SISSA
Scuola
Internazionale
Superiore di
Studi Avanzati

Mathematics Area – PhD course in

Mathematical Analysis, Modelling, and Applications

Reduced order models for parametric bifurcation problems in nonlinear PDEs

Candidate:
Federico Pichi

Advisor:
Prof. Gianluigi Rozza

Co-advisor:
Dr. Francesco Ballarin

Academic Year 2019 - 2020

Acknowledgements

First of all, I would like to thank, Prof. Gianluigi Rozza, for being the supervisor of this thesis, but more importantly a considerate mentor and an example of dedication and kindness as fundamental human values. He gave me the freedom to explore interesting paths and to be part of a stimulating and cooperative research group.

I am thankful to my co-advisor, Dr. Francesco Ballarin, for his extraordinary guidance, invaluable advice and enormous patience.

I would like to thank the members of the committee who revised this work, Prof. Jan S. Hesthaven, for his very insightful and constructive feedbacks during our collaboration, and Prof. Anthony T. Patera, for allowing me to visit MIT and benefit from his inspiring viewpoint and fruitful discussions.

A special thank goes to all the new friends and colleagues that I found in Trieste - Matteo, Saddam, Monica, Veronica, Maria, Luca, Emanuele, Daniele - which taught me a lot, giving me support and good memories.

I would also like to thank all the people who care about me, from the high school mates (plus additions) - Antonio, Francesco, Gabriele, Massimiliano, Riccardo, Giulia, Elisabetta - to the university colleagues - Luca, Pietro, Francesco - for their true friendship despite the distance. I sincerely thank Gloria for entering my life and being by my side during this time.

Infine, vorrei ringraziare dal profondo i miei genitori, mio fratello e in generale tutta la nostra grande famiglia allargata, per i valori trasmessi, per il supporto ricevuto e per l'amore che mi avete sempre dimostrato.

Roma, August 2020

Abstract

This work is concerned with the analysis and the development of efficient Reduced Order Models (ROMs) for the numerical investigation of complex bifurcating phenomena held by nonlinear parametrized Partial Differential Equations (PDEs) in many physical contexts, from Continuum Mechanics to Quantum Mechanics passing through Fluid Dynamics. Indeed, the reconstruction of the bifurcation diagrams, which highlight the singularities of the equations and the possible coexisting states, requires a huge computational effort, especially in the multi-parameter context.

To overcome this issue, we developed a reduced order branch-wise algorithm for the efficient investigation of such complex behaviour, with a focus on the stability properties of the solutions. We applied our approach to the Von Kármán equations for buckling plates, the Gross-Pitaevskii equations in Bose-Einstein condensates, the Hyperelastic models for bending beams and the Navier-Stokes model for the flow in a channel.

Several issues and questions arise when dealing with the approximation and the reduction of bifurcating phenomena, we addressed them by considering new models and emerging methodologies. In particular, we developed a reduced order approach to deflated continuation method, to efficiently discover new solution branches. We proposed and discussed different Optimal Control Problems (OCPs) to steer the bifurcating behaviour towards desired states.

Finally, we exploited a Neural Network approach based on the Proper Orthogonal Decomposition (POD-NN), as an alternative to the Empirical Interpolation Method (EIM), to develop a reduced manifold based algorithm for the efficient detection of the bifurcation points.

Keywords: reduced order models, reduced basis methods, bifurcation theory, stability analysis, continuum mechanics, fluid dynamics, quantum mechanics, optimal control problems, artificial neural network.

Contents

Chapter 1

Introduction and motivations

A beam under compression, the flow in heart's chambers and the evolution of a particle's wave-function. We are surrounded by these phenomena in our daily life, and a strategic task for the progress is to develop a deep understanding of their complex behaviour. The above physical processes can be modelled by the so called Partial Differential Equations (PDEs), that describe through differential operators the relations between the temporal/spatial unknowns of the problem under investigation.

Thinking about how differently a rubber object responds to compression with respect to a steel one, it is clear that the analysis of the models has to carry out some information about their physical properties. This is generally achieved by employing a parameter dependence of the PDEs.

Solving analytically such complex equations is in most cases impossible, thus advanced numerical methodologies can be implemented to approximate their solutions. However, despite the increasing computational resources, accurate enough numerical simulations are usually too much time-demanding, such that a complete analysis of their behaviour with respect to the parameter considered becomes impracticable.

The aforementioned physical phenomena are characterized by their complexity, which reflects into nonlinear relations that worsen the investigation. Indeed, such nonlinearities usually originate the so called bifurcating phenomena, i.e. sudden changes in the behaviour of the solutions, and could represent a critical source of issues for their numerical approximation. In fact, while from the physical point of view this translates into the non-uniqueness of the state for a given parameter, on the other hand the mathematical model becomes singular in such critical points where the bifurcation occurs.

These considerations actually constitute the backbone of this thesis. Indeed, this work is concerned with the analysis and the development of efficient Reduced Order Models (ROMs) for the investigation of complex bifurcating phenomena held by nonlinear parametrized PDEs in many physical contexts, from Continuum Mechanics to Quantum Mechanics passing through Fluid Dynamics.

Let us describe more in detail the main motivations of this work. First of all, despite the huge discrepancies among the models we treated, we provide a unique abstract

framework where to cast the analysis of all the different physical contexts. This way we can develop an algorithmic procedure that can be applied to each model.

The numerical investigation of complex physical behaviour, which entails bifurcating phenomena, usually required fine discretization of the domain in which the problem is posed. The presence of nonlinearities and the parameter dependence lead to a multi-query context that could represent a bottleneck for a suitable study. In fact, when dealing with bifurcations, one needs to compute the numerical approximation of the solution for many instances of the parameter in order to discover the critical points of the model and its post-bifurcating behaviour. Moreover, the analysis becomes even more complicated if the goal is a complete reconstruction of all the possible solutions that the model could admit. Finally, a key aspect in these contexts is the stability analysis of the discovered solutions, which can be studied by means of specific eigenvalue problems.

The numerical approximation of such models requires a combination of different methodologies. To obtain an high fidelity approximation of the problem, we considered as Full Order Model (FOM) the Finite Element method, which we combined with a continuation method, to follow the solution varying the parameter, and to the Newton method which serves to linearize the equation.

We tackled the huge computational cost that such investigation would require by means of the ROMs. In particular, we developed a branch-wise algorithm based on the POD technique which aims at efficiently reconstructing the bifurcating behaviour projecting the governing equations into a low dimensional manifold spanned by global basis functions (w.r.t. the parameter space).

The key ingredient to recover the efficiency and allow for a complete real-time investigation of the parameter space in a multi-query context is the offline-online paradigm. Indeed, one assumes to decouple the computation such that the high fidelity quantities can be precomputed offline during an expansive phase, while the actual investigation of the parameter dependence is studied online, assembling reduced quantities with cheaper costs. Unfortunately, the nonlinear terms appearing in the models we considered, compromise this decoupling. For this reason, affine-recovery techniques as the Empirical Interpolation Method (EIM) are fundamental to obtain rapid evaluations during the online phase. Despite this, sometimes they can add a further impracticable cost, even for the offline stage: different methodologies have to be considered.

A wide variety of models have been studied with the twofold motivation of investigating the physics of their bifurcation phenomena and show the applicability of the ROMs strategies we have developed in the non-uniqueness context. An example of a sudden change of the state solution with respect to a (compression) parameter is the buckling of plates, which is modelled through the Von Kármán equations. Remaining in the context of deformation of solids, one can also study the buckling of beams, which is governed by a class of models called Hyperelastic equations. A different perspective is given by the analysis of the density of particle's wave-function in a Bose-Einstein condensate, indeed while varying the chemical potential different

configurations can coexist due to the complex nonlinear Schrödinger equation which describes their behaviour. Finally, interesting bifurcating phenomena can be discovered in fluid dynamics applications governed by the Navier-Stokes equations.

Of course, many issues and questions arise during the analysis of these complex models. It is possible to develop an automatic methodology that guides our algorithm towards all the coexisting solutions for a given instance of the parameter? To answer this question we analysed and developed a reduced deflation method which prevents the system to converge to already known solutions, enabling the discovery of new ones.

Can we somehow act on the system to steer the bifurcating solutions towards states with desired properties? We deeply investigated how Optimal Control Problems (OCPs) can be designed to this goal, while studying their stability properties.

Which approaches can we consider as alternatives to the affine-recovery techniques? Novel machine learning methodologies can be adapted to the ROMs paradigm by means of Artificial Neural Networks (ANNs), constituting a non-intrusive way to recover the online efficiency while discovering manifold's characteristic features.

The numerical results presented in this thesis have been performed with RBniCS [135] library, developed at SISSA mathLab, which is an implementation in FEniCS [57] of several reduced order modelling techniques; we acknowledge developers and contributors to both libraries.

The presentation of this work is original, however it is based on results which are ongoing works, submitted preprints or already published journal articles. In particular, Chapter 4 is mainly based upon published work with G. Rozza [120]. Chapter 5 is mainly based upon accepted work with A. Quaini [119]. Chapter 6 is based upon ongoing work with J. Eftang and A. T. Patera [118]. Chapters 7 and 9 are entirely based on the preprint done in collaboration with M. Strazzullo [121]. Chapter 8 is entirely based on the submitted paper in collaboration with M. Pintore whom I supervised for his master thesis [122]. Chapter 10 is entirely based on the ongoing work with J. S. Hesthaven [117].

Finally, the ongoing work with F. Ballarin on the development of a new technique called Successive Partition Method (SPM) [11] for the efficient evaluation of parametrized stability factors, was not presented in this thesis for the sake of consistency.

Thesis outline

This thesis is divided in three parts. The first part concerns a theoretical and methodological introduction to bifurcation theory and its numerical approximation. Subsequently, we discuss, analyse and (efficiently) numerically investigate four different models, each one with its specific bifurcating behaviour. In the third part, we take a step forward towards more advanced models and methodologies. In particular:

- **Part I** - Theoretical and numerical background for bifurcation analysis
  - In **Chapter 2** we present the mathematical framework of parametrized nonlinear PDEs holding bifurcating phenomena and non-uniqueness of solutions, with a focus on their stability properties through the eigenvalue analysis.
  - In **Chapter 3** we review the existing high fidelity and reduced order methods that we combined to develop our branch-wise algorithm for the efficient reconstruction of the bifurcation diagram.
- **Part II** - Approximation of parametrized bifurcating PDEs models
  - In **Chapter 4** we discuss the Von Kármán equations as a model for the buckling of plates. We show the theoretical connection between buckling points and eigenvalues by means of generalized eigenvalue problems. Two different geometries are studied, leading to qualitatively different diagrams. Moreover, we investigate the presence of secondary bifurcations, we study a multi-parameter test case varying the type of compression and finally we show an application of empirical interpolation strategies.
  - In **Chapter 5** we study the rich branching behaviour of the Gross-Pitaevskii model in Bose-Einstein condensates. We reconstruct efficiently the multiple bifurcations in the diagram, comparing the performance of the affine-recovery techniques. A multi-parameter application is presented where we investigate the densities of the particle varying the strength of the confinement region.
  - In **Chapter 6** we review the class of hyperelastic models to study the buckling of beams. The Kirchhoff–Saint Venant and the neo-Hookean constitutive relations are presented and their buckling investigated with respect to different type of compressions (imposed by boundary conditions). We discussed different multi-parameter test cases involving physical quantities and geometrical parametrization for the length of the beam, for bi-dimensional and three-dimensional geometries. Finally, we present a real 3-D test case scenario motivated by industrial application.
  - In **Chapter 7** we investigate the Coanda effect, a bifurcating phenomena in fluid dynamics governed by the Navier-Stokes equations. The stability features of the coexisting solutions are studied and discussed in detail, since this model serves as a starting point for the last part.
- **Part III** - Towards more complex problems and emerging methodologies
  - In **Chapter 8** we briefly present the deflation method which we developed for the automated reconstruction of bifurcation diagrams in a reduced order context. Together with advanced continuation method we show how the reduced deflation algorithm properly reconstruct the full bifurcation diagram for the Navier-Stokes test case.

- In **Chapter 9** we model different optimal control problems to steer the bifurcating phenomena. A general framework for the reduction of nonlinear OCPs is presented and investigated in connection to distributed and boundary controls. We compare four different models, describing the way in which they influence the state profiles and the stability of the corresponding solutions.
- In **Chapter 10** we apply a non-intrusive strategy based on the POD-NN methodology to recover efficiently the bifurcation diagram, without the help of hyper-reduction strategies. We present a geometrical parametrization variant of the Coanda effect and a preliminary investigation of the bifurcating phenomena for flows inside a varying triangular domain. Finally, we develop an efficient detection tool for the discovery of the bifurcation points.

Part I

Theoretical and numerical background for bifurcation analysis

Chapter 2

Nonlinear analysis of PDEs and bifurcation problems

In this chapter, we will introduce the mathematical setting needed for the analysis performed throughout the thesis. This will serve us later to consider a unified framework where to cast all the different models that we have studied. Therefore, here we will review the main definitions for the functional setting of Partial Differential Equations (PDEs) in a general, parametric and nonlinear context. Moreover, we will focus on their singular points, which are the keystones of the bifurcation analysis, and on the stability theory, that we investigated in order to understand the properties of the physical phenomena considered.

2.1 Brief overview of Functional Analysis

In this section, we recall the basic definitions for the standard functional setting in nonlinear analysis. We are interested in the case of parametrized PDEs, thus let $\mathcal{P} \subset \mathbb{R}^P$ be the parameter space, a closed and bounded subset of the space $\mathbb{R}^P$ with $P \geq 1$. A parameter $\boldsymbol{\mu}$ is a point in the space $\mathcal{P}$ that can contain physical or geometrical information about the system. Then, we introduce the (reference) domain $\Omega \subset \mathbb{R}^d$, where $d = 2, 3$ is the spatial dimension.

Although time-dependent problems are of great interest, we will focus on steady problems, investigating the features of their equilibrium solutions. We can thus define the Hilbert space $\mathbb{X} := \mathbb{X}(\Omega)$ and its dual space $\mathbb{X}'$, as the space of linear and continuous functional over $\mathbb{X}$. The Hilbert space is equipped with the norm $\|V\|_{\mathbb{X}} = (V, V)_{\mathbb{X}}^{1/2}$ for any $V \in \mathbb{X}$, induced by its inner product $(\cdot, \cdot)_{\mathbb{X}}$. Within this setting, we denote the duality pairing between $\mathbb{X}'$ and $\mathbb{X}$ by means of

$$\langle G, V \rangle := \langle G, V \rangle_{\mathbb{X}'\mathbb{X}} \quad \forall\, G \in \mathbb{X}', V \in \mathbb{X}.$$

We have now all the ingredients to introduce the general framework of PDEs that we will investigate throughout this thesis. Let us denote with $G : \mathbb{X} \times \mathcal{P} \to \mathbb{X}'$ the

parametrized mapping which represents the nonlinear PDE, and thus the model under investigation.

The strong form of the parametric PDE problem read as: given $\boldsymbol{\mu} \in \mathcal{P}$, find $X(\boldsymbol{\mu}) \in \mathbb{X}$ such that

$$G(X(\boldsymbol{\mu}); \boldsymbol{\mu}) = 0 \quad \in \mathbb{X}'. \tag{2.1}$$

To study the solutions of the abstract equation above, a fundamental step is the analysis of its variational formulation, which will play a key role in the numerical approximation of the problem. Thus, we introduce the parametrized variational form $g(\cdot, \cdot; \boldsymbol{\mu}) : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ as

$$g(X, Y; \boldsymbol{\mu}) = \langle G(X; \boldsymbol{\mu}), Y \rangle \quad \forall\, X, Y \in \mathbb{X}, \tag{2.2}$$

and the weak formulation of equation (2.1) reads as: given $\boldsymbol{\mu} \in \mathcal{P}$, find $X(\boldsymbol{\mu}) \in \mathbb{X}$ such that

$$g(X(\boldsymbol{\mu}), Y; \boldsymbol{\mu}) = 0 \quad \forall\, Y \in \mathbb{X}. \tag{2.3}$$

In order to have a complete description of the problem formulation in (2.1), we need some suitable *boundary conditions* (BCs) on the domain Ω. Here, we will cover both the cases of homogeneous/inhomogeneous Dirichlet and Neumann conditions, depending on the test case at hand. Moreover, these BCs will be automatically embedded into (2.3) by an appropriate choice of the space where to seek the solution.

Having introduced the mathematical formulation of the problem, a starting point for the analysis is the investigation of its well-posedness. A PDE is said to be a *well-posed* problem if it is characterized by the existence and the uniqueness of the solution. Indeed, these properties are fundamental also for the numerical approximation of the model. Although these assumptions are often required, dealing with more realistic models can complicate the setting. In fact, one has to find a trade-off between the complexity of the model and its effective tractability. As we will see later on, complex nonlinear and parametric PDEs are proved to be more accurate in describing the physics, but the lack of good mathematical properties makes them difficult to approach, both theoretically and numerically.

Hence, in the following, we investigate when such properties are conserved and thus the problem admits a unique solution. On the contrary, in the next section, we will discuss the presence of critical points in which the well-posedness is lost, understanding how the model behaves and evolves in such cases.

Let us assume that the map G in equation (2.1) is continuously differentiable with respect to X and $\boldsymbol{\mu}$. Let $(\overline{X}; \overline{\boldsymbol{\mu}}) \in \mathbb{X} \times \mathcal{P}$ be a known solution, i.e. $G(\overline{X}; \overline{\boldsymbol{\mu}}) = 0$. Then, we denote by $D_X G(X; \boldsymbol{\mu}) : \mathbb{X} \to \mathbb{X}'$ and $D_{\boldsymbol{\mu}} G(X; \boldsymbol{\mu}) : \mathcal{P} \to \mathbb{X}'$ the partial derivatives of G on a generic point $(X, \boldsymbol{\mu}) \in \mathbb{X} \times \mathcal{P}$. When dealing with the approximation of (2.1), the strong assumption usually found in literature, which ensures the well-posedness of the problem, is that $D_X G(\overline{X}; \overline{\boldsymbol{\mu}}) : \mathbb{X} \to \mathbb{X}'$ is bijective. Indeed, the following result holds.

Theorem 2.1.1. *Let $B_r(\overline{X}), B_{\overline{r}}(\overline{\boldsymbol{\mu}})$ be two balls of radius r and $\overline{r}$ around $\overline{X}$ and $\overline{\boldsymbol{\mu}}$ respectively. Let $G : \mathbb{X} \times \mathcal{P} \to \mathbb{X}'$ be a C^1 map and assume that*

(1) $G(\overline{X};\overline{\boldsymbol{\mu}}) = 0$, *for the solution* $(\overline{X},\overline{\boldsymbol{\mu}}) \in \mathbb{X} \times \mathcal{P}$

(2) $D_X G(\overline{X};\overline{\boldsymbol{\mu}}) : \mathbb{X} \to \mathbb{X}'$ *is bijective*

then, there exist $r, \overline{r} > 0$ *and a unique solution* $X(\boldsymbol{\mu}) \in B_r(\overline{X}) \cap \mathbb{X}$ *such that*

$$G(X(\boldsymbol{\mu});\boldsymbol{\mu}) = 0 \quad \forall\, \boldsymbol{\mu} \in B_{\overline{r}}(\overline{\boldsymbol{\mu}}) \cap \mathcal{P}.$$

While this is a straightforward application of the Implicit Function Theorem [43, 158], it ensures the well-posedness of the problem and thus the existence of a *local branch of non-singular solutions*. Furthermore, we can rewrite the Frechét partial derivatives of G on $(Z,\boldsymbol{\mu}) \in \mathbb{X} \times \mathcal{P}$ with respect to X as

$$\mathrm{d}g[Z](X,Y;\boldsymbol{\mu}) = \langle D_X G(Z;\boldsymbol{\mu})X, Y\rangle \quad \forall\, X, Y \in \mathbb{X} \tag{2.4}$$

where we have introduced the parametrized variational form $\mathrm{d}g[Z](\cdot,\cdot;\boldsymbol{\mu})$. We now observe that assumption (2) in the previous theorem, can be reformulated in terms of the variational form $\mathrm{d}g[Z](\cdot,\cdot;\boldsymbol{\mu})$ itself, which is said to be

(2a) *continuous* on $\mathbb{X} \times \mathbb{X}$, if there exists a continuity constant $\overline{\gamma} > 0$ such that

$$\gamma(\boldsymbol{\mu}) = \sup_{X\in\mathbb{X}} \sup_{Y\in\mathbb{X}} \frac{\mathrm{d}g[Z](X,Y;\boldsymbol{\mu})}{\|X\|_{\mathbb{X}}\|Y\|_{\mathbb{X}}} > \overline{\gamma} \qquad \forall\, \boldsymbol{\mu} \in \mathcal{P}, \tag{2.5}$$

(2b) *inf-sup stable* on $\mathbb{X} \times \mathbb{X}$, if there exists an inf-sup constant $\overline{\beta} > 0$ such that

$$\beta(\boldsymbol{\mu}) = \inf_{X\in\mathbb{X}} \sup_{Y\in\mathbb{X}} \frac{\mathrm{d}g[Z](X,Y;\boldsymbol{\mu})}{\|X\|_{\mathbb{X}}\|Y\|_{\mathbb{X}}} \geq \overline{\beta} \qquad \forall\, \boldsymbol{\mu} \in \mathcal{P}, \tag{2.6}$$

and

$$\inf_{Y\in\mathbb{X}} \sup_{X\in\mathbb{X}} \frac{\mathrm{d}g[Z](X,Y;\boldsymbol{\mu})}{\|X\|_{\mathbb{X}}\|Y\|_{\mathbb{X}}} > 0 \qquad \forall\, \boldsymbol{\mu} \in \mathcal{P}. \tag{2.7}$$

Remark 2.1.1. The conditions (2.6) and (2.7) are equivalent to the injectivity and the surjectivity of the Frechét derivative $D_X G(Z;\boldsymbol{\mu})$, respectively. In particular, condition (2.6) can be reformulated as

$$\exists \beta(\boldsymbol{\mu}) > 0 \quad : \quad \|D_X G(Z;\boldsymbol{\mu})X\|_{\mathbb{X}'} \geq \beta(\boldsymbol{\mu})\|X\|_{\mathbb{X}} \tag{2.8}$$

that is, the Frechét derivative of G is *bounded below*. It is clear that (2.8) holds if and only if $D_X G(Z;\boldsymbol{\mu})$ is *injective*. On the other hand, condition (2.7) can be easily restated as the injectivity of the *adjoint* operator of $D_X G(Z;\boldsymbol{\mu})$, which can be expressed in terms of its variational form $\mathrm{d}g$ as

$$Y \in \mathbb{X}, \quad \mathrm{d}g[Z](X,Y;\boldsymbol{\mu}) = 0 \quad \forall\, X \in \mathbb{X} \quad \Rightarrow \quad Y = 0\,. \tag{2.9}$$

Therefore, we observe that condition (2.7) is actually stating that $D_X G(Z;\boldsymbol{\mu})$ is *surjective*, since the injectivity of the adjoint of an operator is equivalent to the surjectivity of the operator itself.

Remark 2.1.2. When dealing with linear parametrized PDEs, it is clear that (e.g. for an unforced problem) it holds

$$\mathrm{d}g[Z](X, Y; \boldsymbol{\mu}) = g(X, Y; \boldsymbol{\mu}) \qquad \forall X, Y, Z \in \mathbb{X}, \forall \boldsymbol{\mu} \in \mathcal{P}. \tag{2.10}$$

Hence, the general assumptions in (2.5),(2.6) and (2.7) fall back on the standard ones required by the well-known theorems of Lax-Milgram and Nečas [54, 140].

In conclusion, the bijectivity assumption in Theorem 2.1.1, and equivalently the injectivity and surjectivity conditions (2.8) and (2.9), are the key ingredients to have a well-posed problem. In the following, we will see what happens when these hypotheses are no longer valid at some point in the parameter space.

2.2 A glimpse on Bifurcation Theory

In this section, we want to provide a brief description of bifurcating behaviours and their mathematical investigation. Before moving to the theory, let us start with a non-mathematical introductory explanation. In the world around us, we can observe many phenomena that involve either gradual or sudden changes, and we characterize them as quantitative or qualitative, respectively. As an example, we can think about the supported beam under compression depicted in Figure 2.1.

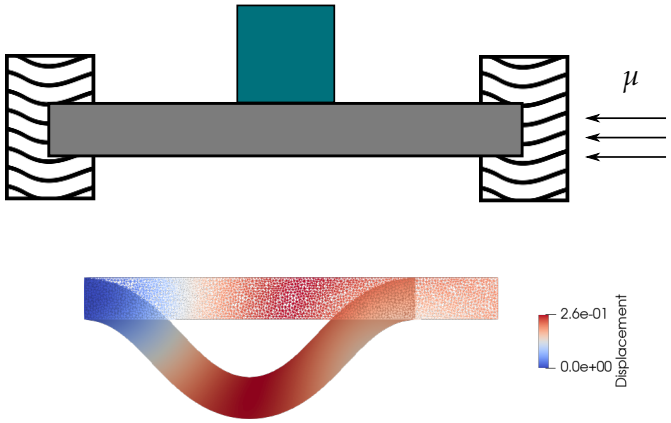


Figure 2.1: A beam subject to a load and its buckled configuration.

When the load applied to the beam, denoted with μ, is sufficiently small, we observe a slight deformation that corresponds to a quantitative change. Of course, this deformation depends on the magnitude of the load and on the specific material properties of the beam. Hence, we expect that a small perturbation of μ leads to a configuration with the same qualitative feature.

At the same time, we empirically know that it exists a critical value for the load, that we call μ^*, after which the model does not conserve the original qualitative be-

haviour. Practically speaking, if the load is big enough, the beam can not sustain the compression and it buckles. This is a clear example of a qualitative change.

Therefore, we will denote the qualitative changes that are not stable under small perturbations of a parameter μ as *bifurcation phenomena*, and the points μ^* in which they occur as *bifurcation points* [142, 89, 93]. Usually, we can distinguish between different qualitative states by means of their geometrical shape or pattern configuration. Indeed, two of the most relevant features that may change in presence of a bifurcation phenomenon are the stability and the symmetry of the resulting states.

Now, let us translate the example above in a more precise mathematical language. The study of the solution set for complex nonlinear PDEs such as (2.1) is usually very complicated both theoretically and numerically. The aim is to investigate the solutions of a specific model varying a one dimensional parameter, i.e. the *bifurcation parameter*, which is the responsible for the bifurcating phenomena. Correlated to the qualitative change of the solutions is the non-uniqueness behaviour, in fact, as we will see, in such situations the model admits different solutions for the same value of the parameter. We will refer to the set of solutions with the same qualitative properties as *branch*. As we said before, when the parameter changes slightly, we expect that a stable solution evolves continuously in a unique manner. In fact, this is usually guaranteed by Theorem 2.1.1 and follows as a corollary of the Implicit Function Theorem.

The situation changes drastically if the inf-sup stability of the model, ensured by (2.6) and (2.7), is lost. This happens when the bifurcation parameter reaches a critical value μ^*, for which the system admits the existence of a qualitatively different solution that bifurcates from the previous stable one. Within this thesis we will only treat codimension one bifurcation, which means that the difference between the dimension of the parameter space and the dimension of the corresponding bifurcation boundary (the manifold generated by the critical values) is one. Hence, for the sake of clarity, during the rest of this section we will restrict our exposition to the case in which the first component of $\boldsymbol{\mu} \in \mathcal{P} \subset \mathbb{R}^P$ is the bifurcation parameter μ, while the remaining $P-1$ parameters are considered fixed. For this reason, with a little abuse of notation, here we will consider $P = 1$ and the bifurcating parameter μ varying in $\mathcal{P} \subset \mathbb{R}$.

Following [4, 3], we consider the problem (2.1) and assume that $X = 0$ is a solution for every parameter $\mu \in \mathcal{P}$, namely

$$G(0;\mu) = 0, \quad \forall\, \mu \in \mathcal{P}$$

which will be referred to as the *trivial solution*. Then, we can define the set of *non-trivial solutions* of (2.1) as

$$\mathcal{S} = \{(X,\mu) \in \mathbb{X} \times \mathcal{P} \mid X \neq 0,\ G(X;\mu) = 0\}\,. \tag{2.11}$$

Remark 2.2.1. For the sake of simplicity, we chose to present here the bifurcations that originate form the trivial solution. Indeed, many times, and in particular among the test cases we treated, one has a non-trivial *ground state* solution, which loses its uniqueness branching to another non-trivial state. It is remarkable to note that such situation

can happen in a nested fashion. Increasing in complexity, one can find a bifurcation emerging from an already bifurcated configuration. This phenomenon is usually called *secondary bifurcation*, and an example of that will be shown in Section 4.2.3.

As we understood from the example reported at the beginning, there could exist some values for the parameter μ such that one or more non-trivial solutions branching off from the trivial one. These are the critical points that we have introduced before, and in the context of bifurcation theory we define them as in [4, 3].

Definition 2.2.1. A parameter value $\mu^* \in \mathcal{P}$ is a *bifurcation point* for (2.1) if there exists a sequence $(X_n, \mu_n) \in \mathbb{X} \times \mathcal{P}$, with $X_n \neq 0$ such that

- $G(X_n; \mu_n) = 0$
- $(X_n, \mu_n) \to (0, \mu^*)$.

Remark 2.2.2. Equivalently, one can ask that the couple $(0, \mu^*)$ belongs to the closure of $\mathcal{S}$ in $\mathbb{X} \times \mathcal{P}$, which means that in any neighbourhood of $(0, \mu^*)$ there exists a point $(X, \mu) \in \mathcal{S}$.

Bifurcation is thus a paradigm for non-uniqueness in nonlinear analysis, and a necessary condition is the failure of the Implicit Function Theorem, as stated by the next proposition.

Proposition 2.2.1. *A necessary condition for μ^* to be a bifurcation point for G is that the partial Frechét derivative $D_X G(0; \mu^*)$ is not invertible.*

Proof. By contradiction, if $D_X G(0; \mu^*)$ is invertible, then a straightforward application of the Implicit Function Theorem states that it exists a neighbourhood $\mathbb{U} \times \mathcal{O}$ of $(0, \mu^*)$ such that

$$G(X; \mu) = 0, \text{ with } (X, \mu) \in \mathbb{U} \times \mathcal{O} \quad \text{if and only if} \quad X = 0.$$

Therefore, μ^* is not a bifurcation point for G. □

Remark 2.2.3. We highlight that the inf-sup stability property, that should guarantee the well-posedness of the PDE, is actually no longer valid for bifurcating phenomena. Indeed, in the bifurcation point μ^* the inf-sup constant $\overline{\beta}$ becomes zero, and the Frechét derivative $D_X G(X; \mu)$ fails to be invertible. Therefore, unless specified otherwise, we will talk about well posed problem as long as $\mu \neq \mu^*$.

Remark 2.2.4. Consider a differentiable $F : \mathbb{X} \to \mathbb{X}$ and the PDE of the form

$$G(X; \mu) = \mu X - F(X),$$

and hence such that $D_X G(0; \mu^*) = \mu^* I - F'(0)$, with I the identity operator. It is clear that the result in Proposition 2.2.1 becomes: if μ^* is a bifurcation point for $\mu X - F(X)$, then it belongs to the spectrum of $F'(0)$. In general, the converse is not true. Indeed, if μ^* belongs to the spectrum of $F'(0)$, then μ^* is not necessarily a bifurcation point.

Remark 2.2.5 (connected to the previous). In the particular case in which

$$F(X) = AX \quad \text{with} \quad A : \mathbb{X} \to \mathbb{X}$$

a linear and continuous map, we can obtain a more precise characterization of the bifurcation points. Indeed, since $G'(0) = A$, all the eigenvalues of G are clearly bifurcation points. Moreover, it is straightforward to prove that μ^* is a bifurcation point for G if and only if μ^* belongs to the closure of the eigenvalues of A.

The aim of the *bifurcation theory* is thus to provide a mathematical description of the bifurcating scenario that can be observed in physical systems and experiments. The understanding of the global behaviour, both from the theoretical and numerical viewpoints, is usually supported by the reconstruction of some scalar characteristic value, which we will denote by $s(X(\mu))$. Of course, this highly depends on the physics under investigation. A simple choice can be a representative norm of the solution, while other examples are e.g.: (1) the maximum displacement of a compressed beam, (2) a point-wise evaluation of the velocity for a jet-flow, (3) the number of bosons in a Bose–Einstein condensate.

Given $(X(\mu), \mu)$ a solution of (2.1) for each instance of the parameter, one can draw a plot of $s(X(\mu))$ versus μ which in literature is known as *bifurcation diagram*. Therefore, the existence of different solutions for the same values of the parameter will result in the presence of multiple branch depicted on the diagram. Different choices of $s(\cdot)$ result in possibly different plots, thus, to avoid confusion, such a function has to be chosen carefully and has to represent the main features while changing the qualitative behaviour of the resulting configuration.

It is clear that there exist many different bifurcation phenomena, and usually complex nonlinear problems are characterized by one (or more) of those. In fact, the literature is full of examples of physical systems with qualitatively changing behaviour. Each one of these bifurcating phenomena has a specific peculiarity, but in the field of bifurcation theory they are usually categorized by some common features. Among the others, the most known ones are: *turning point* (or fold) bifurcations, *transcritical* bifurcations, *pitchfork* bifurcations and *Hopf* bifurcations. While the latter are usually of interest when dealing with time-dependent problems, in this work we will focus on models that exhibit pitchfork bifurcations.

Here, a much simpler one dimensional benchmark can be used to provide a synthetic and clear example. Let us consider the Ordinary Differential Equation (ODE),

$$\frac{du}{dt} = G(u;\mu) \quad \text{with} \quad G(u;\mu) = \mu u - u^3 \tag{2.12}$$

introduced by Landau to describe the effects of hydrodynamical instabilities. From the standard ODEs stability theory we can impose $G(u;\mu) = 0$, and find the following stationary solutions

$$\overline{u}_1 = 0, \quad \overline{u}_2 = \sqrt{\mu}, \quad \overline{u}_3 = -\sqrt{\mu}.$$

It is clear that the number of the existing solutions depends on the parameter μ. Indeed, for $\mu \leq 0$ we have the existence of a unique (actually trivial) solution, while for

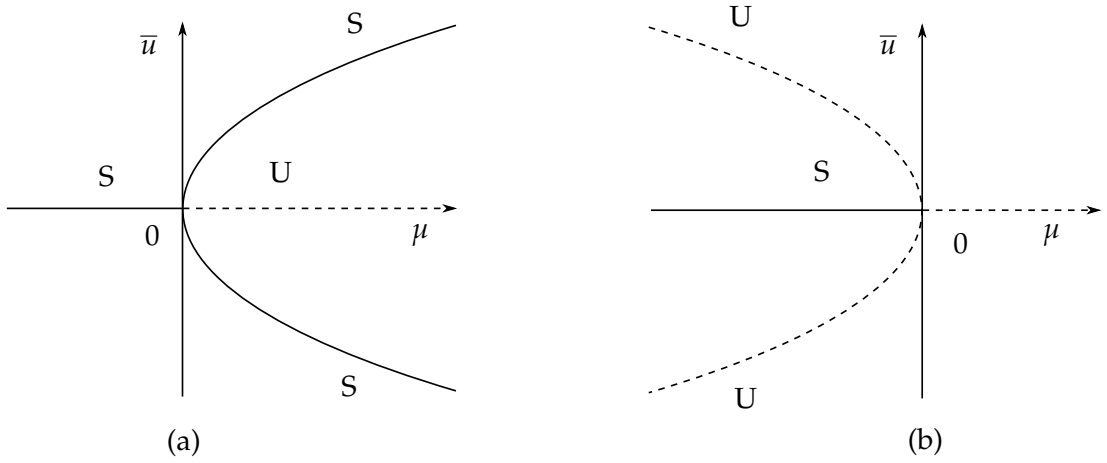


Figure 2.2: Supercritical (a) and subcritical (b) pitchfork bifurcation, with stable and unstable branches denoted by S and U respectively.

$\mu > 0$ we have three different branches, corresponding to the three non-trivial solutions $\overline{u}_1$, $\overline{u}_2$, and $\overline{u}_3$. From the previous analysis, we can thus identify the bifurcation point as $\mu^* = 0$. Indeed, we can observe that the necessary condition in Proposition 2.2.1 is satisfied, being $D_u G|_{(0,\mu^*)} = (\mu - 3u^2)|_{(0,\mu^*)} = 0$.

We sketch this scenario in Figure 2.2, that represents an example of what is called *supercritical pitchfork bifurcation*. The word supercritical refers to the fact that the branches evolve to the right of the critical point μ^*. A similar configuration, called *subcritical pitchfork bifurcation*, can be obtained changing the sign of the cubic term in (2.12). Conversely, the latter is characterized by the non-uniqueness behaviour for $\mu < \mu^*$.

Furthermore, it is interesting to note the symmetry of the two plots in Figure 2.2 with respect to the μ-axis. This is due to the fact that the two possible bifurcated states describe the same configuration in two different reference systems. This will be clear considering the physical setting in Sections 4 and 7. Another key feature of pitchfork bifurcations is that usually branching states have less symmetry than the ground one. For this reason such phenomena are typically called symmetry-breaking bifurcations.

Last but not least, a fundamental property to be studied during the investigation of the bifurcation diagram is the stability property of a branch. As we can see in Figure 2.2, already in this simple setting the qualitative change in the solution corresponds to a stability change for the trivial branch. More specifically, the branch corresponding to the solution $u = 0$ remains stable (solid line) until it exists as unique solution. Once passing μ^*, the two symmetric branches inherit the stability feature, while the trivial one becomes unstable (dotted line).

The stability properties of the model can be mathematically investigated through an extensive study of its spectral features, as we will describe in the next section.

2.3 Stability analysis and eigenvalue problems

In this section, we review the basic notions for the stability analysis of a general PDE of the form (2.1), and its connection with the eigenvalues problem [83, 54, 93]. The word "stability" is widely used in many mathematical contexts, here we want to investigate the physical stability of all the possible configurations existing as solutions for a certain value of the parameter. For this reason, we can consider for the rest of the section a parameter independent version of the problem (2.1), i.e. once fixed the value $\boldsymbol{\mu} = \overline{\boldsymbol{\mu}}$ we want to investigate the behaviour of a stationary solution $X = X(\overline{\boldsymbol{\mu}})$ with respect to small perturbations. Nevertheless, it is clear that this is equivalent to the study of the equilibrium solution of a time dependent PDE of the form

$$\frac{\partial X}{\partial t} = G(X). \tag{2.13}$$

Stability theory for PDE mimics classical finite dimensional ODEs theory, with the fundamental difference that the former inherits a state space that is infinite dimensional. Therefore, one has to carefully translate the concepts from these two settings, but since we will deal with the stability of PDEs in a discretized setting, a review of the finite dimensional setting is enlightening.

As an example, let us set $\mathbb{X} = \mathbb{R}^d$, $X = u$ and consider the non-parametric ODE given by

$$\frac{\partial u}{\partial t} = G(u) \tag{2.14}$$

where $u = u(t) \in \mathbb{R}^d$, and let us denote with u_0 its stationary solution, so that $G(u_0) = 0$. This means that if we start from the initial condition $u(0) = u_0$, the autonomous system evolves always as the initial configuration u_0. On the contrary, to study the stability of u_0 as solution of (2.14), it is standard to introduce a perturbation $v(t)$ on the initial value u_0, investigating the behaviour of the solution $u(t) = u_0 + v(t)$ with initial condition $u(0) = u_1$.

We can characterize the evolution of the perturbation, and thus the stability property of the solution u_0, by means of the following linear system

$$\frac{\partial v}{\partial t} = D_u G(u_0) v, \quad \text{with } v_0 := v(0) = u_1 - u_0, \tag{2.15}$$

which solution can be expressed by the well known Duhamel's formula [54]. Hence, since $v(t) = e^{D_v G(u_0) t} v_0$ is the unique solution to (2.15), the stability of the stationary solution u_0 is completely determined by the eigenvalues of the matrix $D_v G(u_0)$, as stated by the following classical theorem [83].

Theorem 2.3.1 (Stable Manifold Theorem). *Assume that it holds* $\mathrm{Re}(\sigma) < -\kappa$ *for every* $\sigma \in \mathbb{C}$ *eigenvalue of* $D_v G(u_0)$ *with* $\kappa > 0$. *Then there exists* $\delta > 0$ *and* $C \geq 1$ *such that if* $|v_0| < \delta$ *then the solution* v *of* (2.15) *exists for all* $t > 0$ *and satisfies the exponential decay estimate* $|v(t)| \leq C e^{-\kappa |v_0|}$ *for all* $t \geq 0$. *Furthermore,* u_0 *is a stable solution of* (2.14).

Remark 2.3.1. Conversely, it is straightforward to see from the previous theorem that, if $D_v G(u_0)$ has an eigenvalue with positive real part, then u_0 is an unstable solution of (2.14). More precisely, there exists an $\varepsilon > 0$ such that for every $\delta \in (0, \varepsilon)$ there exists an initial data v_0 with $|v_0| < \delta$ such that the associated solution $v(t)$ satisfies $|v(T)| > \varepsilon$ for some finite $T > 0$.

As in the case of ODEs theory, also the PDEs dynamic for the general nonlinear problem (2.13) is mainly governed by the behaviour of the linearized system. Indeed, as we have seen from the Proposition 2.2.1 and successive remarks, the spectral properties of the linearized operator play a key role in this context. We can thus go back to (2.13), and similarly to what we have done before, we consider a perturbation $V(t)$ of the equilibrium solution X_0. Again, if we take a small initial data X_1 and let $X(t) = X_0 + V(t)$ be the associated local solution, then as $\|V(t)\|_{\mathbb{X}}$ remains sufficiently small we expect the dynamic of (2.13) to be well approximated by the linear evolution equation

$$\frac{\partial V}{\partial t} = D_X G(X_0) V, \quad \text{with } V_0 := V(0) = X_1 - X_0. \tag{2.16}$$

Analogously, we can give a proper definition of stable solution for the infinite dimensional setting.

Definition 2.3.1. Let X_0 be an equilibrium of (2.13). We say that X_0 is a (linearly) stable solution for (2.13) if $V = 0$ is a stable solution of (2.16), i.e. if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that for every $V_0 \in \mathbb{X}$ with $\|V_0\|_{\mathbb{X}} < \delta$, the solution $V(t)$ of (2.16) with $V(0) = V_0$ satisfies $\|V(t)\|_{\mathbb{X}} < \varepsilon$ for all $t \geq 0$.

Moreover, recalling the Proposition 2.2.1, we can give the following definition

Definition 2.3.2. Let L, I be a linear and the identity operators, respectively. We denote by $\rho(L)$ the *resolvent* of L on $\mathbb{X}$ as the set of all $\sigma \in \mathbb{C}$ such that $L - \sigma I$ is invertible with bounded inverse. Furthermore, the *spectrum* of L on $\mathbb{X}$, denoted by $\Sigma(L)$ is defined as $\Sigma(L) := \mathbb{C} \setminus \rho(L)$. Finally, $\sigma \in \Sigma(L)$ is an eigenvalue of L if $\mathrm{Ker}(L - \sigma I) \neq 0$.

Therefore, the main difference passing from the finite to the infinite dimensional space is that the concept of stability does not depend only on the eigenvalues, but rather on the entire spectrum. In fact, in view of the Remark 2.2.5, it is now clear that, even considering $G(X; \mu) = \mu X - AX$ with A linear and continuous, there might be points μ in the spectrum of A that are not bifurcation points for G. On the other hand, μ^* can be a bifurcation point for G without being an eigenvalue of A.

We can thus conclude this section by remarking a necessary condition, based on the above considerations, for the stability of an equilibrium solution of (2.13).

Proposition 2.3.1. *An equilibrium solution X_0 of* (2.13) *is said to be stable if $\Sigma(L) \cap \{\sigma \in \mathbb{C} : \mathrm{Re}(\sigma) > 0\} = \emptyset$. Else, X_0 is said to be unstable.*

Therefore, the stability properties of the models considered in this thesis will be analysed through the eigenvalues problems presented.

It is clear that in the context of nonlinear parametrized PDEs, such investigation, together with the analysis of the bifurcation diagrams, can only be achieved by means

of numerical approximation methods. Therefore, after having presented in this chapter the mathematical background for the analysis of bifurcating parametrized problems and the existing relation between the stability analysis and the investigation of the spectral properties of the models, now we can finally move to the presentation of the numerical methodology we have developed.

In the next chapter, we will describe the different techniques, with a focus on the computational complexity, that will allow us to discover numerically the singularities of the physical phenomena under investigation.

Chapter 3

Numerical approximation of bifurcating phenomena

As we said in the previous chapter, we are interested in nonlinear and parametric PDEs where the solution for a given parameter $\boldsymbol{\mu}$ exist, but may not be unique [24, 27]. Indeed, the local (w.r.t. the parameter) well-posedness of the problem (2.1) relies on the local spectral property of its Frechét derivative $D_X G(\cdot;\cdot)$, and if it fails to fulfill certain assumptions, the behaviour of the model has to be investigated more carefully.

In the bifurcating context, we can mimic the definition of the set of non-trivial solutions in (2.11), introducing the terminology of *solution branch* as the set of the solutions with the same qualitative features for different values of the parameter $\boldsymbol{\mu}$. Hence, we will denote a solution branch as

$$\mathcal{M} = \{X(\boldsymbol{\mu}) \in \mathbb{X} \mid G(X(\boldsymbol{\mu});\boldsymbol{\mu}) = 0,\ \boldsymbol{\mu} \in \mathcal{P}\}. \tag{3.1}$$

Since we deal with the non-uniqueness of the solution at fixed $\boldsymbol{\mu}$, we expect the system to be characterized by the existence of multiple solution branches. We can thus define the set of all the existing solutions to (2.1) as the set of the solution branches

$$\mathcal{X} = \bigcup_{i=1}^{k} \mathcal{M}_i := \bigcup_{i=1}^{k} \{X_i(\boldsymbol{\mu}) \in \mathbb{X} \mid G(X_i(\boldsymbol{\mu});\boldsymbol{\mu}) = 0,\ \boldsymbol{\mu} \in \mathcal{P}\}. \tag{3.2}$$

We remark that when $\boldsymbol{\mu} \notin \mathcal{P}$, the case $k = 1$ can be seen as the well-posed problem and $\mathcal{X}$ is often referred to as the *solution manifold* [74].

Remark 3.0.1. Although we are considering in (3.1) only the set of a single qualitative state, we highlight that since we are interested in the bifurcating behaviour, the set $\mathcal{X}$ usually contains the singularity $\boldsymbol{\mu}^*$. Hence, contrarily to what is usually found in literature, we are not restricting ourselves to the case of a non-singular branch of solution. For this reason, we will extend the definition of each solution branch to the pre-bifurcation range of the parameter, where the solution is unique.

In this chapter, we want to describe the numerical approximation of a general nonlinear bifurcating equation. In the first part, we will review the branch-wise approach we developed in order to recover the bifurcation diagram. The stability properties of the solutions are investigated as well, with the help of different strategies involving generalized eigenvalue problems. We will refer to these steps as the *high fidelity* approximation, since we deal with high dimensional and time consuming procedures. In order to avoid this computational complexity, in the second part, we will present the Reduced Order Models that we applied to reduce the dimensionality of the system, while allowing for an efficient approximation of the bifurcating phenomena through the reduced branch-wise algorithm we developed.

3.1 High fidelity approximation

The standard techniques we will review in this section are associated to the context of "high fidelity" methods, in the sense that they are representing the first level of discretization, where a lot of degrees of freedom have to be involved, possibly causing a too costly computation.

The strategy we have developed for the reconstruction and the analysis of the whole bifurcation diagram is based on the combination of three well-studied methodologies: the Galerkin Finite Element (FE) method, the Newton-Kantorovich method and a Continuation method. These will serve us to discretize, linearize and continue each branch. A discretization method is certainly needed for the numerical analysis of a PDEs, therefore in the next subsections we will review the Galerkin Finite Element method in the context of the general nonlinear PDE in (2.1).

Through the rest of this section we will always be interested in recovering the full bifurcation diagram for the model under consideration, thus the numerical approximation will involve the computation of the discretized counterpart of (3.2). To this end, we pursued an high fidelity *branch-wise* approach. In other words: we aimed at reconstructing one fixed branch $\mathcal{M} \subset \mathcal{X}$, corresponding to some $i \in \{1, \dots, k\}$, at time. The global behaviour is thus obtained approximating its solutions for all the values in $\mathcal{P}$ and then moving to the next solution branch in $\mathcal{X}$.

3.1.1 Galerkin Finite Element method

We introduce the standard notion for the discretization of a generic problem of the form (2.1) with the Galerkin FE method [133, 129].

Let $\mathbb{X}_h$ be a family of finite dimensional spaces, dependent from the discretization parameter h, such that $\mathbb{X}_h \subset \mathbb{X}$, with inherited inner product $(X, Y)_{\mathbb{X}_h} = (X, Y)_{\mathbb{X}}$ and norm $\|X\|_{\mathbb{X}_h} = \|X\|_{\mathbb{X}}$ for $X, Y \in \mathbb{X}_h$. Furthermore, we will assume that once fixed the discretization parameter h, the computational domain $\Omega_h = \text{int}\left(\bigcup_{K \in \mathcal{T}_h} K\right)$, i.e. the interior of the union of the elements K of the triangulation $\mathcal{T}_h$, coincides exactly with the real domain Ω, an open and bounded regular domain in $\mathbb{R}^d$, with $d = 2, 3$. Then, let us consider the discretized spatial space $\mathbb{X}_{\mathcal{N}} = \mathbb{X}_h \cap \mathbb{K}_r$ of dimension $\dim(\mathbb{X}_{\mathcal{N}}) =$

$\mathcal{N} < \infty$, which denotes the high fidelity degrees of freedom, where

$$\mathbb{K}_r = \{V \in C^0(\overline{\Omega}) \ : \ V|_K \in \mathbb{P}_r, \ \ \forall K \in \mathcal{T}_h\},$$

and $\mathbb{P}_r$ is the space of all the polynomials of degree at most equal to r on the single element of the triangulation $\mathcal{T}_h$.

Then, we can go back to the weak form (2.3) of the equation (2.1), and project it into the finite dimensional vector space $\mathbb{X}_\mathcal{N}$. The Galerkin FE method reads as: given $\boldsymbol{\mu} \in \mathcal{P}$, we seek $X_\mathcal{N} := X_\mathcal{N}(\boldsymbol{\mu}) \in \mathbb{X}_\mathcal{N}$ that satisfies

$$\langle G(X_\mathcal{N}; \boldsymbol{\mu}), Y_\mathcal{N} \rangle = g(X_\mathcal{N}, Y_\mathcal{N}; \boldsymbol{\mu}) = 0 \ , \quad \forall \, Y_\mathcal{N} \in \mathbb{X}_\mathcal{N}. \tag{3.3}$$

The well-posedness of the discrete problem (3.3) is not inherited from the continuum one, also assuming the inclusion property $\mathbb{X}_\mathcal{N} \subset \mathbb{X}$. Indeed, while it is clear that the continuity hypothesis (2.5) is still satisfied also at the finite dimensional level, the same is not true for the inf-sup stability. Indeed, the discrete counterpart of the continuity assumption (2.5) can be reformulated as the existence of a constant $\overline{\gamma}_\mathcal{N} > 0$ such that

$$\gamma_\mathcal{N}(\boldsymbol{\mu}) = \sup_{X_\mathcal{N} \in \mathbb{X}_\mathcal{N}} \sup_{Y_\mathcal{N} \in \mathbb{X}_\mathcal{N}} \frac{\mathrm{d}g[Z_\mathcal{N}](X_\mathcal{N}, Y_\mathcal{N}; \boldsymbol{\mu})}{\|X_\mathcal{N}\|_{\mathbb{X}_\mathcal{N}} \|Y_\mathcal{N}\|_{\mathbb{X}_\mathcal{N}}} \leq \overline{\gamma}_\mathcal{N} \ , \tag{3.4}$$

while the inclusion property is only a necessary but not sufficient condition for (2.6) and (2.7) to hold at the discrete level. Hence, an additional assumption has to be required for the inf-sup discrete stability of the variational form $\mathrm{d}g$. Relying on the Babuška theorem [6], we will also assume that at the discrete level $\mathrm{d}g[Z_\mathcal{N}](\cdot, \cdot; \boldsymbol{\mu})$ is inf-sup stable at $\boldsymbol{\mu}$, i.e. there exists a constant $\overline{\beta}_\mathcal{N} > 0$ such that

$$\beta_\mathcal{N}(\boldsymbol{\mu}) = \inf_{X_\mathcal{N} \in \mathbb{X}_\mathcal{N}} \sup_{Y_\mathcal{N} \in \mathbb{X}_\mathcal{N}} \frac{\mathrm{d}g[Z_\mathcal{N}](X_\mathcal{N}, Y_\mathcal{N}; \boldsymbol{\mu})}{\|X_\mathcal{N}\|_{\mathbb{X}_\mathcal{N}} \|Y_\mathcal{N}\|_{\mathbb{X}_\mathcal{N}}} \geq \overline{\beta}_\mathcal{N} \ . \tag{3.5}$$

Once again, we highlight that also at the discrete level the well-posedness is lost at the bifurcation points. This means that $\beta_\mathcal{N}(\boldsymbol{\mu})$ reaches the value zero as $\boldsymbol{\mu}$ approaches $\boldsymbol{\mu}^*$, causing the non invertibility of the Frechét derivative.

