|||^2 := \varepsilon_h(h, \boldsymbol{\mu})||\nabla w||^2 + \delta_h(h, \boldsymbol{\mu})||\mathbf{V}(\boldsymbol{\mu}) \cdot \nabla w||^2 + ||w||^2, \tag{6.32}$$

where with $||\cdot||$ we indicate the usual $L^2(\Omega)$ norm ([109]).

It is possible to show that the stabilized Problem 1(bis) admits an unique solution, being the form $a_h(\cdot, \cdot; \boldsymbol{\mu})$ bilinear, continuous and coercive and the functional $F_h(\cdot; \boldsymbol{\mu})$ linear and continuous. In fact, from Eq.s (6.3) and (6.19) and by integration by parts of the term $\int_\Omega (\mathbf{V}(\boldsymbol{\mu}) \cdot \nabla w_h)\, \phi_h \, d\Omega$, we obtain the following estimate:

$$\begin{aligned} |a_h(w_h, \phi_h; \boldsymbol{\mu})| \quad \leq \quad & \varepsilon_h(h, \boldsymbol{\mu})\|\nabla w_h\| \, \|\nabla \phi_h\| \\ & + \delta_h(h, \boldsymbol{\mu}) \, \|\mathbf{V}(\boldsymbol{\mu}) \cdot \nabla w_h\| \, \|\mathbf{V}(\boldsymbol{\mu}) \cdot \nabla \phi_h\| \\ & + ||\gamma(\boldsymbol{\mu})||_\infty \, \|w_h\| \, \|\phi_h\| \\ & + \delta_h(h, \boldsymbol{\mu})\|w_h\| \, \|\mathbf{V}(\boldsymbol{\mu}) \cdot \nabla \phi_h\| \\ & + \|w_h\| \, \|\mathbf{V}(\boldsymbol{\mu}) \cdot \nabla \phi_h\|, \end{aligned} \tag{6.33}$$

being $\Gamma_D \equiv \partial\Omega$ and $\nabla \cdot \mathbf{V}(\boldsymbol{\mu}) = 0 \ \forall \boldsymbol{\mu} \in \mathcal{D}$. Let us observe that, for the sake of simplicity, $||\gamma(\boldsymbol{\mu})||_\infty$ indicates $||\gamma(\boldsymbol{\mu})||_{L^\infty(\Omega)}$. By observing from Eq.(6.32) that $\|\mathbf{V}(\boldsymbol{\mu}) \cdot \nabla\phi_h\| \leq 1/\delta_h(h,\boldsymbol{\mu})^{1/2}|||\phi_h|||$ and that $\delta_h(h,\boldsymbol{\mu}) < \delta_h(h,\boldsymbol{\mu})^{1/2}$, being $\delta_h(h,\boldsymbol{\mu})$ “small”, we obtain:

$$|a_h(w_h, \phi_h; \boldsymbol{\mu})| \leq \Big[\max\{1, ||\gamma(\boldsymbol{\mu})||_\infty\} \ |||w_h||| + (\delta_h(h,\boldsymbol{\mu}))^{-1/2} \, ||w_h||\Big] \, |||\phi_h|||, \tag{6.34}$$

from which the continuity of the form $a_h(\cdot,\cdot;\boldsymbol{\mu})$ follows:

$$|a_h(w_h, \phi_h; \boldsymbol{\mu})| \leq \Big[\max\{1, ||\gamma(\boldsymbol{\mu})||_\infty\} + (\delta_h(h,\boldsymbol{\mu}))^{-1/2}\Big] \ |||w_h||| \ |||\phi_h||| \qquad \forall w_h, \phi_h \in \mathcal{V}_h, \tag{6.35}$$

In the same manner it is possible to show that the form $a_h(\cdot,\cdot;\boldsymbol{\mu})$ is coercive:

$$a_h(\phi_h, \phi_h; \boldsymbol{\mu}) \geq \alpha(\boldsymbol{\mu}) \ |||\phi_h|||^2 \qquad \forall \phi_h \in \mathcal{V}_h, \tag{6.36}$$

where $\alpha(\boldsymbol{\mu})$ is the coercivity constant:

$$\alpha(\boldsymbol{\mu}) := \min\{1, ||\gamma(\boldsymbol{\mu})||_\infty\}. \tag{6.37}$$

The same considerations hold also for the corresponding stabilized dual problem, which admits an unique solution.

By defining the error $e_h^{pr}(\boldsymbol{\mu}) \in \mathcal{V}_h$ as $e_h^{pr}(\boldsymbol{\mu}) := v(\boldsymbol{\mu}) - v_h(\boldsymbol{\mu})$ and by introducing the linear nodal interpolant $I_h v(\boldsymbol{\mu})$ of the exact solution $v(\boldsymbol{\mu})$ (see e.g. [153]), we can write the error as $e_h^{pr}(\boldsymbol{\mu}) = e^{pr,I_h}(\boldsymbol{\mu}) + e_h^{pr,I_h}(\boldsymbol{\mu})$, where $e^{pr,I_h}(\boldsymbol{\mu}) := v(\boldsymbol{\mu}) - I_h v(\boldsymbol{\mu})$ and $e_h^{pr,I_h}(\boldsymbol{\mu}) := I_h v(\boldsymbol{\mu}) - v_h(\boldsymbol{\mu})$; the error $|||e_h^{pr}(\boldsymbol{\mu})|||$ can be estimated as it follows:

$$|||e_h^{pr}(\boldsymbol{\mu})||| \leq |||e^{pr,I_h}(\boldsymbol{\mu})||| + |||e_h^{pr,I_h}(\boldsymbol{\mu})|||. \tag{6.38}$$

If we assume $v(\boldsymbol{\mu}) \in H^2(\Omega) \cap \mathcal{V}$, the following estimates hold $\forall \boldsymbol{\mu} \in \mathcal{D}$ ([153]):

$$\begin{aligned} ||e^{pr,I_h}(\boldsymbol{\mu})|| &\leq C_1 h^2 |v(\boldsymbol{\mu})|_{H^2(\Omega)}, \\ ||\nabla e^{pr,I_h}(\boldsymbol{\mu})|| &\leq C_2 h |v(\boldsymbol{\mu})|_{H^2(\Omega)}) \end{aligned} \tag{6.39}$$

where $C_1, C_2 \in \mathbb{R}$. Due to $||\mathbf{V}(\boldsymbol{\mu}) \cdot \nabla\phi|| \leq ||\mathbf{V}(\boldsymbol{\mu})|| \ ||\nabla\phi||$ and Eq.(6.32), we have, $\forall \boldsymbol{\mu} \in \mathcal{D}$:

$$|||e^{pr,I_h}(\boldsymbol{\mu})||| \leq C_1 h \left\{ h + \frac{C_2}{C_1} \Big[(\varepsilon_h(h,\boldsymbol{\mu}))^{1/2} + (\delta_h(h,\boldsymbol{\mu}))^{1/2} ||\mathbf{V}(\boldsymbol{\mu})|| \Big] \right\} |v(\boldsymbol{\mu})|_{H^2(\Omega)}. \tag{6.40}$$

According with the choice made for the coefficients $\varepsilon_h(h,\boldsymbol{\mu})$ and $\delta_h(h,\boldsymbol{\mu})$ in Eq.(6.20), the previous inequality can be finally written as:

$$|||e^{pr,I_h}(\boldsymbol{\mu})||| \leq \widetilde{C}(\boldsymbol{\mu}) \ h^{3/2} \ |v(\boldsymbol{\mu})|^2_{H^2(\Omega)}, \tag{6.41}$$

where the terms of order greater than 3/2 in h have been neglected and the constant $\widetilde{C}(\boldsymbol{\mu}) \in \mathbb{R}$ depends on $C_\varepsilon(\boldsymbol{\mu})$, $C_\delta(\boldsymbol{\mu})$, $||\mathbf{V}(\boldsymbol{\mu})||$ and C_1, C_2.

In order to complete the a priori FE error estimate (6.38), we need to bound the term $|||e_h^{pr,I_h}(\boldsymbol{\mu})|||$; with this aim, let us recall the following inequality:

$$a_h(e_h^{pr,I_h}(\boldsymbol{\mu}), e_h^{pr,I_h}(\boldsymbol{\mu}); \boldsymbol{\mu}) \leq |a_h(e_h^{pr}(\boldsymbol{\mu}), e_h^{pr,I_h}(\boldsymbol{\mu}); \boldsymbol{\mu})| + |a_h(e^{pr,I_h}(\boldsymbol{\mu}), e_h^{pr,I_h}(\boldsymbol{\mu}); \boldsymbol{\mu})|. \tag{6.42}$$

For the problem (6.18) the Galerkin "skew orthogonality" property holds, i.e.:

$$a_h(e_h^{pr}(\boldsymbol{\mu}), \phi_h; \boldsymbol{\mu}) = \varepsilon_h(h, \boldsymbol{\mu}) \int_\Omega \nabla v(\boldsymbol{\mu}) \cdot \nabla \phi_h \; d\Omega \qquad \forall \phi_h \in \mathcal{V}_h, \quad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.43}$$

where we have considered $\varepsilon = 0$, being $\varepsilon << \varepsilon_h(h, \boldsymbol{\mu})$ (see Remark 6.3). By integration by parts and being $\Gamma_D \equiv \partial\Omega$, from Eq.(6.43) it follows:

$$a_h(e_h^{pr}(\boldsymbol{\mu}), \phi_h; \boldsymbol{\mu}) = -\varepsilon_h(h, \boldsymbol{\mu}) \int_\Omega \Delta v(\boldsymbol{\mu}) \; \phi_h \; d\Omega \qquad \forall \phi_h \in \mathcal{V}_h, \tag{6.44}$$

from which:

$$|a_h(e_h^{pr}(\boldsymbol{\mu}), e_h^{pr,I_h}(\boldsymbol{\mu}); \boldsymbol{\mu})| \leq \varepsilon_h(h, \boldsymbol{\mu}) |v(\boldsymbol{\mu})|_{H^2(\Omega)} ||e_h^{pr,I_h}(\boldsymbol{\mu})||; \tag{6.45}$$

according with Eq.(6.20) and being $||\cdot|| < |||\cdot|||$ (see Eq.(6.32)), the previous estimate reads:

$$|a_h(e_h^{pr}(\boldsymbol{\mu}), e_h^{pr,I_h}(\boldsymbol{\mu}); \boldsymbol{\mu})| \leq C_\varepsilon(\boldsymbol{\mu}) h^{3/2} |v(\boldsymbol{\mu})|_{H^2(\Omega)} |||e_h^{pr,I_h}(\boldsymbol{\mu})|||. \tag{6.46}$$

Then, by using the estimate (6.34), the second term of Eq.(6.42) can be bounded as:

$$\begin{aligned} |a_h(e^{pr,I_h}(\boldsymbol{\mu}), e_h^{pr,I_h}(\boldsymbol{\mu}); \boldsymbol{\mu})| \;\; \leq \;\; & \Big[\max\{1, ||\gamma(\boldsymbol{\mu})||_\infty\} \; |||e^{pr,I_h}(\boldsymbol{\mu})||| \\ & + (\delta_h(h, \boldsymbol{\mu}))^{-1/2} \, ||e^{pr,I_h}(\boldsymbol{\mu})|| \Big] \;\; |||e_h^{pr,I_h}(\boldsymbol{\mu})|||, \end{aligned} \tag{6.47}$$

which, combined with the definitions (6.20) and the estimates (6.39) and (6.41), reads:

$$|a_h(e^{pr,I_h}(\boldsymbol{\mu}), e_h^{pr,I_h}(\boldsymbol{\mu}); \boldsymbol{\mu})| \leq \overline{C}(\boldsymbol{\mu}) h^{3/2} |v(\boldsymbol{\mu})|_{H^2(\Omega)} |||e_h^{pr,I_h}(\boldsymbol{\mu})|||, \tag{6.48}$$

where the constant $\overline{C}(\boldsymbol{\mu}) \in \mathbb{R}$ depends on $C_\varepsilon(\boldsymbol{\mu})$, $C_\delta(\boldsymbol{\mu})$, $||\mathbf{V}(\boldsymbol{\mu})||$, $||\gamma(\boldsymbol{\mu})||_\infty$ and C_1, C_2. Finally, by combining Eq.s (6.42), (6.46) and (6.48) and by considering the coercivity property (6.36), we obtain the following a priori error estimate:

$$|||e_h^{pr}(\boldsymbol{\mu})||| \leq C(\boldsymbol{\mu}) h^{3/2} |v(\boldsymbol{\mu})|_{H^2(\Omega} \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.49}$$

with $C(\boldsymbol{\mu})$ depending on $C_\varepsilon(\boldsymbol{\mu})$, $C_\delta(\boldsymbol{\mu})$, $||\mathbf{V}(\boldsymbol{\mu})||$, $||\gamma(\boldsymbol{\mu})||_\infty$, C_1, C_2; the estimate (6.49) shows the convergence rate $3/2$ in h.

Let us observe that, if we consider the Problem 1 endowed with the Problem 1(bis) properties, the estimate (6.49) reads:

$$|||e_h^{pr}(\boldsymbol{\mu})||| \leq c \; \mu^{1/2} \; h^{3/2} |v(\boldsymbol{\mu})|_{H^2(\Omega)} \qquad \forall \mu \in \mathcal{D}, \tag{6.50}$$

for some $c \in \mathbb{R}$, which does not depend on μ.

By using similar arguments to those considered previously for the primal problem, it is possible to provide an a priori FE error estimate also for the dual problem (6.30), which, if $z(\boldsymbol{\mu}) \in H^2(\Omega) \cap \mathcal{V}$, reads:

$$|||e_h^{du}(\boldsymbol{\mu})||| \leq C^{du}(\boldsymbol{\mu}) h^{3/2} |z(\boldsymbol{\mu})|_{H^2(\Omega)} \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.51}$$

where $e_h^{du}(\boldsymbol{\mu}) := z(\boldsymbol{\mu}) - z_h(\boldsymbol{\mu})$ and $C^{du}(\boldsymbol{\mu})$ depends on the data of the problem.

In order to provide the a priori FE error estimate for the output $s(\boldsymbol{\mu})$, let us observe that, by using duality principles (see e.g. [22, 134] and Proposition 4.1) applied to Eq.s (6.5) and (6.29), the following result holds:

$$s(\boldsymbol{\mu}) - s_h(\boldsymbol{\mu}) = L(e_h^{pr}(\boldsymbol{\mu}); \boldsymbol{\mu}) = -a_h(e_h^{pr}(\boldsymbol{\mu}), z(\boldsymbol{\mu}); \boldsymbol{\mu}) \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.52}$$

s.t. the error on the output can be rewritten as:

$$s(\boldsymbol{\mu}) - s_h(\boldsymbol{\mu}) = -a_h(e_h^{pr}(\boldsymbol{\mu}), e_h^{du}(\boldsymbol{\mu}); \boldsymbol{\mu}) - a_h(e_h^{pr}(\boldsymbol{\mu}), z_h(\boldsymbol{\mu}); \boldsymbol{\mu}) \qquad \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{6.53}$$

The first term of (6.53) can be bounded as:

$$|a_h(e_h^{pr}(\boldsymbol{\mu}), e_h^{du}(\boldsymbol{\mu}); \boldsymbol{\mu})| \leq \left[\max\{1, ||\gamma(\boldsymbol{\mu})||_\infty\} + (\delta_h(h, \boldsymbol{\mu}))^{-1/2}\right] |||e_h^{pr}(\boldsymbol{\mu})||| \; |||e_h^{du}(\boldsymbol{\mu})|||, \tag{6.54}$$

due to Eq.(6.35); the second term of (6.53) can be rewritten as:

$$\begin{aligned} -a_h(e_h^{pr}(\boldsymbol{\mu}), z_h(\boldsymbol{\mu}); \boldsymbol{\mu}) &= -\varepsilon_h(h, \boldsymbol{\mu}) \int_\Omega \nabla e_h^{pr}(\boldsymbol{\mu}) \cdot \nabla z_h \; d\Omega \\ &\quad -\varepsilon_h(h, \boldsymbol{\mu}) \int_\Omega \nabla v_h(\boldsymbol{\mu}) \cdot \nabla z_h \; d\Omega \\ &= \varepsilon_h(h, \boldsymbol{\mu}) \int_\Omega e_h^{pr}(\boldsymbol{\mu}) \; \Delta z_h \; d\Omega \\ &\quad -\varepsilon_h(h, \boldsymbol{\mu}) \int_\Omega \nabla v_h(\boldsymbol{\mu}) \cdot \nabla z_h \; d\Omega, \end{aligned} \tag{6.55}$$

due to the Galerkin "skew orthogonality" property (6.43), integration by parts (Eq.(6.44)) and $\Gamma_D \equiv \partial\Omega$. By observing that $\Delta z_h(\boldsymbol{\mu}) = 0$, *a.e.* in Ω, due to $z_h(\boldsymbol{\mu}) \in \mathcal{V}_h$, and by defining the following (computable) correction term:

$$\delta s_h(\boldsymbol{\mu}) := -\varepsilon_h(h, \boldsymbol{\mu}) \int_\Omega \nabla v_h(\boldsymbol{\mu}) \cdot \nabla z_h \; d\Omega, \tag{6.56}$$

and the corrected output:

$$\widetilde{s}_h(\boldsymbol{\mu}) := s_h(\boldsymbol{\mu}) - \delta s_h(\boldsymbol{\mu}), \tag{6.57}$$

it follows that the error on the corrected output can be bounded as Eq.(6.54). By recalling the estimates (6.49) and (6.51) and the definitions (6.20), we obtain the following a priori FE error estimate for the corrected output:

$$|s(\boldsymbol{\mu}) - \widetilde{s}_h(\boldsymbol{\mu})| \leq C^s(h, \boldsymbol{\mu}) h^{5/2} |v(\boldsymbol{\mu})|_{H^2(\Omega)} |z(\boldsymbol{\mu})|_{H^2(\Omega)}, \tag{6.58}$$

where the constant $C^s(h, \boldsymbol{\mu}) \in \mathbb{R}$ does not affect the convergence order in h, being:

$$C^s(h, \boldsymbol{\mu}) := C(\boldsymbol{\mu}) C^{du}(\boldsymbol{\mu}) [C_\delta(\boldsymbol{\mu})]^{-1/2} \left\{1 + \max\{1, ||\gamma(\boldsymbol{\mu})||_\infty\} [C_\delta(\boldsymbol{\mu})]^{1/2} h^{1/2}\right\}. \tag{6.59}$$

By recalling the particular case of Problem 1 endowed with the Problem 1(bis) properties, we have from Eq.s (6.50) and (6.59) and for $h \to 0$:

$$|s(\boldsymbol{\mu}) - \widetilde{s}_h(\boldsymbol{\mu})| \leq c_s \mu^{3/2} h^{5/2} |v(\boldsymbol{\mu})|_{H^2(\Omega)} |z(\boldsymbol{\mu})|_{H^2(\Omega)}, \tag{6.60}$$

for some $c_s \in \mathbb{R}$, which does not depend neither on h nor on μ.

The general case

We consider now the general problem introduced in Sec.6.1.1; even if we observe that in this case it is not straightforward to provide a priori FE error estimates for the primal and dual solutions and the output in the same manner as made for the Problem 1(bis), being $\Gamma_D \neq \partial\Omega$. However, we highlight here the existence and uniqueness of the solutions of the primal and dual problems and some estimates that we use in the following Sections for the analysis of the RB problem.
For the general case, we introduce the following norm:

$$|||w|||^2 := \varepsilon_h(h,\boldsymbol{\mu})||\nabla w||^2 + \delta_h(h,\boldsymbol{\mu})||\mathbf{V}(\boldsymbol{\mu})\cdot\nabla w||^2 + ||w||^2 + \frac{1+\delta_h(h,\boldsymbol{\mu})}{2}||(\mathbf{V}(\boldsymbol{\mu})\cdot\hat{\mathbf{n}})^{1/2}w||^2_{\Gamma_N}, \tag{6.61}$$

where $||\cdot||_{\Gamma_N} := ||\cdot||_{L(\Gamma_N)}$; let us observe that the norm (6.32) is a particular case of the norm defined in Eq.(6.61), when $\Gamma_N \equiv \emptyset$. We recall that $|||\cdot|||$ depends on the parameters vector $\boldsymbol{\mu}$.
In analogy with Sec.6.2.2 as made for the Problem 1(bis), it is simple to prove the existence and uniqueness of the solution of the problem (6.18); in fact, the form (6.19) is coercive (see Eq.(6.36)) and continuous (see Eq.(6.35)) even with the new norm. This last property follows from the following inequality:

$$|a_h(w_h,\phi_h;\boldsymbol{\mu})| \leq \left[2\max\{1,||\gamma(\boldsymbol{\mu})||_\infty\}\ |||w_h||| + (\delta_h(h,\boldsymbol{\mu}))^{-1/2}\ ||w_h||\right]\ |||\phi_h|||, \tag{6.62}$$

which is the analogous of (6.34) for the general problem under consideration endowed with the norm (6.61). Eq.(6.62) can finally be rewritten as:

$$|a_h(w_h,\phi_h;\boldsymbol{\mu})| \leq \left[2\max\{1,||\gamma(\boldsymbol{\mu})||_\infty\} + (\delta_h(h,\boldsymbol{\mu}))^{-1/2}\right]\ |||w_h|||\ |||\phi_h|||, \tag{6.63}$$

which shows the continuity of the bilinear form $a_h(\cdot,\cdot;\boldsymbol{\mu})$.
In the same manner, existence and uniqueness of the dual problem solution are ensured. With this aim, it is useful to express the inequality (6.62) as:

$$|a_h(w_h,\phi_h;\boldsymbol{\mu})| \leq |||w_h|||\ \left[2\max\{1,||\gamma(\boldsymbol{\mu})||_\infty\}\ |||\phi_h||| + (\delta_h(h,\boldsymbol{\mu}))^{-1/2}\ ||\phi_h||\right], \tag{6.64}$$

which leads to the same inequality written in Eq.(6.63).

6.3 The Reduced Basis method

In this Section we consider the RB method for the solution of the stabilized parametrized advection–reaction problem described in Sec.6.2 according with the "primal–dual" RB approach. Moreover, we provide both the a priori and a posteriori RB error estimates and an adaptive procedure for the choice of the samples for the basis generation (adaptive algorithm). Finally, we provide a sensitivity analysis for the output functional w.r.t. the parameters.

6.3.1 The RB method for the stabilized problem

We recall here the RB method by referring to the stabilized problem described in Sec.6.2.1; in particular, we consider the "primal–dual" RB approach. For further details about the RB

method and applications to other problems described by PDEs, see [70, 72, 129, 136, 145, 150, 162, 165, 179, 180].
Let us introduce the following set of parameters $\mathcal{S}_N^{pr} := \{\boldsymbol{\mu}_1^{pr}, \ldots, \boldsymbol{\mu}_N^{pr}\}$, with $\boldsymbol{\mu}_i^{pr} \in \mathcal{D}$, $\forall i = 1, \ldots, N^{pr}$, $N^{pr} \in \mathbb{N}$; we observe that the superscript "pr" is used to refer to the primal problem. The set of parameters $\boldsymbol{\mu}_i^{pr}$ and their number N^{pr} can be chosen in the space $\mathcal{D}$ according with the adaptive algorithm that we discuss in Sec.6.3.4; however, other choices are possible. E.g., the most simple choice corresponds to an equidistributed selection of samples in $\mathcal{D}$; in the case of symmetric elliptic problems, it is also possible to select the parameters according with a priori estimates (see [113, 114]). For each of the samples $\boldsymbol{\mu}_i^{pr} \in \mathcal{S}_N^{pr}$, we solve, by means of the FE element method, the stabilized primal problem (6.18), obtaining the corresponding approximated primal solutions $v_h(\boldsymbol{\mu}_i^{pr}) \in \mathcal{V}_h$. Then, we define the RB space associated with the primal problem:

$$\mathcal{V}_N^{pr} := \operatorname{span}\{\xi_i := v_h(\boldsymbol{\mu}_i^{pr}) \quad i = 1, \ldots, N^{pr}\} \tag{6.65}$$

and the RB primal problem:

$$\text{find } v_N(\boldsymbol{\mu}) \in \mathcal{V}_N^{pr} \;:\; a_h(v_N(\boldsymbol{\mu}), \phi_N; \boldsymbol{\mu}) = F_h(\phi_N; \boldsymbol{\mu}) \qquad \forall \phi_N \in \mathcal{V}_N^{pr}, \quad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.66}$$

where the bilinear form $a_h(\cdot, \cdot; \boldsymbol{\mu})$ and the functional $F_h(\cdot; \boldsymbol{\mu})$ are defined in Eq.(6.19); we suppose that the solution $\phi_N(\boldsymbol{\mu})$ of the RB primal problem could be expressed as:

$$v_N(\boldsymbol{\mu}) = \sum_{j=1}^{N^{pr}} v_{Nj}(\boldsymbol{\mu}) \xi_j. \tag{6.67}$$

Let us observe that, for the sake of simplicity, we have used the notation $v_N(\boldsymbol{\mu})$ to indicate the RB solution $v_{hN}(\boldsymbol{\mu})$ of the primal problem (6.66).

Remark 6.4. *From a numerical point of view, the approximated problem associated with (6.66) becomes more and more ill–conditioned with N^{pr} increasing in size, so affecting the accuracy of the RB solution. To overcome this difficulty, it is convenient to adopt an orthonormal basis for the generation of the RB space $\mathcal{V}_N^{pr}$; in particular we consider the Gram–Schmidt orthormalization w.r.t. the inner product $(\cdot, \cdot)$ induced by the norm $||\cdot||$ (see e.g. [136]). Let us indicate the new orthonormal basis for the RB space $\mathcal{V}_N^{pr}$ as $\{\varrho_i\}_{i=1}^{N^{pr}}$, which we compute according with the following procedure:*

$$\begin{gathered} \varrho_1 = \xi_1 / ||\xi_1||, \\ \psi_i = \xi_i - \sum_{j=1}^{i-1} (\varrho_j, \xi_i) \varrho_j, \qquad \varrho_i = \psi_i / ||\psi_i|| \qquad i = 2, \ldots, N^{pr}. \end{gathered} \tag{6.68}$$

We obtain that $\mathcal{V}_N^{pr} = \operatorname{span}\{\xi_i,\ i = 1, \ldots, N^{pr}\} = \operatorname{span}\{\varrho_i,\ i = 1, \ldots, N^{pr}\}$. Let us notice that we will consider the orthonormal basis for the definition of our RB problem; however, for the sake of simplicity, we identify the basis $\{\xi_i\}_{i=1}^{N^{pr}}$ used in Eq.s (6.65) and (6.66) as the orthonormal basis $\{\varrho_i\}_{i=1}^{N^{pr}}$.

Owing to the assumption made in Sec.6.2.1 concerning the affine decomposition of the bilinear form $a_h(\cdot, \cdot; \boldsymbol{\mu})$ and the functional $F_h(\cdot; \boldsymbol{\mu})$ (see Eq.(6.23)) and Eq.(6.67), Eq.(6.66) reads:

$$\text{find } v_{Nj}(\boldsymbol{\mu}) \quad j = 1, \dots, N^{pr} \ :$$

$$\sum_{q=1}^{Q_h} \vartheta_{hq}(\boldsymbol{\mu}) a_{hq}(\xi_j, \xi_i) v_{Nj}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_h^F} \vartheta_{hq}^F(\boldsymbol{\mu}) F_{hq}(\xi_i) \tag{6.69}$$
$$\forall i = 1, \dots, N^{pr}, \quad \forall \boldsymbol{\mu} \in \mathcal{D}.$$

which can be rewritten in matricial notation as:

$$\text{find } \boldsymbol{v}_N(\boldsymbol{\mu}) \in \mathbb{R}^{N^{pr}} \ : \ A_N^{pr}(\boldsymbol{\mu}) \boldsymbol{v}_N(\boldsymbol{\mu}) = \mathbf{F}_N^{pr}(\boldsymbol{\mu}), \tag{6.70}$$

being the matrix $A_N^{pr}(\boldsymbol{\mu}) \in \mathbb{R}^{N^{pr} \times N^{pr}}$ and the vector $\mathbf{F}_N^{pr}(\boldsymbol{\mu}) \in \mathbb{R}^{N^{pr}}$ defined as:

$$\begin{aligned} A_N^{pr}(\boldsymbol{\mu}) &:= \sum_{q=1}^{Q_h} \vartheta_{hq}(\boldsymbol{\mu}) A_{Nq}^{pr} \qquad \text{with } \left(A_{Nq}^{pr}\right)_{i,j} := a_{hq}(\xi_j, \xi_i), \\ \mathbf{F}_N^{pr}(\boldsymbol{\mu}) &:= \sum_{q=1}^{Q_h^F} \vartheta_{hq}^F(\boldsymbol{\mu}) \mathbf{F}_{Nq} \qquad \text{with } \left(\mathbf{F}_{Nq}^{pr}\right)_{i} := F_{hq}(\xi_i). \end{aligned} \tag{6.71}$$

By recalling Eq.s (6.5), (6.8) and (6.67), it follows that:

$$s_N(\boldsymbol{\mu}) := L(v_N(\boldsymbol{\mu}); \boldsymbol{\mu}) = \boldsymbol{v}_N(\boldsymbol{\mu}) \cdot \mathbf{L}_N^{pr}(\boldsymbol{\mu}), \tag{6.72}$$

where the vector $\mathbf{L}_N^{pr}(\boldsymbol{\mu}) \in \mathbb{R}^{N^{pr}}$ is defined as:

$$\mathbf{L}_N^{pr}(\boldsymbol{\mu}) := \sum_{q=1}^{Q^L} \vartheta_q^L(\boldsymbol{\mu}) \mathbf{L}_{Nq}^{pr} \qquad \text{with } \left(\mathbf{L}_{Nq}^{pr}\right)_{i} := L_q(\xi_i). \tag{6.73}$$

In analogous manner it is possible to afford the dual RB problem, which is required for the “primal–dual” RB formulation.
Let us select a set of parameters $\mathcal{S}^{du} := \{\boldsymbol{\mu}_1^{du}, \dots, \boldsymbol{\mu}_N^{du}\}$, with $\boldsymbol{\mu}_i^{du} \in \mathcal{D}$, $\forall i = 1, \dots, N^{du}$, $N^{du} \in \mathbb{N}$; in this case the superscript ’du’ refers to the dual problem.

Remark 6.5. *Let us observe that we consider the non–integrated primal and dual RB approach [136], for which, not only $N^{pr} \neq N^{du}$, but also the sets $\mathcal{S}^{pr}$ and $\mathcal{S}^{du}$ are composed by different parameters; this issue will be recalled and discussed in Sec.6.3.4.*

By defining the RB space for the dual problem as:

$$\mathcal{V}_N^{du} := \operatorname{span}\left\{\zeta_i := z_h(\boldsymbol{\mu}_i^{du}) \ \ i = 1, \dots, N^{du}\right\}, \tag{6.74}$$

where $z_h(\boldsymbol{\mu}_i^{du})$ is the solution of the FE dual problem (6.30) corresponding to the parameter $\boldsymbol{\mu}_i^{du}$ and $\{\zeta_i\}_{i=1}^{N^{du}}$ is an orthonormal basis (see Remark 6.4). The dual RB problem reads:

$$\text{find } z_N(\boldsymbol{\mu}) \in \mathcal{V}_N^{du} \ : \ a_h(\phi_N, z_N(\boldsymbol{\mu}); \boldsymbol{\mu}) = -L(\phi_N; \boldsymbol{\mu}) \qquad \forall \phi_N \in \mathcal{V}_N^{du}, \quad \forall \boldsymbol{\mu} \in \mathcal{D}; \tag{6.75}$$

the dual RB solution can be written as:

$$z_N(\boldsymbol{\mu}) = \sum_{j=1}^{N^{du}} z_{Nj}(\boldsymbol{\mu}) \zeta_j. \tag{6.76}$$

According with Eq.s (6.7), (6.23), (6.30), Eq.(6.75) can be expressed in matricial notation as:

$$\text{find } \boldsymbol{z}_N(\boldsymbol{\mu}) \in \mathbb{R}^{N^{du}} \;:\; A_N^{du}(\boldsymbol{\mu})\boldsymbol{z}_N(\boldsymbol{\mu}) = -\mathbf{L}_N^{du}(\boldsymbol{\mu}), \tag{6.77}$$

being the matrix $A_N^{du}(\boldsymbol{\mu}) \in \mathbb{R}^{N^{du} \times N^{du}}$ and the vector $\mathbf{L}_N^{du}(\boldsymbol{\mu}) \in \mathbb{R}^{N^{du}}$ defined as:

$$\begin{aligned} A_N^{du}(\boldsymbol{\mu}) &:= \sum_{q=1}^{Q_h} \vartheta_{hq}(\boldsymbol{\mu}) A_{Nq}^{du} \qquad \text{with } \left(A_{Nq}^{du}\right)_{i,j} := a_{hq}(\zeta_i, \zeta_j), \\ \mathbf{L}_N^{du}(\boldsymbol{\mu}) &:= \sum_{q=1}^{Q^L} \vartheta_q^L(\boldsymbol{\mu}) \mathbf{L}_{Nq}^{du} \qquad \text{with } \left(\mathbf{L}_{Nq}^{du}\right)_i := L_q(\zeta_i). \end{aligned} \tag{6.78}$$

6.3.2 A priori RB error estimate

In this Section we discuss the a priori RB error estimates both for the primal and dual solutions and the output. With this aim we make use of some of the estimates provided in Sec.6.2.2 for the FE problem. For further details about the a priori approximation theory refer to [113, 114, 136] where the case of symmetric elliptic problems with single parameteric dependence is discussed; for multiple parametric dependence, work is still in progress [31].

First of all, let us observe that the solution of the primal RB problem $v_N(\boldsymbol{\mu})$ belongs not only to $\mathcal{V}_N^{pr}$, but also to the FE space $\mathcal{V}_h$ ($v_N(\boldsymbol{\mu}) \in \mathcal{V}_N^{pr} \subset \mathcal{V}_h$); in the same manner the dual RB solution $z_N(\boldsymbol{\mu}) \in \mathcal{V}_N^{du} \subset \mathcal{V}_h$.

Let us recall the coercivity property (6.36) of the bilinear form $a_h(\cdot, \cdot; \boldsymbol{\mu})$ with respect to the norm $|||\cdot|||$ defined in Eq.(6.61) (or in Eq.(6.32) for the Problem 1(bis)), from which it follows:

$$|||e_N^{pr}(\boldsymbol{\mu})|||^2 \leq \frac{1}{\min\{1, ||\gamma(\boldsymbol{\mu})||_\infty\}} a_h(e_N^{pr}(\boldsymbol{\mu}), e_N^{pr}(\boldsymbol{\mu}); \boldsymbol{\mu}) \tag{6.79}$$

where $e_N^{pr}(\boldsymbol{\mu}) := v_h(\boldsymbol{\mu}) - v_N(\boldsymbol{\mu})$ is the primal RB error, s.t. $e_N^{pr}(\boldsymbol{\mu} \in \mathcal{V}_h$. By observing from Eq.s (6.18) and (6.66) that, for the primal RB problem, the Galerkin orthogonality property holds:

$$a_h(e_N^{pr}(\boldsymbol{\mu}), \phi_N; \boldsymbol{\mu}) = 0 \qquad \forall \phi_N \in \mathcal{V}_N^{pr}, \tag{6.80}$$

the bilinear form $a_h(e_N^{pr}(\boldsymbol{\mu}), e_N^{pr}(\boldsymbol{\mu}); \boldsymbol{\mu})$ can be written as:

$$a_h(e_N^{pr}(\boldsymbol{\mu}), e_N^{pr}(\boldsymbol{\mu}); \boldsymbol{\mu}) = a_h(e_N^{pr}(\boldsymbol{\mu}), v_h(\boldsymbol{\mu}) - \phi_N; \boldsymbol{\mu}) \qquad \forall \phi_N \in \mathcal{V}_N^{pr}, \tag{6.81}$$

being $v_h(\boldsymbol{\mu}) - \phi_N \in \mathcal{V}_N^{pr}$. By applying the inequality (6.63) to Eq.(6.81), recalling Eq.(6.79) and simplifying the term $|||e_N^{pr}(\boldsymbol{\mu})|||$ both on the l.h.s and r.h.s of the inequality, we have:

$$|||e_N^{pr}(\boldsymbol{\mu})||| \leq \frac{2\max\{1, ||\gamma(\boldsymbol{\mu})||_\infty\} + (\delta_h(h, \boldsymbol{\mu}))^{-1/2}}{\min\{1, ||\gamma(\boldsymbol{\mu})||_\infty\}} |||v_h(\boldsymbol{\mu}) - \phi_N||| \qquad \forall \phi_N \in \mathcal{V}_N^{pr}. \tag{6.82}$$

From Eq.s (6.20) and (6.82) it follows the a priori primal RB error estimate:

$$|||e_N^{pr}(\boldsymbol{\mu})||| \leq \Xi(\boldsymbol{\mu}) \inf_{\phi_N \in \mathcal{V}_N^{pr}} |||v_h(\boldsymbol{\mu}) - \phi_N||| \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.83}$$

with:

$$\Xi(\boldsymbol{\mu}) := \frac{2\max\{1, ||\gamma(\boldsymbol{\mu})||_\infty\} + (C_\delta(\boldsymbol{\mu})h)^{-1/2}}{\min\{1, ||\gamma(\boldsymbol{\mu})||_\infty\}}. \tag{6.84}$$

Let us observe that, in the special case of Problem 1, the estimate (6.83) holds with:

$$\Xi(\mu) = 2 + \left(\frac{\mu}{c_\delta h}\right)^{1/2}. \tag{6.85}$$

In the same manner, the following a priori RB error estimate for the dual problem holds:

$$|||e_N^{du}(\boldsymbol{\mu})||| \leq \Xi(\boldsymbol{\mu}) \inf_{\phi_N \in \mathcal{V}_N^{du}} |||z_h(\boldsymbol{\mu}) - \phi_N||| \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.86}$$

where $e_N^{du}(\boldsymbol{\mu}) := z_h(\boldsymbol{\mu}) - z_N(\boldsymbol{\mu})$.

By means of Eq.s (6.5), (6.22), (6.30) and (6.72) and the Galerkin orthogonality property (6.80), we have:

$$s_h(\boldsymbol{\mu}) - s_N(\boldsymbol{\mu}) = L(e_N^{pr}(\boldsymbol{\mu}); \boldsymbol{\mu}) = -a_h(e_N^{pr}(\boldsymbol{\mu}), e_N^{du}(\boldsymbol{\mu}); \boldsymbol{\mu}), \tag{6.87}$$

which, according with the inequality (6.63), leads to the following a priori RB error estimate for the output:

$$|s_h(\boldsymbol{\mu}) - s_N(\boldsymbol{\mu})| \leq \Xi(\boldsymbol{\mu})^3 \left(\inf_{\phi_N \in \mathcal{V}_N^{pr}} |||v_h(\boldsymbol{\mu}) - \phi_N|||\right)\left(\inf_{w_N \in \mathcal{V}_N^{du}} |||z_h(\boldsymbol{\mu}) - w_N|||\right). \tag{6.88}$$

6.3.3 A posteriori RB error estimate

In this Section we discuss the a posteriori RB error estimate for the output; for further details refer to [129, 136, 168]. We will use this result in Sec.6.3.4 for the choice of the samples sets $\mathcal{S}_N^{pr}$ and $\mathcal{S}_N^{du}$.

Let us define the RB residuals associated with the primal and dual problems by recalling respectively Eq.s (6.18) and (6.30):

$$R_N^{pr}(\phi_h; \boldsymbol{\mu}) := F_h(\phi_h; \boldsymbol{\mu}) - a_h(v_N(\boldsymbol{\mu}), \phi_h; \boldsymbol{\mu}), \tag{6.89}$$

$$R_N^{du}(\phi_h; \boldsymbol{\mu}) := -L(\phi_h; \boldsymbol{\mu}) - a_h(\phi_h, z_N(\boldsymbol{\mu}); \boldsymbol{\mu}), \tag{6.90}$$

with $\phi_h \in \mathcal{V}_h$, which, equivalently, can be rewritten respectively as:

$$R_N^{pr}(\phi_h; \boldsymbol{\mu}) = a_h(e_N^{pr}(\boldsymbol{\mu}), \phi_h; \boldsymbol{\mu}), \tag{6.91}$$

$$R_N^{du}(\phi_h; \boldsymbol{\mu}) = a_h(\phi_h, e_N^{du}(\boldsymbol{\mu}); \boldsymbol{\mu}). \tag{6.92}$$

Remark 6.6. *We observe that, by using the notation of Sec.4.1.2, we ideally have $R_N^{pr}(\phi_h; \boldsymbol{\mu}) = R^{pr}(v_N)(\phi_h)$ and $R_N^{du}(\phi_h; \boldsymbol{\mu}) = R^{du}(z_N)(\phi_h)$, being $R^{pr}(\cdot)(\cdot)$ and $R^{du}(\cdot)(\cdot)$ defined in Eq.s (4.7) and (4.8) for the parameter vector $\boldsymbol{\mu}$, respectively.*

Moreover, we define the *corrected RB output* (deflated) [168] (see also Remark 4.4):

$$\widetilde{s}_N(\boldsymbol{\mu}) := s_N(\boldsymbol{\mu}) - R_N^{pr}(z_N(\boldsymbol{\mu}); \boldsymbol{\mu}) \tag{6.93}$$

and we express the error w.r.t. the corrected RB output as:

$$\begin{aligned} s_h(\boldsymbol{\mu}) - \widetilde{s}_N(\boldsymbol{\mu}) &= L(e_N^{pr}(\boldsymbol{\mu}); \boldsymbol{\mu}) + R_N^{pr}(z_N(\boldsymbol{\mu}); \boldsymbol{\mu}) \\ &= -a_h(e_N^{pr}(\boldsymbol{\mu}), z_h(\boldsymbol{\mu})) + R_N^{pr}(z_N(\boldsymbol{\mu}); \boldsymbol{\mu}) \\ &= -R_N^{pr}(e_N^{du}(\boldsymbol{\mu}); \boldsymbol{\mu}), \end{aligned} \tag{6.94}$$

similarly to the result (4.37) of Proposition 4.1. By introducing the dual norm of $|||\cdot|||$ (see e.g. [109]), which we indicate as $|||\cdot|||_*$, the following inequality holds:

$$|s_h(\boldsymbol{\mu}) - \widetilde{s}_N(\boldsymbol{\mu})| \leq \varepsilon_N^{pr}(\boldsymbol{\mu})\ |||e_N^{du}(\boldsymbol{\mu})|||, \tag{6.95}$$

where:

$$\varepsilon_N^{pr}(\boldsymbol{\mu}) := \sup_{\phi_h \in \mathcal{V}_h \setminus \{0\}} \frac{R_N^{pr}(\phi_h; \boldsymbol{\mu})}{|||\phi_h|||} \qquad \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{6.96}$$

For the computation of the dual norms of the residual refer to [136, 168].
By introducing the parametrized Babŭska *inf–sup* stability constant (see [11]) $\beta(\boldsymbol{\mu})$:

$$\beta(\boldsymbol{\mu}) := \inf_{w_h \in \mathcal{V}_h \setminus \{0\}} \sup_{\phi_h \in \mathcal{V}_h \setminus \{0\}} \frac{a_h(w_h, \phi_h; \boldsymbol{\mu})}{|||w_h|||\ |||\phi_h|||} \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.97}$$

and by observing from Eq.(6.92) that:

$$\beta(\boldsymbol{\mu})|||e_N^{du}(\boldsymbol{\mu})|||^2 \leq R_N^{du}(e_N^{du}(\boldsymbol{\mu}); \boldsymbol{\mu}) \leq \varepsilon_N^{du}(\boldsymbol{\mu})\ |||e_N^{du}(\boldsymbol{\mu})||| \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.98}$$

with:

$$\varepsilon_N^{du}(\boldsymbol{\mu}) := \sup_{\phi_h \in \mathcal{V}_h \setminus \{0\}} \frac{R_N^{du}(\phi_h; \boldsymbol{\mu})}{|||\phi_h|||} \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{6.99}$$

it follows:

$$|||e_N^{du}(\boldsymbol{\mu})||| \leq \frac{1}{\beta(\boldsymbol{\mu})} \varepsilon_N^{du}(\boldsymbol{\mu}) \qquad \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{6.100}$$

Due to Eq.s (6.95) and (6.100), the following a posteriori RB error estimate for the corrected output holds:

$$|s_h(\boldsymbol{\mu}) - \widetilde{s}_N(\boldsymbol{\mu})| \leq \Delta_N(\boldsymbol{\mu}) \qquad \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{6.101}$$

where $\Delta_N(\boldsymbol{\mu}) := \Delta_N^{pr}(\boldsymbol{\mu})\ \Delta_N^{du}(\boldsymbol{\mu})$, with:

$$\begin{aligned} \Delta_N^{pr}(\boldsymbol{\mu}) &:= \frac{1}{\sqrt{\beta(\boldsymbol{\mu})}} \varepsilon_N^{pr}(\boldsymbol{\mu}), \\ \Delta_N^{du}(\boldsymbol{\mu}) &:= \frac{1}{\sqrt{\beta(\boldsymbol{\mu})}} \varepsilon_N^{du}(\boldsymbol{\mu}). \end{aligned} \tag{6.102}$$

Remark 6.7. *Being the bilinear form $a_h(\cdot, \cdot; \boldsymbol{\mu})$ (6.19) coercive (Eq.(6.36)), the parametrized inf–sup constant (6.97) is positive $\beta(\boldsymbol{\mu}) > 0$ $\forall \boldsymbol{\mu} \in \mathcal{D}$, for which the a posteriori estimate (6.101) holds. However, some particular RB problems could show a singular behavior for some choices of the paramateres, i.e. $\beta(\bar{\boldsymbol{\mu}}) \leq 0$ (or simply $\beta(\bar{\boldsymbol{\mu}}) \simeq 0$), for some $\bar{\boldsymbol{\mu}} \in \bar{\mathcal{D}} \subset \mathcal{D}$; this is the case, e.g. of the Helmholtz acoustic problem in resonance conditions (see [168]). In these cases the a posteriori RB error estimate (6.101) is not any longer effective; however it is still possible to adopt the estimate (6.101) by replacing $\beta(\boldsymbol{\mu})$ by a positive definite lower bound. For further details refer to [90, 168, 181].*

Remark 6.8. *In order to make use of the a posteriori RB error estimate (6.101), we need the expression of $\beta(\boldsymbol{\mu})$ $\forall \boldsymbol{\mu} \in \mathcal{D}$; however, this is possible only in some particular cases. A possible solution consists in replacing the inf–sup constant $\beta(\boldsymbol{\mu})$ by a lower–bound $\beta_{LB}(\boldsymbol{\mu}) \leq \beta(\boldsymbol{\mu})$ $\forall \boldsymbol{\mu} \in \mathcal{D}$, which allows sufficiently and accurate approximations of $\beta(\boldsymbol{\mu})$ (see [168]). An other* *possibility arises by observing that, for a coercive problem, $\beta(\boldsymbol{\mu}) \geq \alpha(\boldsymbol{\mu})$ $\forall \boldsymbol{\mu} \in \mathcal{D}$, being $\alpha(\boldsymbol{\mu})$ the coercivity constant (6.37), for which we could assume $\beta_{LB}(\boldsymbol{\mu}) = \alpha(\boldsymbol{\mu}) = \min\{1, ||\gamma(\boldsymbol{\mu})||_\infty\}$. However, for problems with dominating transport regimes, the choice $\beta_{LB}(\boldsymbol{\mu}) = \alpha(\boldsymbol{\mu})$ could lead to excessively pessimistic approximations of $\beta(\boldsymbol{\mu})$, which affect the sharpness of the a posteriori RB estimate (6.101).*

As anticipated, the a posteriori RB estimate (6.101) should be reliable, in the sense that the error is bounded by the estimator, but also sharp, i.e. the error and the bound should be close each other, in order to avoid too pessimistic RB error estimates. To evaluate the sharpness property of the estimator $\Delta_N(\boldsymbol{\mu})$, let us introduce the following indicator (see [136]), regarded as the *effectivity index*:

$$\eta_N := \frac{\sup\limits_{\boldsymbol{\mu}\in\mathcal{D}} \Delta_N(\boldsymbol{\mu})}{\sup\limits_{\boldsymbol{\mu}\in\mathcal{D}} |s_h(\boldsymbol{\mu}) - \widetilde{s}_N(\boldsymbol{\mu})|}, \tag{6.103}$$

which, according with the reliability property, should be greater than (or equal to) one ($\eta_N \geq 1$), but as close as possible to 1 due to the sharpness property.

6.3.4 Numerical assessment: offline–online computational procedure and adaptive algorithm

In this Section we discuss the details concerning the numerical solution of the RB problem described in the previous Sections. In particular, we highlight the importance of the offline–online decomposition of the computational procedure for the RB problem and we provide an adaptive algorithm for the choice of the primal and dual samples sets by using a minimization criterium of online computational costs. For further details about the solution of RB problems by means of the offline–online decomposition strategy and other possible adaptive algorithms, refer to [129, 136, 145].

Numerical issues

Before to describe the offline–online computational procedure and the adaptive algorithm, let us introduce some issues related to the numerical solution of the RB problem.
The norm $|||\cdot|||$ introduced in Eq.s (6.32) or (6.61) depends on the parameter $\boldsymbol{\mu} \in \mathbf{D}$; however, for computational reasons, it is convenient to fix a sample $\overline{\boldsymbol{\mu}} \in \mathcal{D}$ for which the norm is evaluated, s.t. the norms (6.32) and (6.61) become respectively:

$$|||w|||^2_{\overline{\boldsymbol{\mu}}} := \varepsilon_h(h,\overline{\boldsymbol{\mu}})||\nabla w||^2 + \delta_h(h,\overline{\boldsymbol{\mu}})||\mathbf{V}(\overline{\boldsymbol{\mu}})\cdot\nabla w||^2 + ||w||^2,$$

$$|||w|||^2_{\overline{\boldsymbol{\mu}}} := \varepsilon_h(h,\overline{\boldsymbol{\mu}})||\nabla w||^2 + \delta_h(h,\overline{\boldsymbol{\mu}})||\mathbf{V}(\overline{\boldsymbol{\mu}})\cdot\nabla w||^2 + ||w||^2 + \frac{1+\delta_h(h,\overline{\boldsymbol{\mu}})}{2}||(\mathbf{V}(\overline{\boldsymbol{\mu}})\cdot\hat{\mathbf{n}})^{1/2}w||^2_{\Gamma_N}. \tag{6.104}$$

This does not affect the a priori FE and RB error estimates (see Sec.s 6.2.2, 6.2.2 and 6.3.2) and the a posteriori RB estimate (see Sec.6.3.3): simply, the quantities related to the parametrical "fixed" norm should be recomputed according with $|||\cdot|||_{\overline{\boldsymbol{\mu}}}$. For example, the inf–sup constant $\beta(\boldsymbol{\mu})$ (6.97), defined w.r.t. the "fixed" norm reads:

$$\overline{\beta}(\boldsymbol{\mu}) := \inf_{w_h\in\mathcal{V}_h\backslash\{0\}} \sup_{\phi_h\in\mathcal{V}_h\backslash\{0\}} \frac{a_h(w_h,\phi_h;\boldsymbol{\mu})}{|||w_h|||_{\overline{\boldsymbol{\mu}}}\ |||\phi_h|||_{\overline{\boldsymbol{\mu}}}} \qquad \forall\boldsymbol{\mu}\in\mathcal{D}. \tag{6.105}$$

As anticipated in Remark 6.8, we need the expression of $\overline{\beta}(\boldsymbol{\mu})$ or of a lower bound. However, for the evaluation of $\overline{\beta}(\boldsymbol{\mu})$ we use a composite linear polinomial interpolant ([152]) defined on the parameter space $\mathcal{D}$:

$$\widetilde{\beta}(\boldsymbol{\mu}) := \sum_{m=1}^{K} \overline{\beta}(\boldsymbol{\mu}_m)\Pi_m(\boldsymbol{\mu}), \tag{6.106}$$

being $\{\Pi_m(\boldsymbol{\mu})\}_{m=1}^{K}$ the basis functions associated with the linear interpolant and $\{\boldsymbol{\mu}_m\}_{m=1}^{K}$ the samples, chosen in $\mathcal{D}$, being K the number of samples. The choice of the samples $\{\boldsymbol{\mu}_m\}_{m=1}^{K}$ and K is made by means of inspection of the function $\overline{\beta}(\boldsymbol{\mu})$ for $\boldsymbol{\mu} \in \mathcal{D}$.
From this it follows that the a posteriori RB estimate (6.101) reads now:

$$|s_h(\boldsymbol{\mu}) - \widetilde{s}_N(\boldsymbol{\mu})| \leq \overline{\Delta}_N(\boldsymbol{\mu}) \qquad \forall \boldsymbol{\mu} \in \mathcal{D}. \tag{6.107}$$

where $\overline{\Delta}_N(\boldsymbol{\mu}) := \overline{\Delta}_N^{pr}(\boldsymbol{\mu})\ \overline{\Delta}_N^{du}(\boldsymbol{\mu})$, with:

$$\begin{aligned} \overline{\Delta}_N^{pr}(\boldsymbol{\mu}) &:= \frac{1}{\sqrt{\widetilde{\beta}(\boldsymbol{\mu})}} \varepsilon_{N\overline{\boldsymbol{\mu}}}^{pr}(\boldsymbol{\mu}), \\ \overline{\Delta}_N^{du}(\boldsymbol{\mu}) &:= \frac{1}{\sqrt{\widetilde{\beta}(\boldsymbol{\mu})}} \varepsilon_{N\overline{\boldsymbol{\mu}}}^{du}(\boldsymbol{\mu}), \end{aligned} \tag{6.108}$$

being $\varepsilon_{N\overline{\boldsymbol{\mu}}}^{pr}(\boldsymbol{\mu})$ and $\varepsilon_{N\overline{\boldsymbol{\mu}}}^{du}(\boldsymbol{\mu})$ the dual norms associated with $|||\cdot|||_{\overline{\boldsymbol{\mu}}}$.
The RB problems described in the previous Sections refers to the parameter $\boldsymbol{\mu}$ chosen in the space $\mathcal{D}$. However, for computational reasons, the research of the RB samples sets $\mathcal{S}_N^{pr}$, $\mathcal{S}_N^{du}$ must be restricted to a approximated subset $\overline{\mathcal{D}} \subset \mathcal{D}$ with a number of samples $\overline{N}$ "sufficiently great". In order to evaluate the RB error, it is convenient to define the following indicators for the maximum and minimum output errors for any given N^{pr} and N^{du}:

$$\begin{aligned} E_N^{max} &:= \max_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} |s_h(\boldsymbol{\mu}) - \widetilde{s}_N(\boldsymbol{\mu})|, \\ E_N^{mean} &:= \left(\sum_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} |s_h(\boldsymbol{\mu}) - \widetilde{s}_N(\boldsymbol{\mu})| \right) / \overline{N}, \end{aligned} \tag{6.109}$$

In similar way, the effectivity index η_N (6.103) should be modified in:

$$\overline{\eta}_N := \frac{\max\limits_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} \Delta_N(\boldsymbol{\mu})}{\max\limits_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} |s_h(\boldsymbol{\mu}) - \widetilde{s}_N(\boldsymbol{\mu})|}. \tag{6.110}$$

Offline–online computational procedure

The rapid answer of the RB method in a many query input–output context is highlighted if the RB problem is solved by means of an appropriate the *offline–online decomposition* procedure.

In the *offline* step we define the RB spaces once time the primal and dual samples sets $\mathcal{S}_N^{pr}$, $\mathcal{S}_N^{du}$ are given. This choice is performed according with the adaptive algorithm, which we describe in Sec.6.3.4. Then we build the RB orthonormal basis $\{\xi_i\}_{i=1}^{N^{pr}}$, $\{\zeta_i\}_{i=1}^{N^{du}}$ by solving the primal and dual FE problems and finally we assemble the parameter independent RB matrices A^{pr}_{Nq}, A^{du}_{Nq} and vectors $\mathbf{F}^{pr}_{Nq}$, $\mathbf{L}^{pr}_{Nq}$, $\mathbf{L}^{du}_{Nq}$. Finally, we prepare the matrices and vectors for the evaluation of the a posteriori RB error estimate. These operations have, in general, relevant computational costs being dependent on the dimension of the FE problems N_h, which is greater than that of the primal and dual RB problems, i.e. $N^{pr} << N_h$ and $N^{du} << N_h$.

In the *online* step, for any given $\boldsymbol{\mu} \in \mathcal{D}$, we assemble the parameter dependent RB matrices $A^{pr}_N(\boldsymbol{\mu})$, $A^{du}_N(\boldsymbol{\mu})$ and vectors $\mathbf{F}^{pr}_N(\boldsymbol{\mu})$, $\mathbf{L}^{pr}_N(\boldsymbol{\mu})$, $\mathbf{L}^{du}_N(\boldsymbol{\mu})$, we solve the primal and dual RB problems (6.70) and (6.77), we compute the corrected output $\widetilde{s}_N(\boldsymbol{\mu})$ and, if it is requested, we provide the a posteriori RB error estimate (6.101) associated with the computed output. The online computational costs can be accounted in the following manner (for more details see e.g. [129, 136, 145]):

- assembling of the RB matrices and vectors: $O\left(Q_h\left(N^{pr\,2}+N^{du\,2}\right)\right)$;
- solving the primal and dual RB linear systems[1]: $O\left(N^{pr\,3}+N^{du\,3}\right)$;
- computing the corrected output: $O\left(Q_h N^{pr} N^{du}\right)$;
- evaluating the a posteriori RB estimate (if requested): $O\left(Q_h^2\left(N^{pr\,2}+N^{du\,2}\right)\right)$;

let us observe that we have indicated only the orders of the dominating costs for each online step. We refer to the total online computational costs by means of the following indicator:

$$\Lambda_N = \Lambda_N(N^{pr}, N^{du}; Q_h) := N^{pr\,3} + N^{du\,3} + Q_h\left(N^{pr\,2} + N^{du\,2}\right) + Q_h N^{pr} N^{du}, \qquad (6.111)$$

which depends on N^{pr}, N^{du} and Q_h; let us observe that Q_h depends only on the problem under consideration; this implies that, given a parametrized problem, the online computational cost depends only on the dimension of the RB problem. We chose to do not take into account in the indicator Λ_N for the costs associated with the evaluation of the a posteriori RB error estimate; this choice is due to the adaptive algorithm which we adopt and which we describe in the following Section.

Remark 6.9. *The previous offline–online decomposition is allowed by the affine decomposition hypothesis made in Sec.s 6.1.1 and 6.2.1 concerning the bilinear form $a_h(\cdot,\cdot;\boldsymbol{\mu})$ and the functionals $F_h(\cdot;\boldsymbol{\mu})$ and $L(\cdot;\boldsymbol{\mu})$. However, in general, for other parametric problems, these hypotheses could not be guaranteed; this is the case, e.g., of geometrical parametrizations which do not admit affine mappings to a reference domain ([150, 162, 164, 165]) or physical parametrizations with coefficients depending both on the parameters and the spatial coordinates in a not separable manner ([17]). In this cases the offline–online computational procedure can not be used in the manner described till now, thus leading to loose the property of rapid and reliable evaluations of input–output relationships. To overcome this difficulty, an empirical, stable and computationally cheap interpolation method has been proposed in [17] to decouple the dependence on the parameters and the spatial coordinates, thus restoring the affine decomposition properties the offline–online computational procedure.*

[1] The RB matrices are in general full [136, 145]; for this reason, the linear RB systems (6.70) and (6.77) can be conveniently solved by means of direct methods (see e.g. [152]) which shows a computational cost of order $O(\mathcal{N}^3)$, being $\mathcal{N}$ the dimension of the system.

Adaptive algorithm

We propose now an adaptive algorithm for the choice of the samples associated with the primal and dual samples sets $\mathcal{S}_N^{pr}$, $\mathcal{S}_N^{du}$, which we use in the offline context. The proposed strategy is based on the minimization of the computational costs (6.111) associated with the online step and the samples are set according with an adaptive procedure; with this aim we use the a posteriori RB error estimate (6.107).

The adaptive algorithm can be summarized as it follows:

1. choose a tolerance *tol* on the absolute error associated with the RB corrected output $\widetilde{s}_N(\boldsymbol{\mu})$ (6.93);

2. choose randomly in $\overline{\mathcal{D}}$ a sample, $\boldsymbol{\mu}_1^{pr}$, for the primal problem and another one, $\boldsymbol{\mu}_1^{du}$, for the dual one; set $N^{pr} = 1$, $N^{du} = 1$, initialize the sets $\mathcal{S}_N^{pr}$, $\mathcal{S}_N^{du}$ and build the RB spaces $\mathcal{V}_N^{pr}$, $\mathcal{V}_N^{du}$;

3. assemble the RB matrices and vectors for the primal and dual problems and the a posteriori RB output error estimate (6.107) (when updating the RB matrices and vectors, simply add the new entries to the old ones);

4. evaluate the primal and dual RB error bounds $\overline{\Delta}_N^{pr}(\boldsymbol{\mu})$ and $\overline{\Delta}_N^{du}(\boldsymbol{\mu})$ (see Eq.(6.108)), $\forall \boldsymbol{\mu} \in \overline{\mathcal{D}}$;

5. if $\max_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} \overline{\Delta}_N^{pr}(\boldsymbol{\mu}) < \sqrt{tol}$ or $\max_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} \overline{\Delta}_N^{du}(\boldsymbol{\mu}) < \sqrt{tol}$, jump to step 7, otherwise go to step 6 (let us observe that, in general, $N^{pr} \neq N^{du}$ and this "stopping" criterium can be fulfilled separately for the primal and dual problem);

6. set $N^{pr} = N^{pr} + 1$ and $N^{du} = N^{du} + 1$, choose the primal and dual samples as:

$$\begin{aligned} \boldsymbol{\mu}_i^{pr} &= \underset{\boldsymbol{\mu}\in\overline{\mathcal{D}}}{\operatorname{argmax}}\, \overline{\Delta}_{i-1}^{pr}(\boldsymbol{\mu}) \qquad i = N^{pr}, \\ \boldsymbol{\mu}_i^{du} &= \underset{\boldsymbol{\mu}\in\overline{\mathcal{D}}}{\operatorname{argmax}}\, \overline{\Delta}_{i-1}^{du}(\boldsymbol{\mu}) \qquad i = N^{du}, \end{aligned} \tag{6.112}$$

   update the samples sets $\mathcal{S}_N^{pr}$, $\mathcal{S}_N^{du}$ and the RB spaces $\mathcal{V}_N^{pr}$, $\mathcal{V}_N^{du}$ and go back to step 3;

7. set $N_{max}^{pr} = N^{pr}$ and $N_{max}^{du} = N^{du}$ and build a matrix (table), say $D_N \in \mathbb{R}^{N_{max}^{pr} \times N_{max}^{du}}$, whose entries are:

$$(D_N)_{i,j} := \max_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} \left(\overline{\Delta}_i^{pr}(\boldsymbol{\mu})\, \overline{\Delta}_j^{du}(\boldsymbol{\mu}) \right) \qquad i = 1, \ldots, N_{max}^{pr}, \quad j = 1, \ldots, N_{max}^{du}; \tag{6.113}$$

8. set a vector of prescribed "error levels", say $E_{lev} \in \mathbb{R}^Z$, s.t. $E_{lev} = \{tol_1, tol_2, \ldots, tol_Z\}$, for some $Z \in \mathbb{N}$ and $tol_1 > tol_2 > \ldots > tol_Z$, with $tol_Z \leq tol$;

9. for each error level $\{tol_m\}_{m=1}^Z$ identify the entries i_m, j_m of the matrix D_N s.t. $(D_N)_{i_m,j_m} < E_{lev}$; among these, select the coordinates N_m^{pr} and N_m^{du} s.t.:

$$\left(N_m^{pr}, N_m^{du} \right) = \underset{i_m, j_m}{\operatorname{argmin}}\, \Lambda_N(i_m, j_m; Q_h) \qquad \forall m = 1, \ldots, Z, \tag{6.114}$$

   where Λ_N is the indicator of the online computational cost (6.111);

10. build a matrix (table), say E_N, with Z rows and 4 columns in order to summarize the results of the whole procedure; in the first column we report the error levels vector E_{lev}, while in the following ones the corresponding number of primal and dual RB basis N_m^{pr}, N_m^{du} and the online computational costs indicator $\Lambda_N(N_m^{pr}, N_m^{du}; Q_h)$, respectively.

Remark 6.10. *The previous adaptive algorithm allows us to avoid the computation of the a posteriori RB error estimate (6.107) in the online context: in fact, once time we have decided an error level, it is immediate to provide the corresponding maximum error bound and the number of basis for primal and dual RB, simply by accessing to the matrix (table)* E_N*. Moreover, in the offline context, we assemble the RB matrices and vectors* A_{Nq}^{pr}*,* A_{Nq}^{du}*,* $\mathbf{F}_N^{pr}$*,* $\mathbf{L}_N^{pr}$ *and* $\mathbf{L}_N^{du}$ *for* N_{max}^{pr} *and* N_{max}^{du}*; in the online context, as observed at step 3 of the adaptive algorithm, for given* $N^{pr} < N_{max}^{pr}$ *and* $N^{du} < N_{max}^{du}$*, the corresponding RB matrices and vectors are obtained simply by eliminating the exceeding rows and columns from the "complete" RB matrices and vectors (the online computational costs associated with these operations are negligible w.r.t. the other dominating costs).*
If a "local" error bound $\forall \boldsymbol{\mu} \in \mathcal{D}$ *it is requested, the computational costs associated with the a posteriori RB error estimate should be added to the indicator (6.111) (see Sec.6.3.4).*

Remark 6.11. *As anticipated in Remark 6.5, we have used the non integrated approach for the choice of the sets* $\mathcal{S}_N^{pr}$ *and* $\mathcal{S}_N^{du}$*. This is the consequence of the adaptive algorithm used: the crucial point it is at step 6, where the samples are chosen according with the primal and dual RB error indicators* $\overline{\Delta}_N^{pr}(\boldsymbol{\mu})$ *and* $\overline{\Delta}_N^{du}(\boldsymbol{\mu})$*. In fact, even if the primal and dual problems assume the same behavior, they can show very different solutions depending also on the functionals* $F_h(\cdot;\boldsymbol{\mu})$ *and* $L(\cdot;\boldsymbol{\mu})$*. For this reason it is evident that the samples for these problems can, and should, be different. Moreover, the adaptive algorithm adopted allows to fit the samples as best as possible, both for the primal and dual problem, even if the ultimate goal consists in computing the output functional.*

6.3.5 Sensitivity analysis

In this Section we provide a sensitivity analysis for the output functional w.r.t. the parameters. We afford this issue directly in the RB context in order to inherit the benefits related to the RB approach for a parametrized problem.

Let us suppose that the functions $\vartheta_{hq}(\boldsymbol{\mu})$, $q = 1, \dots, Q_h$, $\vartheta_{hq}^F(\boldsymbol{\mu})$, $q = 1, \dots, Q_h^F$ and $\vartheta_q^L(\boldsymbol{\mu})$, $q = 1, \dots, Q^L$ (defined in Eq.s (6.7) and (6.23), respectively) belong to $C^1(\mathcal{D})$. For the sake of simplicity, we introduce the following RB space $\mathcal{V}_N := \mathcal{V}_N^{pr} \cup \mathcal{V}_N^{du}$ and we look for the primal and dual RB solutions $v_N(\boldsymbol{\mu})$, $z_N(\boldsymbol{\mu})$ in $\mathcal{V}_N$; let us observe that, by definition, $\mathcal{V}_N^{pr} \subseteq \mathcal{V}_N$ and $\mathcal{V}_N^{du} \subseteq \mathcal{V}_N$. Moreover, for a generic function $\phi_N(\boldsymbol{\mu}) \in \mathcal{V}_N$, we define $\partial \phi_N / \partial \mu_i(\boldsymbol{\mu}) \in \mathcal{V}_N$, for $i = 1, \dots, P$, as the derivative of $\phi_N(\boldsymbol{\mu})$ w.r.t. the i–th component of $\boldsymbol{\mu} \in \mathcal{D} \subset \mathbb{R}^P$ (see [136]). By recalling from Eq.s (6.72) and (6.75) that:

$$s_N(\boldsymbol{\mu}) = L(v_N(\boldsymbol{\mu}); \boldsymbol{\mu}) = -a_h(v_N(\boldsymbol{\mu}), z_N(\boldsymbol{\mu}); \boldsymbol{\mu}), \tag{6.115}$$

with $v_N(\boldsymbol{\mu}), z_N(\boldsymbol{\mu}) \in \mathcal{V}_N$, we obtain:

$$\frac{\partial s_N}{\partial \mu_i}(\boldsymbol{\mu}) = \frac{\partial L}{\partial \mu_i}\left(v_N(\boldsymbol{\mu}); \boldsymbol{\mu}\right) - a_h\left(\frac{\partial v_N}{\partial \mu_i}(\boldsymbol{\mu}), z_N(\boldsymbol{\mu}); \boldsymbol{\mu}\right). \tag{6.116}$$

By derivation of Eq.(6.66) w.r.t. μ_i and by using once time the same identity (6.66), it is simple to show that:

$$a_h\left(\frac{\partial v_N}{\partial \mu_i}(\boldsymbol{\mu}), z_N(\boldsymbol{\mu}); \boldsymbol{\mu}\right) = -\frac{\partial a_h}{\partial \mu_i}\left(v_N(\boldsymbol{\mu}), z_N(\boldsymbol{\mu}); \boldsymbol{\mu}\right) + \frac{\partial F_h}{\partial \mu_i}\left(z_N(\boldsymbol{\mu}); \boldsymbol{\mu}\right); \tag{6.117}$$

finally, by combining Eq.s (6.115) and (6.117), we have:

$$\frac{\partial s_N}{\partial \mu_i}(\boldsymbol{\mu}) = \frac{\partial L}{\partial \mu_i}(v_N(\boldsymbol{\mu}); \boldsymbol{\mu}) - \frac{\partial F_h}{\partial \mu_i}\left(z_N(\boldsymbol{\mu}); \boldsymbol{\mu}\right) + \frac{\partial a_h}{\partial \mu_i}\left(v_N(\boldsymbol{\mu}), z_N(\boldsymbol{\mu}); \boldsymbol{\mu}\right). \tag{6.118}$$

Due to the affine decomposition assumptions (6.7) and (6.23), the sensitivity of the output w.r.t. μ_i reads:

$$\begin{aligned}\frac{\partial s_N}{\partial \mu_i}(\boldsymbol{\mu}) \quad = \quad & \sum_{q=1}^{Q^L} \frac{\partial \vartheta_q^L}{\partial \mu_i}(\boldsymbol{\mu}) L(v_N(\boldsymbol{\mu})) - \sum_{q=1}^{Q_h^F} \frac{\partial \vartheta_{hq}^F}{\partial \mu_i}(\boldsymbol{\mu}) F_{hq}(z_N(\boldsymbol{\mu})) \\ & + \sum_{q=1}^{Q_h} \frac{\partial \vartheta_{hq}}{\partial \mu_i}(\boldsymbol{\mu}) a_{hq}(v_N(\boldsymbol{\mu}), z_N(\boldsymbol{\mu})) \qquad i = 1, \ldots, P.\end{aligned} \tag{6.119}$$

It is simple and computationally not expensive to evaluate the sensitivity at the online step: in fact, as usual, the required RB matrices and vectors calculations are performed in the offline context.

In order to develop the sensitivity analysis, we have assumed the primal and dual RB solutions in $\mathcal{V}_N$; however, in practise, we will continue to search $v_N(\boldsymbol{\mu}) \in \mathcal{V}_N^{pr}$ and $z_N(\boldsymbol{\mu}) \in \mathcal{V}_N^{du}$. This could affect the sensitivity analysis (6.119), even if we suppose that this perturbation is not significant; this can be considered true if N^{pr} and N^{du} are sufficiently "large".

6.4 Combining the Finite Element and Reduced Basis approximations

In the Sec.s 6.2 and 6.3 we have discussed the FE and RB methods for the numerical solution of the parametrized problem introduced in Sec.6.1.1. However these approaches should not be seen as separate for the solution of our problem, but rather as complementary.

Let us recall that our continuous problem is hyperbolic as shown in Sec.6.1.1. For the numerical approximation of this problem we have considered the FE method with stabilization; with this aim, we have provided the stabilized problem in Sec.6.2.1 by introducing the stabilized versions of the bilinear form $a_h(\cdot, \cdot; \boldsymbol{\mu})$ and functional $F_h(\cdot; \boldsymbol{\mu})$. Let us observe that at this step, we have the following error on the output:

$$|s(\boldsymbol{\mu}) - s_h(\boldsymbol{\mu})| \qquad \text{with } \boldsymbol{\mu} \in \mathcal{D}, \tag{6.120}$$

where $s(\boldsymbol{\mu})$ is the exact value of the output, while $s_h(\boldsymbol{\mu})$ is the output computed by means of the FE method corresponding to the parameter value $\boldsymbol{\mu} \in \mathcal{D}$ (i.e. $\forall \boldsymbol{\mu} \in \mathcal{D}$ we solve a computational expensive FE problem). As shown by the a priori error estimate (6.60), the error $|s(\boldsymbol{\mu}) - s_h(\boldsymbol{\mu})| \to 0$, as $h \to 0$, $\forall \boldsymbol{\mu} \in \mathcal{D}$.

The RB method is based on the FE approximations, which are used to set the RB basis; with this aim, the stabilized Galerkin problem is used in order to define the RB method in Sec.6.3.

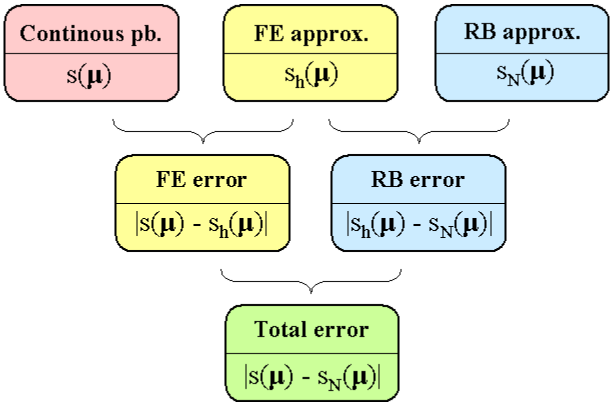

Figure 6.1: Scheme for numerical solution of the parameterized problem: total, FE and RB errors.

For this reason the RB error on the output is evaluated w.r.t. the output corresponding to the FE approximation for a given parameter $\boldsymbol{\mu} \in \mathcal{D}$, i.e.:

$$|s_h(\boldsymbol{\mu}) - s_N(\boldsymbol{\mu})| \qquad \text{with } \boldsymbol{\mu} \in \mathcal{D}, \tag{6.121}$$

being $s_N(\boldsymbol{\mu})$ the output computed by means of the RB method. The a priori RB error estimate (6.88) shows that the error $|s_h(\boldsymbol{\mu}) - s_N(\boldsymbol{\mu})| \to 0$, as N increases ($N \to \infty$), where N indicates generically the size of the primal and dual RB problems.
The total error on the output is composed by two terms, given form Eq.s (6.120) and (6.121):

$$|s(\boldsymbol{\mu}) - s_N(\boldsymbol{\mu})| \leq |s(\boldsymbol{\mu}) - s_h(\boldsymbol{\mu})| + |s_h(\boldsymbol{\mu}) - s_N(\boldsymbol{\mu})| \qquad \text{with } \boldsymbol{\mu} \in \mathcal{D}; \tag{6.122}$$

in Fig.6.1 the previous issues are outlined by means of a block diagram. However, even if $N \to \infty$ the total error reduces, due to the reduction of the RB part of the error, this is not the case for $h \to 0$. In fact, by referring e.g. to the Problem 1(bis), if $h \to 0$, the FE error reduces according with the a priori FE error estimate (6.60), but, for a prescribed N, the RB error increases; this is due to the term $\Xi(\boldsymbol{\mu})$ (6.84) of the a priori RB estimate (6.88), which is of order $h^{-1/2}$. This shows that, as $h \to 0$, for a prescribed $\boldsymbol{\mu} \in \mathcal{D}$ and a fixed N, the RB error could increase; in other words, if the stabilized problem tends to assume an hyperbolic nature (for $h \to 0$), the "complexity" of the RB problem enhances, i.e. the number of RB basis should increase in order to obtain a prescribed accuracy on the RB error.
For further details and examples about the RB method for hyperbolic problems, see [135, 104].

6.5 Numerical tests

In this Section we report two numerical tests, which refer to the problems outlined in Sec.s 6.1.2 and 6.1.3, concerning the advection–reaction problem with physical and geometrical parametrizations. In particular, we highlight the role of the dual problem in the RB method, specifically in the selection of the samples for the basis and in the reduction of the online computational costs; moreover, we provide a sensitivity analysis for the output functional w.r.t. the parameters of the problem. Both the numerical tests are inspired by environmental problems concerning air pollution (see Chapter 1), for which the goal consists in evaluating

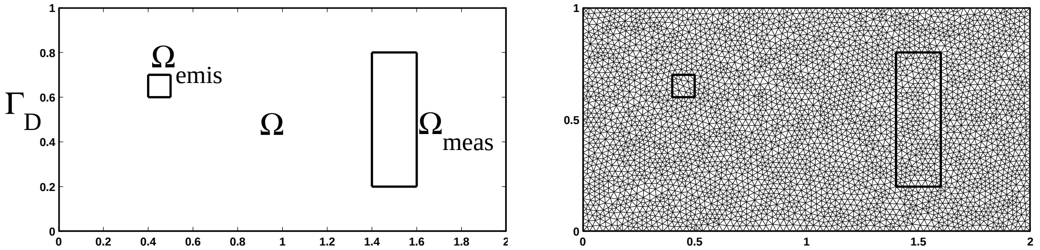

Figure 6.2: PB1. Domain, subdomains and Dirichlet boundary (left); quasi–uniform mesh with 7, 556 triangles (right).

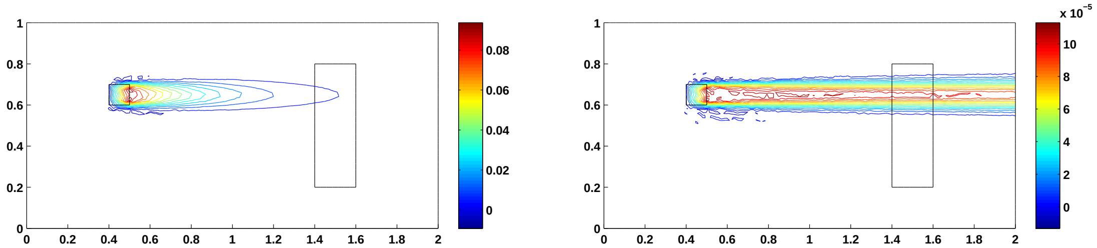

Figure 6.3: PB1. Primal FE solutions for $\mu = 1$ (left) and $\mu = 10^3$ (right).

the mean concentration of a pollutant in an area of interest (e.g. a city) emitted by a source (e.g. an industrial chimney) (see [44, 149, 151]). Such a pollutant is transported by a wind field and can react in air; diffusion processes are considered negligible.

6.5.1 Problem 1: physical parametrization

In this Section we report the numerical results concerning the Problem 1 outlined in Sec.6.1.2. By referring to the problem (6.8), we consider the domain Ω reported in Fig.6.2(left), where we have defined two internal subdomains Ω_{emis} and Ω_{meas}. We assume $\mu \in \mathcal{D} = [1, 10^3]$, $\mathbf{b} = \hat{\mathbf{x}}$, $g = \chi_{emis}(\mathbf{x})$ and $\delta = 1/|\Omega_{meas}|\ \chi_{meas}(\mathbf{x})$, where $\chi_{emis}(\mathbf{x})$ and $\chi_{meas}(\mathbf{x})$ are the characteristic functions of the subdomains Ω_{emis} and Ω_{meas}, respectively; $|\Omega_{meas}|$ is the area of Ω_{meas}. From this, it follows that $\Gamma_D = \{\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2) \in \partial\Omega\ :\ \mathrm{x}_1 = 0\}$, as reported in Fig.6.2(left). Let us recall that this problem admits an affine decomposition on the bilinear form and functionals that define the weak problem; in particular, by referring to Sec.6.2.1, we have $Q_h = 2$, $Q_h^F = 1$ and $Q^L = 1$ (see Eq.s (6.23) and (6.7)). In order to evaluate the norm $|||\cdot|||$, which depends on μ as reported in Sec.6.3.4, we choose $\overline{\mu} = 1$; for the sake of simplicity we indicate $|||\cdot|||_{\overline{\mu}}$ simply by $|||\cdot|||$.

In Fig.6.2(right) we report the quasi–uniform mesh with 7, 556 triangles ($h = 0.0265$) which we use to solve the FE problem for different values of the parameter μ. Moreover, we choose the constants (6.20) as $\varepsilon_h = c_\varepsilon \mu h^{3/2}$ and $\delta_h = c_\delta h/\mu$, with $c_\varepsilon = 5 \cdot 10^{-2}$ and $c_\delta = 5 \cdot 10^{-2}$,

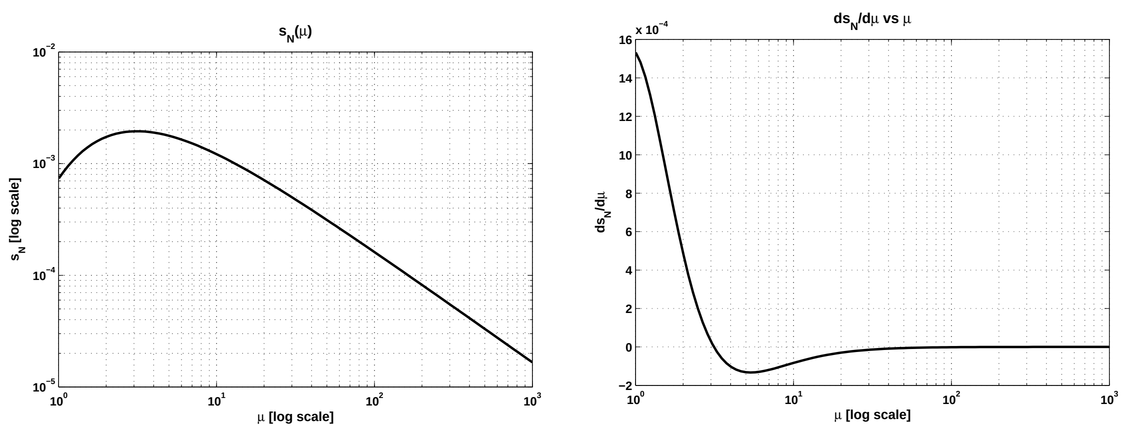

Figure 6.4: PB1. Output $s_N(\mu)$ (left) and $(\partial s_N/\partial\mu)(\mu)$ vs $\mu \in \mathcal{D}$; logarithmic scale on both axis.

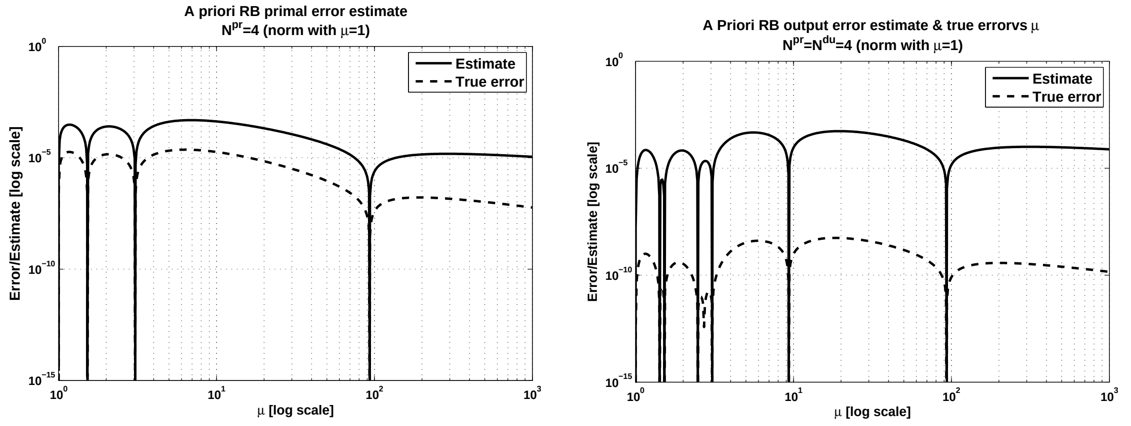

Figure 6.5: PB1. A priori RB error estimates (continuous) and true RB errors (dashed) vs $\mu \in \mathcal{D}$: primal RB error with $N^{pr} = 4$ (left) and output RB error with $N^{pr} = N^{du} = 4$ (right); logarithmic scales on both axis.

which allows to contain the under and over–shooting of the FE solution $\forall \mu \in \mathcal{D}$. In Fig.6.3 we report the FE solutions of the primal problem for the choices $\mu = 1$ (left) and $\mu = 10^3$ (right); the corresponding FE dual solutions show similar behaviors even if the solutions arise from Ω_{meas} and the flow direction is opposite to that of the primal one.

We solve now the parametrized problem by means of the RB method outlined in Sec.6.3 and we provide some results concerning the a priori and a posteriori RB error estimates.

First of all, we report in Fig.6.4(left) the output $s_N(\mu)$ vs $\mu \in \mathcal{D}$ computed by means of the RB method, with a number of basis "sufficiently" large s.t. $s_N(\mu)$ is a good approximation of $s_h(\mu)$ for all $\mu \in \mathcal{D}$. In Fig.6.4(right) we report the sensitivity analysis $(\partial s_N/\partial\mu)(\mu)$ computed by means of the RB method according with Eq.(6.119).

In Fig.6.5(left) we compare the a priori RB primal error estimate (6.83) with the true RB error on the primal solution $|||e_N^{pr}(\mu)|||$ for different values of $\mu \in \mathcal{D}$, being $N^{pr} = 4$ with

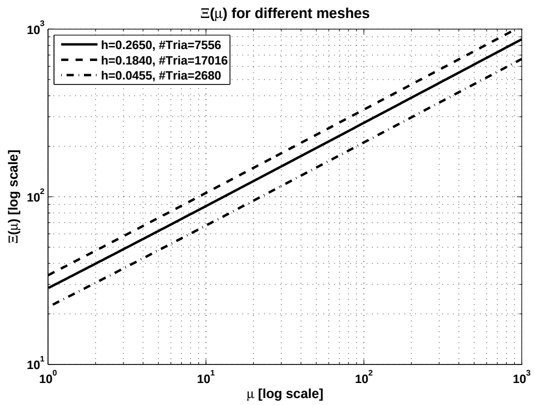

Figure 6.6: PB1. $\Xi(\mu)$ (6.85) vs $\mu \in \mathcal{D}$ for meshes with $2,680$ (dot–dashed), $7,556$ (continuous) and $17,016$ (dashed) triangles; logarithmic scales on both axis.

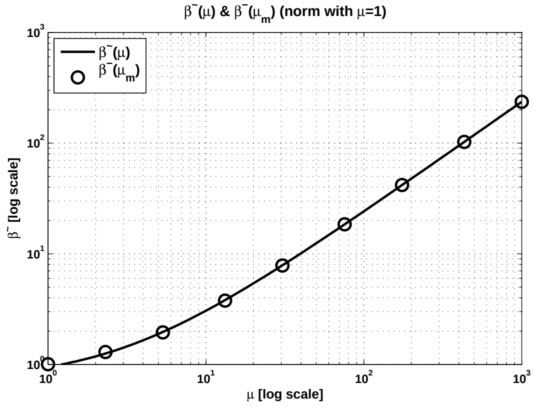

Figure 6.7: PB1. $\widetilde{\beta}(\mu)$ vs $\mu \in \mathcal{D}$ (continuous) and the pairs $\{(\mu_m, \overline{\beta}(\mu_m)\}_{m=1}^{9}$ (circle); logarithmic scales on both axis.

$\mathcal{S}_N^{pr} = \{1.00,\ 1.52,\ 3.06,\ 93.29\}$. In similar way, in Fig.6.5(right) the a priori RB error estimate for the output (6.88) is compared with the true output error $|s_h(\mu) - s_N(\mu)|$ for $\mu \in \mathcal{D}$; in this case $N^{pr} = N^{du} = 4$, with the set $\mathcal{S}_N^{pr}$ chosen as previously mentioned and $\mathcal{S}_N^{du} = \{1.00,\ 1.42,\ 2.48,\ 9.33\}$. In Fig.6.6 we show the effect of the FE mesh on $\Xi(\mu)$ (6.84) (see Eq.(6.85) for Problem 1) by comparing three quasi–uniform meshes with $2,680$, $7,556$ and $17,016$ triangles; we observe that as h decreases (i.e. the FE solution "improves"), the parametrized constant $\Xi(\mu)$ increases for any fixed values of $\mu \in \mathcal{D}$, thus affecting the sharpness property of the a priori RB estimates (6.83), (6.87) and (6.88). This implies that, as $h \rightarrow 0$, N^{pr} and N^{du} should increase in order to contain the RB errors (primal, dual and output) for any fixed $\mu \in \mathcal{D}$; this confirms the considerations of Sec.6.4 about the effectivity of the RB method for hyperbolic problems.

We deals now with the a posteriori RB error estimate outlined in Sec.6.3.3 which we use in the adaptive algorithm of Sec.6.3.4 for the choice of the RB samples and basis. As first issue for the computation of the a posteriori RB error estimate for the output (6.101), we need to

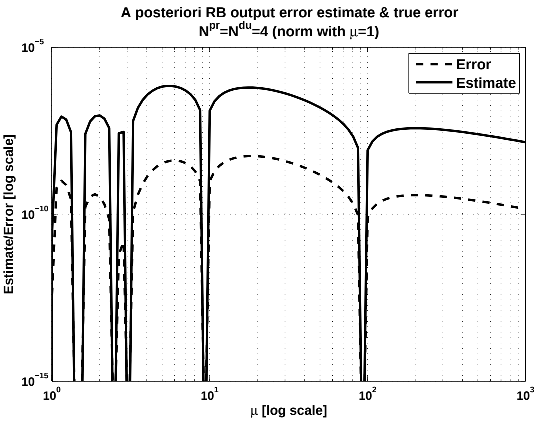

Figure 6.8: PB1. A posteriori RB error estimate for the output (6.107) with $N^{pr} = N^{du} = 4$ (continuous) and true error (dashed) vs $\mu \in \mathcal{D}$; logarithmic scales on both axis.

evaluate the parametrized inf–sup stability constant $\beta(\mu)$ (6.97); with this aim, we use the numerical procedure outlined in Sec.6.3.4 (Eq.s (6.105) and (6.106)). Being $\mathcal{D} \subset \mathbb{R}$, we adopt for $\{\Pi_m(\mu)\}_{m=1}^{K}$ a polinomial approximation in the least–squares sense of degree 3 obtained by means of the logarithmic–equally spaced pairs $(\mu_m, \bar{\beta}(\mu_m))$, $m = 1, \ldots, K$ with $K = 9$; in Fig.6.7 we plot the computed $\widetilde{\beta}(\mu)$ for $\mu \in \mathcal{D}$. In particular we observe that, according with the choice made for $\overline{\mu} = 1$ ($|||\cdot|||_{\overline{\mu}}$), we have, for our problem, $\widetilde{\beta}(1) = 1$; moreover, we remark that $\widetilde{\beta}(\mu) \gg \alpha$, for $\mu \gg 1$, being $\alpha = 1$ the coercivity constant (6.37). In Fig.6.8 we report the a posteriori RB error estimate for the output (6.101), which we evaluate by means of Eq.(6.107), and the true error for $N^{pr} = N^{du} = 4$; the samples sets $\mathcal{S}_N^{pr}$ and $\mathcal{S}_N^{du}$ are the same chosen for the a priori RB error estimates of Fig.s 6.5(left) and (right).