From the algebraic point of view, we denote with $\{E^j\}_{j=1}^\mathcal{N}$ a Lagrangian basis for $\mathbb{X}_\mathcal{N}$, such that we can write every element $X_\mathcal{N}(\boldsymbol{\mu}) \in \mathbb{X}_\mathcal{N}$ as

$$X_\mathcal{N}(\boldsymbol{\mu}) = \sum_{j=1}^{\mathcal{N}} X_\mathcal{N}^{(j)}(\boldsymbol{\mu}) E^j \ , \tag{3.6}$$

and denote the solution vector as $\mathsf{X}_\mathcal{N}(\boldsymbol{\mu}) = \{X_\mathcal{N}^{(j)}(\boldsymbol{\mu})\}_{j=1}^\mathcal{N}$.

We remark that usually the solution $X_\mathcal{N}$ contains multiple variable/components, in which case $\mathcal{N}$ has to be considered as the sum of the degrees of freedom of each variable/component. Moreover, unless otherwise specified, we will assume that all the components are discretized on the same portion of the computational domain, with the same degrees of freedom. Despite this, we will still denote the dimension of the

FE space as $\mathcal{N}$, keeping in mind that this actually is representing the dimension of each component.

We then lead back to the study of the solution $\mathsf{X}_{\mathcal{N}}(\boldsymbol{\mu}) \in \mathbb{R}^{\mathcal{N}}$ of the nonlinear system

$$g\left(\sum_{j=1}^{\mathcal{N}} X_{\mathcal{N}}^{(j)}(\boldsymbol{\mu}) E^j, E^i; \boldsymbol{\mu}\right) = 0\,, \quad \forall\, i = 1, \dots, \mathcal{N} \tag{3.7}$$

which corresponds to the solution of the algebraic form of (3.3) given by

$$\mathsf{G}_{\mathcal{N}}(\mathsf{X}_{\mathcal{N}}(\boldsymbol{\mu}); \boldsymbol{\mu}) = 0 \quad \text{in } \mathbb{R}^{\mathcal{N}} \tag{3.8}$$

where the *high fidelity residual vector* $\mathsf{G}_{\mathcal{N}}$ is defined as

$$(\mathsf{G}_{\mathcal{N}}(\mathsf{X}_{\mathcal{N}}(\boldsymbol{\mu}); \boldsymbol{\mu}))_i = g(\mathsf{X}_{\mathcal{N}}(\boldsymbol{\mu}), E^i; \boldsymbol{\mu}) \quad \forall\, i = 1, \dots, \mathcal{N}.$$

In the next subsection, we will describe how to handle the nonlinearity appearing in equation (3.8).

3.1.2 Newton-Kantorovich method

Since the discretized weak problem (3.3) still involves a nonlinear structure, we have to deal with a method that linearizes it, in order to be effectively treated and solved by means of the Galerkin FE method. The nonlinear solver chosen to linearize the weak formulation in (3.3) is the well-known Newton-Kantorovich method [43, 130], which reads as follows: given $\boldsymbol{\mu} \in \mathcal{P}$ and chosen an initial guess $X_{\mathcal{N}}^0(\boldsymbol{\mu}) \in \mathbb{X}_{\mathcal{N}}$, for every $k = 0, 1, \dots$ we seek the variation $\delta X_{\mathcal{N}} \in \mathbb{X}_{\mathcal{N}}$ such that

$$\mathrm{d}g[X_{\mathcal{N}}^k(\boldsymbol{\mu})](\delta X_{\mathcal{N}}, Y_{\mathcal{N}}; \boldsymbol{\mu}) = g(X_{\mathcal{N}}^k(\boldsymbol{\mu}), Y_{\mathcal{N}}; \boldsymbol{\mu})\,, \quad \forall\, Y_{\mathcal{N}} \in \mathbb{X}_{\mathcal{N}}, \tag{3.9}$$

then we update the solution at the iteration $k+1$ as

$$X_{\mathcal{N}}^{k+1}(\boldsymbol{\mu}) = X_{\mathcal{N}}^k(\boldsymbol{\mu}) - \delta X_{\mathcal{N}}$$

and repeat these steps until an appropriate stopping criterion is verified.

Remark 3.1.1. Among the others, the most relevant stopping criteria are:

- the *increment* control on the relative norm of the variation $\delta X_{\mathcal{N}}$, which checks for small steps of the method;
- the *residual* control on the absolute/relative norm of the high fidelity residual $\mathsf{G}_{\mathcal{N}}$.

Regarding the convergence of the Newton-Kantorovich method, we can rely on the following theorem [158, 43].

Theorem 3.1.1. *Let $X_{\mathcal{N}}(\boldsymbol{\mu})$ a solution of the equation* (3.3)*, such that the continuity* (3.4) *and inf-sup* (3.5) *properties are satisfied. Let $\mathrm{d}g[X_{\mathcal{N}}(\boldsymbol{\mu})](\cdot,\cdot;\boldsymbol{\mu})$ be locally Lipschitz continuous at $X_{\mathcal{N}}(\boldsymbol{\mu})$. Then, if the initial condition $X^0_{\mathcal{N}}(\boldsymbol{\mu})$ is sufficiently near the solution, the sequence $\{X^k_{\mathcal{N}}(\boldsymbol{\mu})\}_k$ generated by the Newton method converges quadratically to $X_{\mathcal{N}}(\boldsymbol{\mu})$.*

Recalling the unique decomposition of the solution $X_{\mathcal{N}}(\boldsymbol{\mu})$ in terms of the basis functions $\{E^j\}_{j=1}^{\mathcal{N}}$, we can proceed with the algebraic form of the linearised system in (3.9). Hence, the k-th step of the Newton method combined with the Galerkin FE method and applied to (3.8) reads: fixed $\boldsymbol{\mu} \in \mathcal{P}$, find $\delta \mathsf{X}_{\mathcal{N}} \in \mathbb{R}^{\mathcal{N}}$ such that

$$\mathsf{J}_{\mathcal{N}}(\mathsf{X}^k_{\mathcal{N}}(\boldsymbol{\mu});\boldsymbol{\mu})\delta\mathsf{X}_{\mathcal{N}} = \mathsf{G}_{\mathcal{N}}(\mathsf{X}^k_{\mathcal{N}}(\boldsymbol{\mu});\boldsymbol{\mu}) \,, \tag{3.10}$$

where the Jacobian matrix in $\mathbb{R}^{\mathcal{N}\times\mathcal{N}}$ is defined as

$$(\mathsf{J}_{\mathcal{N}}(\mathsf{X}^k_{\mathcal{N}}(\boldsymbol{\mu});\boldsymbol{\mu}))_{ij} = \mathrm{d}g[\mathsf{X}^k_{\mathcal{N}}(\boldsymbol{\mu})](E^j, E^i;\boldsymbol{\mu}), \quad \text{for all} \quad i,j = 1,\dots,\mathcal{N}\,. \tag{3.11}$$

It is clear that as long as $\boldsymbol{\mu} \neq \boldsymbol{\mu}^*$ the invertibility of the Jacobian matrix is guaranteed under the discrete inf-sup assumptions required before.

Remark 3.1.2. In view of the Remark 2.1.1, we can observe that, for the discrete inf-sup stability (3.5), the condition on the surjectivity of the form $\mathrm{d}g[Z_{\mathcal{N}}](\cdot,\cdot;\boldsymbol{\mu})$ is no longer needed. Indeed, while the assumption (3.5) corresponds to the non-singularity of the jacobian matrix $\mathsf{J}_{\mathcal{N}}$, a discrete counterpart of the assumption (2.7) would require its surjectivity, or equivalently the injectivity of the transpose matrix $\mathsf{J}^T_{\mathcal{N}}$, which being square would be the same as requiring (3.5).

At this point we can approximate a solution $X(\boldsymbol{\mu})$ to (2.1) for a fixed value of the parameter $\boldsymbol{\mu}$. In the following, we aim at investigate its stability properties through the spectral analysis.

3.1.3 Eigenvalue problem

Dealing with the existence of many possible configurations for the system, a natural question is to understand which one inherits the stability of the unique solution, when it exists. As we remarked in Section 2.3, to perform the stability analysis of the model under investigation, we can rely on its spectral analysis, which consists in the study of the eigenvalues of the linearized PDE. Indeed, this connection allows us to understand the stability property of a specific configuration, by looking at the sign of the spectrum of the linearised operator. Of course, this kind of analysis is not always feasible at the continuous level, hence the investigation has to be performed at the discrete one. In particular, for a general nonlinear problem, one linearizes the equation (3.3) around the solution $\overline{X} = X(\overline{\boldsymbol{\mu}})$ to be examined, and then solve the eigenvalue problem given by

$$D_X G(\overline{X};\overline{\boldsymbol{\mu}})X_e = \sigma_{\overline{\boldsymbol{\mu}}} X_e, \tag{3.12}$$

where $(\sigma_{\overline{\boldsymbol{\mu}}}, X_e)$ is the eigenpair formed by the $\boldsymbol{\mu}$-dependent eigenvalues $\sigma_{\overline{\boldsymbol{\mu}}}$ and the eigenvectors X_e.

From the considerations in Section 2.3, we aim at observing that the eigenvalues $\sigma_{\boldsymbol{\mu}}$ vary continuously w.r.t. the parameter $\boldsymbol{\mu}$, and all have negative real parts when the corresponding unique solution $X(\boldsymbol{\mu})$ exists and is stable. On the other hand, when approaching to the bifurcation point $\boldsymbol{\mu}^*$, the model shows different stability properties and thus it exhibits a change of the sign for the rightmost eigenvalue. Looking at the complex plane and following the path traced by the eigenvalues, we will discover at the same time the crossing of the imaginary axis for the biggest eigenvalue and the presence of a new solution, that will inherits the stability of the former one. Hence, the old stable solution will be now characterized by an eigenvalue with positive real part, denoting its instability.

Studying the behaviour of the eigenvalues can provide useful informations, but it usually results in a complex and computationally expansive analysis. For this reason another possibility is to consider a different way to study the spectral properties. This involves the construction of a non-parametrized generalized eigenproblem, where the parameter is an unknown which plays the role of the eigenvalue. This allows to find the actual values of the bifurcation points, overcoming the need for a complete investigation of the eigenvalues detecting the crossing of the imaginary axis. We will present an application of this strategy in Section 4.2.1. With the application of this strategy we will also be able to detect more precisely the phenomenon of multiple bifurcation, that occurs in connection to eigenvalue with algebraic multiplicity strictly greater than one.

Therefore, once fixed $\boldsymbol{\mu} \in \mathcal{P}$, we are able to approximate a solution $X_N(\boldsymbol{\mu})$ and discover its stability properties, but our aim is to investigate the evolution of the solutions varying such a parameter. For this reason we will introduce in the next subsection a class of methods to follow the branches.

3.1.4 Continuation methods

To understand how the stability features of a solution change with respect to the variation of the parameter $\boldsymbol{\mu}$, it is fundamental to be able to follow the branch which it belongs. Moreover, the detection of the solution path is the key to recover the bifurcation diagram. In fact, as we said, it provides a global picture of the model behaviour, which can also be enriched with its stability analysis. In view of this need, we introduce the *continuation methods* [2, 52, 51, 86], that serve us to follow the branching behaviour of the model.

A continuation method allows to generate a sequence of solutions, corresponding to the selected values of the parameter, in order to construct branches of possible configurations. For this reason, they are usually called *branch tracing* or *path following* methods.

The main ingredients for a continuation method are:

- the selection of a good initial guess;
- a (possibly adaptive) rule to select the step size $\Delta\boldsymbol{\mu}$;

- a detection tool for the bifurcation point;
- a stability check for the approximated solution.

While the last two items are already embedded in the eigenvalue analysis we presented before, the first one is strongly related to the choice of the methodology. In particular, we are interested in *predictor-corrector* methods, which represent a wide class of techniques that helps the branch tracing by splitting the approximation in two steps, as follows. Let us start from a solution $\mathsf{X}_{\mathcal{N}}(\boldsymbol{\mu}_j)$ to (3.8) for the parameter value $\boldsymbol{\mu}_j$, we want to approximate the solution $\mathsf{X}_{\mathcal{N}}(\boldsymbol{\mu}_{j+1})$ for the corresponding next value $\boldsymbol{\mu}_{j+1}$. Instead of computing directly the pair $(\mathsf{X}_{\mathcal{N}}(\boldsymbol{\mu}_{j+1}), \boldsymbol{\mu}_{j+1})$, a predictor-corrector method acts as follows: in the first step the pair $(\widetilde{\mathsf{X}}_{\mathcal{N}}(\widetilde{\boldsymbol{\mu}}_j), \widetilde{\boldsymbol{\mu}}_j)$ is constructed from $(\mathsf{X}_{\mathcal{N}}(\boldsymbol{\mu}_j), \boldsymbol{\mu}_j)$, and then the former, without being a solution for $\boldsymbol{\mu}_{j+1}$, merely serves as initial guess in the second step.

As concerns the second item of the previous list, it is connected to the choice of the parameter values at which compute a solution to (3.8). Usually this represents a bottleneck for these methodologies, indeed there is plenty of literature that investigate complex techniques, which performances highly depend on the bifurcation phenomenon and the resulting bifurcation diagram [35, 138, 128].

Following a branch is not an easy task, in fact, during the investigation of the parameter space, frequent issues dealing with the continuation step are:

- $\Delta\boldsymbol{\mu}$ is too large, such that we can possibly skip the bifurcation point, without noticing the branching behaviour;
- $\Delta\boldsymbol{\mu}$ is too small, causing a waste of computational resources, mostly when approximating regions far from the critical point.

The identification of such locations is thus related to the kind of parameterization strategy chosen to trace the branch. Moreover, we highlight that when $\Delta\boldsymbol{\mu}$ is chosen to be too large, the previously computed solution will be far from the one that we want to approximate, possibly causing the non-convergence of the Newton method.

In the following, unless specified otherwise, since we always deal with pitchfork bifurcations we will restrict ourselves to the simplest continuation method available. This is called *simple continuation* method and can be seen as a basic predictor-corrector scheme where the pair $(\widetilde{\mathsf{X}}_{\mathcal{N}}(\widetilde{\boldsymbol{\mu}}_j), \widetilde{\boldsymbol{\mu}}_j)$ is actually given by $(\mathsf{X}_{\mathcal{N}}(\boldsymbol{\mu}_j), \boldsymbol{\mu}_j + \Delta\boldsymbol{\mu}_j)$. In fact, within this range of applications, we will never encounter fold bifurcations, which usually cause the non-convergence of the simple continuation method.

When dealing with more complex phenomena, one has to change the continuation method in order to help the Newton iterations to converge. For this reason, one can change the parametrization strategy, considering e.g. an arc-length like procedure. This exploits the Taylor expansion of the solution as $\boldsymbol{\mu}$-dependent function, in order to add an equation to (3.8) at each step, that automatically find the step-length $\Delta\boldsymbol{\mu}$. Such procedure is usually called *pseudo arc-length* technique [2, 142] and it aims at linearizing

the branch through the arc-lengh parametrization. An example of the application of such technique will be analysed in Chapter 8.

We conclude this section by illustrating the algorithm for the high fidelity reconstruction of a branch, which is the result of the combination of all the methodologies we have presented until now.

3.1.5 A branch-wise algorithm

We can finally present the algorithm we have developed in order to deal with the branch-wise reconstruction of the bifurcating behaviour of the models. The approach that we have implemented in this high fidelity context, requires the combination of all the different methodologies we presented in the previous subsections. Within the same setting, we will analyse also the stability of the discovered solutions through the spectral study depicted above.

The Algorithm 1 is the implementation result of the building blocks needed to linearize, discretize and continue each solution branch. More precisely, we combine respectively:

(i) Newton method, as the nonlinear solver,

(ii) Galerkin FE method, as the discretization phase,

(iii) simple continuation method, as the bifurcation path tracer,

(iv) eigenvalue problem, as the stability and bifurcation detection tool.

These are the main steps for the reconstruction of a bifurcation branch of a general non-linear parametric PDE. We remark that here we are pursuing a branch-wise approach, in the sense that we are fixing the branch $\mathcal{M} \subset \mathcal{X}$ to reconstruct, while later we will highlight the tricks that can be helpful to discover different branching solutions. Moreover, as already remarked in Section 2.2, in the following we will consider the parameter $\boldsymbol{\mu}$ as a one dimensional object. In fact, thanks to the codimension one assumption, we can always consider that the first component of $\boldsymbol{\mu} \in \mathbb{R}^P$ is the actual bifurcation parameter, while the other $P-1$ are fixed. Finally, in the algorithm we omitted for ease of notation the subscript $\mathcal{N}$ of the high fidelity solution vector $\mathsf{X}_{\mathcal{N}}$.

Let us now review the combination of the methodologies (i)-(iv). At the very beginning one implicitly chooses the branch to approximate, by choosing the initial guess to start the iteration of the Newton method. Indeed, the simplest way to "guide" the non-linear solver to a desired branch is through the choice of the initial guess. We remark that a more involved but less problem-specific strategy will be analysed in Section 8.2.2.

Hence, in order to fully recover the branch behaviour, when dealing with the simple continuation method (in which the parameter step-length is fixed and already prescribed) we consider a discrete version of the parameter space $\mathcal{P}_K = [\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K] \subset \mathcal{P}$ of cardinality K and loop over this ordered set. The loop serves to mimic the predictor-corrector method, assigning the solution obtained for a given parameter $\boldsymbol{\mu}_j$ as the

initial guess for the non-linear solver at next iteration for $\boldsymbol{\mu}_{j+1}$. This allows us to follow the bifurcating behaviour of the model. As we said, this simplest version of the predictor-corrector methodology works well with pitchfork like bifurcation, while a more involved methodology has to be implemented when dealing with turning points or secondary bifurcations, as we will see later in Section 4.2.3 where a variant of the simple continuation method will be presented.

For the actual linearization and discretization of the equation (3.8), we adopt the Newton-Kantorovich method (3.10) in combination with the Galerkin FE method. The former helps to linearize the system around the approximation of the solution at the k-th iteration, while the latter projects the problem into the finite dimensional space $\mathbb{X}_{\mathcal{N}}$.

Remark 3.1.3. Performing a projection onto a space which dimension is given by the number of degrees of freedom $\mathcal{N}$ of the problem can be a computational bottleneck. Indeed, complex phenomena require a fine discretization of the discrete domain, which cause a possibly huge linear system that has to be solved repeatedly until a convergence criterion is satisfied (here we chose a threshold tolerance ε for the space norm of the high fidelity residual). As a matter of fact, this means that we have to solve a linear system with $\mathcal{N}$ degrees of freedom, for each iteration of the Newton method and for each parameter in $\mathcal{P}_K$, assembling repeatedly the Jacobian matrix and thus causing an almost impracticable computational effort. This will be the starting point for the next section.

Finally, having computed a solution X_j of the problem for the parameter $\boldsymbol{\mu}_j$, we can investigate its stability properties solving the eigenvalue problem, which involves the Jacobian matrix $\mathsf{J}_{\mathcal{N}}$ and the inner product matrix $\mathsf{M}_{\mathcal{N}}$. This will allow us to understand the physical stability of the approximated solutions and to detect the bifurcation points connected to the qualitative changes of the model.

Algorithm 1 A pseudo-code for the reconstruction of a branch

$\mathsf{X}_0 = \mathsf{X}_{guess}$ ▷ Initial guess
for $\boldsymbol{\mu}_j \in \mathcal{P}_K$ **do** ▷ Continuation loop
  $\mathsf{X}_j^{(0)} = \mathsf{X}_{j-1}$ ▷ Continuation guess
  while $\|\mathsf{G}_{\mathcal{N}}(\mathsf{X}_j^{(k)};\boldsymbol{\mu}_j)\|_{\mathbb{X}_{\mathcal{N}}} > \varepsilon$ **do** ▷ Newton method
    $\mathsf{J}_{\mathcal{N}}(\mathsf{X}_j^{(k)};\boldsymbol{\mu}_j)\delta\mathsf{X} = \mathsf{G}_{\mathcal{N}}(\mathsf{X}_j^{(k)};\boldsymbol{\mu}_j)$ ▷ Galerkin FE method
    $\mathsf{X}_j^{(k+1)} = \mathsf{X}_j^{(k)} - \delta\mathsf{X}$
  end while
  $\mathsf{J}_{\mathcal{N}}(\mathsf{X}_j;\boldsymbol{\mu}_j)\mathsf{X}_e = \sigma_{\boldsymbol{\mu}_j}\mathsf{M}_{\mathcal{N}}\mathsf{X}_e$ ▷ Eigenproblem for stability
end for

The algorithm we have just presented can easily reconstruct a branch while depicting its stability properties. Depending on the application, the investigation of the

bifurcation phenomenon with its peculiarities is evident already during the approximation of one single branch, discovering the qualitatively changes that the solutions exhibit when reaching the bifurcation point. As we showed in the simpler case of the ODE, for pitchfork like bifurcation we pass from a stable trivial configuration to another stable but non-trivial one. This is an example of what usually happens when the algorithm is used for the approximation of a branch, therefore with a little abuse of terminology, as before we will call branch the unique extension of the bifurcating behaviour to the pre-bifurcation regime.

Moreover, it is evident that the choice of the initial guess is fundamental, but it is not sufficient to recover the full bifurcation diagram. We will explain case by case how to properly set the continuation method in conjunction with the initial guess, since we analysed different methodologies depending on the availability of the information that we had or we recovered from the discrete approximation of the systems.

In particular, such tricks for the choice of the initial guess include

- the discretized version of analytic expressions, which have the main properties of the sought solution (see Sections 4.2.2 and 5.2.1);
- the eigenvectors of the global eigenvalue problem, to obtain the direction of the bifurcation branch in a neighbourhood of the bifurcation points (see Section 4.2.4);
- a deflation method, which requires only one initial guess, and then reconstructs the full diagram preventing the convergence to already discovered solution, helping the solver to find new branches (see Section 8.2.2).

In the next section, we will start from the consideration in Remark 3.1.3 to review the Reduced Order Models, in connection to bifurcation problems, in order to present an efficient version of the Algorithm 1.

3.2 Reduced Order Models

Dealing with the approximation of parametrized problems can be a critical task from the computational viewpoint. Such difficulties even increase when nonlinearities are taken into account. Therefore, in the following sections, we will review the basic notions of the so called *Reduced Order Models* (ROMs) [115, 74, 130, 17, 15], by which we are able reduce the computational cost of our simulations. With this aim in the past years many authors, to mention few works [80, 127, 75, 65, 32, 154, 38, 131, 16], developed and applied this collection of methodologies used to replace the original high dimensional problem, called *high fidelity approximation*, with a reduced one that is easy to manage. After a brief introduction to the reduced techniques we adopted, together with the additional methodologies, to tackle the presence of nonlinear terms, we will present the efficient counterpart of Algorithm 1 for the reconstruction of a bifurcation diagram. We remark that despite the great interest behind these complex

phenomena, only few works treated them by means of reduction strategies [113, 149, 71, 72, 124, 112].

3.2.1 The Reduced Basis method

As we have seen during the previous section, the preliminary step for the discretization of (2.1) is the projection of the weak formulation (2.3) in a finite dimensional setting, which results in the Galerkin problem (3.3). Despite the increasing computational resources available, finding the solutions to this problem is very challenging because of the potential high number of degrees of freedom $\mathcal{N}$. Moreover, we recall that in addition to the multi-query context and the nonlinearity of the models considered, the existence of multiple solutions complicates the setting, requiring a deep investigation of the system behaviour with respect to the parameter space.

For this reason, here we focus on a specific methodology in the class of the ROMs, called *Reduced Basis* (RB) method. It consists in a projection like technique, and therefore it shares many features with the Galerkin FE method. Roughly speaking, this method consists in a projection of the high fidelity problem on a subspace of smaller dimension, constructed with some properly chosen basis functions.

In practice, the key feature of the RB method is the adoption of the offline-online paradigm. Indeed, in order to retrieve a solution to (3.3), for a given parameter $\boldsymbol{\mu} \in \mathcal{P}$, the computation is divided in two steps:

- An *offline phase*: approximated solutions to (3.3) are computed with an high fidelity method (Finite Element, Spectral Element, Finite Volume, Finite Difference), corresponding to selected representative parameters values/system configurations, and stored together with other information about the parametrized problem. This is the computationally expensive step, and it can benefit from the possibly available high performance computing (HPC) facilities.

- An *online phase*: the pre-processed information obtained during the offline phase is assembled and used to compute the solution for each new instance of the parameter in a short amount of time (ideally in real-time), even on a relatively low power device such as a laptop or a smartphone.

This split in the computational procedures is built in such a way that new parameter dependent quantities can be easily and quickly computed online, while representative basis functions for selected parameter values and more demanding quantities are precomputed offline.

Thus, the main goal of the offline phase is the construction of a low dimensional basis for a discrete manifold $\mathbb{X}_N \subset \mathbb{X}_{\mathcal{N}}$, called *reduced manifold*, which we assume to well approximate the high fidelity manifold $X_{\mathcal{N}}$.

This entails solving N_{train} times the Galerkin high fidelity problem associated to N_{train} values of $\boldsymbol{\mu}$ in $\mathcal{P}$. The obtained solutions $\{X(\boldsymbol{\mu}_i)\}_{i=1}^{N_{train}}$ are usually called *snapshots*. Having recovered the information about the $\boldsymbol{\mu}$-dependence of the solutions, we need to process these snapshots in order to obtain a basis that span the reduced manifold $\mathbb{X}_N$.

Among the others, the *Proper orthogonal decomposition* (POD) and the *Greedy* techniques [74, 115, 130] are the most known and used ones. Let us assume for a while that such a basis is already built, and see how the reduced problem is constructed through the projection on the subspace spanned by it.

As concerns the online phase, it is the efficient and reliable part where the solutions are computed through the projection on $\mathbb{X}_N$. The complexity reduction, which enables the efficiency, is based on two main assumptions:

(i) the *affine decomposition* holds, i.e. one can rewrite the weak formulation as a linear combination of $\boldsymbol{\mu}$-independent forms and $\boldsymbol{\mu}$-dependent coefficients. Hence, the contribution of the parameter is entirely encoded by the coefficients, which usually allow to rapidly assemble the system, relying on the precomputed parameter independent quantities.

(ii) we can accurately approximate the discrete manifold $\mathbb{X}_{\mathcal{N}}$ with a much lower dimension space $\mathbb{X}_N$, i.e. we need only a small number $N \ll \mathcal{N}$ of basis functions.

Therefore, the reduced computational cost mainly comes from avoiding to project on the large FE manifold, while relying on the small RB one. It is clear that, in the our context, the assumption (i) is difficult to fulfill, because of the $\boldsymbol{\mu}$-dependence of the solution around which we linearise the nonlinear weak formulation (3.8). The assumption (ii) is instead usually linked with the concept of Kolmogorov n-width [103, 45], which expresses the capability of a reduced manifold to approximate the high fidelity one.

Let us now describe the details of the online phase. Since it is a projection like method, we can follow the same steps described in Section 3.1. Hence, we consider the discrete weak formulation (3.3) and projecting it into the reduced space $\mathbb{X}_N$ we obtain the following problem: given $\boldsymbol{\mu} \in \mathcal{P}$, we seek $X_N(\boldsymbol{\mu}) \in \mathbb{X}_N$ such that

$$g(X_N(\boldsymbol{\mu}), Y_N; \boldsymbol{\mu}) = 0, \quad \forall\, Y_N \in \mathbb{X}_N. \tag{3.13}$$

Also here we remark that the obtained reduced weak formulation can not be directly solved, since it involves nonlinear terms. Therefore, following Section 3.1.2 we apply the Newton-Kantorovich method that reads as: chosen an initial guess $X_N^0(\boldsymbol{\mu}) \in \mathbb{X}_N$, for every $k = 0, 1, \dots$, find the variation $\delta X_N \in \mathbb{X}_N$ such that:

$$\mathrm{d}g[X_N^k(\boldsymbol{\mu})](\delta X_N, Y_N; \boldsymbol{\mu}) = g(X_N^k(\boldsymbol{\mu}), Y_N; \boldsymbol{\mu}), \quad \forall\, Y_N \in \mathbb{X}_N, \tag{3.14}$$

and then we update the solution as $X_N^{k+1}(\boldsymbol{\mu}) = X_N^k(\boldsymbol{\mu}) - \delta X_N$, until an appropriate stopping criterion is verified.

It is clear now that the system we constructed is based on a low dimensional manifold, so we can finally show the counterpart of the high fidelity quantities in (3.10).

Let $\{\Sigma^m\}_{m=1}^N$ be a given orthonormal basis for $\mathbb{X}_N$, with respect to the inner product defined on the space $\mathbb{X}_{\mathcal{N}}$. Then, the reduced manifold is defined as

$$\mathbb{X}_N = \mathrm{span}\{\Sigma^1, \dots, \Sigma^N\},$$

therefore we can express every element $X_N(\boldsymbol{\mu}) \in \mathbb{X}_N$ in the following way

$$X_N(\boldsymbol{\mu}) = \sum_{m=1}^{N} X_N^{(m)}(\boldsymbol{\mu})\Sigma^m. \tag{3.15}$$

Unless specified otherwise, as for the high fidelity setting, we remark that when X_N contains more variables/components, the dimension of the reduced basis space N has to be considered as the sum of the basis used for each variable/component. From the algebraic standpoint, we can denote with $\mathsf{X}_N(\boldsymbol{\mu}) = \{X_N^{(m)}(\boldsymbol{\mu})\}_{m=1}^N \in \mathbb{R}^N$ the *reduced solution vector*. Plugging (3.15) into (3.13) and choosing $Y_N = \Sigma^n \in \mathbb{X}_N$, for $1 \leq n \leq N$, we obtain the following algebraic system

$$g\left(\sum_{m=1}^{N} X_N^{(m)}(\boldsymbol{\mu})\Sigma^m, \Sigma^n; \boldsymbol{\mu}\right) = 0\,, \quad n = 1, \dots, N\,. \tag{3.16}$$

We will denote by G_N the reduced counterpart of $\mathsf{G}_{\mathcal{N}}$ defined by

$$(\mathsf{G}_N(\mathsf{X}_N(\boldsymbol{\mu}); \boldsymbol{\mu}))_n = g\left(\sum_{m=1}^{N} X_N^{(m)}(\boldsymbol{\mu})\Sigma^m, \Sigma^n; \boldsymbol{\mu}\right), \tag{3.17}$$

which we will refer to as the *reduced residual vector*. Moreover, we will denote with $\mathsf{V} \in \mathbb{R}^{\mathcal{N} \times N}$ the transformation matrix whose elements

$$(\mathsf{V})_{jm} = \Sigma^m_{(j)} \tag{3.18}$$

are the nodal evaluation of the m-th basis function at the j-th node. With this notation, we can rewrite problem (3.16) as

$$\mathsf{V}^T \mathsf{G}_N(\mathsf{V}\mathsf{X}_N(\boldsymbol{\mu}); \boldsymbol{\mu}) = 0. \tag{3.19}$$

Finally, the algebraic form of the Newton method, combined with the RB technique, provides the following formulation: at every iteration k we seek $\delta\mathsf{X}_N \in \mathbb{R}^N$ such that

$$\mathsf{J}_N(\mathsf{X}_N^k(\boldsymbol{\mu}); \boldsymbol{\mu})\delta\mathsf{X}_N = \mathsf{G}_N(\mathsf{X}_N^k(\boldsymbol{\mu}); \boldsymbol{\mu}), \tag{3.20}$$

where J_N is the $\mathbb{R}^{N \times N}$ *reduced Jacobian matrix*

$$\mathsf{J}_N(\mathsf{X}_N^k(\boldsymbol{\mu}); \boldsymbol{\mu}) = \mathsf{V}^T \mathsf{J}(\mathsf{V}\mathsf{X}_N^k(\boldsymbol{\mu}); \boldsymbol{\mu})\mathsf{V}. \tag{3.21}$$

Remark 3.2.1. For a general nonlinear problem equation (3.20) still involves the degrees of freedom of the high fidelity problem $\mathcal{N}$. Because of this, the repeated assembly of the reduced Jacobian compromises also at the low fidelity level the efficiency of the reduced order method during the online phase. As we will see later, this issue can be overcome by adopting a class of affine-recovery techniques, which allows consistent speed-ups by interpolating the nonlinear part of the variational form.

Having described the online simulation, we now go back on the building process of the reduced manifold $\mathbb{X}_N$. Hence, in the next subsections, we will analyse the main techniques to construct a basis and the strategies to recover the affinity assumption when, due to the nonlinearity, it does not hold.

3.2.2 The Proper Orthogonal Decomposition

The keypoint for the application of the Reduced Basis method is the construction of the basis. Here we focus on the Proper Orthogonal Decomposition (POD) [10, 82], which is a compression strategy, closely related to the Singular Value Decomposition (SVD) and the Principal Component Analysis (PCA), that serves to reduce the dimensionality of a given dataset. Indeed, starting from a sufficiently rich information about the system, it allows to extract the main features by a lower dimensional representation given by the first computed modes, which ideally retain most of its energy. In our applications, we use the POD to generate the Reduced Basis space $\mathbb{X}_N$. Moreover, this is proved to be optimal in the $l^2(\mathbb{R}^N)$ sense, in fact the choice of the basis is given by the reduced space $\mathbb{X}_N$ which minimizes among all the the N-dimensional subspaces $\mathbb{Z}_N \subset \mathbb{X}_\mathcal{N}$, the following quantity

$$\sqrt{\frac{1}{N_{train}} \sum_{\boldsymbol{\mu} \in \mathcal{P}_{train}} \inf_{Z_N \in \mathbb{Z}_N} \| X_\mathcal{N}(\boldsymbol{\mu}) - Z_N \|_\mathbb{X}^2}, \tag{3.22}$$

where we denoted with $\mathcal{P}_{train} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{train}\}$ a finite sampling of $\mathcal{P}$ of dimension N_{train}. Therefore, in order to construct $\mathbb{X}_N$ one considers $\mathcal{P}_{train}$ and build from the correspondent solutions $\{X_\mathcal{N}(\boldsymbol{\mu}_j)\}_{j=1}^{\mathcal{N}_{train}}$ the symmetric and linear *correlation* operator $C : \mathbb{X}_\mathcal{N} \to \mathbb{X}_\mathcal{N}$ defined as

$$C(V_\mathcal{N}) = \frac{1}{N_{train}} \sum_{j=1}^{N_{train}} \left(V_\mathcal{N}, X_\mathcal{N}(\boldsymbol{\mu}_j) \right)_\mathbb{X} X_\mathcal{N}(\boldsymbol{\mu}_j), \qquad V_\mathcal{N} \in \mathbb{X}_\mathcal{N} . \tag{3.23}$$

Then, one computes the eigenvalue-eigenfunction pairs $(\sigma_i, \Psi_i) \in \mathbb{R} \times \mathbb{X}_\mathcal{N}$, such that $\|\Psi_i\|_\mathbb{X} = 1$ for any $i = 1, \dots, N_{train}$, by means of the following eigenproblem

$$\left(C\left(\Psi_i\right), X_\mathcal{N}(\boldsymbol{\mu}_j) \right)_\mathbb{X} = \sigma_i \left(\Psi_i, X_\mathcal{N}(\boldsymbol{\mu}_j) \right)_\mathbb{X}, \qquad 1 \le j \le N_{train} . \tag{3.24}$$

Sorting the eigenvalues and the associated eigenfunctions in descending order, we can use the first N eigenfunctions to built the reduced space $\mathbb{X}_N$. Moreover, it is possible to prove that the error obtained approximating the solutions of $\mathbb{X}_\mathcal{N}$ with the ones in $\mathbb{X}_N$ is given by

$$\sqrt{\frac{1}{N_{train}} \sum_{j=1}^{N_{train}} \| X_\mathcal{N}(\boldsymbol{\mu}_j) - P_N \left(X_\mathcal{N}(\boldsymbol{\mu}_j) \right) \|_\mathbb{X}^2} = \sqrt{\sum_{j=N+1}^{N_{train}} \sigma_j}, \tag{3.25}$$

where $P_N : \mathbb{X}_\mathcal{N} \to \mathbb{X}_N$ is the projection operator over $\mathbb{X}_N$ defined as

$$P_N(V_\mathcal{N}) = \sum_{i=1}^{N} (V_\mathcal{N}, \Psi_i)_\mathbb{X} \Psi_i . \tag{3.26}$$

The result in (3.25) is often used as a criterion to select the dimension of the POD basis. Indeed, one chooses the minimal integer N such that the retained energy from the last

$(N_{train} - N)$ snapshots, which is the right-hand side term in (3.25) normalized with respect to the total energy, is less than a fixed tolerance ε_{POD}. This means that we are finding the basis which minimizes over all possible N dimensional orthonormal bases $\mathbb{Z}_N$ the error between the snapshots and their projection through $\mathbb{Z}_N$. Therefore, from the discrete point of view, in order to obtain a reduced space with such desired properties, one has to solve the eigenproblem (3.24) for the correlation matrix C. To perform efficiently this task, one usually consider the SVD of the *snapshots matrix*

$$\mathsf{S} = [X_{\mathcal{N}}(\boldsymbol{\mu}_1), \dots, X_{\mathcal{N}}(\boldsymbol{\mu}_{N_{train}})] \in \mathbb{R}^{\mathcal{N} \times N_{train}}, \tag{3.27}$$

which columns are the degrees of freedom of the N_{train} snapshots. This way, the POD basis of dimension N is given by the first N left singular vectors.

It is clear that one of the disadvantages of this technique is the huge number of high fidelity solutions that have to be computed to obtain a fairly accurate representation of the high fidelity manifold. Indeed if we miss some information at the first discretization level, also the reduced basis, after the compression step given by the POD, usually will not contain them.

3.2.3 The Greedy algorithm

As an alternative to POD, one can implement a Greedy algorithm [74, 130, 25], an iterative technique which is constructed in order to increase the precision of the basis at each iteration. Instead of building a huge dataset, corresponding to a given parameter sample $\mathcal{P}_{train}$, the Greedy algorithm requires only one high fidelity solution per iteration, and a total number of N solution for a basis of dimension N. Hence, the key ingredient is the selection of the parameters for which compute the solutions. Such choice is usually guided by an error estimator $\Delta_N(\boldsymbol{\mu})$ such that

$$\|X_{\mathcal{N}}(\boldsymbol{\mu}) - X_N(\boldsymbol{\mu})\|_{\mathbb{X}_{\mathcal{N}}} \leq \Delta_N(\boldsymbol{\mu}) \quad \forall\, \boldsymbol{\mu} \in \mathcal{P}, \tag{3.28}$$

which is used at the generic n-th iteration, to find the worst approximated solution by the reduced basis space of dimension n in the whole parameter space $\mathcal{P}$ as

$$\boldsymbol{\mu}_{n+1} = \arg\max_{\boldsymbol{\mu} \in \mathcal{P}} \Delta_n(\boldsymbol{\mu}). \tag{3.29}$$

Then, the corresponding snapshot is orthonormalized with the Gram-Schmidt algorithm [80, 130], in order to preserve the orthonormality of the basis, and successively added to it. The iterative procedure ends when a prescribed tolerance ε_{Gr} is reached by the error estimator.

The bottleneck of this methodology is clearly the individuation of the parameter which produces the worst approximation, since it would involve a complete description of the high fidelity manifold to build the error estimator. For this reason one usually need the estimator to be efficiently treated. To obtain this property one can again use the affine decomposition assumption, which allows for an inexpensive computation of the estimator [115, 80].

We remark that due to the lack of the affine decomposition assumption and the general unavailability of an a priori error estimator, the Greedy method is rarely applied to nonlinear problems. For this reason, we decided to build the reduced basis for our applications by means of the POD strategy, while the Greedy algorithm will be only utilized as a part of the empirical interpolation strategies (as we will see in the next subsection). Moreover, even if the POD strategy increases the computational cost in the offline part, it gives us a reliable representation of the reduced manifold (that we will exploit with new approaches in Section 10), keeping track of the energy information that we are discarding.

Linked with the energy information carried by the system, we also highlight that the POD is usually considered in connection with time dependent problem, in which the evolution behaviour is strictly connected to the energy of the configuration observed. As a consequence, in some of the applications we will present, we can consider that the bifurcation parameter $\boldsymbol{\mu}$ represents a quasi-static evolution in time.

On the contrary, to build an efficient a posteriori error estimator $\Delta_N(\boldsymbol{\mu})$ in the nonlinear (bifurcating) context, one should rely on the Brezzi-Rappaz-Raviart (BRR) theory [24, 27, 136, 101, 50]. Despite the fundamental importance of the BRR theory, such approach is usually difficult to perform and mainly problem specific.

Finally, we remark that a promising approach could be the application of the *POD-Greedy* method [70, 111, 69], where one combines the two sampling procedures presented, in order to decrease the number of snapshots to be computed during the offline phase.

As we said, the reduced quantities we have discussed are $\boldsymbol{\mu}$-dependent and some affine-recovery techniques, also called hyper-reduction strategies have to be taken in consideration to recover the efficiency of the whole reduced approach, this will be the topic of the next subsection.

3.2.4 Hyper-reduction strategies

As we said previously, one of the main assumption to obtain an efficient reduction method is that the forms in (3.3) have an affine parametric dependence. This is crucial since it serves to obtain a $\mathcal{N}$-independent online phase, in which we can efficiently compute the solution of the parametrized PDE at each new instance of the parameter $\boldsymbol{\mu}$. Despite this, when dealing with nonlinear problems, such assumption is still not sufficient, due to the fact that the residual (3.8) depends on the parametrized solution itself.

Indeed, even assuming that we can write

$$\mathsf{G}_{\mathcal{N}}(\mathsf{X}_{\mathcal{N}};\boldsymbol{\mu}) = \sum_{q=1}^{Q_G} \theta_G^q(\boldsymbol{\mu})\mathsf{G}_{\mathcal{N}}^q(\mathsf{X}_{\mathcal{N}}), \tag{3.30}$$

which translates at the reduced level as

$$\mathsf{G}_N(\mathsf{X}_N;\boldsymbol{\mu}) = \sum_{q=1}^{Q_G} \theta_G^q(\boldsymbol{\mu})\mathsf{V}^T\mathsf{G}_{\mathcal{N}}^q(\mathsf{V}\mathsf{X}_N), \tag{3.31}$$

it is clear that we can not recover an efficient offline-online decomposition, since the assembly and the projection still involves the $\mathcal{N}$ degrees of freedom, compromising the efficiency. An easy way to overcome this issue is an efficient assembly of the nonlinear residual in a tensor form, but this is only practicable when the nonlinearity taken into account is polynomial.

While this is true e.g. for the Navier-Stokes equation (as we will see in Chapter 8), unfortunately this is not the case for many other applications. Furthermore, even when the tensor formulation is applicable, it requires a huge data structure that can potentially slow further the offline phase.

Among the others, reliable alternatives to achieve an efficient online phase come from an *hyper-reduction* approach. These techniques include the Empirical Interpolation Method (EIM) [12, 97] and its variant the Discrete Empirical Interpolation Method (DEIM) [36], which aim at obtaining the online computations independent from the number of degrees of freedom of the chosen full order discretization method.

The hyper-reduction techniques serve at approximating a general parametrized function $g : \Omega \times \mathcal{P} \to \mathbb{R}$ by the sum of affine terms

$$g(\mathbf{x}, \boldsymbol{\mu}) \approx \mathcal{I}[g_{\boldsymbol{\mu}}](\mathbf{x}) = \sum_{q=1}^{Q} c^q(\boldsymbol{\mu}) h^q(\mathbf{x}) \tag{3.32}$$

where $\mathcal{I}$ is the interpolation operator, the set of basis functions $\{h^q\}_{q=1}^{Q}$ are obtained by means of a linear combination of Q snapshots $\{g_{\boldsymbol{\mu}_q}\}_{q=1}^{Q}$ and the sample points are chosen through a Greedy approach. The coefficients $c_q(\boldsymbol{\mu})$ of such expansion in (3.32) are found by solving

$$\mathcal{I}[g_{\boldsymbol{\mu}}](\mathbf{x}_j) = g_{\boldsymbol{\mu}}(\mathbf{x}_j)$$

in some particular points $\{\mathbf{x}_j\}_{j=1}^{Q}$ of the domain Ω, usually called *magic points* [97]. Therefore, the Empirical Interpolation strategies provide a discrete version of $\mathcal{I}[g_{\boldsymbol{\mu}}](\mathbf{x})$, that is the interpolation matrix $\mathsf{H} \in \mathbb{R}^{Q \times Q}$ such that

$$g_Q(\boldsymbol{\mu}) = \mathsf{H} c(\boldsymbol{\mu}), \qquad \mathsf{H} = \{h^q(\mathbf{x}_j)\}_{(j,q)}.$$

Despite the similarity of the two hyper-reduction techniques, the main difference between EIM and DEIM is in the construction of H. In fact, while EIM embeds the construction of the basis inside a Greedy procedure, DEIM exploits a POD on a set of snapshots. Moreover, the DEIM strategy starts with discretizing the nonlinearity, while EIM constructs the set of the magic points and the basis functions before the discretization step.

Finally, the hyper-reduction techniques described above provide the following affine approximation of the reduced residual vector

$$\mathsf{G}_N(\mathsf{X}_N; \boldsymbol{\mu}) \approx \sum_{q=1}^{Q_G} c_\theta^q(\mathsf{X}_N; \boldsymbol{\mu}) \mathsf{V}^T h^q$$

where $\{h^q\}_{q=1}^{Q_G}$ represent a suitable basis and c_θ^q are the interpolation coefficients, which include the contribute of the $\boldsymbol{\mu}$-dependent terms. Finally, the Jacobian matrix $\mathsf{J}_N(\mathsf{X}_N;\boldsymbol{\mu})$ can be assembled in a similar fashion.

We now have all the reduced ingredients to efficiently reconstruct a solution branch online, thus in the next subsection we will present the reduced counterpart of the Algorithm 1.

3.2.5 A reduced branch-wise algorithm

We want to end this chapter, concerning the numerical approximation of bifurcating phenomena, with the reduced algorithm we developed for the efficient reconstruction of the bifurcation diagram. Since this presents many similarities with the high fidelity Algorithm 1, we will describe briefly the main differences between the two.

Before starting with the description of the online approximation, we clarify some details about the offline phase. First of all, an hyper-reduction technique as EIM or DEIM can be set up for the approximation of the parametrized nonlinear form as linear combination of the interpolation basis functions. Then, to construct the reduced manifold we implemented the POD approach described in Section 3.2.2.

In principle, one can choose a random training set $\mathcal{P}_{train}$ for the computation of the snapshots. Despite the simplicity of this approach, this could lead to wasted computational time. Indeed, a smarter way to help the convergence and to prevent unnecessary Newton iterations, is to select the snapshots location through the continuation method, which in its simplest variant results in an equispaced sampling of $\mathcal{P}$. Of course, here one has to take care of all the possible issues depicted in subsection 3.1.4.

Until now we restricted ourselves to the case in which we deal with a one dimensional bifurcating parameter, while the other $P-1$ are considered fixed. Of course, this does not create issues for the branch-wise approach, but if one wants to build a unique reduced manifold which contains all the bifurcating modes of the model, then the snapshots matrix should contain such information. Thus, as before, the continuation loop over the bifurcation parameter space in Algorithm 1 has to be performed for each physical/geometrical configuration in the training set chosen for the corresponding last $P-1$ components of $\boldsymbol{\mu}$.

More remarkably, the same considerations hold also for the case $P=1$, where we select the snapshots, again through the continuation method, belonging to different branches. In this view, we remark that the aforementioned POD-Greedy algorithm could improve the accuracy results when $P \gg 1$.

Remark 3.2.2. A consequence of the previous observation is that a huge number of snapshots have to be computed (as usual when dealing with POD, but here it depends also on the number of branches to be approximated). Moreover, the POD space in such cases will contain a much larger number of basis functions, since it has to encode the properties of all the recovered branches. Despite these drawbacks, the main advantages are: (i) the possibility to efficiently reconstruct online the whole bifurcation diagram; (ii) the construction of a single ROM in contrast to a reduced basis for each

branch. Thus, within the global approach, a single POD compression can be used to recover all the branches.

Once finished the offline phase we have computed a (global or branch specific) basis for the reduced manifold, and we can adapt all the methodologies to the reduced setting. Thus, we combine the projection step (this time in the reduced basis space $\mathbb{X}_N$), the Newton method (3.14) and the simple continuation method for the reduced vector solution X_N. The reduced initial guess can be obtained as the projection of the high fidelity one, while the continuation method proceeds as before, assigning the new guess as the reduced vector of dimension N corresponding to the solution for the previous value of the parameter. For the for the sake of clarity, as before, in the Algorithm 2 we will omit the subscript N from the reduced solution X_N.

Algorithm 2 A pseudo-code for the efficient reconstruction of a branch

Offline phase

EIM/DEIM hyper-reduction strategies for the computation of H

POD applied to the snapshot matrix S to built the basis encoded in V

Online phase

$\mathsf{X}_0 = \mathsf{X}_{guess}$ ▷ Initial guess

for $\boldsymbol{\mu}_j \in \mathcal{P}_K$ **do** ▷ Continuation loop

  $\mathsf{X}_j^{(0)} = \mathsf{X}_{j-1}$ ▷ Continuation guess

  while $||\mathsf{G}_N(\mathsf{X}_j^{(k)}; \boldsymbol{\mu}_j)||_{\mathbb{X}_N} > \varepsilon$ **do** ▷ Newton method

    $\mathsf{J}_N(\mathsf{X}_j^{(k)}; \boldsymbol{\mu}_j)\delta\mathsf{X} = \mathsf{G}_N(\mathsf{X}_j^{(k)}; \boldsymbol{\mu}_j)$ ▷ Galerkin RB method

    $\mathsf{X}_j^{(k+1)} = \mathsf{X}_j^{(k)} - \delta\mathsf{X}$

  end while

end for

Since the reduced cost of each evaluation $\boldsymbol{\mu}_j \to \mathsf{X}_N(\boldsymbol{\mu}_j)$, the set $\mathcal{P}_K$ can be chosen as a much refined version of the one in the high fidelity context. Therefore, the reduced approach allows to obtain a better investigation of the region near the critical points, capturing the dynamics of the bifurcation with smaller steps.

Remark 3.2.3. As concerns the theoretical rationale behind the reduced basis method [130, 74], it was proved that one of the main ingredients to have good approximation properties is the regularity of the solution as a function of the parameter, rather then its regularity in space. In fact, this can be an issue for bifurcation problems, where e.g. for pitchfork phenomena the critical points represent a discontinuity in the *parametric sensitivity* $\partial X(\boldsymbol{\mu})/\partial\boldsymbol{\mu}$.

Hence, we expect that the error analysis in the parameter space $\mathcal{P}$ will show higher peaks at bifurcation points. We remark that in a black-box approach, one can utilize this statistic to forecast the location of a bifurcating phenomena.

It is clear that, in the presented pseudo-code, one can embed the efficient computation of an output s that can represent the scalar measure used for the description of the bifurcation diagram (an example of this will be showed in Section 4.2.4).