The a posteriori RB error estimate (6.107) is used in the adaptive algorithm introduced in Sec.6.3.4 for the choice of the samples sets $\mathcal{S}_N^{pr}$ and $\mathcal{S}_N^{du}$. The results of the adaptive algorithm are summarized in Table 6.1 by means of the matrix (table) E_N defined at point 10 of Sec.6.3.4; in order to show the effectivity of the "primal–dual" RB approach we compare the results with those of an "only primal" RB approach, evidencing the savings of computational costs allowed by the former one. In Table 6.1 we report, for each error level E_{lev} for the output, the selected number of basis N^{pr}, N^{du} (primal and dual) and the computational costs associated with the "primal–dual" ($\Lambda_N^{pr,du} = \Lambda_N$, see Eq.(6.111)) and the "only primal" (Λ_N^{pr}) RB approaches respectively; moreover, we report the ratio $\Lambda_N^{pr}/\Lambda_N^{pr,du}$ for each error level. Let us notice that in the case for which $N^{pr} > 0$ and $N^{du} = 0$, the "primal–dual" RB approach is equivalent ot the "only primal" one. As we can observe from Table 6.1, $\Lambda_N^{pr} \geq \Lambda_N^{pr,du}$ $\forall E_{lev}$; this shows the effectiveness of the "primal–dual" RB approach w.r.t. the "only primal" one, which is more expensive even for this simple test problem with a single parameter μ. For example, if we require an error level $E_{lev} = 10^{-8}$ (i.e. the relative error on $s(\mu)$ is inferior to 0.1% $\forall \mu \in \mathcal{D}$, see Fig.6.4(left)) the "primal–dual" RB approach selects $N^{pr} = N^{du} = 5$ w.r.t. $N^{pr} = 8$ requested by the "only primal" RB approach and it allows a saving of the online computational costs of about the 60%. We notice that even larger computational costs savings can be obtained; e.g. for $E_{lev} = 1.778 \cdot 10^{-10}$, the saving is about the 130%. Let us observe that the adaptive algorithm does not tend to select the case $N^{du} = 0$ as E_{lev} decreases; in fact,

	“primal–dual” RB			“only primal” RB		
E_{lev}	N^{pr}	N^{du}	$\Lambda_N^{pr,du}$	N^{pr}	Λ_N^{pr}	$\Lambda_N^{pr}/\Lambda_N^{pr,du}$
$1.000e+01$	1	0	$3.000e+00$	1	$3.000e+00$	1.000
$5.623e+00$	1	0	$3.000e+00$	1	$3.000e+00$	1.000
$3.162e+00$	1	0	$3.000e+00$	1	$3.000e+00$	1.000
$1.778e+00$	1	0	$3.000e+00$	1	$3.000e+00$	1.000
$1.000e+00$	1	0	$3.000e+00$	1	$3.000e+00$	1.000
$5.623e-01$	1	0	$3.000e+00$	1	$3.000e+00$	1.000
$3.162e-01$	1	0	$3.000e+00$	1	$3.000e+00$	1.000
$1.778e-01$	1	0	$3.000e+00$	1	$3.000e+00$	1.000
$1.000e-01$	1	0	$3.000e+00$	1	$3.000e+00$	1.000
$5.623e-02$	1	1	$8.000e+00$	2	$1.600e+01$	2.000
$3.162e-02$	2	0	$1.600e+01$	2	$1.600e+01$	1.000
$1.778e-02$	1	2	$2.300e+01$	3	$4.500e+01$	1.957
$1.000e-02$	2	2	$4.000e+01$	3	$4.500e+01$	1.125
$5.623e-03$	3	0	$4.500e+01$	3	$4.500e+01$	1.000
$3.162e-03$	3	0	$4.500e+01$	3	$4.500e+01$	1.000
$1.778e-03$	1	3	$5.400e+01$	4	$9.600e+01$	1.778
$1.000e-03$	2	3	$7.300e+01$	4	$9.600e+01$	1.315
$5.623e-04$	2	3	$7.300e+01$	4	$9.600e+01$	1.315
$3.162e-04$	4	0	$9.600e+01$	4	$9.600e+01$	1.000
$1.778e-04$	4	1	$1.070e+02$	5	$1.750e+02$	1.636
$1.000e-04$	4	1	$1.070e+02$	5	$1.750e+02$	1.636
$5.623e-05$	4	2	$1.280e+02$	5	$1.750e+02$	1.367
$3.162e-05$	4	3	$1.650e+02$	5	$1.750e+02$	1.061
$1.778e-05$	4	3	$1.650e+02$	6	$2.880e+02$	1.745
$1.000e-05$	4	3	$1.650e+02$	6	$2.880e+02$	1.745
$5.623e-06$	4	3	$1.650e+02$	6	$2.880e+02$	1.745
$3.162e-06$	5	2	$2.110e+02$	6	$2.880e+02$	1.365
$1.778e-06$	4	4	$2.240e+02$	6	$2.880e+02$	1.286
$1.000e-06$	4	4	$2.240e+02$	6	$2.880e+02$	1.286
$5.623e-07$	3	5	$2.500e+02$	7	$4.410e+02$	1.764
$3.162e-07$	5	4	$3.110e+02$	7	$4.410e+02$	1.418
$1.778e-07$	5	4	$3.110e+02$	7	$4.410e+02$	1.418
$1.000e-07$	5	4	$3.110e+02$	7	$4.410e+02$	1.418
$5.623e-08$	5	4	$3.110e+02$	7	$4.410e+02$	1.418
$3.162e-08$	5	4	$3.110e+02$	8	$6.400e+02$	2.058
$1.778e-08$	5	4	$3.110e+02$	8	$6.400e+02$	2.058
$1.000e-08$	5	5	$4.000e+02$	8	$6.400e+02$	1.600
$5.623e-09$	5	5	$4.000e+02$	9	$8.910e+02$	2.228
$3.162e-09$	6	4	$4.320e+02$	9	$8.910e+02$	2.062
$1.778e-09$	6	5	$5.230e+02$	9	$8.910e+02$	1.704
$1.000e-09$	6	5	$5.230e+02$	9	$8.910e+02$	1.704
$5.623e-10$	6	5	$5.230e+02$	9	$8.910e+02$	1.704
$3.162e-10$	6	5	$5.230e+02$	10	$1.200e+03$	2.294
$1.778e-10$	6	5	$5.230e+02$	10	$1.200e+03$	2.294
$1.000e-10$	6	6	$6.480e+02$	10	$1.200e+03$	1.852
$5.623e-11$	6	6	$6.480e+02$	10	$1.200e+03$	1.852
$3.162e-11$	6	6	$6.480e+02$	10	$1.200e+03$	1.852
$1.778e-11$	5	7	$6.860e+02$			
$1.000e-11$	6	7	$8.130e+02$			
$5.623e-12$	6	7	$8.130e+02$			
$3.162e-12$	6	7	$8.130e+02$			
$1.778e-12$	6	7	$8.130e+02$			
$1.000e-12$	5	8	$8.950e+02$			

Table 6.1: PB1. Adaptive algorithm: results of samples selection procedure and comparison between the ‘primal–dual” and the “only primal” RB approach.

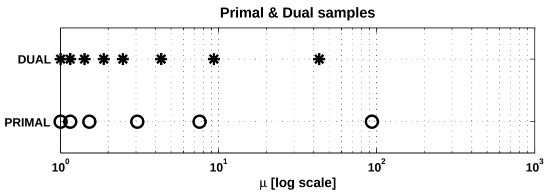

Figure 6.9: PB1. Samples sets $\mathcal{S}_N^{pr}$ (circle) and $\mathcal{S}_N^{du}$ (starred) selected by the adaptive algorithm of Sec.6.3.4; axis in logarithmic scale.

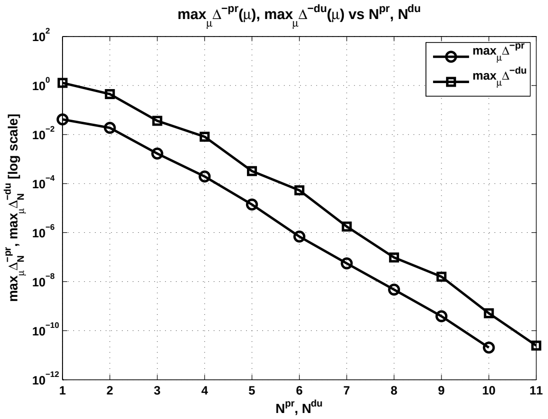

Figure 6.10: PB1. RB error indicators $\max_{\mu\in\mathcal{D}} \overline{\Delta}_N^{pr}(\mu)$ (circle) and $\max_{\mu\in\mathcal{D}} \overline{\Delta}_N^{du}(\mu)$ (square) vs N^{pr} and N^{du}; error axis in logarithmic scale.

$\Lambda_N^{pr,du}$ increases if $N^{pr} \neq N^{du}$. In Fig.6.9 we report the samples selected for the sets $\mathcal{S}_N^{pr}$ and $\mathcal{S}_N^{du}$; we observe that the inferior extremum of the set $\mathcal{D}$ $(\mu = 1)$ is selected both for the primal and dual problem and that the upper part of the parameters domain $\mathcal{D}$ $(\mu\tilde{>}\mu_\star = 10^2)$ is not "crucial" for the problem under consideration, being all the samples inferior to $\mu_\star$. In Fig.6.10 we report the behavior of the RB error indicators $\max_{\mu\in\mathcal{D}} \overline{\Delta}_N^{pr}(\mu)$ vs N^{pr} and $\max_{\mu\in\mathcal{D}} \overline{\Delta}_N^{du}(\mu)$ vs N^{du}, where $\overline{\Delta}_N^{pr}(\mu)$ and $\overline{\Delta}_N^{du}(\mu)$ compose the a posteriori RB estimate (6.107) and are defined in Eq.(6.108); the plot shows that the convergence rate for these primal and dual error indicators is at most the same, thus indicating that also the "complexity" of the primal and the dual problem is the same.

Finally, we compare the true RB errors for the output obtained by using the adaptive algorithm for the "primal–dual" RB approach with those of the "only primal" one; with this aim we recall the RB error indicators for the output E_N^{max} and E_N^{mean} (see Eq.(6.109)). In Fig.6.11(left) we report E_N^{max} and E_N^{mean} vs $N^{pr} + N^{du}$ for both the "primal–dual" and "only primal" RB approaches, while in Fig.6.11(right) the same RB error indicators vs the online computational cost $\Lambda_N^{pr,du}$ and Λ_N^{pr}. We observe that the first plot does not give any useful indication on which could be the best RB approach, even if the convergence of the RB error as N^{pr} and N^{du} increase is shown. On the contrary Fig.6.11(right) shows that the "primal–dual" RB approach allows to minimize the online computational costs for any

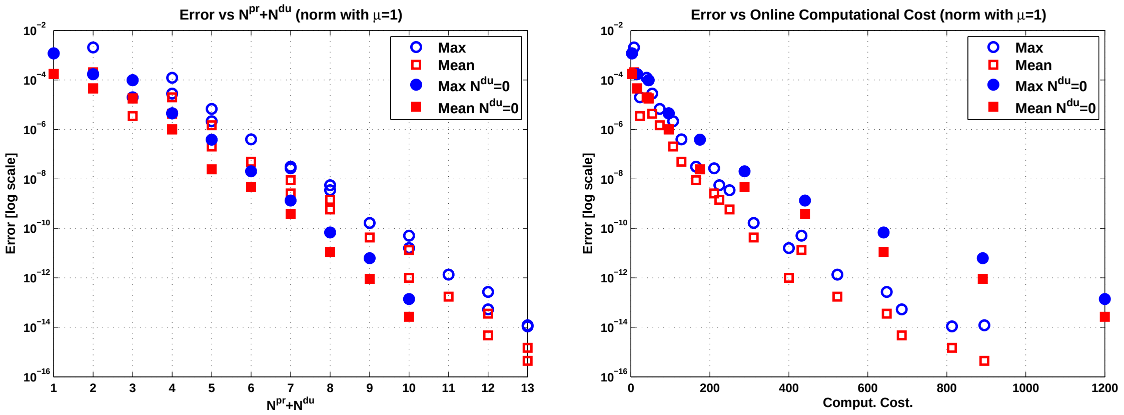

Figure 6.11: PB1. E_N^{max} (empty circle) and E_N^{mean} (empty square) for the "primal–dual" RB approach vs $N^{pr} + N^{du}$ (left) and online computational costs $\Lambda_N^{pr,du}$, Λ_N^{pr} (right); full circles and squares refer to the indicators E_N^{max}, E_N^{mean} for the "only primal" RB approach. Error axis in logarithmic scale.

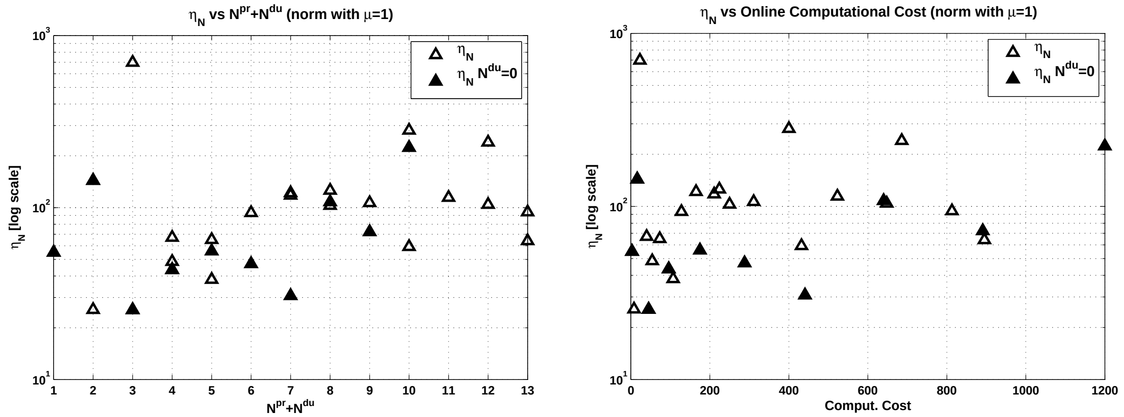

Figure 6.12: PB1. $\overline{\eta}_N$ (empty triangle) for the "primal–dual" RB approach vs $N^{pr} + N^{du}$ (left) and online computational costs $\Lambda_N^{pr,du}$, Λ_N^{pr} (right); full triangles refer to the indicator $\overline{\eta}_N$ for the "only primal" RB approach. Effectivity index axis in logarithmic scale.

given error tolerance on the output; for example if we fix the tolerance error at 10^{-12} we see that the online computational costs associated with the "primal–dual" approach is about the half of that corresponding to the "only primal" approach. This confirms, a posteriori, the validity of the adaptive algorithm and the criterium of the "minimum online computational costs". In similar way we report Fig.6.12(left) and (right) the effectivity index indicator $\overline{\eta}_N$ vs $N^{pr} + N^{du}$ and $\Lambda_N^{pr,du}$, Λ_N^{pr} for both the RB approaches. We notice that $\overline{\eta}_N$ is always greater than 10, but no particular differences are shown on the effectivity index indicator between the "primal–dual" and "only primal" RB approaches.

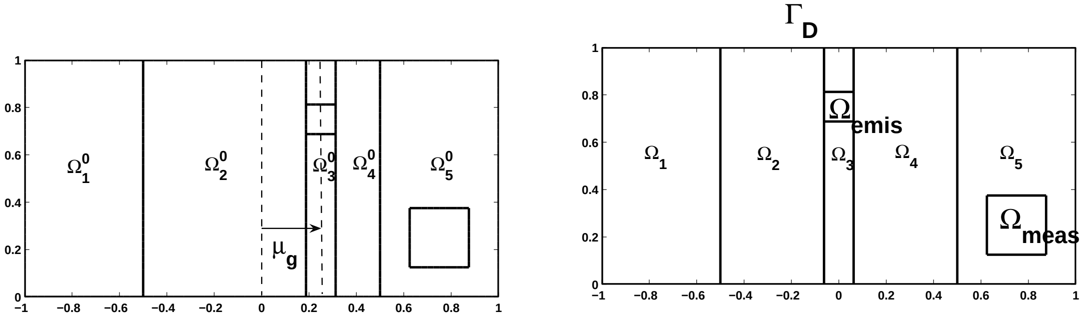

Figure 6.13: PB2. Real Ω_0 (left) and reference Ω (right) domains; geometrical parameter and Dirichlet boundary.

6.5.2 Problem 2: physical and geometrical parametrization

We consider now the Problem 2 outlined in Sec.6.1.3; in this case we deals now with both physical and geometrical parameters.

As anticipated in Sec.6.1.3, we introduce the geometrical parameter, μ_g s.t. the real domain Ω_0 could be affinely mapped into a reference one Ω. In Fig.6.13(left) we report the real domain Ω_0, which is partitioned into 5 subdomains Ω_{0i}, s.t. $\cup_{i=1}^5 \Omega_{0i} = \Omega_0$. The geometrical parameter μ_g (with signum) measures the distance between the mid abscissa of the subdomain Ω_{03} and the coordinate $\mathrm{x}_0 = 0$; the subdomains Ω_{01} and Ω_{05} are fixed, while Ω_{02} and Ω_{04} deform in affine manner according with the moving of Ω_{03}. All the subdomains Ω_{0i} can be mapped in Ω by means of affine maps in the form (6.16); the reference domain $\Omega = \cup_{i=1}^5 \Omega_i$ is reported in Fig.6.13(right). We remark that the subdomains Ω_{emis} and Ω_{meas} move fixed with the subdomains Ω_{03} and Ω_{05} respectively; the Dirichlet boundary, which corresponds to the upper boundary of Ω, is shown in Fig.6.13(right).

The parameter vector $\boldsymbol{\mu} = (\mu_g, \mu_p)$ is taken in the parameter domain $\mathcal{D} = [\mu_g^{min}, \mu_g^{max}] \times [\mu_p^{min}, \mu_p^{max}]$, where we choose $\mu_g^{min} = -3/8$, $\mu_g^{max} = 3/8$, $\mu_p^{min} = 1$ and $\mu_p^{max} = 10^3$. We observe that the choices made for μ_g^{min} and μ_g^{max} do not allow the degeneration of the subdomains Ω_{02} and Ω_{04}. By recalling the notation of Sec.s 6.1.3 and 6.5.1, we choose $\mathbf{b}_0 = \mathrm{x}_{1,0}\hat{\mathbf{x}} - \mathrm{x}_{2,0}\hat{\mathbf{y}}$, $g_0 = \chi_{meas}(\mathbf{x}_0)$ and $\delta_0 = 1/|\Omega_{meas}|\ \chi_{meas}(\mathbf{x}_0)$. By considering the data reported on the reference domain (see Eq.(6.17)), we obtain the problem in the general weak form (6.2). It is simple to show that, for the problem under consideration, the hypotheses on the data requested in Sec.6.1.1 are satisfied; in particular the property $\nabla \cdot \mathbf{V}(\boldsymbol{\mu}) = 0\ \forall \boldsymbol{\mu} \in \mathcal{D}$ holds. The weak problem admits an affine decomposition for which we have $Q_h = 16$, $Q_h^F = 2$ and $Q^L = 1$ (see Eq.s (6.20) and (6.7)). For the evaluation of the norm $|||\cdot|||$ we choose $\overline{\boldsymbol{\mu}} = (\overline{\mu}_g, \overline{\mu}_p) = (0, 1)$ as reported in Sec.6.3.4.

For the FE solution of the problem we choose a quasi–uniform mesh for the reference domain Ω with 7,504 triangles and the constants (6.20) as $\varepsilon_h(h, \mu_p) = c_\varepsilon \mu_p h^{3/2}$ and $\delta_h = c_\delta h/\mu_p$ with $c_\varepsilon = 2{\cdot}10^{-1}$ and $c_\delta = 2{\cdot}10^{-1}$. In Fig.s 6.14(left) and (right) we report the primal FE solutions, mapped from the reference domain Ω into the real one Ω_0, corresponding to the choices $\boldsymbol{\mu} = (0, 1)$ and $\boldsymbol{\mu} = (0, 10^3)$ respectively. We observe that, if $\mu_g = 0\ \forall \mu_p \in [\mu_p^{min}, \mu_p^{max}]$, the real domain Ω_0 coincides with the reference one Ω. In similar way, Fig.s 6.15(left) and (right) show the primal FE solutions for $\boldsymbol{\mu} = (-3/8, 1)$ and $\boldsymbol{\mu} = (-3/8, 10^3)$ respectively; finally, in

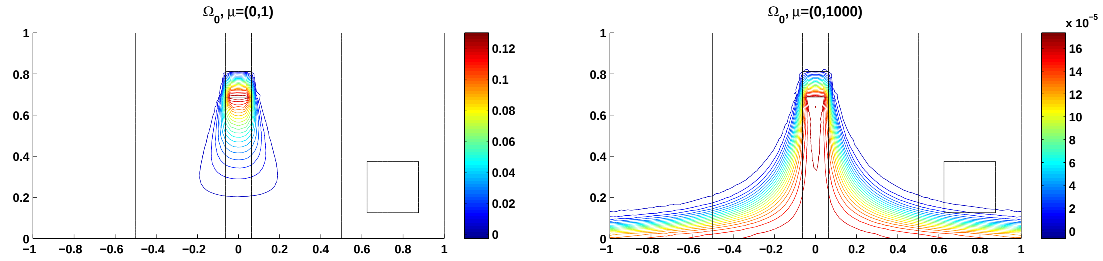

Figure 6.14: PB2. Primal FE solutions in the real domain Ω_0 for $\boldsymbol{\mu} = (0, 1)$ (left) and $\boldsymbol{\mu} = (0, 10^3)$ (right).

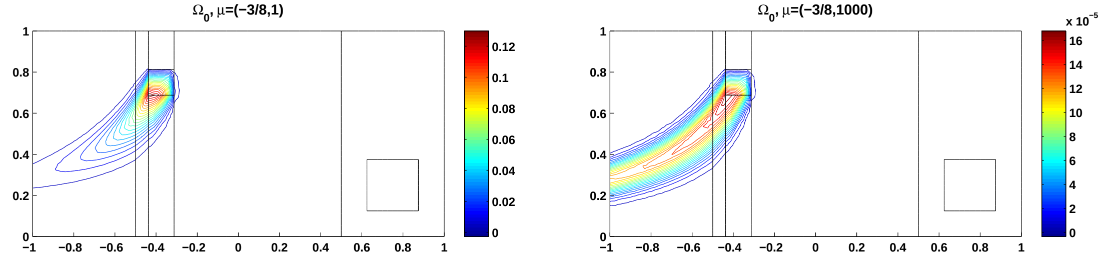

Figure 6.15: PB2. Primal FE solutions in the real domain Ω_0 for $\boldsymbol{\mu} = (-3/8, 1)$ (left) and $\boldsymbol{\mu} = (-3/8, 10^3)$ (right).

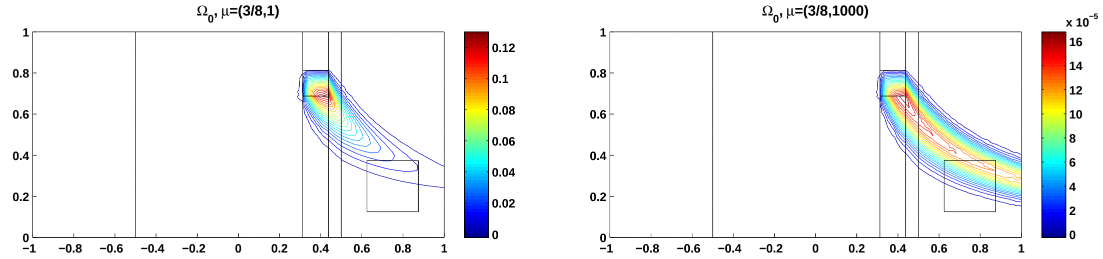

Figure 6.16: PB2. Primal FE solutions in the real domain Ω_0 for $\boldsymbol{\mu} = (3/8, 1)$ (left) and $\boldsymbol{\mu} = (3/8, 10^3)$ (right).

Fig.s 6.16(left) and (right) the cases $\boldsymbol{\mu} = (3/8, 1)$ and $\boldsymbol{\mu} = (3/8, 10^3)$ are reported. As shown, the primal solution strongly depends on the values assumed by the geometrical parameter μ_p and not only by the physical one μ_p. On the contrary, for this particular problem, the

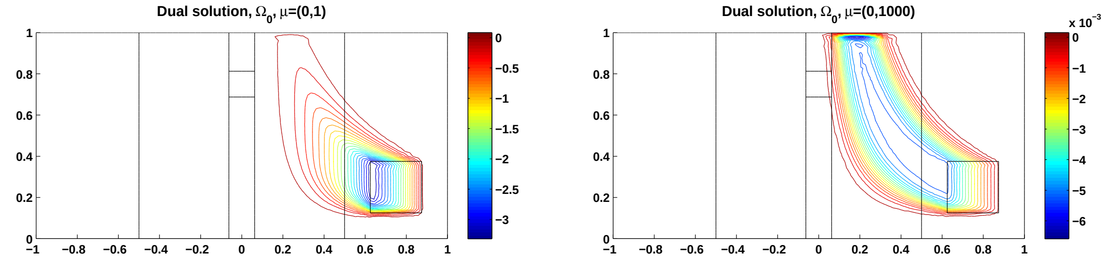

Figure 6.17: PB2. Dual FE solutions in the real domain Ω_0 for $\boldsymbol{\mu} = (0, 1)$ (left) and $\boldsymbol{\mu} = (0, 10^3)$ (right).

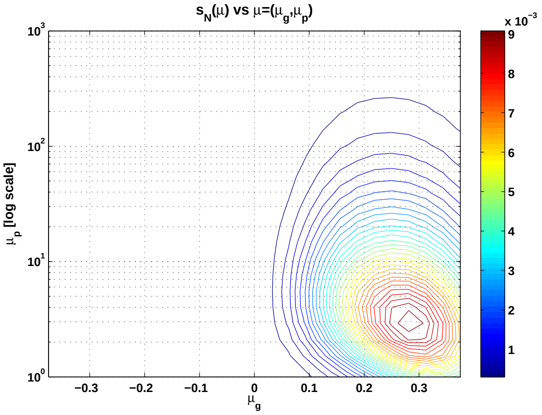

Figure 6.18: PB2. Contourlines of $s_N(\boldsymbol{\mu})$ vs $\boldsymbol{\mu} = (\mu_g, \mu_p) \in \mathcal{D}$; axis μ_p in logarithmic scale.

behavior of the dual solution in the real domain Ω_0 is not influenced by the geometrical parameter, being the "dual source" subdomain Ω_{meas} fixed. However, this is not true in the reference domain Ω, being the weak form of the stabilized dual problem (6.30) depending on μ_g. In Fig.s 6.17(left) and (right) the dual FE solutions, mapped into the real domain Ω_0, are reported for $\boldsymbol{\mu} = (0, 1)$ and $\boldsymbol{\mu} = (0, 10^3)$ respectively.

We solve the parametrized problem by means of the RB method and, in similar way as for the numerical test Problem 1, we discuss the results about the adaptive algorithm and the a posteriori RB estimate.

In Fig.6.18 we report the output $s_N(\boldsymbol{\mu})$ vs $\boldsymbol{\mu} \in \mathcal{D}$, where we notice that $s_N(\boldsymbol{\mu})$ is maximum for about $\boldsymbol{\mu} = (0.28, 3.0)$; as expected by observing the Fig.s (6.14), (6.15) and (6.16), if $\mu_g < 0$ the subdomain Ω_{meas} is not interested by appreciable values of the primal solution. We recall that the RB output $s_N(\boldsymbol{\mu})$ is a good approximation of $s_h(\boldsymbol{\mu})$ if N is "sufficiently" large. In Fig.6.19(left) and (right) we report the sensitivity analysis of the output (see Eq.(6.119)). We confirm that, in the part of the parameter domain $\mathcal{D}$ s.t. $\mu_g < 0$, no substantial variations of $s_N(\boldsymbol{\mu})$ occur.

We deals now with the a posteriori RB error estimate for the output of Sec.6.3.3 and the adap-

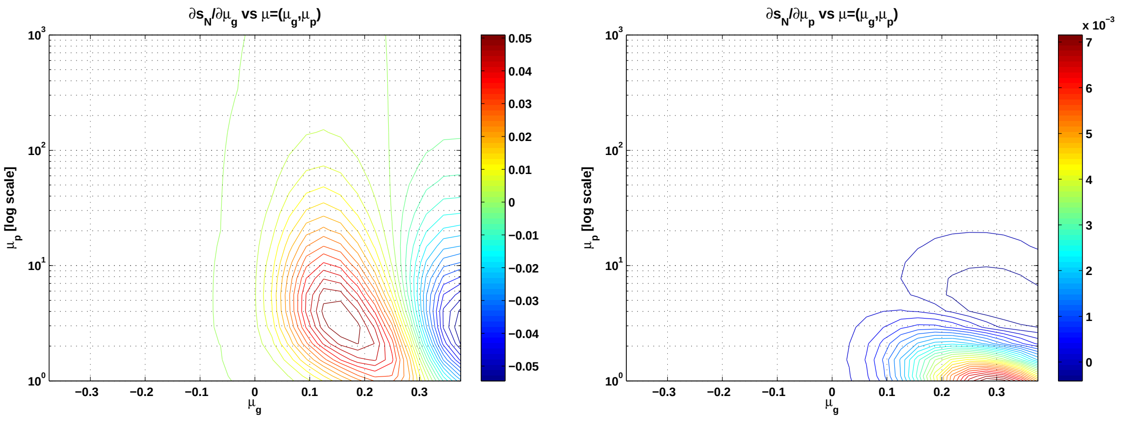

Figure 6.19: PB2. Contourlines of $(\partial s_N/\partial \mu_g)(\boldsymbol{\mu})$ (left) and $(\partial s_N/\partial \mu_p)(\boldsymbol{\mu})$ (right) vs $\boldsymbol{\mu} = (\mu_g, \mu_p) \in \mathcal{D}$; axis μ_p in logarithmic scale.

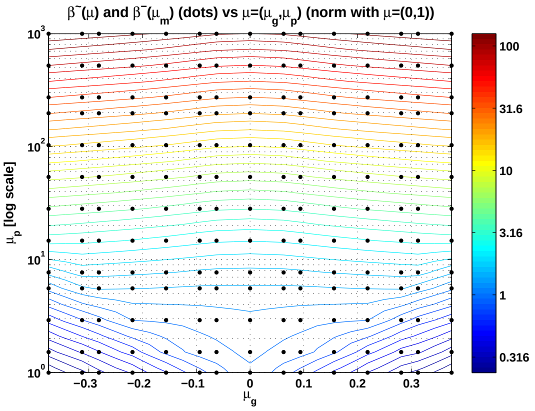

Figure 6.20: PB2. Contourlines of $\widetilde{\beta}(\boldsymbol{\mu})$ vs $\boldsymbol{\mu} = (\mu_g, \mu_p) \in \mathcal{D}$ and the pairs $\{(\boldsymbol{\mu}_m, \overline{\beta}(\boldsymbol{\mu}_m))\}_{m=1}^{195}$ (dots); axis μ_p in logarithmic scale.

tive algorithm for the samples selection (see Sec.6.3.4). In order to use the estimate (6.107), we need to evaluate the inf–sup constant; with this aim, we adopt the procedure outlined in Sec.6.3.4 obtaining the parametrized constant $\widetilde{\beta}(\boldsymbol{\mu})$ (6.106), with $K = 195$. In Fig.6.20 we report the contourlines of $\widetilde{\beta}(\boldsymbol{\mu})$ vs $\boldsymbol{\mu} \in \mathcal{D}$; we notice that, according with the choice made for the norm $|||\cdot|||$ with $\overline{\boldsymbol{\mu}} = (0, 1)$, $\widetilde{\beta}(\boldsymbol{\mu}) = 1$.

The a posteriori RB error estimate (6.107) is used in the adaptive algorithm in order to define the samples sets $\mathcal{S}_N^{pr}$ and $\mathcal{S}_N^{du}$. The results are reported in Table 6.2 by using the matrix E_N defined at point 10 of Sec.6.3.4; as for Problem 1, a comparison between the "primal–dual" and the "only primal" RB approaches, is reported. The same notation of Table 6.1 (see Sec.6.5.1) is used for Table 6.2. We recall that the "primal–dual" RB approach in the case $N^{pr} > 0$, $N^{du} = 0$, is equivalent to the "only primal" one. Once time, we observe that the computational cost associated with the "only primal" RB approach is always greater than that of the "primal–dual" one, i.e. $\Lambda_N^{pr} \geq \Lambda_N^{pr,du}$ $\forall E_{lev}$. E.g., if we fix $E_{lev} = 10^{-5}$, we see from Table 6.2 that

	"primal–dual" RB			"only primal" RB		
E_{lev}	N^{pr}	N^{du}	$\Lambda_N^{pr,du}$	N^{pr}	Λ_N^{pr}	$\Lambda_N^{pr}/\Lambda_N^{pr,du}$
$1.000e+01$	1	0	$1.700e+01$	1	$1.700e+01$	1.000
$5.623e+00$	1	0	$1.700e+01$	1	$1.700e+01$	1.000
$3.162e+00$	1	0	$1.700e+01$	1	$1.700e+01$	1.000
$1.778e+00$	1	0	$1.700e+01$	1	$1.700e+01$	1.000
$1.000e+00$	1	0	$1.700e+01$	1	$1.700e+01$	1.000
$5.623e-01$	1	0	$1.700e+01$	1	$1.700e+01$	1.000
$3.162e-01$	2	2	$2.080e+02$	4	$3.200e+02$	1.538
$1.778e-01$	4	0	$3.200e+02$	4	$3.200e+02$	1.000
$1.000e-01$	1	5	$6.220e+02$	11	$3.267e+03$	5.252
$5.623e-02$	1	6	$9.050e+02$	15	$6.975e+03$	7.707
$3.162e-02$	7	5	$2.212e+03$	22	$1.839e+04$	8.315
$1.778e-02$	7	7	$3.038e+03$	25	$2.562e+04$	8.435
$1.000e-02$	5	11	$4.672e+03$	33	$5.336e+04$	11.421
$5.623e-03$	7	12	$6.503e+03$	39	$8.366e+04$	12.864
$3.162e-03$	5	17	$1.142e+04$	46	$1.312e+05$	11.486
$1.778e-03$	10	17	$1.486e+04$	52	$1.839e+05$	12.376
$1.000e-03$	14	17	$1.922e+04$	55	$2.148e+05$	11.172
$5.623e-04$	14	20	$2.476e+04$	68	$3.884e+05$	15.687
$3.162e-04$	20	21	$3.744e+04$	72	$4.562e+05$	12.186
$1.778e-04$	14	29	$5.022e+04$	79	$5.929e+05$	11.806
$1.000e-04$	21	28	$6.022e+04$	86	$7.544e+05$	12.527
$5.623e-05$	20	33	$7.832e+04$	92	$9.141e+05$	11.671
$3.162e-05$	26	33	$9.548e+04$	94	$9.720e+05$	10.180
$1.778e-05$	29	34	$1.114e+05$	107	$1.408e+06$	12.639
$1.000e-05$	27	41	$1.449e+05$	113	$1.647e+06$	11.370
$5.623e-06$	26	45	$1.706e+05$	120	$1.958e+06$	11.477
$3.162e-06$	29	47	$1.988e+05$			
$1.778e-06$	30	50	$2.304e+05$			
$1.000e-06$	39	50	$2.799e+05$			
$5.623e-07$	49	47	$3.321e+05$			
$3.162e-07$	50	50	$3.700e+05$			
$1.778e-07$	44	61	$4.456e+05$			
$1.000e-07$	50	61	$5.003e+05$			
$5.623e-08$	47	67	$5.621e+05$			
$3.162e-08$	54	66	$6.183e+05$			
$1.778e-08$	60	66	$6.942e+05$			
$1.000e-08$	55	76	$8.130e+05$			
$5.623e-09$	50	83	$9.134e+05$			
$3.162e-09$	61	83	$1.050e+06$			
$1.778e-09$	57	89	$1.150e+06$			
$1.000e-09$	74	83	$1.273e+06$			
$5.623e-10$	73	87	$1.356e+06$			
$3.162e-10$	79	87	$1.482e+06$			
$1.778e-10$	77	95	$1.670e+06$			
$1.000e-10$	78	98	$1.789e+06$			
$5.623e-11$	77	102	$1.905e+06$			
$3.162e-11$	85	104	$2.169e+06$			
$1.778e-11$	93	102	$2.322e+06$			
$1.000e-11$	94	105	$2.464e+06$			
$5.623e-12$	107	100	$2.739e+06$			
$3.162e-12$	110	105	$3.043e+06$			
$1.778e-12$	116	107	$3.383e+06$			
$1.000e-12$	126	101	$3.652e+06$			

Table 6.2: PB2. Adaptive algorithm: results of samples selection procedure and comparison between the 'primal–dual" and the "only primal" RB approach.

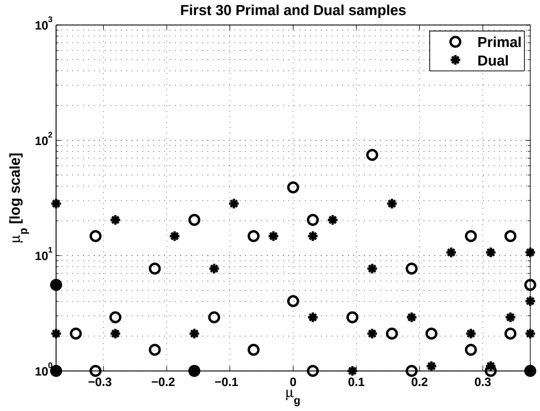

Figure 6.21: PB2. First 30 samples of the sets $\mathcal{S}_N^{pr}$ (circle) and $\mathcal{S}_N^{du}$ (starred) selected by the adaptive algorithm of Sec.6.3.4; axis μ_p in logarithmic scale.

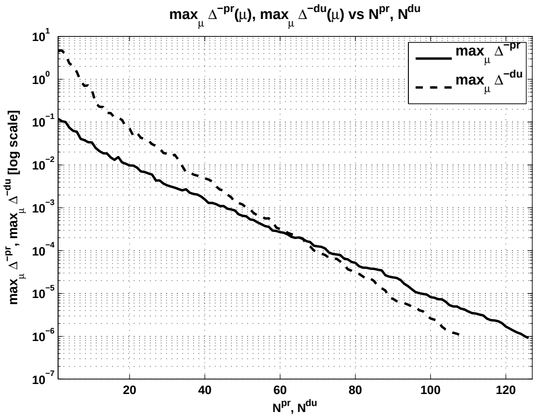

Figure 6.22: PB2. RB error indicators $\max_{\mu\in\mathcal{D}} \overline{\Delta}_N^{pr}(\mu)$ (continuous) and $\max_{\mu\in\mathcal{D}} \overline{\Delta}_N^{du}(\mu)$ (dashed) vs N^{pr} and N^{du}; error axis in logarithmic scale.

the "primal–dual" RB approach selects $N^{pr} = 27$ and $N^{du} = 41$ w.r.t. $N^{pr} = 113$ chosen by the "only primal" one; morevover, the saving of online computational cost is considerable, being of about 11 times inferior. Even larger savings are allowed for the problem under consideration: for example, if $E_{lev} = 5.623 \cdot 10^{-4}$ we have $\Lambda_N^{pr}/\Lambda_N^{pr,du} = 15.687$. In Fig.6.21 we report the samples selected for the definition of the sets $\mathcal{S}_N^{pr}$ and $\mathcal{S}_N^{du}$; in oder to make simple the visualization, we report only the first 30 samples for the primal and dual problem. We notice that the part of $\mathcal{D}$ for which $\mu_p \overset{\sim}{>} \mu_\star = 10^2$ is not so "crucial" both for the primal and dual problem; on the contrary, primal and dual samples are set $\forall \mu_g \in [\mu_g^{min}, \mu_g^{max}]$ even if $s_N(\boldsymbol{\mu}) \simeq 0$, $(\partial s_N/\partial \mu_g)(\boldsymbol{\mu}) \simeq 0$ and $(\partial s_N/\partial \mu_p)(\boldsymbol{\mu}) \simeq 0$ for $\mu_g < 0$ (see Fig.s 6.18 and 6.19). Fig.6.22 shows the behavior of the RB error indicators $\max_{\mu\in\mathcal{D}} \overline{\Delta}_N^{pr}(\mu)$ and $\max_{\mu\in\mathcal{D}} \overline{\Delta}_N^{du}(\mu)$ vs N^{pr} and N^{du}, respectively (see Eq.(6.108)). We observe that the convergence rate of the dual RB error is greater than that of the primal RB error, thus indicating a minor "complexity"

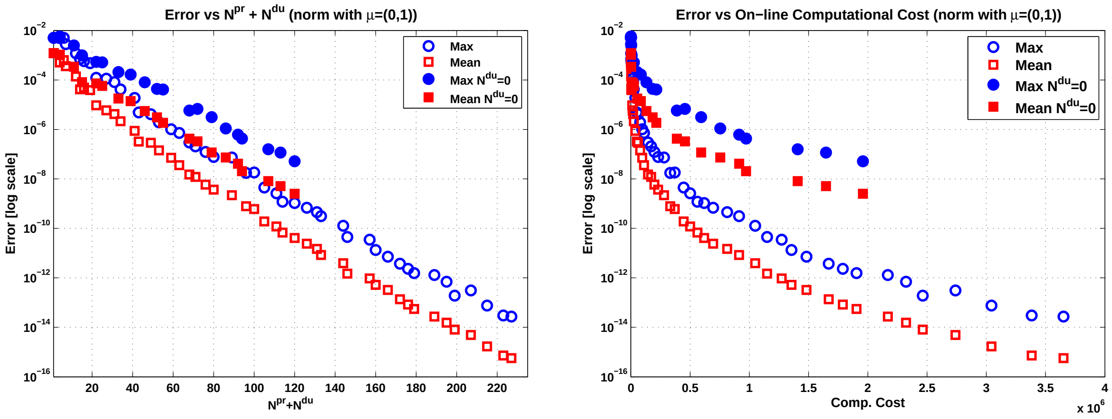

Figure 6.23: PB2. E_N^{max} (empty circle) and E_N^{mean} (empty square) for the "primal–dual" RB approach vs $N^{pr} + N^{du}$ (left) and online computational costs $\Lambda_N^{pr,du}$, Λ_N^{pr} (right); full circles and squares refer to the indicators E_N^{max}, E_N^{mean} for the "only primal" RB approach. Error axis in logarithmic scale.

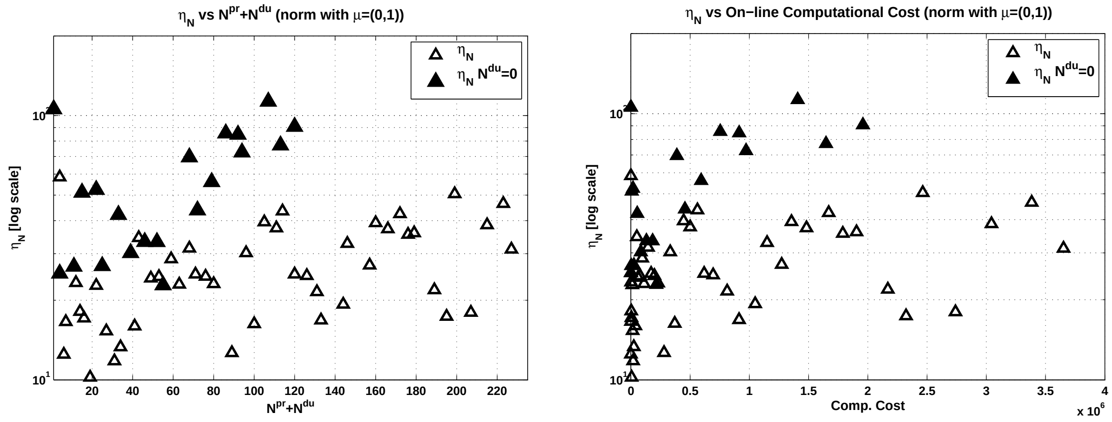

Figure 6.24: PB2. $\overline{\eta}_N$ (empty triangle) for the "primal–dual" RB approach vs $N^{pr} + N^{du}$ (left) and online computational costs $\Lambda_N^{pr,du}$, Λ_N^{pr} (right); full triangles refer to the indicator $\overline{\eta}_N$ for the "only primal" RB approach. Effectivity index axis in logarithmic scale.

of the dual problem w.r.t. the primal one. However, as previuosly mentioned, even if the continuous dual solution mapped on the real domain Ω_0 is not dependent on the geometrical parameter μ_g, this is not true in the reference domain Ω. For this reason, the convergence rate of the indicator $\max_{\mu\in\mathcal{D}} \overline{\Delta}_N^{du}(\mu)$ vs N^{du} is not only related to the physical parameter μ_p, but it takes also into account for the dependence on μ_g. We could say that the dual RB problem is more "complicated" than that waited simply by inspection of the parametrized dual problem on the real domain Ω_0.
In similar way as made for Problem 1, we compare the true RB errors for the output obtained on the basis of the adaptive algorithm for the "primal–dual" and the "only primal" RB approaches. In Fig.6.23(left) we report the error indicators E_N^{max} and E_N^{mean} (see Eq.(6.109))

vs $N^{pr}+N^{du}$ for both the "primal–dual" and "only primal" RB approaches; in Fig.6.23(right) the same RB error indicators vs the corresponding online computational costs $\Lambda_N^{pr,du}$ and Λ_N^{pr} are shown. As usual, the best indications arise from the plot of Fig.6.23(right), which clearly shows that the "primal–dual" RB allows great savings of online computational costs for any given error tolerance on the output; for example, if we fix the tolerance error at 10^{-8} we see that $\Lambda_N^{pr,du}$ is about 8 times inferior than Λ_N^{pr}. This confirms the validity of the adaptive algorithm and the indications given in Table 6.2. Finally, we report in Fig.6.24(left) and (right) the effectivity index indicator $\overline{\eta}_N$ (see Eq.(6.110)) vs $N^{pr}+N^{du}$ and $\Lambda_N^{pr,du}$, Λ_N^{pr}. We observe that $\overline{\eta}_N$ is greater than 10 and that the effectivity indicator for the "primal–dual" RB approach tends to be inferior to that of the "only primal" far large values of $N^{pr}+N^{du}$, or $\Lambda^{pr,du}$ and Λ_N^{pr}.