In this chapter we presented the approaches we have developed to investigate bifurcating PDEs, from the high fidelity setting to the reduced basis context. In particular, since we have concluded the exposition of the general framework of our study, in the next part we will start the examination of complex physical phenomena through the methodology analysed until now. More precisely, we will begin with the study of problems in Continuum Mechanics as the buckling of plates and beams and then we will move to Fluid Dynamics applications as channel flows, passing also through Quantum Mechanics investigations.

Part II

Approximation of parametrized bifurcating PDEs models

Chapter 4

Von Kármán equations for structural buckling of plates

Starting from the very well known theory of continuum mechanics, Von Kármán proposed a mathematical model in order to describe all the possible configurations that a plate under compression can assume [155]. The Von Kármán model is often used to describe buckling phenomena. A *buckling phenomenon* is indeed an example of bifurcating behaviour, where the system changes suddenly its configuration. As an example, we can think to have a thin rectangular plate at rest between our hands, and compress it until we reach a critical point, at which the plate takes a deformed configuration, or it buckles. This will be exactly what we will simulate. Hence, in the following, we will describe the mathematical formulation of this model, highlighting the difficulties that we encounter approximating its solutions. This part is mainly based on the work done in [120].

4.1 Von Kármán model

Let us consider an elastic and rectangular plate $\Omega = [0, L] \times [0, 1]$ in its undeformed state, subject to a μ-parametrized external load acting on its edges, depicted in Figure 4.1.

Then, the *displacement* from its flat state and the *Airy stress potential*, respectively u and φ, satisfy the Von Kármán equations

$$\begin{cases} \Delta^2 u = [\mu h + \varphi, u] + f & \text{in } \Omega, \\ \Delta^2 \varphi = -[u, u] & \text{in } \Omega, \end{cases} \tag{4.1}$$

where h and f are some given functions, that we can set to specify the external forces acting on our plate, while Δ^2 is the biharmonic operator in Cartesian coordinates and

$$[u, \varphi] := \frac{\partial^2 u}{\partial x^2}\frac{\partial^2 \varphi}{\partial y^2} - 2\frac{\partial^2 u}{\partial x \partial y}\frac{\partial^2 \varphi}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2}\frac{\partial^2 \varphi}{\partial x^2},$$

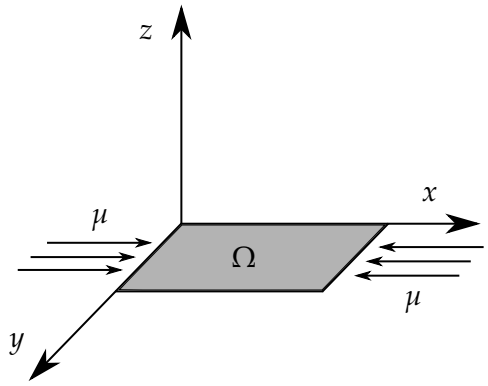


Figure 4.1: An elastic and rectangular plate compressed along the edges parallel to the y-direction.

is the *bracket of Monge-Ampére*. Thus, we aim at finding the pair given by the displacement and the *Airy stress potential*, which is physically linked to the second derivatives of the Piola-Kirchhoff stress tensor, that solves the system (4.1). The Von Kármán model is of fourth order, due to the presence of the biharmonic operator, nonlinear due to the product of second derivatives in the bracket, and parametric due to the buckling coefficient μ varying in a proper range of real numbers. Moreover, we are presenting for the sake of simplicity a non-dimensional model where all the physical quantities, except for the compression parameter μ, are set to unity.

In order to have a complete description of the physics involved, we must supplement to the system of partial differential equations some opportune boundary conditions for both the unknowns. Although they can be imposed in many different ways [42], in order to be coherent with the experiments of interest, we present only the most used ones. The first option is to impose *totally clamped* boundary conditions of the form

$$\begin{cases} u = \partial_n u = 0 & \text{in } \partial\Omega, \\ \varphi = \partial_n \varphi = 0 & \text{in } \partial\Omega, \end{cases} \tag{4.2}$$

whose meaning is that the plate is completely blocked on its sides. Note that here we denoted with ∂_n the directional derivative along the normal n to the boundary $\partial\Omega$. A second possible option is represented by the so called *simply supported* boundary conditions

$$\begin{cases} u = \Delta u = 0 & \text{in } \partial\Omega, \\ \varphi = \Delta \varphi = 0 & \text{in } \partial\Omega, \end{cases} \tag{4.3}$$

which are physically complex to reproduce, but also the most used ones for the simulations because of their versatility in the weak formulation. So, from now on, we will consider the system (4.1) with simply supported boundary conditions (4.3). We remark that, despite the simple boundary conditions chosen, the goal of this work is understanding the bifurcation behaviour for the Von Kármán plate equation, regard-

less the numerical constraints that a conforming method for more involved boundary conditions could impose.

A key observation, coming from such a choice of the BCs, is that we can split the system of two fourth order nonlinear elliptic equations, into a system of four second order nonlinear elliptic equations. In order to carry out this trick, introduced in [44], we add two new unknowns $U = \Delta u$ and $\Phi = \Delta\varphi$, so that we can rewrite the system (4.1) as

$$\begin{cases} \Delta U = [\mu h + \varphi, u] + f & \text{in } \Omega, \\ \Delta u = U & \text{in } \Omega, \\ \Delta \Phi = -[u, u] & \text{in } \Omega, \\ \Delta \varphi = \Phi & \text{in } \Omega, \end{cases} \tag{4.4}$$

to which we assign the homogeneous Dirichlet boundary conditions, derived from the simply supported ones

$$\begin{cases} u = 0 & \text{in } \partial\Omega, \\ U = 0 & \text{in } \partial\Omega, \\ \varphi = 0 & \text{in } \partial\Omega, \\ \Phi = 0 & \text{in } \partial\Omega. \end{cases} \tag{4.5}$$

We know that (4.1) and (4.4) are equivalent [159] when the boundary is sufficiently regular and the solution is smooth enough, so from now on we just consider the latter.

Finally, since we are interested in the behaviour of the plate at rest under compression, we can set the external body force $f = 0$ restricting ourselves to the study of the homogeneous system. Moreover, through the function h, we can model different kind of stresses at the boundaries. Indeed, if we choose $h(x, y) = -\frac{1}{2}y^2$, the linear part of the Monge-Ampére bracket becomes $[\mu h, u] = -\mu u_{xx}$ and the system reads as

$$\begin{cases} \Delta U + \mu u_{xx} = [\varphi, u] & \text{in } \Omega, \\ \Delta u = U & \text{in } \Omega, \\ \Delta \Phi = -[u, u] & \text{in } \Omega, \\ \Delta \varphi = \Phi & \text{in } \Omega, \end{cases} \tag{4.6}$$

where we are assuming that the compression is acting on the edges parallel to the y direction (see Figure 4.1). We also remark that if instead we chose $h(x, y) = -\frac{1}{2}(x^2 + y^2)$, we would have the stress component given by $[\mu h, u] = -\mu\Delta u$, in which case the compression have to be considered to act on the whole boundary.

4.1.1 Weak formulation

Starting from the PDE setting described in the previous chapters, we are now able to set the model in the variational mathematical framework, which we will consider for the numerical investigation. So let us consider $\mu \in \mathcal{P} \subset \mathbb{R}$, where $\mathcal{P}$ here is the one dimensional bifurcation parameter space, $\Omega \subset \mathbb{R}^2$ is the rectangular domain, that we

identify with the plate, whereas $\mathbb{X} = \mathbb{X}(\Omega) = \left(H_0^1(\Omega)\right)^4$ is the Hilbert space in which we will seek the solution and $\mathbb{X}'$ its dual space.

Hence, we can represent the nonlinear PDE given by the Von Kármán system (4.4) as the parametrized mapping $G : \mathbb{X} \times \mathcal{P} \to \mathbb{X}'$ we have defined in Section 2.1. Thus, its weak formulation reads: given $\mu \in \mathcal{P}$, find $X(\mu) := (u(\mu), U(\mu), \varphi(\mu), \Phi(\mu)) \in \mathbb{X}$ such that (omitting the μ-dependence)

$$\begin{cases} (\nabla u, \nabla w)_{L^2(\Omega)} + (U, w)_{L^2(\Omega)} = 0 & \forall\, w \in H_0^1(\Omega), \\ (\nabla U, \nabla v)_{L^2(\Omega)} + ([\mu h + \varphi, u]\,, v)_{L^2(\Omega)} = 0 & \forall\, v \in H_0^1(\Omega), \\ (\nabla \varphi, \nabla \theta)_{L^2(\Omega)} + (\Phi, \theta)_{L^2(\Omega)} = 0 & \forall\, \theta \in H_0^1(\Omega), \\ (\nabla \Phi, \nabla \psi)_{L^2(\Omega)} - ([u, u]\,, \psi)_{L^2(\Omega)} = 0 & \forall\, \psi \in H_0^1(\Omega), \end{cases} \tag{4.7}$$

in which we embed the simply supported boundary conditions in the choice of the space $H_0^1(\Omega)$, where each component of the test function $Y := (w, v, \theta, \psi)$ resides. We have denoted with $(\cdot, \cdot)_{L^2(\Omega)}$ the usual inner product in the Hilbert space $L^2(\Omega)$. Moreover, we note that all the boundary terms vanish due to the choice of simply supported boundary conditions (4.3).

In this case, the parametrized variational form $g(\cdot, \cdot; \mu)$ introduced in (2.2) is defined as follows

$$\begin{aligned} g(X(\mu), Y; \mu) = {} & a(u(\mu), w) + b(U(\mu), w) + a(U(\mu), v) \\ & + \mu c(h, u(\mu), v) + c(\varphi(\mu), u(\mu), v) + a(\varphi(\mu), \theta) \\ & + b(\Phi(\mu), \theta) + a(\varphi(\mu), \psi) - c(u(\mu), u(\mu), \psi) \\ \forall\, Y \in \mathbb{X},\ \forall \mu \in \mathcal{P}, & \end{aligned} \tag{4.8}$$

where the following bilinear and trilinear forms have been introduced

$$a(\eta, \omega) = \int_\Omega \nabla \eta \cdot \nabla \omega \, d\Omega\,, \qquad b(\eta, \omega) = \int_\Omega \eta \omega \, d\Omega\,, \qquad c(\eta, \omega, \zeta) = \int_\Omega [\eta, \omega]\, \zeta \, d\Omega\,.$$

The numerical treatment of the variational form including the bracket of Monge-Ampére obviously needs a nonlinear solver as depicted in Section 3.1.2. To this end, we compute the partial Fréchet derivative of $g(Z, \cdot; \mu)$ at $Z \in \mathbb{X}$ that can be expressed as

$$\begin{aligned} \mathrm{d}g[Z](X(\mu), Y; \mu) = {} & a(u(\mu), w) + b(U(\mu), w) + a(U(\mu), v) \\ & + \mu c(h, u(\mu), v) + c(\varphi(\mu), Z_1, v) + c(Z_3, u(\mu), v) + a(\varphi(\mu), \theta) \\ & + b(\Phi(\mu), \theta) + a(\Phi(\mu), \psi) - c(u(\mu), Z_1, \psi) - c(Z_1, u(\mu), \psi) \\ \forall\, Z, Y \in \mathbb{X},\ \forall \mu \in \mathcal{P}, & \end{aligned} \tag{4.9}$$

where we denoted with $Z = (Z_1, Z_2, Z_3, Z_4)$ the components of the point in which we are computing the derivative.

4.1.2 Bifurcation analysis of the buckling problem

The focus of this application is the detection of the multiple solutions of the equations (4.4). In order to better explain the lack of uniqueness in this setting, let us go back to the physics behind the Von Kármán model.

We notice that, due to the symmetry of the problem, when the plate starts buckling we can expect at least two different configurations. In fact, for a given value of the compression parameter μ, if a given displacement is a solution of the system, then of course the displacement which can be obtained reflecting this configuration with respect to the plane where the plate lies, is again a solution of the system. Moreover, for each value of $\mu \in \mathcal{P}$, the equation (4.1) with zero external force (namely $f = 0$) always admits the trivial solution $(u, \varphi) = (0, 0)$, which corresponds to the original undeformed configuration.

From these considerations, we understand that also for a fixed value of the compression, the problem can be ill-posed. Which solution should we expect from our numerical solver? How can we know that the solution we found is unique in the interval considered?

We understood from Section 2.2 that an answer to these questions is strictly linked with the location of the bifurcation points. We can numerically investigate the equations for each value of the parameter $\mu \in \mathcal{P}$ observing when the buckling phenomena occur, despite in principle we have no a priori information about the parameter space $\mathcal{P}$. Unfortunately, this is a computationally very expensive task and thus we can apply the detection tool based on the eigenvalue analysis that we presented in Section 3.1.3.

From the mathematical standpoint, we have analysed the path pursued in [19, 4, 14], where they highlighted the link between the bifurcation points and the behaviour of the eigenvalues of the linearized problem. Therefore, if we linearize the equations in (4.4) around the trivial solution $X = 0$, the system we obtain is simply given by

$$\begin{cases} \Delta U = \mu\,[h, u] & \text{in } \Omega, \\ \Delta u = U & \text{in } \Omega, \end{cases} \qquad \text{with} \qquad \begin{cases} u = 0 & \text{in } \partial\Omega, \\ U = 0 & \text{in } \partial\Omega. \end{cases} \tag{4.10}$$

This connection is not surprising, in fact, as said before, from ODE's theory we know that the stability of the solutions is linked to eigenvalues that change sign, i.e. cross the imaginary axis varying μ. In particular, in connection to Proposition 2.2.1, we present the results from [19, 14] which apply directly to the Von Kármán model.

Theorem 4.1.1. *Bifurcation points of Von Kármán equations* (4.4) *with respect to the trivial solution* $X = 0$ *can occur only at the eigenvalues of the linearized problem* (4.10).

Thus, for Von Kármán equations we know that every bifurcation point is an eigenvalue of the linearized problem, but this assertion can also be inverted. Indeed, if we assume that all the eigenvalues are real, positive and ordered in such a way that $0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq \ldots$, from [19] we can also reverse the statement of Theorem 4.1.1 as follows.

Theorem 4.1.2. *From each eigenvalue of the system* (4.10) *at least one branch of non-trivial solution of* (4.1) *bifurcates. In particular, one branch bifurcates from a simple eigenvalue and*

at least two branches bifurcate from a multiple eigenvalue. Furthermore, if μ_1 is the smallest eigenvalue then for every value $\mu \leq \mu_1$ the unforced system has no non-trivial solutions.

Now we have all the definitions needed to set the Von Kármán model in finite dimension, that is finding the numerical solution which efficiently approximates the real one, investigating the bifurcation diagram and of course exploring the eigenvalue properties of the system.

4.2 Numerical approximation

Here we want to recast the weak formulation (4.8) in the finite dimensional space $\mathbb{X}_{\mathcal{N}}$ of dimension $\mathcal{N}$, as we did in general in (3.3). Moreover, in order to embed the simply supported boundary conditions also at the discrete level, we set $\mathbb{X}_{\mathcal{N}} = \mathbb{X}_h \cap \mathbb{K}_r^0$ where we have denoted with

$$\mathbb{K}_r^0 = \{V \in \mathbb{K}_r : V|_{\partial\Omega} = 0\} \tag{4.11}$$

the space of globally continuous functions that are polynomials of degree r on the single element of the triangulation $\mathcal{T}_h$ of the domain, which vanish on the boundary.

To provide a clear matrix representation of the application of the Galerkin FE method, we present the projected weak formulation. In this case the Newton method (3.10) reads: given $\mu \in \mathcal{P}$ and an initial guess $X_{\mathcal{N}}^0 = (u_{\mathcal{N}}^0, U_{\mathcal{N}}^0, \varphi_{\mathcal{N}}^0, \Phi_{\mathcal{N}}^0) \in \mathbb{X}_{\mathcal{N}}$ for $k = 0, 1, \dots$ until convergence we seek $\delta X_{\mathcal{N}} = (\delta u_{\mathcal{N}}, \delta U_{\mathcal{N}}, \delta\varphi_{\mathcal{N}}, \delta\Phi_{\mathcal{N}}) \in \mathbb{X}_{\mathcal{N}}$ such that

$$\begin{cases} a(\delta u_{\mathcal{N}}, w_{\mathcal{N}}) + b(\delta U_{\mathcal{N}}, w_{\mathcal{N}}) = a(u_{\mathcal{N}}^k, w_{\mathcal{N}}) + b(U_{\mathcal{N}}^k, w_{\mathcal{N}}) & \forall\, w_{\mathcal{N}} \in \mathbb{X}_{\mathcal{N}}, \\ a(\delta U_{\mathcal{N}}, v_{\mathcal{N}}) + c(\delta\varphi_{\mathcal{N}}, u_{\mathcal{N}}^k, v_{\mathcal{N}}) + c(\varphi_{\mathcal{N}}^k, \delta u_{\mathcal{N}}, v_{\mathcal{N}}) & \\ \qquad +\mu c(h, \delta u_{\mathcal{N}}, v_{\mathcal{N}}) = a(U_{\mathcal{N}}^k, v_{\mathcal{N}}) + c(\varphi_{\mathcal{N}}^k, u_{\mathcal{N}}^k, v_{\mathcal{N}}) & \\ \qquad +\mu c(h, u_{\mathcal{N}}^k, v_{\mathcal{N}}) & \forall\, v_{\mathcal{N}} \in \mathbb{X}_{\mathcal{N}}, \\ a(\delta\varphi_{\mathcal{N}}, \theta_{\mathcal{N}}) + b(\delta\Phi_{\mathcal{N}}, \theta_{\mathcal{N}}) = a(\varphi_{\mathcal{N}}^k, \theta_{\mathcal{N}}) + b(\Phi_{\mathcal{N}}^k, \theta_{\mathcal{N}}) & \forall\, \theta_{\mathcal{N}} \in \mathbb{X}_{\mathcal{N}}, \\ a(\delta\Phi_{\mathcal{N}}, \psi_{\mathcal{N}}) - c(\delta u_{\mathcal{N}}, u_{\mathcal{N}}^k, \psi_{\mathcal{N}}) & \\ \qquad -c(u_{\mathcal{N}}^k, \delta u_{\mathcal{N}}, \psi_{\mathcal{N}}) = a(\Phi_{\mathcal{N}}^k, \psi_{\mathcal{N}}) - c(u_{\mathcal{N}}^k, u_{\mathcal{N}}^k, \psi_{\mathcal{N}}) & \forall\, \psi_{\mathcal{N}} \in \mathbb{X}_{\mathcal{N}}, \end{cases} \tag{4.12}$$

and then set $X_{\mathcal{N}}^{k+1} = X_{\mathcal{N}}^k - \delta X_{\mathcal{N}}$.

We can finally present the correspondent matrix formulation, that follows directly from (4.12). Denoting the solution vector with $\mathsf{X}_{\mathcal{N}} = (\mathsf{u}_{\mathcal{N}}, \mathsf{U}_{\mathcal{N}}, \varphi_{\mathcal{N}}, \Phi_{\mathcal{N}})$ we can write the algebraic form (3.10) of the system as

$$\begin{pmatrix} \mathsf{A}_{\mathcal{N}} & \mathsf{B}_{\mathcal{N}} & 0 & 0 \\ \mathsf{C}_{\mathcal{N}}^2 + \mu\mathsf{C}_{\mathcal{N}}^0 & \mathsf{A}_{\mathcal{N}} & \mathsf{C}_{\mathcal{N}}^1 & 0 \\ 0 & 0 & \mathsf{A}_{\mathcal{N}} & \mathsf{B}_{\mathcal{N}} \\ -\mathsf{C}_{\mathcal{N}}^1 - \mathsf{C}_{\mathcal{N}}^3 & 0 & 0 & \mathsf{A}_{\mathcal{N}} \end{pmatrix} \begin{pmatrix} \delta\mathsf{u}_{\mathcal{N}} \\ \delta\mathsf{U}_{\mathcal{N}} \\ \delta\varphi_{\mathcal{N}} \\ \delta\Phi_{\mathcal{N}} \end{pmatrix} = \begin{pmatrix} \mathsf{A}_{\mathcal{N}}\mathsf{u}_{\mathcal{N}}^k + \mathsf{B}_{\mathcal{N}}\mathsf{U}_{\mathcal{N}}^k \\ \mathsf{A}_{\mathcal{N}}\mathsf{U}_{\mathcal{N}}^k + \mathsf{C}_{\mathcal{N}}^1\mathsf{u}_{\mathcal{N}}^k + \mu\mathsf{C}_{\mathcal{N}}^0\mathsf{u}_{\mathcal{N}}^k \\ \mathsf{A}_{\mathcal{N}}\varphi_{\mathcal{N}}^k + \mathsf{B}_{\mathcal{N}}\Phi_{\mathcal{N}}^k \\ \mathsf{A}_{\mathcal{N}}\Phi_{\mathcal{N}}^k - \mathsf{C}_{\mathcal{N}}^1\mathsf{u}_{\mathcal{N}}^k \end{pmatrix}, \tag{4.13}$$

with the high fidelity Jacobian $\mathsf{J}_{\mathcal{N}}$ and residual $\mathsf{G}_{\mathcal{N}}$ are given by

$$\mathsf{J}_{\mathcal{N}}(\mathsf{X}^k_{\mathcal{N}}(\boldsymbol{\mu});\boldsymbol{\mu}) = \begin{pmatrix} \mathsf{A}_{\mathcal{N}} & \mathsf{B}_{\mathcal{N}} & 0 & 0 \\ \mathsf{C}^2_{\mathcal{N}} + \mu \mathsf{C}^0_{\mathcal{N}} & \mathsf{A}_{\mathcal{N}} & \mathsf{C}^1_{\mathcal{N}} & 0 \\ 0 & 0 & \mathsf{A}_{\mathcal{N}} & \mathsf{B}_{\mathcal{N}} \\ -\mathsf{C}^1_{\mathcal{N}} - \mathsf{C}^3_{\mathcal{N}} & 0 & 0 & \mathsf{A}_{\mathcal{N}} \end{pmatrix} \quad \text{and}$$

$$\mathsf{G}_{\mathcal{N}}(\mathsf{X}^k_{\mathcal{N}}(\boldsymbol{\mu});\boldsymbol{\mu}) = \begin{pmatrix} \mathsf{A}_{\mathcal{N}}\mathsf{u}^k_{\mathcal{N}} + \mathsf{B}_{\mathcal{N}}\mathsf{U}^k_{\mathcal{N}} \\ \mathsf{A}_{\mathcal{N}}\mathsf{U}^k_{\mathcal{N}} + \mathsf{C}^1_{\mathcal{N}}\mathsf{u}^k_{\mathcal{N}} + \mu \mathsf{C}^0_{\mathcal{N}}\mathsf{u}^k_{\mathcal{N}} \\ \mathsf{A}_{\mathcal{N}}\boldsymbol{\varphi}^k_{\mathcal{N}} + \mathsf{B}_{\mathcal{N}}\boldsymbol{\Phi}^k_{\mathcal{N}} \\ \mathsf{A}_{\mathcal{N}}\boldsymbol{\Phi}^k_{\mathcal{N}} - \mathsf{C}^1_{\mathcal{N}}\mathsf{u}^k_{\mathcal{N}} \end{pmatrix},$$

where we introduced the matrices as follows

$$(\mathsf{A}_{\mathcal{N}})_{ij} = a(E^j, E^i)\,, \quad (\mathsf{B}_{\mathcal{N}})_{ij} = b(E^j, E^i)\,, \quad (\mathsf{C}^0_{\mathcal{N}})_{ij} = c(h, E^j, E^i)\,,$$
$$(\mathsf{C}^1_{\mathcal{N}})_{ij} = c(E^j, \mathsf{u}^k_{\mathcal{N}}, E^i)\,, \quad (\mathsf{C}^2_{\mathcal{N}})_{ij} = c(\varphi^k_{\mathcal{N}}, E^j, E^i)\,, \quad (\mathsf{C}^3_{\mathcal{N}})_{ij} = c(\mathsf{u}^k_{\mathcal{N}}, E^j, E^i)\,.$$

Note that because of the symmetry of the bracket of Monge-Ampére, we easily obtain that it holds $\mathsf{C}^1_{\mathcal{N}} \equiv \mathsf{C}^3_{\mathcal{N}}$.

As we said previously, the reduced basis method shares the projection properties of the Galerkin method. Indeed, the weak formulation that we obtain from the application of the Newton method at the reduced level, presented in (3.14), reads as: given $\mu \in \mathcal{P}$ and an initial guess $X^0_N \in \mathbb{X}_N$ for $k = 0, 1, \dots$ until convergence we seek $\delta X_N = (\delta u_N, \delta U_N, \delta \varphi_N, \delta \Phi_N) \in \mathbb{X}_N$ such that

$$\begin{cases} a(\delta u_N, w_N) + b(\delta U_N, w_N) = a(u^k_N, w_N) + b(U^k_N, w_N) & \forall\, w_N \in \mathbb{X}_N, \\ a(\delta U_N, v_N) + c(\delta \varphi_N, u^k_N, v_N) + c(\varphi^k_N, \delta u_N, v_N) & \\ \qquad +\mu c(h, \delta u_N, v_N) = a(U^k_N, v_N) + c(\varphi^k_N, u^k_N, v_N) & \\ \qquad +\mu c(h, u^k_N, v_N) & \forall\, v_N \in \mathbb{X}_N, \\ a(\delta \varphi_N, \theta_N) + b(\delta \Phi_N, \theta_N) = a(\varphi^k_N, \theta_N) + b(\Phi^k_N, \theta_N) & \forall\, \theta_N \in \mathbb{X}_N, \\ a(\delta \Phi_N, \psi_N) - c(\delta u_N, u^k_N, \psi_N) & \\ \qquad -c(u^k_N, \delta u_N, \psi_N) = a(\Phi^k_N, \psi_N) - c(u^k_N, u^k_N, \psi_N) & \forall\, \psi_N \in \mathbb{X}_N, \end{cases} \tag{4.14}$$

and then set $X^{k+1}_N = X^k_N - \delta X_N$. Also in this case, the matrix formulation follows directly from the expressions above. Indeed, we can obtain the compact version (3.20), with the reduced Jacobian

$$\mathsf{J}_N(\mathsf{X}^k_N(\boldsymbol{\mu});\boldsymbol{\mu}) = \begin{pmatrix} \mathsf{A}_N & \mathsf{B}_N & 0 & 0 \\ \mathsf{C}^2_N + \mu \mathsf{C}^0_N & \mathsf{A}_N & \mathsf{C}^1_N & 0 \\ 0 & 0 & \mathsf{A}_N & \mathsf{B}_N \\ -\mathsf{C}^1_N - \mathsf{C}^3_N & 0 & 0 & \mathsf{A}_N \end{pmatrix}, \tag{4.15}$$

having the same structure of the finite element one. If we introduce the transformation matrices with respect to the different components of the solution, V_u and V_φ,

respectively for u and φ, we can define the reduced matrices in the following way:

$$\mathsf{C}_N^0 = \mathsf{V}^T \mathsf{C}_{\mathcal{N}}^0 \mathsf{V} \,, \quad \mathsf{C}_N^1 = \sum_{n=1}^{N} u_N^{(n)} \mathsf{V}_u^T \mathsf{C}_{\mathcal{N}}^1(\Sigma_u^n) \mathsf{V}_u \,, \quad \mathsf{C}_N^2 = \sum_{n=1}^{N} \varphi_N^{(n)} \mathsf{V}_\varphi^T \mathsf{C}_{\mathcal{N}}^2(\Sigma_\varphi^n) \mathsf{V}_\varphi \,,$$
$$\mathsf{C}_N^3 = \sum_{n=1}^{N} \varphi_N^{(n)} \mathsf{V}_\varphi^T \mathsf{C}_{\mathcal{N}}^3(\Sigma_\varphi^n) \mathsf{V}_\varphi \,, \quad \mathsf{A}_N = \mathsf{V}^T \mathsf{A}_{\mathcal{N}} \mathsf{V} \,, \quad \mathsf{B}_N = \mathsf{V}^T \mathsf{B}_{\mathcal{N}} \mathsf{V} \,.$$

Moreover, we highlight that also the reduced residual vector has the same form of the finite element one, indeed it is given by

$$\mathsf{G}_N(\mathsf{X}_N^k(\boldsymbol{\mu}); \boldsymbol{\mu}) = \begin{pmatrix} \mathsf{A}_N \mathsf{u}_N^k + \mathsf{B}_N \mathsf{U}_N^k \\ \mathsf{A}_N U_N^k + \mathsf{C}_N^1 \mathsf{u}_N^k + \mu \mathsf{C}_N^0 \mathsf{u}_N^k \\ \mathsf{A}_N \varphi_N^k + \mathsf{B}_N \Phi_N^k \\ \mathsf{A}_N \Phi_N^k - \mathsf{C}_N^1 \mathsf{u}_N^k \end{pmatrix} . \tag{4.16}$$

We have now illustrated the main ingredients of the online phase, that allows an efficient evaluation of the solution and possibly related outputs for each choice of the parameter $\mu \in \mathcal{P}$. Despite this, as we already specified, one of the key point of this time savings is the affine decomposition assumption. While the model shows an affine dependence on the parameter μ, the reduced matrices we have just presented are clearly μ-dependent due to the presence of the solution which we used to linearize the system. Thus, since the nonlinearity complicates the setting, the affine-recovery techniques already presented in Section 3.2.4 are needed to recover efficiency. We will present an application of these in Section 4.2.5.

4.2.1 Spectral analysis

In the previous sections we discussed about the issue of the computational complexity of the problem itself, that we try to avoid using the ROM. It is clear that drawing the bifurcation diagram is still a difficult task. Indeed how can we investigate the parameter space $\mathcal{P}$ without having any information on the position of these points?

Taking some inspiration from [125, 124] and supported by the theoretical results given in Section 4.1.2, we tried to locate more precisely the buckling points. Our goal in this section is thus to investigate the stability properties with the help of the spectral problems built upon the Von Kármán system.

Let us consider the case of the plate compressed along the edges parallel to the y-direction (see Figure 4.1), and thus assuming the compression term to be of the form $h(x, y) = -\frac{1}{2}y^2$. Hence, we construct the eigenvalue problem for the linearized (around the trivial solution) parametrized operator

$$\begin{cases} \Delta^2 u + \mu u_{xx} = \sigma_\mu u & \text{in } \Omega = [0, L] \times [0, 1], \\ u = \Delta u = 0 & \text{in } \partial\Omega, \end{cases} \tag{4.17}$$

where we want to find, varying the buckling parameter μ, the pair $(u, \sigma_\mu) \in H_0^1(\Omega) \times \mathbb{R}$, whose components are respectively the eigenfunction and eigenvalue. We will

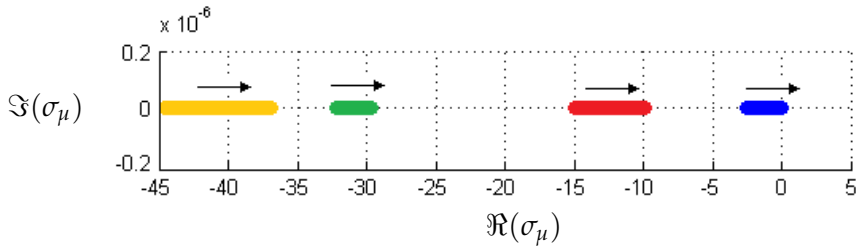


Figure 4.2: Behaviour of the first four eigenvalues σ_μ for $\mu \in [30, 40]$.

restrict our investigations to the square plate with $L = 1$ and the rectangular one with $L = 2$.

We are interested in the behaviour of σ_μ with respect to μ, in fact since the sign of the eigenvalues is strictly linked with the stability property of the solution, we aim at observing that the first eigenvalue crosses the y-axis when the plate is buckling. This is exactly what we found, indeed in Figure 4.2 for $L = 1$ we can see the behaviour of the first four eigenvalues σ_μ for $\mu \in [30, 40]$ and if we use a μ-step equals to $\Delta\mu = 0.5$ the crossing happens for the value $\mu^* \approx 39.5$.

This should tell us that after this value we have a change of the stability properties in connection to the presence of a new solution branch. Indeed, as we will see, because of the symmetry, there will be at least two new branches for each simple eigenvalue crossing the origin. Investigating greater values of the compression parameter μ in the space $\mathcal{P}$, we also observed the crossing of the successive eigenvalues.

Moreover, if we solve the eigenvalue problem (4.17) for the case with $L = 2$, we note that instead of having a single eigenvalue crossing the origin at $\mu^* \approx 62$ (as for the square plate) we found a double crossing (see Figure 4.3). This fact has a relevant consequence from the physical point of view, and its consequence will be clear in the bifurcation diagram for the rectangular plate.

What we just showed is computationally heavy to perform, so to keep in mind the efficiency as keyword of the whole analysis, we tried two other ways that helps to validate the results and at the same time reduce the computational time. In the first approach, we consider the linear problem (4.10) but in its original form

$$\begin{cases} \Delta^2 u + \mu u_{xx} = 0 & \text{in } \Omega = [0, L] \times [0, 1], \\ u = \Delta u = 0 & \text{in } \partial\Omega, \end{cases} \tag{4.18}$$

that has non-trivial solutions for $m, n = 1, 2, \dots$ given by

$$u_{m,n} = sin\left(\frac{m\pi x}{L}\right) sin\left(n\pi y\right) \quad \text{if and only if} \quad \mu_{m,n} = \left(\frac{\pi}{L}\right)^2 \left[m + \frac{n^2 L^2}{m}\right]^2, \tag{4.19}$$

where $u_{m,n}$ and $\mu_{m,n}$, with a little abuse of notation, can be considered as the eigenfunctions and eigenvalues for this new generalized eigenvalue problem. Now we have a much simpler problem, indeed since μ plays the role of the eigenvalue, the system is

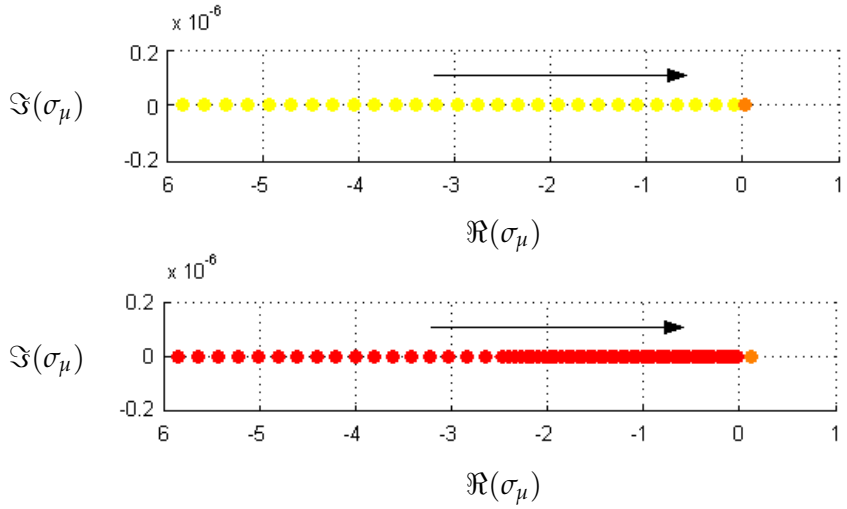


Figure 4.3: Double eigenvalue for the rectangular plate crossing for $\mu \in [50, 62]$.

no longer parametrized. This provides us also an explicit expression for the spectrum, that we can use to validate the results.

Using the formula (4.19) we find the analytical values of the eigenvalues of the problem (4.18), which turn out to be the buckling parameters, i.e. the bifurcation points. For the square and rectangular plate case we obtain respectively

$$L = 1: \quad \mu_{1,1} = 4\pi^2,\ \mu_{2,1} = \frac{25}{4}\pi^2,\ \mu_{3,1} = \frac{100}{9}\pi^2,\ \mu_{4,1} = \frac{289}{16}\pi^2,$$
$$L = 2: \quad \mu_{2,1} = 4\pi^2,\ \mu_{3,1} = \frac{169}{36}\pi^2,\ \mu_{1,1} = \mu_{4,1} = \frac{25}{4}\pi^2.$$

Indeed, we found the value $\mu^* \approx \mu_{1,1} \simeq 39.47$ predicted in Figure 4.2 for the square plate, but we also notice the presence of the double eigenvalue $\mu_{1,1} = \mu_{4,1} \simeq 61.68$ that confirms what we saw in Figure 4.3 for the rectangular one.

Furthermore, we can investigate analytically the multiplicity of the eigenvalues while varying the length of the plate. We just have to impose the condition $\mu_{m,n} = \mu_{m+k,n}$ for some $k \in \mathbb{N}$, from which we can deduce the relation $L = \frac{\sqrt{m(m+k)}}{n}$. If we plug in the values $m = n = 1$ and $k = 3$, which characterize the rectangular plate, we find the value $L = 2$ as we expected.

Finally, using the techniques in [7], is an easy task to prove the following theorem [106] that provides us a tool to better understand how good is our approximation.

Theorem 4.2.1. *There exists a strictly positive constant C such that*

$$|\mu - \mu_h| \leq Ch^2,$$

where μ_h is an approximation, dependent on the sparsity of the grid, of the true eigenvalue μ.

L=1	h = 1.e-1	h = 6.e-2	h = 1.e-2	h = 6.e-3	Order	Exact
$\mu_{1,1}$	39.91	39.59	39.48	39.47	1.98	39.47841
$\mu_{2,1}$	63.70	62.20	61.70	61.69	1.99	61.68502
$\mu_{3,1}$	116.63	111.54	109.73	109.68	1.97	109.66227

Table 4.1: Buckling coefficients for the square plate ($L = 1$) with the average order of convergence.

L=2	h = 1.e-1	h = 6.e-2	h = 1.e-2	h = 6.e-3	Order	Exact
$\mu_{2,1}$	40.74	39.76	39.48	39.48	2.05	39.47841
$\mu_{3,1}$	49.15	46.97	46.35	46.33	2.34	46.33230
$\mu_{1,1}$	62.08	61.79	61.68	61.68	1.98	61.68502
$\mu_{4,1}$	67.44	63.02	61.73	61.69	2.05	61.68502

Table 4.2: Buckling coefficients for the rectangular plate ($L = 2$) with the average order of convergence.

The theorem above is crucial when we are dealing with problems for which we do not know an explicit expression of the eigenvalues. For the sake of completeness we provide in Tables 4.1 and 4.2 the order of convergence results respectively for the square and rectangular plate, which agree with the theoretical ones.

To conclude this section we briefly discuss also the second straightforward way to reduce the computational complexity of solving multiple times a full order eigenvalue problem. Coming back to the parametrized eigenproblem (4.17), we can apply also to this system the Reduced Basis method [96, 99, 59, 78], and thanks to the affine decomposition, we easily obtain in a more efficient way the same behaviour of the results discussed before, as we can see from Figures 4.4 and 4.5. Of course, the question on how to built the ROM for this problem can have multiple answers. We projected the eigenvalue problem on the solution of the Von Kármán system, but as we can see form the plots we are only able to detect the crossing instead of a good approximation of the first eigenvalue. A more precise and accurate way is to project over the space spanned by a basis built upon some selected eigenvectors. Since for this test case the characterization given in (4.18) completely remove the parameter dependence, we do not discuss further this last approach. In fact we can easily embed the computation for the eigenproblem (4.18) in the offline phase and have all the informations to build the ROM.

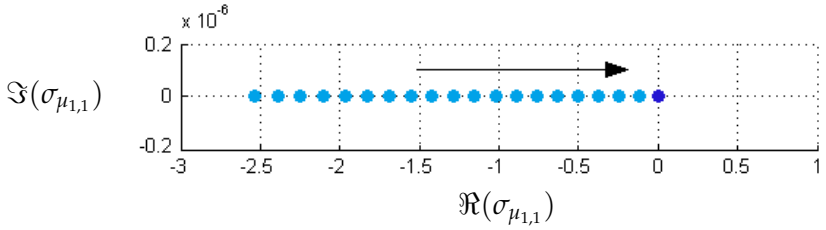


Figure 4.4: Full order first eigenvalue $\sigma_{\mu_{1,1}}$ for $L = 1$ crossing at $\mu^* = 39.5$.

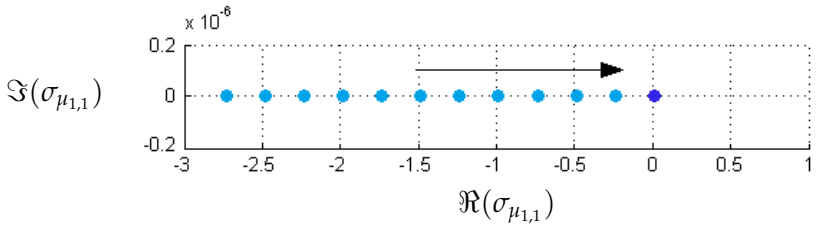


Figure 4.5: Reduced order first eigenvalue $\sigma_{\mu_{1,1}}$ for $L = 1$ crossing at $\mu^* = 39.5$.

4.2.2 Results and test problems

In this section, we will show how the buckling phenomena, i.e. the loss of uniqueness of the solution,can be analysed through the bifurcation diagram, both in the square and rectangular plate cases.

Thus, in order to recap, let us consider the Von Kármán plate equations, with simply supported BCs, in the bi-dimensional domain $\Omega = [0, L] \times [0, 1]$ given by

$$\begin{cases} \Delta^2 u + \mu u_{xx} = [\varphi, u] & \text{in } \Omega, \\ \Delta^2 \varphi = -[u, u] & \text{in } \Omega, \\ u = \Delta u = 0 & \text{in } \partial\Omega, \\ \varphi = \Delta \varphi = 0 & \text{in } \partial\Omega, \end{cases} \tag{4.20}$$

and we are interested in the study of the solution while varying the buckling parameter μ, which describes the compression along the edges parallel to the y-axis.

This model was previously numerically investigated by many authors [23, 39, 138], but as we already said the biggest issue was the computational complexity, that we overcome by means of the Reduced Basis method.

The investigation done with the eigenproblem give us the necessary information that the parameters responsible of the buckling are in the interval $\mathcal{P} = [35, 65]$, which we chose as our parameter domain. The main goal is the reconstruction of the bifurcation diagram, which represents for every value of $\mu \in \mathcal{P}_K$ on the x-axis, the correspondent value of the displacement u in its point of maximum modulo. Thus, in this case, the scalar measure output which allows to draw the diagram, will be the infinite norm of the displacement with its sign, namely $s(u) = (\operatorname{sgn} u) \max_{x \in \Omega} \|u\|_\infty$. As concerns the continuation technique, for these test cases we utilized the simple one described in

the Algorithm 1. Instead, for the initial guess, we implemented here the most straightforward strategy, where analytical sinusoidal functions, derived from considerations involving the eigenproblem (4.18), are used. Indeed, the modes of the plate will be easily characterized, together with the branch they belongs, by the number/type of cells that are showed by the contour plot of the displacement. Thus, we expected that properly selecting the initial guess, by the wavelength of the sinusoidal functions, we could recover the whole branching behaviour. As we will see, this approach worked well for both geometries.

4.2.2.1 Square plate test case

Here we present the results for the square plate case with $L = 1$. The FE space $\mathbb{X}_{\mathcal{N}}$, defined on the domain $\Omega = [0,1] \times [0,1]$ and discretized with quadratic elements $\mathbb{P}_2$, has dimension $\mathcal{N} = 9840$. We solved $N_{train} = 40$ high fidelity problems, obtaining information by the snapshots of both (e.g. upper) branches, for the offline phase, while the online approximation is constructed employing $N = 8$ basis. Moreover, for the branch reconstruction we chose $K = 60$ equispaced points (with continuation step $\Delta\mu = 0.5$), repeating the continuation loop in Algorithm 1 for each prescribed guess, that for this test case is chosen as $X_{guess} = \sin(\frac{\pi x}{O})\sin(\pi y)$ with the wavelength $O \in \{1, 0.5\}$ depending on the different Von Kármán branch to recover. We can finally present the bifurcation diagram in Figure 4.6. As we previously predicted, we can observe the buckling phenomena from the trivial solution. Moreover, we note that the first bifurcation happens for values of μ near $\mu_{1,1} \simeq 39.47$. This confirms the eigenvalue analysis that we considered in the previous section.

We did not stop at the first bifurcation, in fact choosing properly the initial guess we have been able to detect also the second bifurcation for the square plate. Again, this result is confirming what we predicted, since for values of μ near $\mu_{2,1} \simeq 61.68$ we obtain two other branches.

The physical symmetry issue is evident in both buckling points and the same will hold for the rectangular plate. Indeed, once we chose a bifurcation point, a solution from the upper branch is identical to the corresponding solution belonging to the lower one, but it is reflected with respect to the plane of the plate. Thus, for the first bifurcation emerging near $\mu_{1,1}$, looking at the contour plot, we observe the one cell like displacement depicted in Figure 4.7. While if we look at the second branch, so the one near $\mu_{2,1}$, we find the two cells like displacement as in Figure 4.8. So for the square plate case we obtained four different solutions for each $\mu \geq \mu_{2,1}$.

As concerns the RB method, we can now present the good results we have obtained. Indeed, as we can see in Figure 4.9, the reduced basis solution approximates perfectly not only the behaviour but also the order of magnitude. The remarkable point is that in order to obtain the solution on the right in Figure 4.9 we just solved a linear system of order 10 instead of the one given by the Galerkin FE method of order 10^4.

To be more precise we present in Table 4.3 a convergence result: the error between the truth approximation and the reduced one as a function of N. The error reported, $\varepsilon_N = \max_{\mu \in \mathcal{P}} \|u_{\mathcal{N}}(\mu) - u_N(\mu)\|_{H_0^1(\Omega)}$ is the maximum of the approximation

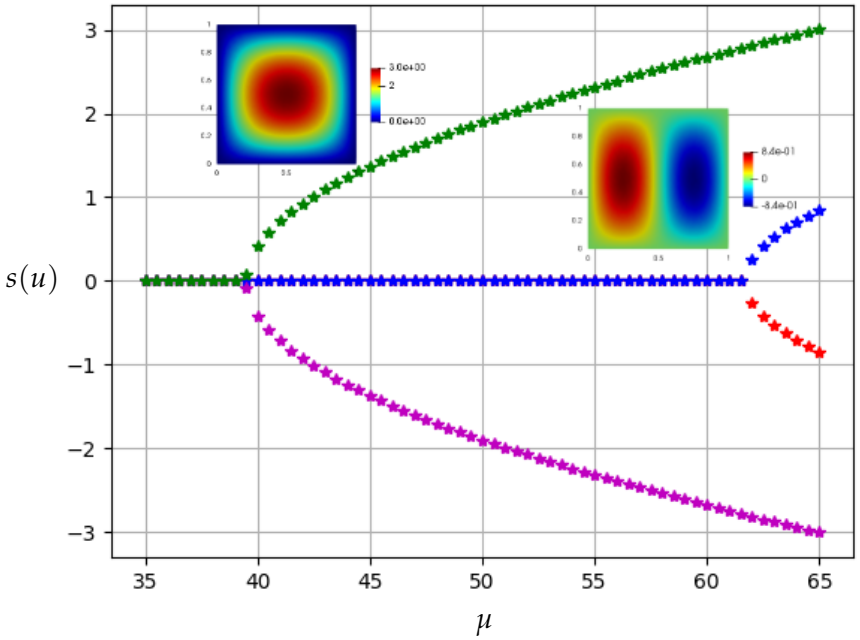


Figure 4.6: High fidelity bifurcation diagram for the square plate with $L = 1$.

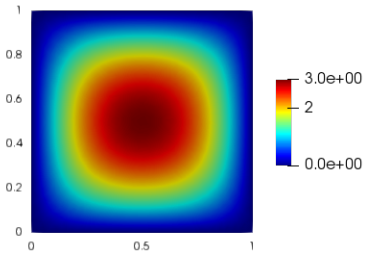


Figure 4.7: Full order one cell solution for the displacement u at $\mu = 65$ (green branch).

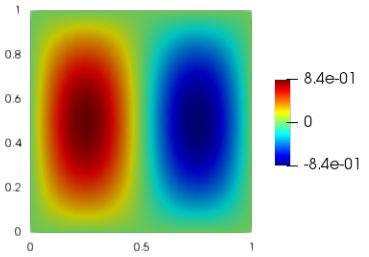


Figure 4.8: Full order two cells solution for the displacement u at $\mu = 65$ (blue branch).

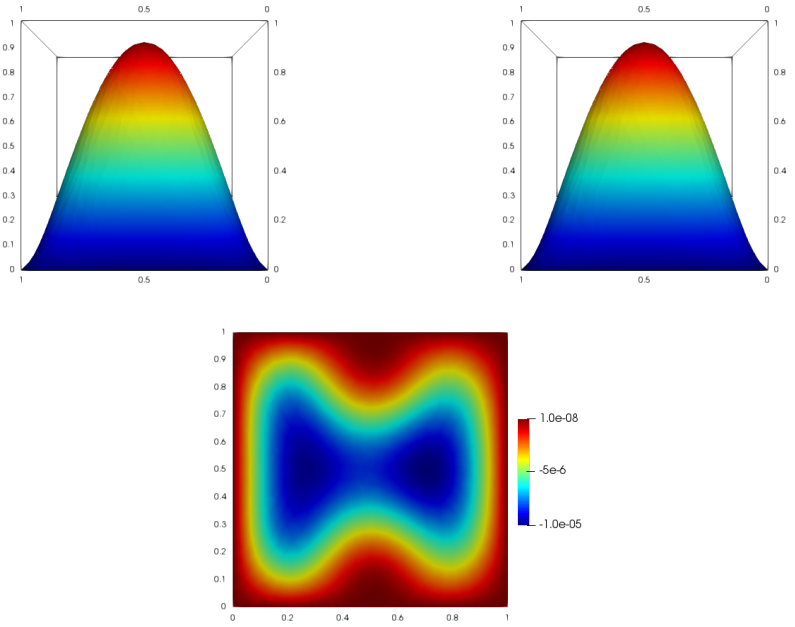


Figure 4.9: Comparison between the full order solution (on the left) and the reduced order one (on the right) for the displacement u at $\mu = 46$, belonging to the green branch. Below the RB contour error plot on the displacement u.

error for the displacement over a uniformly chosen test sample. Here we selected only the displacement, since the other variables are performing similarly and because the displacement is the physical variable we are interested in to draw the bifurcation diagram.

We highlight that we present here just the full order bifurcation diagram, in fact also in view of Table 4.3 and Figure 4.9, its reduced order version looks exactly the same. Furthermore, as we can see in Figure 4.10, the RB error for the reconstruction of the first and second branches shows good approximation results, having the maximum peaks at their bifurcation points $\mu_{1,1}$ and $\mu_{2,1}$.

We remark that we did not implemented the Greedy algorithm here, because it would be needed a suitable application of the BRR theory for the a posteriori error estimate [24, 154, 65, 32]. However, as we said, applying BRR theory at reduced level is not straightforward and we leave it for further future investigation.

Finally, as regards computational times, a RB evaluation $\mu \to u_N(\mu)$ requires almost $t_{RB} = 0.51$(s) for $N = 8$; while a FE solution $\mu \to u_{\mathcal{N}}(\mu)$ requires $t_{FE} = 1.21$(s): thus our RB online speed-up is only 2.37. This is of course not completely satisfactory, but it is obviously due to the lack of hyper-reduction techniques.

N	ε_N
1	6.61e+00
2	6.90e–01
3	7.81e–02
4	2.53e–02
5	1.88e–02
6	1.24e–02
7	9.02e–03
8	8.46e–03

Table 4.3: The Reduced Basis convergence with respect to the number of the basis N for the square plate case.

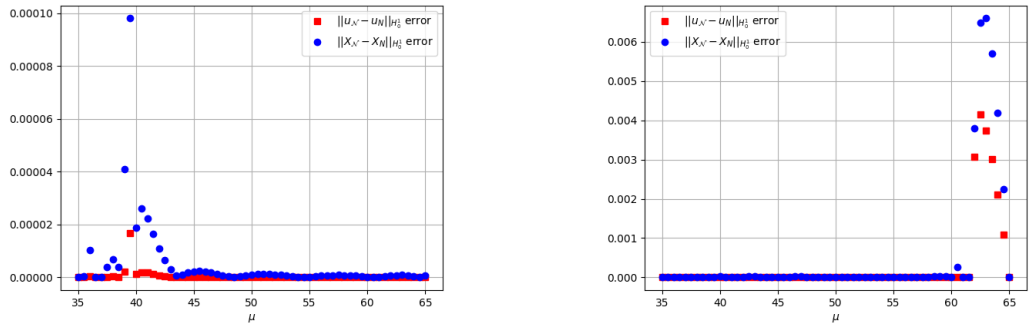


Figure 4.10: Reduced Basis error for the green and blue solution branches, emerging from $\mu_{1,1}$ and $\mu_{2,1}$, left and right respectively, with $L = 1$.