6.6 Concluding remarks

In this Chapter we have considered the RB method for the approximation of parametrized advection–reaction PDEs. We have generated the basis by means of the FE method applied to the stabilized version of the weak problem. For the RB method, in the "primal–dual" formulation, we have provided both a priori and a posteriori RB error estimates. We have shown that the RB "complexity" increases as the FE mesh size reduces, thus requiring a larger number of basis in order to satisfy a prescribed tolerance on the error. An adaptive algorithm for the selection of the sample sets, based on a criterium of minimization of the online computational costs, has been proposed. We have proved, by means of numerical tests, the effectiveness of this algorithm and the savings of computational costs allowed by the "primal–dual" RB approach w.r.t. those of the "only primal" one. The numerical tests also show that if the FE approximation is stable, then so it is the RB one.

Chapter 7

Reduced Basis Method for Parametrized Optimal Control Problems

Many design problems in fluid dynamics, structural mechanics, thermal analysis and environmental sciences, can be afforded by means of optimal control or shape optimization techniques applied to the system of PDEs which describe the problem under investigation; see Chapter 2. Possible examples include shape optimization of bypass anastomosis in blood vessels, airfoils in Aerodynamics and regulation of thermal sources for heat exchange. In the environmental field, an application consists in controlling the emissions of atmospheric pollutants in order to preserve air quality, as highlighted in Sec.1.3. Since the characterization of a system depends, in general, on a set of parameters, its response will be parameter dependent as well as, in optimization contexts, the optimal shape or control. Several engineering instances can be provided in the field of *parametrized optimal control problems.* For environmental applications, this could be the case of the regulation of the pollutant emissions from an industrial chimney for which the "optimal" emission rate depends on the meteorological conditions, the latter to be regarded as parameters.

The numerical solution of optimal control problems can be very expensive in terms of computational costs, as e.g. in the case of large scale environmental problems. This could make unaffordable optimization techniques for large classes of practical problems, especially in the many–query and real time contexts for parametrized time dependent PDEs. Hence, it arises the necessity to reduce the complexity of the original problem, while preserving its main features and without loosing the accuracy on the reduced solution.

The idea of using *model order reduction* techniques for the numerical solution of optimal control and optimization problems has been already faced in [69, 78, 80, 91, 92, 93] even if without considering the multi–parameter case. In this context, a possible approach for model order reduction is represented by the Proper Orthogonal Decomposition (POD) method; see [103] and e.g. [9, 102, 155] for applications to optimal control problems. For parameter independent contexts, an other possible approach consists in coupling the Finite Element (FE) approximation with mesh adaption strategies led by a posteriori error estimates for optimal control problems; see [18, 19, 20, 22, 44, 151].

As already pointed out in Chapter 6, the Reduced Basis (RB) method represents a rapid and reliable approach for the solution of parametrized PDEs and the evaluation of the associated outputs of interest, thus allowing to contain the computational costs in the many–query and real time frameworks. The main features of the RB method can be reasonably extended to the solution of parametrized optimal control problems, as emphasized e.g. in [71] for the design and optimization of engineering systems and devices.

In [70, 72] the RB method is presented for the solution of parametrized parabolic PDEs and, in the case of linear time–invariant systems, extended to the case of optimal control problems by means of the "impulse response" approach. Parameter identification and inverse scattering problems have been afforded in [127] by means of the RB method. In [149, 151] the RB method is proposed for the approximation of the parametrized PDEs (first order necessary conditions) describing a steady optimal control problem in order to speed up the iterative optimization process.

In this Chapter we deal with the RB method for parametrized optimal control problems described by linear time–invariant parabolic PDEs. In particular, we define and use the RB method for reducing the order of the original optimal control problem and not just for speeding up the optimization process as done in [149, 151]. This approach allows us to inherit the features of the goal oriented analysis for the RB approximation, to treat coherently the RB approximated primal and dual PDEs and, finally, to derive an "ad hoc" a posteriori estimate for the RB error on the optimal solution. Such an estimate, developed for quadratic costs functionals, is used to evaluate the accuracy of the RB approximation and to lead an adaptive procedure for the choice of the samples set, on which the RB basis is defined. This basis, obtained by means of an RB "integrated" approach, is set according with the solution of the optimal control problem and not just on the usual solution of the primal and dual equations (see [70, 72]). The computational procedure is based on the offline–online decomposition, thus allowing at the online step a rapid and reliable solution of the optimal control problem for any given parameter. Both the cases of unconstrained and constrained optimal control problems are afforded.

In Sec.7.1 we report the RB method for the solution of linear time–invariant parabolic PDEs, the a posteriori RB error estimate for output functionals and the adaptive procedure for the choice of the samples set and RB basis. In Sec.7.2 we provide the formulation of the RB approach for the solution of parametrized optimal control problems, both unconstrained and constrained. In particular, we put into evidence the offline–online decomposition for the solution of parametrized optimal control problems by means of the FE and RB methods. Then, after having recalled the role of the "predictability" and "optimality" errors, we provide an a posteriori error estimate for the "optimality" error. Finally, our "integrated" RB approach is reported, as well as the adaptive procedure used for the choice of the RB basis and space. The effectivity of the procedure is highlighted by numerical tests. Concluding remarks follow.

7.1 Parametrized linear parabolic PDEs

In this Section we report the RB method for the solution of parametrized parabolic PDEs, in particular for linear–time invariant equations (see Sec.7.1.1 for this definition). After having introduced the problem in Sec.7.1.1 and the RB method in Sec.7.1.2, we report in Sec.7.1.3 the a posteriori RB error estimate for the evaluation of linear output functionals and the adaptive procedure for the choice of the basis. Finally, in Sec.7.1.4 a numerical test, referring to an environmental problem, is discussed.

Throughout this Section we refer principally to [128]; for a further deepening we refer the reader to [70, 72] and also to [89].

7.1.1 Problem definition and FE approximation

By recalling the unsteady problem introduced in Sec.4.1.3, we consider the following parametrized linear parabolic PDE in weak form:

$$\begin{aligned}\text{find } v(\boldsymbol{\mu}) = v(t,\boldsymbol{\mu}) \in \mathcal{V} \quad : \quad & m\left(\frac{\partial v(t,\boldsymbol{\mu})}{\partial t}, \phi; \boldsymbol{\mu}\right) + a(v(t,\boldsymbol{\mu}), \phi; \boldsymbol{\mu}) = b(\phi;\boldsymbol{\mu}) g(t) \\ & \qquad\qquad\qquad \forall \phi \in \mathcal{Z},\ t \in (0,T), \qquad (7.1) \\ \text{with } v(0;\boldsymbol{\mu}) = v_0(\boldsymbol{\mu}), &\end{aligned}$$

where the parameters vector $\boldsymbol{\mu} \in \mathcal{D}$, being $\mathcal{D}$ the parameters set (see Sec.6.1.1). The spaces $\mathcal{Z}$ and $\mathcal{V}$ are appropriate functional spaces s.t. $H_0^1(\Omega) \subset \mathcal{Z} \subset H^1(\Omega)$ and $\mathcal{V} = \{w \in L^2(0,T;\mathcal{Z}) : \frac{\partial w}{\partial t} \in L^2(0,T;\mathcal{Z}^*)\}$; moreover, we define a functional space $\mathcal{Y}$ s.t. $\mathcal{Y} \subset L^2(\Omega)$. We assume the forms $a(\cdot,\cdot;\boldsymbol{\mu})$ and $m(\cdot,\cdot;\boldsymbol{\mu})$ as bilinear, continuous and coercive ($m(\cdot,\cdot;\boldsymbol{\mu})$ positive) in the norms induced by the spaces $\mathcal{Z}$ and $\mathcal{Y}$, respectively; $b(\cdot;\boldsymbol{\mu})$ denotes a linear bounded functional, while $g(t) \in \mathcal{G}$ is the input function. The space $\mathcal{G}$ is chosen s.t. $\int_0^T b(w(t);\boldsymbol{\mu})g(t)\ dt$ is bounded $\forall w(t) \in \mathcal{V}$. For example, if $b(\cdot;\boldsymbol{\mu}) \in L^2(0,T;H^{-1}(\Omega))$, we require $\mathcal{G} = L^2(0,T)$; however, if $b(\cdot;\boldsymbol{\mu}) \in C^0([0,T];L^2(\Omega))$, we can assume $\mathcal{G} = \left(C^0([0,T])\right)^*$, thus allowing to handle with an input function $g(t)$ in the form of a Dirac delta distribution.
Let us introduce, similarly as done in Sec.6.1.1, an output functional $L(\cdot;\boldsymbol{\mu})$ s.t.:

$$s(\boldsymbol{\mu}) = L(v(\boldsymbol{\mu});\boldsymbol{\mu}) = \int_0^T l(v(t,\boldsymbol{\mu});\boldsymbol{\mu})\ dt, \qquad (7.2)$$

being $l(\cdot;\boldsymbol{\mu}) \in L^1(0,T;H^{-1}(\Omega))$ a linear and bounded functional (see also Eq.(4.14)). We observe that, even if we focus on output functionals in the form (7.2), it is also possible to consider output functionals as $s(\overline{t},\boldsymbol{\mu}) = l(v(\overline{t},\boldsymbol{\mu});\boldsymbol{\mu})$, for $\overline{t} \in (0,T)$; see [128].
On this basis, it is possible to afford the problem by means of the goal–oriented analysis for output functionals already discussed in Sec.4.1. In particular, without introducing explicitly the Lagrangian functional formalism, the dual equation in weak form reads as in Eq.(4.17), where the parametric dependence has been introduced.
Similarly to Sec.6.2.1, we assume that $a(\cdot,\cdot;\boldsymbol{\mu})$, $m(\cdot,\cdot;\boldsymbol{\mu})$, $b(\cdot;\boldsymbol{\mu})$, $l(\cdot;\boldsymbol{\mu})$ and $v_0(\boldsymbol{\mu})$ are affinely dependent on the parameters vector $\boldsymbol{\mu}$. This allows to decompose the solution procedure of both the FE and RB problems into an offline and an online step as discussed in Sec.6.3.4. In particular, for some integers $Q_{a,m,b,l,v}$, $a(w,\phi;\boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \vartheta_q^a(\boldsymbol{\mu}) a_q(w,\phi)$, $m(w,\phi;\boldsymbol{\mu}) = \sum_{q=1}^{Q_m} \vartheta_q^m(\boldsymbol{\mu}) m_q(w,\phi)$, $b(w;\boldsymbol{\mu}) = \sum_{q=1}^{Q_b} \vartheta_q^b(\boldsymbol{\mu}) b_q(w)$, $l(w;\boldsymbol{\mu}) = \sum_{q=1}^{Q_l} \vartheta_q^l(\boldsymbol{\mu}) l_q(w)$ and $v_0(\boldsymbol{\mu}) = \sum_{q=1}^{Q_v} \vartheta_q^v(\boldsymbol{\mu}) v_{0q}$ $\forall w,\phi \in \mathcal{Z}$, $\forall \boldsymbol{\mu} \in \mathcal{D}$; the functions $\vartheta^q_{a,m,b,l,v}(\boldsymbol{\mu}) : \mathcal{D} \to \mathbb{R}$ depend only on the parameter vector $\boldsymbol{\mu}$, while the forms and functionals $a^q(\cdot,\cdot)$, $m^q(\cdot,\cdot)$, $b^q(\cdot)$, $l^q(\cdot)$ and v^q are parameter independent.
Finally, we assume from now that all the bilinear forms and linear functionals are time independent, for which the problem is *linear time–invariant* (LTI).

As already highlighted in Chapter 6, the RB method passes through the approximation of the continuous parametrized problem. In particular, the FE method is used for the spatial

approximation (yielding what is typically called the "truth" approximation), while the Euler–Backward is used for the temporal one; see [153]. Other approximation methods are possible; for example a different temporal scheme could be considered [70].
With this aim, let us subdivide $[0,T]$ into K subintervals of equal length $\Delta t = T/K$ and define $t_k := k\Delta T$, with $k = 0, 1, \ldots, K$, and $\mathbb{K} := \{1, 2, \ldots, K\}$; moreover, let us introduce the FE approximation and the corresponding spaces $\mathcal{Z}_h \subset \mathcal{Z}$ and $\mathcal{Y}_h \subset \mathcal{Y} \subset L^2(\Omega)$. Typically, the dimension N_h of the space $\mathcal{Z}_h$ is very "large". The FE solution $v_h^k(\boldsymbol{\mu})$ at the time step t_k of the FE approximated primal equation (7.1) reads:

$$
\begin{aligned}
&\text{find } v_h^k(\boldsymbol{\mu}) \in \mathcal{Z}_h \quad : \\
&\qquad m\left(v_h^k(\boldsymbol{\mu}), \phi_h; \boldsymbol{\mu}\right) + \Delta t \; a\left(v_h^k(\boldsymbol{\mu}), \phi_h; \boldsymbol{\mu}\right) = \\
&\qquad\qquad m\left(v_h^{k-1}(\boldsymbol{\mu}), \phi_h; \boldsymbol{\mu}\right) + \Delta t \; b(\phi_h; \boldsymbol{\mu}) g^k \qquad \forall \phi_h \in \mathcal{Z}_h, \forall k \in \mathbb{K}, \\
&\text{with } v_h^0(\boldsymbol{\mu}) = v_{0h}(\boldsymbol{\mu}),
\end{aligned}
\tag{7.3}
$$

being $v_{0h}(\boldsymbol{\mu})$ the $\mathcal{Y}_h$ projection of $v_0(\boldsymbol{\mu})$ and $g^k = g(t^k)$. The FE approximated output functional (7.2) is then computed as:

$$
s_h(\boldsymbol{\mu}) = \Delta t \left(\frac{1}{2} l\left(v_h^0(\boldsymbol{\mu}); \boldsymbol{\mu}\right) + \sum_{j=1}^{K-1} l\left(v_h^j(\boldsymbol{\mu}); \boldsymbol{\mu}\right) + \frac{1}{2} l\left(v_h^K(\boldsymbol{\mu}); \boldsymbol{\mu}\right) \right) \tag{7.4}
$$

and, from the goal–oriented analysis, the associated FE dual problem reads [128]:

$$
\begin{aligned}
&\text{find } z_h^k(\boldsymbol{\mu}) \in \mathcal{Z}_h \quad : \\
&\qquad m\left(\phi_h, z_h^k(\boldsymbol{\mu}); \boldsymbol{\mu}\right) + \Delta t \; a\left(\phi_h, z_h^k(\boldsymbol{\mu}); \boldsymbol{\mu}\right) = \\
&\qquad\qquad m\left(\phi_h, z_h^{k+1}(\boldsymbol{\mu}); \boldsymbol{\mu}\right) + \Delta t \; l(\phi_h; \boldsymbol{\mu}) \qquad \forall \phi_h \in \mathcal{Z}_h, \; \forall k \in \{K-1, \ldots, 1\}, \\
&\text{with } m\left(\phi_h, z_h^K(\boldsymbol{\mu}); \boldsymbol{\mu}\right) + \Delta t \; a\left(\phi_h, z_h^K(\boldsymbol{\mu}); \boldsymbol{\mu}\right) = \frac{1}{2} \Delta t \; l(\phi_h; \boldsymbol{\mu}).
\end{aligned}
\tag{7.5}
$$

Similarly to Sec.4.1.3, we define, from Eq.s (7.3) and (7.5), the following primal and dual residuals, respectively:

$$
\begin{aligned}
R^{pr,k}\left(w_h^k, w_h^{k-1}\right)(\phi_h; \boldsymbol{\mu}) := {} & b(\phi_h, \boldsymbol{\mu}) g^k - a\left(w_h^k, \phi_h; \boldsymbol{\mu}\right) \\
& - \frac{1}{\Delta t} m\left(w_h^k - w_h^{k-1}, \phi_h; \boldsymbol{\mu}\right) \qquad \forall \phi_h \in \mathcal{Z}_h, \; \forall k \in \mathbb{K},
\end{aligned}
\tag{7.6}
$$

for some $w_h^k \in \mathcal{Z}_h \; \forall k \in \{0, 1, \ldots, K\}$ and:

$$
R^{du,k}\left(w_h^k, w_h^{k+1}\right)(\phi_h; \boldsymbol{\mu}) := \begin{cases}
l(\phi_h; \boldsymbol{\mu}) - a\left(\phi, w_h^k; \boldsymbol{\mu}\right) - \dfrac{1}{\Delta t} m\left(\phi_h, w_h^k - w_h^{k+1}; \boldsymbol{\mu}\right) \\
\qquad\qquad \forall \phi_h \in \mathcal{Z}, \; \forall k \in \{K-1, \ldots, 1\}, \\
\dfrac{1}{2} l(\phi_h, \boldsymbol{\mu}) - a\left(\phi_h, w_h^K; \boldsymbol{\mu}\right) - \dfrac{1}{\Delta t} m\left(\phi_h, w_h^K; \boldsymbol{\mu}\right) \\
\qquad\qquad \forall \phi_h \in \mathcal{Z}_h, \; k = K,
\end{cases}
\tag{7.7}
$$

for some $w_h^k \in \mathcal{Z}_h \ \forall k \in \{K, \dots, 1\}$. In order to make simpler the notation, we define the FE primal and dual residuals as, respectively:

$$R_h^{pr,k}(\phi_h; \boldsymbol{\mu}) := R^{pr,k}\left(v_h^k(\boldsymbol{\mu}), v_h^{k-1}(\boldsymbol{\mu})\right)(\phi_h; \boldsymbol{\mu}) \qquad \forall \phi_h \in \mathcal{Z}_h,\ \forall k \in \mathbb{K}, \tag{7.8}$$

$$R_h^{du,k}(\phi_h; \boldsymbol{\mu}) := R^{du,k}\left(z_h^k(\boldsymbol{\mu}), z_h^{k+1}(\boldsymbol{\mu})\right)(\phi_h; \boldsymbol{\mu}) \qquad \forall \phi_h \in \mathcal{Z}_h,\ \forall k \in \{K, \dots, 1\}. \tag{7.9}$$

Then, in analogy with Eq.s (4.22) and (4.23), Eq.s (7.3) and (7.5) read, respectively:

$$\begin{aligned} &\text{find } v_h^k(\boldsymbol{\mu}) \in \mathcal{Z}_h \quad : \quad \Delta t \ R_h^{pr,k}(\phi_h; \boldsymbol{\mu}) = 0 \qquad \forall \phi_h \in \mathcal{Z}_h,\ \forall k \in \mathbb{K} \\ &\text{with } v_h^0(\boldsymbol{\mu}) = v_{0h}(\boldsymbol{\mu}), \end{aligned} \tag{7.10}$$

$$\text{find } z_h^k(\boldsymbol{\mu}) \in \mathcal{Z}_h \quad : \quad \Delta t \ R_h^{du,k}(\phi_h; \boldsymbol{\mu}) = 0 \qquad \forall \phi_h \in \mathcal{Z}_h,\ \forall k \in \{K, \dots, 1\}, \tag{7.11}$$

As done in Sec.6.2.1, the FE approximated primal and dual problems are then rewritten in matricial formulation. In particular, by writing $v_h^k(\boldsymbol{\mu}) = \sum_{j=1}^{N_h} v_{hj}^k(\boldsymbol{\mu})\varphi_j$, being φ_j the FE lagrangian basis function associated to the FE space $\mathcal{Z}_h$, Eq.(7.3) reads:

$$\begin{aligned} &\text{find } \mathbf{v}_h^k(\boldsymbol{\mu}) \in \mathbb{R}^{N_h} \quad : \\ &\quad \left(M_h(\boldsymbol{\mu}) + \Delta t \ A_h(\boldsymbol{\mu})\right) \mathbf{v}_h^k(\boldsymbol{\mu}) = \\ &\qquad\qquad M_h(\boldsymbol{\mu}) \mathbf{v}_h^{k-1}(\boldsymbol{\mu}) + \Delta t \ \mathbf{B}_h(\boldsymbol{\mu}) g^k \qquad \forall k \in \{2, \dots, K\}, \\ &\text{with } \left(M_h(\boldsymbol{\mu}) + \Delta t \ A_h(\boldsymbol{\mu})\right) \mathbf{v}_h^1(\boldsymbol{\mu}) = \mathbf{v}_h^0(\boldsymbol{\mu}) + \Delta t \ \mathbf{B}_h(\boldsymbol{\mu}) g^1, \end{aligned} \tag{7.12}$$

with $\mathbf{v}_h^k(\boldsymbol{\mu}) := [v_{h1}^k(\boldsymbol{\mu}), \dots, v_{hN_h}^k(\boldsymbol{\mu})]^T$ and $\mathbf{v}_h^0(\boldsymbol{\mu}) = \mathbf{v}_{0h}(\boldsymbol{\mu})$. Similarly, by writing $z_h^k(\boldsymbol{\mu}) = \sum_{j=1}^{N_h} z_{hj}^k(\boldsymbol{\mu})\varphi_j$, Eq.(7.5) corresponds to:

$$\begin{aligned} &\text{find } \mathbf{z}_h^k(\boldsymbol{\mu}) \in \mathbb{R}^{N_h} \quad : \\ &\quad \left(M_h(\boldsymbol{\mu}) + \Delta t \ A_h^T(\boldsymbol{\mu})\right) \mathbf{z}_h^k(\boldsymbol{\mu}) = \\ &\qquad\qquad M_h(\boldsymbol{\mu}) \mathbf{z}_h^{k+1}(\boldsymbol{\mu}) + \Delta t \ \mathbf{L}_h(\boldsymbol{\mu}) \qquad \forall k \in \{K-1, \dots, 1\}, \\ &\text{with } \left(M_h(\boldsymbol{\mu}) + \Delta t \ A_h^T(\boldsymbol{\mu})\right) \mathbf{z}_h^K(\boldsymbol{\mu}) = \tfrac{1}{2} \Delta t \ \mathbf{L}_h(\boldsymbol{\mu}). \end{aligned} \tag{7.13}$$

Finally, the FE approximated output functional (7.4) reads:

$$s_h(\boldsymbol{\mu}) = \Delta t \ \mathbf{L}_h(\boldsymbol{\mu})^T \left(\frac{1}{2} \mathbf{v}_{0h}(\boldsymbol{\mu}) + \sum_{j=1}^{K-1} \mathbf{v}_h^j(\boldsymbol{\mu}) + \frac{1}{2} \mathbf{v}_h^K(\boldsymbol{\mu}) \right). \tag{7.14}$$

The FE matrices and vectors defined in Eq.s (7.12)–(7.14) read as $A_h(\boldsymbol{\mu}) = \sum_{q=1}^{Q_a} \vartheta_q^a(\boldsymbol{\mu}) A_{qh}$, $M(\boldsymbol{\mu}) = \sum_{q=1}^{Q_m} \vartheta_q^m(\boldsymbol{\mu}) M_{qh}$, $\mathbf{B}_h(\boldsymbol{\mu}) = \sum_{q=1}^{Q_b} \vartheta_q^b(\boldsymbol{\mu}) \mathbf{B}_{qh}$, $\mathbf{L}_h(\boldsymbol{\mu}) = \sum_{q=1}^{Q_l} \vartheta_q^l(\boldsymbol{\mu}) \mathbf{L}_{qh}$ and $\mathbf{v}_{0h}(\boldsymbol{\mu}) = \sum_{q=1}^{Q_m} \sum_{r=1}^{Q_v} \vartheta_q^m(\boldsymbol{\mu}) \vartheta_r^v(\boldsymbol{\mu}) \mathbf{v}_{0qrh}$, according with the affine decomposition assumptions, being $(A_{qh})_{i,j} = a_q(\varphi_j, \varphi_i)$, $(M_{qh})_{i,j} = m_q(\varphi_j, \varphi_i)$, $(\mathbf{B}_{qh})_i = b_q(\varphi_i)$, $(\mathbf{L}_{qh})_i = l_q(\varphi_i)$ and $(\mathbf{v}_{0qrh})_i = m_q(\varphi_i, v_{0r})$ for a generic q.

7.1.2 RB approximation

The Galerkin–RB approximation of the primal and dual equations (7.3) and (7.5) is obtained by defining the RB spaces $\mathcal{Z}_N^{pr}$, $\mathcal{Z}_N^{du} \subset \mathcal{Z}_h$, as already discussed in Sec.6.3.1. In particular, we have:

$$\begin{aligned} \mathcal{Z}_N^{pr} &= \text{span}\left\{\xi_i \quad i = 1, \dots, N^{pr}\right\}, \\ \mathcal{Z}_N^{du} &= \text{span}\left\{\zeta_i \quad i = 1, \dots, N^{du}\right\}, \end{aligned} \tag{7.15}$$

being the primal basis $\{\xi_i\}_{i=1}^{N_{pr}}$ and the dual basis $\{\zeta_i\}_{i=1}^{N_{du}}$ defined according with an adaptive procedure (see Sec.7.1.3). These bases depend on the FE primal solutions $v_h^k(\boldsymbol{\mu})$ and dual solutions $z_h^k(\boldsymbol{\mu})$ corresponding to selected parameters vectors $\boldsymbol{\mu} \in \mathcal{D}$. However, we observe that we are interested in using the RB method not only for evaluating the response of the system for a prescribed input function $g(t)$ and for any $\boldsymbol{\mu} \in \mathcal{D}$, but also for any $g(t) \in \mathcal{G}$. For this reason, the input function $g(t)$ (or g^k) is not given a priori; hence, we can not compute directly the corresponding FE primal solution $v_h^k(\boldsymbol{\mu})$.

To overcome this difficulty, the LTI hypothesis is used, thus leading to an *impulse response* approach [70, 72, 128]. In fact, the solution of LTI systems can be rewritten in terms of the convolution of impulse responses. Given an input function g^k $\forall k \in \mathbb{K}$, the solution of the FE primal equation (7.3) reads:

$$v_h^k(\boldsymbol{\mu}) = \sum_{j=1}^{k} v_h^{\delta,k-j+1}(\boldsymbol{\mu}) g^j \qquad \forall k \in \mathbb{K}, \tag{7.16}$$

where $v_h^{\delta,k}(\boldsymbol{\mu})$ is the so–called unit impulse response, corresponding to the solution of Eq.(7.3) for an impulse input function $g^k = \delta_{1k}$, being δ_{ij} the Kronecker function s.t. $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ for $i \neq j$. For this reason, it is required that the primal RB space be able to approximate sufficiently well only the parametrized impulse response and not necessarily the solutions corresponding to any input function g^k.

The RB primal solution $v_N^k(\boldsymbol{\mu}) \in \mathcal{Z}_N^{pr}$ is obtained by solving the following problem, derived from Eq.(7.3):

$$\begin{aligned} &\text{find } v_N^k(\boldsymbol{\mu}) \in \mathcal{Z}_N^{pr} \quad : \\ &\quad m\left(v_N^k(\boldsymbol{\mu}), \phi_N; \boldsymbol{\mu}\right) + \Delta t\ a\left(v_N^k(\boldsymbol{\mu}), \phi_N; \boldsymbol{\mu}\right) = \\ &\qquad\qquad m\left(v_N^{k-1}(\boldsymbol{\mu}), \phi_N; \boldsymbol{\mu}\right) + \Delta t\ b(\phi_N; \boldsymbol{\mu}) g^k \qquad \forall \phi_N \in \mathcal{Z}_N^{pr}, \forall k \in \mathbb{K}, \\ &\text{with } v_N^0(\boldsymbol{\mu}) = v_{0h}(\boldsymbol{\mu}). \end{aligned} \tag{7.17}$$

We deduce that, after having defined the RB primal error as $e_N^{pr,k}(\boldsymbol{\mu}) := v_h^k(\boldsymbol{\mu}) - v_N^k(\boldsymbol{\mu})$ $\forall k \in \{0, 1, \dots, K\}$, $e_N^{pr,0}(\boldsymbol{\mu}) = 0$ and, in general $v_N^0(\boldsymbol{\mu}) \notin \mathcal{Z}_N^{pr}$.

Similarly, from Eq.(7.5), the RB dual solution $z_N^k(\boldsymbol{\mu}) \in \mathcal{Z}_N^{du}$ is obtained by solving the following problem:

$$\begin{aligned}
&\text{find } z_N^k(\boldsymbol{\mu}) \in \mathcal{Z}_N^{du} \quad : \\
&\quad m\left(\phi_N, z_N^k(\boldsymbol{\mu}); \boldsymbol{\mu}\right) + \Delta t \ a\left(\phi_N, z_N^k(\boldsymbol{\mu}); \boldsymbol{\mu}\right) = \\
&\qquad\qquad m\left(\phi_N, z_N^{k+1}(\boldsymbol{\mu}); \boldsymbol{\mu}\right) + \Delta t \ l(\phi_N; \boldsymbol{\mu}) \qquad \forall \phi_N \in \mathcal{Z}_N^{du}, \ k \in \{K-1, \dots, 1\}, \\
&\text{with } m\left(\phi_N, z_N^K(\boldsymbol{\mu}); \boldsymbol{\mu}\right) + \Delta t \ a\left(\phi_N, z_N^K(\boldsymbol{\mu}); \boldsymbol{\mu}\right) = \frac{1}{2}\Delta t \ l(\phi_N; \boldsymbol{\mu}),
\end{aligned} \tag{7.18}$$

for which we define the RB dual error $e_N^{du,k}(\boldsymbol{\mu}) := z_h^k(\boldsymbol{\mu}) - z_N^k(\boldsymbol{\mu})$, $\forall k \in \{K, \dots, 1\}$. In analogy with Sec.s 4.1.3, 6.3.3 and 7.1.1, we define from Eq.s (7.6) and (7.7) the RB primal and dual residuals as, respectively:

$$R_N^{pr,k}(\phi_h; \boldsymbol{\mu}) := R^{pr,k}\left(v_N^k(\boldsymbol{\mu}), v_N^{k-1}(\boldsymbol{\mu})\right)(\phi_h; \boldsymbol{\mu}) \qquad \forall \phi_h \in \mathcal{Z}_h, \ \forall k \in \mathbb{K}, \tag{7.19}$$

$$R_N^{du,k}(\phi_h; \boldsymbol{\mu}) := R^{du,k}\left(z_N^k(\boldsymbol{\mu}), z_N^{k+1}(\boldsymbol{\mu})\right)(\phi_h; \boldsymbol{\mu}) \qquad \forall \phi_h \in \mathcal{Z}_h, \ \forall k \in \{K, \dots, 1\}, \tag{7.20}$$

similarly to what done in Eq.s (7.8) and (7.9) for the FE residuals. Then, in analogy with Eq.s (7.10) and (7.11), Eq.s (7.17) and (7.18) read, respectively:

$$\begin{aligned}
&\text{find } v_N^k(\boldsymbol{\mu}) \in \mathcal{Z}_N^{pr} \quad : \quad \Delta t \ R_N^{pr,k}(\phi_N; \boldsymbol{\mu}) = 0 \qquad \forall \phi_N \in \mathcal{Z}_N^{pr}, \ \forall k \in \mathbb{K} \\
&\text{with } v_N^0(\boldsymbol{\mu}) = v_{0h}(\boldsymbol{\mu}),
\end{aligned} \tag{7.21}$$

$$\text{find } z_N^k(\boldsymbol{\mu}) \in \mathcal{Z}_N^{du} \quad : \quad \Delta t \ R_N^{du,k}(\phi_N; \boldsymbol{\mu}) = 0 \qquad \forall \phi_N \in \mathcal{Z}_N^{du}, \ \forall k \in \{K, \dots, 1\}. \tag{7.22}$$

Moreover, from Eq.(7.4) and in analogy with Eq.(6.93), we define the corrected output, say $\widetilde{s}_N(\boldsymbol{\mu})$, s.t.:

$$\begin{aligned}
\widetilde{s}_N(\boldsymbol{\mu}) \quad := \quad & \Delta t \left(\frac{1}{2} l(v_N^0(\boldsymbol{\mu}); \boldsymbol{\mu}) + \sum_{k=1}^{K-1} l\left(v_N^k(\boldsymbol{\mu}); \boldsymbol{\mu}\right) + \frac{1}{2} l\left(v_N^K(\boldsymbol{\mu}); \boldsymbol{\mu}\right) \right) \\
& + \Delta t \sum_{k=1}^{K} R_N^{pr,k}\left(z_N^k(\boldsymbol{\mu}); \boldsymbol{\mu}\right),
\end{aligned} \tag{7.23}$$

which, in the "primal–dual" RB approach discussed in Chapter 6, allows an high accuracy in the approximation of the output functional.

As done for the FE approximations, also the RB ones are then rewritten in matricial formulation. In particular, by writing $v_N^k(\boldsymbol{\mu}) = \sum_{j=1}^{N^{pr}} v_{Nj}^k(\boldsymbol{\mu})\xi_j$, being ξ_j the basis function associated to the RB primal space $\mathcal{Z}_N^{pr}$, Eq.(7.17) reads:

$$\begin{aligned}
&\text{find } \mathbf{v}_N^k(\boldsymbol{\mu}) \in \mathbb{R}^{N^{pr}} \quad : \\
&\quad \left(M_N^{pr}(\boldsymbol{\mu}) + \Delta t \ A_N^{pr}(\boldsymbol{\mu})\right) \mathbf{v}_N^k(\boldsymbol{\mu}) = \\
&\qquad\qquad M_N^{pr}(\boldsymbol{\mu}) \mathbf{v}_N^{k-1}(\boldsymbol{\mu}) + \Delta t \ \mathbf{B}_N^{pr}(\boldsymbol{\mu}) g^k \qquad \forall k \in \{2, \dots, K\}, \\
&\text{with } \left(M_N^{pr}(\boldsymbol{\mu}) + \Delta t \ A_N^{pr}(\boldsymbol{\mu})\right) \mathbf{v}_N^1(\boldsymbol{\mu}) = \mathbf{v}_N^0(\boldsymbol{\mu}) + \Delta t \ \mathbf{B}_N^{pr}(\boldsymbol{\mu}) g^1,
\end{aligned} \tag{7.24}$$

being $\mathbf{v}_N^k(\boldsymbol{\mu}) := [v_{N1}^k(\boldsymbol{\mu}), \ldots, v_{NN^{pr}}^k(\boldsymbol{\mu})]^T$ and $\mathbf{v}_N^0(\boldsymbol{\mu}) = \mathbf{v}_{0h}(\boldsymbol{\mu})$. In this same manner, by imposing $z_N^k(\boldsymbol{\mu}) = \sum_{j=1}^{N^{du}} z_{Nj}^k(\boldsymbol{\mu})\zeta_j$, being ζ_j the basis function associated to the RB dual space $\mathcal{Z}_N^{du}$, Eq.(7.18) corresponds to:

$$\begin{aligned} &\text{find } \mathbf{z}_N^k(\boldsymbol{\mu}) \in \mathbb{R}^{N^{du}} \quad : \\ &\quad \left(M_N^{du}(\boldsymbol{\mu}) + \Delta t\ A_N^{du}(\boldsymbol{\mu})\right) \mathbf{z}_N^k(\boldsymbol{\mu}) = \\ &\qquad\qquad M_N^{du}(\boldsymbol{\mu})\mathbf{z}_N^{k+1}(\boldsymbol{\mu}) + \Delta t\ \mathbf{L}_N^{du}(\boldsymbol{\mu}) \qquad \forall k \in \{K-1, \ldots, 1\}, \\ &\text{with } \left(M_N^{du}(\boldsymbol{\mu}) + \Delta t\ A_N^{du}(\boldsymbol{\mu})\right) \mathbf{z}_N^K(\boldsymbol{\mu}) = \tfrac{1}{2}\Delta t\ \mathbf{L}_N^{du}(\boldsymbol{\mu}). \end{aligned} \tag{7.25}$$

Then, the corrected RB output (7.23) reads:

$$\begin{aligned} \widetilde{s}_N(\boldsymbol{\mu}) \quad = \quad & \Delta t\ \mathbf{L}_N^{pr}(\boldsymbol{\mu})^T \left(\frac{1}{2}\mathbf{v}_N^0(\boldsymbol{\mu}) + \sum_{k=1}^{K-1} \mathbf{v}_N^k(\boldsymbol{\mu}) + \frac{1}{2}\mathbf{v}_N^K(\boldsymbol{\mu}) \right) \\ & + \sum_{k=1}^{K} \left(\mathbf{R}_N^{du,pr,k}(\boldsymbol{\mu}) \right)^T \mathbf{z}_N^k(\boldsymbol{\mu}), \end{aligned} \tag{7.26}$$

with, from Eq.(7.19):

$$\begin{aligned} \mathbf{R}_N^{du,pr,k}(\boldsymbol{\mu}) := \mathbf{B}_N^{du}(\boldsymbol{\mu}) g^k - A_N^{du,pr}(\boldsymbol{\mu})\mathbf{v}_N^k(\boldsymbol{\mu}) \\ -\frac{1}{\Delta t} M_N^{du,pr}(\boldsymbol{\mu}) \left(v_N^k(\boldsymbol{\mu}) - v_N^{k-1}(\boldsymbol{\mu}) \right) \qquad \forall k \in \mathbb{K}. \end{aligned} \tag{7.27}$$

The RB matrices and vectors are defined as:

$$\begin{aligned} A_N^{pr}(\boldsymbol{\mu}) &= \sum_{q=1}^{Q_a} \vartheta_q^a(\boldsymbol{\mu}) A_{qN}^{pr}, & A_N^{du}(\boldsymbol{\mu}) &= \sum_{q=1}^{Q_a} \vartheta_q^a(\boldsymbol{\mu}) A_{qN}^{du}, & A_N^{du,pr}(\boldsymbol{\mu}) &= \sum_{q=1}^{Q_a} \vartheta_q^a(\boldsymbol{\mu}) A_{qN}^{du,pr}, \\ M_N^{pr}(\boldsymbol{\mu}) &= \sum_{q=1}^{Q_m} \vartheta_q^m(\boldsymbol{\mu}) M_{qN}^{pr}, & M_N^{du}(\boldsymbol{\mu}) &= \sum_{q=1}^{Q_m} \vartheta_q^m(\boldsymbol{\mu}) M_{qN}^{du}, & M_N^{du,pr}(\boldsymbol{\mu}) &= \sum_{q=1}^{Q_m} \vartheta_q^m(\boldsymbol{\mu}) M_{qN}^{du,pr}, \\ \mathbf{B}_N^{pr}(\boldsymbol{\mu}) &= \sum_{q=1}^{Q_b} \vartheta_q^b(\boldsymbol{\mu}) \mathbf{B}_{qN}^{pr}, & \mathbf{B}_N^{du}(\boldsymbol{\mu}) &= \sum_{q=1}^{Q_b} \vartheta_q^b(\boldsymbol{\mu}) \mathbf{B}_{qN}^{du}, & \mathbf{L}_N^{pr}(\boldsymbol{\mu}) &= \sum_{q=1}^{Q_l} \vartheta_q^l(\boldsymbol{\mu}) \mathbf{L}_{qN}^{pr}, \\ \mathbf{L}_N^{du}(\boldsymbol{\mu}) &= \sum_{q=1}^{Q_l} \vartheta_q^l(\boldsymbol{\mu}) \mathbf{L}_{qN}^{du}, & \mathbf{v}_{0N}(\boldsymbol{\mu}) &= \sum_{q=1}^{Q_m} \sum_{r=1}^{Q_v} \vartheta_q^m(\boldsymbol{\mu}) \vartheta_r^v(\boldsymbol{\mu}) \mathbf{v}_{0qrN}. \end{aligned}$$

The parameter independent RB matrices and vectors are:

$$\begin{aligned} \left(A_{qN}^{pr}\right)_{i,j} &= a_q(\xi_j, \xi_i), & \left(A_{qN}^{du}\right)_{i,j} &= a_q(\zeta_i, \zeta_j), & \left(A_{qN}^{du,pr}\right)_{i,j} &= a_q(\zeta_j, \xi_i), \\ \left(M_{qN}^{pr}\right)_{i,j} &= m_q(\xi_j, \xi_i), & \left(M_{qN}^{du}\right)_{i,j} &= m_q(\zeta_j, \zeta_i), & \left(M_{qN}^{du,pr}\right)_{i,j} &= m_q(\zeta_j, \xi_i), \\ \left(\mathbf{B}_{qN}^{pr}\right)_i &= b_q(\xi_i), & \left(\mathbf{B}_{qN}^{du}\right)_i &= b_q(\zeta_i), & \left(\mathbf{L}_{qN}^{pr}\right)_i &= l_q(\xi_i), \\ \left(\mathbf{L}_{qN}^{du}\right)_i &= l_q(\zeta_i), & \left(\mathbf{v}_{0qrN}\right)_i &= m_q(\xi_i, v_{0r}). \end{aligned}$$

As already explained in Chapter 6, the offline–online decomposition holds. Moreover, as in the steady case, the online computational costs are independent on the "large" dimension N_h of the FE problem. In particular, the computational costs associated to the corrected output (7.23) is of order
$O\left(N^{pr3} + N^{du3} + K(N^{pr2} + N^{pr2}) + (N^{pr2} + N^{du2})(Q_a + Q_m) + K(K+1)N^{pr}N^{du}\right)$ ([128]), for which, being N^{pr}, $N^{du} \ll N_h$, significative costs savings are made possible by the RB method.
The computation of $\widetilde{s}_N(\boldsymbol{\mu})$ is conducted by using the "primal–dual" approach for which, in general, $N^{pr} \neq N^{du}$ and $\mathcal{Z}_N^{pr} \neq \mathcal{Z}_N^{du}$, as already highlighted in Chapter 6. Hence, at the online step, the choices of N^{pr} and N^{du} could be performed on the basis of a table E_N similar to that introduced in Sec.6.3.4.

7.1.3 A posteriori error estimate and adaptive procedure

A posteriori RB error estimates are provided also in the case of parametrized parabolic PDEs both for the RB primal and dual solutions and output functional. The purpose of such estimates is twofold: firstly, to evaluate the RB errors at the online step, secondly, to lead, at the offline step, adaptive procedures for the definition of the RB samples sets $\mathcal{S}_N^{pr}$, $\mathcal{S}_N^{du}$ and the corresponding RB spaces $\mathcal{Z}_N^{pr}$ and $\mathcal{Z}_N^{du}$; see Chapter 6.

With this aim, let us introduce the following definitions, on which the a posteriori error estimates are based. By indicating with $\alpha(\boldsymbol{\mu}) : \mathcal{D} \to \mathbb{R}$ the coercivity constant of the bilinear form $a(\cdot, \cdot, \boldsymbol{\mu})$, then we define with $\widetilde{\alpha}(\boldsymbol{\mu}) : \mathcal{D} \to \mathbb{R}$ its lower bound, s.t.:

$$\alpha(\boldsymbol{\mu}) \geq \widetilde{\alpha}(\boldsymbol{\mu}) \geq \alpha^0 > 0 \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{7.28}$$

with $\alpha^0 > 0$. We observe that the lower bound $\widetilde{\alpha}(\boldsymbol{\mu})$ plays the analogous role of the inf–sup constant $\widetilde{\beta}(\boldsymbol{\mu})$ of Eq.(6.106). Different recipes can be adopted in order to evaluate this lower bound, while maintaining $\widetilde{\alpha}(\boldsymbol{\mu})$ as closer as possible to $\alpha(\boldsymbol{\mu})$; see e.g. [90, 136]. By recalling Eq.s (6.96) and (6.99), we indicate the dual norms of the primal residual (7.19) and the dual residual (7.20) as, respectively:

$$\varepsilon_N^{pr,k}(\boldsymbol{\mu}) := \sup_{\phi_h \in \mathcal{Z}_h \setminus \{0\}} \frac{R_N^{pr,k}(\phi_h; \boldsymbol{\mu})}{||\phi_h||_{\mathcal{Z}_h}} \qquad \forall k \in \mathbb{K}, \tag{7.29}$$

$$\varepsilon_N^{du,k}(\boldsymbol{\mu}) := \sup_{\phi_h \in \mathcal{Z}_h \setminus \{0\}} \frac{R_N^{du,k}(\phi_h; \boldsymbol{\mu})}{||\phi_h||_{\mathcal{Z}_h}} \qquad \forall k \in \mathbb{K}. \tag{7.30}$$

Finally, we define the primal and dual "spatial–temporal" energy norms as, respectively:

$$|||w_h^k(\boldsymbol{\mu})|||^{pr} := \left(m\left(w_h^k(\boldsymbol{\mu}), w_h^k(\boldsymbol{\mu}); \boldsymbol{\mu}\right) + \sum_{j=1}^{k} \Delta t \ a\left(w_h^j(\boldsymbol{\mu}), w_h^j(\boldsymbol{\mu}); \boldsymbol{\mu}\right) \right)^{1/2} \qquad \forall w_h \in \mathcal{Z}_h, \tag{7.31}$$

$$|||w_h^k(\boldsymbol{\mu})|||^{du} := \left(m\left(w_h^k(\boldsymbol{\mu}), w_h^k(\boldsymbol{\mu}); \boldsymbol{\mu}\right) + \sum_{j=k}^{K} \Delta t \; a\left(w_h^j(\boldsymbol{\mu}), w_h^j(\boldsymbol{\mu}); \boldsymbol{\mu}\right) \right)^{1/2} \qquad \forall w_h \in \mathcal{Z}_h. \tag{7.32}$$

Hence, the following results follow. For the proof we refer the reader to [128].