4.2.2.2 Rectangular plate test case

Now we can analyse the case of the rectangular plate with $L = 2$. The FE space $\mathbb{X}_{\mathcal{N}}$, defined on the domain $\Omega = [0, 2] \times [0, 1]$ is discretized again with $\mathbb{P}_2$ quadratic elements and has similar dimension. We computed $N_{train} = 20$ snapshots for each branch during the offline phase, while for the reduced manifold we chose $N = 8$ basis. As concerns the branch reconstruction, we design the continuation method with $K = 60$ equispaced points as before. Regardless the similar situation, from the eigenvalue problem analysis performed in Section 4.2.1, we understood that the guesses for the continuation loop had to be chosen as $X_{guess} = \sin(\frac{\pi x}{OL}) \sin(\pi y)$ with $O \in \{2, 1, \frac{2}{3}, 0.5\}$ in order to capture both the newly appeared bifurcating phenomenon and the multiple branching behaviour. Therefore, it is clear that in this case a greater number of computations had to be performed to recover the bifurcation diagram in Figure 4.11.

We have a different situation, in fact, varying the length of the domain Ω, we obtained a new bifurcation point within the same parameter space $\mathcal{P}$, but also the third branch changed its properties. Note that this sensitivity with respect to the

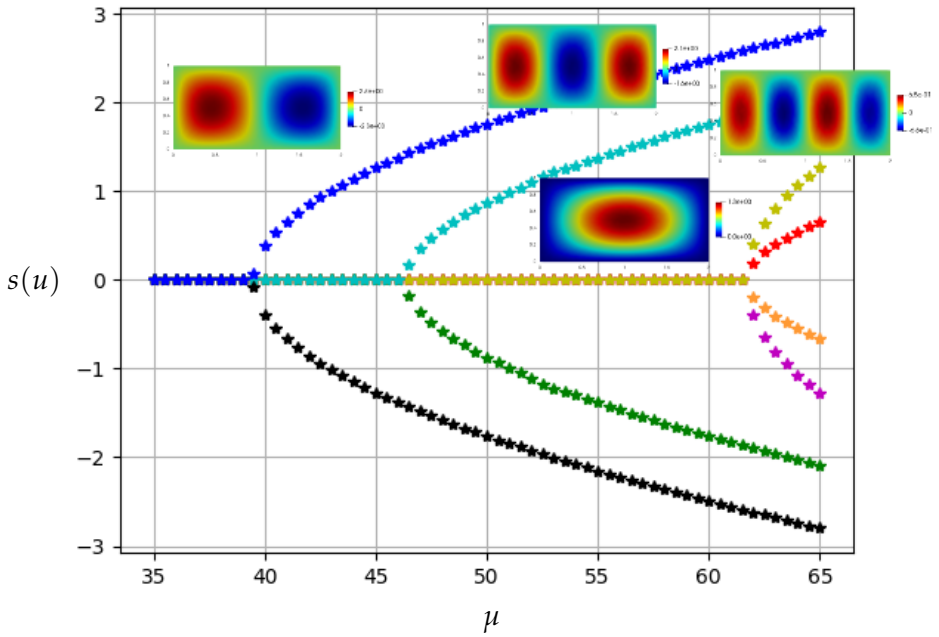


Figure 4.11: High fidelity bifurcation diagram for the rectangular plate with $L = 2$.

dimension of the plate is a big challenge from the computational point of view. We did not treat here L as a second *geometrical parameter* along with μ, but we will see an application of this for the models we will investigate later on.

In this case, the solution starts branching as before from the smallest eigenvalue $\mu_{2,1} \simeq 39.47$. Obviously the number of cells that are formed in the contour plot strictly depends on the length of the domain, for example here the first bifurcation is linked with the two cells configuration as shown in Figure 4.12.

We observed also a new bifurcation for values of μ near $\mu^* \approx 46.5$, as depicted in Figure 4.13, which is characterized by three cells and corresponds to the eigenvalue $\mu_{3,1} \simeq 46.33$. Finally, we comment the last buckling point, which appears to be qualitatively different from the previous ones. As we noticed from Figure 4.11, we have again a buckling for values of μ near $\mu^* \approx 61.68$, but this time the bifurcation is reflected by a pair of emerging branches. In fact, for the rectangular plate we have a multiple eigenvalue $\mu_{1,1} = \mu_{4,1}$ with multiplicity equals to two, which is the responsible of this double bifurcation. In practice what we obtained is a point from which start branching two sets of different solutions with one and four cells, respectively in Figure 4.14 and Figure 4.15. The same conclusions regarding the convergence error ε_N and computational savings can be established also in this case, with very similar results. Finally we show in Figure 4.16 that, also for the rectangular plate, the RB method works well approximating efficiently the solution. In fact, as before, we can see in Figure 4.17, the RB error for the reconstruction of the first two branches, having again the maximum peaks at their bifurcation points $\mu_{2,1}$ and $\mu_{3,1}$.

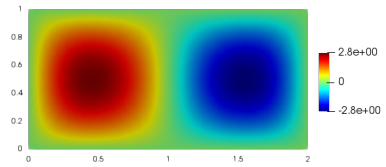


Figure 4.12: Full order two cells solution for the displacement u at $\mu = 65$ (blue branch).

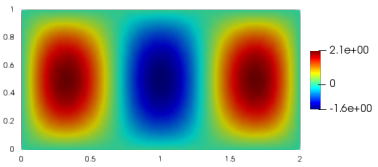


Figure 4.13: Full order three cells solution for the displacement u at $\mu = 65$ (cyan branch).

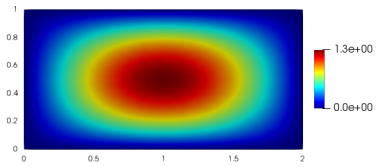


Figure 4.14: Full order one cell solution for the displacement u at $\mu = 65$ (yellow branch).

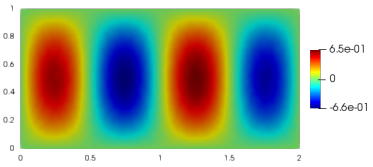


Figure 4.15: Full order four cells solution for the displacement u at $\mu = 65$ (red branch).

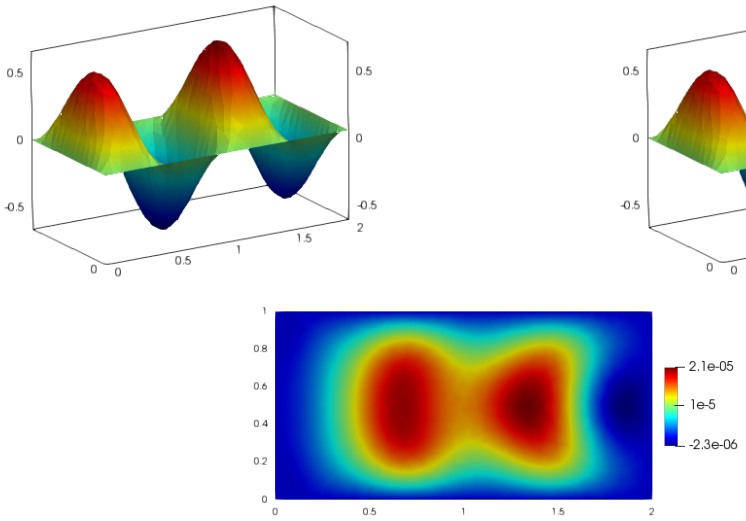


Figure 4.16: Comparison between the full order solution (on the left) and the reduced order one (on the right) for the displacement u at $\mu = 65$, belonging to the red branch. Below the RB contour error plot on the displacement u.

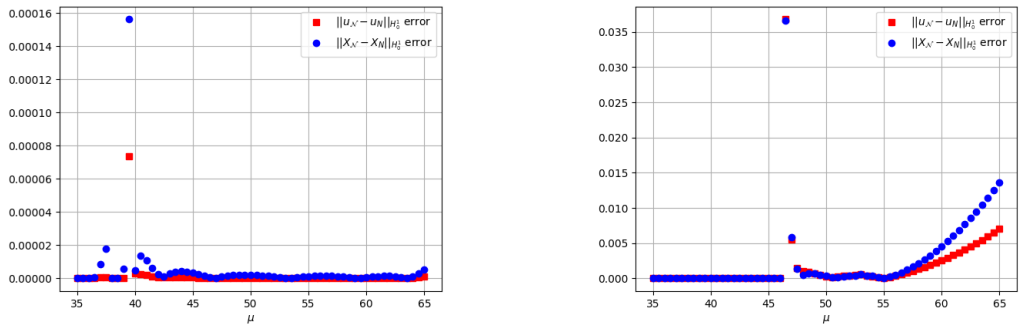


Figure 4.17: Reduced Basis error for the blue and cyan solution branches, emerging from $\mu_{2,1}$ and $\mu_{3,1}$, left and right respectively, with $L = 2$.

4.2.3 An investigation of secondary bifurcations

An interesting property of the Von Kármán model, that we did not investigate until now, is the so called *mode jumping* [40, 141]. As we observed in the previous bifurcation diagrams, each solution branch can be described by the number of cells which are visible from the contour plots of the displacement u. Thus, while we are reconstructing a bifurcating branch, increasing slowly the compression parameter, the wave number (or equivalently the number of cells) remains constant.

Actually, both experiments and theory [144, 102, 13] have shown that the wave number does not always remain constant increasing gradually the loading parameter after the buckling point. In fact, there exist some values of μ at which the buckling behaviour suddenly change.

From the description of this phenomenon it is evident that, mathematically speaking, this corresponds to secondary bifurcation. Indeed, a sudden change in the behaviour of the solution is a clear evidence of a bifurcation, but the key difference is that here it occurs for an already buckled state.

Therefore, the aim of this section is to numerically investigate the mode jumping for the Von Kármán model. The detection of the values at which such phenomenon happens is a complex task, since we can not use the parameter independent eigenvalue problem. Indeed, here we should linearize the system around the buckled non-trivial solution, which of course depends on μ.

Regardless this difficulty, a guess on their location can be obtained through the theoretical analysis performed in [102, 13], where the connection between multiple eigenvalues and secondary bifurcations was studied. In particular, they showed that a multiple bifurcation point may cause a splitting between primary bifurcation points and several secondary bifurcation points at different locations. Furthermore, these secondary critical points move along one or more of the primary branches.

For this reason we decided to analyse the rectangular plate with $L = 2$, where we observed the multiple eigenvalue given by $\mu_{1,1} = \mu_{4,1}$. We highlight that in the

previous analysis we did not find any evidence of this mode jumping phenomena thus, recalling the observation in Section 3.1.4, we adjusted the choice of the continuation method, which has guided our nonlinear solver in the bifurcation reconstruction. Due to the very complex behaviour, we restrict the approximation of this phenomenon only at the high fidelity level, but we would expect the same reduced error behaviour of the other branches, once that the secondary bifurcation is actually recovered during the offline phase.

We fixed the parameter space $\mathcal{P}_K$, with $K = 100$ equispaced point in $\mathcal{P}$. Here we did not consider a more difficult approach such as the pseudo arc-length algorithm, instead we pursued a much simpler route [142] based on a variant of the simple continuation method employing the Implicit Function Theorem. Indeed, in the following we will take advantage of the well-known formula that expresses the sensitivity of the solution to the parameter: given $\mu_j \in \mathcal{P}_K$ we can obtain

$$\frac{\partial X(\mu_j)}{\partial \mu} = -D_X G(X_j; \mu_j) / D_\mu G(X_j; \mu_j). \tag{4.21}$$

This results in a predictor-corrector algorithm where the pair $(\widetilde{\mathsf{X}}(\widetilde{\mu}_j), \widetilde{\mu}_j)$, which represents the guess, is actually given by $(\mathsf{X}(\mu_j) + \Delta\mu_j \mathsf{dX}_j, \mu_j + \Delta\mu_j)$ where we indicated with dX_j the discrete version of the expression (4.21).

With the previous approach we were only able to find primary bifurcations, while the modification we adopted allowed us to recover also the secondary bifurcation. In particular, the new branching behaviour emerge from the configuration with three cells, for a value of the compression parameter $\mu^* \approx 52.25$. As it is clearly visible in Figure 4.18 on the right, the secondary branch has lost any kind of symmetry with respect to the midline $\{x = 1\}$. In fact, the first two cells form the left changed their shape elongating towards the third one, which corresponding points lost displacement magnitude (see the contour plot on the left in Figure 4.18). We remark that it is straightforward to expect that also the displacements obtained through reflection with respect to the axis $\{x = 1\}$ and to the plane of the plate, are admissible solution coming from the secondary branches. Finally, we can present the full bifurcation diagram for the rectangular plate in Figure 4.19, with the secondary bifurcation represented by the cyan branch (and its reflected counterpart w.r.t. the plane of the plate).

Therefore, here we avoided to use the previous solution as a guess for the next iteration, while instead we solved the linear problem involving the Frechét derivatives to compute a better guess.

In the next section we will generalize the whole Von Kármán model adding a new parameter which is able to change the bifurcation points.

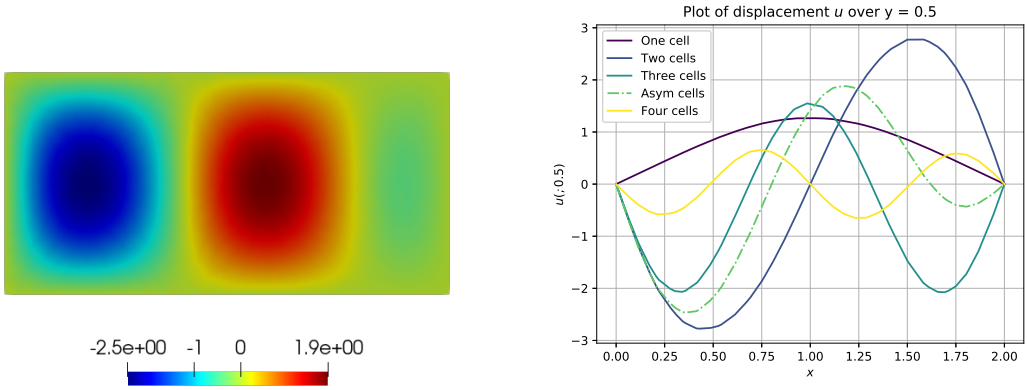


Figure 4.18: High fidelity displacement u for secondary bifurcation and comparison of displacements for all branches, left and right respectively, for $\mu = 65$ and $L = 2$.

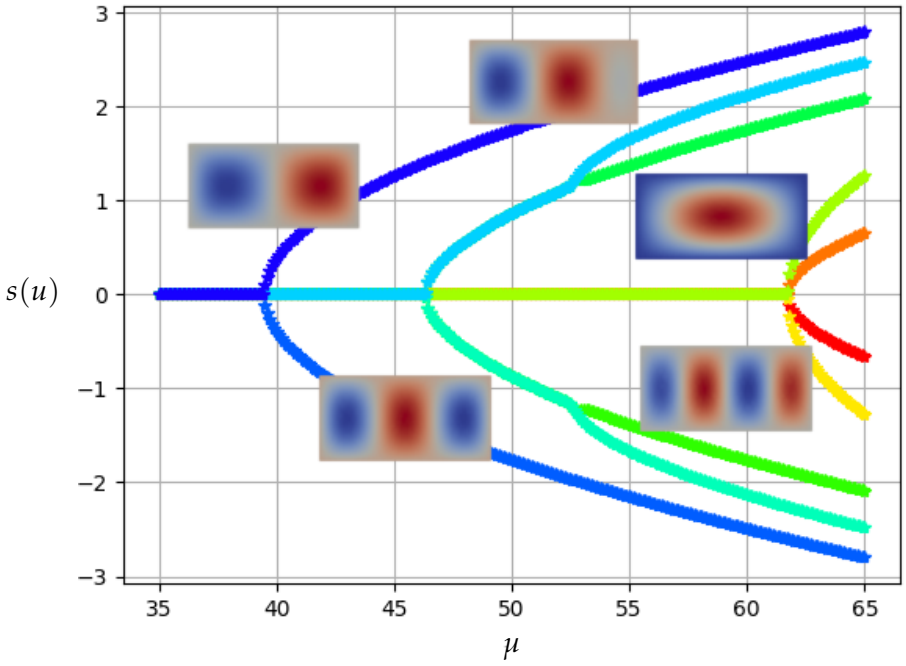


Figure 4.19: High fidelity bifurcation diagram with secondary bifurcation for the rectangular plate with $L = 2$.

4.2.4 Multi-parameter analysis with non-uniform compression

In this section, with motivations coming from some practical applications of buckling plates in naval engineering [62], we want to extend the previous analysis to the multi-parameter case [106] . In particular, we want to consider the same physical phenomenon, while adding a new parameter which leads to a more complicated behaviour, that have to be studied through a 3-D bifurcation diagram. Here, we aim at modelling a compression along the shorter sides of the plate which is no longer uniform on the boundaries. Thus, we parametrized the action of the compression introducing a new parameter ψ. In fact, as we said before, the shape of the compression is determined by the function h appearing in (4.1), thus we can characterize the in-plane load along $\partial\Omega$ generalizing the corresponding term μu_{xx} in (4.18) obtaining the following system

$$\begin{cases} \Delta^2 u + \mu \operatorname{div}(S\nabla u) = 0 & \text{in } \Omega, \\ u = \Delta u = 0 & \text{on } \partial\Omega, \end{cases} \tag{4.22}$$

where $S : \Omega \to \mathbb{R}^{2\times 2}$, $S \neq 0$ is the plane stress tensor field, which is assumed to satisfy the equilibrium equation:

$$\begin{cases} S^T = S & \text{in } \Omega, \\ \operatorname{div}(S) = 0 & \text{in } \Omega. \end{cases} \tag{4.23}$$

It is easy to observe, that we can recover the standard compression problem considered before with the following choice of the stress tensor

$$S = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Furthermore, the compression on the whole boundary, linked to the laplacian operator, is obtained imposing $S = I$, where I is the 2-dimensional identity. Here we are interested in the more general case, in which the uni-axial non-uniform compression is given by the parametrized stress tensor

$$S(\psi) = \begin{bmatrix} \left(1 - \psi\frac{y}{L}\right) & 0 \\ 0 & 0 \end{bmatrix},$$

where $\psi \in [0, 2]$ is the parameter that takes care of the linearly varying in-plane load. Finally, the choice $\psi = 0$ falls back to the aforementioned case.

Therefore, we want to test the strategy developed in the previous sections for a multi-parameter application with $\boldsymbol{\mu} = (\mu, \psi)$, in which two parameters act in the bifurcation phenomenon. As explained in Section 2.2, we are assuming a codimension one bifurcating phenomena, thus the model does not show any other bifurcations while varying the slope ψ, once fixed the compression parameter μ. Therefore, we restrict ourselves to the study of the most physically relevant behaviour, i.e. the evolution with respect to ψ of the first buckling for the generalized system

$$\begin{cases} \Delta^2 u + \mu \operatorname{div}(S(\psi)\nabla u) = [\varphi, u] & \text{in } \Omega, \\ \Delta^2 \varphi = -[u, u] & \text{in } \Omega, \\ u = \Delta u = 0 & \text{in } \partial\Omega, \\ \varphi = \Delta \varphi = 0 & \text{in } \partial\Omega. \end{cases} \tag{4.24}$$

We expect that a change in the slope of the compression could affect the location of the bifurcation points, since the compressive action is no longer uniform along the y-axis. Another consequence of this asymmetry is that the choice of the guesses has to be changed. In fact, the (symmetric) sinusoidal ones can be far from the (asymmetric) buckling states. Moreover, in this case we have no clues about the position of the first buckling for $\psi \neq 0$. For these reasons, following the analysis in Section 4.2.1, we exploited the generalized eigenvalue problem based on (4.24) in which the parameter μ is considered as the eigenvalue. Hence, given $\psi \in [0, 2]$ we want to find the pair (u_ψ, μ_ψ), (again with a little abuse of notation) representing the eigenvector and eigenvalue respectively, which solves

$$\begin{cases} \Delta^2 u_\psi = \mu_\psi \operatorname{div}(S(\psi)\nabla u_\psi) & \text{in } \Omega, \\ u_\psi = \Delta u_\psi = 0 & \text{in } \partial\Omega. \end{cases} \tag{4.25}$$

We remark that the main difference with the analysis in Section 4.2.1 is that now we deal with a parametrized eigenvalue problem, but its computation for different instances of ψ can be embedded in the offline phase. This way we overcome to the first issue about the location of the buckling points for each slope ψ, but we still need to adjust the choice of the guesses. As discussed before, the previous selection of sinusoidal functions was motivated by the analytical expression of the eigenvectors (4.19). Thus, an automatic way to proceed is to utilize the information about the discretized eigenvector associated to the first (smallest) eigenvalue for each fixed ψ. Indeed such eigenvector, opportunely normalized to have $\|\cdot\|_{\mathbb{X}_\mathcal{N}}$-norm equal to one, is only needed in a neighbourhood of the first bifurcation point. In fact, unlike the previous section with secondary bifurcation, after the buckling we can always proceed using the previous iteration as a guess for the successive one.

We focused here on the behaviour of the first buckling for the system (4.24), once fixed the length of the plate as $L = 1$. Some of the representative one cell solutions are presented in Figures 4.20-4.23, from which is evident the progressively lack of symmetry along the horizontal midline with $\{y = 0.5\}$. Moreover, we notice also the completely different behaviour of the displacement configuration with $\psi = 2$, where a two cells like configuration appears. This was indeed an unexpected side effect, that confirms again the complexity of the model. We guess that the compression parameter is too large for the small plate geometry, causing the system to not converge to the one cell configuration.

We can now describe the high fidelity and the reduced approximation of the bifurcation diagram with respect to μ for different choice of the parameter ψ. Within this application to the multi-parameter setting, the Reduced Basis method shows all its

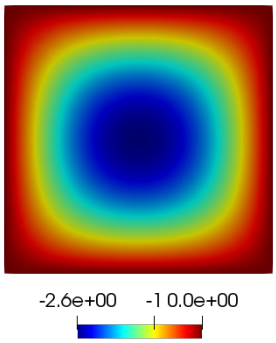


Figure 4.20: High fidelity displacement u at $\mu = 49$ and $\psi = 0$.

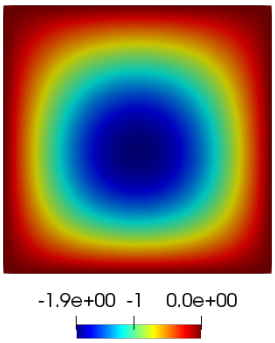


Figure 4.21: High fidelity displacement u at $\mu = 87$ and $\psi = 1$.

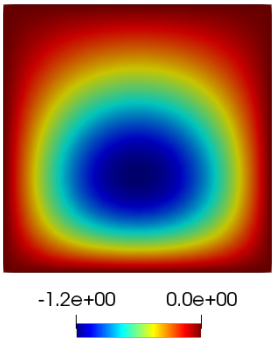


Figure 4.22: High fidelity displacement u at $\mu = 197$ and $\psi = 1.75$.

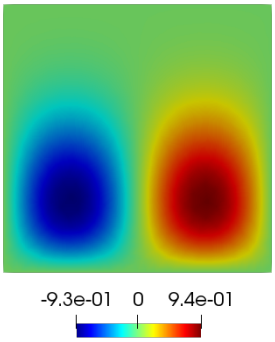


Figure 4.23: High fidelity displacement u at $\mu = 261$ and $\psi = 2$.

potentiality when applied to bifurcating phenomena. Let us review the construction of the reduced manifold in this case. We remark that as before we are considering a global approach, in which we construct a unique ROM for all the buckling modes. The difference here is that instead of compressing the snapshots coming from all the primary branches in e.g. Figure 4.6, we built the ROM from the high fidelity solutions belonging only to the first buckling branches at different values of ψ. In practice, for the FE offline phase, we applied the Algorithm 1 for each fixed $\psi \in [0, 2]$ in the parameter set $\mathcal{P}_K$ defined as the interval centred in the smallest eigenvalue μ_ψ of (4.25), with amplitude equals to 10, using the fundamental information obtained from the eigenvalue problem in (4.25), for both the locations of the buckling and the initial guesses. This choice was taken in order to not consider a too high value of the compression for small values of ψ. Thus, we computed $N_{train} = 15$ snapshots for each of the 5

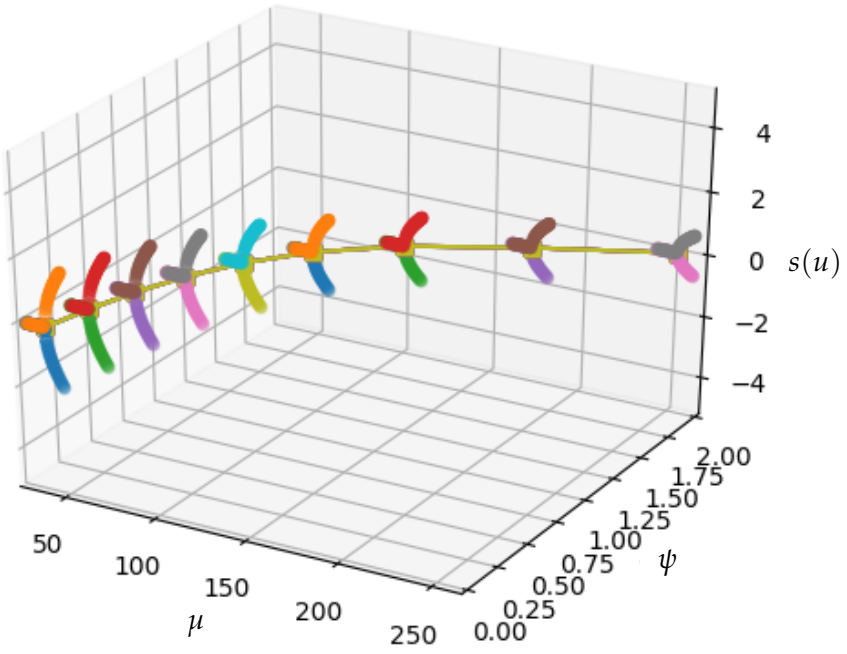


Figure 4.24: 3-D reduced basis bifurcation diagram for the square plate.

equispaced value of the parameter $\psi \in [0,2]$, and we compress them with a unique POD. The resulting basis of dimension $N = 11$, was obtained imposing a tolerance for the POD equals to $\varepsilon_{POD} = 10^{-8}$.

As concerns the online reduced reconstruction of the bifurcation diagram, we chose $\mathcal{P}_K$ with $K = 41$, computing the solution branches for 9 equispaced values of ψ. Thus, we can finally show in the Figure 4.24 the 3-D reduced basis bifurcation plot for the square plate, in which we are describing the first bifurcation point and the post-buckling behaviour, for each one of the nine uniformly sampled $\psi \in [0,2]$. For the sake of clarity, we also show in Figure 4.25 the 2-D projected version of the 3-D plot just presented.

Moreover, as we can see in Figure 4.26 the approximation yields good results from the reduction standpoint. A crucial observation is that we were able to reconstruct the complete branching behaviour at the reduced level, also for values of ψ about which we did not collect any information during the offline phase. This is clearly visible in the right plot of Figure 4.26 for $\psi = 1.75$, where the maximum relative error (always near the buckling point) is of order $1.e-2$.

For this reason, the Algorithm 2 that we implemented for the RB approximation, can be also used in connection to multi-parameter bifurcation phenomena. In fact, collecting information about the buckling for some fixed parameter ψ, we are able to follow efficiently the buckling behaviour for any other ψ in the same range.

As concerns the efficiency, since we are still not using any hyper-reduction tech-

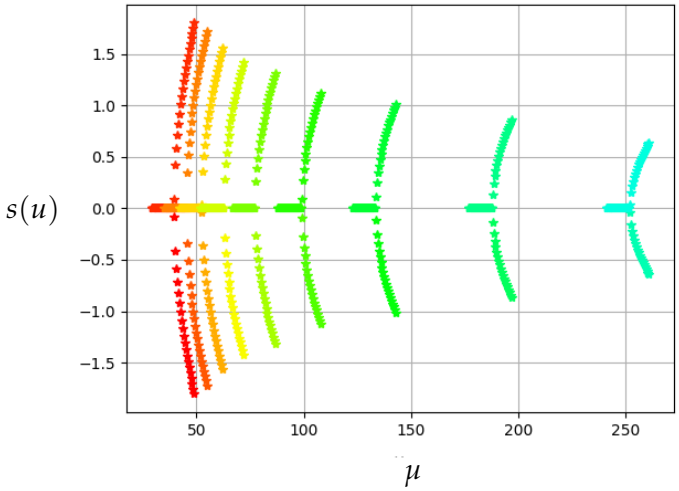


Figure 4.25: 2-D projected bifurcation diagram for the square plate varying $\psi \in [0, 2]$.

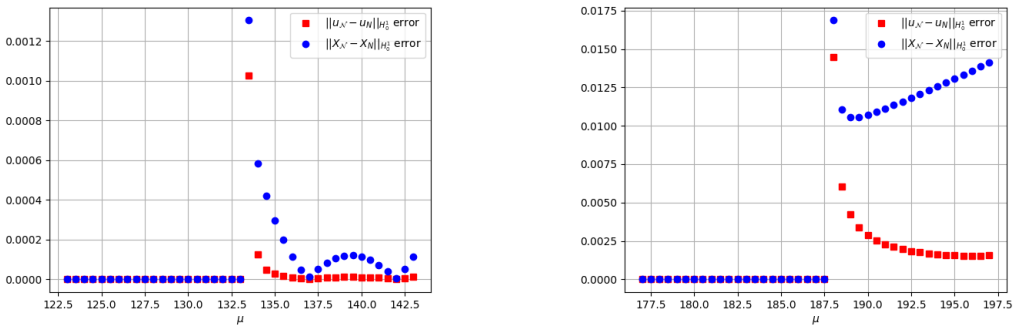


Figure 4.26: Reduced basis errors of the first buckling branch for $\psi = 1.5$ and $\psi = 1.75$, left and right respectively, for $L = 1$.

nique, the speed-up is only 1.58 with $t_{RB} = 320$(s) needed for the RB reconstruction of the full 3-D diagram in contrast to the $t_{FE} = 510$(s) of the FE method. Moreover, despite the chosen output $s(u)$ well reflects the physical event, being linked with the maximum displacement, it involves the high fidelity degrees of freedom $\mathcal{N}$ for the computation of the $\|\cdot\|_\infty$-norm. In a realistic and practical setting one should always choose the output s to be affinely decomposable, in such a way that the projection on the high fidelity space is not needed. As an example in Figure 4.27 we measured the energy of the plate selecting the output as

$$\mathcal{E}(X) = \int_\Omega (\Delta u)^2 - \mu\left(1 - \frac{\psi y}{L}\right) u_x^2 + (\Delta\varphi)^2 d\Omega \tag{4.26}$$

This implementation can be thus used in a real time context, e.g. in a webserver, as we did for the tutorial in ARGOS [126] the Advanced Reduced Groupware Online Sim-

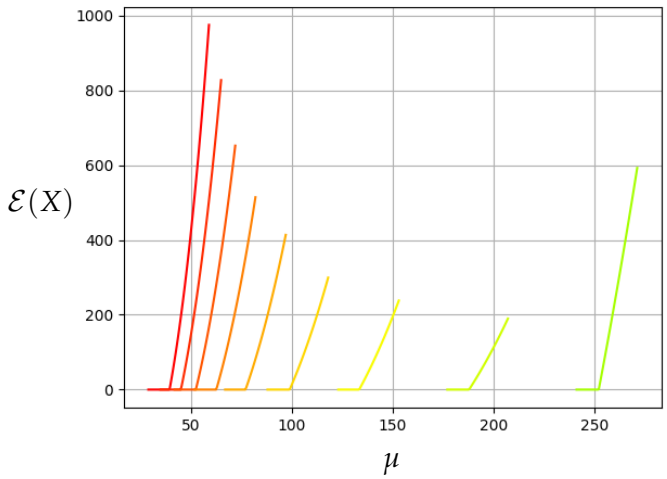


Figure 4.27: 2-D projected bifurcation diagram for the square plate with energy output varying $\psi \in [0, 2]$.

ulation Platform developed by our group. As we can see, the methodology presented in the previous section allowed us to detect in the reduced phase the first bifurcation points with respect to the new parameter ψ that we have introduced. Moreover, we were able to capture correctly the post buckling behaviour, with results validated by the former analysis with $\psi = 0$.

Finally, we highlight that the same computations can be performed also for the rectangular plate with $L = 2$. We can see from the plot in Figure 4.28 the 3-D RB bifurcation diagram with output $s(u)$, and the 2-D projected one with outputs $s(u)$ and $\mathcal{E}(X)$, in Figures 4.29 and 4.30 respectively.

Also here the reduced approximation yields good results, as we can see from Figure 4.31. Furthermore, from the plot on the right we obtain another evidence that through the POD we are able to built a RB basis capable of approximating branches correspondent to values of ψ that we did not collect in the sampling stage.

Finally, we want to remark that here the need for ROMs is even more evident. In fact, considering only the full order problem, we had to solve a huge linear system as many times as the following nested iteration: for each step of the Newton method, for each μ in the parameter domain, for each ψ and finally for each initial guess, if one is interested on multiple branches. For this reason, we conclude the analysis of the Von Kármán model with the results concerning an hyper-reduction approach to recover the efficiency.

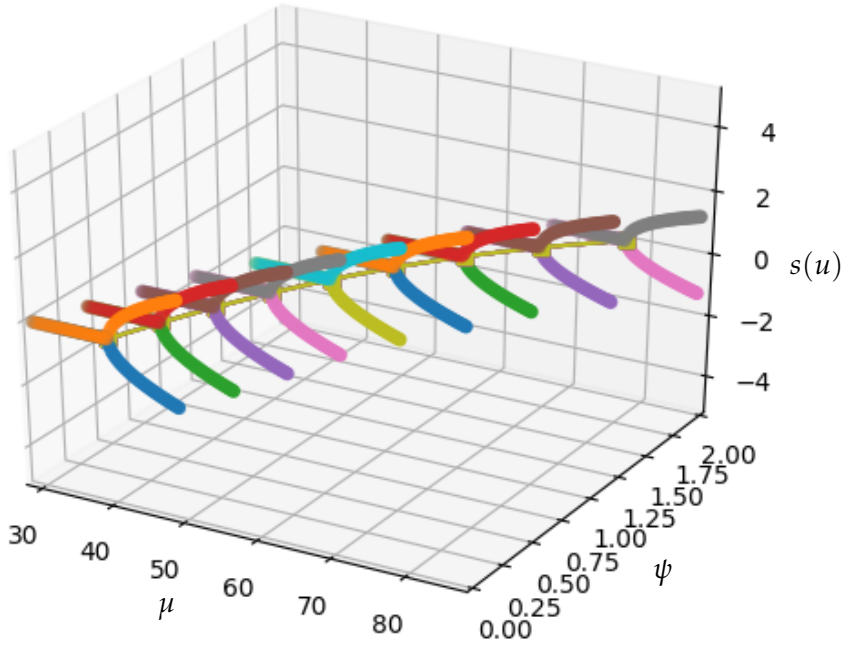


Figure 4.28: 3-D bifurcation reduced basis diagram for the rectangular plate.

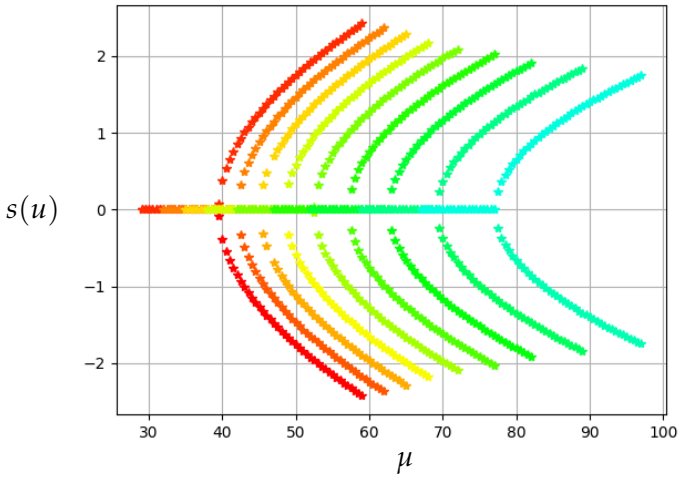


Figure 4.29: 2-D projected bifurcation diagram for the rectangular plate varying $\psi \in [0, 2]$.

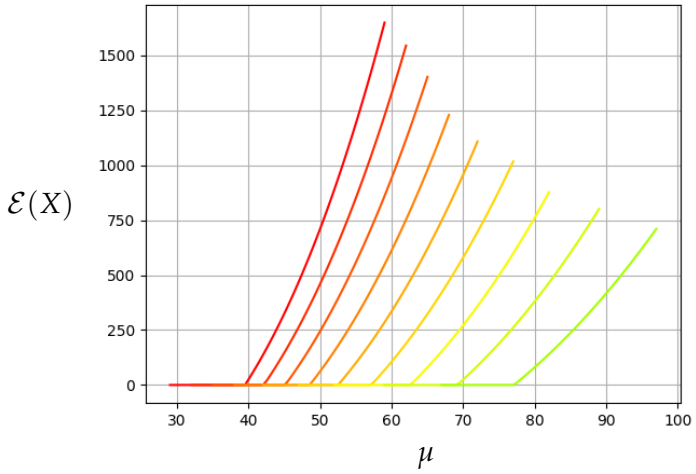


Figure 4.30: 2-D projected bifurcation diagram for the rectangular plate with energy output varying $\psi \in [0, 2]$.

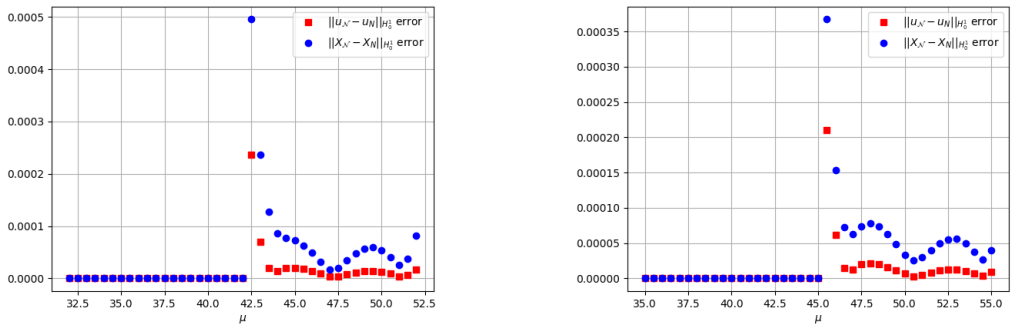


Figure 4.31: Reduced basis errors of the first buckling branch for $\psi = 0.25$ and $\psi = 0.5$, left and right respectively, for $L = 2$.

4.2.5 An empirical interpolation approach

In this last section, we want to present the application of the Empirical Interpolation strategies analysed in Section 3.2.4 to the Von Kármán model with two parameters described in Section 4.2.4. Indeed, we already remarked that the RB methodology performs well in approximating the branching behaviour also for non-sampled values of the parameter ψ, but the computational speed-ups are still low, because of the nonlinearity which does not allow for a $\mathcal{N}$-independent online phase.

Thus, we first consider the square plate with $L = 1$, trying to efficiently recover the 3-D bifurcation diagram in Figure 4.24. We recall that the POD with tolerance $\varepsilon_{POD} = 10^{-8}$ produces a basis for the reduced manifold of dimension $N = 11$. Within the same setting as before, we consider the EIM with Greedy tolerance $\varepsilon_{Gr} = 10^{-7}$ which splits the form into an affine decomposition made up by 8 terms, each one approximated

by almost 15 interpolation basis functions. Conversely, the DEIM approach with the same Greedy tolerance ε_{Gr} produces 3 affine terms each one approximated by almost 25 interpolation basis functions.

We can now show in Figure 4.32 the reduced basis errors, between FE and RB solutions, computed with EIM (left) and DEIM (right). We observe slightly larger error peaks w.r.t. the case of RB with no affine-recovery technique, especially for the DEIM technique, as we can see comparing the plot on the left in Figure 4.26 with the plots in Figure 4.32. These errors are caused by the interpolation error during the hyper-reduction approach in the offline phase. Therefore, EIM is performing better in terms of accuracy, but we also remark that the higher number of terms in the affine decomposition slows down the speed-up with respect to the DEIM, in fact we have $t_{RB,DEIM} = 46$(s) with a speed-up of almost 11, while $t_{RB,EIM} = 135$(s) with a speed-up of almost 4.

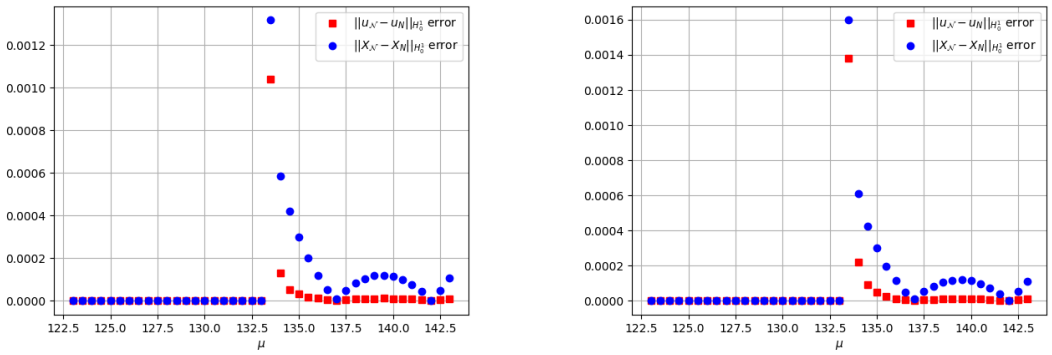


Figure 4.32: Reduced basis errors of the first bifurcating branch computed with EIM (left) and DEIM (right), for $\psi = 1.5$ and $L = 1$.

We can draw similar conclusions when considering the rectangular plate with $L = 2$. We recall that here the POD with tolerance $\varepsilon_{POD} = 10^{-8}$ produces a basis for the reduced manifold of dimension $N = 10$. As expected, the two affine-recovery techniques provide similar affine decompositions with respect to the square plate case. Indeed, keeping the same reduced setting, we chose the EIM (DEIM) with Greedy tolerance as $\varepsilon_{Gr} = 10^{-7}$ which splits the form into an affine decomposition made up by 8 (3) terms, each one approximated by ~ 15 (~ 25) interpolation basis functions.

We can thus end this chapter showing in Figure 4.33 the reduced basis errors, computed with EIM (left) and DEIM (right). Once again, slightly larger error peaks occur, especially for the DEIM technique, as we can see comparing the plot on the right in Figure 4.31 with the plots in Figure 4.33. Finally, the same computational savings were observed also for the rectangular plate.

We have now completely characterized the bifurcation phenomenon occurring

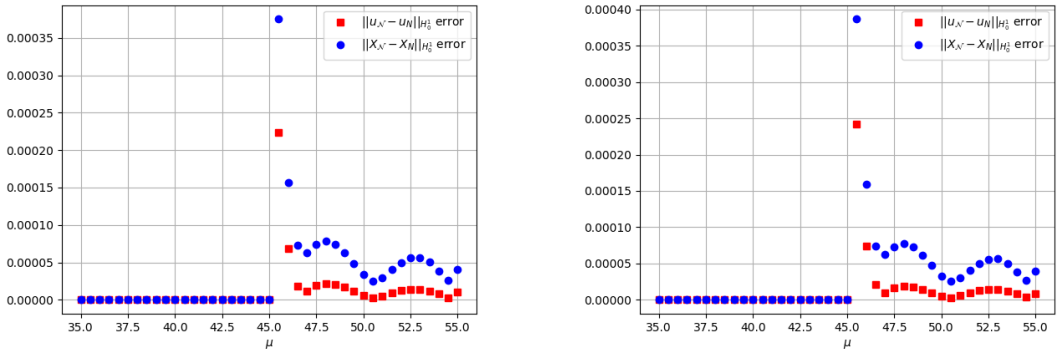


Figure 4.33: Reduced basis errors of the first bifurcating branch computed with EIM (left) and DEIM (right), for $\psi = 0.5$ and $L = 2$.

when dealing with the structural investigation of buckling plates. Remarkably, the reduced order algorithm we have developed was able to recover the bifurcation diagrams in all the physical settings we have considered. Supported by these results, we have taken a step towards more complex models. In particular, in the next chapter, we will discuss the bifurcating behaviour of the Gross-Pitaevskii model in Quantum Mechanics.

Chapter 5

Gross-Pitaevskii equations in Bose-Einstein condensates

In this chapter, we consider the Gross-Pitaevskii equation in Quantum Mechanics [123]. Often referred to as a nonlinear Schrödinger equation, the Gross-Pitaevskii equation models certain classes of Bose-Einstein condensates (BECs), a special state of matter formed by identical bosons at ultra-low temperatures. It is well known that the solutions of the Gross-Pitaevskii equation with a parabolic trap in two dimensions exhibits a rich bifurcating behavior [104, 105, 35], which includes symmetry-breaking bifurcations and vortex-bearing states when a (sufficiently strong) rotational angular momentum term is added [60]. The bifurcating behavior becomes even richer for the two-dimensional coupled Gross-Pitaevskii equations [34]. However, for simplicity we stick to the simple Gross-Pitaevskii equation and present a one parameter study (the chemical potential being the only varying parameter) and a two parameter case (varying chemical potential and the normalized trap strength). For both cases, we will describe how our approach leads to an efficient reconstruction of the bifurcation diagram, while providing good approximation accuracy. This part is mainly based on the work done in collaboration with A. Quaini [119].

5.1 Gross-Pitaevskii model

A BEC is a special state of matter formed by an unlimited number of bosons that "condense" into the same energy state at low temperatures. It is formed by cooling a gas of extremely low density, about one-hundred-thousandth the density of normal air, to ultra-low temperatures (close to absolute zero).

A quantum system is the environment to be studied in terms of wave-particle duality (i.e., all particles exhibit a wave nature and viceversa) and it involves the wave-function and its constituents, such as the momentum and wavelength. The Gross-Pitaevskii equation describes the ground state of a quantum system of identical bosons using two simplifications: the Hartree-Fock approximation and the pseudo-potential interaction model. In the Hartree-Fock approximation, the total wave-function Φ_{tot} of

a system of N_B bosons is taken as a product of single-particle functions Φ

$$\Phi_{tot}(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{N_B}) = \prod_{i=1}^{N_B} \Phi(\mathbf{r}_i),$$

where $\mathbf{r}_i$ is the coordinate of the i-th boson. If the single-particle wave-function satisfies the Gross-Pitaevskii equation, the total wave-function minimizes the expectation value (i.e., the probabilistic expected value of the result of an experiment) of the pseudopotential model Hamiltonian under normalization condition

$$N_B = \int_\Omega \rho \, d\boldsymbol{r}, \quad \rho = |\Phi|^2, \tag{5.1}$$

where $\Omega \subseteq \mathbb{R}^2$ is the domain under consideration and ρ is interpreted as the *particle density*. The Gross-Pitaevskii equation reads: Find the single-particle wave-function $\Phi(\mathbf{r}, t) : \overline{\Omega \times \mathbb{R}^+} \to \mathbb{C}$ such that

$$i\partial_t \Phi = -\frac{1}{2}\Delta\Phi + |\Phi|^2\Phi + W(r)\Phi \quad \text{in } \Omega, \tag{5.2}$$

where i is the imaginary unit, $r = |\mathbf{r}| = \sqrt{x^2 + y^2}$ is the radial coordinate, and $W(r) = \frac{1}{2}\tau^2 r^2$ is the external potential, with τ being the normalized *trap strength*, i.e. the ratio of trappings along and transverse to the plane. For the analysis of the model we will restrict ourselves to a fixed trap strength equals to $\tau = 0.2$. Notice that we consider a single well potential of parabolic type. Equation (5.2) is similar in form to the Ginzburg–Landau equation and is sometimes referred to as a nonlinear Schrödinger equation. Obviously, equation (5.2) needs to be supplemented with suitable boundary conditions.

The construction of the steady solution is based on the ansatz:

$$\Phi(\mathbf{r}, t) = \phi(r) \exp(-i\mu t), \quad \phi(r) : \overline{\Omega} \to \mathbb{C}, \tag{5.3}$$

where μ is the so called *chemical potential*, which can be seen as a measure of the density at the center of the trap and has to satisfy $\mu \geq \tau$. By plugging (5.3) into (5.2), we obtain nonlinear problem

$$G(\phi; \mu) := -\frac{1}{2}\Delta\phi + |\phi|^2\phi + W(r)\phi - \mu\phi = 0. \tag{5.4}$$

It is well known that the solutions of the one-dimensional version of problem (5.4) exhibit a bifurcating behavior [91, 88, 1, 46], which is not particularly rich though. The bifurcations occurring in the two-dimensional problem (5.4) are far more interesting [104, 105, 35]. Indeed, several secondary bifurcations appear, which include symmetry-breaking bifurcations and vortex-bearing states [60]. As before, we are interested in the complex behaviour which originates from the Gross-Pitaevskii model. We will analyse it through its bifurcation diagram, which plots some characteristic quantities

of the solution, such as the number of bosons N_B in the BEC (5.1) or a different norm of the particle density ρ defined in (5.1), as a function of the chemical potential μ.

When $N_B \to 0$, the nonlinearity of the problem becomes irrelevant and the states bifurcate from the respective linear limit. Starting from this low-density context, we are interested in exploring the solution modes for larger values of μ, which make the problem strongly nonlinear. Since an arbitrary potential can be approximated as a harmonic potential at the vicinity of a stable equilibrium point, when $N_B \to 0$ we can decompose the linear eigenfunction $\phi_{m,n}$ in Cartesian form as being proportional to

$$|m,n\rangle \;:=\; \phi_{m,n} \sim H_m(\sqrt{\tau}x)H_n(\sqrt{\tau}y)e^{-\frac{r^2}{2}\tau}, \tag{5.5}$$

where H_j is the Hermite polynomial with j being the associated quantum number of the harmonic oscillator. The critical value of the eigenvalue corresponding to linear eigenfunction $\phi_{m,n}$ is $\mu^* = \mu_{m,n} := (m+n+1)\tau$. Notice that, thanks to the chosen notation, we are highlighting again the connection between the state equation and its spectral properties, since it has important consequences on the bifurcation behaviour of the model itself. Indeed here, as we will see, the characteristic eigenvalue parameter, i.e. the eigenvalue responsible for the bifurcation, is exactly the chemical potential. Thus, given an initial energy μ at the linear limit, we increase the chemical potential (and therefore the number of atoms N_B) in order to approach to the strongly nonlinear regime, that can lead to the discovery of new states originating also from secondary bifurcations.

For the numerical characterization of the stability (and possible classification of the instability) for each state, we refer to [35].

5.1.1 Weak formulation

Given the Gross-Pitaevskii model, expressed through the PDE (5.4), we prepare now the setting to numerically study its branching behaviour. For this reason, here we consider the variational framework to obtain its weak formulation, which is the first step to apply the Algorithm 1 developed in Chapter 3 for its numerical investigation. So let us consider $\mu \in \mathcal{P} \subset \mathbb{R}$ the chemical parameter and $\Omega \subset \mathbb{R}^2$ the domain in which we will approximate the solution.

We recall that the solution ϕ to equation (5.4) is a complex function, thus we will denoting its *real* and *imaginary* parts respectively as φ and ψ, which we will treat as the two components of the solution. Having represented the model with the parametrized mapping $G : \mathbb{X} \times \mathcal{P} \to \mathbb{X}'$ with $\mathbb{X} = \left(H_0^1(\Omega)\right)^2$, its weak formulation reads: given $\mu \in \mathcal{P}$, find $X(\mu) := (\varphi(\mu), \psi(\mu)) \in \mathbb{X}$ such that (omitting the μ-dependency)

$$g(X, Y; \mu) := a(X, Y) + n(X, Y) + b(X, Y) - d(X, Y; \mu) = 0, \quad \forall\, Y \in \mathbb{X}, \tag{5.6}$$

where we have introduced the following variational forms

$$a(X,Y) = \frac{1}{2}\int_\Omega \nabla X \cdot \nabla Y d\Omega, \quad b(X,Y) = \frac{1}{2}\tau \int_\Omega |\boldsymbol{r}|^2 X \cdot Y d\Omega,$$
$$d(X,Y;\mu) = \mu \int_\Omega X \cdot Y d\Omega, \quad n(X,Y) = \int_\Omega |X|^2 X \cdot Y d\Omega.$$

To deal with the nonlinear terms in (5.6), we compute the Frechét derivative of the parametrized variational form g, which can be expressed as

$$\mathrm{d}g[Z](X,Y;\mu) = a(X,Y) + b(X,Y) - d(X,Y;\mu) + c(X,Z,Y) \quad \forall\, Y \in \mathbb{X}, \tag{5.7}$$

where the term

$$c(X,Z,Y) = \int_\Omega [2(X\cdot Z)Z + |Z|^2 X] \cdot Y d\Omega \tag{5.8}$$

represents the linearization of the nonlinear term n at a point $Z \in \mathbb{X}$.

5.2 Numerical approximation

The projection of the weak formulation (5.6) in the finite dimensional space $\mathbb{X}_\mathcal{N} \subset \mathbb{X}$ of dimension $\mathcal{N}$ is needed for the numerical investigation of the branching behaviour. Thus, we defined the FE space $\mathbb{X}_\mathcal{N} = \mathbb{X}_h \cap \mathbb{K}_r^0$ as in (4.11), consisting of polynomials of degree r on the triangulation $\mathcal{T}_h$ of Ω, which vanish on the boundary.

Then, generic k-th iteration of the Galerkin FE method combined with the Newton method reads as: given $\mu \in \mathcal{P}$ and an initial guess $X_\mathcal{N}^0 = (\varphi_\mathcal{N}^0, \psi_\mathcal{N}^0) \in \mathbb{X}_\mathcal{N}$ for $k = 0, 1, \ldots$, until convergence we seek $\delta X_\mathcal{N} = (\delta\varphi_\mathcal{N}, \delta\psi_\mathcal{N}) \in \mathbb{X}_\mathcal{N}$, such that

$$\begin{aligned} &a(\delta X_\mathcal{N}, Y_\mathcal{N}) + b(\delta X_\mathcal{N}, Y_\mathcal{N}) - d(\delta X_\mathcal{N}, Y_\mathcal{N};\mu) + c(\delta X_\mathcal{N}, X_\mathcal{N}^k, Y_\mathcal{N}) = \\ &a(X_\mathcal{N}^k, Y_\mathcal{N}) + b(X_\mathcal{N}^k, Y_\mathcal{N}) - d(X_\mathcal{N}^k, Y_\mathcal{N};\mu) + n(X_\mathcal{N}, Y_\mathcal{N}) \quad \forall\, Y_\mathcal{N} \in \mathbb{X}_\mathcal{N}, \end{aligned} \tag{5.9}$$

where again for ease of notation, we have not specified that the solution $X_\mathcal{N}$ depends on parameter μ.

As concerns the corresponding matrix formulation, we start by introducing the finite element discretization matrices

$$\begin{aligned} &(\mathsf{A}_\mathcal{N})_{ij} = a(E^j, E^i)\ , \quad (\mathsf{B}_\mathcal{N})_{ij} = b(E^j, E^i)\ , \\ &(\mathsf{C}_\mathcal{N})_{ij} = c(E^j, X_\mathcal{N}^k, E^i)\ , \quad (\mathsf{D}_\mathcal{N}(\mu))_{ij} = c(E^j, E^i;\mu)\ . \end{aligned} \tag{5.10}$$

Therefore, the Jacobian matrix $\mathsf{J}_\mathcal{N}(\mathsf{X}_\mathcal{N}^k(\mu);\mu)$ of dimension $\mathbb{R}^{\mathcal{N}\times\mathcal{N}}$ can be written as

$$\mathsf{J}_\mathcal{N}(\mathsf{X}_\mathcal{N}^k(\mu);\mu) = \mathsf{A}_\mathcal{N} + \mathsf{B}_\mathcal{N} - \mathsf{D}_\mathcal{N}(\mu) + \mathsf{C}_\mathcal{N}.$$

Once again, we remark that the high fidelity Jacobian matrix has to be assembled at each iteration of the Newton method and for each new instance of the parameter that controls the chemical potential. Moreover, also in this case, the very rich bifurcating

phenomenon, which characterizes the model, makes the computational complexity heavier and a reduction strategy is needed. Indeed, we relied on the ROM technique presented in Algorithm 2, which is used for the online approximation of each branch of the Gross-Pitaevskii equation.

Given $\mu \in \mathcal{P}$ and an initial guess $X_N^0 \in \mathbb{X}_N$ for $k = 0, 1, \dots$ until convergence we seek $\delta X_N = (\delta\varphi_N, \delta\psi_N) \in \mathbb{X}_N$ such that

$$\begin{aligned} a(\delta X_N, Y_N) + b(\delta X_N, Y_N) - d(\delta X_N, Y_N; \mu) + c(\delta X_N, X_N^k, Y_N) = \\ a(X_N^k, Y_N) + b(X_N^k, Y_N) - d(X_N^k, Y_N; \mu) + n(X_N^k, Y_N) \quad \forall\, Y_N \in \mathbb{X}_N, \end{aligned} \tag{5.11}$$

and then we set $X_N^{k+1} = X_N^k - \delta X_N$.

Finally, the reduced Jacobian $\mathsf{J}_N(\mathsf{X}_N^k(\mu); \mu) \in \mathbb{R}^{N \times N}$ can be written as

$$\mathsf{J}_N(\mathsf{X}_N^k(\mu); \mu) = \mathsf{A}_N + \mathsf{B}_N - \mathsf{D}_N(\mu) + \mathsf{C}_N \tag{5.12}$$

where

$$\mathsf{A}_N = \mathsf{V}^T \mathsf{A}_{\mathcal{N}} \mathsf{V}\,, \quad \mathsf{B}_N = \mathsf{V}^T \mathsf{B}_{\mathcal{N}} \mathsf{V}\,,$$

$$\mathsf{D}_N(\mu) = \mathsf{V}^T \mathsf{D}_{\mathcal{N}}(\mu) \mathsf{V}\,, \quad \mathsf{C}_N = \sum_{n=1}^{N} X_N^{(n)} \mathsf{V}^T \mathsf{C}_{\mathcal{N}}(\Sigma^n) \mathsf{V}$$

are the reduced matrices written in terms of finite element matrices (5.10) and transformation matrix (3.18).

The time saving promised by the online-offline strategy, usually enabled by the affine decomposition, can not be obtained due to the nonlinearity, that as we will see later on, it will force us to adopt the Empirical Interpolation techniques described in Section 3.2.4 in order to obtain substantial savings of the computational time during the online phase. Indeed, the reduced matrix C_N introduced above depends on μ through the solution computed at each step of Newton method. Let us now present the high fidelity benchmark, validating our algorithm with the results in literature.

5.2.1 Validation of the high fidelity approximation

In order to validate the Full Order Method for the approximation of the Gross-Pitaevskii equation, by means of the Finite Element method, as described in Section 5.1.1, we consider a setting proposed in [35]. We approximate the solution to equation (5.4) in domain $\Omega = (-12, 12)^2$, with homogeneous Dirichlet boundary conditions on the entire boundary of Ω. The choice of a square domain with homogeneous BCs serves to guarantee that the solution is not influenced by the boundary of the domain itself, since the support of the solutions is focused at the center.

We recall that we set $\tau = 0.2$, while the chemical potential varies in $\mathcal{P} = [0, 1.2]$. For the space discretization, we use $\mathbb{P}_2$ finite elements and a refined mesh in the ball of radius 5 centred in the origin, with 6889 elements, which produces a finite element space dimension equals to $\mathcal{N} = 27908$.

We could investigate the model through its spectral properties, as we did before, since the existing connection with the eigenproblem (5.5). While this gives us information about the location of the bifurcation points, a more extensive use of the eigenproblem to also provide the initial guesses is not possible here. Indeed, the numerical eigenvalue analysis of such problems is a difficult task [53, 28, 29], preventing a straightforward application of the methodology utilized before. The main reason is that, as studied in [21, 34], the Gross-Pitaevskii equation (5.4) has an infinite symmetry group that acts on the space of solutions. Therefore, any rotation of the solution is still an admissible solution, i.e. for any $\theta \in \mathbb{R}$

$$\text{if} \quad G(\phi;\mu) = 0 \qquad \text{then} \quad G(e^{i\theta}\phi;\mu) = 0. \tag{5.13}$$

We remark that a similar situation occurred for the Von Kármán plate equation in the previous chapter, where the symmetry taken into account was the simple reflection through the plane of the plate. This of course is a computational issue, but can be tackled with the implementation of properly chosen continuation methods in conjunction to the deflation method [56, 35, 122]. We will present an application of this in Chapter 8.

Thus, having utilized the eigenvalues information to fix the parameter space $\mathcal{P}$ where to perform the investigation, we choose its discretization $\mathcal{P}_K$ for the reconstruction of the bifurcation diagram. As we anticipated, the output scalar measures s will be in this case the number of boson N_B and the $\|\cdot\|_\infty$-norm of the density ρ.

The simple continuation method was applied within the framework described in the Algorithm 1. We chose $K = 960$ equispaced points with continuation step $\Delta\mu = 1.25 \cdot 10^{-3}$, in fact to properly follow the branches with the simple continuation technique a huge number of solutions have to be computed. As concerns the initial guess, here we took advantage of analytical expressions for the linear limit in (5.5). In particular, we assign an initial guess proportional to the product of Hermite polynomials H_m and H_n, where m and n chosen according to the value of the critical point $\mu^* = (m + n + 1)\tau$. We thus used different Hermite functions to follow different branches, always keeping in mind the location of their occurrence by looking at the corresponding eigenvalues, i.e. the bifurcation points. This approach worked well due to the approximation near the linear limit for low value of the chemical potential μ.

We start by showing in Figures 5.1 and 5.2 the high fidelity bifurcation diagrams in the μ-$N_{B\mathcal{N}}$ plane and in the μ-$\|\rho_{\mathcal{N}}\|_\infty$ plane, respectively. These diagrams show the first three bifurcation points and the corresponding six branches, showing the non-uniqueness of the solution with respect to the chemical parameter μ.

As μ is increased, the sequence of events is as follows. The ground state $|0,0\rangle$ is the system simplest state. Its linear eigenfunction $\phi_{0,0}$ has corresponding eigenvalue $\mu = \tau$. The ground state is generically stable [87], thus no further bifurcations occur from this state [87]. As expected, a unique solution branch departs from $\mu = \tau$ in Figures 5.1 and 5.2. A representative density function for this branch is shown in Figure 5.3a. We can easily see that no further bifurcation originates for $\tau \leq \mu < 2\tau$. The first interesting events in terms of bifurcation analysis occur for $\mu = 2\tau$, with

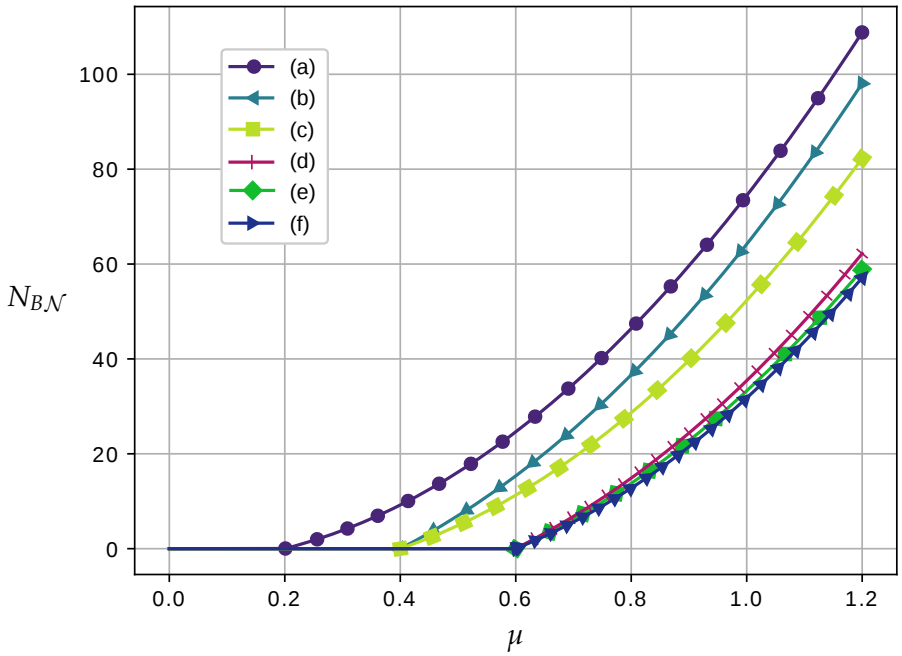


Figure 5.1: High fidelity bifurcation diagram with the number of bosons $N_{B\mathcal{N}}$ plotted against the chemical potential μ. The labels (a)-(f) are referred to solutions in Figure 5.3.

$n+m=1$: two branches, associated to $|0,1\rangle$ and $|1,0\rangle$, bifurcate from point $(2\tau,0)$ in the μ-$N_{B\mathcal{N}}$ and μ-$\|\rho_{\mathcal{N}}\|_\infty$ planes [104, 47]. Indeed, from point $(2\tau,0)$ in Figures 5.1 and 5.2 we observe the two expected branches. Representative density functions for these two branches are reported in Figures 5.3b and 5.3c. The next, more complicated, case of bifurcations emanates from point $(3\tau,0)$, with $n+m=2$. In Figures 5.1 and 5.2, we see that three branches depart from this point, associated to $|1,1\rangle$, $|0,2\rangle$, and $|2,0\rangle$. The corresponding representative densities are shown in Figures 5.3f, 5.3e, and 5.3d. Finally, all the points without marker in Figures 5.1 and 5.2 correspond to the non-physical solution $\phi=0$ that exists since there are no external forces in equation (5.4).

Figure 5.3 displays the density functions associated to $\mu=1.2$ and all the six solution branches in Figures 5.1 and 5.2. We observe the richness of density patterns in order of decreasing N_B. In particular, we see the ground state $|0,0\rangle$ in Figure 5.3a, the single charge vortex $|0,1\rangle$ in Figure 5.3b, the 1-dark soliton stripe $|1,0\rangle$ in Figure 5.3c, the dark soliton cross $|1,1\rangle$ in Figure 5.3f, the ring dark soliton $|0,2\rangle$ in Figure 5.3e, the 2-dark soliton stripe $|2,0\rangle$ in Figure 5.3d. Notice that the six branches in Figures 5.1 and 5.2 are related to the first three eigenvalues. For example, the second bifurcation stems from a double eigenvalue and thus we have two branches. This is again an example of multiple bifurcation. The stability property of these branches are different

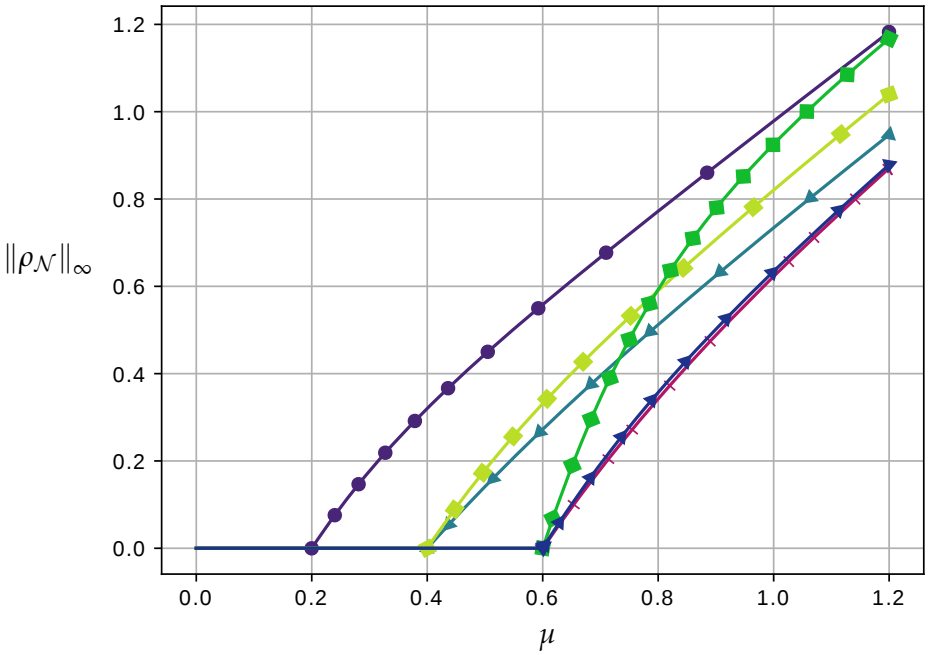
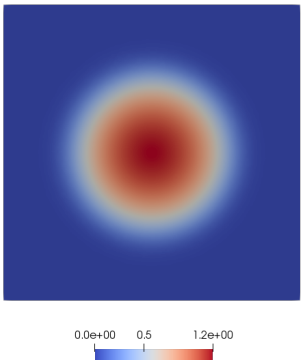


Figure 5.2: High fidelity bifurcation diagram with the infinity norm of density $\rho_{\mathcal{N}}$ plotted against the chemical potential μ. The labels (a)-(f) are referred to solutions in Figure 5.3.

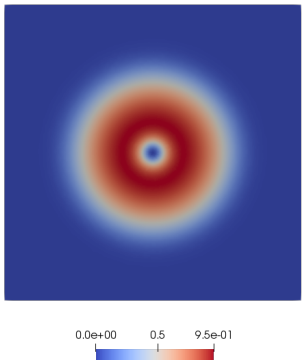
for each case, i.e the single charge vortex is always stable while the 1-dark soliton stripe is subject to multiple secondary bifurcations. These properties can be easily studied using the standard techniques presented in Section 3.1.3, as did in [35], but also due to the complexity mentioned before, this analysis is beyond our scope. The results in Figures 5.1, 5.2 and 5.3 are in excellent agreement with the results reported in [35], indicating that the FE methodology together with the choice of the mesh that we used provide a good starting point for the implementation of the reduction strategies.

The overall simulation time required to complete the diagrams in Figures 5.1 and 5.2 is roughly $t_{HF} = 96$(min).

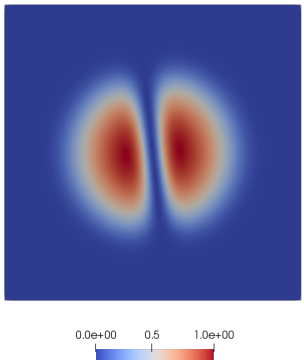
We can finally move to the reduced order approximation in which we aim at reducing the computational time to built the bifurcation diagrams.



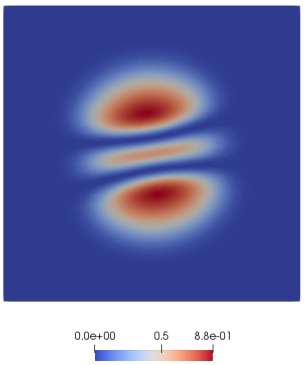
(a) ground state $|0,0\rangle$



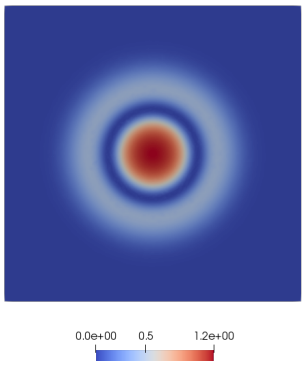
(b) single charge vortex $|0,1\rangle$



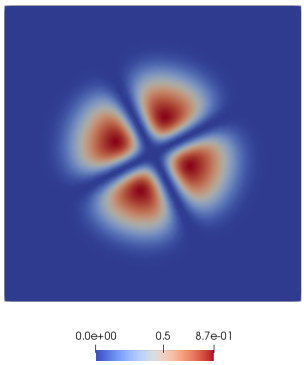
(c) 1-dark soliton stripe $|1,0\rangle$



(d) 2-dark soliton stripe $|1,1\rangle$



(e) ring dark soliton $|0,2\rangle$



(f) dark soliton cross $|2,0\rangle$

Figure 5.3: Density functions computed with the high fidelity method at $\mu = 1.2$. Each plot is associated to one of the six solution branches in Figures 5.1 and 5.2. From (a) to (f) the number of bosons N_B is decreasing.

5.2.2 ROM approach to multiple bifurcations

In this section, we present the results obtained with the Reduced Order Method embedded in Algorithm 2, as described in Section 3.2.5, and compare them with the FOM results reported in the previous section. For this first test case, the only parameter that varies is the chemical potential μ in interval $\mathcal{P} = [0, 1.2]$.

Concerning the construction of the reduced manifold, we employed a training set for the POD with cardinality $N_{train} = 160$ for each one of the six branches. Setting the POD tolerance to $\varepsilon_{POD} = 10^{-9}$, we obtain a global basis of dimension $N = 51$. During the online phase, we reconstruct the reduced bifurcation diagram for all the 961 equally spaced points in $\mathcal{P} = [0, 1.2]$ used in the high-fidelity bifurcation diagram shown in Figures 5.1 and 5.2. Such points correspond to continuation step $\Delta\mu = 1.25 \cdot 10^{-3}$.

Figure 5.4 shows reduced order errors

$$E_{N_B} = |N_{B\mathcal{N}} - N_{BN}| \quad \text{and} \quad E_\rho = |\|\rho_\mathcal{N}\|_\infty - \|\rho_N\|_\infty|, \tag{5.14}$$

i.e. the difference in absolute value between the branches of the bifurcation diagram computed with FOM ($N_{B\mathcal{N}}$ and $\|\rho_\mathcal{N}\|_\infty$) and ROM ($N_{BN}$ and $\|\rho_N\|_\infty$) in the μ-N_B plane (top) and in the μ-$\|\rho\|_\infty$ plane (bottom). In Figure 5.4, we see that the largest peaks are associated to the $|1,1\rangle$ branch at $\mu = 0.6$. As before, it is expected to have larger errors at the bifurcation points where differentiability with respect to the parameter μ is lost. We also infer that the errors are largest at $\mu = 0.6$ due to the more complicated solution structure (compare Figure 5.3f to the other panels in Figure 5.3). From Figure 5.4 (bottom), we see the largest error is of the order of 10^{-4} also for the infinity norm of the density. However, we observe larger errors over interval $[0.2, 1.2]$ for μ, as opposed to localized at $\mu = 0.6$ as in Figure 5.4 (top).

As further evidence of the accuracy of our ROM approach, we plot in Figure 5.5 the difference between $X_\mathcal{N}$ (FOM solution) and X_N (ROM solution) in the L^2 and H_0^1 norms. Recall that $X_\mathcal{N}$ and X_N consists of both real and imaginary part of the computed solution of the Gross-Pitaevskii equation. As expected, for all the branches the largest errors in both norms occur at the point where each branch departs from the horizontal axis, i.e. $\mu = 0.2$ for branch $|0,0\rangle$, $\mu = 0.4$ for branches $|1,0\rangle$ and $|0,1\rangle$, and $\mu = 0.6$ for branches $|1,1\rangle$, $|0,2\rangle$, and $|2,0\rangle$. In addition, just like in Figure 5.4 the largest errors are associated with branch $|1,1\rangle$.

Figure 5.6 reports the difference between the density function ρ computed with FOM and ROM in the L^2 and H_0^1 norms. In the case of the density, the largest errors for each branch occur for μ larger than the critical value where the branch departs from the horizontal axis. For a better understanding of how the ROM density compares with the FOM density, Figure 5.7 displays the difference $\rho_\mathcal{N} - \rho_N$ at $\mu = 1.2$. We observe larger errors for the 1- and 2-dark soliton stripe and the dark soliton cross, i.e. for the solutions that do not have central symmetry.

Figures 5.4–5.7 show the ability of our ROM approach to accurately reconstruct bifurcation diagrams as parameter μ varies. Clearly, it makes sense to set up the ROM machinery if there is a substantial gain in terms of computational time. Because of the nonlinearity in the Gross-Pitaevskii equation, which makes the computations in the online phase dependent on the number of FEM degrees of freedom, the computational speed-up enabled by our ROM approach is only 1.1: it took $t_{RB} = 86$(min) to generate the data needed for the bifurcation diagrams, using continuation step $\Delta\mu = 1.25 \cdot 10^{-3}$, accounting only for the online phase computations. Recall that it takes 96 minutes with FOM. These computational savings are not satisfactory, especially if we were to include the cost for the offline phase. Before introducing the affine-recovery techniques to drastically improve computational efficiency, we present a two-parameter study.

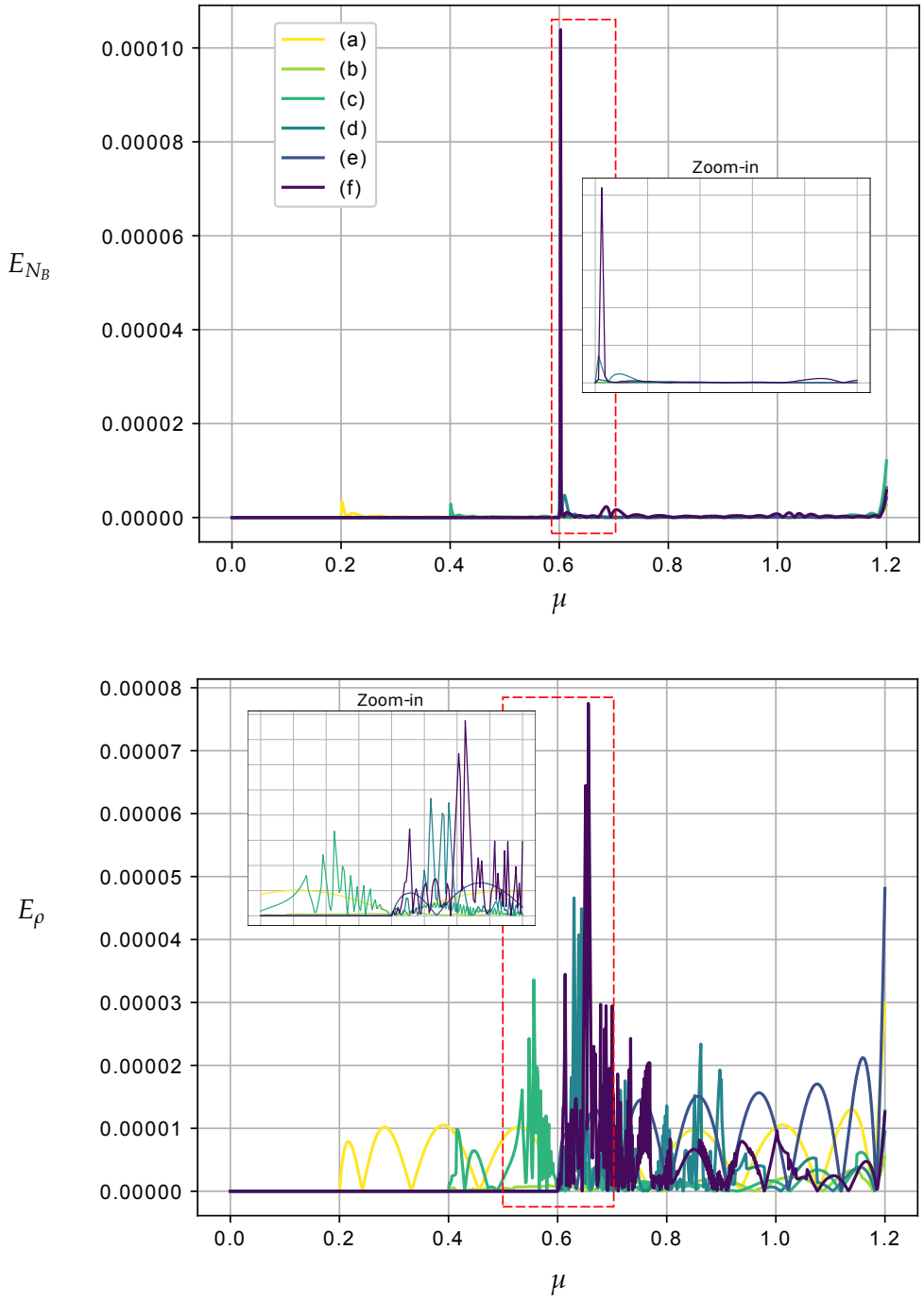
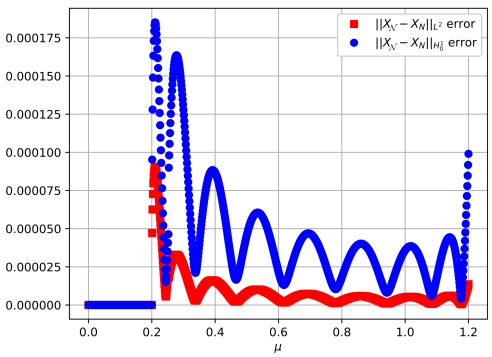
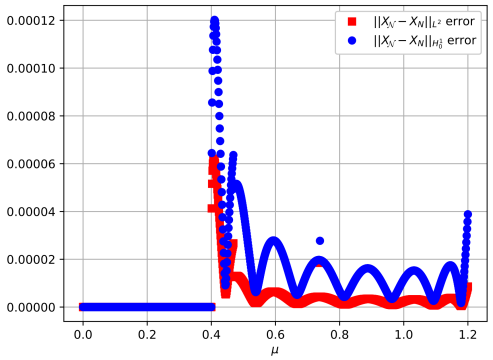


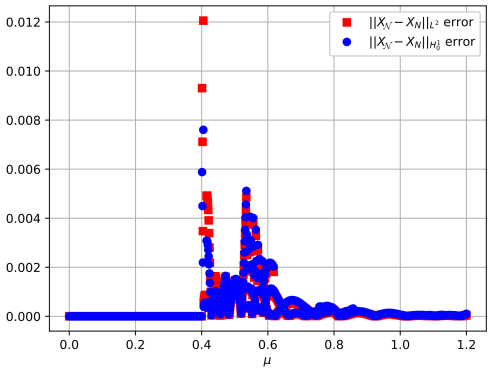
Figure 5.4: Difference in absolute value between the branches of the bifurcation diagram computed with FOM and ROM in the μ-N_B plane (top) and in the μ-$\|\rho\|_\infty$ plane (bottom), i.e. reduced order errors E_{N_B} (top) and E_ρ (bottom) defined in (5.14). The labels (a)-(f) refer to the solutions reported in Figure 5.3.



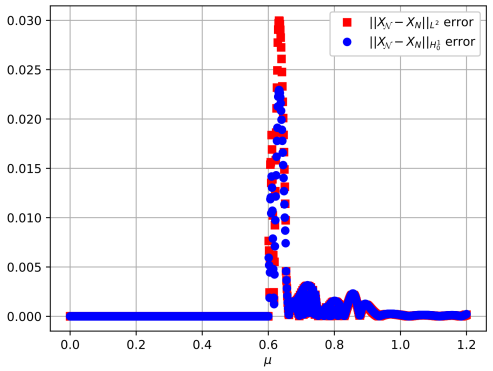
(a) branch $|0,0\rangle$



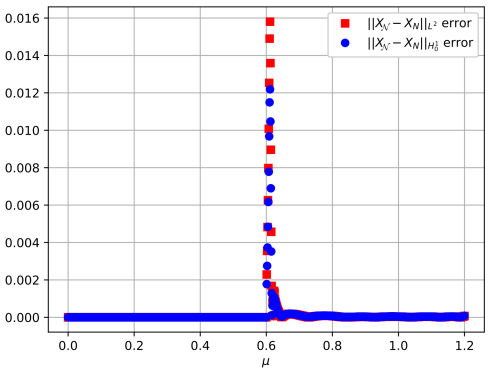
(b) branch $|0,1\rangle$



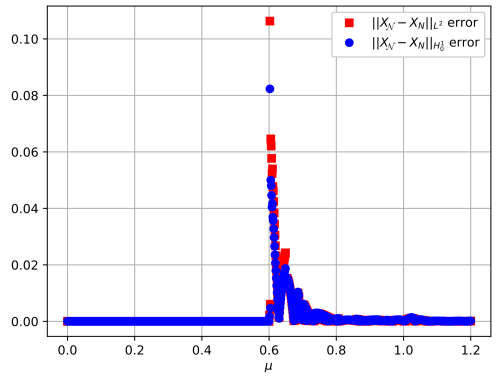
(c) branch $|1,0\rangle$



(d) branch $|1,1\rangle$

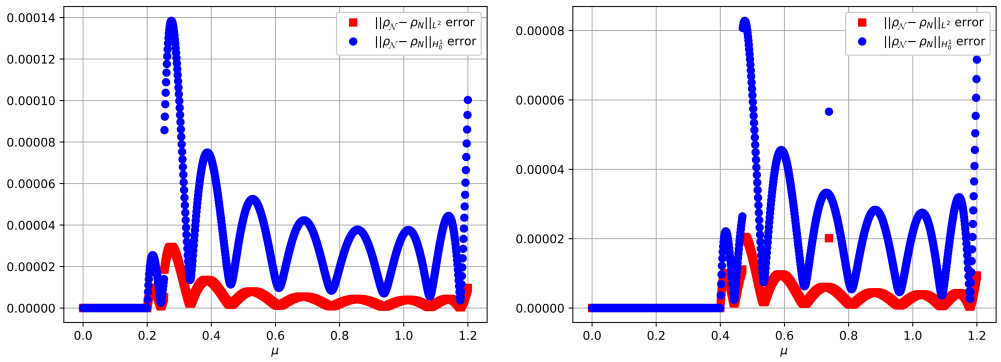


(e) branch $|0,2\rangle$

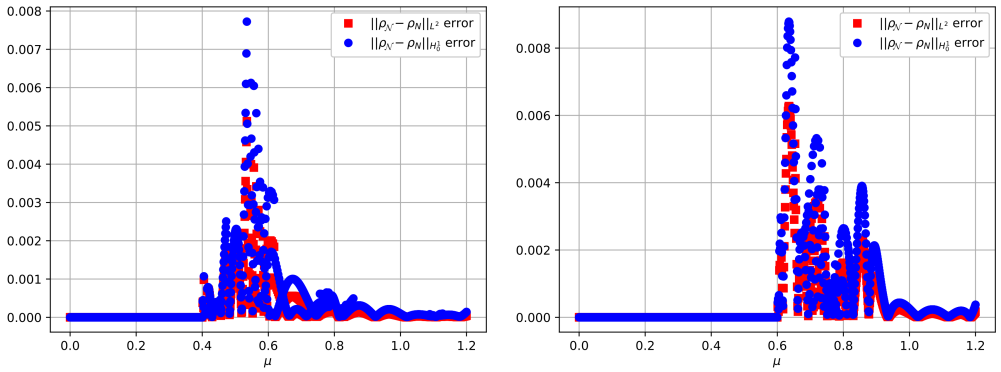


(f) branch $|2,0\rangle$

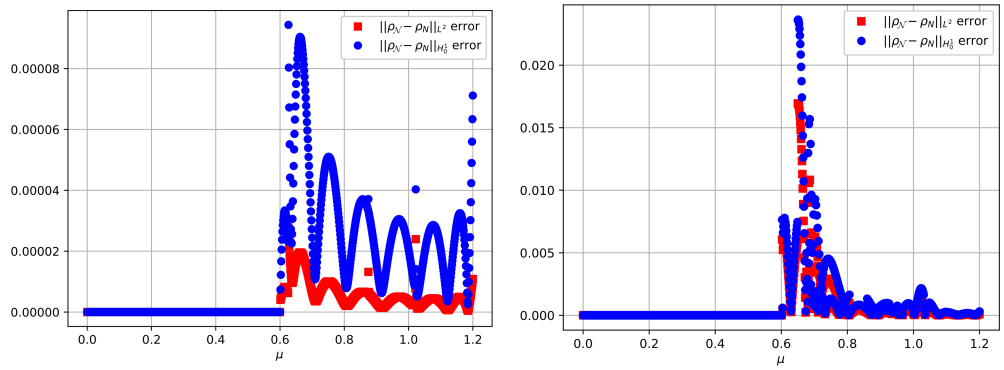
Figure 5.5: Difference between $X_{\mathcal{N}}$ (FOM solution) and X_N (ROM solution) in the L^2 and H_0^1 norms for each one of the six solution branches.



(a) branch $|0,0\rangle$ (b) branch $|0,1\rangle$

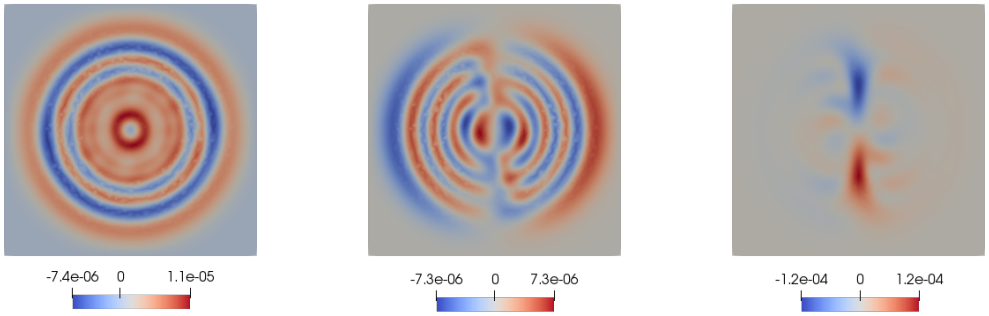


(c) branch $|1,0\rangle$ (d) branch $|1,1\rangle$



(e) branch $|0,2\rangle$ (f) branch $|2,0\rangle$

Figure 5.6: Difference between $\rho_{\mathcal{N}}$ (density computed with FOM) and ρ_N (density computed with ROM) in the L^2 and H_0^1 norms for each one of the six solution branches.



(a) branch $|0,0\rangle$ (b) branch $|0,1\rangle$ (c) branch $|1,0\rangle$

(d) branch $|1,1\rangle$ (e) branch $|0,2\rangle$ (f) branch $|2,0\rangle$

Figure 5.7: Difference between the density computed with FOM and ROM for the first six branches at $\mu = 1.2$.

5.2.3 Multi-parameter analysis with trap strength

In this section, we want to investigate the Gross-Pitaevskii model in the multi-parameter case, in which the chemical potential μ varies in interval $[0, 0.8]$ and the trap strength τ varies in interval $[0.1, 0.3]$. We are thus acting on the concavity of the parabolic trap, considering a varying strength depending on the value τ.

First, we focus on the first bifurcation, i.e. branch $|0,0\rangle$. As concerns the reduced manifold, we applied the Algorithm 1 for 3 equispaced points in $[0.1, 0.3]$, in order to recover the branching state from $\mu^* = \tau$ for $K = 160$ equispaced points in $[0, 0.8]$. The resulting basis in this case consists of only $N = 12$ modes, with a POD tolerance equals to $\varepsilon_{POD} = 10^{-8}$.

Figure 5.8 shows such branch in a two-parameter bifurcation diagram obtained with the Reduced Order Model as μ and τ are varied. As expected from the theory (5.5), we see that as τ increases the critical value of μ for the first bifurcation increases linearly. Recall that $\mu^* = \tau$, for $m = n = 0$, and Figure 5.8 clearly shows it (see the black dotted line in the highlighted red rectangle). For Figure 5.8, we used 21 equispaced points in $[0.1, 0.3]$ with an increment $\Delta\tau = 0.01$ and continuation step $\Delta\mu = 1.25 \cdot 10^{-3}$.

As concerns the efficiency, due to the lack of an hyper-reduction strategy, the speed-up is only 1.52 with almost $t_{RB} = 2.73$(h) needed for the RB reconstruction of the full 3-D diagram in Figure 5.8, in contrast to the $t_{FE} = 4.15$(h) of the FE method. This represents an improvement over the 1.1 speed-up for the one parameter test case, but it is still not enough to justify the computational costs of the offline phase.

From these considerations, since we used a small number of basis functions, we can highlight that the still high computational time for the reduced approach is mainly caused by the assembly of the Jacobian matrix, which involves the projection on the high fidelity space of dimension $\mathcal{N}$.

Despite this, it is clear that in general through the multi-parameter approach one can really gain a lot of information about the model in an efficient way (once that one has an $\mathcal{N}$-independent online phase, as we will see in the next section). Indeed, also for this rich and complex bifurcating model, through the POD of three branches (namely for $\tau = 0.1, 0.2, 0.3$) we were able to built a unique ROM which allows to approximate with good accuracy all the first branches emerging from 21 equispaced values of $\tau \in [0.1, 0.3]$. In fact as we can see from Figure 5.10 the reduced errors for the density show a maximum error of the order $1.e-3$, with peaks located at the current value of τ. The corresponding representative solutions at $\mu = 0.8$ are depicted in Figure 5.9, where it is clear that the more one increases the trap strength the more the solutions remain confined in a smaller region around the center of the domain.

Next, we focus on the first two bifurcations but we also restrict the study to only 5 equispaced value in $[0.1, 0.3]$. In Figure 5.11 we show the reduced order bifurcation plot for the first three branches, i.e. branches $|0,0\rangle$, $|0,1\rangle$, and $|1,0\rangle$. The reduced manifold was obtained again by sampling 3 equispaced trap values in $[0.1, 0.3]$, for $K = 160$ equispaced points $\mu \in [0, 0.8]$, but also for each one of first the three emerging branch (one for $\mu^* = \tau$ and two for $\mu^* = 2\tau$). Of course, the resulting basis in this case

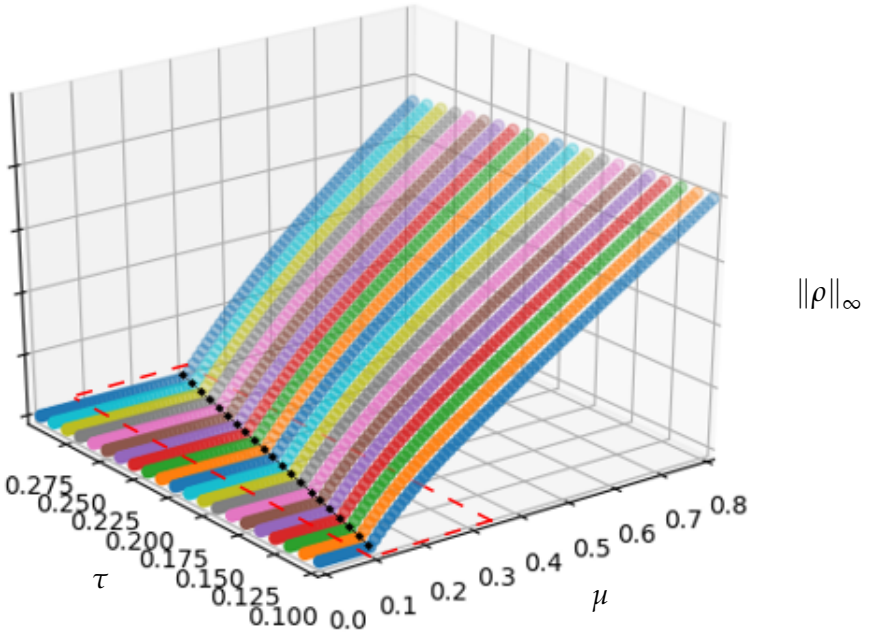


Figure 5.8: First bifurcation in a two-parameter bifurcation diagram obtained with the Reduced Order Model: infinity norm of the density as chemical potential μ and trap strength τ vary. The black dotted line in the highlighted red rectangle shows the critical value of μ for the first bifurcation: $\mu^* = \tau$.

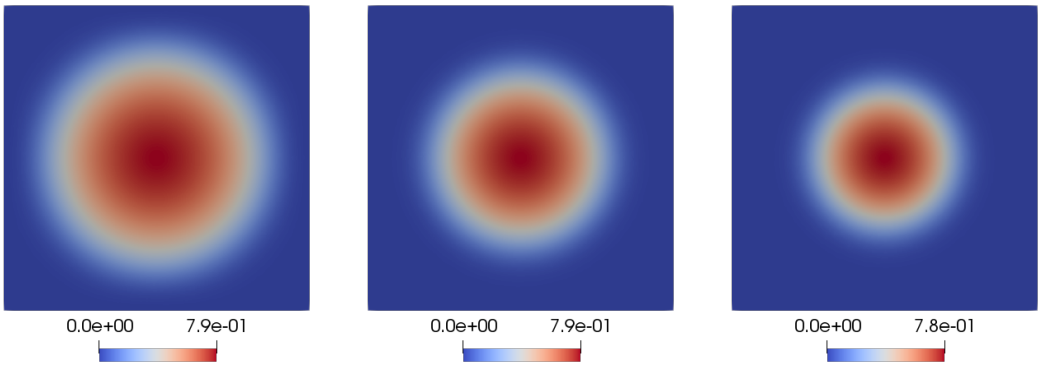
has a much greater dimension, in order to capture all the possible states. As a matter of fact, we obtained $N = 41$ modes with a POD tolerance equals to $\varepsilon_{POD} = 10^{-8}$.

We see that our ROM approach successfully captures also the critical μ for the second bifurcation: branches $|0,1\rangle$ and $|1,0\rangle$ depart from $\mu^* = 2\tau$. Just like for Figure 5.8, we used continuation step $\Delta\mu = 1.25 \cdot 10^{-3}$.

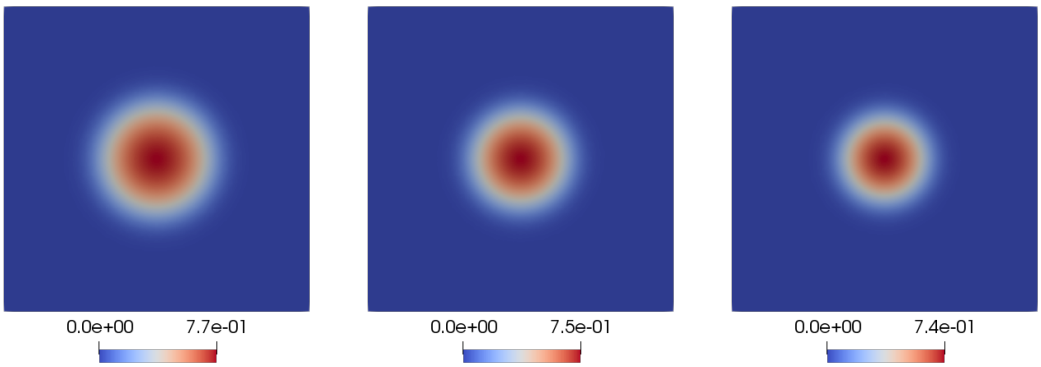
As concerns the efficiency, due to the lack of an hyper-reduction strategy, the speed-up is only 1.12 with almost $t_{RB} = 2.55$(h) needed for the RB reconstruction of the full 3-D diagram in Figure 5.11, in contrast to the $t_{FE} = 2.86$(h) of the FE method.

Here, we highlight that the computational speed-up is worse than the previous case since we are using a much bigger number of basis functions (11 vs 41). Once again, we highlight the remarkable fact that the ROM we built by compressing the branches for three equispaced values of τ, is able to detect and properly follow the bifurcating states emerging for different and not sampled values of the trap strength. Thus, also introducing qualitatively different solutions in the POD stage, the ROM is able to capture the energy coming from all the branches and to provide basis functions which well approximate the high fidelity manifold of solutions.

As we can see from Figure 5.12 the reduced errors for the density show again a maximum error of the order $1.e-3$, with peaks near the corresponding value of τ.



(a) $\tau = 0.12$ (b) $\tau = 0.15$ (c) $\tau = 0.18$

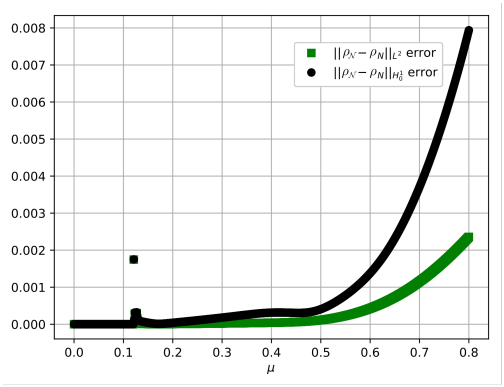


(d) $\tau = 0.22$ (e) $\tau = 0.25$ (f) $\tau = 0.28$

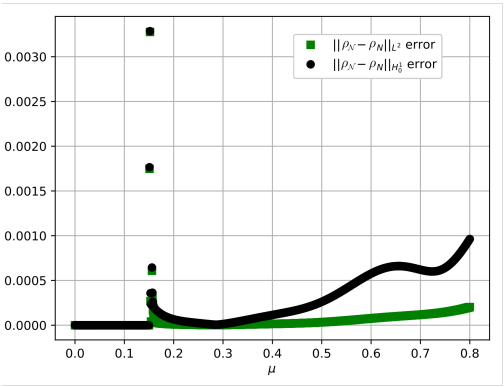
Figure 5.9: Density functions computed with the reduced basis method at $\mu = 0.8$ varying the trap strength $\tau \in [0.1, 0.3]$. From (a) to (f) the trapping parameter is confining the solutions in a smaller region.

Some representative solutions at $\mu = 0.8$ are shown in Figure 5.13, where we can observe the same effect that we noted before, with the density solutions that shrinks towards the center of the domain.

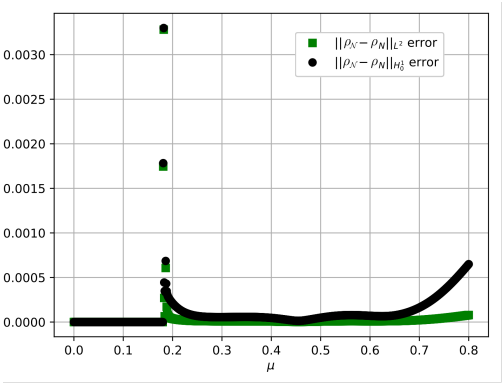
Finally, we can now present the application of the hyper-reduction strategies, that will allow us to fully decouple the online phase from the degrees of freedom $\mathcal{N}$ of the high fidelity problem.



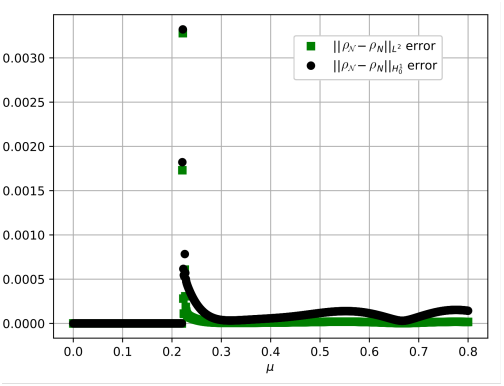
(a) $\tau = 0.12$



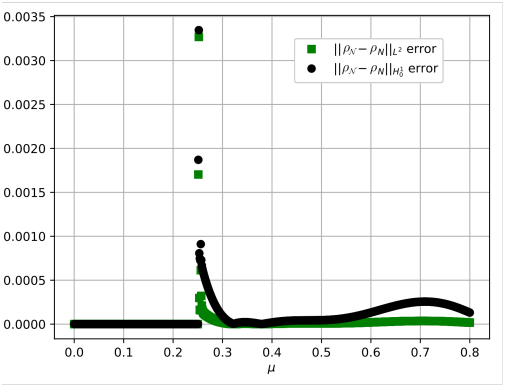
(b) $\tau = 0.15$



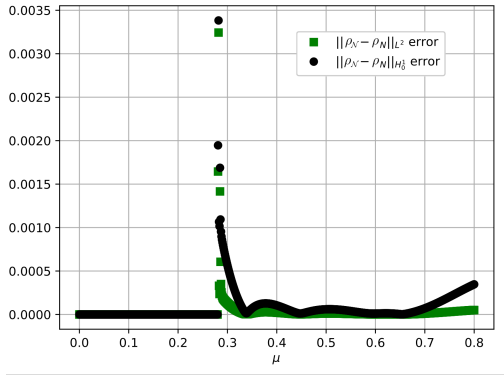
(c) $\tau = 0.18$



(d) $\tau = 0.22$



(e) $\tau = 0.25$



(f) $\tau = 0.28$

Figure 5.10: Difference between $\rho_{\mathcal{N}}$ and ρ_N in the L^2 and H_0^1 norms for in the multi-parameter case.