Proposition 7.1. *The RB primal error* $e_N^{pr,k}(\boldsymbol{\mu}) = v_h^k(\boldsymbol{\mu}) - v_N^k(\boldsymbol{\mu})$ *at the time step* t_k *is bounded as:*

$$|||e_N^{pr,k}(\boldsymbol{\mu})|||^{pr} \leq \Delta_N^{pr,k}(\boldsymbol{\mu}), \tag{7.33}$$

with:

$$\Delta_N^{pr,k}(\boldsymbol{\mu}) := \left(\frac{\Delta t}{\widetilde{\alpha}(\boldsymbol{\mu})} \sum_{j=1}^{k} \left(\varepsilon_N^{pr,j}(\boldsymbol{\mu}) \right)^2 \right)^{1/2}, \tag{7.34}$$

being $\widetilde{\alpha}(\boldsymbol{\mu})$ *and* $\varepsilon_N^{pr,j}(\boldsymbol{\mu})$ *given in Eq.s (7.28) and (7.29), respectively.*

Proposition 7.2. *The RB dual error* $e_N^{du,k}(\boldsymbol{\mu}) = z_h^k(\boldsymbol{\mu}) - z_N^k(\boldsymbol{\mu})$ *at the time step* t_k *is bounded as:*

$$|||e_N^{du,k}(\boldsymbol{\mu})|||^{du} \leq \Delta_N^{du,k}(\boldsymbol{\mu}), \tag{7.35}$$

with:

$$\Delta_N^{du,k}(\boldsymbol{\mu}) := \left(\frac{\Delta t}{\widetilde{\alpha}(\boldsymbol{\mu})} \sum_{j=k}^{K} \left(\varepsilon_N^{du,j}(\boldsymbol{\mu}) \right)^2 \right)^{1/2}, \tag{7.36}$$

being $\widetilde{\alpha}(\boldsymbol{\mu})$ *and* $\varepsilon_N^{du,j}(\boldsymbol{\mu})$ *given in Eq.s (7.28) and (7.30), respectively.*

Theorem 7.1. *From Eq.s (7.14) and (7.23) the following a posteriori RB error estimate for the corrected output functional holds:*

$$|s_h(\boldsymbol{\mu}) - \widetilde{s}_N(\boldsymbol{\mu})| \leq \Delta_N^s(\boldsymbol{\mu}) \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{7.37}$$

with:

$$\Delta_N^s(\boldsymbol{\mu}) := \Delta_N^{pr,K}(\boldsymbol{\mu}) \; \Delta_N^{du,1}(\boldsymbol{\mu}), \tag{7.38}$$

being $\Delta_N^{pr,K}(\boldsymbol{\mu})$ *and* $\Delta_N^{du,1}(\boldsymbol{\mu})$ *defined in Eq.s (7.34) and (7.36), respectively.*

We observe that, similarly to Sec.7.1.2 the a posteriori RB error estimates can be implemented by means of RB matrices and vectors, which are independent on the dimension N_h of the FE problem. In fact, the cost associated with the computation of the estimate (7.37) is of order $O(K({N^{pr}}^2 + {N^{du}}^2)(Q_a^2 + Q_a Q_m + Q_m^2))$.

For the choice of the samples sets $\mathcal{S}_N^{pr}$ and $\mathcal{S}_N^{du}$ and the RB basis spaces $\mathcal{Z}_N^{pr}$ and $\mathcal{Z}_N^{du}$ an adaptive procedure is used; see [128]. In particular, on the basis of the a posteriori error estimate (7.37) and the estimator (7.38), we deduce that the primal and dual estimators (7.34) and (7.36) can be computed independently; hence, the sampling procedures for the primal and dual problems can be performed separately amd performed in analogous manner. For this reason, we report only the sampling procedure for the primal problem, from which that of the dual problem can be deduced.

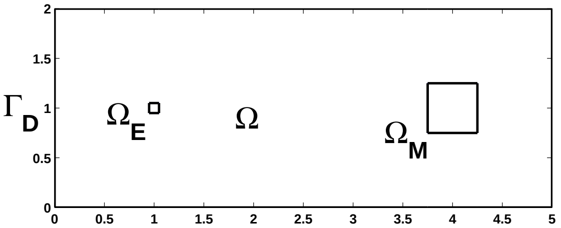

Figure 7.1: Test P. Computational domain.

As already done in Sec.6.3.4, we introduce a discrete subset $\overline{\mathcal{D}} \subset \mathcal{D}$ containing $\overline{N}$ samples. Moreover, we recall that the primal solution depends on the input function g^k. For this reason, the adaptive procedure is defined for a prescribed input function g^k; for example, the impulse function is typically used.

The procedure starts by choosing a first sample $\boldsymbol{\mu}_1 \in \overline{\mathcal{D}}$ and solving the corresponding FE primal problem (7.3), thus obtaining $\{v_h^k(\boldsymbol{\mu}_1)\}_{k=1}^K$. A POD procedure (see [103] and also e.g. [35, 82, 185]) is used to compute the first I_1 eigenmodes $\{\rho_i(\boldsymbol{\mu}_1)\}_{i=1}^{I_1}$ of the set $\{v_h^k(\boldsymbol{\mu}_1)\}_{k=1}^K$. Then, we define $\mathcal{Z}_N^{pr} = \{\xi_1 = \rho_1(\boldsymbol{\mu}_1)\}$, we set $N^{pr} = 1$ and we introduce the set of snapshots $\mathcal{Q} = \{\rho_i(\boldsymbol{\mu}_1)\}_{i=1}^{I_1}$. The successive sample $\boldsymbol{\mu}_{N^{pr}+1}$ is chosen as:

$$\boldsymbol{\mu}_{N^{pr}+1} = \arg\max_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} \left(\max_{k\in\mathbb{K}} \Delta_N^{pr,k}(\boldsymbol{\mu}) \right), \tag{7.39}$$

for which the snapshots set $\mathcal{Q}$ is enriched as $\mathcal{Q} = \mathcal{Q} \cup \{\rho_i(\boldsymbol{\mu}_{N^{pr}+1})\}_{i=1}^{I_{N^{pr}+1}}$, being $\{\rho_i(\boldsymbol{\mu}_{N^{pr}+1})\}_{i=1}^{I_{N^{pr}+1}}$ the first $I_{N^{pr}+1}$ POD eigenmodes obtained from the set $\{v_h^k(\boldsymbol{\mu}_{N^{pr}+1})\}_{k=1}^K$. The POD procedure is used again in order to compute the basis functions $\{\xi_i\}_{i=1}^{N^{pr}+1}$ starting from the set $\mathcal{Q}$; the RB primal space is then defined as $\mathcal{Z}_N^{pr} = \text{span}\{\xi_1, \ldots, \xi_{N^{pr}+1}\}$. Finally, we set $N^{pr} = N^{pr} + 1$ and we repeat the procedure until $\max_{k\in\mathbb{K}} \Delta_N^{pr,k}(\boldsymbol{\mu}) < tol$, being tol a prescribed tolerance.

7.1.4 Numerical test

In this Section we provide a numerical test, inspired by an environmental application (see Chapter 2 and particularly Sec.1.3), for which a parabolic advection–diffusion PDE is considered. We refer to this test as Test P.

With this aim, by recalling Sec.7.1.1, let us consider the following parametrized parabolic PDE:

$$\begin{cases} \dfrac{\partial v(t,\boldsymbol{\mu})}{\partial t} - \nabla \cdot \left(\nu \nabla v(t,\boldsymbol{\mu}) + \mathbf{V}(\boldsymbol{\mu}) v(t,\boldsymbol{\mu}) \right) = g(t) f & \text{in } \Omega,\ t \in (0,T), \\ v(t,\boldsymbol{\mu}) = 0 & \text{on } \Gamma_D,\ t \in (0,T), \\ \nu \nabla v(t,\boldsymbol{\mu}) \cdot \hat{\mathbf{n}} + V^-(\boldsymbol{\mu}) v(t,\boldsymbol{\mu}) = 0 & \text{on } \Gamma_N,\ t \in (0,T), \\ v(0,\boldsymbol{\mu}) = 0 & \text{on } \Omega, \end{cases} \tag{7.40}$$

where the domain Ω is reported in Fig.7.1, $\nu \in \mathbb{R}^+$, $f \in L^2(\Omega)$, $\mathbf{V}(\boldsymbol{\mu}) = [\mu_1, \mu_2]^T$, with $V^-(\boldsymbol{\mu}) = -\mathbf{V}(\boldsymbol{\mu}) \cdot \hat{\mathbf{n}}$ if $\mathbf{V}(\boldsymbol{\mu}) \cdot \hat{\mathbf{n}} < 0$ else $V^-(\boldsymbol{\mu}) = 0$, being $\hat{\mathbf{n}}$ is the outward directed unit

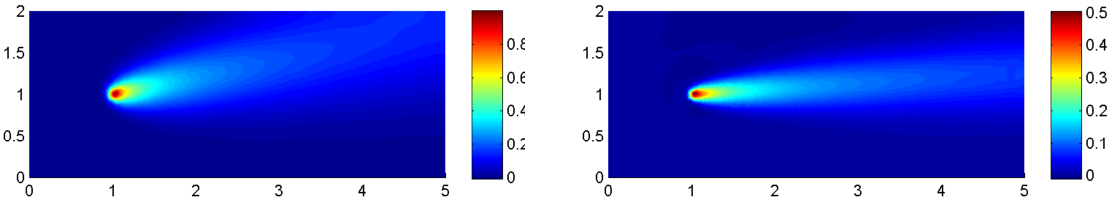

Figure 7.2: Test P. RB primal solutions for $t_K = T$ for $\boldsymbol{\mu}_a$ (left) and $\boldsymbol{\mu}_b$ (right).

vector normal to $\Gamma_N := \partial\Omega\backslash\Gamma_D$. In particular, we assume $\nu = 1$, $f = 1000\chi_E$, being χ_E the characteristic function of the emission subdomain Ω_E reported in Fig.7.1, $T = 0.4$ and $\boldsymbol{\mu} \in \mathcal{D}$, with $\mathcal{D} = [5, 75] \times [0, 5]$. We are interested in evaluating $s(\boldsymbol{\mu})$ (see Eq.(7.2)), for which $l(\cdot; \boldsymbol{\mu}) = \frac{1}{|\Omega_M|}\int_\Omega \cdot\chi_M \, d\Omega$, being χ_M the characteristic function of the subdomain Ω_M.
We observe that Eq.(7.40) could be rewritten in weak form as in Eq.(7.1), for which the affine decomposition hypothesis holds, with $Q_a = 3$, $Q_m = 1$, $Q_b = 1$, $Q_l = 1$ and $Q_v = 1$.
The problem is solved by means of the RB method, after having introduced the FE approximation based on piecewise linear basis functions defined on triangular elements. In particular, we set $\Delta t = 10^{-2}$, for which $K = 40$, and we choose a mesh composed by $7,320$ triangles and $3,723$ nodes; for the definition of the RB primal and dual spaces, we use the adaptive sampling procedure described in Sec.7.1.3 for which we choose $tol = 10^{-3}$.
The simulations are performed by using the *rbMIT©MIT Software*, see [193], and carried out on an Intel®Pentium®M 1.60 GHz processor, with 2 MB of Memory Cache and 512 MB of RAM.

By choosing $g(t) = 0.5$, $\boldsymbol{\mu} = \boldsymbol{\mu}_a = (25, 5)$, $N^{pr} = 54$ and $N^{du} = 43$, we obtain the RB primal solution reported in Fig.7.2(left) at the final time step $t_K = T$, for which the corresponding error in energy norm is bounded by $\Delta_N^{pr,K}(\boldsymbol{\mu}_a) = 6.089 \cdot 10^{-2}$. Moreover, we obtain $\widetilde{s}_N(\boldsymbol{\mu}_a) = 2.661 \cdot 10^{-2}$, whose associated error is bounded by $\Delta_N^s(\boldsymbol{\mu}_a) = 8.186 \cdot 10^{-4}$. We notice that the computational cost associated with the online RB step is 0.355 s w.r.t. the 3.21 s required by the FE solution.
Similarly, by choosing now $\boldsymbol{\mu} = \boldsymbol{\mu}_b = (75, 5)$, we obtain the RB primal solution reported in Fig.7.2(right) at the final time step $t_K = T$, for which the error in energy norm is bounded by $\Delta_N^{pr,K}(\boldsymbol{\mu}_b) = 5.503 \cdot 10^{-2}$. In this case, we have $\widetilde{s}_N(\boldsymbol{\mu}_b) = 2.921 \cdot 10^{-2}$, whose associated error is bounded by $\Delta_N^s(\boldsymbol{\mu}_b) = 2.865 \cdot 10^{-4}$.

7.2 The RB method for parametrized optimal control problems: the "integrated" approach and a posteriori error estimate

In this Section we provide the formulation of the RB method for the solution of parametrized optimal control problems. Firstly, we introduce the RB approximation into the optimal control framework, taking into account for both the unconstrained and constrained cases; the choice of the RB space is discussed in an abstract setting. Then, after having introduced the offline–online decomposition and having discussed the "predictability" and "optimality" errors, we provide the a posteriori error estimate for the "optimality" error on the cost functional. Finally, we propose our "integrated" RB approach, for which we choose the RB basis and space taking into account for the features of both the primal and dual problems. An adaptive procedure for the definition of such basis is provided as well as numerical tests proving the effectivity of the proposed approach.

7.2.1 The general problem and its FE approximation

The parametrized optimal control problem can be introduced by adding the parameter dependence into the formulation provided in Chapter 2.

In particular, by adopting the notation of Sec.7.1, the *parametrized optimal control* problem described by a parabolic PDE reads:

$$\begin{aligned}&\text{find } u(\boldsymbol{\mu}) = u(t,\boldsymbol{\mu}) \in \mathcal{U}_{ad}, \quad u(t,\boldsymbol{\mu}) = \text{argmin}\left(\ J\left(v(\boldsymbol{\mu}), u(\boldsymbol{\mu}); \boldsymbol{\mu}\right) \ \right)\\ &\text{where } v(\boldsymbol{\mu}) = v(t,\boldsymbol{\mu}) \in \mathcal{V} \ \text{ is solution of}\\ &m\left(\frac{\partial v(t,\boldsymbol{\mu})}{\partial t}, \phi; \boldsymbol{\mu}\right) + a(v(t,\boldsymbol{\mu}), \phi; \boldsymbol{\mu}) = b(\phi;\boldsymbol{\mu})u(t,\boldsymbol{\mu}) \quad \forall \phi \in \mathcal{Z},\ t \in (0,T),\\ &\text{with } v(0;\boldsymbol{\mu}) = v_0(\boldsymbol{\mu}).\end{aligned} \tag{7.41}$$

We consider the following cost functional $J(\cdot,\cdot;\boldsymbol{\mu})$, in analogy with Eq.(2.15):

$$\begin{aligned}J(v(\boldsymbol{\mu}), u(\boldsymbol{\mu}); \boldsymbol{\mu}) \quad = \quad & \frac{1}{2}\int_0^T m_d(v(t,\boldsymbol{\mu}) - v_d(t), v(t,\boldsymbol{\mu}) - v_d(t); \boldsymbol{\mu})\ dt\\ &+\frac{1}{2}\gamma \int_0^T \left(u(t,\boldsymbol{\mu}) - u_d(t)\right)^2\ dt,\end{aligned} \tag{7.42}$$

where the form $m_d(\cdot,\cdot;\boldsymbol{\mu})$ is symmetric, continuous, bilinear and positive in the norm induced on the space $\mathcal{Y}_d \subseteq \mathcal{Y}$, $\gamma > 0$ and, for the sake of simplicity, $v_d(t)$ and $u_d(t)$ do not depend on the parameter vector $\boldsymbol{\mu}$. For the form $m_d(\cdot,\cdot;\boldsymbol{\mu})$ we suppose that the affine decomposition hypothesis holds s.t. $m_d(w,\phi) = \sum_{q=1}^{Q_d} \vartheta_q^d(\boldsymbol{\mu}) m_{dq}(w,\phi)\ \forall w,\phi \in \mathcal{Z}$, $\forall \boldsymbol{\mu} \in \mathcal{D}$. The space of admissible controls $\mathcal{U}_{ad}$ is taken s.t. $\mathcal{U}_{ad} = \Big\{u(t,\boldsymbol{\mu}) \in \mathcal{U}\ :\ u_{min} \leq u(t,\boldsymbol{\mu}) \leq u_{max}$ and $\int_0^T u(t,\boldsymbol{\mu})\ dt \leq T\overline{U} \ \ \forall \boldsymbol{\mu} \in \mathcal{D}\Big\}$, with $\mathcal{U} = L^2(0,T)$; hence, we require that $b(\cdot;\boldsymbol{\mu}) \in L^2(0,T;H^{-1}(\Omega))\ \forall \boldsymbol{\mu} \in \mathcal{D}$. If $\mathcal{U}_{ad} \equiv \mathcal{U}$, the parametrized optimal control problem is unconstrained.

The analysis of the continuous optimal control problem mimics the one provided in Sec.s 2.1 and 2.2 for the unconstrained and constrained cases. With this aim, the parameter dependent Lagrangian functional is defined as $\mathcal{L}\left(v(t,\boldsymbol{\mu}), z(t,\boldsymbol{\mu}), u(t,\boldsymbol{\mu}); \boldsymbol{\mu}\right) : \mathcal{X}_{ad} \to \mathbb{R}\ \forall \boldsymbol{\mu} \in \mathcal{D}$.

Then, the first order necessary conditions (2.4)–(2.6) in the unconstrained case or (2.29)–(2.31) in the constrained one, hold $\forall \boldsymbol{\mu} \in \mathcal{D}$.

Let us remark that the optimal solution of problem (7.41) is parameter dependent; i.e. the optimal cost functional is $J^{**}(\boldsymbol{\mu}) = J(v^{**}(t,\boldsymbol{\mu}), u^{**}(t,\boldsymbol{\mu}); \boldsymbol{\mu})$ corresponding to the optimal control function $u^{**}(t,\boldsymbol{\mu}) \in \mathcal{U}_{ad}$. This means that, as the parameters vector $\boldsymbol{\mu}$ changes, also the optimal solution does; for that reason we refer to problem (7.41) as a parametrized optimal control problem.

The FE approximation is represented by the extension of that presented in Sec.7.1.1 for the case of parabolic PDEs. In particular, we make use of the "discretize–then–optimize" approach to optimal control problems, as described in Sec.4.2.4.
The FE primal problem reads:

$$\begin{aligned}&\text{find } v_h^k(\boldsymbol{\mu}) \in \mathcal{Z}_h \quad : \quad \Delta t \; R_h^{pr,k}(\phi_h;\boldsymbol{\mu}) = 0 \qquad \forall \phi_h \in \mathcal{Z}_h, \text{ with } u_h^k(\boldsymbol{\mu}) \in \mathcal{U}_{ad,h}, \; \forall k \in \mathbb{K} \\ &\text{with } v_h^0(\boldsymbol{\mu}) = v_{0h}(\boldsymbol{\mu}),\end{aligned} \tag{7.43}$$

where the FE primal residual is now re–defined as:

$$R_h^{pr,k}(\phi_h;\boldsymbol{\mu}) := R^{pr,k}\left(v_h^k(\boldsymbol{\mu}), v_h^{k-1}(\boldsymbol{\mu}), u_h^k(\boldsymbol{\mu})\right)(\phi_h;\boldsymbol{\mu}) \qquad \forall \phi_h \in \mathcal{Z}_h, \; \forall k \in \mathbb{K}, \tag{7.44}$$

being, in analogy with Eq.(7.6):

$$\begin{aligned}R^{pr,k}\left(w_h^k, w_h^{k-1}, q_h^k\right)(\phi_h;\boldsymbol{\mu}) &:= b(\phi_h,\boldsymbol{\mu}) q_h^k(\boldsymbol{\mu}) - a\left(w_h^k(\boldsymbol{\mu}),\phi_h;\boldsymbol{\mu}\right) \\ &\quad -\frac{1}{\Delta t} m\left(w_h^k(\boldsymbol{\mu}) - w_h^{k-1}(\boldsymbol{\mu}),\phi_h;\boldsymbol{\mu}\right) \qquad \forall \phi_h \in \mathcal{Z}_h, \; \forall k \in \mathbb{K},\end{aligned} \tag{7.45}$$

for some $w_h^k \in \mathcal{Z}_h$ $\forall k \in \{0,1,\dots,K\}$ and $q_h^k \in \mathcal{U}_{ad,h}$ $\forall k \in \mathbb{K}$. Let us observe that with the notation $u_h^k(\boldsymbol{\mu})$ we indicate the control function associated with the FE approximated optimal control problem and not the FE approximation of a given control function. Moreover, with $\mathcal{U}_h$ and $\mathcal{U}_{ad,h}$ we indicate the discrete versions of the control spaces $\mathcal{U}$ and $\mathcal{U}_{ad}$, respectively.
The FE approximated cost functional (7.42) reads:

$$\begin{aligned}J_h(\mathbf{y}_h(\boldsymbol{\mu});\boldsymbol{\mu}) \;=\;& \frac{1}{2}\Delta t \left[\frac{1}{2} m_d\left(v_h^0(\boldsymbol{\mu}) - v_d^0, v_h^0(\boldsymbol{\mu}) - v_d^0;\boldsymbol{\mu}\right)\right. \\ &+ \sum_{k=1}^{K-1} m_d\left(v_h^k(\boldsymbol{\mu}) - v_d^k, v_h^k(\boldsymbol{\mu}) - v_d^k;\boldsymbol{\mu}\right) \\ &\left.+ \frac{1}{2} m_d\left(v_h^K(\boldsymbol{\mu}) - v_d^K, v_h^K(\boldsymbol{\mu}) - v_d^K;\boldsymbol{\mu}\right)\right] \\ &+\frac{1}{2}\gamma\Delta t\left[\sum_{k=1}^{K-1}\left(u_h^k(\boldsymbol{\mu}) - u_d^k\right)^2 + \frac{1}{2}\left(u_h^K(\boldsymbol{\mu}) - u_d^K\right)^2\right],\end{aligned} \tag{7.46}$$

where $v_d^k = v_d(t_k)$ and $u_d^k = u_d(t)$; moreover, $\mathbf{y}_h(\boldsymbol{\mu}) := \{v_h^0(\boldsymbol{\mu}),\dots,v_h^K(\boldsymbol{\mu})\} \times \{u_h^1(\boldsymbol{\mu}),\dots,u_h^K(\boldsymbol{\mu})\}$. Let us define $\mathbf{x}_h(\boldsymbol{\mu}) \in \mathcal{X}_{ad,h}$ as $\mathbf{x}_h(\boldsymbol{\mu}) := \left\{v_h^0(\boldsymbol{\mu}),\dots,v_h^K(\boldsymbol{\mu})\right\} \times \left\{z_h^0(\boldsymbol{\mu}),\dots,z_h^K(\boldsymbol{\mu})\right\} \times \left\{u_h^1(\boldsymbol{\mu}),\dots,u_h^K(\boldsymbol{\mu})\right\}$, with $\mathcal{X}_{ad,h} := (\mathcal{Z}_h)^{2(K+1)} \times (\mathcal{U}_{ad,h})^K$; we observe that if $\mathcal{U}_{ad,h} \equiv \mathcal{U}$, we have $\mathcal{X}_{ad,h} \equiv \mathcal{X}$. Then, by using the primal equation (7.43), the FE based parametrized Lagrangian functional $\mathcal{L}_h(\mathbf{x}_h(\boldsymbol{\mu});\boldsymbol{\mu}) : \mathcal{X}_{ad,h} \to \mathbb{R}$ $\forall \boldsymbol{\mu} \in \mathcal{D}$ reads:

$$\begin{aligned}\mathcal{L}_h(\mathbf{x}_h(\boldsymbol{\mu});\boldsymbol{\mu}) \;:=\;& J_h(\mathbf{y}_h;\boldsymbol{\mu}) + \Delta t \sum_{k=1}^{K} R^{pr,k}\left(v_h^k(\boldsymbol{\mu}), v_h^{k-1}(\boldsymbol{\mu}), u_h^k(\boldsymbol{\mu})\right)\left(z_h^k(\boldsymbol{\mu});\boldsymbol{\mu}\right) \\ &+m(v_h^0(\boldsymbol{\mu}) - v_{0h}(\boldsymbol{\mu}), z_h^0(\boldsymbol{\mu});\boldsymbol{\mu}),\end{aligned} \tag{7.47}$$

or, from Eq.(7.44), in more compact form:

$$\begin{aligned}\mathcal{L}_h(\mathbf{x}_h(\boldsymbol{\mu});\boldsymbol{\mu}) \;:=\;& J_h(\mathbf{y}_h(\boldsymbol{\mu});\boldsymbol{\mu}) + \Delta t \sum_{k=1}^{K} R_h^{pr,k}\left(z_h^k(\boldsymbol{\mu});\boldsymbol{\mu}\right) \\ &+m(v_h^0(\boldsymbol{\mu}) - v_{0h}(\boldsymbol{\mu}), z_h^0(\boldsymbol{\mu});\boldsymbol{\mu}).\end{aligned} \tag{7.48}$$

Let us observe that we indicate with $\mathbf{x}_h^{**}(\boldsymbol{\mu}) := \left\{ v_h^{**,0}(\boldsymbol{\mu}), \dots, v_h^{**,K}(\boldsymbol{\mu}) \right\} \times \left\{ z_h^{**,0}(\boldsymbol{\mu}), \dots, z_h^{**,K}(\boldsymbol{\mu}) \right\} \times \left\{ u_h^{**,1}(\boldsymbol{\mu}), \dots, u_h^{**,K}(\boldsymbol{\mu}) \right\}$ the parameter dependent FE optimal solution, with $\mathbf{x}_h^{**}(\boldsymbol{\mu}) \in \mathcal{X}_{ad,h}$, while the FE approximated optimal cost functional reads $J_h^{**}(\boldsymbol{\mu}) = J_h(\mathbf{y}_h^{**}(\boldsymbol{\mu}); \boldsymbol{\mu})$, being $\mathbf{y}_h^{**}(\boldsymbol{\mu}) := \left\{ v_h^{**,0}(\boldsymbol{\mu}), \dots, v_h^{**,K}(\boldsymbol{\mu}) \right\} \times \left\{ u_h^{**,1}(\boldsymbol{\mu}), \dots, u_h^{**,K}(\boldsymbol{\mu}) \right\}$. We notice that we have used the double superscript $**$ also for the FE approximation, even if in Sec.4.2 it has been reserved to the continuous optimal solution; this is due to the fact that the FE approximation is considered as the "truth" approximation and we have in mind to evaluate the RB errors and not the FE ones.

By differentiating $\mathcal{L}_h(\mathbf{x}_h; \boldsymbol{\mu})$ w.r.t. $v_h^k(\boldsymbol{\mu})$ $\forall k \in \mathbb{K}$, we obtain the FE dual equation:

$$\text{find } z_h^k(\boldsymbol{\mu}) \in \mathcal{Z}_h \quad : \quad \Delta t \; R_h^{du,k}(\phi_h; \boldsymbol{\mu}) = 0 \qquad \forall \phi_h \in \mathcal{Z}_h, \; \text{with } v_h^k(\boldsymbol{\mu}) \in \mathcal{Z}_h, \; \forall k \in \{K, \dots, 1\}, \tag{7.49}$$

with the FE dual residual now re–defined as:

$$R_h^{du,k}(\phi_h; \boldsymbol{\mu}) := R^{du,k}\left(z_h^k(\boldsymbol{\mu}), z_h^{k+1}(\boldsymbol{\mu}), v_h^k(\boldsymbol{\mu}) \right)(\phi_h; \boldsymbol{\mu}) \qquad \forall \phi_h \in \mathcal{Z}_h, \; \forall k \in \{K, \dots, 1\}, \tag{7.50}$$

being:

$$R^{du,k}\left(p_h^k, p_h^{k+1}, w_h^k \right)(\phi_h; \boldsymbol{\mu}) := \begin{cases} m_d\left(w_h^k - v_d^k, \phi_h; \boldsymbol{\mu} \right) - a\left(\phi, p_h^k; \boldsymbol{\mu} \right) - \dfrac{1}{\Delta t} m\left(\phi_h, p_h^k - p_h^{k+1}; \boldsymbol{\mu} \right) \\ \qquad\qquad \forall \phi_h \in \mathcal{Z}, \; \forall k \in \{K-1, \dots, 1\}, \\ \dfrac{1}{2} m_d\left(w_h^K - v_d^K, \phi_h; \boldsymbol{\mu} \right) - a\left(\phi_h, p_h^K; \boldsymbol{\mu} \right) - \dfrac{1}{\Delta t} m\left(\phi_h, p_h^K; \boldsymbol{\mu} \right) \\ \qquad\qquad \forall \phi_h \in \mathcal{Z}_h, \; k = K. \end{cases} \tag{7.51}$$

Finally, by differentiating $\mathcal{L}_h(\mathbf{x}_h; \boldsymbol{\mu})$ w.r.t. $u_h^k(\boldsymbol{\mu})$ $\forall k \in \mathbb{K}$, we obtain the FE approximated optimality constraint, which in the unconstrained case reads:

$$\Delta t \; R_h^{opt,**,k}(\psi_h; \boldsymbol{\mu}) = 0 \qquad \forall \psi_h \in \mathcal{U}_h, \; \forall k \in \mathbb{K}, \tag{7.52}$$

while in the constrained one is:

$$\Delta t \; R_h^{opt,**,k}(\psi_h - u_h^{**,k}(\boldsymbol{\mu}); \boldsymbol{\mu}) \geq 0 \qquad \forall \psi_h \in \mathcal{U}_{ad,h}, \; \forall k \in \mathbb{K}. \tag{7.53}$$

The FE optimality residual $R_h^{opt,**,k}(\cdot; \boldsymbol{\mu})$ is defined as:

$$R_h^{opt,**,k}(\psi_h; \boldsymbol{\mu}) := R^{opt,k}\left(z_h^{**,k}(\boldsymbol{\mu}), u_h^{**,k}(\boldsymbol{\mu}) \right)(\psi_h; \boldsymbol{\mu}) \qquad \forall \psi_h \in \mathcal{U}_h, \; \forall k \in \mathbb{K}, \tag{7.54}$$

and similarly:

$$R_h^{opt,k}(\psi_h; \boldsymbol{\mu}) := R^{opt,k}\left(z_h^k(\boldsymbol{\mu}), u_h^k(\boldsymbol{\mu}) \right)(\psi_h; \boldsymbol{\mu}) \qquad \forall \psi_h \in \mathcal{U}_h, \; \forall k \in \mathbb{K}, \tag{7.55}$$

being:

$$R^{opt,k}\left(p_h^k, q_h^k \right)(\psi_h; \boldsymbol{\mu}) := \begin{cases} \gamma(q_h^k - u_d^k)\psi_h + b(p_h^k; \boldsymbol{\mu})\psi_h \\ \qquad\qquad \forall \psi_h \in \mathcal{U}_h, \; \forall k \in \{1, \dots, K-1\}, \\ \dfrac{1}{2}\gamma(q_h^K - u_d^K)\psi_h + b(p_h^K; \boldsymbol{\mu})\psi_h \\ \qquad\qquad \forall \psi_h \in \mathcal{U}_h, \; k = K. \end{cases} \tag{7.56}$$

Similarly, for the sake of simplicity, we define, from Eq.s (7.44) and (7.50) the following optimal residuals, respectively:

$$R_h^{pr,**,k}(\phi_h;\boldsymbol{\mu}) := R^{pr,k}\left(v_h^{**,k}(\boldsymbol{\mu}), v_h^{**,k-1}(\boldsymbol{\mu}), u_h^{**,k}(\boldsymbol{\mu})\right)(\phi_h;\boldsymbol{\mu}) \qquad \forall \phi_h \in \mathcal{Z}_h,\ \forall k \in \mathbb{K}, \tag{7.57}$$

$$R_h^{du,**,k}(\phi_h;\boldsymbol{\mu}) := R^{du,k}\left(z_h^{**,k}(\boldsymbol{\mu}), z_h^{**,k+1}(\boldsymbol{\mu}), v_h^{**,k}(\boldsymbol{\mu})\right)(\phi_h;\boldsymbol{\mu}) \qquad \forall \phi_h \in \mathcal{Z}_h, \forall k \in \{K,\dots,1\}. \tag{7.58}$$

Synthetically, in analogy with Eq.s (4.50) and (4.51), the first order necessary conditions read, in the unconstrained case:

$$\text{find } \mathbf{x}_h^{**}(\boldsymbol{\mu}) \in \mathcal{X}_h \quad : \quad \mathcal{L}'_h\left(\mathbf{x}_h^{**}(\boldsymbol{\mu});\boldsymbol{\mu}\right)(\boldsymbol{\phi}_h) = 0 \qquad \forall \boldsymbol{\phi}_h \in \mathcal{X}_h, \tag{7.59}$$

while in the constrained one:

$$\text{find } \mathbf{x}_h^{**}(\boldsymbol{\mu}) \in \mathcal{X}_{ad,h} \quad : \quad \mathcal{L}'_h\left(\mathbf{x}_h^{**}(\boldsymbol{\mu});\boldsymbol{\mu}\right)\left(\boldsymbol{\phi}_h - \mathbf{x}_h^{**}(\boldsymbol{\mu})\right) \geq 0 \qquad \forall \boldsymbol{\phi}_h \in \mathcal{X}_{ad,h}, \tag{7.60}$$

where:

$$\begin{aligned}
\mathcal{L}'_h(\mathbf{x}_h(\boldsymbol{\mu});\boldsymbol{\mu})(\boldsymbol{\phi}_h) \quad := \quad & \Delta t \sum_{k=1}^{K} R_h^{pr,k}(p_h^k;\boldsymbol{\mu}) + m\left(v_h^0(\boldsymbol{\mu}) - v_{0h}(\boldsymbol{\mu}), p_h^0;\boldsymbol{\mu}\right) \\
& + \Delta t \sum_{k=1}^{K} R_h^{du,k}(w_h^k;\boldsymbol{\mu}) \\
& + \Delta t \sum_{k=1}^{K} R_h^{opt,k}(q_h^k;\boldsymbol{\mu}),
\end{aligned} \tag{7.61}$$

with $\boldsymbol{\phi}_h := \left\{p_h^0,\dots,p_h^K\right\} \times \left\{w_h^0,\dots,w_h^K\right\} \times \left\{q_h^1,\dots,q_h^K\right\} \in \mathcal{X}_{ad,h}$ and hence:

$$\begin{aligned}
\mathcal{L}'_h(\mathbf{x}_h^{**}(\boldsymbol{\mu});\boldsymbol{\mu})(\boldsymbol{\phi}_h) \quad = \quad & \Delta t \sum_{k=1}^{K} R_h^{pr,**,k}(p_h^k;\boldsymbol{\mu}) + m\left(v_h^{**,0}(\boldsymbol{\mu}) - v_{0h}(\boldsymbol{\mu}), p_h^0;\boldsymbol{\mu}\right) \\
& + \Delta t \sum_{k=1}^{K} R_h^{du,**,k}(w_h^k;\boldsymbol{\mu}) \\
& + \Delta t \sum_{k=1}^{K} R_h^{opt,**,k}(q_h^k;\boldsymbol{\mu}).
\end{aligned} \tag{7.62}$$

Similarly to what done in Sec.7.1.1 the FE approximated optimal control problem can be rewritten in matricial notation. In particular, for the FE primal problem, Eq.(7.12) holds simply by replacing g^k with $u_h^k(\boldsymbol{\mu})$; for the dual problem, we replace the vector $\mathbf{L}_h(\boldsymbol{\mu})$ with the term $M_h(\boldsymbol{\mu})\left(\mathbf{v}_h^k(\boldsymbol{\mu}) - \mathbf{v}_{dh}^k\right)$ in Eq.(7.13), being $M_h(\boldsymbol{\mu}) = \sum_{q=1}^{Q_d} \vartheta_q^d(\boldsymbol{\mu}) M_{dqh}$, with $\left(M_{dqh}\right)_{i,j} = m_{dq}(\varphi_j,\varphi_i)$, and $\mathbf{v}_{dh}^k$ s.t. $m_d\left(\varphi_i, v_{dh}^k;\boldsymbol{\mu}\right) = m_d\left(\varphi_i, v_d;\boldsymbol{\mu}\right)$ $\forall i = 1,\dots,N_h$. Similarly, the cost functional (7.46) is computed. The optimality conditions (7.52) and (7.53) are then rewritten in terms of the FE optimal residual, which reads from Eq.(7.56):

$$R_h^{opt,**,k}(\psi_h;\boldsymbol{\mu}) = \begin{cases} \gamma\left(u_h^{**,k}(\boldsymbol{\mu}) - u_d^k\right)\psi_h + \mathbf{B}_h(\boldsymbol{\mu})^T\mathbf{z}_h^{**,k}(\boldsymbol{\mu})\psi_h \\ \qquad\qquad\qquad \forall\psi_h \in \mathcal{U}_h,\ \forall k \in \{1,\dots,K-1\}, \\ \dfrac{1}{2}\gamma\left(u_h^{**,K}(\boldsymbol{\mu}) - u_d^K\right)\psi_h + \mathbf{B}_h(\boldsymbol{\mu})^T\mathbf{z}_h^{**,k}(\boldsymbol{\mu})\psi_h \\ \qquad\qquad\qquad \forall\psi_h \in \mathcal{U}_h,\ k = K. \end{cases} \tag{7.63}$$

7.2.2 RB approximation

The RB formulation for the solution of the parametrized optimal control problems mimics that presented in Sec.7.1.2 for parabolic PDEs. However, in this case, in order to use the Lagrangian formulation for the analysis of the optimal control problem, we require that the RB primal and dual spaces coincide, $\mathcal{Z}_N \equiv \mathcal{Z}_N^{pr} \equiv \mathcal{Z}_N^{du}$. In particular, the RB space $\mathcal{Z}_N$ is defined as:

$$\mathcal{Z}_N := \operatorname{span}\{\omega_i \ \ i = 1,\dots,N\}, \tag{7.64}$$

with the basis $\{\omega_i\}_{i=1}^N$ determined by an adaptive procedure. In particular, a possible choice consists in assuming $\mathcal{Z}_N$ as the space associated with the primal equation for $u_h^k(\boldsymbol{\mu}) = g^k$ i.e. the primal equation in a control–independent context. It follows that the space $\mathcal{Z}_N$ could be defined according with the adaptive procedure described in Sec.7.1.3 for parametrized parabolic PDEs.

However, we observe that this choice, even if simple and formally correct, does not allow an "optimal" selection the RB space, being related only to the features of the primal problem, as already highlighted in [149]. For this reason, in Sec.7.2.6 we propose an "integrated" RB approach for the selection of such "optimal" samples.

The analysis of the RB approximated optimal control problem is based, as for the FE one, on the "discretize–then–optimize" approach.

The RB primal problem reads:

$$\begin{aligned} &\text{find } v_N^k(\boldsymbol{\mu}) \in \mathcal{Z}_N \quad : \quad \Delta t\ R_N^{pr,k}(\phi_N;\boldsymbol{\mu}) = 0 \qquad \forall\phi_N \in \mathcal{Z}_N,\ \text{with } u_N^k(\boldsymbol{\mu}) \in \mathcal{U}_{ad,h},\ \forall k \in \mathbb{K} \\ &\text{with } v_N^0(\boldsymbol{\mu}) = v_{0h}(\boldsymbol{\mu}), \end{aligned} \tag{7.65}$$

where the RB primal residual is defined from Eq.(7.45) as:

$$R_N^{pr,k}(\phi_h;\boldsymbol{\mu}) := R^{pr,k}\left(v_N^k(\boldsymbol{\mu}), v_N^{k-1}(\boldsymbol{\mu}), u_N^k(\boldsymbol{\mu})\right)(\phi_h;\boldsymbol{\mu}) \qquad \forall\phi_h \in \mathcal{Z}_h,\ \forall k \in \mathbb{K}. \tag{7.66}$$

We observe that $u_N^k(\boldsymbol{\mu})$ indicates the control function corresponding to the RB approximation of the optimal control problem. Moreover, we have $\mathcal{U}_N \equiv \mathcal{U}_h$ and $\mathcal{U}_{ad,N} \equiv \mathcal{U}_{ad,h}$.

Starting from the FE approximation (7.46), the RB approximated cost functional (7.42) reads:

$$\begin{aligned}J_N(\mathbf{y}_N(\boldsymbol{\mu});\boldsymbol{\mu}) \;=\;& \frac{1}{2}\Delta t\left[\frac{1}{2}m_d\left(v_N^0(\boldsymbol{\mu})-v_d^0, v_N^0(\boldsymbol{\mu})-v_d^0;\boldsymbol{\mu}\right)\right.\\ &+\sum_{k=1}^{K-1} m_d\left(v_N^k(\boldsymbol{\mu})-v_d^k, v_N^k(\boldsymbol{\mu})-v_d^k;\boldsymbol{\mu}\right)\\ &\left.+\frac{1}{2}m_d\left(v_N^K(\boldsymbol{\mu})-v_d^K, v_N^K(\boldsymbol{\mu})-v_d^K;\boldsymbol{\mu}\right)\right]\\ &+\frac{1}{2}\gamma\Delta t\left[\sum_{k=1}^{K-1}\left(u_N^k(\boldsymbol{\mu})-u_d^k\right)^2+\frac{1}{2}\left(u_N^K(\boldsymbol{\mu})-u_d^K\right)^2\right],\end{aligned}\tag{7.67}$$

where $\mathbf{y}_N(\boldsymbol{\mu}) := \{v_N^0(\boldsymbol{\mu}),\dots,v_N^K(\boldsymbol{\mu})\}\times\{u_N^1(\boldsymbol{\mu}),\dots,u_N^K(\boldsymbol{\mu})\}$.
We observe that $\mathbf{x}_N(\boldsymbol{\mu})\in\mathcal{X}_{ad,N}$, with $\mathbf{x}_N(\boldsymbol{\mu}) := \left\{v_N^0(\boldsymbol{\mu}),\dots,v_N^K(\boldsymbol{\mu})\right\}\times\left\{z_N^0(\boldsymbol{\mu}),\dots,z_N^K(\boldsymbol{\mu})\right\}\times\left\{u_N^1(\boldsymbol{\mu}),\dots,u_N^K(\boldsymbol{\mu})\right\}$ and $\mathcal{X}_{ad,N} := (\mathcal{Z}_N)^{2(K+1)}\times(\mathcal{U}_{ad,h})^K$; if $\mathcal{U}_{ad,h}\equiv\mathcal{U}_h$, then $\mathcal{X}_{ad,N}\equiv\mathcal{X}_N$. By using the primal equation (7.65), the RB approximated parametrized Lagrangian functional follows from Eq.(7.48):

$$\begin{aligned}\mathcal{L}_h(\mathbf{x}_N(\boldsymbol{\mu});\boldsymbol{\mu}) \;:=\;& J_N(\mathbf{y}_N(\boldsymbol{\mu});\boldsymbol{\mu})+\Delta t\sum_{k=1}^{K}R_N^{pr,k}\left(z_N^k(\boldsymbol{\mu});\boldsymbol{\mu}\right)\\ &+m(v_N^0(\boldsymbol{\mu})-v_{0h}(\boldsymbol{\mu}),z_N^0(\boldsymbol{\mu});\boldsymbol{\mu}),\end{aligned}\tag{7.68}$$

where the FE approximated variables have just been replaced by the RB ones. We indicate with $\mathbf{x}_N^*(\boldsymbol{\mu}) := \left\{v_N^{*,0}(\boldsymbol{\mu}),\dots,v_N^{*,K}(\boldsymbol{\mu})\right\}\times\left\{z_N^{*,0}(\boldsymbol{\mu}),\dots,z_N^{*,K}(\boldsymbol{\mu})\right\}\times\left\{u_N^{*,1}(\boldsymbol{\mu}),\dots,u_N^{*,K}(\boldsymbol{\mu})\right\}$ the parameter dependent RB optimal solution, with $\mathbf{x}_N^*(\boldsymbol{\mu})\in\mathcal{X}_{ad,N}\subset\mathcal{X}_{ad,h}$. Then, the RB approximated cost functional $J_N^*(\boldsymbol{\mu})$ reads $J_N^*(\boldsymbol{\mu}) = J_N(\mathbf{y}_N^*(\boldsymbol{\mu});\boldsymbol{\mu})$, being $\mathbf{y}_N^*(\boldsymbol{\mu}) := \left\{v_N^{*,0}(\boldsymbol{\mu}),\dots,v_N^{*,K}(\boldsymbol{\mu})\right\}\times\left\{u_N^{*,1}(\boldsymbol{\mu}),\dots,u_N^{*,K}(\boldsymbol{\mu})\right\}$.
By differentiating $\mathcal{L}_h(\mathbf{x}_N;\boldsymbol{\mu})$ w.r.t. $v_N^k(\boldsymbol{\mu})$ $\forall k\in\mathbb{K}$, we obtain the RB dual equation:

$$\text{find } z_N^k(\boldsymbol{\mu})\in\mathcal{Z}_N \;:\; \Delta t\; R_N^{du,k}(\phi_N;\boldsymbol{\mu})=0 \qquad \forall\phi_N\in\mathcal{Z}_N,\ \text{with } v_N^k(\boldsymbol{\mu})\in\mathcal{Z}_N,\ \forall k\in\{K,\dots,1\},\tag{7.69}$$

where from Eq.(7.51):

$$R_N^{du,k}(\phi_h;\boldsymbol{\mu}) := R^{du,k}\left(z_N^k(\boldsymbol{\mu}),z_N^{k+1}(\boldsymbol{\mu}),v_N^k(\boldsymbol{\mu})\right)(\phi_h;\boldsymbol{\mu}) \qquad \forall\phi_h\in\mathcal{Z}_h,\ \forall k\in\{K,\dots,1\}.\tag{7.70}$$

Finally, by differentiating $\mathcal{L}_h(\mathbf{x}_N;\boldsymbol{\mu})$ w.r.t. $u_N^k(\boldsymbol{\mu})$ $\forall k\in\mathbb{K}$, we obtain the RB approximated optimality constraint, which in the unconstrained case is:

$$\Delta t\; R_N^{opt,*,k}(\psi_N;\boldsymbol{\mu})=0 \qquad \forall\psi_N\in\mathcal{U}_h,\ \forall k\in\mathbb{K},\tag{7.71}$$

while in the constrained one is:

$$\Delta t\; R_N^{opt,*,k}(\psi_N-u_N^{*,k}(\boldsymbol{\mu});\boldsymbol{\mu})\geq 0 \qquad \forall\psi_N\in\mathcal{U}_{ad,h},\ \forall k\in\mathbb{K}.\tag{7.72}$$

Starting from Eq.(7.56) the residual $R_N^{opt,*,k}(\cdot;\boldsymbol{\mu})$ is defined as:

$$R_N^{opt,*,k}(\psi_h;\boldsymbol{\mu}) := R^{opt,k}\left(z_N^{*,k}(\boldsymbol{\mu}),u_N^{*,k}(\boldsymbol{\mu})\right)(\psi_h;\boldsymbol{\mu}) \qquad \forall\psi_h\in\mathcal{U}_h,\ \forall k\in\mathbb{K},\tag{7.73}$$

and similarly:

$$R_N^{opt,k}(\psi_h;\boldsymbol{\mu}) := R^{opt,k}\left(z_N^k(\boldsymbol{\mu}), u_N^k(\boldsymbol{\mu})\right)(\psi_h;\boldsymbol{\mu}) \qquad \forall\psi_h \in \mathcal{U}_h,\ \forall k \in \mathbb{K}. \tag{7.74}$$

Moreover, we define from Eq.s (7.44) and (7.50) respectively, the following RB optimal residuals:

$$R_N^{pr,*,k}(\phi_h;\boldsymbol{\mu}) := R^{pr,k}\left(v_N^{*,k}(\boldsymbol{\mu}), v_N^{*,k-1}(\boldsymbol{\mu}), u_N^{*,k}(\boldsymbol{\mu})\right)(\phi_h;\boldsymbol{\mu}) \qquad \forall\phi_h \in \mathcal{Z}_h,\ \forall k \in \mathbb{K}, \tag{7.75}$$

$$R_N^{du,*,k}(\phi_h;\boldsymbol{\mu}) := R^{du,*,k}\left(z_N^{*,k}(\boldsymbol{\mu}), z_N^{*,k+1}(\boldsymbol{\mu}), v_N^{*,k}(\boldsymbol{\mu})\right)(\phi_h;\boldsymbol{\mu}) \qquad \forall\phi_h \in \mathcal{Z}_h,\ \forall k \in \{K,\dots,1\}. \tag{7.76}$$

In analogy with Eq.s (7.59) and (7.60), the first order necessary conditions for the RB problem read, in the unconstrained case:

$$\text{find } \mathbf{x}_N^*(\boldsymbol{\mu}) \in \mathcal{X}_N \quad : \quad \mathcal{L}_h'\left(\mathbf{x}_N^*(\boldsymbol{\mu});\boldsymbol{\mu}\right)(\boldsymbol{\phi}_N) = 0 \qquad \forall\boldsymbol{\phi}_N \in \mathcal{X}_N, \tag{7.77}$$

while in the constrained one:

$$\text{find } \mathbf{x}_N^*(\boldsymbol{\mu}) \in \mathcal{X}_{ad,N} \quad : \quad \mathcal{L}_h'\left(\mathbf{x}_N^*(\boldsymbol{\mu});\boldsymbol{\mu}\right)(\boldsymbol{\phi}_N - \mathbf{x}_N^*(\boldsymbol{\mu})) \geq 0 \qquad \forall\boldsymbol{\phi}_N \in \mathcal{X}_{ad,N}, \tag{7.78}$$

where $\mathcal{L}_h'(\cdot;\boldsymbol{\mu})(\cdot)$ is defined in Eq.(7.61) by replacing the FE residuals with the corresponding RB ones and $\boldsymbol{\phi}_N := \left\{p_N^0,\dots,p_N^K\right\} \times \left\{w_N^0,\dots,w_N^K\right\} \times \left\{q_N^1,\dots,q_N^K\right\} \in \mathcal{X}_{ad,N}$.