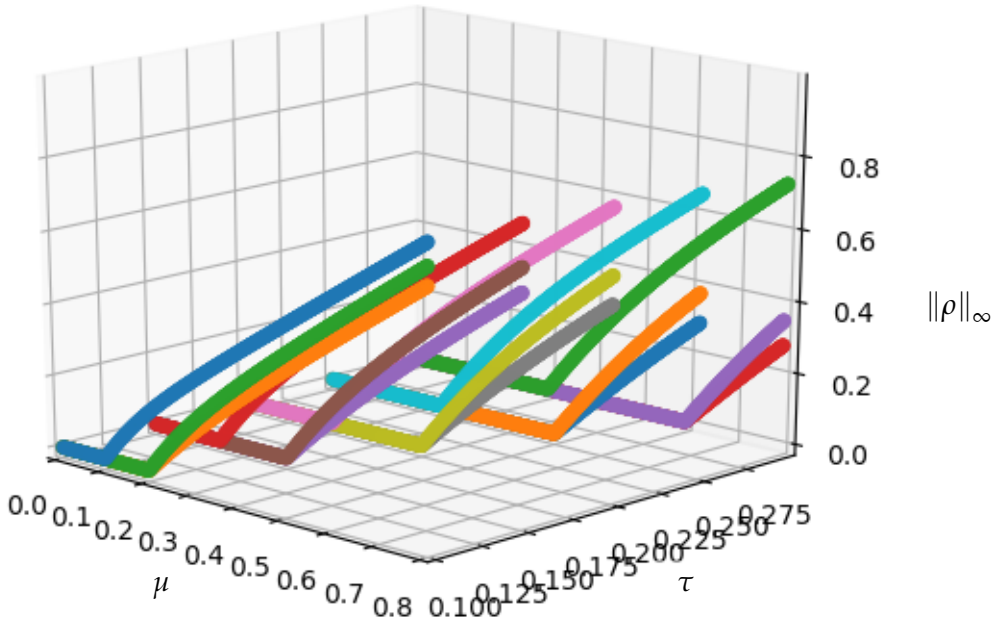
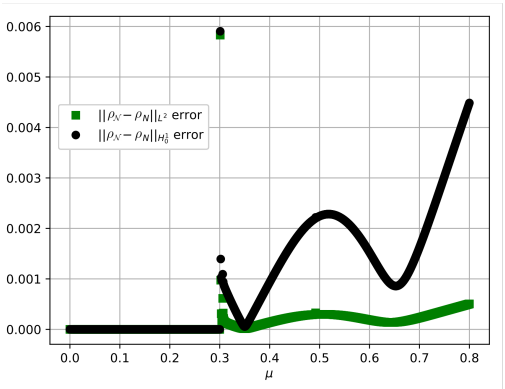
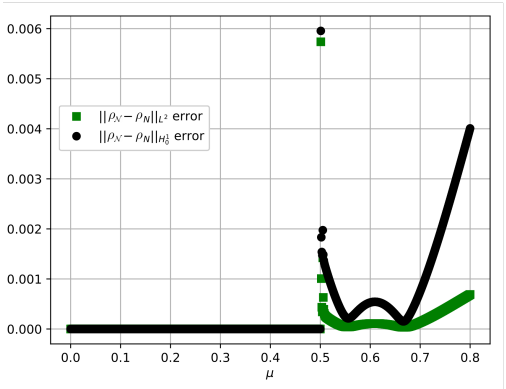


Figure 5.11: First two bifurcations in a two-parameter bifurcation diagram obtained with the Reduced Order Model: infinity norm of the density as chemical potential μ varies for trap strength $\tau = 0.1, 0.2, 0.3$.

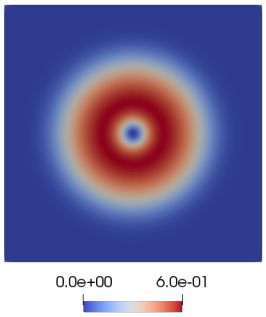


(a) branch $|0,1\rangle$ emerging at $\mu^* = 0.3$

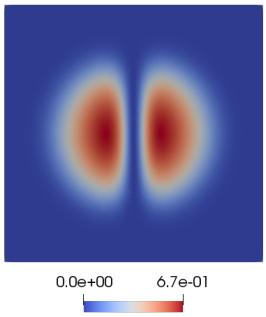


(b) branch $|0,1\rangle$ emerging at $\mu^* = 0.5$

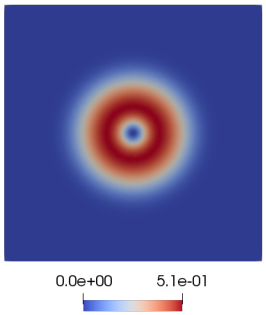
Figure 5.12: Difference between $\rho_{\mathcal{N}}$ and ρ_N in the L^2 and H_0^1 norms for $\tau = 0.15$ and $\tau = 0.25$, left and right respectively, in the multi-parameter case.



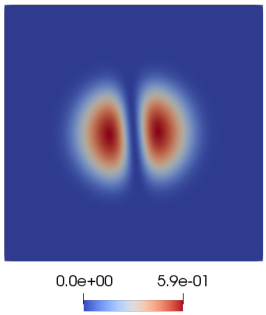
(a) branch $|0,1\rangle$ for $\tau = 0.15$



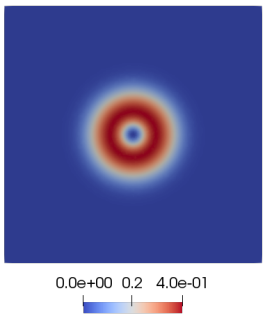
(b) branch $|1,0\rangle$ for $\tau = 0.15$



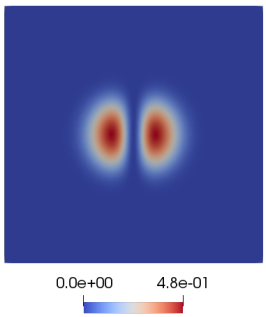
(c) branch $|0,1\rangle$ for $\tau = 0.2$



(d) branch $|1,0\rangle$ for $\tau = 0.2$



(e) branch $|0,1\rangle$ for $\tau = 0.25$



(f) branch $|1,0\rangle$ for $\tau = 0.25$

Figure 5.13: Reduced basis density functions for the branches $|1,0\rangle$ and $|0,1\rangle$, left and right column, at $\mu = 0.8$ varying the trap strength $\tau \in [0.1, 0.3]$. From top to bottom the trapping parameter is confining the solutions in a smaller region.

5.2.4 Hyper-reduction with EIM/DEIM techniques

In this section, we apply the affine-recovery techniques described in Section 3.2.4, EIM and DEIM, in order to efficiently reconstruct the branching behaviour of the model, for the one and two parameter test cases. The purpose of this section is two-fold: show that the branch is reconstructed accurately with a report on the substantial computational time savings enabled by the hyper-reduction strategies and compare the approximation properties of the two techniques.

Let us start with the one parameter test case in Section 5.2.2. We focus here on the reconstruction of a single branch, say $|0,1\rangle$ emerging from $\mu^* = 2\tau$. Within the same setting as before, we consider the EIM with Greedy tolerance $\varepsilon_{Gr} = 10^{-7}$ together with a basis for the reduced manifold of dimension $N = 6$. The EIM involves an affine decomposition made up by 6 terms, each one approximated by almost 10 suitable interpolation basis functions.

On the contrary, the DEIM approach with the same Greedy tolerance ε_{Gr} produces 2 affine terms approximated again by 10 interpolation basis functions. Also in this case the basis for the reduced manifold has dimension $N = 6$.

Thus, we show in Figure 5.14 the reduced order errors E_{N_B} (left) and E_ρ (right) for branch $|0,1\rangle$. We see that the values of E_{N_B} for EIM and DEIM are actually comparable over the entire interval $[0, 1.2]$, with a peak at $\mu = 1.2$ of the order of 10^{-5}. Error E_ρ is slightly larger for EIM, with a peak of the order of 10^{-4}.

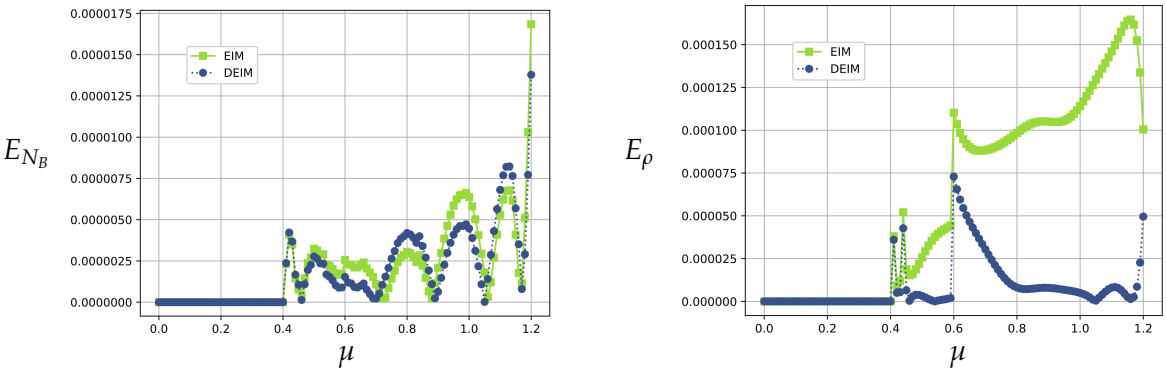


Figure 5.14: Difference in absolute value between the branches of the bifurcation diagram computed with FOM and ROM with EIM/DEIM in the μ-N_B plane (left) and in the μ-$\|\rho\|_\infty$ plane (right), i.e. reduced order errors E_{N_B} (left) and E_ρ (right) for branch $|0,1\rangle$.

Finally, Figure 5.15 shows the difference between FOM solution and ROM solution computed with EIM (left) and DEIM (right) in the L^2 and H^1_0 norms, again for branch $|0,1\rangle$. We observe slightly larger error peaks than in the case of ROM without affine-recovery technique, as we can see comparing Figure 5.15 with Figure 5.5b. These increased errors are the small price to pay for a considerable computational speed-up.

Indeed, with EIM, it takes only $t_{RB,EIM} = 55$(s) to construct branch $|0,1\rangle$ while it takes $t_{HF} = 246$(s) with FOM. So, our ROM approach coupled with EIM is almost five times faster than the FOM. As for DEIM, the computational time savings are even better: it takes only $t_{RB,DEIM} = 7$(s) for the construction of branch $|0,1\rangle$, with corresponds to a speed-up factor of almost 32. This drastic reduction of the computational time is mainly due to the different number of affine terms that we have previously mentioned, but as we can see the order of accuracy is almost the same.

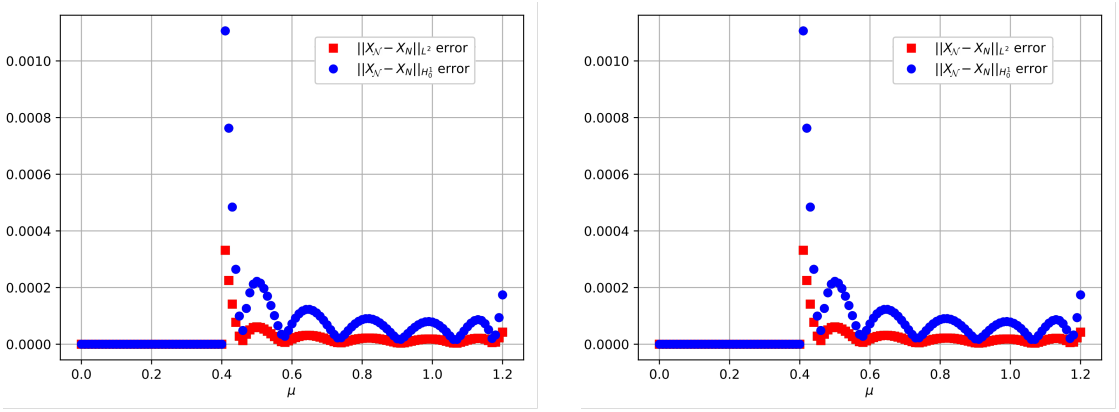


Figure 5.15: Difference between $X_{\mathcal{N}}$ (FOM solution) and X_N (ROM solution) computed with EIM (left) and DEIM (right) in the L^2 and H_0^1 norms for branch $|0,1\rangle$.

The hyper-reduction techniques are especially effective in reducing the computational time needed for the two-parameter studies presented in Section 5.2.3. Before presenting the results for the hyper-reduction approach to the multi-parameter case, we remark that one has to pay attention on the application of the affine-recovery techniques when many different configurations are taken into consideration.

Indeed, let us consider e.g. the DEIM approach for the reconstruction of the bifurcation diagram in Figure 5.1. In order to keep low the number of training points N_{train}, we chose for each branch a parameter space centred in each corresponding bifurcation point with amplitude equals to 0.1. Choosing a tolerance $\varepsilon_{Gr} = 10^{-7}$ for the DEIM strategy, the two affine terms were approximated by means of 50 interpolation basis functions each. As concerns the dimension of reduced manifold, due to the smaller parameter space considered here, we need $N = 18$ basis functions to reach the POD tolerance 10^{-9}.

As we can see from Figure 5.16 strange issues can occur, where for example some branches are properly and efficiently recovered online, while others are completely missed. Of course, this can happen due to the non-convergence of the Newton iterations, but more frequently the Algorithm 2 converges to the "wrong" branch or to the "right" one, but rotated of a certain angle θ. This is actually the case reported in Figure 5.17, where we show the high fidelity real part φ of the solution X and its

reduced approximation belonging to the branch $|1,0\rangle$. It is evident that, in connection to what previously stated in (5.13), the reduced configuration on the right correspond to a rotation of the high fidelity approximation, which is still a solution for the system.

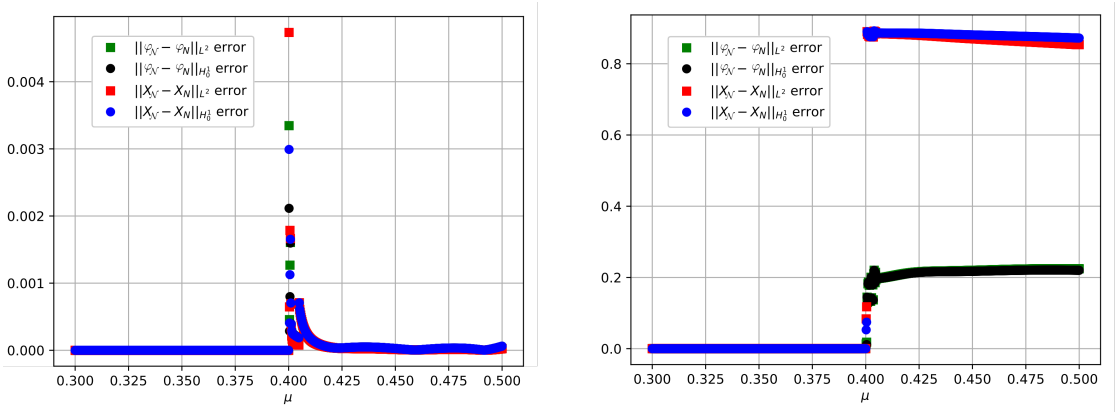


Figure 5.16: Reduced basis errors for the solution X and its real part φ with DEIM approach for $|0,1\rangle$ branch (left) and $|1,0\rangle$ branch (right) in the L^2 and H_0^1 norms.

Figure 5.17: High fidelity and reduced basis real part φ of the solution X for the $|1,0\rangle$ branch, left and right respectively, at $\mu = 0.5$.

For these reasons, we avoided to consider here the reconstruction of the graph in Figure 5.11, while instead we focused on a DEIM application to recover efficiently the bifurcation diagram in Figure 5.8. This choice is motivated by the fact that now we are considering the evolution of a fixed branch with respect to the trapping parameter τ, instead of mixing different branches associated to all the existing states.

Within the same setting as before, we considered a tolerance $\varepsilon_{Gr} = 10^{-7}$, which produces 15 interpolation basis functions for each one of the two terms. The $N = 13$ reduced basis functions with POD tolerance $\varepsilon_{POD} = 10^{-9}$ allowed us to efficiently reconstruct the 3-D bifurcation diagram in Figure 5.8.

Our ROM approach with DEIM took only $t_{RB,DEIM} = 6.1$(min) to reconstruct the reduced diagram in Figure 5.11, obtaining a speed-up factor of 30 with respect to our ROM approach without DEIM and a factor 40 with respect to the FOM.

The reduced algorithm we have developed proved to work well also in connection to richer and more difficult bifurcating phenomena. In particular, we also tested more deeply the application of hyper-reduction strategies to recover the efficiency, remarking how much powerful and fundamental these techniques could be, but also pointing out some critical issues that could be analysed further.

Before moving to Fluid Dynamics applications, we take a step back to Continuum Mechanics problems. Indeed, in the next chapter we will deal with the buckling of beams. Despite the similar context, we will study a different class of models which exhibit a bifurcating behaviour. In particular, we will also consider more complex test cases, including: different physical settings, multi-parameter applications, geometrical parametrization and three-dimensional geometries. All these benchmarks will led us to the investigation of a real test case scenario motivated by industry application.

Chapter 6

Hyperelastic models for bending beams

In this chapter, we deal again with physical phenomena in the context of Continuum Mechanics. Indeed, as the Von Kármán equations describes the buckling of a plate, many different models can be investigated in the study of the deflection of beams. As for the plates, also the beams are subjected to a sudden variation of their equilibrium state when a compression acting on their boundaries exceeds a critical value. The aim here is to provide a brief description of the bifurcating (buckling) phenomena for 2-D and 3-D beams with different constitutive relations, in particular we will focus on the class of the so called hyperelastic models. Moreover, we will analyse a multi-parametric setting in which the bifurcation point is studied also in conjunction to a geometrically parametrized domain. Finally, we will use the information obtained with these toy problems in a real case scenario coming from the Norwegian petroleum industry [114]. This is an ongoing work with J. Eftang, G. Rozza and A. T. Patera [118] in the framework of the MIT-FVG "ROM2S" project (Regione Friuli Venezia Giulia).

6.1 A Continuum Mechanics framework

Let us given $\Omega \subset \mathbb{R}^d$, with $d = 2, 3$, a bounded domain as the reference configuration of a body $\mathcal{B}$ that undergoes deformation. We can capture the *displacement* u of a material point P, from the reference position X to the new deformed location x, through the deformation function $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ defined as $x = \varphi(X)$. In fact, to capture the motion of a point, we can write $u(X) = x - X$, and from this it is immediate to define the deformation gradient as

$$F = \frac{\partial \varphi(X)}{\partial X} = \nabla u + I.$$

Moreover, we can also define the determinant of the deformation gradient, which encodes the volume changes, as $J = \det(F)$.

Once defined these basic quantities, we can present the standard and well-known equilibrium equation derived from the equation of motion [150, 41, 42, 81, 153]

$$-\operatorname{div}(P) = B \quad \text{in } \Omega, \tag{6.1}$$

where B is an external force acting on the whole domain Ω and $P = P(F)$ is the first Piola-Kirchhoff stress tensor, related to the Cauchy stress tensor σ by the formula $P = J\sigma F^{-T}$.

Of course, the computation of the displacement field u of the considered body as a function of the external loads requires to complement equation (6.1) by a suitable law. In particular, if we want to to study the equation (6.1) in the context of elasticity problems, we have to characterize it through a relation between the stresses (forces) and the strains (displacements).

These relationships, known as *constitutive equations*, highly depend on the type of material under consideration. In this context, we relied on the so called *hyperelastic* material (or *Green elastic* material) [68], which are characterized by the existence of a function which defines the Piola-Kirchhoff stress tensor.

More precisely, we can define a *strain energy* function $\psi = \psi(F)$ such that

$$P(F) = \frac{\partial \psi(F)}{\partial F}, \tag{6.2}$$

which is used to define an hyperelastic material by assuming that the stress can be obtained by taking the derivative of the energy ψ with respect to the strain F.

Of course, many simplifications can be adopted when dealing with specific type of materials. As an example, for an hyperelastic material which is also *isotropic* (its properties are independent from the direction of examination), the strain energy function can be expressed only by means of the *principal stretches*. Indeed, from these one can obtain the *principal invariants* of the (left) Cauchy–Green deformation tensor $C = F^T F$, which are defined as

$$\begin{aligned} I_1 &= \mathrm{tr}(C), \\ I_2 &= \frac{1}{2}[\mathrm{tr}(C)^2 - \mathrm{tr}(C^2)], \\ I_3 &= \det(C). \end{aligned} \tag{6.3}$$

Moreover, we remark that these materials are characterized by the fact that the work done by stresses does not depend on the path of deformation, thus they conserve the total energy, and in contrast with linear elasticity, we do not have to require any infinitesimal assumption on the strains.

We can now proceed with the description of the models considered for the study of compressed beams. In particular, we will analyse two different choices for the strain energy function ψ, investigating numerically their properties with respect to the buckling problems.

The simplest constitutive relation is the so called *Saint Venant-Kirchhoff* (SVK) model, which is an extension of the linear elastic material model to the geometrically nonlinear regime. This model is defined through the strain energy function defined as

$$\psi(F) = \lambda_1 \mathcal{E} : \mathcal{E} + \lambda_2 (\mathrm{tr}(\mathcal{E}))^2/2, \tag{6.4}$$

where

$$\mathcal{E} = \frac{1}{2}(F^T F - I)$$

is the Green-Lagrange strain tensor (a measure for varying lengths between points), λ_1 and λ_2 are the Lamé constants, which are related to the material properties through the Young modulus E and the Poisson ratio ν as follows

$$\lambda_1 = \frac{E}{2(1+\nu)}, \quad \lambda_2 = \frac{E\nu}{(1+\nu)(1-2\nu)}.$$

Moreover, from the definition of the hyperelastic material in (6.2), we obtain that the first Piola-Kirchhoff stress tensor for the SVK model is given by

$$P(F) = F(2\lambda_1 \mathcal{E} + \lambda_2 \operatorname{tr}(\mathcal{E}) I).$$

We remark that, as we said before, such model is analogous to the linear elasticity one. Indeed, one can characterize the linear elasticity model by means of the strain energy function

$$\psi(F) = \lambda_1 \varepsilon : \varepsilon + \lambda_2 (\operatorname{tr}(\varepsilon))^2, \tag{6.5}$$

where

$$\varepsilon = \frac{1}{2}(F + F^T - I)$$

is the infinitesimal strain tensor.

The second hyperelastic constitutive relation we will consider is the so called *neo-Hookean* (NH) model, which can be expressed through the the strain energy function defined as

$$\psi(F) = \frac{\lambda_1}{2}(I_1 - 3) - \lambda_1 \ln J + \frac{\lambda_2}{2}(\ln J)^2, \tag{6.6}$$

where I_1 is the first principal invariants in (6.3) and λ_1, λ_2 are the Lamé constants as before. The first Piola-Kirchhoff stress tensor in this case is given by

$$P(F) = F[(\lambda_1(I - C^{-1}) + \lambda_2(\ln J)C^{-1}].$$

We remark that the words buckling and bifurcation will be used here indistinctly, but we will be interested only in the approximation of the first buckling mode, properly following the post-buckling behaviour as the target branch.

In the next section, we will present the mathematical formulation of the problem, focusing on its weak formulation that will allow us to apply our methodology.

6.1.1 Weak formulation

Having presented the constitutive relations that we will use in the rest of the chapter, we can now state the mathematical formulation of the problem. Starting from equation (6.1) we can consider the boundary value problem, in the reference domain Ω, given by

$$\begin{cases} -\operatorname{div}(P(u)) = B & \text{in } \Omega, \\ u = u_D & \text{in } \Gamma_D, \\ P(u)n = T & \text{in } \Gamma_N, \end{cases} \tag{6.7}$$

where $u : \Omega \to \mathbb{R}^d$ is the in-plane displacement, B is the *body force* per unit reference area, and T is the *traction force* per unit reference length. Thus, our aim is to study the deformation of the domain Ω, subjected to a prescribed displacement u_D on the Dirichlet boundary Γ_D together with body and traction forces. We remark that, due to the Dirichlet non-homogeneous BCs, the function space in which we set the problem is the space $H^1_D(\Omega) = \left\{v \in H^1(\Omega) \mid v = u_D \in \Gamma_D\right\}$, where in this case we are assuming that each component of u is in $H^1_D(\Omega)$.

To derive the weak formulation of the problem, we proceed as usual by taking the dot product with a test function $v \in \mathbb{X} = H^1_0(\Omega)$ and integrating over the reference domain Ω, this way we obtain

$$\int_\Omega \operatorname{div}(P(u)) \cdot v \, d\Omega + \int_\Omega B \cdot v \, d\Omega = 0 \quad \forall\, v \in \mathbb{X}. \tag{6.8}$$

Applying the divergence theorem and embedding the traction and displacement boundary conditions (the test function $v \in \mathbb{X}$ satisfies homogeneous BCs on Γ_D), the weak formulation reads as: find $u \in H^1_D(\Omega)$ such that

$$\int_\Omega P(u) : \nabla v \, d\Omega - \int_\Omega B \cdot v \, d\Omega - \int_{\Gamma_N} T \cdot v \, d\Gamma = 0 \quad \forall\, v \in \mathbb{X}. \tag{6.9}$$

In practice, a standard technique to set the problem in the same space is to consider a lifting function $R_D \in H^1(\Omega)$ such that $R_D|_{\Gamma_D} = u_D$. Then, one sets $\mathring{u} = u - R_D$ and, from the immediate consideration that $\mathring{u} \in \mathbb{X}$, the symmetry in spaces between solution and test functions is restored.

Furthermore, the boundary value problem in (6.7) for hyperelastic media can be also expressed as a minimization problem by means of the the *Theorem of Virtual Work* [68]. In fact, we can define the *potential energy* of the beam in term of the strain energy function ψ as

$$\Pi(u) = \int_\Omega \psi(u) \, d\Omega - \int_\Omega B \cdot u \, d\Omega - \int_{\Gamma_N} T \cdot u \, d\Gamma. \tag{6.10}$$

At minimum point of $\Pi(u)$, the directional derivative of Π with respect to the change in u is given by

$$g(u, v) := D_v \Pi(u) = \left.\frac{\mathrm{d}\Pi(u + \delta v)}{\mathrm{d}\delta}\right|_{\delta=0}, \tag{6.11}$$

is equal to zero for all $v \in \mathbb{X}$, that is

$$g(u, v) = 0 \quad \forall\, v \in \mathbb{X}. \tag{6.12}$$

It is easy to observe that evaluating equation (6.11) leads to the weak formulation in (6.9). The constitutive relation we chose are characterized by the fact that equation

(6.12) results in a nonlinear form w.r.t. u. In such cases we need to rely on the Jacobian of g, that is defined by

$$\mathrm{d}g[u](z,v) = \left.\frac{\mathrm{d}g(u+\delta z, v)}{\mathrm{d}\delta}\right|_{\delta=0}. \tag{6.13}$$

Before moving to the numerical approximation of the problem, we want to highlight that the variational forms above were presented without any dependency from a generic (multi-)parameter $\boldsymbol{\mu}$. The reason behind this choice is that we will consider different geometrical/physical parameter for each toy problem, specifying case by case the parametrized quantities.

6.2 Numerical approximation

Here, we will present different toy problems, in which we will study the buckling behaviour of a beam with different constitutive relations, boundary conditions, geometry settings and external forces. Moreover, the parametric investigation by means of the the reduced order strategy developed in Section 3.2 will allow an efficient computation of the bifurcation diagrams.

Since the only variable involved will be the displacement $u \in \mathbb{R}^d$, with $d = 2, 3$, and the forms highly depend on the constitutive equations we will present the high fidelity projection only in an abstract version. Therefore, as we did in (3.3) the projection of the weak formulation (6.9) in the finite dimensional space $\mathbb{X}_{\mathcal{N}}$ of dimension $\mathcal{N}$, reads as: given $\boldsymbol{\mu} \in \mathcal{P}$, seek $u_{\mathcal{N}} := u_{\mathcal{N}}(\boldsymbol{\mu}) \in \mathbb{X}_{\mathcal{N}}$ that satisfies

$$g(u_{\mathcal{N}}, v_{\mathcal{N}}; \boldsymbol{\mu}) = 0 \quad \forall\, v_{\mathcal{N}} \in \mathbb{X}_{\mathcal{N}}, \tag{6.14}$$

where we introduced in the weak formulation the geometrical and/or physical multi-parameter $\boldsymbol{\mu} \in \mathcal{P}$. The application of the Newton method leads us to solve, at the generic k-th step, the algebraic equation

$$\mathsf{J}_{\mathcal{N}}(\mathsf{u}^k_{\mathcal{N}}(\boldsymbol{\mu}); \boldsymbol{\mu})\delta\mathsf{u}_{\mathcal{N}} = \mathsf{G}_{\mathcal{N}}(\mathsf{u}^k_{\mathcal{N}}(\boldsymbol{\mu}); \boldsymbol{\mu}), \tag{6.15}$$

updating the solution as $\mathsf{u}^{k+1}_{\mathcal{N}} = \mathsf{u}^k_{\mathcal{N}} - \delta\mathsf{u}_{\mathcal{N}}$ until convergence.

In order to efficiently recover the solution for repeated instances of the parameter $\boldsymbol{\mu}$ during the online stage, we will rely on the projection in the reduced basis space $\mathbb{X}_N$. Thus, given $\boldsymbol{\mu} \in \mathcal{P}$, we seek $u_N := u_N(\boldsymbol{\mu}) \in \mathbb{X}_N$ that satisfies

$$g(u_N, v_N; \boldsymbol{\mu}) = 0 \quad \forall\, v_N \in \mathbb{X}_N, \tag{6.16}$$

which form the algebraic standpoint translates into

$$\mathsf{J}_N(\mathsf{u}^k_N(\boldsymbol{\mu}); \boldsymbol{\mu})\delta\mathsf{u}_N = \mathsf{G}_N(\mathsf{u}^k_N(\boldsymbol{\mu}); \boldsymbol{\mu}), \tag{6.17}$$

with the reduced residual and the reduced Jacobian matrix defined as in (3.19) and (3.21), respectively.

Now we can finally present the different test cases analysed to investigate the phenomenon of buckling for elastic beams.

6.2.1 2-D toy problem

In this first application, we consider a two-dimensional beam which corresponds to the domain $\Omega = [0,1] \times [0,0.1]$. We built the Finite Element space with $\mathbb{P}_1$ linear elements, resulting in an high fidelity dimension $\mathcal{N} = 4328$. It is well-known that, e.g. for $B = (0,0)$ and $T = 0$, the bifurcation diagram enjoys a $\mathbb{Z}^2$ reflective symmetry and it undergoes a series of pitchfork bifurcations as μ is increased. Despite this, here we focused only on the first branch, being usually the most common cause of failure. In this simple 2-D context, we will analyse different test cases in which the buckling properties will be studied in connection to different compression conditions, materials characteristics and geometries.

6.2.1.1 Dirichlet compression

Here, we consider a beam subjected to a parametrized uniform compression imposed by means of Dirichlet boundary conditions. In particular, we will analyse the Saint Venant-Kirchhoff and the neo-Hookean models, choosing: a null traction force $T = 0$, a Young modulus $E = 10^6$ and a Poisson ratio $\nu = 0.3$, while for the body force we will study either $B = (0,0)$ or $B = (0,-1000)$. In particular, we want to study the buckling of the beam subjected to a compression of magnitude μ on its right. To do so, we rewrite the Dirichlet conditions in (6.7) as

$$\begin{aligned} u &= (0,0) \quad \text{on } \Gamma_D^l = \{0\} \times [0,0.1], \\ u &= (-\mu,0) \quad \text{on } \Gamma_D^r = \{1\} \times [0,0.1], \end{aligned} \tag{6.18}$$

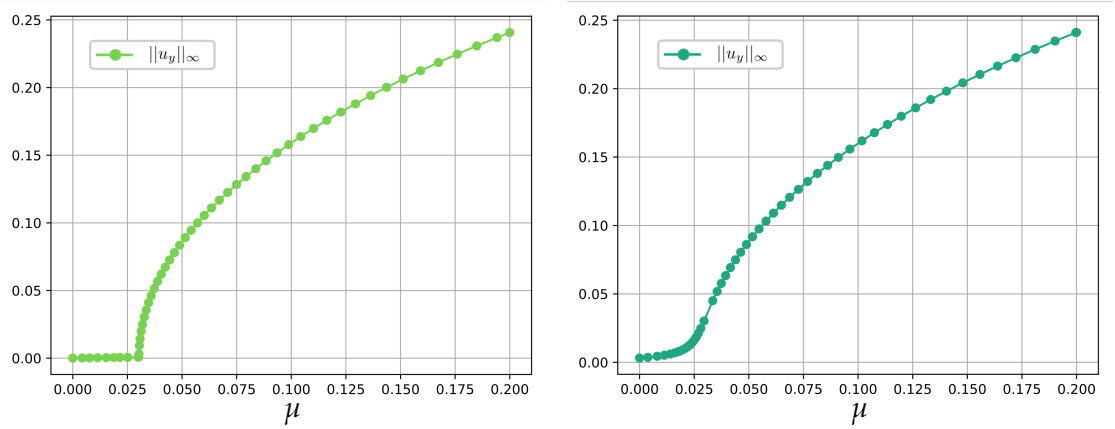
which correspond to a clamped condition on the left end of the beam, and an increasing uniform uni-axial compression on the other end. Thus, for a fixed compression μ, we will consider the functional space

$$\mathbb{X} = \{u \in (H^1(\Omega))^2 : u = (0,0) \text{ on } \Gamma_D^l,\ u = (-\mu,0) \text{ on } \Gamma_D^r\}.$$

Despite the simplicity of the models, the buckling phenomenon makes the analysis of the behaviour of the solution with respect to the compression μ not straightforward, especially as concerns the reduction strategies. Indeed, for the high fidelity setting we had to choose $N_{train} = 1000$ points in the parameter space $\mathcal{P} = [0,0.2]$, in order to properly detect the critical point and follow the post-buckling branch. We applied for all the test cases a POD tolerance $\varepsilon_{POD} = 10^{-8}$ and an online continuation method based on $K = 2000$ equispaced points in $\mathcal{P}$, which corresponds to a continuation step $\Delta\mu = 10^{-4}$.

Let us now consider the reduced order strategy for the SVK model with null body force $B = (0,0)$. We obtained a reduced basis space of dimension $N = 5$, that we used for the online projection in order to reconstruct the bifurcation diagram in Figure 6.1a. The functional considered in this case is the infinite norm of the second component of the displacement, namely $s(u) = \|u_y\|_\infty$. We can clearly observe that the buckling of the beam occurs for the value $\mu^* \approx 0.03$. In fact, in such point the vertical component

of the displacement u changes suddenly from being trivial to causing the buckling. Furthermore, we want to highlight the sharp discontinuity in the sensitivity $\frac{\partial u(\mu)}{\partial \mu}$ at the buckling point.



(a) $B = (0, 0)$ (b) $B = (0, -1000)$

Figure 6.1: Reduced basis bifurcation diagrams for the SVK beam with different body forces.

A representative solution of the post-buckling branch is depicted in Figure 6.2 for $\mu = 0.2$ with respect to the original undeformed configuration (mesh wireframe). We remark that here the beam buckles upwards.

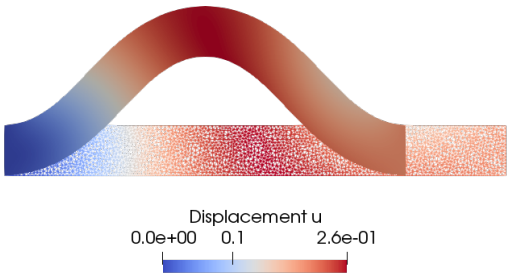
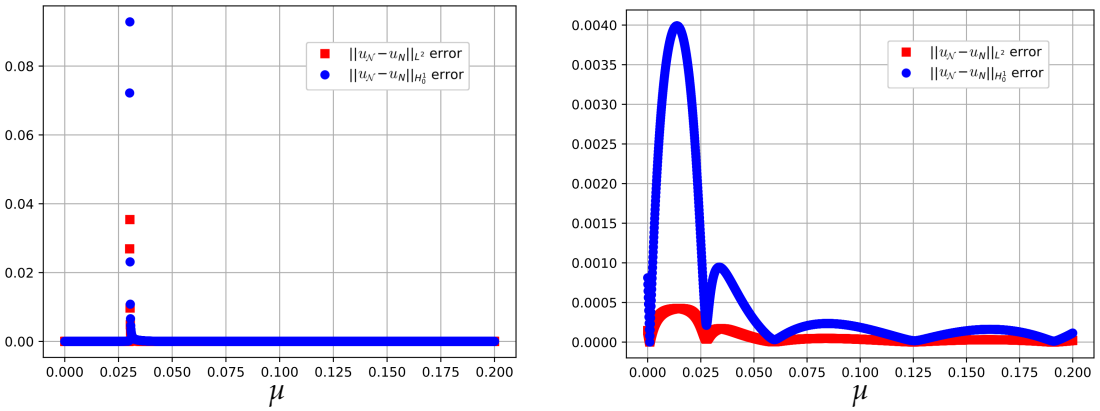


Figure 6.2: High fidelity displacement u for the SVK beam with $B = (0, 0)$ at $\mu = 0.2$.

As we can see from Figure 6.3a, the POD with only $N = 5$ basis functions is able to recover a good approximation of the bifurcation diagram, with an average error on $\mathcal{P}$ of order 10^{-6}. Despite this, we can observe also here, even more clearly, that the RB error has its maximum at the buckling point. Moreover, here the difference between the maximum and average RB error is of almost 4 orders of magnitude, denoting the great difficulty of the reduced manifold at reproducing the bifurcation. We also want

to remark that, when Dirichlet BCs are considered in the PDE, a lifting function has to be added to the POD basis, in order to encode the value at the boundary.

The speed-up, equals to 1.26, is quite low, in fact to plot the high fidelity version of the bifurcation diagram in Figure 6.1 we spent $t_{HF} = 454$(s) while the reduced order one required $t_{RB} = 358$(s). This is due to the fact that: (1) the number of degrees of freedom is kept low by the linear elements, (2) empirical interpolation strategies were not applied, (3) a $\mathcal{N}$-independent output functional could have been chosen.



(a) $B = (0, 0)$ (b) $B = (0, -1000)$

Figure 6.3: Reduced basis errors for the SVK beam with respect to different body forces.

Now we want to study how a perpendicular body force acting on the beam change the solution properties. For this reason we consider a non-trivial $B = (0, -1000)$ and within the same reduced setting we obtained $N = 4$ reduced basis dimension upon which we built the buckling diagram in Figure 6.1b. Is it immediate to observe that the sharp discontinuity in the sensitivity was a bit smoothed by the force B which therefore produces a more gradually buckling behaviour. Moreover, as it is possible to observe from Figure 6.4, the action of a "gravity"-like force causes the branching behaviour to be characterized by a downwards buckling. Hence, the reflective symmetry has been broken by imposing a gravitational body force and the bifurcation diagram behaves more smoothly near the buckling point. Furthermore, due to the elastic properties chosen, the body force is only able to qualitatively change the buckling without consistently affecting the quantitative displacement.

The reduced error plot in Figure 6.3b shows an effect of this smoothness, in fact now the maximum and average error are of the same order. We also notice that here the error is maximum in the region of the unbuckled states. Moreover, we remark that the speed-up is the same as the unforced case. Finally, we highlight that in general a regularization action for the bifurcation diagram could be the key for a good RB strategy, when only a qualitative understanding on the system is required.

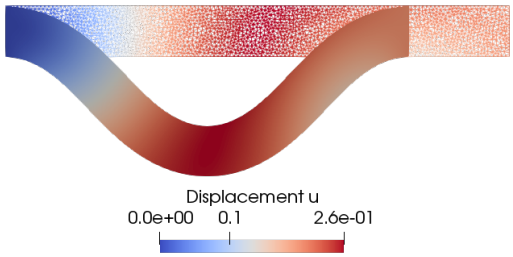
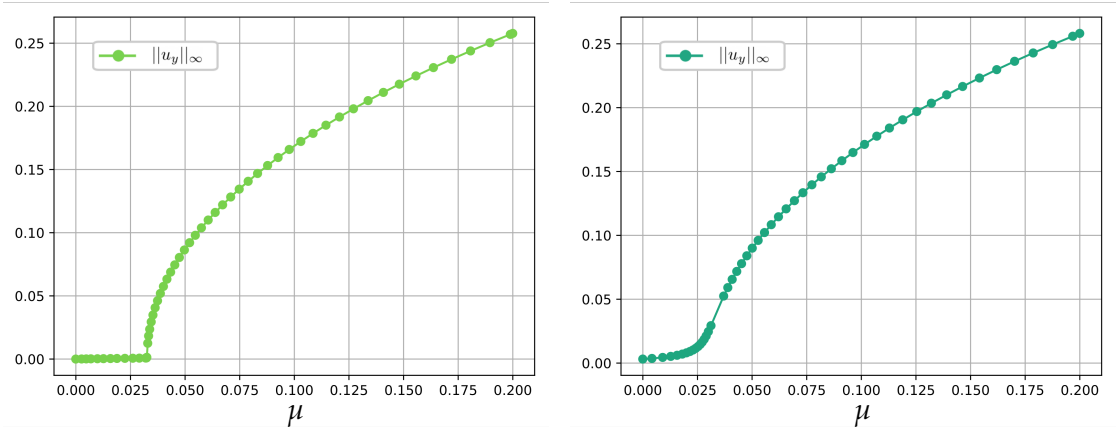


Figure 6.4: High fidelity displacement u for the SVK beam with $B = (0, -1000)$ at $\mu = 0.2$.

We can now move to the investigation of the neo-Hookean constitutive relation in (6.6). Using the same offline setting, we obtained a reduced basis space of dimension $N = 4$ for both body forces test cases. We can show the reduced basis bifurcation diagrams in Figure 6.5, from which it can be observed the same smoothing effect of the body force near the buckling. Despite it is well-known from the literature that the neo-Hookean model is more accurate far from the small displacement regime, we can notice that the detection of buckling point is consistent with the one predicted by the Saint Venant-Kirchhoff model. The results do not differ too much form the previous ones, only a small increment of the maximum displacement can be observed in Figure 6.5.



(a) $B = (0, 0)$ (b) $B = (0, -1000)$

Figure 6.5: Reduced basis bifurcation diagrams for the NH beam with different body forces.

Representative reduced error plots on the displacement of the post-buckling branch for both $B = (0, 0)$ and $B = (0, -1000)$ case are depicted in Figures 6.6 and 6.7 for

$\mu = 0.2$, showing the good approximation accuracy reached by the POD basis for the post-buckling branch.

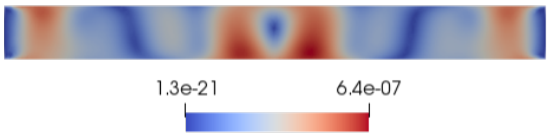


Figure 6.6: Reduced basis error of the displacement u for the NH beam with $B = (0, 0)$ at $\mu = 0.2$.

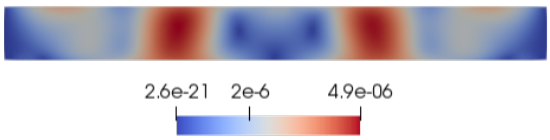
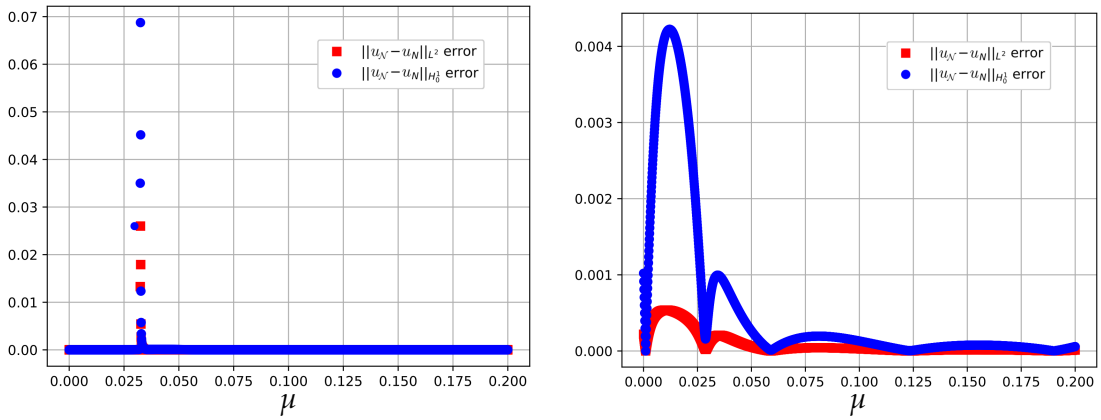


Figure 6.7: Reduced basis error of the displacement u for the NH beam with $B = (0, -1000)$ at $\mu = 0.2$.

As before, we present the plots of the reduced basis error with respect to each μ in the parameter domain $\mathcal{P}$ in Figure 6.8. We can draw the same conclusions about the accuracy of the reduction strategy when both body forces were applied.



(a) $B = (0, 0)$ (b) $B = (0, -1000)$

Figure 6.8: Reduced basis errors for the NH beam with respect to different body forces.

The speed-ups are similar in both forced and unforced cases, of order 1.22, due to the same reasons. The complexity of the model makes the computation of the bifurcation diagrams more expansive, in fact we spent almost $t_{HF} = 763$(s) for the

high fidelity version and $t_{RB} = 624$(s) for the reduced order one.

Before ending the analysis of this test case, we remark that although we were only interested in the first buckling, as we said previously many configurations can coexist for the same values of the compression parameter μ. As an example, in Figures 6.9 and 6.10 we plot the high fidelity displacements u at $\mu = 0.2$, corresponding to the 0-th and 2-th buckling modes.

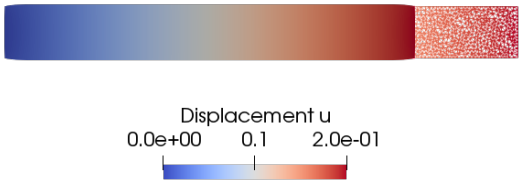


Figure 6.9: High fidelity 0-th mode displacement u for the SVK beam with $B = (0, 0)$ at $\mu = 0.2$.

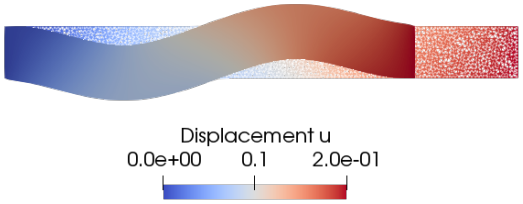


Figure 6.10: High fidelity 2-th mode displacement u for the SVK beam with $B = (0, 0)$ at $\mu = 0.2$.

6.2.1.2 Neumann compression

In this section, we will consider a different type of compression, in which we are no longer fixing the displacement at the boundary, rather we parametrize a traction force T. In literature this is usually called the cantilever beam test case.

In particular, we want to investigate the Saint Venant-Kirchhoff model with a Young modulus $E = 10^6$ and a Poisson ratio $\nu = 0.3$, with the trivial body force $B = (0, 0)$. Hence, the compression is modelled through the traction $T = (-\mu, 0)$ on $\Gamma_N^r = \{1\} \times [0, 0.1]$, while the beam is clamped, homogeneous Dirichlet condition $u = (0, 0)$, at the opposite edge $\Gamma_D^l = \{0\} \times [0, 0.1]$.

Of course, given the changed boundary conditions, we expect to obtain a different buckled configuration, since now the beam, thanks to the Neumann condition, has more freedom to move. As concerns the high fidelity setting, here we were able

to reconstruct the post-buckling branch with much fewer points, indeed we chose $N_{train} = 50$ equispaced points in the parameter space $\mathcal{P} = [2200, 2400]$ which represent the magnitude of the traction force T. For this model we experimented also the POD approximation accuracy with respect to the number of basis functions involved. Hence, we considered two reduced models, the first with tolerance $\varepsilon_{POD} = 10^{-8}$ corresponding to a reduced basis space dimension $N_1 = 4$, and the second one with tolerance $\varepsilon_{POD} = 10^{-10}$ corresponding to a reduced basis space dimension $N_2 = 5$. For the online phase we built the branch over $K = 400$ equispaced points in $\mathcal{P}$, which corresponds to a continuation step $\Delta\mu = 5 \cdot 10^{-1}$.

We considered again the functional $s(u) = \|u_y\|_\infty$ to plot the reduced basis bifurcation diagram in Figure 6.11. Since the body force is trivial, consistently with the previous analysis we have a sharp sensitivity near the buckling point $\mu^* \approx 2267$.

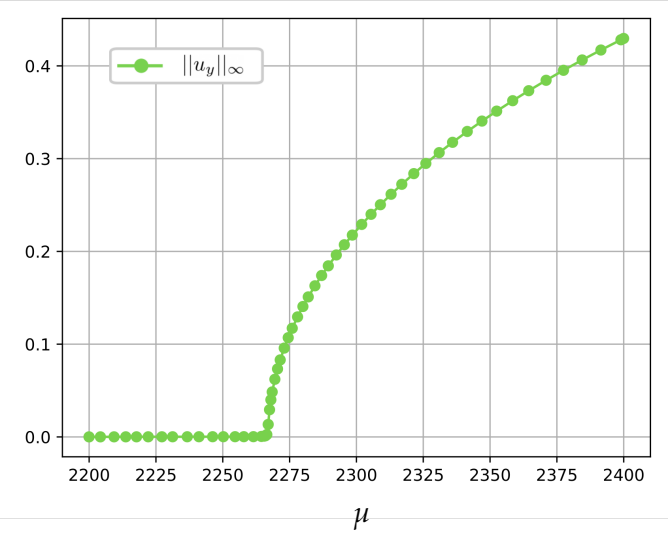


Figure 6.11: Reduced basis bifurcation diagrams for the SVK beam with Neumann compression and $B = (0, 0)$.

As we expected, the displacement solution shows an upwards buckling with a qualitative different behaviour with respect to the Dirichlet compression test case. We plot in Figure 6.12 a representative solution of the post-buckling branch for the Neumann compression at $\mu = 0.2$.

Now we come back to the comparison between the two ROMs, employing respectively $N_1 = 4$ and $N_2 = 5$ basis. As we can see from Figure 6.13, the POD of dimension $N_1 = 4$ (left) produces a not completely satisfactory reduced error at the buckling point μ^* of order $2.e-1$, although it corresponds to a quite low tolerance $\varepsilon_{POD} = 10^{-8}$. Due to the expected exponential convergence of the RB method, adding a single basis function, selected by the tolerance $\varepsilon_{POD} = 10^{-10}$, the maximum error diminishes of almost three orders. Of course, in both cases, the average errors of order below 10^{-7} on $\mathcal{P}$ confirm the good approximation property of the POD. Once again, the peaks of the

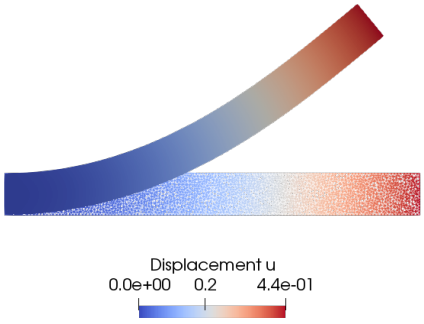
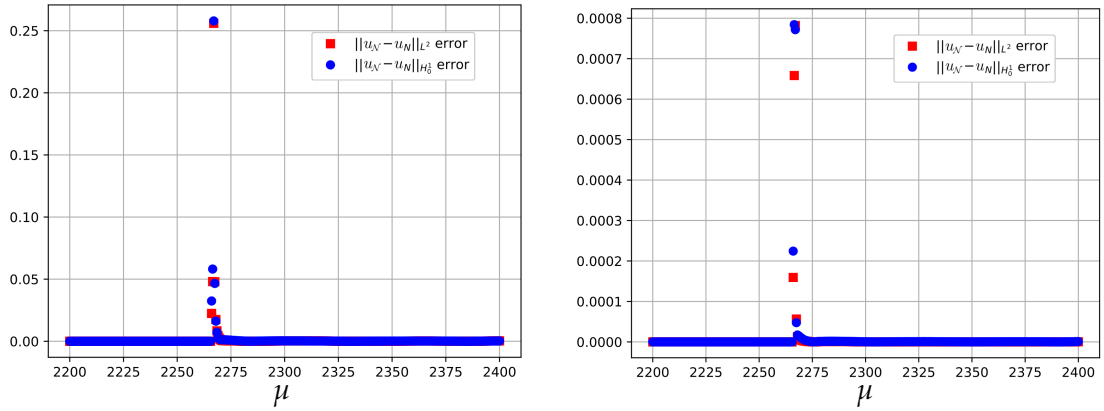


Figure 6.12: High fidelity displacement u for the SVK beam with Neumann compression and $B = (0, 0)$ at $\mu = 0.2$.

error are reached where the sensitivity loses its differentiability.

The speed-up is still very low, almost 1.2, but also the computation of high fidelity version of the bifurcation diagram in Figure 6.11 was less costly, indeed to build it we spent $t_{HF} = 180$(s), while the reduced order one required $t_{RB} = 150$(s).



(a) $N_1 = 4$ (b) $N_2 = 5$

Figure 6.13: Reduced basis errors for the SVK beam with a different number of basis functions.

6.2.1.3 Multi-parameter test case

Here, we want to extend the investigation done in Section 6.2.1.1 by modelling different materials through the parametrization of the elasticity constants. In particular, staying within the physical parametrization (we will see in the next section an example of the geometrical one), we want to understand how different materials behave with respect to an increasing compression imposed by means of Dirichlet boundary condi-

tion. Hence, we can define the multi-parameter $\boldsymbol{\mu} \in \mathcal{P} \subset \mathbb{R}^3$ as the triplet $\boldsymbol{\mu} = (\mu, E, \nu)$, where μ is the bifurcating parameter which controls the Dirichlet compression and E, ν are respectively the Young modulus and the Poisson ratio.

Once again, we considered the SVK model with gravitational body force $B = (0, -1000)$, and $s(u) = \|u_y\|_\infty$ as the output functional for the buckling detection.

For the analysis, we chose a parameter space given by $\mathcal{P} = [0.0, 0.2] \times [10^5, 10^7] \times [0.25, 0.42]$. As concerns the offline phase, we sampled $N_{train} = 1000$ snapshots for equispaced values of μ, for each one of the 4 vertexes of the bidimensional parameter space of the elasticity constants, i.e. $[10^5, 10^7] \times [0.25, 0.42]$.

Hence, as in the multi-parameter test cases in the previous chapters, also here we have sampled the bifurcation diagrams for different physical configuration and then we have used the POD compression to extract the modes needed to recover the buckling of beams made by different materials.

Since the 3-D parameter space, we expected that a much greater number of basis functions are needed, indeed using a POD tolerance $\varepsilon_{POD} = 10^{-8}$ we obtained a reduced basis space of dimension $N = 43$. The online continuation method was based on $K = 201$ equispaced μ values in $\mathcal{P}$, which corresponds to a continuation step $\Delta\mu = 10^{-3}$, and the bifurcation diagram was depicted for 5 random pairs of the elasticity constants $(E, \nu) \in [10^5, 10^7] \times [0.25, 0.42]$. Therefore, let us show in Figure 6.14 the reduced basis bifurcation diagrams with respect to the Young modulus.

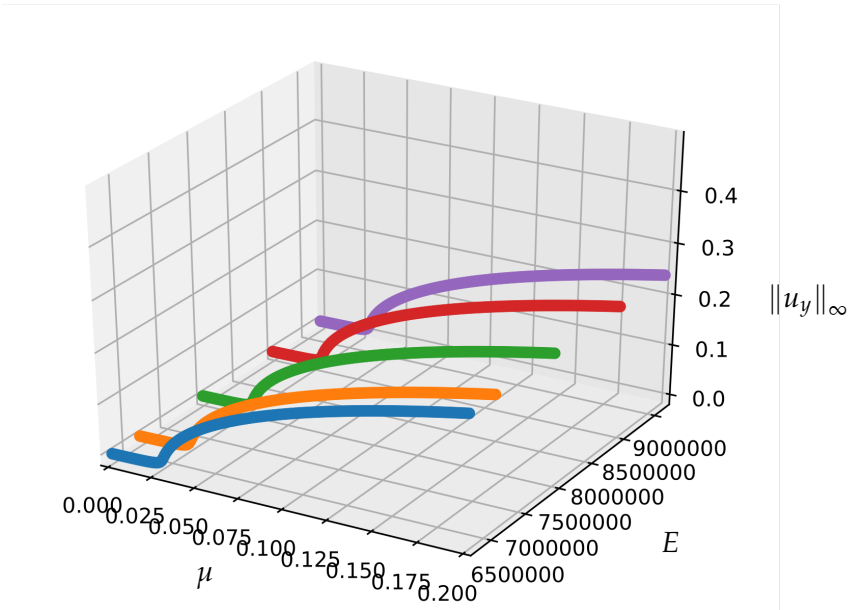
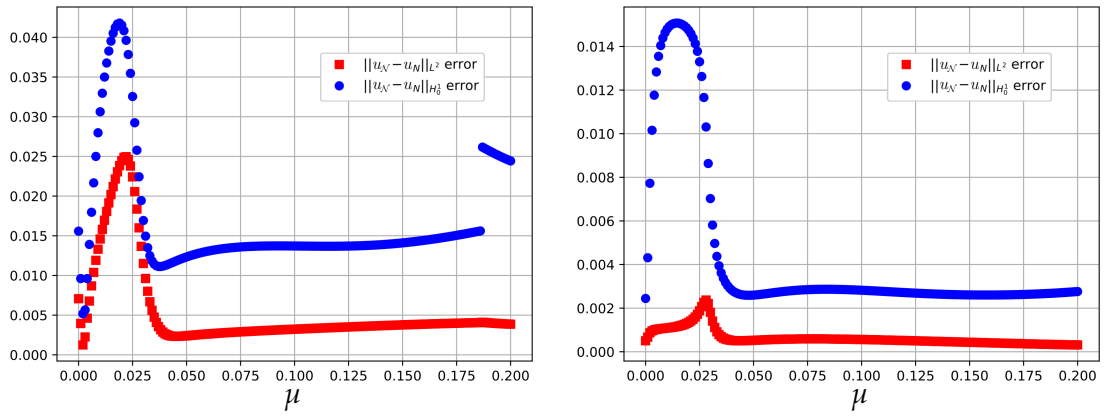


Figure 6.14: 3D reduced bifurcation plot for SVK beam with $B = (0, -1000)$ for five random pairs $(E, \nu) \in [10^5, 10^7] \times [0.25, 0.42]$.

As we can see the reduced model was able to reconstruct the post-buckling be-

haviour for all the pairs. As an example we show in Figure 6.15 the reduced errors for two other pairs $(E_1, \nu_1) = (5.2 \cdot 10^5, 0.352)$ and $(E_2, \nu_2) = (2.6 \cdot 10^6, 0.272)$. We remark that the choice of the non-trivial body force was takes in order to avoid sharp gradients in the bifurcation diagrams and thus larger errors in the reduced approximation.



(a) $(E_1, \nu_1) = (5.2 \cdot 10^5, 0.352)$ (b) $(E_2, \nu_2) = (2.6 \cdot 10^6, 0.272)$

Figure 6.15: Reduced basis errors for the SVK beam with $B = (0, -1000)$ for fixed pairs (E, ν).

Despite the displacement solution for the materials considered during the online phase shows similar properties in the considered parameter range, we can observe from Figure 6.16 that a slight modification of the buckling point can occur. Moreover, we remark that for some choices of the elasticity constants, the body force can become irrelevant leading again to sharp discontinuity in the sensitivity.

As regard the computational speed-up, given the increased dimension of the reduced space, the plot of the reduced diagram costs exactly as it high fidelity version $t_{HF} \approx t_{RB} = 900$(s). Hence, the need for an empirical interpolation approach capable of decoupling the online phase from the high fidelity degrees of freedom is still more evident. Despite this, the interesting point here is that, even through a naive approach for the sampling of the elasticity parameters, we were able to detect the buckling and the related post-buckling behaviour for a wide range of a 3-D parameter space, by means of $N = 43$ basis functions.

6.2.1.4 Geometrical parametrization

In this test case, we will consider for the first time a parametrized geometry, trying to understand how this influences the buckling property of the beam. Thus, let us gives brief introduction to the geometrical parametrization test case [130], which we will use also in subsequent sections.

Given a parameter $\boldsymbol{\mu} \in \mathcal{P}$, we can distinguish between the physical quantities $\boldsymbol{\mu}_p \in$

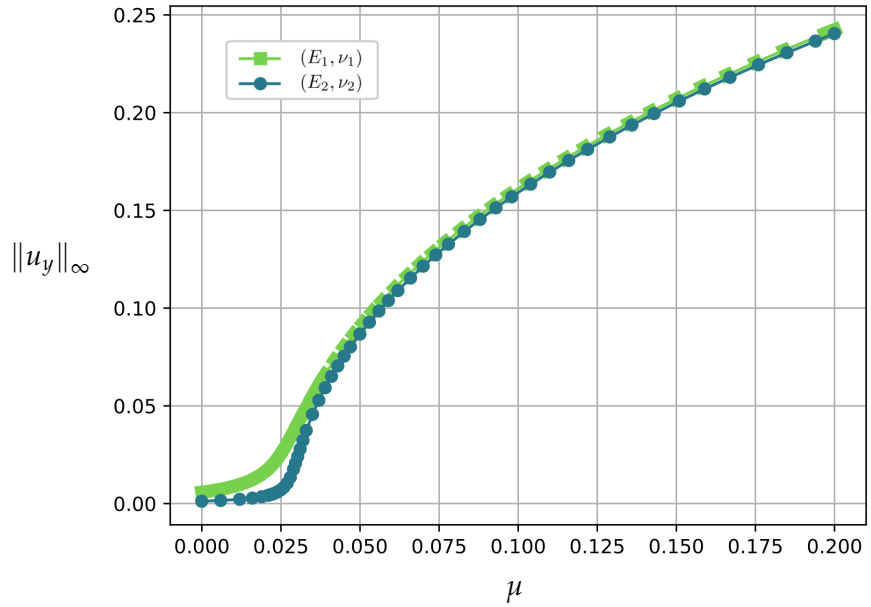


Figure 6.16: Reduced bifurcation diagram for SVK beam with $B = (0, -1000)$ for $(E_1, \nu_1) = (5.2 \cdot 10^5, 0.352)$ and $(E_2, \nu_2) = (2.6 \cdot 10^6, 0.272)$.

$\mathcal{P}_p$ (compression, forces, viscosity, trap strength) and geometrical quantities $\boldsymbol{\mu}_g \in \mathcal{P}_g$ (lengths, angles). Hence, we now consider the parameter dependent domain $\widetilde{\Omega}(\boldsymbol{\mu}_g)$ and its reference configuration Ω, related by the transformation map

$$\Phi : \Omega \times \mathcal{P}_g \to \mathbb{R}^d \quad \text{such that} \quad \widetilde{\Omega}(\boldsymbol{\mu}_g) = \Phi(\Omega; \boldsymbol{\mu}_g), \quad \forall \boldsymbol{\mu}_g \in \mathcal{P}_g. \tag{6.19}$$

The aim of such map is allow us to write a generic weak formulation posed on the parameter dependent domain $\widetilde{\Omega}(\boldsymbol{\mu}_g)$, in the reference configuration Ω, in order to guarantee the assembly of the high fidelity $\boldsymbol{\mu}$-independent quantities during the offline phase.

The key point is the formula for the change of variables, which for any integrable functions $\widetilde{f} : \widetilde{\Omega} \to \mathbb{R}^d$ is given by

$$\int_{\widetilde{\Omega}(\boldsymbol{\mu}_g)} \widetilde{f} \, d\widetilde{\Omega} = \int_\Omega f \det(\mathsf{J}_\Phi) \, d\Omega, \tag{6.20}$$

where $f = \widetilde{f} \circ \Phi$ and J_Φ is the Jacobian of the transformation map Φ. When the forms involve spatial derivatives, one relies on the chain rule to obtain a formula that encodes the parametrization dependence. As an example we can consider the standard H^1 seminorm, which can be transformed as

$$\int_{\widetilde{\Omega}(\boldsymbol{\mu}_g)} \nabla_{\tilde{x}} \widetilde{f} : \nabla_{\tilde{x}} \widetilde{h} \, d\widetilde{\Omega} = \int_\Omega (\nabla_x f) \mathsf{K} : \nabla_x h \, d\Omega, \tag{6.21}$$

where the parametrized tensor $\mathsf{K} : \mathbb{R}^d \times \mathcal{P} \to \mathbb{R}^{d\times d}$ is defined as

$$\mathsf{K}(x;\boldsymbol{\mu}) = \mathsf{J}_\Phi^{-1}(x;\boldsymbol{\mu})\mathsf{J}_\Phi^{-T}(x;\boldsymbol{\mu})\det(\mathsf{J}_\Phi(x;\boldsymbol{\mu})).$$

Having presented the basic notions on geometrical parametrization, we can go back to the our 2-D toy problem to investigate how the length of the beam is related to the buckling point. For this reason we now consider the geometrically parametrized beam, where its semi-length μ_g is added as new parameter. Therefore, the domain depicted in Figure 6.17 can be expressed as $\widetilde{\Omega}(\mu_g) = \widetilde{\Omega}_1 \cup \widetilde{\Omega}_2(\mu_g)$, where $\widetilde{\Omega}_1 = [0, 0.5] \times [0, 0.1]$ and $\widetilde{\Omega}_2(\mu_g) = [0.5, 0.5 + \mu_g] \times [0, 0.1]$. In this case the transformation map is simply given by the affine function

$$\Phi(x;\mu_g) = \begin{bmatrix} 2\mu_g(x - 0.5) + 0.5 \\ y \end{bmatrix} \quad \text{for } x \in \Omega_2 = \widetilde{\Omega}_2(0.5).$$

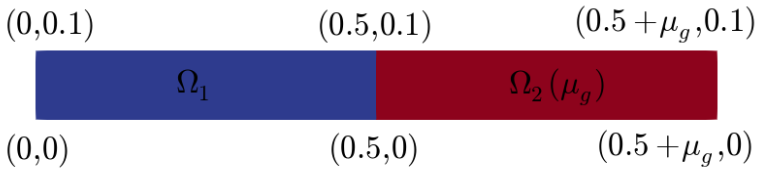


Figure 6.17: 2-D beam with parametrized geometry and $\mu_g \in [0.5, 1]$.

As a consequence of the geometrical parametrization, a consistent number of terms are involved in the affine (w.r.t. the parameter) decomposition of the weak formulation. As an example, for the SVK model one obtains the split of (6.14) in 3 linear and 5 nonlinear terms, already for this simple geometry change.

As regard the physical setting, we consider the SVK model compressed through Dirichlet boundary conditions with null traction force $T = 0$, Young modulus $E = 10^6$, Poisson ratio $\nu = 0.3$ and gravitational body force $B = (0, -1000)$. During the offline phase we sampled $N_{train} = 1000$ points in the parameter space $\mathcal{P}_p = [0, 0.2]$, for each one of the 3 equispaced points in $\mathcal{P}_g = [0.5, 1]$. Using, as always, a POD global compression with tolerance $\varepsilon_{POD} = 10^{-8}$ we obtained a reduced basis space of dimension $N = 12$. On the contrary, the online continuation method to reconstruct the 3-D bifurcation diagram in Figure 6.18 we chose $K = 201$ equispaced points in $\mathcal{P}_p$, for 5 equispaced values of the semi-length $\mu_g \in \mathcal{P}_g$.

Two representative solutions of the post-buckling branches for $\mu_g = 0.625$ and $\mu_g = 0.875$ are depicted in Figure 6.19, respectively left and right, at $\mu = 0.2$.

As we can see from Figure 6.20 the reduced manifold was able to approximate the buckling also for not sampled geometries with good approximation accuracy. This way, we can effectively study the evolution of the buckling varying the length of the beam.

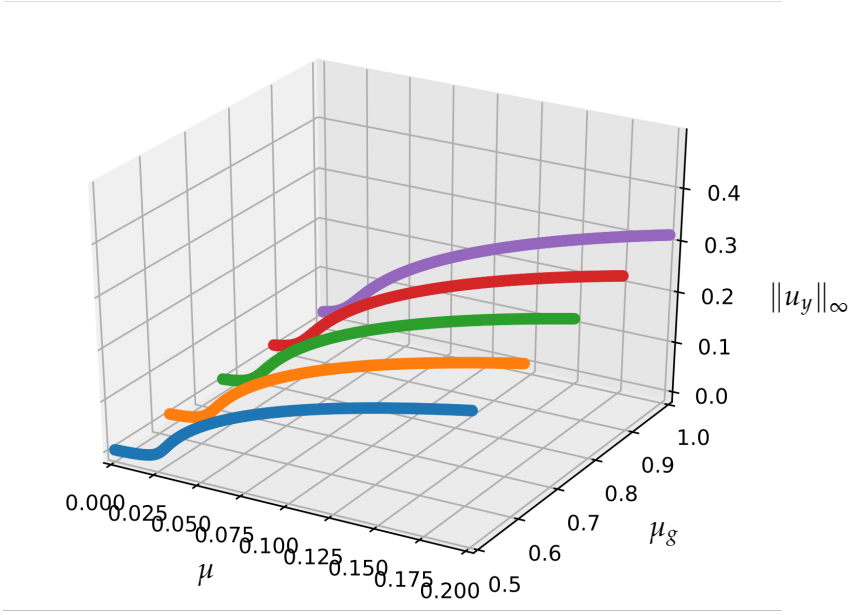


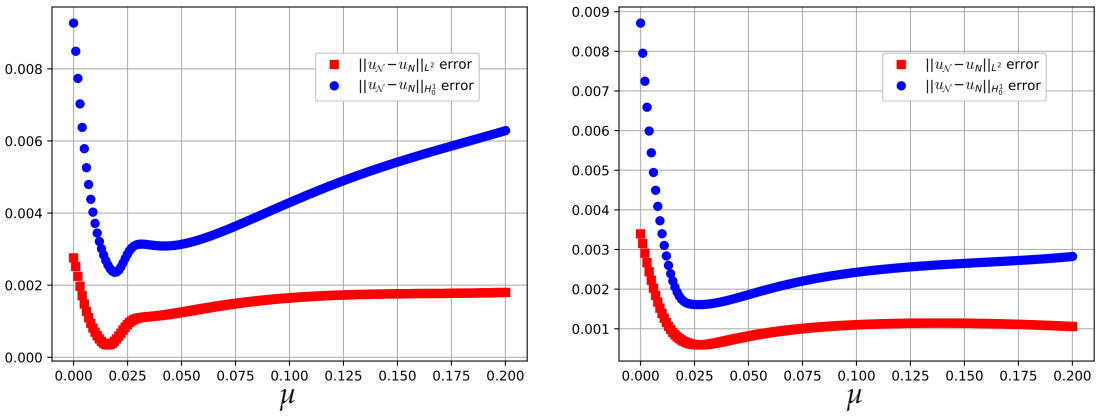
Figure 6.18: 3D bifurcation plot for SVK beam with $B = (0, -1000)$ with $\mu_g \in \mathcal{P}_g$.

(a) $\mu_g = 0.625$ (b) $\mu_g = 0.875$

Figure 6.19: High fidelity displacement u for the SVK beam with $B = (0, -1000)$ at $\mu = 0.2$ for different geometries.

As we expected when the length of the beam varies, also the buckling point changes its position. Indeed, as we can observe from Figure 6.21 the longer is the beam the sooner it buckles. This is still more evident in the 2-D projected diagram where the branches correspondent to $\mu_g = 0.625$ and $\mu_g = 0.875$ are plotted. Indeed, the buckling point for the beam with corresponding semi-length $\mu_g = 0.875$ buckles for a value of $\mu^* = 0.016$, while for the configuration with $\mu_g = 0.625$ it occurs at $\mu^* = 0.022$.

Also here the speed-up is essentially null with $t_{HF} \approx t_{RB} = 3000$(s), in fact the lower number of basis function is balanced by the increased number of terms involved in the weak formulation, which due to the lack of empirical interpolation strategies require a costly projection on the high fidelity dimension space.



(a) $\mu_g = 0.625$ (b) $\mu_g = 0.875$

Figure 6.20: Reduced basis errors for the SVK beam with $B = (0, -1000)$ for fixed geometries.

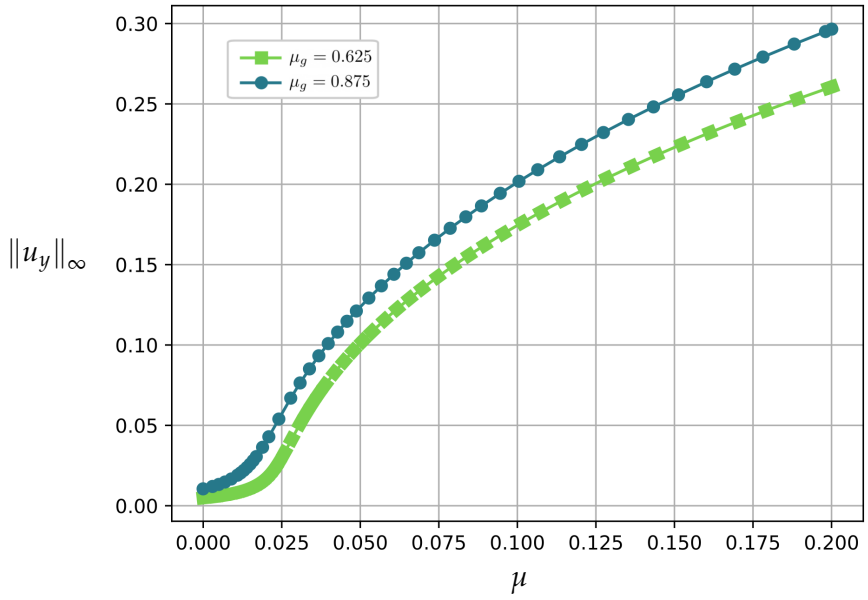


Figure 6.21: Reduced bifurcation diagram for the SVK beam with $B = (0, -1000)$ for $\mu_g = 0.625$ and $\mu_g = 0.875$.

6.2.2 3-D toy problem

Having analysed in the previous section a variety of settings for the buckling of a two-dimensional beam, the main step towards real applications is to consider a 3-D geometry, as an extension of the one studied previously. Following the test case investigated in [157], we fix the domain as $\Omega = [0, 1] \times [0, 0.2] \times [0, 0.079]$, we choose

the body forces as either $B = (0, 0, 0)$ or $B = (0, 0, -1000)$ with trivial traction $T = 0$, and fixed material properties $E = 10^6$ and $\nu = 0.3$.

Hence, we go back to the one parameter case, in which the compression parameter is imposed through Dirichlet boundary conditions. For this reason, we seek the solution displacement u in the functional space

$$\mathbb{X} = \{u \in (H^1(\Omega))^3 : u = (0, 0, 0) \text{ on } \Gamma_D^l,\ u = (-\mu, 0, 0) \text{ on } \Gamma_D^r\},$$

where $\Gamma_D^l = \{0\} \times [0, 0.2] \times [0, 0.079]$ and $\Gamma_D^r = \{1\} \times [0, 0.2] \times [0, 0.079]$.

We built the Finite Element space with $\mathbb{P}_1$ linear elements on a tetrahedral mesh, resulting in an high fidelity dimension $\mathcal{N} = 1734$.

As before the parameter space is given by $\mathcal{P} = [0.0, 0.2]$, and its exploration is performed through the simple continuation method with fixed step $\Delta\mu = 2 \cdot 10^{-4}$, resulting in $N_{train} = 1000$ snapshots. The reduced manifold was built choosing a tolerance $\varepsilon_{POD} = 10^{-8}$ and the bifurcation diagram is recovered with an online continuation method based on $K = 2000$ equispaced points in $\mathcal{P}$.

We remark that since here we added a space dimension, the buckling for the rectangular cross-section beam can happen in both the directions individuated by the cross-section of the beam perpendicular to the compression axis. Despite this, the first buckling usually occurs in the direction of minimum length, hence in this case the z-axis. For this reason, in order to detect the buckling behaviour, we consider as output functional the infinite norm of the z-component of the displacement $\|u_z\|_\infty$.

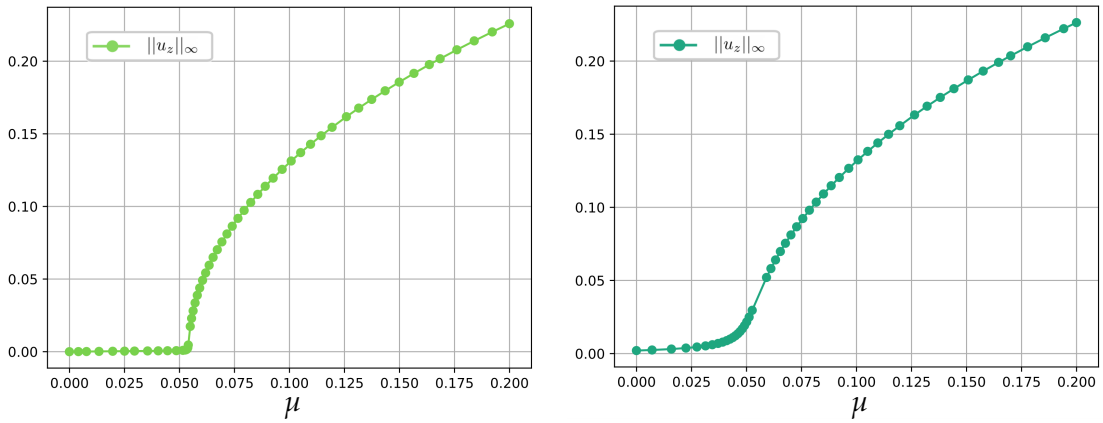
For the three dimensional beam, both constitutive relations, SVK and NH, have been investigated. As concerns the SVK model, we obtained a reduced basis space of dimension $N = 9$, for both the choices of the body force B.

In Figure 6.22 we can see the bifurcation plot for trivial $B = (0, 0, 0)$ and gravitational $B = (0, 0, -1000)$ body forces. As we can see, also in this case the sharp gradient in the sensitivity was smoothed by the external force, but in Figure 6.22a we can clearly observe that the buckling of the beam occurs for the value $\mu^* \approx 0.053$.

As predicted before the buckling occurs along the z-direction and a representative solution of the post-buckling branch is depicted in Figure 6.23 for $\mu = 0.2$ with respect to the original undeformed configuration (mesh wireframe).

As we can see from Figure 6.24, the reduced order model has overall good approximation properties, since it is capable to reconstruct the post-buckling behaviour with a maximum error of order 10^{-2}. We remark that while the peaks in correspondence to the bifurcation point agree with the previous analysis considering different body forces, in this case the maximum errors were obtained for a bigger value of the compression. We did not investigate further the buckling, but from the considerations about the performances of the RB at bifurcation points, we guess that other buckling points are located near $\mu = 0.122$ and $\mu = 0.166$.

Once again, we observe the consistent difference between the maximum error and the average one of order 10^{-8}, remarking that this is caused by the lack of regularity of the solution with respect to the parameter, due to the bifurcation phenomena. Same results as before hold for the speed-up.



(a) $B = (0, 0, 0)$ (b) $B = (0, 0, -1000)$

Figure 6.22: Reduced basis bifurcation diagrams for the 3D SVK beam with respect to different body forces.

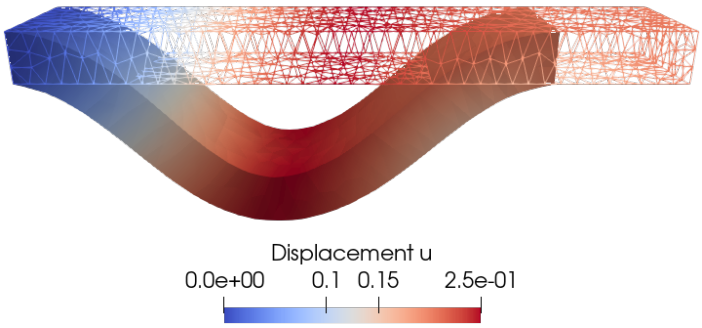
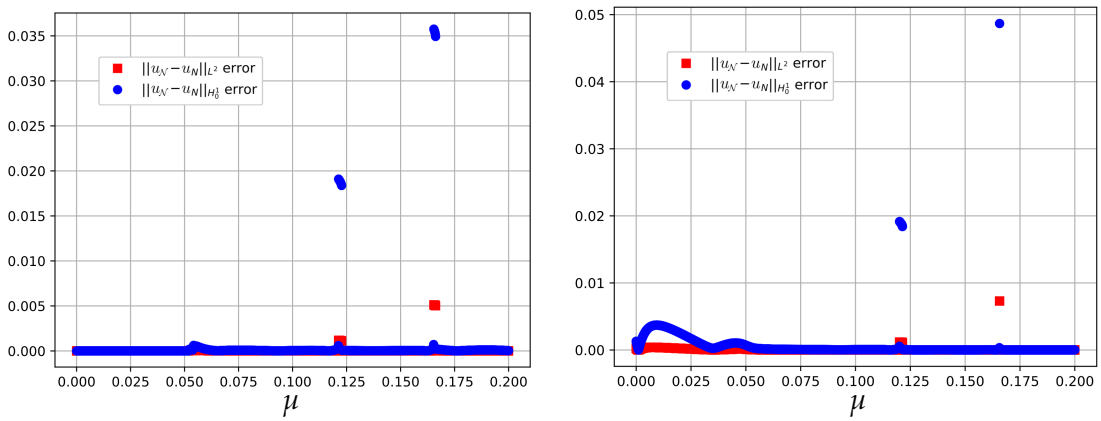


Figure 6.23: High fidelity displacement u for the 3D SVK beam with $B = (0, 0, 0)$ at $\mu = 0.2$.

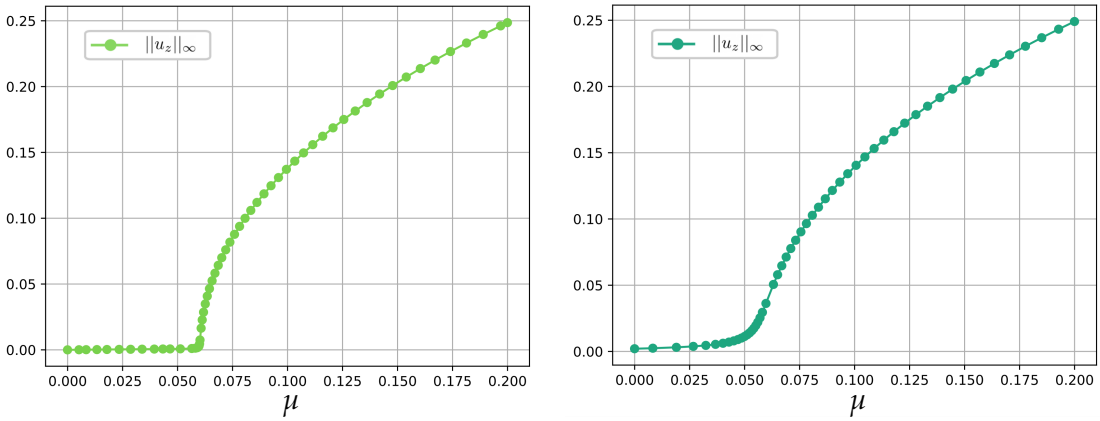
For the sake of comparison let us consider also the NH model within the same setting as before. The same POD tolerance was reached by a fewer number of basis function, namely $N = 5$. In Figure 6.25 we can see the bifurcation plot for trivial $B = (0, 0, 0)$ and gravitational $B = (0, 0, -1000)$ body forces. For the three dimensional beam we can observe that the NH constitutive relation actually predict the buckling in a slightly different location, indeed from Figure 6.25a we can detect the buckling occurring for the value $\mu^* \approx 0.059$ (compare with Figure 6.22a).

Furthermore, as we can see from Figure 6.26, the NH model did not encounter the same accuracy issues as for the SVK model in Figure 6.24, thus showing a better approximation over $\mathcal{P}$. As an example we show in Figure 6.27 the reduced error for the post-buckling displacement solution at $\mu = 0.2$ with gravitational body force.



(a) $B = (0, 0, 0)$ (b) $B = (0, 0, -1000)$

Figure 6.24: Reduced basis errors for the 3D SVK beam with respect to different body forces.



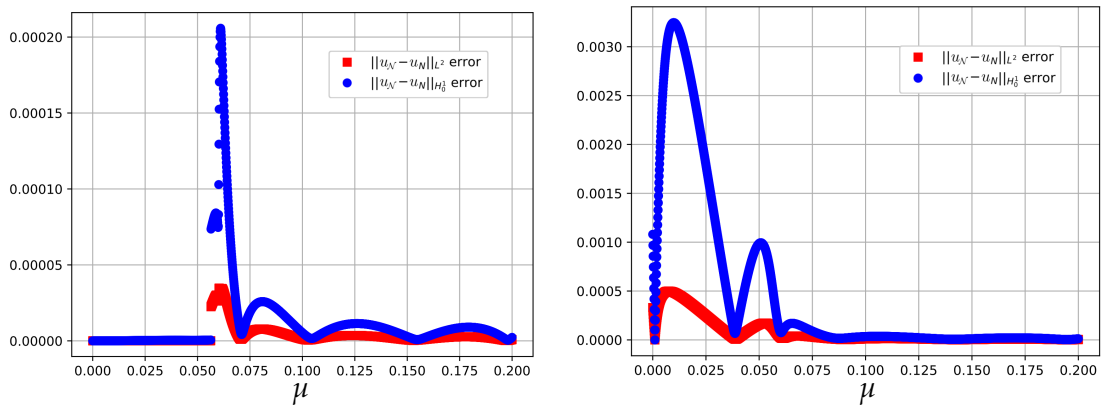
(a) $B = (0, 0, 0)$ (b) $B = (0, 0, -1000)$

Figure 6.25: Reduced basis bifurcation diagrams for the 3D NH beam with respect to different body forces.

Before ending this section we want to remark that as we have seen previously, also in this case the way in which we impose the compression and the boundary conditions chosen had a great influence on the buckling location. In fact, if we allow the right end of the beam to move in the two perpendicular direction to the compression (y and z axis), the resulting buckling mode and its corresponding branching point change consistently. In practice we fixed the functional space as

$$\mathbb{X} = \{u \in (H^1(\Omega))^3 : u = (0, 0, 0) \text{ on } \Gamma_D^l,\ u_x = -\mu \text{ on } \Gamma_D^r\},$$

and the parameter space as $\mathcal{P} = [0, 0.03]$.



(a) $B = (0, 0, 0)$ (b) $B = (0, 0, -1000)$

Figure 6.26: Reduced basis errors for the 3D NH beam with respect to different body forces.

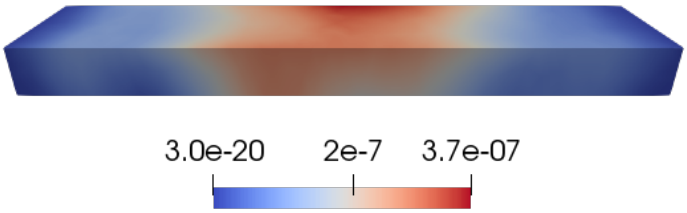


Figure 6.27: Reduced basis error plot of the displacement u for the 3D NH beam with $B = (0, 0, -1000)$ at $\mu = 0.2$.

In the same setting as before, choosing the SVK constitutive relation with the trivial body force we obtained the reduced bifurcation diagram in Figure 6.28, where we can see the buckling occurring at $\mu^* = 0.014$.

The new buckled state solution is depicted in Figure 6.29 for $\mu = 0.2$. As we can see from Figure 6.30, the reduced approximation follows the same behaviour of the previous case, with $N = 4$ basis functions, maximum error of order 10^{-2} and a difference of 4 orders of magnitude with the average one.

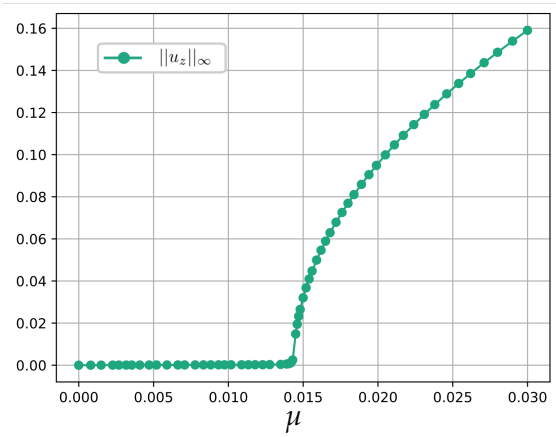


Figure 6.28: Reduced basis bifurcation diagram for the 3D SVK beam with $B = (0, 0, 0)$.

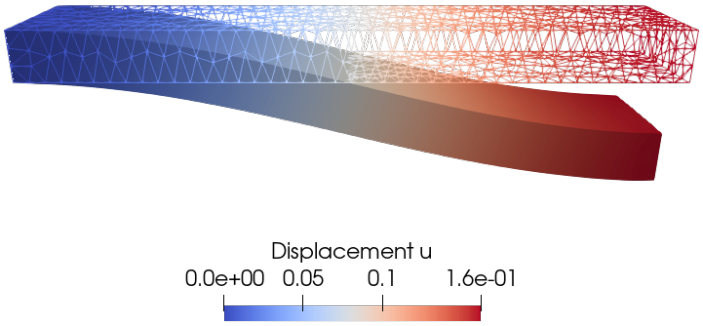


Figure 6.29: High fidelity displacement u for the 3D SVK beam with $B = (0, 0, 0)$ at $\mu = 0.03$.

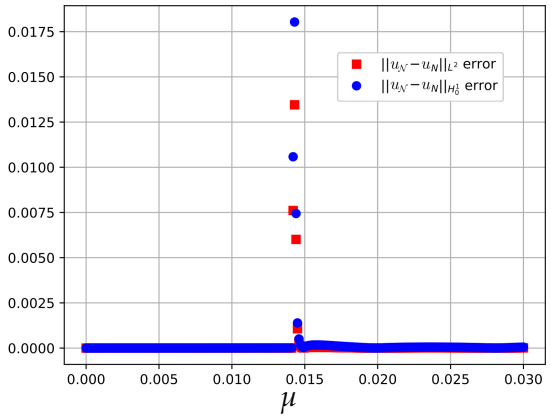


Figure 6.30: Reduced basis errors for the 3D SVK beam with $B = (0, 0, 0)$.

6.2.3 Industrial test case application to tubular element

Finally, we can introduce a real test case scenario which comes from the Norwegian petroleum industry [114]. The results obtained in the previous sections allowed us to investigate the following more complex problem. We want to investigate the deformation of a 3-D tubular geometry with annular section. It is clear that when dealing with real geometries the situation complicates, since many factors have to be taken into consideration. Here, for practical interest the domain is defined as a tubular member characterized by an annular cross section with inner and outer radii $r = 0.28$(m) and $R = 0.30$(m). Thus, the thickness corresponds to $t = 0.02$(m) and the outer diameter is $D = 0.6$(m). This is important since the condition on the ratio $D/t < 120$ has to be satisfied for the correct reconstruction of the physics at hand.

As before, different settings will be analysed in order to understand its buckling properties. Despite the great difficulties and some unexpected behaviour that such model exhibits, in the following we present a successful application of our reduced methodology.

6.2.3.1 A comparison of constitutive relations

In this section we fix the length of the tube as $L = 2$(m) and our focus will be on the behaviour of the displacement in the case of Neumann compression. Hence, we fix the domain as $\Omega = A_r^R \times [0, 2]$ where we denoted the annular section as

$$A_r^R = \{(x, y) \in \mathbb{R}^2 | r^2 \leq x^2 + y^2 \leq R^2\},$$

we choose a trivial body force $B = (0, 0, 0)$ and a compression given by the traction term $T = (0, 0, -\mu)$. We remark that now the compression is acting along the z-axis. As concerns the material properties we fixed $E = 2.1 \cdot 10^5$(MPa) and $\nu = 0.3$.

We start with the one parameter test case, where homogeneous Dirichlet boundary condition is imposed at $\Gamma_D = A_r^R \times \{0\}$ and $\Gamma_N = A_r^R \times \{2\}$.

We built the FE space by means of $\mathbb{P}_1$ linear elements on the tetrahedral mesh in Figure 6.31 with 22521 cells, while the resulting number of degrees of freedom is $\mathcal{N} = 23442$.

Given the complexity of the geometry and since our main interest is to find the first failure mode, here we will focus the investigation only up to the buckling, disregarding the post-buckling behaviour. Both constitutive relations, SVK and NH, have been analysed, producing qualitatively different results also concerning the position of the critical value. For these reasons we considered different parameter spaces for the two hyperelastic models.

For the SVK model the parameter space is given by $\mathcal{P} = [0, 4500]$ and $N_{train} = 200$ snapshots were computed through the continuation method. The reduced manifold was built choosing a tolerance $\varepsilon_{POD} = 10^{-8}$ and this led to a reduced basis space of dimension $N = 5$. The bifurcation diagram was recovered with an online continuation method based on $K = 451$ equispaced points in $\mathcal{P}$.

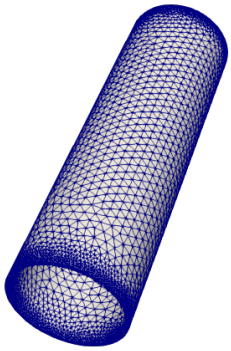


Figure 6.31: Mesh for the Norsok test case.

As concerns the NH material, we computed again $N_{train} = 200$ snapshots, but the parameter space now is defined as $\mathcal{P} = [0, 5000]$. The same reduced setting was applied, providing a reduced basis space of the same dimension.

Similarly to the previous 3-D case, we remark that given the symmetry w.r.t. the tube axis, here the buckling can occur in any direction perpendicular to axis itself. To detect the buckling behaviour, we consider as output functional the sum of the infinite norms of the x and y components of the displacement, namely $s(u) = \|u_x\|_\infty + \|u_y\|_\infty$.

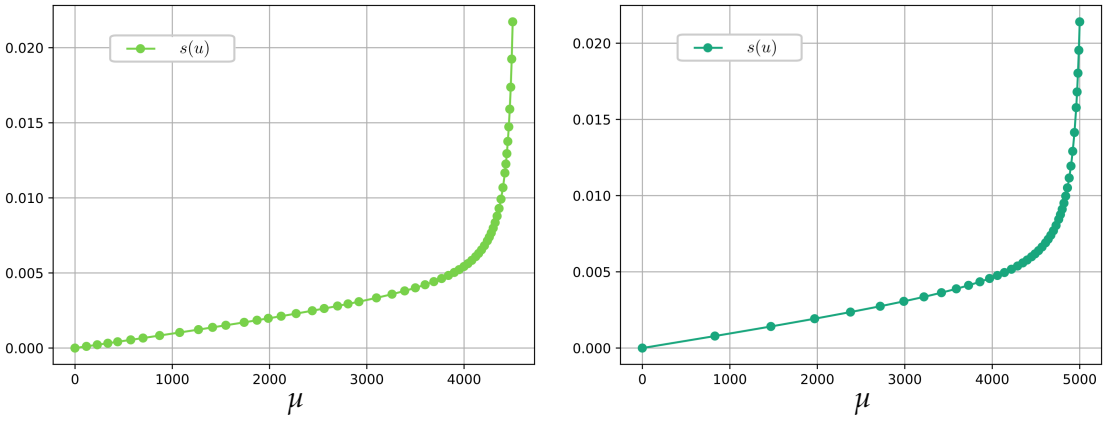
In Figures 6.32a and 6.32b we can see the bifurcation plot for the SVK and NH constitutive relations, respectively. We can clearly observe that the buckling of the beam with different models occurs at different values for the compression μ, in particular the buckling of the SVK beam is slightly anticipated.

As expected, the buckling is qualitatively similar to the one observed in Figure 6.12, with Neumann compression in the 2-D geometry. A representative solution of the buckling mode for the SVK model is depicted in Figure 6.33 for $\mu = 4500$.

As we can see from Figure 6.34, in both cases we were able to reach a good accuracy of the RB solution with respect to μ. Once again the error increases near the buckling point, here the right end of the parameter domains, where the solution changes more rapidly its behaviour. Regarding the computational time, we obtained a low speed-up of order 1.25, in fact to plot the high fidelity versions of the bifurcation diagrams in Figure 6.32b we spent $t_{HF} = 1335$(s) while the reduced order one required $t_{RB} = 1076$(s). Of course this is not satisfactory, for this reason, we now present an application of the hyper-reduction strategies to recover the efficiency.

In fact, especially when dealing with real test cases, a real time evaluation of the solution is a key feature. Thus, we consider the geometry Ω with Neumann compression and we try to efficiently recover the bifurcation diagrams in Figure 6.32 by means of the DEIM.

Within the same setting as before, we consider the DEIM with Greedy tolerance



(a) SVK (b) NH

Figure 6.32: Reduced basis bifurcation diagrams for the 3D tubular geometry with $B = (0, 0, 0)$ and different constitutive relations.

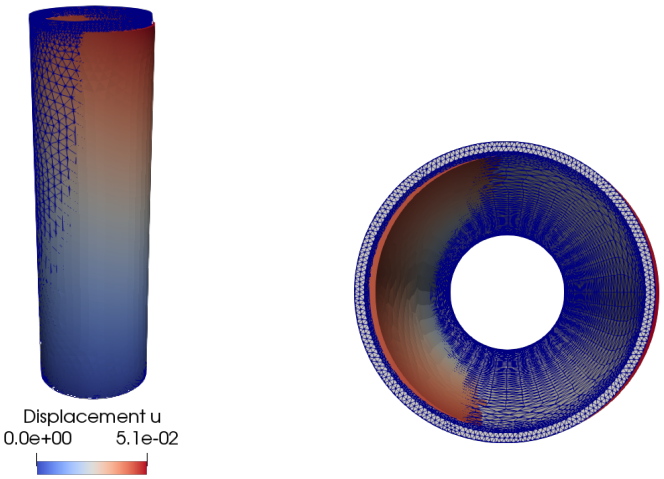
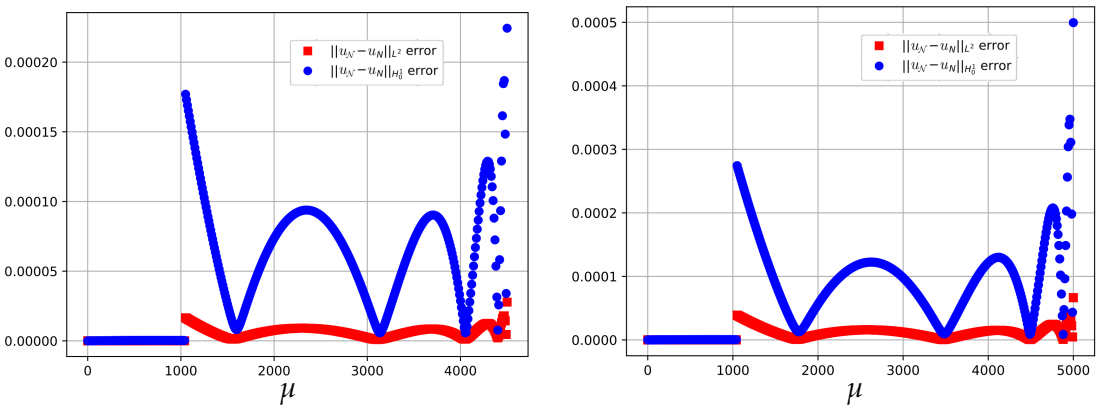


Figure 6.33: High fidelity displacement u for the 3D SVK tubular geometry with $B = (0, 0, 0)$ at $\mu = 4500$.

$\varepsilon_{Gr} = 10^{-10}$ which splits the form into an affine decomposition made up by 2 terms, each one approximated by 15 interpolation basis functions. The reason for such low tolerance is that the complexity of the model, together with its buckling behaviour, makes the approximation of the variational forms a difficult task. Indeed, we observed a non-convergence issue when during the online phase when higher tolerances were chosen.

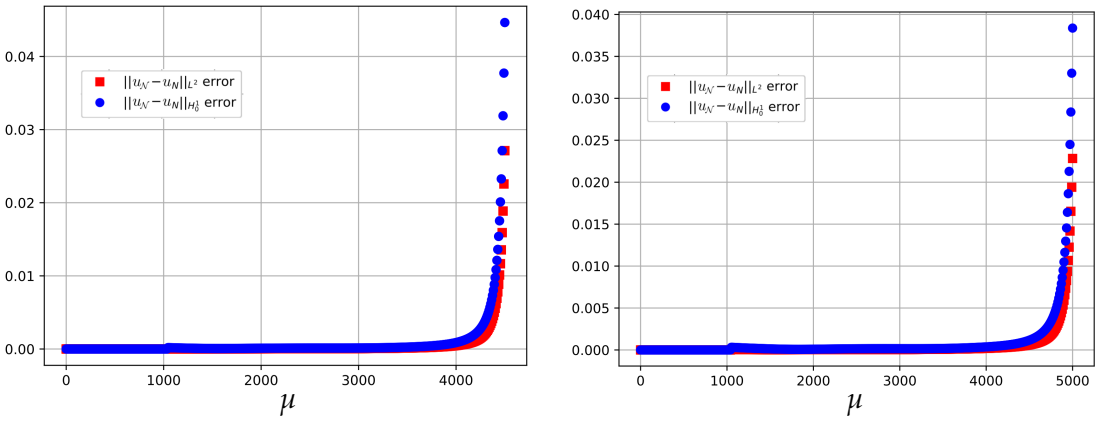
Hence, let us show in Figure 6.35 the reduced basis error, computed with DEIM for the SVK and NH constitutive relations. It is evident that a significant increment of both maximum errors, w.r.t. the straightforward application of Algorithm 2 without



(a) SVK (b) NH

Figure 6.34: Reduced basis errors for the 3D tubular geometry with $B = (0, 0, 0)$ and different constitutive relations.

interpolation strategies, occurs. Thus, the hyper-reduction approach seems to have some difficulties in approximating the forms near critical points. Despite this, the good news is that while the reduced basis error is still acceptable, the speed-up consistently increased. For the e.g. NH material we pass from a speed-up of order 1.25 with $t_{RB} = 1076$(s) to a speed-up of order 38 corresponding to $t_{RB,DEIM} = 35$(s).



(a) SVK (b) NH

Figure 6.35: Reduced basis errors with DEIM for the 3D tubular geometry with $B = (0, 0, 0)$ and different constitutive relations.

As we understood, usually one is not interested in the approximation of the bifurcating phenomenon for a fixed setting, rather the aim is the detection and reconstruction of the buckling modes varying physical or geometrical parameter. Since this is

true especially for real test cases, the latter will be analysed in the next section.

6.2.3.2 Multi-parametric study with varying geometries

Now we want to extend the previous analysis to the multi-parameter context, in which the additional parameter controls the length of the domain Ω. To do so, we recall the investigation done in Section 6.2.1.4. For this reason, we now consider the geometrically parametrized tube, where again its semi-length $\mu_g \in \mathcal{P}_g = [1,2]$ is taken into consideration. Therefore, we can express the 3-D parametrized domain as $\widetilde{\Omega}(\mu_g) = \widetilde{\Omega}_1 \cup \widetilde{\Omega}_2(\mu_g)$, where $\widetilde{\Omega}_1 = A_r^R \times [0,1]$ and $\widetilde{\Omega}_2(\mu_g) = A_r^R \times [1,\mu_g]$. Therefore, the transformation map is simply given by the affine function

$$\Phi(x;\mu_g) = \begin{bmatrix} x \\ y \\ \mu_g(z-1)+1 \end{bmatrix} \quad \text{for } x \in \Omega_2 = \widetilde{\Omega}_2(1).$$

The physical setting with Neumann BCs is the same as before, while for the offline phase we computed $N_{train} = 500$ snapshots. Performing a global POD compression with tolerance $\varepsilon_{POD} = 10^{-8}$ we obtained a reduced basis space of dimension $N = 9$. The online continuation method to reconstruct the 3-D bifurcation diagram in Figure 6.36 is based on $K = 350$ equispaced points in $\mathcal{P}_p = [0,4500]$, for 3 equispaced values of the semi-length $\mu_g \in \mathcal{P}_g$. We remark again that due to the large strains and the complexity of the phenomena, we focus on the solution behaviours only up to their buckling point. For these reasons, the actual parameter space varies for different branches, in fact we truncated it after having detected the buckling. The top-view of three representative solutions of the buckling modes for $\mu_g = \{1, 1.5, 2\}$ are depicted in Figure 6.37 for the last computed value of $\mu = \{4530, 3150, 2300\}$, respectively.