Similarly to Sec.s 7.1.2 and 7.2.1 the RB approximated optimal control problem can be rewritten in matricial notation. We observe that the superscripts "*pr*" and "*du*" are dropped off being $\mathcal{Z}_N^{pr} \equiv \mathcal{Z}_N^{du} \equiv \mathcal{Z}_N$. For the RB primal problem, Eq.(7.24) holds by replacing g^k with $u_N^k(\boldsymbol{\mu})$; for the dual problem in Eq.(7.25), the vector $\mathbf{L}_N^{du}(\boldsymbol{\mu})$ is replaced by the term $M_N(\boldsymbol{\mu})\left(\mathbf{v}_N^k(\boldsymbol{\mu}) - \mathbf{v}_{dN}^k\right)$, being $M_N(\boldsymbol{\mu}) = \sum_{q=1}^{Q_d} \vartheta_q^d(\boldsymbol{\mu}) M_{dqN}$, with $\left(M_{dqN}\right)_{i,j} = m_{dq}(\omega_j,\omega_i)$, and $\mathbf{v}_{dN}^k$ s.t. $m_d\left(\omega_i, v_{dN}^k;\boldsymbol{\mu}\right) = m_d\left(\omega_i, v_d;\boldsymbol{\mu}\right)$ $\forall i = 1,\dots,N$. In the same manner, the cost functional (7.67) is computed. Finally, the optimality conditions (7.72) and (7.73) are then rewritten in terms of the RB approximated optimal residual, which, from Eq.(7.73), is:

$$R_N^{opt,*,k}(\psi_h;\boldsymbol{\mu}) = \begin{cases} \gamma\left(u_N^{*,k}(\boldsymbol{\mu}) - u_d^k\right)\psi_h + \mathbf{B}_N(\boldsymbol{\mu})^T \mathbf{z}_N^{*,k}(\boldsymbol{\mu})\psi_h \\ \qquad\qquad \forall\psi_h \in \mathcal{U}_h,\ \forall k \in \{1,\dots,K-1\}, \\ \dfrac{1}{2}\gamma\left(u_N^{*,K}(\boldsymbol{\mu}) - u_d^K\right)\psi_h + \mathbf{B}_N(\boldsymbol{\mu})^T \mathbf{z}_N^{*,k}(\boldsymbol{\mu})\psi_h \\ \qquad\qquad \forall\psi_h \in \mathcal{U}_h,\ k = K. \end{cases} \tag{7.79}$$

7.2.3 The offline–online decomposition

As done in Sec.6.3.4 for parametrized steady PDEs and anticipated in Sec.7.1 for parabolic PDEs, also in the case of parametrized optimal control problems, the offline–online decomposition holds under the affine decomposition assumptions. We observe that such decomposition is valid both for the FE and the RB approximations and, in particular, it should be used also for the FE approximation in the many query or real time contexts.
The offline–online decomposition for parametrized optimal control problems can be schematized as in Fig.7.3.

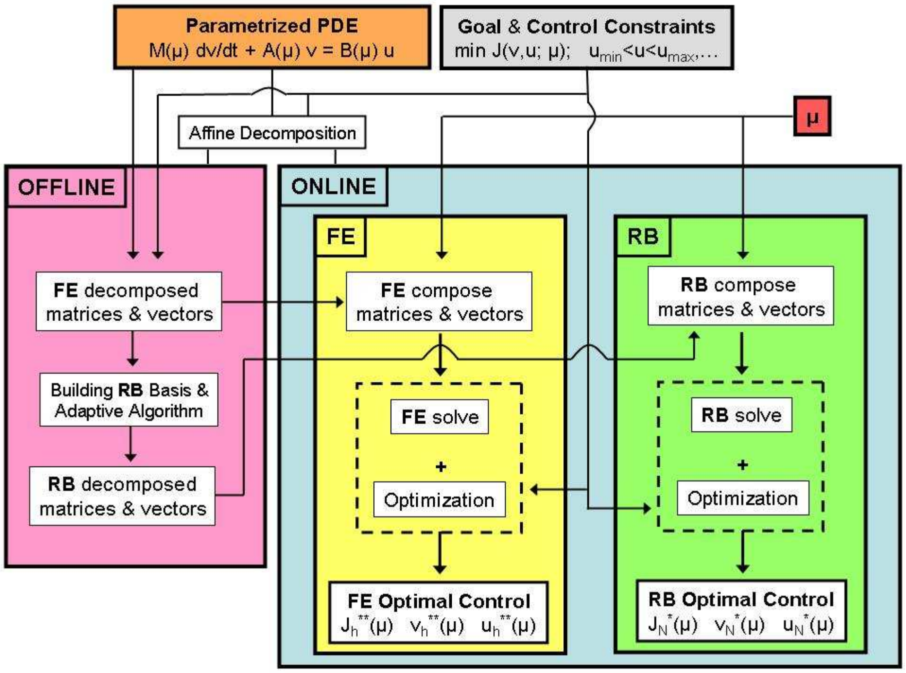
Parametrized PDE
M(μ) dv/dt + A(μ) v = B(μ) u
Goal & Control Constraints
min J(v,u; μ); u_min<u<u_max,...
μ
Affine Decomposition
OFFLINE
ONLINE
FE
RB
FE decomposed matrices & vectors
Building RB Basis & Adaptive Algorithm
RB decomposed matrices & vectors
FE compose matrices & vectors
FE solve
+
Optimization
FE Optimal Control
J_h**(μ) v_h**(μ) u_h**(μ)
RB compose matrices & vectors
RB solve
+
Optimization
RB Optimal Control
J_N*(μ) v_N*(μ) u_N*(μ)

Figure 7.3: Flow chart for the solution of parametrized optimal control problems; offline–online decomposition for either the FE or RB method.

In the offline step we approximate the problem according with the FE method and we assemble the decomposed FE matrices and vectors, see Sec.7.2.1; moreover, we assemble the decomposed RB matrices and vectors according with the RB method as described in Sec.7.2.2. An adaptive procedure, led by a posteriori error estimates, is used for the choice of the basis and the RB space $\mathcal{Z}_N$.

In the online step we solve the optimal control problem for a given value of the parameter vector $\boldsymbol{\mu} \in \mathcal{D}$, taking into account the control constraints. Either the FE and the RB methods could be considered, for which we use the matrices and vectors assembled in the offline step in order to compose the parameter dependent FE and RB matrices and vectors. Then, either the FE and RB approximated optimal control problems are solved by means of an optimization technique such as those described in Sec.2.3; in particular, we consider a quasi–Newton method in the unconstrained case, while a Sequential Quadratic Programming method in the constrained one. The FE and RB optimal solutions $\mathbf{x}_h^{**}(\boldsymbol{\mu})$ and $\mathbf{x}_N^{*}(\boldsymbol{\mu})$ are computed as well as the corresponding optimal cost functionals $J_h^{**}(\boldsymbol{\mu})$ and $J_N^{*}(\boldsymbol{\mu})$. It is immediate to observe that, being $N \ll N_h$ (with N the dimension of the RB space $\mathcal{Z}_N$ and N_h that of the FE one $\mathcal{Z}_h$), the RB method allows considerably savings in terms of computational costs w.r.t. that of the FE method. This fact is made more remarkable in the optimal control context, being the primal and dual problems solved repetitively in the course of the optimization procedure.

7.2.4 "Optimality" and "predictability" errors

As already discussed in Sec.4.2.5, while approximating an optimal control problem, we should care about the "optimality" and "predictability" errors. As anticipated, we assume the FE approximation as the "truth" one, hence our RB errors are evaluated w.r.t. the FE quantities. In particular $\forall \boldsymbol{\mu} \in \mathcal{D}$, the "optimality" error on the cost functional reads:

$$J_h^{**}(\boldsymbol{\mu}) - J_N^*(\boldsymbol{\mu}) = J_h\left(\mathbf{y}_h^{**}(\boldsymbol{\mu}); \boldsymbol{\mu}\right) - J_N\left(\mathbf{y}_N^*(\boldsymbol{\mu}); \boldsymbol{\mu}\right), \tag{7.80}$$

while the "predictability" one is:

$$J_h^*(\boldsymbol{\mu}) - J_N^*(\boldsymbol{\mu}) = J_h\left(\mathbf{y}_h^*(\boldsymbol{\mu}); \boldsymbol{\mu}\right) - J_N\left(\mathbf{y}_N^*(\boldsymbol{\mu}); \boldsymbol{\mu}\right), \tag{7.81}$$

where $\mathbf{y}_h^*(\boldsymbol{\mu}) := \left\{v_h^{*,0}(\boldsymbol{\mu}), \ldots, v_h^{*,K}(\boldsymbol{\mu})\right\} \times \left\{u_N^{*,1}(\boldsymbol{\mu}), \ldots, u_N^{*,K}(\boldsymbol{\mu})\right\}$, being $v_h^{*,k}(\boldsymbol{\mu})$ the solution of the FE primal problem (7.43) at the time step t_k obtained by enforcing $u_h^k(\boldsymbol{\mu}) = u_N^{*,k}(\boldsymbol{\mu})$ $\forall k \in \mathbb{K}$. Similar considerations hold while defining $\mathbf{x}_h^*(\boldsymbol{\mu}) := \left\{v_h^{*,0}(\boldsymbol{\mu}), \ldots, v_h^{*,K}(\boldsymbol{\mu})\right\} \times \left\{z_h^{*,0}(\boldsymbol{\mu}), \ldots, z_h^{*,K}(\boldsymbol{\mu})\right\} \times \left\{u_N^{*,1}(\boldsymbol{\mu}), \ldots, u_N^{*,K}(\boldsymbol{\mu})\right\}$, being $z_h^{*,k}(\boldsymbol{\mu})$ the FE solution of the dual problem (7.49) obtained by enforcing $v_h^k(\boldsymbol{\mu}) = v_h^{*,k}(\boldsymbol{\mu})$ $\forall k \in \mathbb{K}$. We recall that, while the "optimality" error refers to the capability of the RB method to detect an optimal solution $\mathbf{x}_N^*(\boldsymbol{\mu})$ able to approximate the FE one $\mathbf{x}_h^{**}(\boldsymbol{\mu})$, the "predictability" one assesses the capability of the RB method to approximate the FE solution for a prescribed control function, i.e. to bound the error $\mathbf{x}_h^*(\boldsymbol{\mu}) - \mathbf{x}_N^*(\boldsymbol{\mu})$.

By recalling Eq.(4.55), the following rough estimate of the "optimality" error on the cost functional holds:

$$J_h^{**}(\boldsymbol{\mu}) - J_N^*(\boldsymbol{\mu}) \leq J_h^*(\boldsymbol{\mu}) - J_N^*(\boldsymbol{\mu}), \tag{7.82}$$

which is based on the "predictability" one. This last error can be evaluated by means of an a posteriori RB error estimate.

7.2.5 A posteriori RB error estimate

In this Section we derive the a posteriori estimate for the "optimality" RB error; in particular, we assume that the error committed on the RB approximated optimal solution is evaluated by means of the error on the optimal functional cost (see Sec.4.2). With this aim, we use the recipes provided by the goal–oriented analysis in analogy with Sec.4.2; we recall that, in this case, the RB approximation is used and evaluated w.r.t. the FE one ("truth" approximation). The results reported in Sec.7.1.3 are thus extended to the case of a quadratic functional in the parametrized optimal control context. Both the cases of unconstrained and constrained optimal control problems are considered.

In view of the estimate for the "optimality" error $J_h^{**}(\boldsymbol{\mu}) - J_N^*(\boldsymbol{\mu})$, let us introduce the following RB errors on the optimal solution: $\mathbf{e}_N^*(\boldsymbol{\mu}) := \mathbf{x}_h^{**}(\boldsymbol{\mu}) - \mathbf{x}_N^*(\boldsymbol{\mu})$, $e_N^{pr,*,k}(\boldsymbol{\mu}) := v_h^{**,k}(\boldsymbol{\mu}) - v_N^{*,k}(\boldsymbol{\mu})$, $e_N^{du,*,k}(\boldsymbol{\mu}) := z_h^{**,k}(\boldsymbol{\mu}) - z_N^{*,k}(\boldsymbol{\mu})$ and $e_N^{opt,*,k}(\boldsymbol{\mu}) := u_h^{**,k}(\boldsymbol{\mu}) - u_N^{*,k}(\boldsymbol{\mu})$ $\forall k \in \mathbb{K}$. Then, the following Propositions hold.

Proposition 7.3. *For the optimal control problem defined in Eq.s (7.41) and (7.42) in the* unconstrained *case, the RB "optimality" error on the cost functional is:*

$$J_h^{**}(\boldsymbol{\mu}) - J_N^*(\boldsymbol{\mu}) = \frac{1}{2}\Delta t \sum_{k=1}^{K} \left[R_N^{pr,*,k}\left(e_N^{du,*,k}(\boldsymbol{\mu}); \boldsymbol{\mu}\right) + R_N^{du,*,k}\left(e_N^{pr,*,k}(\boldsymbol{\mu}); \boldsymbol{\mu}\right)\right], \tag{7.83}$$

being the RB optimal solution $\mathbf{x}_N^*(\boldsymbol{\mu}) \in \mathcal{X}_N \subset \mathcal{X}_h$ and the residuals defined in Eq.s (7.75) and (7.76).

Proof. By recalling the Proposition 4.2 and by observing that the optimal control problem under consideration is described by a quadratic cost functional and a linear PDE, we have:

$$J_h^{**}(\boldsymbol{\mu}) - J_N^*(\boldsymbol{\mu}) = \frac{1}{2}\mathcal{L}'\left(\mathbf{x}_N^*(\boldsymbol{\mu});\boldsymbol{\mu}\right)\left(\mathbf{e}_N^*(\boldsymbol{\mu})\right). \tag{7.84}$$

By recalling Eq.(7.62) and by introducing the residuals (7.73), (7.75) and (7.76), we obtain:

$$\begin{aligned} J_h^{**}(\boldsymbol{\mu}) - J_N^*(\boldsymbol{\mu}) &= \frac{1}{2}\Delta t \sum_{k=1}^{K}\Big[R_N^{pr,*,k}\left(e_N^{du,*,k}(\boldsymbol{\mu});\boldsymbol{\mu}\right) + R_N^{du,*,k}\left(e_N^{pr,*,k}(\boldsymbol{\mu});\boldsymbol{\mu}\right) \\ &\qquad + R_N^{opt,*,k}\left(e_N^{opt,*,k}(\boldsymbol{\mu});\boldsymbol{\mu}\right)\Big], \end{aligned} \tag{7.85}$$

being $m\left(e_N^{pr,*,0}(\boldsymbol{\mu}), z_N^{*,0}(\boldsymbol{\mu});\boldsymbol{\mu}\right) = 0$, due to $v_N^{*,0}(\boldsymbol{\mu}) = v_h^{**,0}(\boldsymbol{\mu}) = v_{0h}(\boldsymbol{\mu})$. Finally, by observing from Eq.(7.71) that $R_N^{opt,*,k}\left(e_N^{opt,*,k}(\boldsymbol{\mu});\boldsymbol{\mu}\right) = 0$, being $e_N^{du,*,k}(\boldsymbol{\mu}) \in \mathcal{U}_h$, the result (7.83) follows. □

Proposition 7.4. *For the optimal control problem defined in Eq.s (7.41) and (7.42) in the* constrained *case, the RB "optimality" error on the cost functional is estimated as:*

$$\begin{aligned} J_h^{**}(\boldsymbol{\mu}) - J_N^*(\boldsymbol{\mu}) &\leq \frac{1}{2}\Delta t \sum_{k=1}^{K}\Big[R_N^{pr,*,k}\left(e_N^{du,*,k}(\boldsymbol{\mu});\boldsymbol{\mu}\right) + R_N^{du,*,k}\left(e_N^{pr,*,k}(\boldsymbol{\mu});\boldsymbol{\mu}\right) \\ &\qquad + R_N^{opt,*,k}\left(e_N^{opt,*,k}(\boldsymbol{\mu});\boldsymbol{\mu}\right)\Big], \end{aligned} \tag{7.86}$$

being the RB optimal solution $\mathbf{x}_N^*(\boldsymbol{\mu}) \in \mathcal{X}_{ad,N} \subset \mathcal{X}_{ad,h}$ and the residuals defined in Eq.s (7.73), (7.75) and (7.76).

Proof. Result (7.86) follows from Proposition 4.2 in the constrained case and mimicking the proof of Proposition 7.3; the optimality condition (7.72) is used in the constrained case, for which $R_N^{opt,*,k}\left(e_N^{opt,*,k}(\boldsymbol{\mu});\boldsymbol{\mu}\right) \geq 0$, being $u_h^{**,k}(\boldsymbol{\mu}) \in \mathcal{U}_{ad,h}$. □

Let us introduce the continuity constant of the linear functional $b(\cdot;\boldsymbol{\mu})$, say $\lambda(\boldsymbol{\mu}) : \mathcal{D} \to \mathbb{R}$, for which we define its upper bound $\widetilde{\lambda}(\boldsymbol{\mu}) : \mathcal{D} \to \mathbb{R}$, s.t.:

$$\lambda(\boldsymbol{\mu}) \leq \widetilde{\lambda}(\boldsymbol{\mu}) \leq \lambda_0 \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{7.87}$$

with $\lambda_0 \in \mathbb{R}^+$. By recalling Eq.s (6.96) and (6.99), we indicate the dual norms of the optimal primal residual (7.75), the optimal dual residual (7.76) and the optimality residual (7.73) as, respectively:

$$\varepsilon_N^{pr,*,k}(\boldsymbol{\mu}) := \sup_{\phi_h \in \mathcal{Z}_h \setminus \{0\}} \frac{R_N^{pr,*,k}(\phi_h;\boldsymbol{\mu})}{||\phi_h||_{\mathcal{Z}_h}} \qquad \forall k \in \mathbb{K}, \tag{7.88}$$

$$\varepsilon_N^{du,*,k}(\boldsymbol{\mu}) := \sup_{\phi_h \in \mathcal{Z}_h \setminus \{0\}} \frac{R_N^{du,*,k}(\phi_h;\boldsymbol{\mu})}{||\phi_h||_{\mathcal{Z}_h}} \qquad \forall k \in \mathbb{K}, \tag{7.89}$$

$$\varepsilon_N^{opt,*,k}(\boldsymbol{\mu}) := \sup_{\psi_h \in \mathcal{U}_h \setminus \{0\}} \frac{R_N^{opt,*,k}(\psi_h;\boldsymbol{\mu})}{|\psi_h|_{\mathcal{U}_h}} \qquad \forall k \in \mathbb{K}. \tag{7.90}$$

In particular, we observe from Eq.(7.56) that:

$$\varepsilon_N^{opt,*,k}(\boldsymbol{\mu}) = \begin{cases} \gamma\left(u_N^{*,k}(\boldsymbol{\mu}) - u_d^k\right) + b\left(z_N^{*,k}(\boldsymbol{\mu});\boldsymbol{\mu}\right) & \forall k \in \{1,\dots,K-1\}, \\ \frac{1}{2}\gamma\left(u_N^{*,K}(\boldsymbol{\mu}) - u_d^K\right) + b\left(z_h^{*,K}(\boldsymbol{\mu});\boldsymbol{\mu}\right) & k = K. \end{cases} \tag{7.91}$$

These definitions allow us to introduce the following Propositions.

Proposition 7.5. *For the RB approximation of the optimal control problem defined in Eq.s (7.41) and (7.42) in the* unconstrained *case, the RB "optimality" error on the cost functional is estimated as:*

$$|J_h^{**}(\boldsymbol{\mu}) - J_N^*(\boldsymbol{\mu})| \leq \frac{1}{2}\left[\Delta_N^{pr,*,K}(\boldsymbol{\mu})\ |||e_N^{du,*,1}(\boldsymbol{\mu})|||^{du} + \Delta_N^{du,*,1}(\boldsymbol{\mu})\ |||e_N^{pr,*,K}(\boldsymbol{\mu})|||^{pr}\right], \tag{7.92}$$

where, by recalling Eq.s (7.88) and (7.89):

$$\Delta_N^{pr,*,k}(\boldsymbol{\mu}) := \left(\frac{\Delta t}{\widetilde{\alpha}(\boldsymbol{\mu})}\sum_{j=1}^{k}\left(\varepsilon_N^{pr,*,j}(\boldsymbol{\mu})\right)^2\right)^{1/2}, \tag{7.93}$$

$$\Delta_N^{du,*,k}(\boldsymbol{\mu}) := \left(\frac{\Delta t}{\widetilde{\alpha}(\boldsymbol{\mu})}\sum_{j=k}^{K}\left(\varepsilon_N^{du,*,j}(\boldsymbol{\mu})\right)^2\right)^{1/2}, \tag{7.94}$$

with $\widetilde{\alpha}(\boldsymbol{\mu})$ defined in Eq.(7.28), while the energy norms $|||\cdot|||^{pr}$ and $|||\cdot|||^{du}$ in Eq.s (7.31) and (7.32).

Proof. By recalling the result (7.83) of Proposition 7.3 we estimate the "optimality" error on the cost functional as:

$$\begin{aligned} |J_h^{**}(\boldsymbol{\mu}) - J_N^*(\boldsymbol{\mu})| \leq \frac{1}{2}\Bigg[&\left(\Delta t\sum_{k=1}^{K}\left(\varepsilon_N^{pr,*,k}(\boldsymbol{\mu})\right)^2\right)^{1/2}\left(\Delta t\sum_{k=1}^{K}||e_N^{du,*,k}(\boldsymbol{\mu})||^2_{\mathcal{Z}_h}\right)^{1/2} \\ &+\left(\Delta t\sum_{k=1}^{K}\left(\varepsilon_N^{du,*,k}(\boldsymbol{\mu})\right)^2\right)^{1/2}\left(\Delta t\sum_{k=1}^{K}||e_N^{pr,*,k}(\boldsymbol{\mu})||^2_{\mathcal{Z}_h}\right)^{1/2}\Bigg]. \end{aligned} \tag{7.95}$$

By using the coercivity property of the bilinear form $a(\cdot,\cdot;\boldsymbol{\mu})$, the definitions (7.31) and (7.32) and finally the positivity of the bilinear form $m(\cdot,\cdot;\boldsymbol{\mu})$, we have:

$$\begin{aligned} \left(\Delta t\sum_{k=1}^{K}||e_N^{pr,*,k}(\boldsymbol{\mu})||^2_{\mathcal{Z}_h}\right)^{1/2} &\leq \frac{1}{\widetilde{\alpha}(\boldsymbol{\mu})^{1/2}}\left(\Delta t\sum_{k=1}^{K}a\left(e_N^{pr,*,k}(\boldsymbol{\mu}),e_N^{pr,*,k}(\boldsymbol{\mu});\boldsymbol{\mu}\right)\right)^{1/2} \\ &\leq \frac{1}{\widetilde{\alpha}(\boldsymbol{\mu})^{1/2}}|||e_N^{pr,*,K}(\boldsymbol{\mu})|||^{pr}, \end{aligned} \tag{7.96}$$

$$\begin{aligned} \left(\Delta t\sum_{k=1}^{K}||e_N^{du,*,k}(\boldsymbol{\mu})||^2_{\mathcal{Z}_h}\right)^{1/2} &\leq \frac{1}{\widetilde{\alpha}(\boldsymbol{\mu})^{1/2}}\left(\Delta t\sum_{k=1}^{K}a\left(e_N^{du,*,k}(\boldsymbol{\mu}),e_N^{du,*,k}(\boldsymbol{\mu});\boldsymbol{\mu}\right)\right)^{1/2} \\ &\leq \frac{1}{\widetilde{\alpha}(\boldsymbol{\mu})^{1/2}}|||e_N^{du,*,1}(\boldsymbol{\mu})|||^{du}. \end{aligned} \tag{7.97}$$

Hence, by introducing the definitions (7.93) and (7.94), the result (7.92) follows. □

Proposition 7.6. *For the RB approximation of the optimal control problem defined in Eq.s (7.41) and (7.42) in the* constrained *case, the RB "optimality" error on the cost functional is estimated as:*

$$
\begin{aligned}
|J_h^{**}(\boldsymbol{\mu}) - J_N^{*}(\boldsymbol{\mu})| \;\; \leq \;\; & \frac{1}{2}\Bigg[\Delta_N^{pr,*,K}(\boldsymbol{\mu})\ |||e_N^{du,*,1}(\boldsymbol{\mu})|||^{du} + \Delta_N^{du,*,1}(\boldsymbol{\mu})\ |||e_N^{pr,*,K}(\boldsymbol{\mu})|||^{pr} \\
& +2\frac{\widetilde{\lambda}(\boldsymbol{\mu})}{\widetilde{\alpha}(\boldsymbol{\mu})^{1/2}\gamma^{1/2}}\Delta_N^{opt,*,K}(\boldsymbol{\mu})\ |||e_N^{du,*,1}(\boldsymbol{\mu})|||^{du}\Bigg],
\end{aligned}
\tag{7.98}
$$

where $\Delta_N^{pr,*,K}(\boldsymbol{\mu})$ *and* $\Delta_N^{du,*,1}(\boldsymbol{\mu})$ *are defined in Eq.s (7.93) and (7.94), and, from Eq.(7.90):*

$$
\Delta_N^{opt,*,k}(\boldsymbol{\mu}) := \left(\frac{\Delta t}{\gamma}\sum_{j=1}^{k}\left(\varepsilon_N^{opt,*,j}(\boldsymbol{\mu})\right)^2\right)^{1/2}, \tag{7.99}
$$

with $\widetilde{\alpha}(\boldsymbol{\mu})$ *defined in Eq.(7.28),* γ *in Eq.(7.42),* $\widetilde{\lambda}(\boldsymbol{\mu})$ *in Eq.(7.87), while the norms* $|||\cdot|||^{pr}$ *and* $|||\cdot|||^{du}$ *in Eq.s (7.31) and (7.32).*

Proof. The proof mimics that one of Proposition 7.5 by recalling the result (7.86) of Proposition 7.4. By recalling Eq.(7.91), the term related to the optimality error is bounded as:

$$
\Delta t\sum_{k=1}^{K} R_N^{opt,*,k}\left(e_N^{opt,*,k}(\boldsymbol{\mu});\boldsymbol{\mu}\right) \;\; \leq \;\; \left(\Delta t\sum_{k=1}^{K}\left(\varepsilon_N^{opt,*,k}(\boldsymbol{\mu})\right)^2\right)^{1/2}\left(\Delta t\sum_{k=1}^{K}|e_N^{opt,*,k}(\boldsymbol{\mu})|^2\right)^{1/2}. \tag{7.100}
$$

By recalling Eq.s (7.53) and (7.72) and by observing that $\mathcal{U}_{ad,N} = \mathcal{U}_{ad,h}$, we have:

$$
R_h^{opt,**,k}\left(-e_N^{opt,*,k}(\boldsymbol{\mu});\boldsymbol{\mu}\right) \geq 0, \tag{7.101}
$$

$$
R_N^{opt,*,k}\left(e_N^{opt,*,k}(\boldsymbol{\mu});\boldsymbol{\mu}\right) \geq 0. \tag{7.102}
$$

Then, from the definition (7.56) and by subtracting Eq.(7.101) from Eq.(7.102), we deduce that:

$$
\left[-\gamma e_N^{opt,*,k}(\boldsymbol{\mu}) - b\left(e_N^{du,*,k}(\boldsymbol{\mu});\boldsymbol{\mu}\right)\right] e_N^{opt,*,k}(\boldsymbol{\mu}) \geq 0 \qquad \forall k \in \{1,\dots,K-1\}, \tag{7.103}
$$

$$
\left[-\frac{1}{2}\gamma e_N^{opt,*,K}(\boldsymbol{\mu}) - b\left(e_N^{du,*,K}(\boldsymbol{\mu});\boldsymbol{\mu}\right)\right] e_N^{opt,*,K}(\boldsymbol{\mu}) \geq 0 \qquad k = K, \tag{7.104}
$$

and hence from Eq.(7.87):

$$
|e_N^{opt,*,k}(\boldsymbol{\mu})| \leq 2\frac{\widetilde{\lambda}(\boldsymbol{\mu})}{\gamma}||e_N^{du,*,k}(\boldsymbol{\mu})||_{\mathcal{Z}_h} \qquad \forall k \in \mathbb{K}. \tag{7.105}
$$

By replacing Eq.(7.105) into Eq.(7.100), by using the definition (7.99) and by recalling Eq.(7.97), the result (7.98) follows. □

Moreover, we observe that the estimates provided in Propositions 7.5 and 7.6 could not be immediately evaluated, being the "optimality" errors $e_N^{pr,*,K}(\boldsymbol{\mu})$ and $e_N^{du,*,1}(\boldsymbol{\mu})$ depending on the optimal FE primal and dual solutions $v_h^{pr,**,k}(\boldsymbol{\mu})$ and $z_h^{pr,**,k}(\boldsymbol{\mu})$. In order to make effective these estimates, we propose to replace these "optimality" errors with the corresponding "predictability" ones. With this aim, by recalling the definition of $\mathbf{x}_h^*(\boldsymbol{\mu})$ given in Sec.7.2.4, we define the following "predictability" primal and dual errors: $\widetilde{e}_N^{pr,*,k}(\boldsymbol{\mu}) := v_h^{*,k}(\boldsymbol{\mu}) - v_N^{*,k}$ and $\widetilde{e}_N^{du,*,k}(\boldsymbol{\mu}) := z_h^{*,k}(\boldsymbol{\mu}) - z_N^{*,k}$. On these basis, the following Theorems can be stated.

Theorem 7.2. *For the RB approximation of the optimal control problem defined in Eq.s (7.41) and (7.42) in the* unconstrained *case, the RB "optimality" error on the cost functional can be estimated by means of the following estimator:*

$$|J_h^{**}(\boldsymbol{\mu}) - J_N^*(\boldsymbol{\mu})| \lessapprox \Delta_N^{J,*,U}(\boldsymbol{\mu}) \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{7.106}$$

where, by recalling Eq.s (7.93) and (7.94):

$$\Delta_N^{J,*,U}(\boldsymbol{\mu}) := \Delta_N^{pr,*,K}(\boldsymbol{\mu})\ \Delta_N^{du,*,1}(\boldsymbol{\mu}). \tag{7.107}$$

Proof. The result (7.106) follows by recalling Proposition 7.5, by replacing the "optimality" errors $e_N^{pr,*,k}(\boldsymbol{\mu})$ and $e_N^{du,*,k}(\boldsymbol{\mu})$ with the corresponding "predictability" ones $\widetilde{e}_N^{pr,*,k}(\boldsymbol{\mu})$ and $\widetilde{e}_N^{du,*,k}(\boldsymbol{\mu})$ and by using the analogous of the estimates (7.33) and (7.35). □

Theorem 7.3. *For the RB approximation of the optimal control problem defined in Eq.s (7.41) and (7.42) in the* constrained *case, the RB "optimality" error on the cost functional can be estimated by means of the following estimator:*

$$|J_h^{**}(\boldsymbol{\mu}) - J_N^*(\boldsymbol{\mu})| \lessapprox \Delta_N^{J,*,C}(\boldsymbol{\mu}) \qquad \forall \boldsymbol{\mu} \in \mathcal{D}, \tag{7.108}$$

where, by recalling Eq.s (7.94), (7.99) and (7.107):

$$\Delta_N^{J,*,C}(\boldsymbol{\mu}) \ := \ \Delta_N^{J,*,U}(\boldsymbol{\mu}) + \frac{\widetilde{\lambda}(\boldsymbol{\mu})}{\widetilde{\alpha}(\boldsymbol{\mu})^{1/2}\gamma^{1/2}}\ \Delta_N^{opt,*,K}(\boldsymbol{\mu})\ \Delta_N^{du,*,1}(\boldsymbol{\mu}), \tag{7.109}$$

with $\widetilde{\alpha}(\boldsymbol{\mu})$ defined in Eq.(7.28), γ in Eq.(7.42) and $\widetilde{\lambda}(\boldsymbol{\mu})$ in Eq.(7.87).

Proof. The result (7.108) follows by recalling Proposition 7.6 and mimicking the proof of Proposition 7.2. □

Remark 7.1. *We observe that, while the estimators $\Delta_N^{pr,*,K}(\boldsymbol{\mu})$ (7.93) and $\Delta_N^{du,*,1}(\boldsymbol{\mu})$ (7.94) converge to zero as N increases, this is not the case, in general, of $\Delta_N^{opt,*,K}(\boldsymbol{\mu})$ (7.99). In fact, while $\Delta_N^{opt,*,K}(\boldsymbol{\mu}) = 0$ in the unconstrained case, in the constrained one we have $\Delta_N^{opt,*,K}(\boldsymbol{\mu}) \neq 0$ if the control constraints are activated. On this basis, we can infer that the convergence rate of the estimator $\Delta_N^{J,*,C}(\boldsymbol{\mu})$ (7.109) coincides with that of $\Delta_N^{J,*,U}(\boldsymbol{\mu})$ (7.107) when the control constraints are not activated, while resembles that of $\Delta_N^{du,*,1}(\boldsymbol{\mu})$ for N "sufficiently" large.*

In order to evaluate the effectivity of the estimates (7.106) and (7.108) we introduce, similarly to Eq.(6.110), the maximum and mean *effectivity indexes*:

$$\begin{aligned}
\overline{\eta}_N^{max} &:= \frac{\max\limits_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} \Delta_N^{J,*}(\boldsymbol{\mu})}{\max\limits_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} |J_h^{**}(\boldsymbol{\mu}) - J_N^{*}(\boldsymbol{\mu})|}, \\
\overline{\eta}_N^{mean} &:= \frac{\sum\limits_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} \Delta_N^{J,*}(\boldsymbol{\mu})}{\sum\limits_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} |J_h^{**}(\boldsymbol{\mu}) - J_N^{*}(\boldsymbol{\mu})|},
\end{aligned} \tag{7.110}$$

with $\Delta_N^{J,*}(\boldsymbol{\mu}) = \Delta_N^{J,*,U}(\boldsymbol{\mu})$ or $\Delta_N^{J,*,C}(\boldsymbol{\mu})$ depending on the estimator adopted. Moreover, we define the following indicators for the RB errors and estimators:

$$\begin{aligned}
E_N^{J,*,max} &:= \max_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} |J_h^{**}(\boldsymbol{\mu}) - J_N^{*}(\boldsymbol{\mu})|, \\
E_N^{J,*,mean} &:= \left(\sum_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} |J_h^{**}(\boldsymbol{\mu}) - J_N^{*}(\boldsymbol{\mu})| \right) / \overline{N},
\end{aligned} \tag{7.111}$$

$$\begin{aligned}
\Delta_N^{J,*,max} &:= \max_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} \Delta_N^{J,*}(\boldsymbol{\mu}), \\
\Delta_N^{J,*,mean} &:= \left(\sum_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} \Delta_N^{J,*}(\boldsymbol{\mu}) \right) / \overline{N},
\end{aligned} \tag{7.112}$$

being $\overline{N}$ the number of samples in $\overline{\mathcal{D}}$.

7.2.6 The "integrated" RB approach

By recalling Sec.7.2.2 and in particular Eq.(7.64), we need to define the RB space $\mathcal{Z}_N$. With this aim, we propose to assume $\mathcal{Z}_N$ as:

$$\mathcal{Z}_N = \mathcal{Z}_M^{pr,*} \cup \mathcal{Z}_M^{du,*}, \tag{7.113}$$

where:

$$\mathcal{Z}_M^{pr,*} := \left\{ \xi_i^{**} \quad i = 1, \ldots, M \right\}, \tag{7.114}$$

$$\mathcal{Z}_M^{du,*} := \left\{ \zeta_i^{**} \quad i = 1, \ldots, M \right\}, \tag{7.115}$$

being $\{\xi_i^{**}\}_{i=1}^M$ and $\{\zeta_i^{**}\}_{i=1}^M$ the *optimal* primal and dual *basis* obtained by solving the FE approximated parametrized optimal control problem for some values of the parameters vector. We observe that the optimal primal and dual basis are mutually dependent on each other, being ξ_i^{**} and ζ_i^{**} related to the FE optimal primal and dual solutions $v_h^{**,k}(\boldsymbol{\mu}_i)$ and $z_h^{**,k}(\boldsymbol{\mu}_i)$ for $\boldsymbol{\mu}_i \in \mathcal{D}$. It follows that the number of basis in the space $\mathcal{Z}_N$ is always even; in fact, each time we want to add a basis to the space $\mathcal{Z}_M^{pr,*}$, necessarily we add the corresponding one in $\mathcal{Z}_M^{du,*}$. For this reason, we indicate the RB space $\mathcal{Z}_N$ as $\mathcal{Z}_M^*$:

$$\mathcal{Z}_M^* := \mathcal{Z}_M^{pr,*} \cup \mathcal{Z}_M^{du,*}, \tag{7.116}$$

with:

$$\mathcal{Z}^*_M := \text{span}\,\{\omega_i^{**} \quad i = 1,\dots,2M\}\,, \tag{7.117}$$

being $\omega_{2i-1}^{**} = \xi_i^{**}$ and $\omega_{2i}^{**} = \zeta_i^{**}$ for $i = 1,\dots,M$.
The choice made for RB space $\mathcal{Z}^*_M$ allows to take into account of the features of both the optimal primal and dual solutions and, at the same time, to use the RB formulations and results provided in Sec.s 7.2.2 and 7.2.5. In particular, such analysis holds simply by replacing the subscript N with M. As usual, the space $\mathcal{Z}^*_M$ is determined according with an adaptive procedure for which the samples set $\mathcal{S}^*_M$ is selected, as we put into evidence in Sec.7.2.7.
We shall refer to this choice of the space $\mathcal{Z}^*_M$ as the *"integrated" RB approach* for parametrized optimal control problems.

7.2.7 The adaptive procedure

In this Section we report the adaptive procedure for the choice of the "integrated" RB basis and the corresponding space $\mathcal{Z}^*_M$. With this aim, we adapt the procedure reported in Sec.7.1.3 for the case of parametrized parabolic PDEs.
First of all we choose a first sample $\boldsymbol{\mu}_1^* \in \overline{\mathcal{D}}$ and we solve the corresponding FE approximated optimal control problem described in Sec.7.2.1, thus obtaining the FE primal and dual optimal solutions $\{v_h^{**,k}\}_{k=1}^K$ and $\{z_h^{**,k}\}_{k=1}^K$. The POD procedure is used to compute the first I_1 eigenmodes $\{\rho_i^{**}(\boldsymbol{\mu}_1^*)\}_{i=1}^{I_1}$ and $\{\varrho_i^{**}(\boldsymbol{\mu}_1^*)\}_{i=1}^{I_1}$ starting from $\{v_h^{**,k}(\boldsymbol{\mu}_1^*)\}_{k=1}^K$ and $\{z_h^{**,k}(\boldsymbol{\mu}_1^*)\}_{k=1}^K$, respectively. Then, we set $M=1$ and we define $\mathcal{Z}_M^{pr,*} = \{\xi_1^{**} = \rho_i^{**}(\boldsymbol{\mu}_1^*)\}$, $\mathcal{Z}_M^{du,*} = \{\zeta_1^{**} = \rho_i^{**}(\boldsymbol{\mu}_1^*)\}$, $\mathcal{Z}^*_M = \{\xi_1^{**}, \zeta_1^{**}\}$[1] and $\mathcal{S}^*_M = \{\boldsymbol{\mu}_1^*\}$; moreover, the following sets of snapshots are defined $\mathcal{Q}^{pr,**} = \{\rho_i^{**}(\boldsymbol{\mu}_1^*)\}_{i=1}^{I_1}$ and $\mathcal{Q}^{du,**} = \{\varrho_i^{**}(\boldsymbol{\mu}_1^*)\}_{i=1}^{I_1}$. The successive sample $\boldsymbol{\mu}_{M+1}^*$ is chosen as:

$$\boldsymbol{\mu}_{M+1}^* = \arg\max_{\boldsymbol{\mu}\in\overline{\mathcal{D}}} \ \Delta_N^{J,*}(\boldsymbol{\mu}), \tag{7.118}$$

where, by recalling the estimators (7.106) and (7.108), $\Delta_N^{J,*}(\boldsymbol{\mu}) = \Delta_N^{J,*,U}(\boldsymbol{\mu})$ or $\Delta_N^{J,*,C}(\boldsymbol{\mu})$. The FE approximated optimal control problem is solved for the parameters vector $\boldsymbol{\mu}_{M+1}^*$, thus obtaining $\{v_h^{**,k}(\boldsymbol{\mu}_{M+1}^*)\}_{k=1}^K$ and $\{z_h^{**,k}(\boldsymbol{\mu}_{M+1}^*)\}_{k=1}^K$, from which we compute the first I_{M+1} POD eigenmodes $\{\rho_i^{**}(\boldsymbol{\mu}_{M+1}^*)\}_{i=1}^{I_{M+1}}$ and $\{\varrho_i^{**}(\boldsymbol{\mu}_{M+1}^*)\}_{i=1}^{I_{M+1}}$, respectively. On this basis, the snapshots sets $\mathcal{Q}^{pr,**}$ and $\mathcal{Q}^{du,**}$ are then enriched as $\mathcal{Q}^{pr,**} = \mathcal{Q}^{pr,**} \cup \{\rho_i^{**}(\boldsymbol{\mu}_{M+1}^*)\}_{i=1}^{I_{M+1}}$ and $\mathcal{Q}^{du,**} = \mathcal{Q}^{du,**} \cup \{\varrho_i^{**}(\boldsymbol{\mu}_{M+1}^*)\}_{i=1}^{I_{M+1}}$, respectively. The POD procedure is used again in order to compute the basis functions $\{\xi_i^{**}\}_{i=1}^{M+1}$ and $\{\zeta_i^{**}\}_{i=1}^{M+1}$ starting from the snapshots sets $\mathcal{Q}^{pr,**}$ and $\mathcal{Q}^{du,**}$. Then, we define $\mathcal{Z}_{M+1}^{pr,*} = \text{span}\{\xi_1^{**},\dots,\xi_{M+1}^{**}\}$, $\mathcal{Z}_{M+1}^{du,*} = \text{span}\{\zeta_1^{**},\dots,\zeta_{M+1}^{**}\}$, $\mathcal{Z}^*_{M+1} = \text{span}\{\xi_1^{**},\zeta_1^{**},\dots,\xi_{M+1}^{**},\zeta_{M+1}^{**}\}$ and $\mathcal{S}^*_{M+1} = \{\boldsymbol{\mu}_1^*,\dots,\boldsymbol{\mu}_{M+1}^*\}$. Finally, we set $M = M+1$ and we repeat the entire procedure until $\max_{k\in\mathbb{K}} \Delta_M^{J,*}(\boldsymbol{\mu}) < tol$, being *tol* a prescribed tolerance.

7.2.8 Numerical tests

In this Section we provide some numerical tests proving the effectiveness of the RB method for the solution of parametrized optimal control problems. In particular, after having considered

[1] In order to avoid numerical ill–conditioning, the Gram–Schmidt orthonormalization is used, see Remark 6.4.

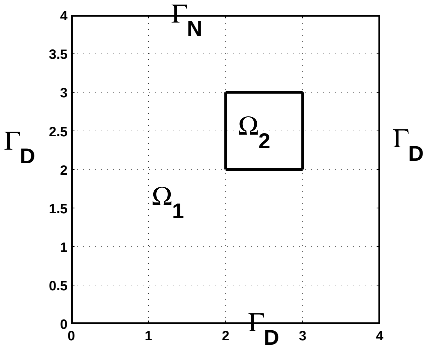

Figure 7.4: Test $C1$. Computational domain.

two academic tests cases (Tests $C1a$ and $C1b$), both unconstrained and constrained, we provide a numerical test inspired by an environmental application (Test $C2$).

For the RB approximation we use the "integrated" RB approach provided in Sec.7.2.6, for which the definition of the RB space $\mathcal{Z}^*_M$ is based on the adaptive procedure described in Sec.7.2.7 and led by the a posteriori RB error estimates on the cost functional proposed in Theorems 7.2 and 7.3.

The numerical simulations are carried out by means of an expanded module based on the *rbMIT©MIT Software* [193] developed in collaboration with the Authors. An Intel®Pentium®M 1.60 *GHz* processor, with 2 *MB* of Memory Cache and 512 *MB* of RAM has been used in order to carry out the computations.

Test $C1$

From Eq.(7.41), we consider the following parametrized parabolic PDE:

$$\begin{cases} \dfrac{\partial v(t,\boldsymbol{\mu})}{\partial t} - \nabla\cdot(\nu(\boldsymbol{\mu})\nabla v(t,\boldsymbol{\mu})) = u(t)f(\boldsymbol{\mu}) & \text{in } \Omega(\boldsymbol{\mu}),\ t\in(0,T), \\ v(t,\boldsymbol{\mu}) = 0 & \text{on } \Gamma_D,\ t\in(0,T), \\ \nu(\boldsymbol{\mu})\nabla v(t,\boldsymbol{\mu})\cdot\hat{\mathbf{n}} = 0 & \text{on } \Gamma_N,\ t\in(0,T), \\ v(0,\boldsymbol{\mu}) = 0 & \text{on } \Omega(\boldsymbol{\mu}), \end{cases} \tag{7.119}$$

where the domain $\Omega(\boldsymbol{\mu}) = \Omega_1(\boldsymbol{\mu})\cup\Omega_2(\boldsymbol{\mu})$ is reported in Fig.7.4, the tensor $\nu(\boldsymbol{\mu})\in[L^\infty(\Omega)]^{2\times 2}$, $f(\boldsymbol{\mu})\in L^2(\Omega)\ \forall\boldsymbol{\mu}\in\mathcal{D}$ and $\hat{\mathbf{n}}$ is the outward directed unit vector normal to $\Gamma_N := \partial\Omega\backslash\Gamma_D$. From Eq.(7.42), the cost functional is:

$$\begin{aligned} J(v(\boldsymbol{\mu}),u(\boldsymbol{\mu});\boldsymbol{\mu}) \;=\; & \frac{1}{2}\int_0^T\int_{\Omega_2(\boldsymbol{\mu})} (v(t,\boldsymbol{\mu})-v_d(t))^2\ d\Omega_2(\boldsymbol{\mu})\ dt \\ & +\frac{1}{2}\gamma\int_0^T (u(t,\boldsymbol{\mu})-u_d(t))^2\ dt. \end{aligned} \tag{7.120}$$

We consider two tests problems: one, say Test $C1a$, with a single parametric dependence, for which $\boldsymbol{\mu}=\mu\in\mathcal{D}\subset\mathbb{R}$; the other, Test $C1b$, with a multiple parametric dependence, precisely $\boldsymbol{\mu}\in\mathcal{D}\subset\mathbb{R}^4$. In this last case a geometrical parametrization is also taken into account. For each of these tests, both the unconstrained and constrained cases are analyzed.

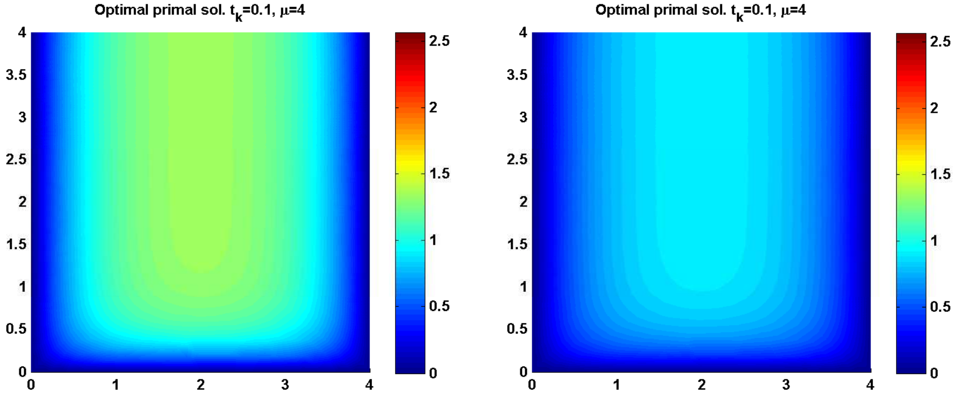

Figure 7.5: Test $C1a$. FE optimal primal solutions $v_h^{**,k}(\mu_s)$ at $t_k = 0.1$ for $\mu_s = 4$ in the unconstrained (left) and constrained (right) cases.

As usual, for the FE approximation we use piecewise linear basis functions defined on a mesh composed by $1,440$ triangular elements and 761 nodes. For the temporal discretization we consider $\Delta t = 0.1$ with $K = 10$, being $T = 1$. Moreover, in the constrained case, we have $\mathcal{U}_{ad,h} = \{u_h^k(\boldsymbol{\mu}) \in \mathbb{R} \ : \ u_{min} \leq u_h^k(\boldsymbol{\mu}) \leq u_{max} \ \forall k \in \mathbb{K} \text{ and } \Delta t \left(\sum_{k=1}^{K-1} u_h^k(\boldsymbol{\mu}) + \frac{1}{2} u_h^K(\boldsymbol{\mu})\right) \leq T\overline{U} \ \ \forall \boldsymbol{\mu} \in \mathcal{D}\}$.

For the adaptive algorithm described in Sec.7.2.7 we choose $tol = 10^{-2}$ and the subset $\overline{\mathcal{D}} \subset \mathcal{D}$ composed by $\overline{N} = 1,000$ samples randomly distributed in $\mathcal{D}$.

Finally, for both the test cases, we assume $\gamma = 1$, $v_d = 2$, $u_d = 0.9$, $u_{min} = 0.5$, $u_{max} = 1$ and $\overline{U} = 0.85$.

Test $C1a$. In this case, we assume $\nu(\mu) = \begin{bmatrix} \mu & 0 \\ 0 & 1 \end{bmatrix}$; moreover, the source term $f = 10$ is parameter independent as well as the domain Ω and subdomains Ω_1, Ω_2 (see Fig.7.4). Finally, we set $\mathcal{D} = [1, 10]$. We obtain $Q_a = 2$, $Q_m = 1$, $Q_b = 1$, $Q_l = 1$ and $Q_v = 1$.

In Fig.7.5 we report the FE optimal primal solutions $v_h^{**,K}(\mu_s)$ at the time step $t_k = 0.1$ for $\mu_s = 4$ in the unconstrained (left) and constrained (right) cases, for which the corresponding FE optimal cost functionals are $J_h^{**}(\mu_s) = 0.1169$ and $J_h^{**}(\mu_s) = 0.1405$, respectively; similarly, the computational costs associated with the FE solution of such optimal control problems are 13.7 s and 9.79 s.

By considering now the RB method in the unconstrained case, we see that the adaptive procedure selects $M = 7$ optimal basis for the definition of RB space $\mathcal{Z}_M^*$; we recall that the number of elements which composes the basis is $N = 2M = 14$. In particular, at the online step, we obtain for $M = 2$ that $|J_h^{**}(\mu_s) - J_M^*(\mu_s)| = 4.758 \cdot 10^{-3}$, which is bounded by the estimator $\Delta_M^{J,*,U}(\mu_s) = 7.780 \cdot 10^{-2}$; similarly, for $M = 6$ we have $|J_h^{**}(\mu_s) - J_M^*(\mu_s)| = 4.066 \cdot 10^{-5}$ bounded by $\Delta_M^{J,*,U}(\mu_s) = 3.926 \cdot 10^{-4}$. The online computational costs associated with the RB approximation of the parametrized optimal control problem are 0.110 s for $M = 2$ and 0.140 s for $M = 6$, thus allowing a remarkable saving w.r.t. to the FE simulation (13.7 s).

In the constrained case, the adaptive procedure selects $M = 23$ optimal basis ($N = 2M = 46$) for the RB space $\mathcal{Z}_M^*$. At the online step, we obtain $|J_h^{**}(\mu_s) - J_M^*(\mu_s)| = 1.490 \cdot 10^{-3}$ for

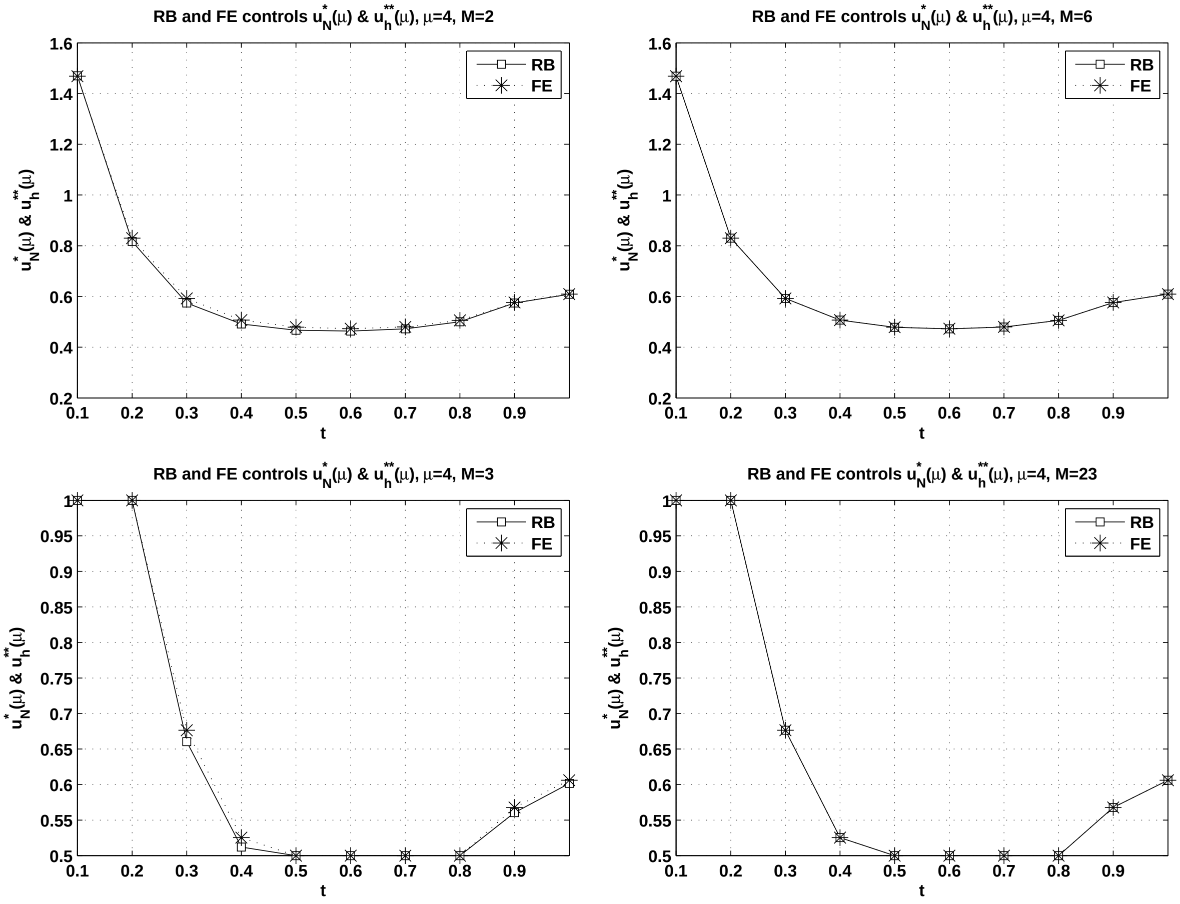

Figure 7.6: Test $C1a$. FE ($*$) and RB ($\Box$) optimal control solutions $u_h^{**,k}(\mu_s)$ and $u_M^{*,k}(\mu_s)$ for $\mu_s = 4$. Unconstrained case (top) with RB optimal controls for $M = 2$ (top–left) and $M = 6$ (top–right); constrained case (bottom) with RB optimal controls for $M = 3$ (bottom–left) and $M = 23$ (bottom–right).

$M = 3$, bounded by $\Delta_M^{J,*,C}(\mu_s) = 5.200 \cdot 10^{-1}$, while, for $M = 23$ we have $|J_h^{**}(\mu_s) - J_M^*(\mu_s)| = 3.493 \cdot 10^{-10}$ and $\Delta_M^{J,*,C}(\mu_s) = 1.313 \cdot 10^{-3}$; the corresponding computational costs are 0.146 s and 0.219 s, respectively, less than the 9.79 s required by the FE approximation.

In Fig.7.6 we provide a comparison of the FE $u_h^{**,k}(\mu_s)$ and $u_M^{*,k}(\mu_s)$ optimal control functions for both the unconstrained and constrained cases. In particular, we notice that, already for "small" M (dimension of the RB space $\mathcal{Z}_M^*$), the RB optimal controls are able to approximate "sufficiently" well the corresponding FE ones.

In Fig.7.7 we compare the RB "optimality" errors on the cost functional $E_M^{J,*,max}$ and $E_M^{J,*,mean}$ with the corresponding estimators $\Delta_M^{J,*,max}$ and $\Delta_M^{J,*,mean}$ for the unconstrained and constrained cases; see Eq.s (7.111) and (7.112). In particular, we highlight their behaviors vs. the number of RB basis M, for which we notice that the true RB errors are effectively bounded by the corresponding RB estimators. The validity of the estimators proposed in Theorems 7.2 and 7.3 is thus verified. Moreover, we observe that in the unconstrained case the convergence rates yielded by the RB estimators and errors are about the same. However, this is not the case of the constrained problem, being, as already discussed in Remark 7.1, $\Delta_M^{opt,*,K}(\boldsymbol{\mu}) \neq 0$

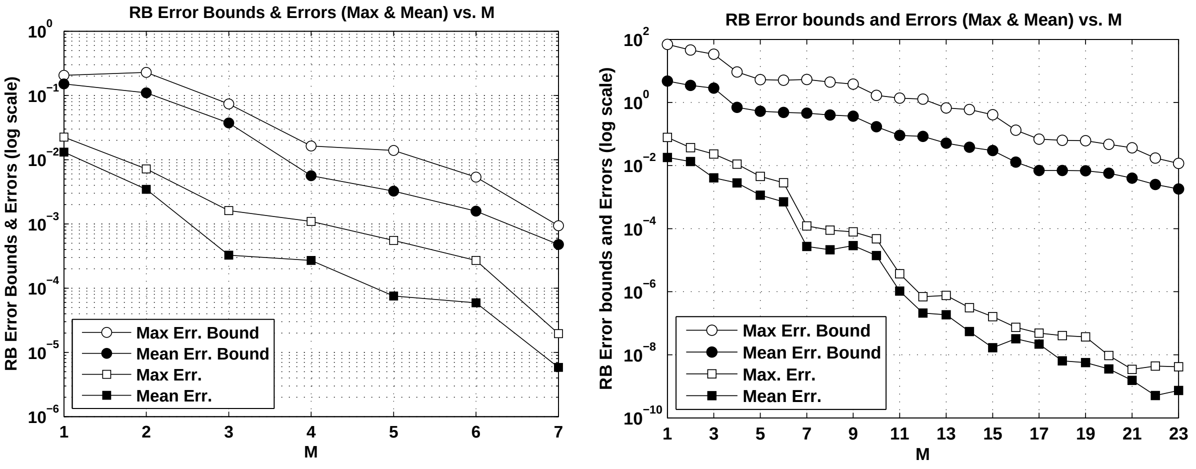

Figure 7.7: Test $C1a$. Error bounds $\Delta_M^{J,*,max}$ (○), $\Delta_M^{J,*,mean}$ (●) and errors $E_M^{J,*,max}$ (□), $E_M^{J,*,mean}$ (■) vs. M; unconstrained (left) and constrained (right) cases. Logarithmic scale on error axis.

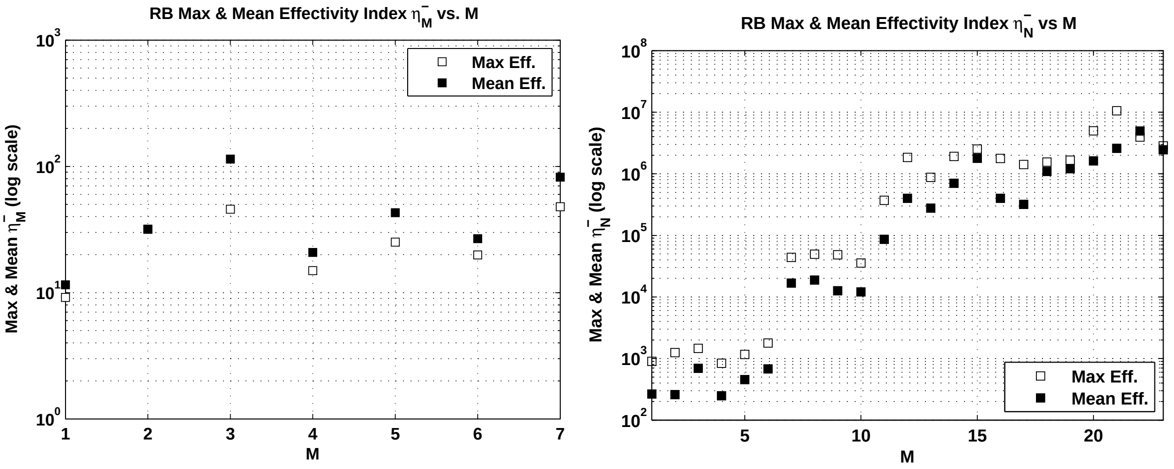

Figure 7.8: Test $C1a$. Effectivity indexes $\overline{\eta}_M^{max}$ (□), $\overline{\eta}_M^{mean}$ (■) vs. M; unconstrained (left) and constrained (right) cases. Logarithmic scale on effectivity index axis.

when the control constraints activated. This motivates also the fact for which, in the constrained case, the dimension M of the RB space $\mathcal{Z}_M^*$, defined at the offline step by the adaptive procedure, is much greater than that of the unconstrained problem. This reflects also on the effectivity indexes $\overline{\eta}_M^{max}$ and $\overline{\eta}_M^{mean}$ (7.110), as highlighted in Fig.7.8. In particular, we observe that for the unconstrained problem, the effectivity indexes evolve in the range $[10^2, 10^3]$ for $M = 1, \dots, 7$. On the contrary, in the constrained case, η_M^{max} and η_M^{mean} considerably increase as M varies from 1 to 43, due to the different convergence rates of the RB estimators and errors.

Finally, in Fig.7.9 we show the behaviors of the errors $E_M^{J,*,max}$ and $E_M^{J,*,mean}$ vs. the RB

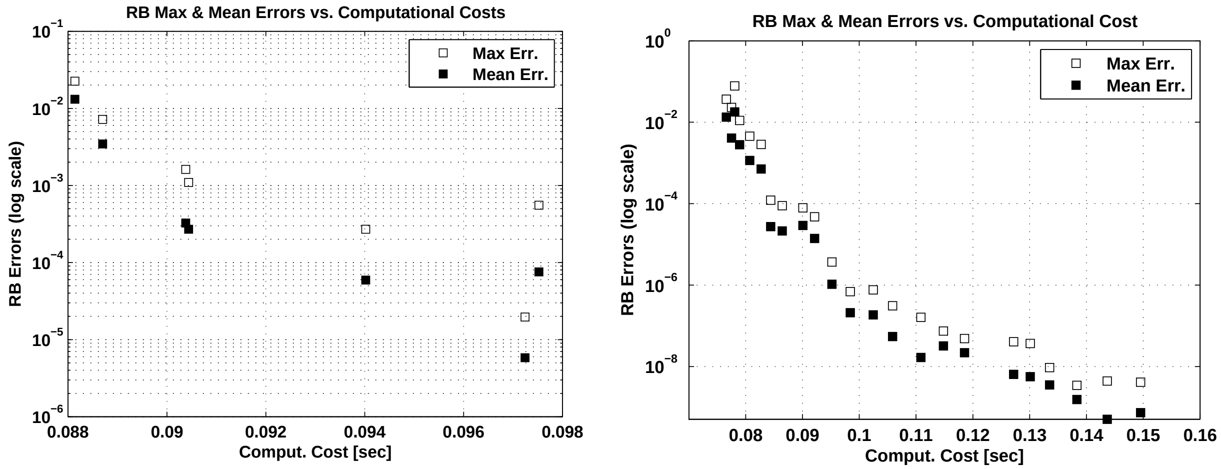

Figure 7.9: Test $C1a$. Errors $E_M^{J,*,max}$ (□) and $E_M^{J,*,mean}$ (■) vs. mean RB computational costs (seconds); unconstrained (left) and constrained (right) cases. Logarithmic scale on error axis.

online computational costs (mean values over $\overline{\mathcal{D}}$ for each M). As expected, the RB errors reduce as the computational costs increase. We remark that the RB computational costs are really competitive w.r.t. those required by the FE approximations. In fact, by solving the FE approximated optimal control problem for each of the $\overline{N}$ samples of $\overline{\mathcal{D}}$, we see that corresponding mean computational costs are 12.3 s in the unconstrained case and 9.04 s in the constrained one, while the RB costs are always smaller than 0.098 s and 0.15 s, respectively.

Test $C1b$. We consider now the multi parametric case, for which $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_4] \in \mathcal{D} \subset \mathbb{R}^4$. In this case, we assume $\nu(\boldsymbol{\mu}) = \begin{bmatrix} \mu_1 & 0 \\ 0 & 1 \end{bmatrix} \chi_1(\mu_3) + \begin{bmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{bmatrix} \chi_2(\mu_3)$ and $f(\boldsymbol{\mu}) = 10\ \chi_1(\mu_3) + \mu_4\ \chi_2(\mu_3)$, being $\chi_i(\boldsymbol{\mu})$ the characteristic functions of the parametrized subdomains $\Omega_i(\boldsymbol{\mu})$ for $i = 1, 2$ (see Fig.7.4). In particular, we choose $\Omega_2(\boldsymbol{\mu}) = (\mu_3, 1 + \mu_3)^2$, for which, being $\Omega(\boldsymbol{\mu}) = (0, 4)^2\ \forall \boldsymbol{\mu}$, we can deduce $\Omega_1(\boldsymbol{\mu}) = \Omega_1(\mu_3)$. Finally, we set $\mathcal{D} = [1, 10] \times [1, 10] \times [1, 2] \times [1, 10]$. In this case we have $Q_a = 14$, $Q_m = 2$, $Q_b = 3$, $Q_l = 1$ and $Q_v = 1$.

In Fig.7.10 we highlight the FE optimal primal solutions $v_h^{**,K}(\mu_s)$ at the final time step $t_K = T = 1$ for $\boldsymbol{\mu}_s = [10, 8, 2, 1]$ in the unconstrained (left) and constrained (right) cases. We notice that these solutions show a similar behavior even if the values are scaled. The computational costs associated with the FE solution of such optimal control problems are 4.10 s and 3.69 s for the unconstrained and constrained cases, respectively; the corresponding FE optimal cost functionals are $J_h^{**}(\boldsymbol{\mu}_s) = 0.4197$ and $J_h^{**}(\boldsymbol{\mu}_s) = 0.4675$.

Let us consider now the RB method in the unconstrained case, for which the adaptive procedure selects $M = 24$ elements ($N = 2M = 48$) for the definition of the RB space $\mathcal{Z}_M^*$. We observe that, at the online step, we obtain for $M = 3$ that $|J_h^{**}(\boldsymbol{\mu}_s) - J_M^*(\boldsymbol{\mu}_s)| = 1.233 \cdot 10^{-2}$, which is bounded by the estimator $\Delta_M^{J,*,U}(\boldsymbol{\mu}_s) = 1.7235$. For $M = 24$ we obtain $|J_h^{**}(\boldsymbol{\mu}_s) - J_M^*(\boldsymbol{\mu}_s)| = 5.132 \cdot 10^{-5}$, bounded by $\Delta_M^{J,*,U}(\boldsymbol{\mu}_s) = 1.97 \cdot 10^{-2}$. For the RB approximation of the parametrized optimal control problem we obtain the following computational costs: 0.0317 s and 0.0781 s for $M = 3$ and $M = 24$, respectively.

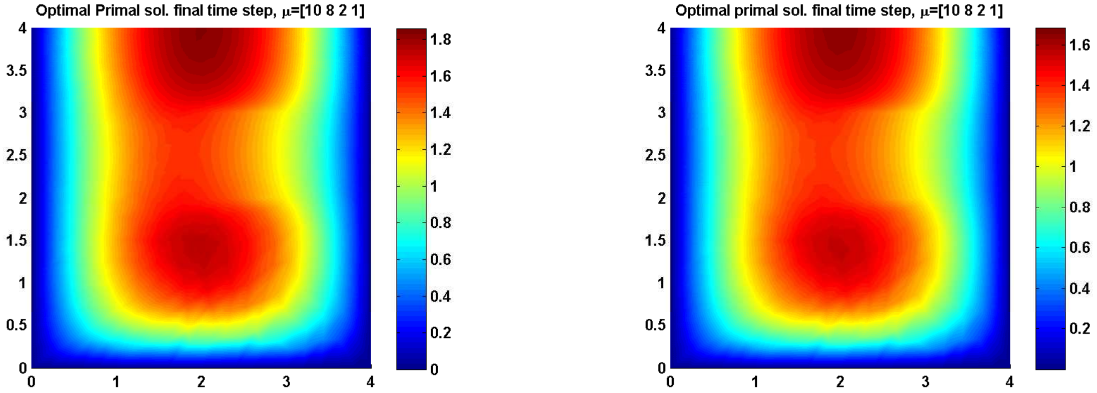

Figure 7.10: Test $C1b$. FE optimal primal solutions $v_h^{**,K}(\boldsymbol{\mu}_s)$ at $t_k = t_K = 1$ for $\boldsymbol{\mu}_s = [10, 8, 2, 1]$ in the unconstrained (left) and constrained (right) cases.

In the constrained case, the adaptive sampling process fixes $M = 44$ optimal basis ($N = 2M = 88$) for the definition of the RB space $\mathcal{Z}_M^*$. At the online step, we obtain $|J_h^{**}(\boldsymbol{\mu}_s) - J_M^*(\boldsymbol{\mu}_s)| = 3.102 \cdot 10^{-2}$ for $M = 3$, which is bounded by $\Delta_M^{J,*,C}(\boldsymbol{\mu}_s) = 7.844 \cdot 10^{-1}$; for $M = 44$ we have $|J_h^{**}(\boldsymbol{\mu}_s) - J_M^*(\boldsymbol{\mu}_s)| = 8.440 \cdot 10^{-6}$, bounded by $\Delta_M^{J,*,C}(\boldsymbol{\mu}_s) = 7.721 \cdot 10^{-3}$. The corresponding RB computational costs are 0.0780 s and 0.281 s, respectively.

We notice that, for both the unconstrained and constrained cases, the dimensions M required by the adaptive procedure in order to set the RB space $\mathcal{Z}_M^*$ is larger than in the corresponding cases of Test $1a$. This is due to the multiple parametric dependence of Test $1b$, being $\boldsymbol{\mu} \in \mathcal{D} \subset \mathbb{R}^4$, instead of $\mu \in \mathcal{D} \subset \mathbb{R}$ for Test $C1a$.

In Fig.7.11 we compare the FE $u_h^{**,k}(\boldsymbol{\mu}_s)$ and $u_M^{*,k}\boldsymbol{\mu}_s)$ optimal control functions for the unconstrained and constrained cases. In analogy with Test $C1a$, we notice that, even if M is "small", the RB optimal control functions approximate "sufficiently" well the corresponding FE ones.

In Fig.7.12 we provide the comparison of the RB "optimality" errors on the cost functional $E_M^{J,*,max}$ and $E_M^{J,*,mean}$ with the estimators $\Delta_M^{J,*,max}$ and $\Delta_M^{J,*,mean}$ for the unconstrained and constrained cases. We observe that $E_M^{J,*,max}$ and $E_M^{J,*,mean}$ are bounded by the corresponding estimators $\Delta_M^{J,*,max}$ and $\Delta_M^{J,*,mean}$, thus proving the reliability of the estimators provided in Theorems 7.2 and 7.3. As already discussed for Test $1a$, we notice that in the unconstrained case the convergence rates in M of the RB estimators and errors are about the same. On the contrary, for the constrained problem, the convergence rates of $E_M^{J,*,max}$ and $E_M^{J,*,mean}$ are greater than those of $\Delta_M^{J,*,max}$ and $\Delta_M^{J,*,mean}$, being in general $\Delta_M^{opt,*,K}(\boldsymbol{\mu}) \neq 0$ and the convergence rate of $\Delta_M^{J,*,C}(\boldsymbol{\mu})$ limited by that of $\Delta_M^{du,*,1}(\boldsymbol{\mu})$. This also motivates the fact that the dimension M selected by the adaptive procedure is much greater in the constrained case than in the unconstrained one. In Fig.7.13 we report the behaviors of the effectivity indexes $\overline{\eta}_M^{max}$ and $\overline{\eta}_M^{mean}$ vs. M. We observe that, in the unconstrained case, the effectivity indexes vary in the range $[10, 10^3]$ for $M = 1, \dots, 24$; specifically $\overline{\eta}_M^{mean}$ assumes values in the range $[10, 10^2]$. However, due to the previous considerations, $\overline{\eta}_M^{max}$ and $\overline{\eta}_M^{mean}$ increase with M in the constrained case; in particular, $\overline{\eta}_M^{mean}$ varies from about 20 for $M = 2$ to about $2,000$ for $M = 44$.

In Fig.7.14 we report the behaviors of the errors $E_M^{J,*,max}$ and $E_M^{J,*,mean}$ vs. the RB online

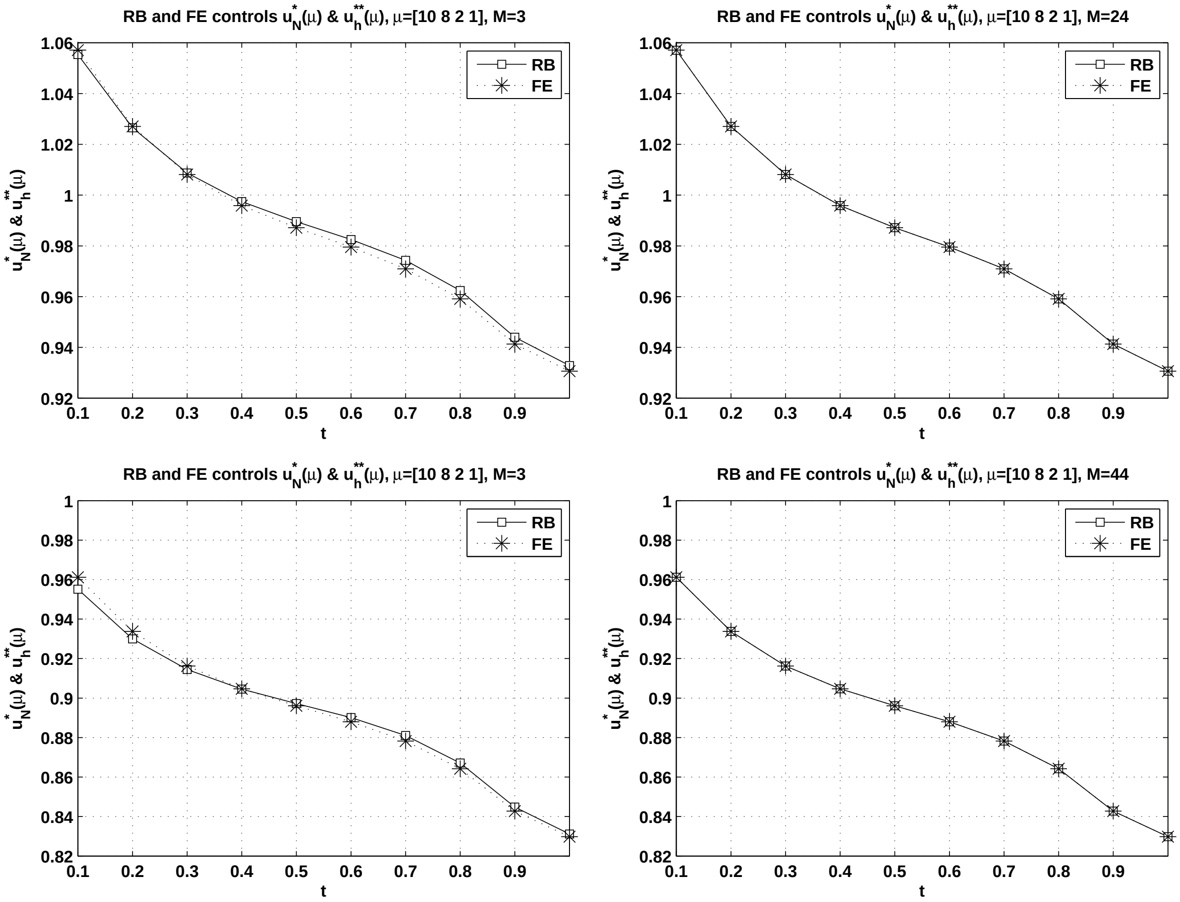

Figure 7.11: Test $C1b$. FE ($*$) and RB ($\Box$) optimal control solutions $u_h^{**,k}(\boldsymbol{\mu}_s)$ and $u_M^{*,k}(\boldsymbol{\mu}_s)$ for $\boldsymbol{\mu}_s = [10, 8, 2, 1]$. Unconstrained case (top) with RB optimal controls for $M = 3$ (top–left) and $M = 24$ (top–right); constrained case (bottom) with RB optimal controls for $M = 3$ (bottom–left) and $M = 44$ (bottom–right).

computational costs (mean values over $\overline{\mathcal{D}}$ for each M). As expected, the RB errors reduce as the computational costs increase. Once again, we stress the fact that the RB computational costs, less than 0.075 s for the unconstrained case and than 0.21 s in the constrained one, are much more inferior w.r.t. those required by the FE approximation, whose mean (over $\overline{\mathcal{D}}$) values are 4.65 s and 4.37 s, respectively.

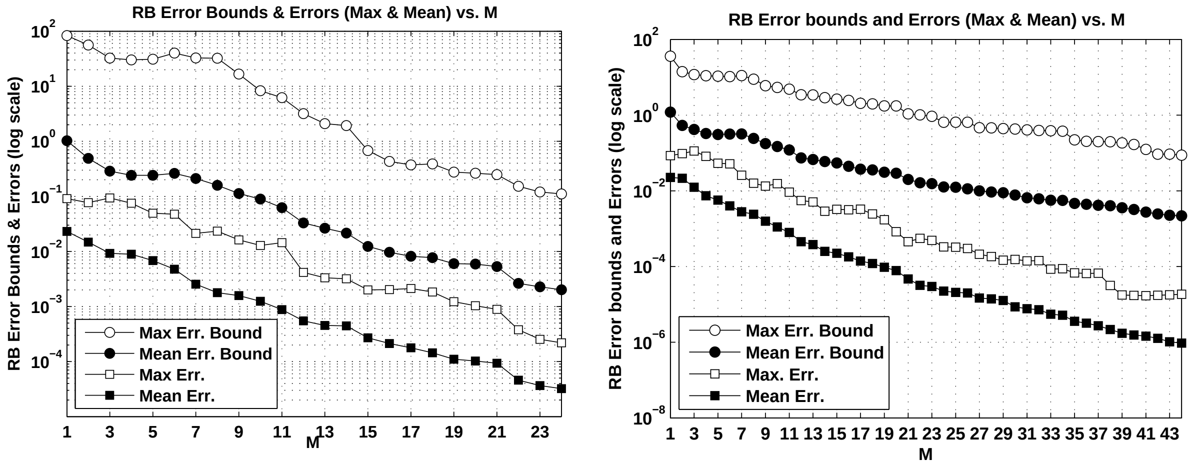

Figure 7.12: Test $C1b$. Error bounds $\Delta_M^{J,*,max}$ (○), $\Delta_M^{J,*,mean}$ (●) and errors $E_M^{J,*,max}$ (□), $E_M^{J,*,mean}$ (■) vs. M; unconstrained (left) and constrained (right) cases. Logarithmic scale on error axis.

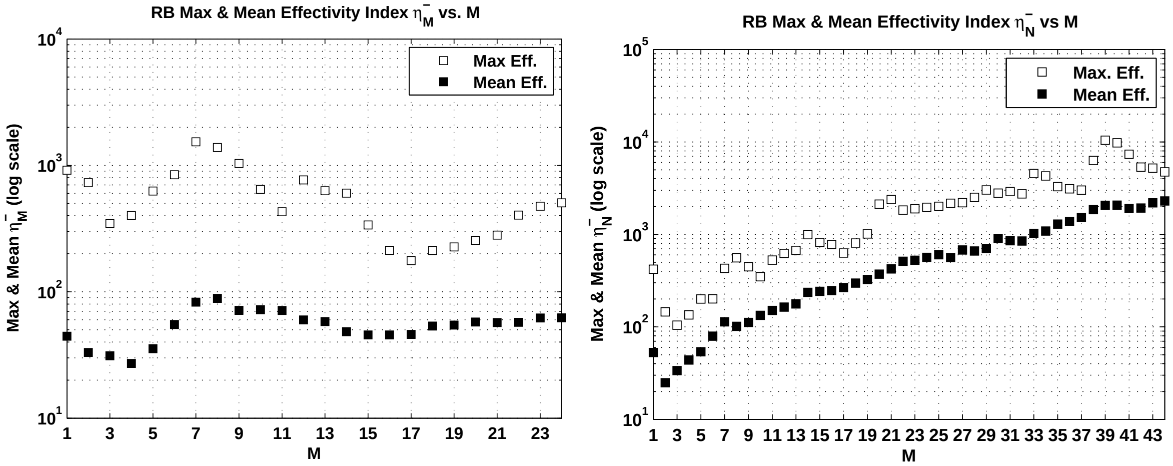

Figure 7.13: Test $C1b$. Effectivity indexes $\overline{\eta}_M^{max}$ (□), $\overline{\eta}_M^{mean}$ (■) vs. M; unconstrained (left) and constrained (right) cases. Logarithmic scale on effectivity index axis.

Test $C2$: pollutant emissions control

We provide now a numerical test inspired by an environmental application. In particular, as anticipated in Sec.1.3, we consider a problem concerning the regulation of the pollutant emissions from an industrial chimney. The goal consists in minimizing the pollutant concentration (the primal solution $v(t;\boldsymbol{\mu})$) over a certain area, e.g. a city, by regulating the emissions rate (the control $u(t;\boldsymbol{\mu})$). The ideal emission rate of the industrial plant and the emissions constraints are also taken into account. As parameters we consider the meteorological conditions, in particular, the wind direction and intensity.

We deal with a constrained optimal control problem described by the parabolic PDE reported

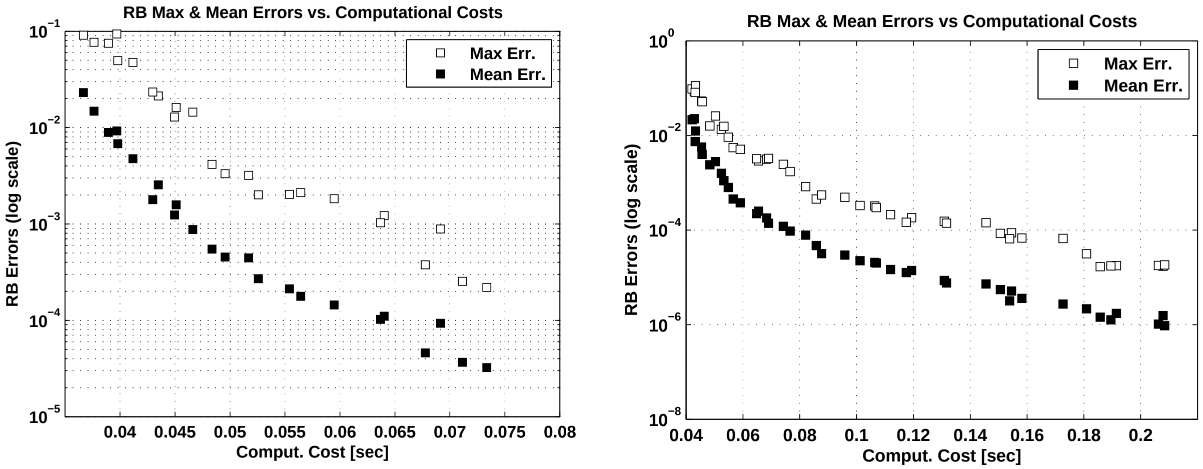

Figure 7.14: Test $C1b$. Errors $E_M^{J,*,max}$ (□) and $E_M^{J,*,mean}$ (■) vs. mean RB computational costs (seconds); unconstrained (left) and constrained (right) cases. Logarithmic scale on error axis.

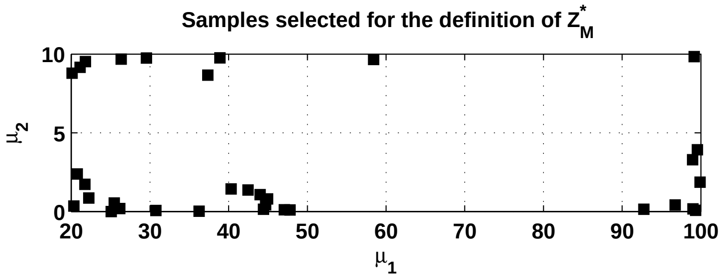

Figure 7.15: Test $C2$. Samples $\boldsymbol{\mu} = [\mu_1, \mu_2] \in \mathcal{D}$ (■) selected by the adaptive procedure for the definition of the set $\mathcal{S}_M^*$.

in Eq.(7.40) (with the control function $u(t; \boldsymbol{\mu})$ in place of $g(t)$) and endowed with the cost functional (7.42). The data provided in Sec.7.1.4 are considered; moreover, we use the same domain represented in Fig.7.1. The bilinear form $m_d(\cdot, \cdot; \boldsymbol{\mu})$ is then defined as $m_d(w, \phi; \boldsymbol{\mu}) := \frac{1}{|\Omega_M|} \int_\Omega w \phi \chi_M \, d\Omega \; \forall w, \phi \in \mathcal{Z}$. Finally, we assume $\mathcal{D} = [20, 100] \times [0, 20]$, $v_d = 0$, $u_d = 0.8$ and $\gamma = 0.5$; moreover, for the control constraints defining the space $\mathcal{U}_{ad,h}$, we set $u_{min} = 0.5$, $u_{max} = 1.0$ and $\overline{U} = 0.75$.

The FE approximation is based on piecewise linear basis functions defined on triangular elements. We set $\Delta t = 10^{-2}$ and $K = 40$, with $T = 0.4$; a mesh composed by $7,320$ triangles and $3,723$ nodes is considered. For the definition of the RB space $\mathcal{Z}_M^*$ we use the adaptive sampling procedure described in Sec.7.2.7 with the tolerance $tol = 10^{-3}$. The number of elements chosen for the definition of $\overline{\mathcal{D}} \subset \mathcal{D}$ is $\overline{N} = 1,000$.

At the offline step, the adaptive procedure selects $M = 33$ optimal basis for the RB space $\mathcal{Z}_M^*$; we recall that the total number of elements of $\mathcal{Z}_M^*$ is $N = 2M = 66$. In Fig.7.15 we report the samples of the set $\mathcal{S}_M^*$ selected by the adaptive procedure.

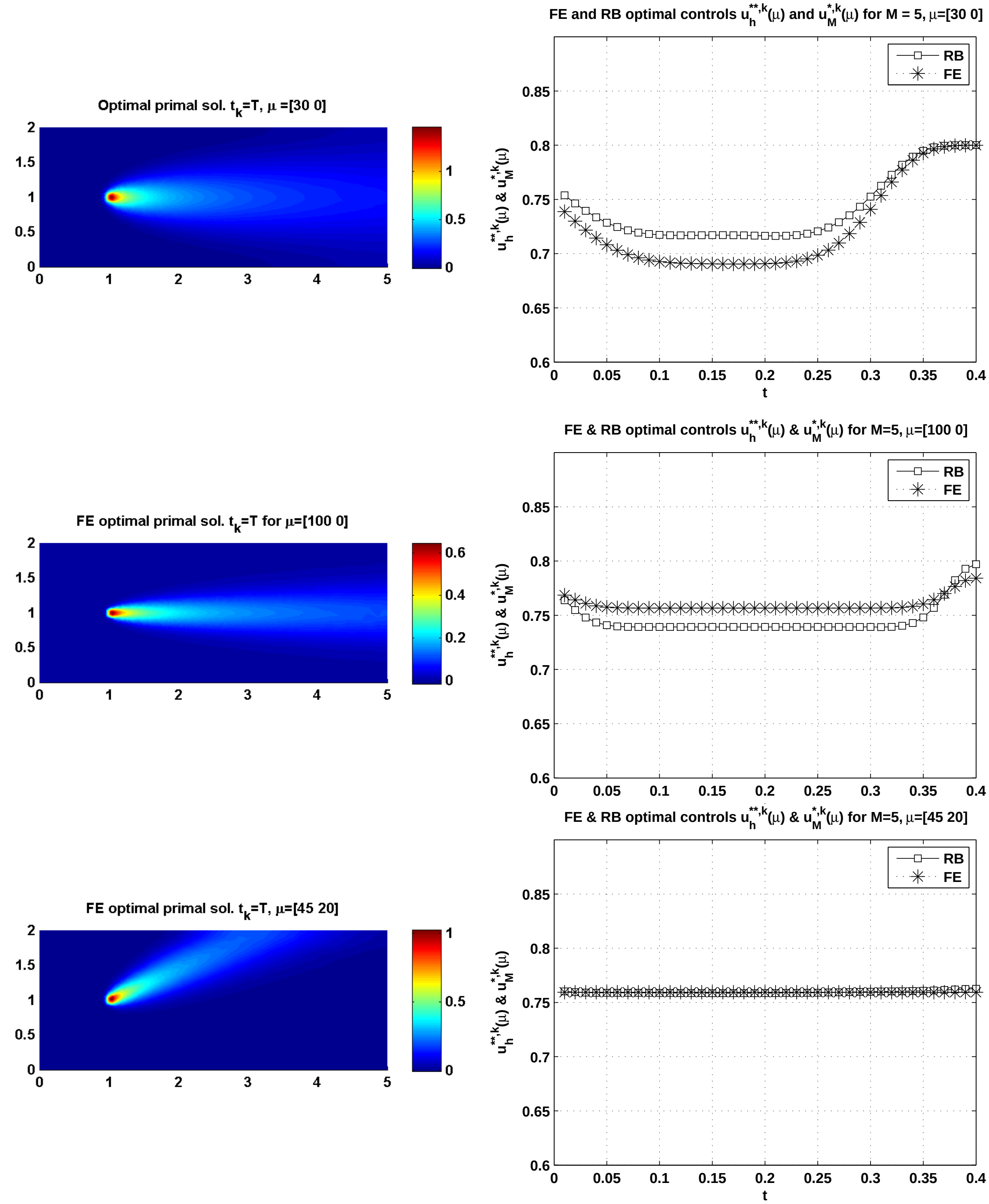

Figure 7.16: Test $C2$. FE optimal primal solutions $v_h^{**}(\boldsymbol{\mu})$ at the final time step $t_K = T$ (left) together with FE ($*$) and RB ($\Box$) optimal control solutions ($u_h^{**,k}(\boldsymbol{\mu})$ and $u_M^{*,k}(\boldsymbol{\mu})$) for $M = 5$ (right); cases with $\boldsymbol{\mu} = \boldsymbol{\mu}_a$ (top), $\boldsymbol{\mu} = \boldsymbol{\mu}_b$ (mid) and $\boldsymbol{\mu} = \boldsymbol{\mu}_c$ (bottom).

In order to evaluate the effectivity of the RB method at the online step, we consider three cases corresponding to different choices of the parameters vector $\boldsymbol{\mu} \in \mathcal{D}$; in particular $\boldsymbol{\mu}_a = [30, 0]$,

	FE		RB $M = 5$			RB $M = 33$		
Case	$J_h^{**}(\boldsymbol{\mu})$	CPU [s]	$E_M^{J,*}(\boldsymbol{\mu})$	$\Delta_M^{J,*,C}(\boldsymbol{\mu})$	CPU [s]	$E_M^{J,*}(\boldsymbol{\mu})$	$\Delta_M^{J,*,C}(\boldsymbol{\mu})$	CPU [s]
$\boldsymbol{\mu}_a$	$6.104 \cdot 10^{-3}$	1109	$1.393 \cdot 10^{-3}$	$8.923 \cdot 10^{-2}$	0.25	$4.336 \cdot 10^{-6}$	$2.873 \cdot 10^{-4}$	0.86
$\boldsymbol{\mu}_b$	$2.132 \cdot 10^{-3}$	1509	$2.186 \cdot 10^{-3}$	$1.479 \cdot 10^{-1}$	0.25	$1.250 \cdot 10^{-6}$	$6.207 \cdot 10^{-4}$	0.61
$\boldsymbol{\mu}_c$	$1.622 \cdot 10^{-4}$	189	$2.587 \cdot 10^{-4}$	$3.228 \cdot 10^{-2}$	0.14	$3.713 \cdot 10^{-6}$	$2.131 \cdot 10^{-4}$	0.38

Table 7.1: Test $C2$. Optimal FE cost functionals $J_h^{**}(\boldsymbol{\mu})$, RB errors $E_M^{J,*}(\boldsymbol{\mu}) = |J_h^{**}(\boldsymbol{\mu}) - J_M^{*}(\boldsymbol{\mu})|$ and RB estimators $\Delta_M^{J,*,C}(\boldsymbol{\mu})$ with $M = 5$ and $M = 33$ for test cases $\boldsymbol{\mu} = \boldsymbol{\mu}_a = [30, 0]$, $\boldsymbol{\mu}_b = [100, 0]$ and $\boldsymbol{\mu}_c = [45, 20]$; comparison of the FE and RB computational costs (CPU times) at the online step.

$\boldsymbol{\mu}_b = [100, 0]$ and $\boldsymbol{\mu}_c = [45, 20]$.

In Fig.7.16(left) we report the FE optimal primal solutions at the final time step $t_K = T$ corresponding to the three cases $\boldsymbol{\mu} = \boldsymbol{\mu}_a$, $\boldsymbol{\mu}_b$ and $\boldsymbol{\mu}_c$. Similarly, in Fig.7.16(right) we compare the FE optimal control functions $u_h^{**,k}(\boldsymbol{\mu})$ with the corresponding RB ones $u_M^{*,k}(\boldsymbol{\mu})$ for $M = 5$. It is evident that, already in this case for which M is "small", the RB optimal controls are able to approximate sufficiently well the corresponding FE ones; for example, for $\boldsymbol{\mu} = \boldsymbol{\mu}_a$ the maximum relative RB error on the FE optimal control is 4.65%. We observe that, as the number of optimal basis M chosen at the online step increases, then $u_M^{*,k}(\boldsymbol{\mu})$ tends to $u_h^{**,k}(\boldsymbol{\mu})$ $\forall k \in \mathbb{K}$.

In Table 7.1 we report the optimal FE cost functionals $J_h^{**}(\boldsymbol{\mu})$ computed for $\boldsymbol{\mu} = \boldsymbol{\mu}_a$, $\boldsymbol{\mu}_b$ and $\boldsymbol{\mu}_c$ together with the RB "optimality" errors $E_M^{J,*}(\boldsymbol{\mu}) = |J_h^{**}(\boldsymbol{\mu}) - J_M^{*}(\boldsymbol{\mu})|$ and estimators $\Delta_M^{J,*,C}(\boldsymbol{\mu})$ for $M = 5$ and $M = 33$. We notice that, for all the cases considered, the estimators are able to bound the corresponding true RB errors. Moreover, we compare the computational costs associated with the FE and RB solutions of the parametrized optimal control problems at the online step. We stress the fact that the RB method allows a considerable costs saving w.r.t. the FE method. For example, for $\boldsymbol{\mu} = \boldsymbol{\mu}_b$, the RB method needs only 0.61 s for $M = 33$ w.r.t. the 1509 s required by the FE one.

7.3 Concluding remarks

In this Chapter, after having introduced the RB method for parabolic PDEs, we have considered the RB method for the solution of parametrized optimal control problems in the unsteady case. In particular, we have provided the formulation of the RB method for the solution of such problems, both in the unconstrained and constrained cases, and we have discussed the advantages offered by the offline–online procedure. A posteriori estimates for the "optimality" RB errors on the quadratic cost functional have been proposed. Moreover, we have introduced an "integrated" RB approach for the definition of the RB space, together with the adaptive procedure for the selection of the RB basis. Numerical tests have highlighted the reliability of the RB estimators in bounding the true RB errors on the cost functional and the capability of the RB method to provide accurate approximations of the FE optimal control functions. Moreover, we have discussed and pointed out the saving of computational costs allowed by the RB method for the solution of parametrized optimal control problems.

Conclusions and Future Developments

In this Thesis we have discussed anisotropic mesh adaption procedures, driven by a posteriori error estimates, for the numerical solution of PDEs based on the FE method. Moreover, we have considered the RB method for the approximation of parametrized PDEs and optimal control problems.

We have shown, via numerical tests inspired by environmental applications, the effectivity of the anisotropic mesh adaption procedure for the solution of advection–diffusion–reaction PDEs. In particular, we have proposed an anisotropic, residual based, a posteriori error estimate for the computation of output functionals. This estimate, which exhibits the good property to depend on the error itself, is made effective by approximating this error via the Zienkiewicz–Zhu gradient recovery procedure [189, 190]. The estimator so obtained highlights better convergence properties w.r.t. both its isotropic version and the error independent estimator. This allows to use a lower number of elements while building the anisotropic meshes and hence to reduce the computational costs associated with the whole adaptive procedure and the numerical solution of the PDEs.

A possible future development in this sense consists in embedding the anisotropic setting into the context of optimal control problems described by PDEs, thus aiming at extending the good properties of the proposed a posteriori error estimate to the case of FE approximation of optimal control problems.

Concerning the numerical solution of parametrized PDEs, we have considered the RB method for the approximation of FE stabilized advection–reaction PDEs. Moving from the a priori RB estimate, we observe that, as the mesh size reduces (i.e. the FE approximation improves), an increasing number of elements for the definition of the RB basis is required in order to bound the RB errors. The stability of the RB solution, if based on stable FE approximations, is also deduced on the basis of numerical tests. Then, we have highlighted the computational costs savings obtained when the problem is solved by means of the "primal–dual" RB formulation based on the goal–oriented analysis for output functionals. Numerical tests show remarkable gains, even with factors greater than 10, for this approach w.r.t. that making use only of the "primal" one.

The RB method has finally been used for the approximation of parametrized optimal control problems described by parabolic PDEs. The proposed a posteriori RB error estimates, both for the unconstrained and constrained cases, allow to extend the good reliability property of the RB method into the optimal control framework. Moreover, the performances in terms of computational costs and accuracy largely justify the adoption of the RB method for the solution of parametrized optimal control problems in many query contexts. Numerical tests

highlight considerable costs savings w.r.t. the FE approximations. For example a gain factor of order greater than $2,000$ is obtained for a pollutant emission control problem with 40 temporal steps, $3,723$ degrees of freedom for the FE approximation and 33 for the RB one, which guarantees a relative error on the RB approximated optimal cost functional lower than 0.06%.
On this basis, the RB method seems to be really promising also for solution of $3D$ parametrized optimal control problems, for which even larger savings w.r.t. the FE approximation are expected. Another possible extension consists in developing a more effective a posteriori RB error estimate in the case of constrained optimal control problems; this in order to contain the effectivity index associated with the RB estimator and to make the estimator sharper.

Bibliography

[1] R.A. Adams. *Sobolev Spaces.* Academic Press: New York, 1975.

[2] V.I. Agoshkov. *Optimal Control Methods and Adjoint Equations in Mathematical Physics Problems.* Institute of Numerical Mathematics, Russian Academy of Science: Moscow, 2003.

[3] V.I. Agoshkov, D. Ambrosi, V. Pennati, A. Quarteroni, F. Saleri. Mathematical and numerical modelling of shallow water flow. *Computational Mechanics* 1993; **11**(5–6):280-299.

[4] M. Al–Baali. Improved Hessian approximations for the limited memory BFGS method. *Numerical Algorithms* 1999; **22**:99–112.

[5] M. Al–Baali. On the behaviour of a combined extra–updating/self–scaling BFGS method. *Journal of Computational and Applied Mathematics* 2001; **134**:269–281.

[6] V.M. Alekseev, V.M. Tikhominov, S.V. Fomin. *Optimal Control.* Consultants Bureau: New York, 1987.

[7] T. Apel. Anisotropic Finite Elements: Local Estimates and Applications. Advances in Numerical Mathematics. Teubner: Stuttgart, 1999.

[8] S.P. Arya. *Air Pollution Meteorology and Dispersion.* Oxford University Press: New York, 1999.

[9] J. Atwell, B. King. Proper orthogonal decomposition for reduced basis feedback controllers for parabolic equations. *Mathematical and Computer Modeling* 2001; **33**(1–3):1–19.

[10] A.K. Aziz, J.W. Wingate, M.J. Balas. *Control Theory of Systems Governed by Partial Differential Equations.* Academic Press: New York, 1977.

[11] I. Babŭska. Error–bounds for finite element method. *Numerische Mathematik* 1971; **16**:322–333.

[12] I. Babŭska, A.D. Miller. The post–processing approach in the finite element method. Part I: calculation of displacements, stresses and other higher derivatives of the displacements. *International Journal for Numerical Methods in Engineering* 1984; **20**:1085–1109.

[13] I. Babŭska, A.D. Miller. The post–processing approach in the finite element method. Part II: the calculation of stress intensity factors. *International Journal for Numerical Methods in Engineering* 1984; **20**:1111–1129.

[14] I. Babŭska, A.D. Miller. The post–processing approach in the finite element method. Part III: a posteriori error estimation and adaptive mesh selection. *International Journal for Numerical Methods in Engineering* 1984; **20**:2311–2324.

[15] I. Babuška, F. Nobile, R. Tempone. Worst–case scenario analysis for elliptic PDE's with uncertainty. In *Proceedings EURODYN 2005*, 889–894 *Structural Dynamics*, MillPress, Rotterdam, 2005.

[16] W. Bangerth, R. Rannacher. *Adaptive Finite Element Methods for Differential Equations.* Birkhauser Verlag: Basel, 2003.

[17] M. Barrault, Y. Maday, N.C. Nguyen, A.T. Patera. An 'empirical interpolation' method: application to efficient reduced–basis discretization of partial differential equations. *Comptes Rendus Mathématique. Acadḿie des Sciences. Paris* 2004; **339**(9):667–672.

[18] R. Becker. Mesh adaptation for Dirichlet flow control via Nitsche's method. *Communications in Numerical Methods in Engineering* 2002; **18**:669–680.

[19] R. Becker. Mesh adaptation for stationary flow control. *Journal of Mathematical Fluid Mechanics* 2001; **3**:317–341.

[20] R. Becker, H. Kapp, R. Rannacher. Adaptive finite element methods for optimal control of partial differential equations: basic concepts. *SIAM Journal on Control and Optimization* 2000; **39**(1):113–132.

[21] R. Becker, R. Rannacher. A feed–back approach to error control in finite elements methods: basic analysis and examples. *East–West Journal of Numerical Mathematics* 1996; **4**(4):237–264.

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

[23] R. Becker, B. Vexler. Optimal control of the convection–diffusion equation using stabilized finite element methods. *Numerische Mathematik* 2007; **106**(3):349–367.

[24] S. Berrone, M. Verani. An adaptive gradient–DWR finite element algorithm for an optimal control constrained problem. *MOX report 2.2008* 2008. `http://mox.polimi.it/`

[25] L. Bonaventura. A semi–implicit, semi–Lagrangian scheme using the height coordinate for a nonhydrostatic and fully elastic model of atmospheric flows. *Journal of Computational Physics* 2000; **158**(2):186–213.

[26] L. Bonaventura, T. Ringler. Analysis of discrete shallow-water models on geodesic Delaunay grids with C–type staggering. *Monthly Weather Review* 2005; **133**:2351–2373.

[27] H. Brezis. *Analyse Fonctionelle, Théorie et Applications.* Masson: Paris, 1983.

[28] F. Brezzi, R.S. Falk. Stability of higher–order Hood–Taylor methods. *SIAM Journal on Numerical Analysis* 1991; **28**:581–590.

[29] F. Brezzi, G. Gilardi. *Functional Analysis and Functional Spaces.* McGraw Hill: New York, 1987.

[30] A.N. Brooks, T.J.R. Hughes. Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations. *Computer Methods in Applied Mechanics and Engineering* 1982; **32**(1–3):199–259.

[31] A. Buffa, Y. Maday, A.T. Patera, C. Prud'homme, G. Turinici. A priori convergence of multi–dimensional parameterized reduced–basis approximations. In progress, 2008.

[32] É. Cancès, C. Le Bris, Y. Maday, N.C. Nguyen, A.T. Patera, G.S.H. Pau. Feasibility and competitiveness of a reduced basis approach for rapid electronic structure calculations in quantum chemistry. *Centre de Recherches Mathèmatiques, CRM Proceedings and Lecture Notes* 2007. To appear.

[33] C. Canuto, M.Y. Hussaini, A. Quarteroni, T.A. Zang. *Spectral Methods. Fundamentals in Single Domains.* Springer–Verlag: Berlin, 2006.

[34] M.J. Castro–Díaz, F. Hecht, B. Mohammadi, O. Pironneau. Anisotropic unstructured mesh adaption for flow simulations. *International Journal for Numerical Methods in Fluids* 1997; **25**(4):475–491.

[35] E.A. Christensen, M. Brøns, N.J. Sørensen. Evaluation of proper orthogonal decomposition–based decomposition techniques applied to parameter–dependent nonturbulent flows. *SIAM Journal on Scientific Computing* 2000; **21**(4):1419–1434.

[36] Ph.G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North–Holland: Amsterdam, 1978.

[37] Ph. Clément. Approximation by finite element functions using local regularization. *RAIRO Analyse Numérique* 1975; **2**:77–84.

[38] S. Collis, M. Heinkenschloss. Analysis of the Streamline Upwind/Petrov Galerkin method applied to the solution of optimal control problems. *CAAM report TR02–01* 2002. `http://www.caam.rice.edu`

[39] J.C. De Los Reyes, F. Tröltzsch. Optimal control of the stationary Navier–Stokes equations with mixed control–state constraints. *SIAM Journal on Control and Optimization* 2007; **46**(2):604–629.

[40] L. Dedè. Controllo ottimale e adattività per equazioni alle derivate parziali e applicazioni. Master Thesis, Politecnico di Milano, 2004. `http://mox.polimi.it/`

[41] L. Dedè. Optimal control for Navier–Stokes equations: drag minimization. *International Journal for Numerical Methods in Fluids* 2007; **55**(4): 347–366.

[42] L. Dedè. Reduced Basis method for parametrized advection–reaction problems. *MOX report 14.2007* 2007. `http://mox.polimi.it/`

[43] L. Dedè, S. Micheletti, S. Perotto. Anisotropic error control for environmental applications. *Applied Numerical Mathematics* 2007. To appear.
Available on line `http://sciencedirect.com/science/journal/01689274`

[44] L. Dedè, A. Quarteroni. Optimal control and numerical adaptivity for advection–diffusion equations. *M2AN. Mathematical Modelling and Numerical Analysis* 2005; **39**(5):1019–1040.

[45] J.E. Dennis Jr., R.B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* SIAM: Philadelphia, 1996.

[46] M. Discacciati, E. Miglio, A. Quarteroni. Mathematical and numerical models for coupling surface and groundwater flows. *Applied Numerical Mathematics* 2002; **43**(1–2):57–74.

[47] K. Eriksson, C. Johnson. Adaptive finite element methods for parabolic problems. Part I: a linear model problem. *SIAM Journal of Numerical Analysis* 1991; **28**(1):43–77.

[48] K. Eriksson, C. Johnson. An adaptive finite element method for linear elliptic problems. *Mathematics of Computation* 1988; **50**(182):361–383.

[49] A. Ern, J.L. Guermond. *Elément Finis: Théorie, Applications, Mise en Oeuvre.* Springer–Verlag: Heidelberg, 2002.

[50] E. Fernández–Cara, E. Zuazua. On the history and perspectives of control theory. *Matapli* 2004; **74**:47–73.

[51] G. Finzi, M. Pirovano, M. Volta. *Gestione della Qualità dell'Aria. Modelli di Simulazione e Previsione.* Mc Graw–Hill: Milano, 2001.

[52] L. Formaggia, S. Micheletti, S. Perotto. Anisotropic mesh adaptation in computational fluid dynamics: application to the advection–diffusion–reaction and the Stokes problems. *Applied Numerical Mathematics* 2004; **51**:511–533.

[53] L. Formaggia, S. Perotto. Anisotropic error estimates for elliptic problems. *Numerische Mathematik* 2003; **94**:67–92.

[54] L. Formaggia, S. Perotto. New anisotropic a priori error estimates. *Numerische Mathematik* 2001; **89**:641–667.

[55] L. Formaggia, S. Perotto, P. Zunino. An anisotropic a–posteriori error estimate for a convection–diffusion problem. *Computing and Visualization in Science* 2001; **4**(2):99–104.

[56] L. Formaggia, V. Selmin. Simulation of hypersonic flows on unstructured grids. *International Journal for Numerical Methods in Engineering* 1992; **34**:569–606.

[57] G. Fourestey, M. Moubachir. Solving inverse problems involving the Navier–Stokes equations discretized by a Lagrange–Galerkin method. *Computer Methods in Applied Mechanics and Engineering* 2005; **194**(6–8):877–906.

[58] G. Fourestey, M. Moubachir. Optimal control of Navier–Stokes equations using Lagrange–Galerkin methods. *INRIA report RR–4609* 2002. `http://www.inria.fr`.

[59] R. Ghanem, H. Sissaoui. A posteriori error estimate by a spectral method of an elliptic optimal control problem. *Journal of Computational Mathematics and Optimization* 2006; **2**(2):111–125.

[60] M.B. Giles, M.G. Larson, J.M. Levenstam, E. Süli. Adaptive error control for finite element approximations of the lift and drag coefficients in viscous flow. *OUCL Numerical Analysis Group report NA-97/06* 1997. `http://web.comlab.ox.ac.uk/oucl`.

[61] M.B. Giles, E. Süli. Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality. *Acta Numerica* 2002; **11**:145–236.

[62] M.B. Giles, N.A. Pierce. Adjoint error correction for integral outputs. In *Error estimation and adaptive discretization methods in computational fluid dynamics*, 47–95, *Lecture Notes in Computational Science and Engineering*, **25**, Springer, Berlin, 2003.

[63] M.B. Giles, N.A. Pierce. An introduction to the adjoint approach to design. *Flow, Turbulence and Combustion* 2000; **65**(3–4):393–415.

[64] P.E. Gill, W. Murray, M.H. Wright. *Practical Optimization.* Academic Press: New York, 1989.

[65] V. Girault, P.A. Raviart. *Finite Element Methods for Navier–Stokes Equations. Theory and Algorithms.* Springer–Verlag: Berlin, 1986.

[66] M. Giuliano, S. Cernuschi. Trattamento degli effluenti gassosi. In *Manuale dell'Ingegnere*, 392–443, A. Guadagni (ed.s), **3**(Q), Hoepli, 2003.

[67] M.S. Gockenbach. Introduction to sequential quadratic programming. Lecture Notes, Michigan Technological University, 2003. `http://www.math.mtu.edu/∼msgocken`

[68] N.I.M. Gould, D. Orban, P.L. Toint. Numerical methods for large scale nonlinear optimization. *Acta Numerica* 2005; **14**:299–361.

[69] W.R. Graham, J. Peraire, K.Y. Tang. Optimal control of vortex shedding using low–order models. Part I: open–loop model development. *International Journal of Numerical Methods in Engineering* 1999; **44**(7):945–972.

[70] M.A. Grepl. Reduced–basis approximations and a posteriori error estimation for parabolic partial differential equations. PhD Thesis, Massachusetts Institute of Technology, 2005. `http://augustine.mit.edu/`

[71] M.A. Grepl, N.C. Nguyen, K. Veroy, A.T. Patera, G.R. Liu. Certified rapid solution of partial differential equations for real–time parameter estimation and optimization. In *Real–time PDE–Constrained Optimization*, 197–215, L. Biegler, O. Ghattas, M. Heinkenschloss, D. Keyes, B. van Bloemen Waanders (ed.s), *Computational Science and Engineering Book Series*, SIAM, Philadelphia, 2007.

[72] M.A. Grepl, A.T. Patera. A posteriori error bounds for reduced–basis approximations of parametrized parabolic partial differential equations. *M2AN Mathematical Modelling and Numerical Analysis* 2005; **39**(1):157–181.

[73] P.M. Gresho, R.L. Sani. *Incompressible Flow and the Finite Elements Method.* J. Wiley: New York, 2000.

[74] R. Griesse. Parametric sensitivity analysis for control–constrained optimal control problems governed by systems of parabolic partial differential equations. PhD Thesis, University of Bayreuth, 2003.

[75] R. Griesse. Parametric sensitivity analysis in optimal control of a reaction diffusion system. Part I: solution differentiability. *Numerical Functional Analysis and Optimization* 2004; **25**(1–2):93-117.

[76] R. Griesse, B. Vexler. Numerical sensitivity analysis for the quantity of interest in PDE–constrained optimization. *SIAM Journal on Scientific Computing* 2007; **29**(1):22–48.

[77] G. Guariso, M. Hitz, H. Werthner. An integrated simulation and optimization modelling Environment for decision support. *Decision Support Systems* 1996; **16**:103–117.

[78] M.D. Gunzburger. Adjoint equation–based methods for control problems in incompressible, viscous flows. *Flow, Turbolence and Combustion* 2000; **65**: 249–272.

[79] M.D. Gunzburger. *Finite Element Method for Viscous Incompressible Flows: a Guide to Theory, Practice and Algorithms*. Academic Press: Boston, 1989.

[80] M.D. Gunzburger. *Perspectives in Flow Control and Optimization, Advances in Design and Control*. SIAM: Philadelphia, 2003.

[81] M.D. Gunzburger, S. Manservisi. The velocity tracking problem for Navier–Stokes flows with boundary control. *SIAM Journal on Control and Optimization* 2000; **39**(2):594–634.

[82] M.D. Gunzburger, J. Peterson, J. Shadid. Reduced–order modeling of time–dependent PDEs with multiple parameters in the boundary data. *Computer Methods in Applied Mechanics and Engineering* 2007; **196**(4–6):1030–1047.

[83] W.G. Habashi, M. Fortin, J. Dompierre, M.G. Vallet, Y. Bourgault. Anisotropic mesh adaptation: a step towards a mesh–independent and user–independent CFD. In *Barriers and Challenges in Computational Fluid Dynamics*, 99–117, V. Venkatakrishnan, M.D. Salas, S.R. Chakravarthy (ed.s), Kluwer Academic, Dordrecht, 1998.

[84] J.L. He, R. Glowinski, R. Metcalfe, A. Nordlander, J. Periaux. Active control and drag optimization for flow past a circular cilinder, I. Oscillatory cylinder rotation. *Journal of Computational Physics* 2000; **163**(1):83–117.

[85] F. Hecht. BAMG: bidimensional anisotropic mesh generator. *INRIA report* 1998. `http://www.inria.fr/`

[86] V. Heuveline, R. Rannacher. Duality–based adaptivity in the *hp*–finite element method. *Journal of Numerical Mathematics* 2003; **11**(2):95–113.

[87] L.S. Hou, S.S. Ravindran. A penalized Neumann control approach for solving an optimal Dirichlet control problem for the Navier–Stokes equations. *SIAM Journal on Control and Optimization* 1998; **36**(5):1795–1814.

[88] L.S. Hou, S.S. Ravindran. Numerical approximation of optimal flow control problems by a penality method: error estimates and numerical results. *SIAM Journal on Scientific Computing* 1999; **20**(5):1753–1777.

[89] D.B.P. Huynh, N.C. Nguyen, A.T. Patera, G. Rozza. Documentation for rbMIT©MIT software package. Part II: Time–dependent problems. *Software Notes*, Massachusetts Institute of Technology, 2007. `http://augustine.mit.edu/`

[90] D.B.P. Huynh, G. Rozza, S. Sen, A.T. Patera. A successive constraint linear optimization method for lower bounds of parametric coercivity and inf–sup stability constants. Submitted to *Comptes Rendus Mathématique. Acadḿie des Sciences. Paris* 2007.

[91] K. Ito, S.S. Ravindran. A reduced basis method for control problems governed by PDEs. In *International Series of Numerical Mathematics*, 153–168, *International Series of Numerical Mathematics*, **126**, Birkäuser, Basel, 1998.

[92] K. Ito, S.S. Ravindran. A reduced–order method for simulation and control of fluid flow. *Journal of Computational Physics* 1998; **143**(2):403–425.

[93] K. Ito, S.S. Ravindran. Reduced basis methods for optimal control of unsteady viscous flows. *International Journal of Computational Fluid Dynamics* 2001; **15**(2):97–113.

[94] A. Jameson. CFD for Aerodynamics design and optimization: its evolution over the last three decades. *16th AIAA CFD Conference*, June 23–26 2003, Orlando, FL, USA. *AIAA paper 2003–3438* 2003.

[95] A. Jameson. Optimum aerodynamic design using CFD and control theory. *AIAA paper* 1988; **95–1729**: 233–260.

[96] C. Johnson. *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Cambridge University Press: Cambridge, 1987.

[97] C. Johnson, A.H. Schatz, L.B. Wahlbin. Crosswind smear and pointwise errors in streamline diffusion finite element methods. *Mathematics of Computation* 1987; **49**(179):25–38.

[98] E. Kalnay. *Atmospheric Modelling, Data Assimilation and Predictability.* Cambridge University Press: 2003.

[99] A.N. Kolmogorov, S.V. Fomin. *Elements of Theory of Functions and Functional Analysis.* V.M. Tikhomirov, Nauka: Moscow, 1989.

[100] G. Kunert. A Posteriori error estimation for anisotropic tetrahedral and triangular finite element meshes. PhD Thesis, Fakultät für Mathematik der Technischen Universität Chemnitz, 1999.

[101] K. Kunisch, A. Rösch. Primal–dual active set strategy for a general class of constrained optimal control problems. *SIAM Journal on Optimization* 2002; **13**(2):321–334.

[102] K. Kunisch, S. Volkwein. Control of Burgers' equation by a reduced–order approach using proper orthogonal decomposition. *Journal of Optimization Theory and Applications* 1999; **102**(2):345–371.

[103] K. Kunisch, S. Volkwein. Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. *SIAM Journal on Numerical Analysis* 2002; **40**(2):492–515.

[104] A.T.Y. Kwang. Reduced basis method for 2*nd* order wave equation: application to one–dimension seismic problem. Master Thesis, Singapore–MIT Alliance, National University of Singapore, 2006. `http://augustine.mit.edu/`

[105] R.J. LeVeque. *Numerical Methods for Conservation Laws*. Birkhäuser Verlag: Basel, 2002.

[106] R. Li, W. Liu, H. Ma, T. Tang. Adaptive finite element approximation for distributed elliptic optimal control problems. *SIAM Journal on Control and Optimization* 2003; **41**(5):1321–1349.

[107] J.L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations.* Springer–Verlag: New York, 1971.

[108] J.L. Lions. *Some Aspects of the Optimal Control of Distributed Parameter Systems.* SIAM: Philadelphia, 1972.

[109] J.L. Lions, E. Magenes. *Non–homogeneous Boundary Value Problems and Applications*, Vol.I. Springer–Verlag: New York and Heidelberg, 1972.

[110] D.C. Liu, J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming* 1989; **45**:503–528.

[111] A.E. Løvgren, Y. Maday, E. Rønquist. A reduced basis element method for the steady Stokes problem. *M2AN. Mathematical Modelling and Numerical Analysis* 2006; **40**(3):529–552.

[112] L. Machiels, J. Peraire, A.T. Patera. A posteriori finite element output bounds for the incompressible Navier–Stokes equations: application to a natural convection problem. *MIT–FML technical report 99–4* 1999. `http://mit.edu/`

[113] Y. Maday, A.T. Patera, G. Turinici. A priori convergence theory for reduced–basis approximations of single–parameter elliptic partial differential equations. *Journal of Scientific Computing* 2002; **17**(1–4):437–446.

[114] Y. Maday, A.T. Patera, G. Turinici. Global a priori convergence theory for reduced–basis approximations of single–parameter symmetric coercive elliptic partial differential equations. *Comptes Rendus Mathématique. Académie des Sciences. Paris* 2002; **335**(3):289–294.

[115] G. Maisano, S. Micheletti, S. Perotto, C.L. Bottasso. On some new recovery based a posteriori error estimators. *Computer Methods in Applied Mechanics and Engineering* 2006; **195**(37–40):4794–4815.

[116] H. Maurer. First and second order sufficient optimality conditions in mathematical programming and optimal control. *Mathematical Programming Study* 1981; **14**:163–177.

[117] H. Maurer, H.D. Mittelmann. Optimization techniques for solving elliptic control problems with control and state constraints: Part 1. Boundary control. *Computational Optimization and Applications* 2000; **16**(1):29–55.

[118] H. Maurer, J. Zowe. First and second order necessary and sufficient optimality conditions for infinite–dimensional programming problems. *Mathematical Programming* 1979; **16**:98–110.

[119] D. Meider, B. Vexler. Adaptive space–time Finite Element methods for parabolic optimization problems. *SIAM Journal on Control and Optimization* 2007; **46**(1):116–142.

[120] S. Micheletti, S. Perotto. An anisotropic recovery-based aposteriori error estimator. In *Numerical Mathematics and Advanced Applications*, 731–741, F. Brezzi, A. Buffa, S. Corsaro, A. Murli (ed.s), ENUMATH2001 4th European International Conference on Numerical Mathematics and Advanced Applications, Springer–Verlag Italia, 2003.

[121] S. Micheletti, S. Perotto. Reliability and efficiency of an anisotropic Zienkiewicz–Zhu error estimator. *Computer Methods in Applied Mechanics and Engineering* 2006; **195**(9–12):799–835.

[122] S. Micheletti, S. Perotto, M. Picasso. Stabilized finite elements on anisotropic meshes: a priori error estimates for the advection–diffusion and the Stokes problems. *SIAM Journal on Numerical Analysis* 2003; **41**(3):1131–1162.

[123] E. Miglio, A. Quarteroni, F. Saleri. Finite element approximation of quasi-3D shallow water equations. *Computer Methods in Applied Mechanics and Engineering* 1999; **174**(3–4):355–369.

[124] E. Miglio, A. Quarteroni, F. Saleri. Mathematical modelling of free surface flows. In *Topics in Mathematical Fluid Mechanics*, 95–123, *Quaderni di Matematica* **10**, Seconda Università di Napoli, Caserta, 2002.

[125] B. Mohammadi, O. Pironneau. *Applied Shape Optimization for Fluids.* Clarendon Press: Oxford, 2001.

[126] Networking and Information Technology Research and Development, President's Information Technology Advisory Committee. Computational Science: ensuring America's competitiveness. *NITRD, PITAC report*, 2005. `http://www.nitrd.gov/`

[127] N.C. Nguyen. Reduced Basis approximation and a posteriori error bounds for nonaffine and nonlinear partial differential equations: application to inverse analysis. PhD Thesis, Singapore–MIT Alliance, National University of Singapore, 2005. `http://augustine.mit.edu/`

[128] N.C. Nguyen, A.T. Patera. Reduced basis approximation and a posteriori error estimation for linear parabolic problems. *MIT report* 2008. In progress.

[129] N.C. Nguyen, K. Veroy, A.T. Patera. Certified real–time solution of parametrized partial differential equations. In *Handbook of Materials Modeling*, 1523–1558, S. Yip (ed.s), Springer, 2005.

[130] J. Nocedal, S.J. Wright. *Numerical Optimization.* Springer: New York, 2006.

[131] J.T. Oden, S. Prudhomme. Goal–oriented estimation and adaptivity for the finite element method. *Computers & Mathematics with Applications* 2001; **41**(5–6):735–756.

[132] J.T. Oden, S. Prudhomme. On goal–oriented error estimation for elliptic problems: application to the control of pointwise errors. *Computer Methods in Applied Mechanics and Engineering* 1999; **176**:313–331.

[133] J.T. Oden, J.N. Reddy. *Variational Methods in Theoretical Mechanics.* Springer: Berlin and Heidelberg, 1983.

[134] A.T. Patera, E. Rønquist. A general output bound result: application to discretization and iteration error estimation and control. *Mathematical Models & Methods in Applied Sciences* 2001; **11**(4):685–712.

[135] A.T. Patera, E. Rønquist. Reduced basis approximations and a posteriori error estimation for a Boltzmann model. *Computer Methods in Applied Mechanics and Engineering* 2007. To appear.
Available on line `http://www.sciencedirect.com/science/journal/00457825`

[136] A.T. Patera, G. Rozza. *Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations.* Version 1.0, Copyright MIT 2006–2007. To appear in (tentative rubric) MIT Pappalardo Graduate Monographs in Mechanical Engineering. `http://augustine.mit.edu/`

[137] J. Pedlosky. *Geophysical Fluid Dynamics.* Springer–Verlag: New York, 1987.

[138] J. Peraire, A.T. Patera. Bounds for linear–functional outputs of coercive partial differential equations: local indicators and adaptive refinement. In *Advances in Adaptive Computational Methods in Mechanics*, 199–215, P. Ladeveze, J.T. Odens (ed.s), Elsevier, Amsterdam, 1998.

[139] J. Peraire, M. Vahadati, K. Morgan, O.C. Zienkiewicz. Adaptive remeshing for compressible flow computations. *Journal of Computional Physics* 1987; **72**:449–466.

[140] M. Picasso. An anisotropic error indicator based on Zienkiewicz–Zhu error estimator: application to elliptic and parabolic problems. *SIAM Journal on Scientific Computing* 2003; **24**(4):1328–1355.

[141] M. Picasso. Anisotropic a posteriori error estimate for an optimal control problem governed by the heat equation. *Numerical Methods for Partial Differential Equations* 2006; **22**(6):1314–1336.

[142] R.A. Pielke. *Mesoscale Meteorological Modeling.* Academic Press: New York, 1984.

[143] O. Pironneau, E. Polak. Consistent approximation and approximate functions and gradients in optimal control. *SIAM Journal on Control and Optimization* 2002; **42**(2):487–510.

[144] E. Polak. *Optimization: Algorithms and Consistent Approximations.* Springer: New York, 1997.

[145] C. Prud'homme, D. Rovas, K. Veroy, Y. Maday, A.T. Patera, G. Turinici. Reliable real–time solution of parametrized partial differential equations: reduced–basis output bound methods. *Journal of Fluids Engineering* 2002; **124**(1):70–80.

[146] L. Quartapelle. *Numerical Solution of the Incompressible Navier–Stokes Equations.* Birkhäuser Verlag: Basel, 1993.

[147] A. Quarteroni. *Modellistica Numerica per Problemi Differenziali.* Springer–Verlag: Milano, 2007.

[148] A. Quarteroni, L. Bonaventura, L. Dedè, E. Miglio, A. Quaini, M. Restelli, G. Rozza, F. Saleri. Modellistica matematica in problemi ambientali. *Istituto Lombardo Scienze e Lettere, Quaderni Incontri di Studio* 2006; **42**. In press.

[149] A. Quarteroni, A. Quaini, G. Rozza. Reduced basis methods for advection–diffusion optimal control problems. In *Advances in Numerical Mathematics*, 193–216, W. Fitzgibbon, R. Hoppe, J. Periaux, O. Pironneau, Y. Vassilevski (ed.s), Moscow, Russian Academy of Science and Houston, University of Houston, 2006.
Also *MOX report 79.2006* `http://mox.polimi.it`

[150] A. Quarteroni, G. Rozza. Numerical solutions of parametrized Navier–Stokes equations by reduced basis method. *Numerical Methods for Partial Differential Equations*, 2007. To appear.
Available on line `http://www3.interscience.wiley.com/cgi-bin/jhome/35979`

[151] A. Quarteroni, G. Rozza, L. Dedè, A. Quaini. Numerical approximation of a control problem for advection-diffusion processes. In *System Modeling and Optimization*, 261–273, *IFIP International Federation for Information Processing*, **199**, Springer, New York, 2006.
Available on line `http://www.springerlink.com/content/100525/`

[152] A. Quarteroni, R. Sacco, F. Saleri. *Numerical Mathematics.* Springer–Verlag: Berlin, 2007.

[153] A. Quarteroni, A. Valli. *Numerical Approximation of Partial Differential Equations.* Springer–Verlag: Berlin and Heidelberg, 1994.

[154] R. Rannacher. A posteriori error estimation in least–squares stabilized finite element schemes. *Computer Methods in Applied Mechanics and Engineering* 1998; **166**:99–114.

[155] S.S. Ravindran. A reduced–order approach for optimal control of fluids using proper orthogonal decomposition. *International Journal of Numerical Methods for Fluids* 2000; **34**(5):425–448.

[156] M. Restelli. Semi–lagrangian and semi–implicit discontinuous Galerkin methods for atmospheric modeling applications. PhD Thesis, Politecnico di Milano, 2007. `http://mox.polimi.it/`

[157] M. Restelli, L. Bonaventura, R. Sacco. A semi–Lagrangian discontinuous Galerkin method for scalar advection by incompressible flows. *Journal of Computational Physics* 2006; **216**(1):195–215.

[158] S. Rinaldi, W. Sanderson, A. Gragnani. Pollution control policies and natural resource dynamics: a theoretical analysis. *Journal of Environmental Management* 1996; **48**:357–373.

[159] R. Rodriguez. Some remark on the Zienkiewicz–Zhu estimator. *Numerical Methods for Partial Differential Equations* 1994; **10**(5):625–635.

[160] J.M. Roussel, A. Haro, R.A. Cunjak. Field test of a new method for tracking small fishes in shallow rivers using passive integrated transponders. *Canadian Journal of Fisheries and Aquatic Sciences* 2000; **57**:1326–1329.

[161] G. Rozza. Controllo ottimale e ottimizzazione di forma in fluidodinamica computazionale. Master Thesis, Politecnico di Milano, 2002. `http://mox.polimi.it`

[162] G. Rozza. Optimal flow control and reduced basis techniques in shape design with applications in haemodynamics. PhD Thesis, École Polytechnique Fédérale de Lausanne, 2005. `http://library.epfl.ch/theses/`

[163] G. Rozza. Reduced–basis methods for elliptic equations in sub–domains with a posteriori error bounds and adaptivity. *Applied Numerical Mathematics* 2005; **55**(4):403–424.

[164] G. Rozza. Reduced basis methods for Stokes equations in domains with non affine parameter dependence. *Computing and Visualization in Science* 2007. To appear. Available on line `http://springerlink.metapress.com/content/100525/`

[165] G. Rozza, K. Veroy. On the stability of the reduced basis method for Stokes equations in parametrized domains. *Computer Methods in Applied Mechanics and Engineering* 2007; **196**(7):1244–1260.

[166] M. Schäfer, S. Turek, R. Rannacher. Evaluation of a CFD benchmark for laminar flows. In *Proceedings ENUMATH 1997*, 549–563, September 28 – October 3 1997, Heidelberg, Germany, World Scientific Publications, Singapore, 1998.

[167] K. Schittkowski. Solving nonlinear programming problems with very many constraints. *Optimization* 1992; **25**(2–3):179–196.

[168] S. Sen, K. Veroy, D.B.P. Huynh, S. Deparis, N.C. Nguyen, A.T. Patera. "Natural norm" a posteriori error estimators for reduced basis approximations. *Journal of Computional Physics* 2006; **217**(1):37–62.

[169] K.G. Siebert. An a posteriori error estimator for anisotropic refinement. *Numerische Mathematik* 1996; **73**(3):373–398.

[170] J. Sokolowski, J.P. Zolesio. *Introduction to Shape Optimization (Shape Sensitivity Analysis)*. Springer–Verlag: New York, 1991.

[171] R.B. Stull. *An Introduction to Boundary Layer Meteorology.* Kluver Academic Publishers: Dordrecht, 1988.

[172] J. Sun, J. Zhang. Global convergence of conjugate gradient methods without line search. *Annals of Operations Research* 2001; **103**:161–173.

[173] C. Taylor, P. Hood. A numerical solution of the Navier–Stokes equations using the finite element technique. *Computers & Fluids* 1973; **1**(1):73–100.

[174] F. Tröltzsch. *Optimale Steuerung Partieller Differentialgleichungen.* Friedr. Vieweg & Sohn Verlag: Wiesbaden, 2005.

[175] D.B. Turner. *Workbook of Atmospheric Dispersion Estimates: an Introduction to Dispersion Modeling.* Lewis Publishers: Boca Raton, 1994.

[176] F.P. Vasil'ev. *Methods for Solving the Extremal Problems*. Nauka: Moscow, 1981.

[177] D.A. Venditti, D.L. Darmofal. Grid adaption for functional outputs: application to two–dimensional inviscid fluids. *Journal of Computational Physics* 2002; **176**(1):40–69.

[178] R. Verfürth. *A Review of A Posteriori Error Estimation and Adaptive Mesh–refinement Techniques*. Wiley–Teubner: New York, 1996.

[179] K. Veroy. Reduced–basis methods applied to problems in elasticity: analysis and applications. PhD Thesis, Massachusetts Institute of Technology, 2003. `http://augustine.mit.edu/`

[180] K. Veroy, A.T. Patera. Certified real–time solution of the parametrized steady incompressible Navier–Stokes equations: rigorous reduced–basis a posteriori error bounds. *International Journal for Numerical Methods in Fluids* 2005; **47**(8–9):773–788.

[181] K. Veroy, C. Prud'homme, D.V. Rovas, A.T. Patera. 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, *AIAA Paper* 2003–3847.

[182] B. Vexler, W. Wollner. Adaptive finite elements for elliptic optimization problems with control constraints. *SIAM Journal on Control and Optimization* 2007. To appear.

[183] C. Wang. Some remarks on conjugate gradient methods without line search. *Applied Mathematics and Computation* 2006; **181**:370–379.

[184] Z. Wei, G. Li, L. Qi. New nonlinear conjugate gradient formulas for large–scale unconstrained optimization problems. *Applied Mathematics and Computation* 2006; **179**(2):407–430.

[185] K. Willcox, J. Peraire. Balanced model reduction via the proper orthogonal decomposition. *AIAA Journal 0001–1452* 2002; **40**(11):2323–2330.

[186] K. Yosida. *Functional Analysis*. Springer–Verlag: Berlin, 1974.

[187] G. Zhou. How accurate is the streamline diffusion finite element method? *Mathematics of Computation* 1997; **66**(217):31–44.

[188] O.C. Zienkiewicz, J.Z. Zhu. A simple error estimator and adaptive procedure for practical engineering analysis. *International Journal for Numerical Methods in Engineering* 1987; **24**(2):337–357.

[189] O.C. Zienkiewicz, J.Z. Zhu. The superconvergent patch recovery and a posteriori error estimates. Part I: the recovery technique. *International Journal for Numerical Methods in Engineering* 1992; **33**:1331–1364.

[190] O.C. Zienkiewicz, J.Z. Zhu. The superconvergent patch recovery and a posteriori error estimates. Part II: error estimates and adaptivity. *International Journal for Numerical Methods in Engineering* 1992; **33**:1365–1382.

[191] *freeFEM++*. freeFEM.org. `http://www.freefem.org`.

[192] *Matlab®*. The MathWorks. `http://www.mathworks.com/`

[193] *rbMIT©MIT Software*. Massachusetts Institute of Technology. `http://augustine.mit.edu/`