We present in Figure 6.38 the reduced basis error. We highlight that, as we can see from Figure 6.38a, also in this case the reduced manifold was able to approximate the buckling for unsampled geometries.

The study of the evolution of the buckling, varying the length of the beam, confirm as expected that longer beams need smaller forces to buckle. We highlight that since the importance of the buckling location, already during the offline stage, one can not perform an extensive analysis on the parametrized geometries. Moreover, to prevent non-convergence issues due to bigger strains one should consider a much refined mesh, increasing the high fidelity dimension.

Once again the speed-up equals to 1.4 is not that much relevant, with $t_{HF} = 2363$(s) and $t_{RB} = 1677$(s). Despite this, we remark that given the high dimensionality of the finite element space, especially for 3-D geometry with refined mesh and polynomial of order $P > 1$, the naive RB approach without empirical interpolation strategies can still provide consistent speed-up.

Up to now, we have highlighted the difficulties while modelling the buckling through the Neumann compression, in the next section the aim will be the investigation of the Dirichlet compression.

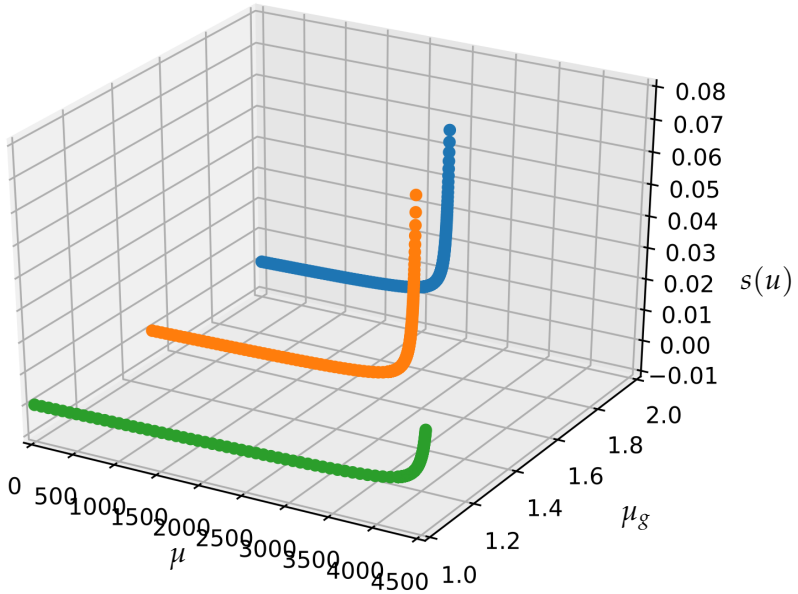
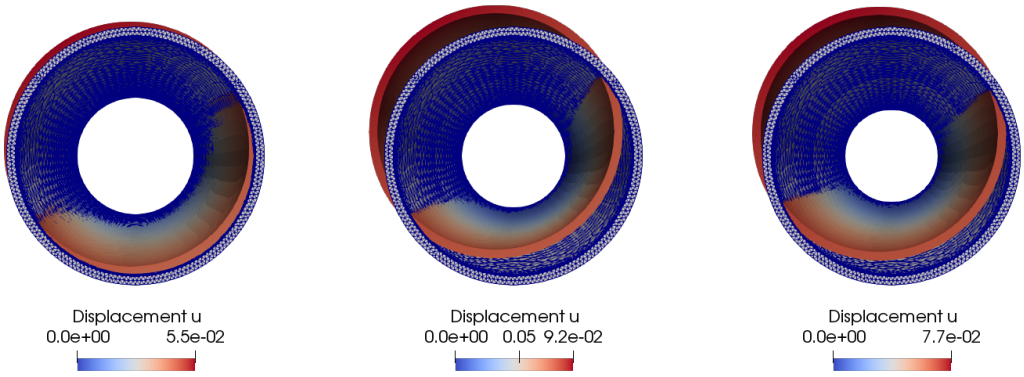


Figure 6.36: 3D bifurcation plot for SVK beam with $B = (0, 0, 0)$ and $\mu_g \in \mathcal{P}_g$.

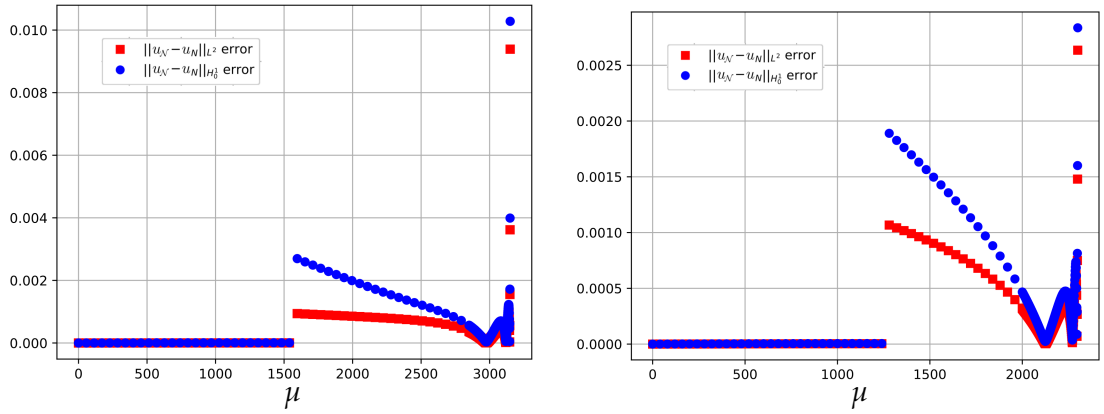


(a) $\mu_g = 1.$ (b) $\mu_g = 1.5$ (c) $\mu_g = 2.$

Figure 6.37: Top-view of the high fidelity displacement u for the SVK beam with $B = (0, 0, 0)$ for different geometries.

6.2.3.3 Dirichlet compression

Here, we want to extend the study of the Dirichlet compression for the tubular 3-D geometry in Figure 6.31. So far, we have understood that the choice of compression which has to be modelled is a complex task. Another confirmation of this comes from



(a) $\mu_g = 1.5$ (b) $\mu_g = 2.0$

Figure 6.38: Reduced Basis errors with respect to $\mu \in \mathcal{P}$ for the SVK beam with $B = (0,0,0)$ for different geometries.

the following scenario. In fact, when we tried to apply a Dirichlet compression over $\Gamma_D = A_r^R \times \{2\}$, we encountered many difficulties.

In particular, the cavity inside the member and the need for huge refinement of the mesh (connected to the remark done previously), made the reconstruction of the bifurcation diagram a too much difficult task. For these reasons, we were not able to fully recover the buckling within this compression context, but we only found the three different modes depicted in Figure 6.39.

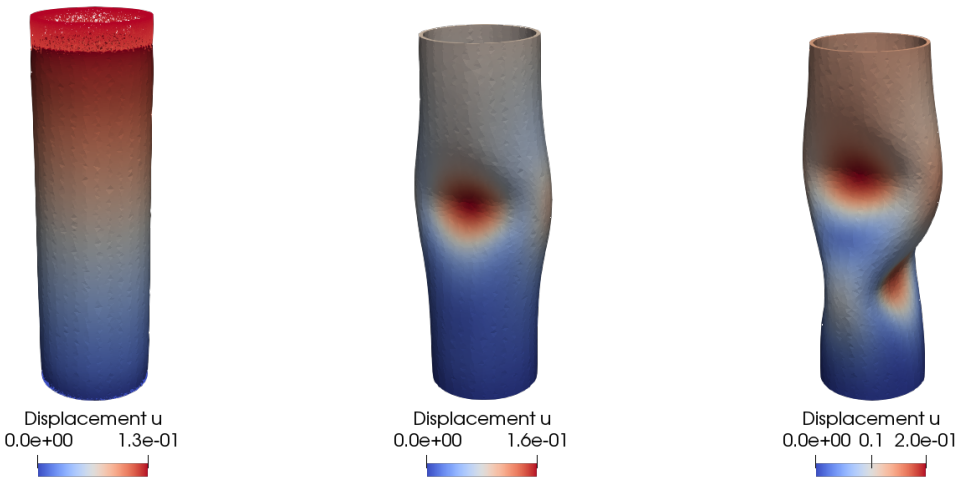


Figure 6.39: Representative solutions of the 3D SVK model for the tubular geometry with Dirichlet compression.

From the analysis of the parametrized geometries we understood that the length of

the tube plays a fundamental role in the buckling detection, thus we decided to investigate the tubular geometry with the same ratio, while considering a much increased length.

Therefore, let us consider the tube represented by the domain $\Omega = A_r^R \times [0, 20]$. We chose SVK material, trivial body and traction forces and we fixed the material properties $E = 2.1 \cdot 10^5$ and $\nu = 0.3$. We first consider the one parameter test case, in which the parameter controls the compression through Dirichlet BC on $\Gamma_D = A_r^R \times \{20\}$. We remark that in this case the tetrahedral mesh consists in 147133 cells, resulting in an high fidelity dimension $\mathcal{N} = 147852$ when $\mathbb{P}_1$ linear elements are used.

Here we fixed the parameter space as $\mathcal{P} = [0, 0.13]$, and its exploration is performed through the simple continuation method with step $\Delta\mu = 10^{-3}$, resulting in $N_{train} = 130$ snapshots. The reduced manifold was built choosing a tolerance $\varepsilon_{POD} = 10^{-8}$, which provides a reduced basis space of dimension $N = 5$.

In Figure 6.40 we can see the bifurcation plot for the SVK tube with trivial body force $B = (0, 0, 0)$. This is recognizable since also here the bifurcation phenomenon is characterized by a sharp gradient in the sensitivity. This has the effect of compromising the RB accuracy in the buckling point, as we can observe from Figure 6.41. Moreover, we can clearly observe that the buckling of the tube occurs for the value $\mu^* = 0.082$. Despite this, the reduced order model is able to reconstruct the post-buckling behaviour with a maximum error of order 10^{-3} and an average one of order 10^{-5}. As we remarked earlier the huge number of cells produces a slightly better speed-up of order 2, which is far from being a good result but highlights how much the empirical interpolation strategies could be useful.

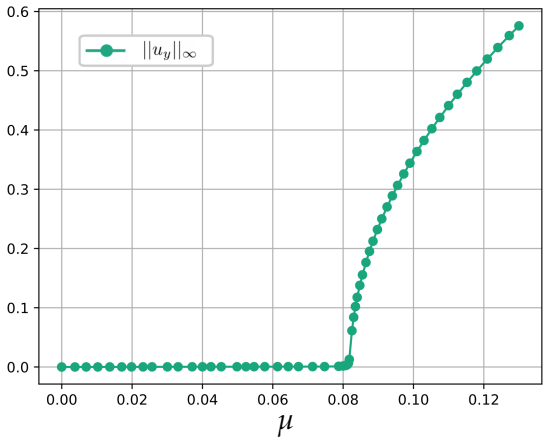


Figure 6.40: Reduced bifurcation diagram for the 3D SVK tubular geometry with $B = (0, 0, 0)$.

A representative solution of the post-buckling behaviour is depicted in Figure 6.42 for $\mu = 0.13$ with respect to the original undeformed configuration (mesh wireframe). Moreover, we plot in Figure 6.43 the top-view sliced with respect to its axis at $z =$

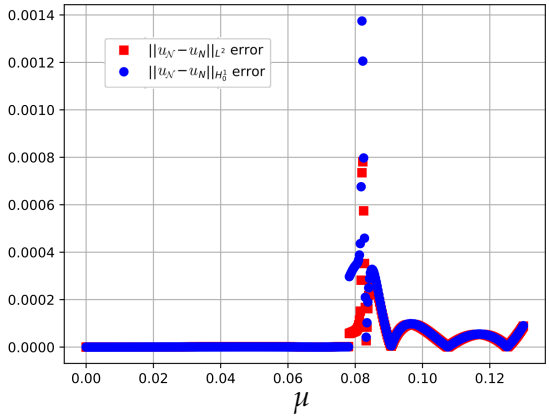


Figure 6.41: Reduced basis error for the 3D SVK tubular geometry with $B = (0, 0, 0)$.

10, where we can observe that the displacement of the cross section for $\mu = 0.13$ completely exit from its (wireframe) original configuration.

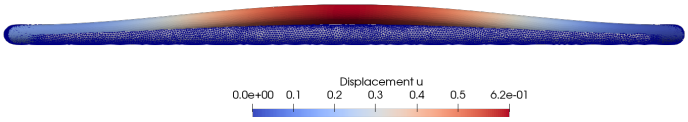


Figure 6.42: High fidelity displacement u for the 3D SVK tubular geometry with $B = (0, 0, 0)$ at $\mu = 0.13$.

Now that we have a complete overview of the model, we can finally present the a multi-parameter test case with geometrical parametrization of the Dirichlet compressed tubular beam. The aim here is to investigate the buckling behaviour of longer geometries.

The domain is defined, consistently with Section 6.2.3.2, as $\widetilde{\Omega}(\mu_g) = \widetilde{\Omega}_1 \cup \widetilde{\Omega}_2(\mu_g)$, where $\widetilde{\Omega}_1 = A_r^R \times [0, 20]$ and $\widetilde{\Omega}_2(\mu_g) = A_r^R \times [20, 20 + \mu_g]$ where the geometrical parameter μ_g varies in $\mathcal{P}_g = [10, 20]$.

Keeping fixed the setting as before, we computed a global number of $N_{train} = 800$ snapshots, divided as equispaced points in $\mathcal{P}_p$ for each one of the three equispaced values in $\mathcal{P}_g$ and we obtained a reduced basis space of dimension $N = 12$ with POD tolerance $\varepsilon_{POD} = 10^{-10}$. We present the 3-D bifurcation diagram in Figure 6.44, in which we reconstructed the buckling behaviour for 5 equispaced values of the semi-length $\mu_g \in \mathcal{P}_g$. From the figure it is evident the effect of the length on the buckling location, indeed we pass from $\mu^* = 0.048$ at $\mu_g = 10$ to $\mu^* = 0.056$ for $\mu_g = 20$.

The top-view of five representative solutions of the buckling modes for $\mu_g = \{10, 12.5, 15, 17.5, 20\}$ (colored by their magnitude) are depicted in Figure 6.45 for the same value of the Dirichlet compression $\mu = 0.06$.

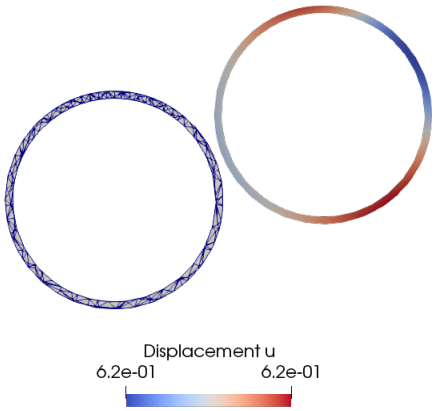


Figure 6.43: Top view of the high fidelity displacement u at $z = 10$ for the 3D SVK tubular geometry with $B = (0, 0, 0)$ at $\mu = 0.13$.

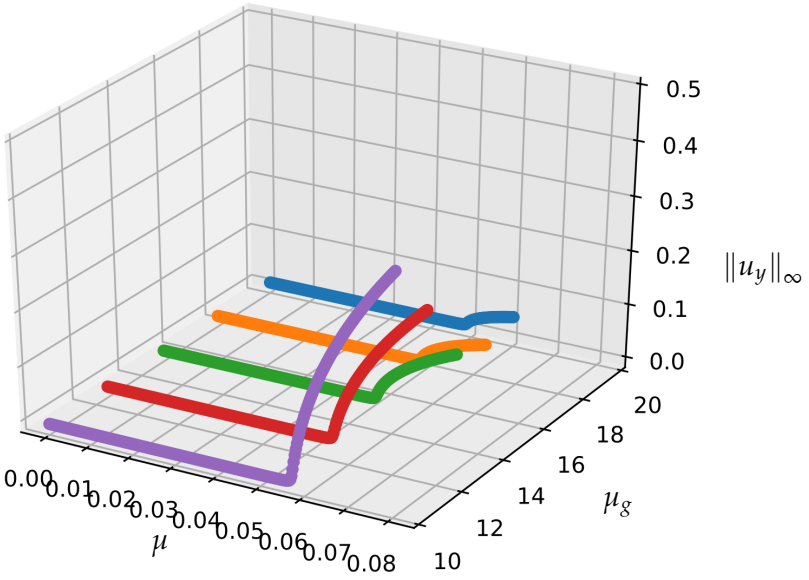
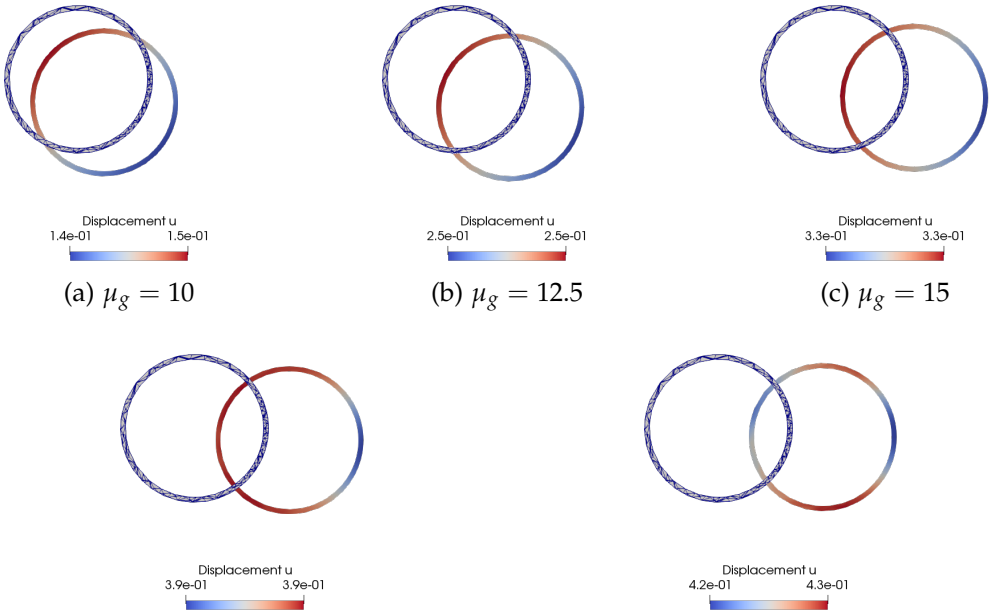


Figure 6.44: 3D bifurcation plot for 3D SVK tubular geometries with $B = (0, 0, 0)$ and $\mu_g \in \mathcal{P}_g$.

We present in Figure 6.46 the reduced basis error for two values of μ_g, remarking that, also in this more complex context, the reduced manifold was able to approximate with good accuracy the buckling for unsampled values of the geometrical parameter space $\mathcal{P}_g$.

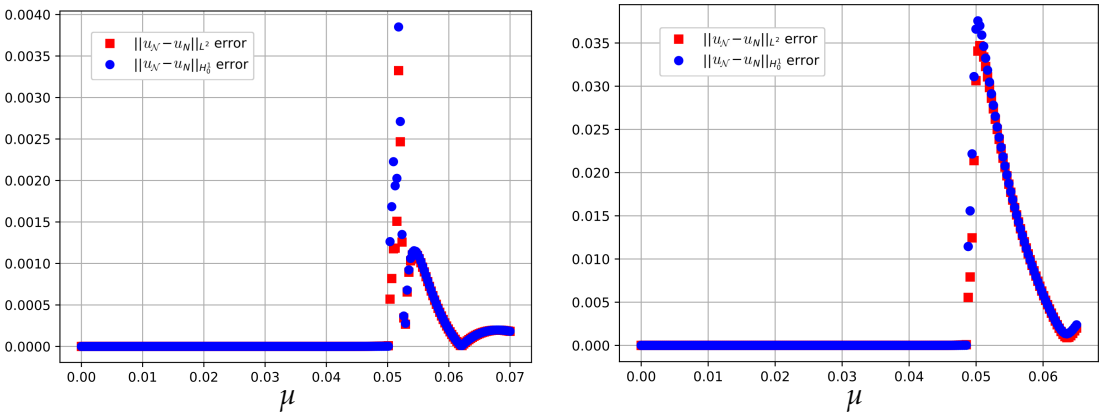
The study of the evolution of the buckling, varying the length of the beam, confirms as expected that longer beams need smaller forces to buckle. We highlight that since



(a) $\mu_g = 10$ (b) $\mu_g = 12.5$ (c) $\mu_g = 15$

(d) $\mu_g = 17.5$ (e) $\mu_g = 20$

Figure 6.45: Top-view of the high fidelity displacement u for the SVK beam with $B = (0, 0, 0)$ for different geometries at $\mu = 0.06$.



(a) $\mu_g = 15$ (b) $\mu_g = 17.5$

Figure 6.46: Reduced Basis errors with respect to $\mu \in \mathcal{P}$ for the SVK beam with $B = (0, 0, 0)$ for different geometries.

the importance of the buckling location, already during the offline stage, one can not perform an extensive analysis on the parametrized geometries. Moreover, to prevent non convergence issues due to bigger strains one should consider a much refined mesh,

increasing the high fidelity dimension.

Same conclusions on the speed-up hold here, where it increases up to 2.5, mainly due to the computational time $t_{HF} = 51102$(s) required to recover the high fidelity version of Figure 6.44.

Finally we remark that, due to the very high number of degrees of freedom within this test case we did not apply any empirical interpolation strategies. Indeed, it would cause an impracticable and too costly offline phase. In Chapter 10 we will present an approach that could help when facing with this kind of demanding tasks.

Within this setting we analysed many different test cases for the investigation of buckling beams. These allowed us to study how the critical points of the models vary in multi-parameter settings. Moreover, we understood that despite the simpler bifurcating behaviour, if interested only in the first buckling mode, unexpected phenomena can occur and a deep investigation is needed.

In the next chapter we will pss to a fluid dynamics application, analysing the stability properties of the coexisting flow profiles.

Chapter 7

Navier-Stokes flow in a channel: the Coanda effect

In this chapter, we are interested in a Fluid Dynamics application, analysing a bifurcating phenomenon deriving from Navier-Stokes (NS) equations in a sudden-expansion channel flow problem.

Among all the possible physical applications we mention combustion chambers, mixing vessel and heat exchangers; here we consider a simplified version for a model of a cardiac disease, called *mitral valve regurgitation*. The latter phenomenon, which can be clinically detected through echocardiography, is called the *Coanda effect* [151], and expresses the tendency of a fluid jet to be attracted to a nearby surface. This represents an issue from the medical point of view, because a wall-hugging jet might lead to inaccurate echocardiography measurements.

For this reason, we consider the channel geometry depicted in Figure 7.1 and consider the following physical context. A fluid, characterized by an high viscosity, presents a jet which is symmetric w.r.t. the horizontal axis. Furthermore, a pair of vortices, called *Moffatt eddies* [108], form downstream of the expansion. Lowering the viscosity, the inertial effects of fluid become more important and the two symmetric recirculation regions break the symmetry. Indeed, as the length of the recirculation zones increases, one can observe a non-uniform decrease of the pressure along the vertical axis. Thus, when we reach the aforementioned critical value, one recirculation zone expands whereas the other shrinks, giving rise to an asymmetric jet, which represents the wall hugging behaviour.

In the next section, we introduce the mathematical formulation of Navier-Stokes equations describing the flow in a channel. This will serve us to investigate the bifurcated behaviour of the system and its stability properties. Moreover, it will be fundamental for the analysis that we will present in next chapters, where the NS system will be the starting point for the application and development of different models and numerical methodologies.

7.1 Navier-Stokes viscous model

From the mathematical point of view, the problem described above translates to a generic parametrized PDE of the form (2.1) which, decreasing the viscosity μ below a certain critical value μ^*, admits the existence of more solutions for the same value of $\mu \in \mathcal{P}$. During the study of the solution for different viscosity values, we expect the system to show two qualitatively different configurations:

- a physically unstable configuration with a symmetric jet flow, which we will refer to as the *symmetric solution*,
- a physically stable configuration with a wall-hugging jet, which we will refer to as the *asymmetric solution*.

These solutions, that coexist for parameter values below the critical one μ^*, belong to different branches that intersect in the bifurcation point, forming a pitchfork bifurcation.

Here we consider a simplified setting with a two-dimensional planar straight channel with a narrow inlet and a sudden expansion, depicted in Figure 7.1, which represent a simplification of the left atrium and the mitral valve, respectively. We define $\Gamma_{in} = \{0\} \times [2.5, 5]$ and $\Gamma_{out} = \{50\} \times [0, 7.5]$, where inflow and outflow boundary conditions are imposed, respectively. We indicate with Γ_{wall}, the boundaries representing the walls, in this case $\Gamma_{wall} = \Gamma_D \cup \Gamma_0$, where $\Gamma_D = \{\{0\} \times [0, 2.5]\} \cup \{\{0\} \times [5, 7.5]\}$ and $\Gamma_0 = \partial\Omega \setminus \{\Gamma_{in} \cup \Gamma_D \cup \Gamma_{out}\}$.

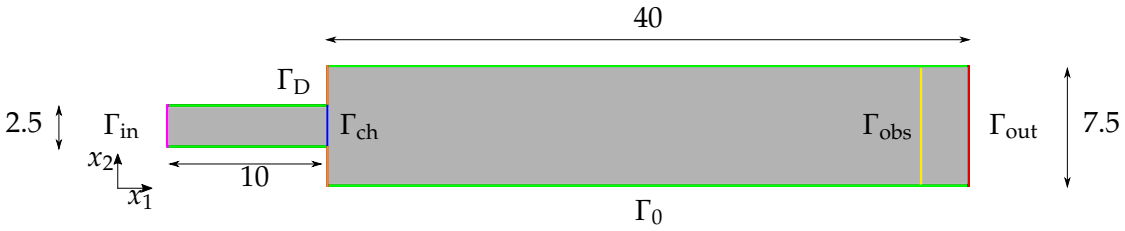


Figure 7.1: Domain Ω which represents a straight channel with a narrow inlet.

The steady and incompressible Navier-Stokes equations for a viscous flow in Ω read as:

$$\begin{cases} -\mu\Delta v + v \cdot \nabla v + \nabla p = 0 & \text{in } \Omega, \\ \nabla \cdot v = 0 & \text{in } \Omega, \\ v = v_{in} & \text{on } \Gamma_{in}, \\ v = 0 & \text{on } \Gamma_{wall}, \\ -pn + (\mu\nabla v)n = 0 & \text{on } \Gamma_{out}, \end{cases} \tag{7.1}$$

where $v = (v_{x_1}, v_{x_2})$ is the velocity of the fluid, p is its pressure normalized over a constant density and μ represents the kinematic viscosity. We supplement the system (7.1) with proper boundary conditions: a stress free boundary condition on the velocity

at the outlet, Γ_{out} with outer normal n, a no-slip (homogeneous) Dirichlet boundary condition on Γ_{wall} and a non-homogeneous Dirichlet boundary conditions v_{in} at the inlet Γ_{in} given by

$$v_{\text{in}}(x_2) = \begin{bmatrix} 20(5 - x_2)(x_2 - 2.5) \\ 0 \end{bmatrix}.$$

For later convenience, we introduce the dimensionless *Reynolds number* as $\text{Re} = Uw/\mu$, which represents the ratio between inertial and viscous forces, where U and w are characteristic velocity (i.e., maximum inlet velocity, $U = 31.25$) and the characteristic length of the domain (i.e., length of the inlet section, $w = 2.5$), respectively. In the following we will consider μ as the parameter. However, especially while commenting the results, we will often refer to the associated Reynolds number.

Fixed the domain Ω, the flow regime varies as we consider different values for the viscosity μ in $\mathcal{P} \subset \mathbb{R}$. We remark that this model exhibits a bifurcating behaviour. Indeed, we have the existence and uniqueness of the solution only above a certain critical value for the viscosity, that for this test case corresponds to $\mu^* \approx 0.96$. Thus the problem for higher values of the corresponding Reynolds number loses the well-posedness and we have to refer to solution branches of a pitchfork bifurcation. Since our aim is to investigate the loss of uniqueness of the solution in a neighbourhood of this pitchfork bifurcation, we set the parameter space as $\mathcal{P} = [0.5, 2.0]$, such that the first critical point μ^* is included. These values for the viscosity correspond to Re in the interval [39.0, 156.0].

7.1.1 Weak formulation

Let $\mathbb{V} = \left(H^1(\Omega)\right)^2$, $\mathbb{V}_{\text{in}} = \{v \in \mathbb{V} \mid v = v_{\text{in}} \text{ on } \Gamma_{\text{in}}, v = 0 \text{ on } \Gamma_{\text{wall}}\}$, $\mathbb{V}_0 = \{v \in \mathbb{V} \mid v = 0 \text{ on } \Gamma_{\text{in}} \cup \Gamma_{\text{wall}}\}$ be the function spaces for velocity. Furthermore, let $\mathbb{Q} = L^2(\Omega)$ be the function space for pressure. The weak formulation of (7.1) reads as: given $\mu \in \mathcal{P}$, find $v \in \mathbb{V}_{\text{in}}$ and $p \in \mathbb{Q}$ such that

$$\begin{cases} \mu \displaystyle\int_\Omega \nabla v \cdot \nabla \psi \, d\Omega + \int_\Omega (v \cdot \nabla v)\, \psi \, d\Omega - \int_\Omega p \nabla \cdot \psi \, d\Omega = 0 & \forall\, \psi \in \mathbb{V}_0, \\ \displaystyle\int_\Omega \pi \nabla \cdot v \, d\Omega = 0 & \forall\, \pi \in \mathbb{Q}. \end{cases} \tag{7.2}$$

We can rewrite the formulation of (7.2) in an equivalent way as: given $\mu \in \mathcal{P}$, find $v \in \mathbb{V}_{\text{in}}$ and $p \in \mathbb{Q}$ such that

$$\begin{cases} a(v, \psi; \mu) + s(v, v, \psi) + b(\psi, p) = 0 & \forall\, \psi \in \mathbb{V}_0, \\ b(v, \pi) = 0 & \forall\, \pi \in \mathbb{Q}, \end{cases} \tag{7.3}$$

having introduced the following bilinear and trilinear forms

$$\begin{aligned} a(v,\psi;\mu) &= \mu \int_\Omega \nabla v \cdot \nabla \psi \, d\Omega \qquad && \forall\, v,\psi \in \mathbb{V}, \\ b(v,p) &= -\int_\Omega (\nabla \cdot v)\, p \, d\Omega \qquad && \forall\, v \in \mathbb{V},\ \forall\, p \in \mathbb{Q}, \\ s(v,\bar{v},\psi) &= \int_\Omega (v \cdot \nabla \bar{v})\, \psi \, d\Omega \qquad && \forall\, v,\bar{v},\psi \in \mathbb{V}. \end{aligned} \tag{7.4}$$

In particular, if we define as $\mathbb{X} = \mathbb{V}_{\text{in}} \times \mathbb{Q}$ the space of solution $X = (v,p)$, then we can rewrite the weak formulation in terms of the parametrized variational form $g : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ as

$$g(X,Y;\mu) := a(v,\psi;\mu) + s(v,v,\psi) + b(\psi,p) + b(v,\pi) = 0, \quad \forall\, Y \in \mathbb{X}, \tag{7.5}$$

where we denoted with $Y = (\psi,\pi) \in \mathbb{X}$ the test function. Furthermore, we can compute the Frechét derivative with respect to v at $Z = (z_v, z_s) \in \mathbb{X}$ of the parametrized variational form g, which can be expressed as

$$\mathrm{d}g[Z](X,Y;\mu) = a(v,\psi;\mu) + s(v,z_v,\psi) + s(z_v,v,\psi) + b(\psi,p) + b(v,\pi) \quad \forall\, Y \in \mathbb{X}. \tag{7.6}$$

As we already specified in Section 6.1.1, we remark that when a lifting function is taken into account to preserve the symmetry between the test and trial function spaces, one has also to incorporate the linear terms coming from nonlinear ones evaluated at the lifting. In the following we will present the numerical approximation of the model.

7.2 Numerical approximation

The first step to understand numerically the bifurcating behaviour of the model by means of the FE method, is the projection of its weak formulation in (7.3) into the finite dimensional space $\mathbb{X}_\mathcal{N} = \mathbb{V}_{\mathcal{N}_v} \times \mathbb{Q}_{\mathcal{N}_p}$ of dimension $\mathcal{N} = \mathcal{N}_v + \mathcal{N}_p$.

Thus, the combination of the Galerkin FE and the generic k-th step of Newton method reads as: given $\mu \in \mathcal{P}$ and an initial guess $X^0_\mathcal{N} = (v^0_\mathcal{N}, p^0_\mathcal{N}) \in \mathbb{X}_\mathcal{N}$ for $k = 0, 1, \dots$ until convergence we seek $\delta X_\mathcal{N} = (\delta v_\mathcal{N}, \delta p_\mathcal{N}) \in \mathbb{X}_\mathcal{N}$ such that

$$\begin{cases} a(\delta v_\mathcal{N}, \psi_\mathcal{N};\mu) + s(v^k_\mathcal{N}, \delta v_\mathcal{N}, \psi) + s(\delta v_\mathcal{N}, v^k_\mathcal{N}, \psi_\mathcal{N}) + b(\psi_\mathcal{N}, \delta p_\mathcal{N}) = & \\ \qquad + a(v^k_\mathcal{N}, \psi_\mathcal{N};\mu) + s(v^k_\mathcal{N}, v^k_\mathcal{N}, \psi) + b(\psi_\mathcal{N}, p^k_\mathcal{N}) & \forall\, \psi_\mathcal{N} \in \mathbb{V}_{0,\mathcal{N}}, \\ b(\delta v_\mathcal{N}, \pi_\mathcal{N}) = b(v^k_\mathcal{N}, \pi_\mathcal{N}) & \forall\, \pi_\mathcal{N} \in \mathbb{Q}_\mathcal{N}, \end{cases} \tag{7.7}$$

and then set $X^{k+1}_\mathcal{N} = X^k_\mathcal{N} - \delta X_\mathcal{N}$.

Thus, we can present the corresponding algebraic form of the system (7.7) as in (3.10). In fact, denoting the solution vector with $\mathsf{X}_\mathcal{N} = (\mathsf{v}_\mathcal{N}, \mathsf{p}_\mathcal{N})$ we can write

$$\begin{pmatrix} \mathsf{A}_\mathcal{N}(\mu) + \mathsf{S}_{1,\mathcal{N}}(\mathsf{v}^k_\mathcal{N}) + \mathsf{S}_{2,\mathcal{N}}(\mathsf{v}^k_\mathcal{N}) & \mathsf{B}^T_\mathcal{N} \\ \mathsf{B}_\mathcal{N} & 0 \end{pmatrix} \begin{pmatrix} \delta \mathsf{v}_\mathcal{N} \\ \delta \mathsf{p}_\mathcal{N} \end{pmatrix} = \begin{pmatrix} \mathsf{A}_\mathcal{N} \mathsf{v}^k_\mathcal{N} + \mathsf{S}_{1,\mathcal{N}}(\mathsf{v}^k_\mathcal{N}) \mathsf{v}^k_\mathcal{N} + \mathsf{B}^T_\mathcal{N} \mathsf{p}_\mathcal{N} \\ \mathsf{B}_\mathcal{N} \mathsf{v}_\mathcal{N} \end{pmatrix}, \tag{7.8}$$

where we denoted the high fidelity matrices as follows

$$(\mathsf{A}_{\mathcal{N}}(\mu))_{ij} = a(E^j_{\mathcal{N}}, E^i_{\mathcal{N}}; \mu) \ , \quad (\mathsf{B}_{\mathcal{N}})_{ij} = b(E^j_{\mathcal{N}}, E^i_{\mathcal{N}}) \ ,$$
$$(\mathsf{S}_{1,\mathcal{N}})_{ij} = s(\mathsf{v}^k_{\mathcal{N}}, E^j_{\mathcal{N}}, E^i_{\mathcal{N}}) \ , \quad (\mathsf{S}_{2,\mathcal{N}})_{ij} = s(E^j_{\mathcal{N}}, \mathsf{v}^k_{\mathcal{N}}, E^i_{\mathcal{N}}) \ .$$

The same structure is inherited by the reduced order model, obtained through the projection on the space $\mathbb{X}_N = \mathbb{V}_{N_v} \times \mathbb{Q}_{N_p}$ of dimension $N = N_v + N_p$.

Indeed, the k-th reduced step of the Newton method reads as: given $\mu \in \mathcal{P}$ and an initial guess $\mathsf{X}^0_N \in \mathbb{X}_N$ for $k = 0, 1, \dots$ until convergence we seek $\delta\mathsf{X}_N = (\delta\mathsf{v}_N, \delta\mathsf{p}_N) \in \mathbb{X}_N$ that satisfies (3.20), with the reduced Jacobian and the reduced residual are given respectively by

$$\mathsf{J}_N(\mathsf{X}^k_N(\mu); \mu) = \begin{pmatrix} \mathsf{A}_N(\mu) + \mathsf{S}_{1,N}(\mathsf{v}^k_N) + \mathsf{S}_{2,N}(\mathsf{v}^k_N) & \mathsf{B}^T_N \\ \mathsf{B}_N & 0 \end{pmatrix} \ , \quad \text{and}$$

$$\mathsf{G}_N(\mathsf{X}^k_N(\mu); \mu) = \begin{pmatrix} \mathsf{A}_N(\mu)\mathsf{v}^k_N + \mathsf{S}_{1,N}(\mathsf{v}^k_N)\mathsf{v}^k_N + \mathsf{B}^T_N\mathsf{p}_N \\ \mathsf{B}_N\mathsf{v}_N \end{pmatrix} \ .$$

Once again, in the expression above we have introduced the transformation matrices with respect to the different components of the solution, V_v and V_p, respectively for velocity and pressure, defining the reduced matrices in the following way:

$$\mathsf{A}_N(\mu) = \mathsf{V}_v^T \mathsf{A}_{\mathcal{N}}(\mu)\mathsf{V}_v \ , \quad \mathsf{B}_N = \mathsf{V}_p^T \mathsf{B}_{\mathcal{N}}\mathsf{V}_v \ ,$$
$$\mathsf{S}_{1,N} = \sum_{n=1}^{N_v} v_N^{(n)} \mathsf{V}_v^T \mathsf{S}_{1,\mathcal{N}}(\Sigma_v^n)\mathsf{V}_v \ , \quad \mathsf{S}_{2,N} = \sum_{n=1}^{N_v} v_N^{(n)} \mathsf{V}_v^T \mathsf{S}_{2,\mathcal{N}}(\Sigma_v^n)\mathsf{V}_v \ .$$

As concerns the well-posedness analysis, we postpone the discussion to Section 9.1.3, where it will be treated in much more generality.

We remark that the these matrices prevent the offline-online decomposition needed for an efficient investigation of the bifurcation diagram. Of course, we could rely on the empirical interpolation strategies adopted in the previous sections, but in the next section we will focus on a different perspective. The numerical (reduced) study of the physical phenomenon will serve as the starting point for the modelling and the methodological development which we will presented in the next part.

7.2.1 Bifurcating approach to the Coanda effect

We can now discuss the numerical approximation of the branching behaviour for the Navier-Stokes equations. We consider a mesh on the domain Ω with 2785 cells, corresponding to $\mathcal{N} = 24301$ degrees of freedom associated to a Taylor-Hood $\mathbb{P}^2$-$\mathbb{P}^1$ discretization of $\mathbb{V} \times \mathbb{Q}$. This choice is motivated by the well-known stability results of the Taylor-Hood Finite Element pair [133].

For the high fidelity setting we chose $N_{train} = 51$ viscosity points in the parameter space $\mathcal{P} = [0.5, 2]$. Within this setting we slightly changed the reduced approach.

In fact, we considered also at the reduced setting a branch-wise reconstruction, by considering the snapshots coming only from a single branch. Thus we build a ROM for each dynamics, the symmetric and the asymmetric one.

A POD compression is applied to each branch, extracting a different bases of the same dimension $N_v = N_p = 20$. As regard the online continuation method, it is based on $K = 151$ equispaced points in $\mathcal{P}$, which corresponds to a continuation step $\Delta\mu = 10^{-2}$.

In order to plot the bifurcation diagram, we choose an output value that results in a symmetry indicator of the approximated solution. This function is given by the value of the vertical component of the velocity, in a point lying on the middle axis of the channel, namely v_{x_2} evaluated at $(x_1, x_2) = (14, 4)$. In Figure 7.2 we plot the bifurcation diagram with all the solution branches of the system (7.3) in the viscosity range chosen.

The numerical approximation clearly shows that a supercritical pitchfork bifurcation occurs around the critical viscosity value $\mu^* \approx 0.96$. It is evident that we have a unique solution for all $\mu > \mu^*$, thus when the fluid behaves like a Stokes one, while we find three qualitatively different solutions increasing the Reynolds number. The bifurcation point μ^* is also the one responsible for the change in stability properties of the model. Indeed, the unique symmetric solution remains stable until it encounters the critical value μ^*, where it becomes unstable. Moreover, this feature is inherited by the bifurcating solutions, which evolve as a physically stable branch. Finally, we remark that the diagram evolves in "opposite" direction with respect to the previous cases because interesting phenomena occur at higher Reynolds numbers, i.e. lower viscosities.

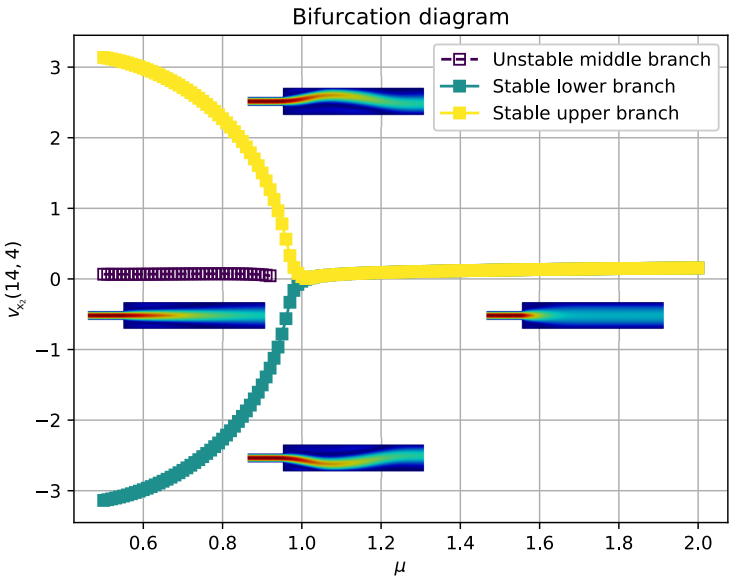


Figure 7.2: Bifurcation diagram for the Navier-Stokes system.

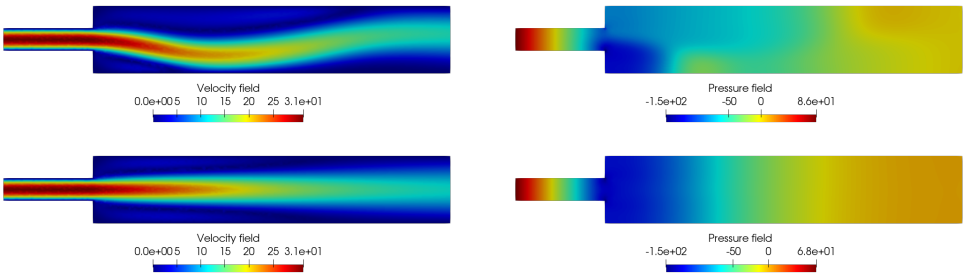


Figure 7.3: Representative solutions for the Navier-Stokes system at $\mu = 0.5$, velocity and pressure fields, lower and middle branch, top and bottom respectively.

Some representative solutions for the lower and middle branch are presented in Figure 7.3. Velocity and pressure fields belonging to different branches, for the same viscosity value $\mu = 0.5$, present qualitatively dissimilar behaviour. Indeed, the pressure for the lower branch decreases near the bottom-left corner of the expansion, causing the velocity to deflect, hugging the lower wall. Finally, thanks to the no-slip boundary condition the flux goes back to mid line, ending with a non-axis-symmetric outflow.

As we can see from Figure 7.4, both ROMs were able to recover online a good approximation of the bifurcating branches for both velocity and pressure fields. We can observe that in both cases the RB error has its maximum at the critical value μ^*. Actually, the higher error for the symmetric branch is caused by the convergence of the FE solution to the bifurcated asymmetric configuration that coexist at the bifurcation point. Moreover, we remark that the average error on $\mathcal{P}$ for the symmetric branch of order 10^{-9} is two order lower than the one for the asymmetric branch, which corresponds to 10^{-7}. This is obviously due to the fact that the latter contains both profiles: the symmetric one in the pre-bifurcation range and the asymmetric one in the bifurcating regime.

The stability analysis is performed through the eigenvalue analysis depicted in Section 3.1.3, where Algorithm 1 has been applied to the NS equation in (7.1). In particular, we analysed the behaviour of the first $N_{eig} = 100$ eigenvalues of (3.12), by means of the Krylov-Schur algorithm, varying the viscosity of the fluid. Such eigenvalues are plotted in Figure 7.5 for the stable lower branch (left panel) and unstable middle branch (right panel). Note that since the Navier-Stokes operator is not symmetric, we have the presence of both real and complex eigenvalues. Here we are just interested in pitchfork bifurcation, thus in the zoom we follow the behaviour of the biggest real eigenvalue and its sign. When investigating the stability of the lower branch, all the eigenvalues of the Navier-Stokes system, linearized around this stable solution, have negative real part. From the consideration of Section 3.1.3, we can assert the stability of the wall-hugging branch. Indeed, the zoom in the left plot of Figure 7.5 shows no crossing of negative-real part eigenvalues. On the contrary, the close up in the right plot, which corresponds to the symmetric flow, shows the sign change of the biggest eigenvalue, thus characterizing a physically unstable solution.

The analysis of the Coanda effect we performed in this chapter, provided a com-

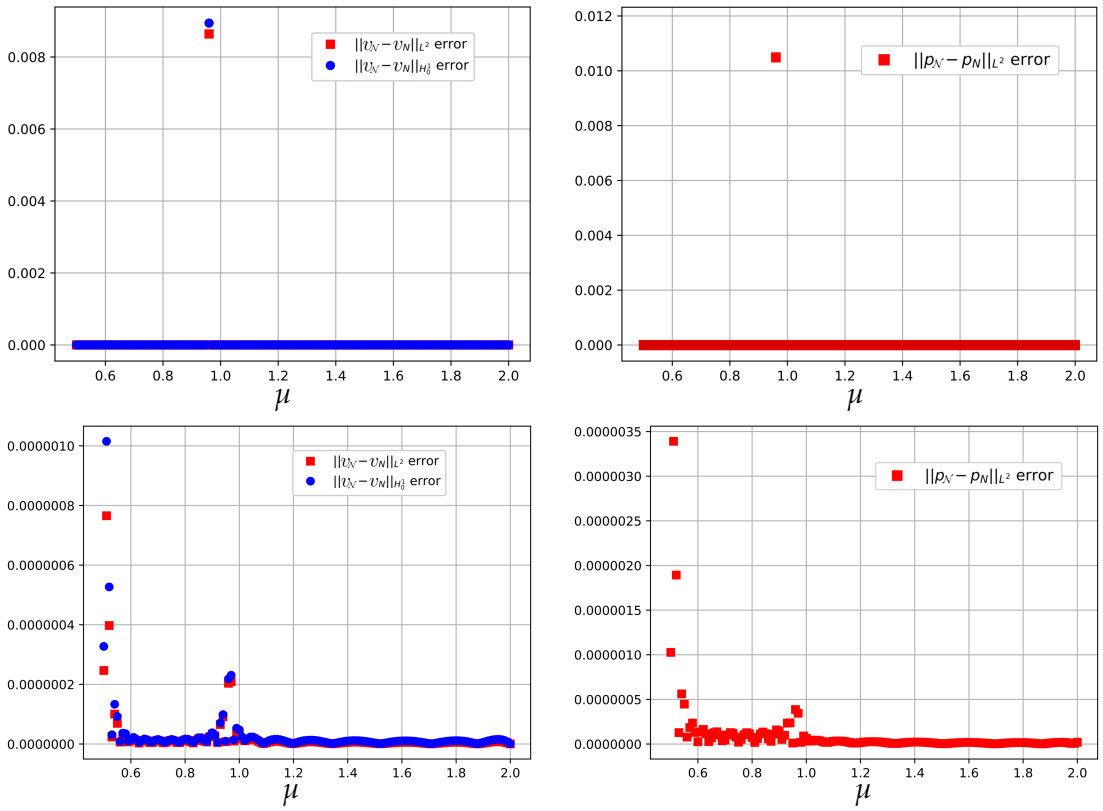


Figure 7.4: Reduced basis errors with respect to $\mu \in \mathcal{P}$ for velocity and pressure, symmetric (top row) and asymmetric (bottom row) branches.

plete description of the bifurcating phenomena shown by the Navier-Stokes equations. Here, we discussed the stability of the wall-hugging profile an the instability of the straight flow by means of the eigenvalue analysis.

Moreover, with this chapter we concluded the second part of the thesis, where we presented different models and their corresponding branching behaviour. In the following, the Navier-Stokes test case investigated here, will be the starting point for the development and analysis of more complex models, by means of new methodologies which aim at overcoming some of the issues we encountered so far.

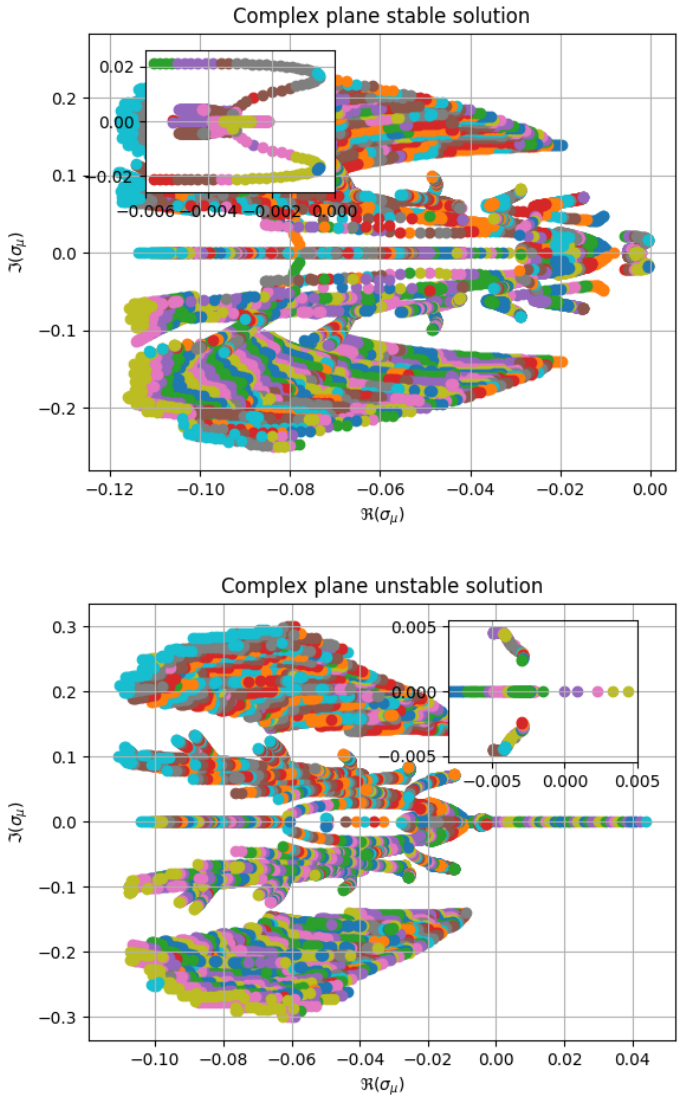


Figure 7.5: Eigenvalues of the state eigenproblem in the complex plane for the Navier-Stokes equation: stable and unstable solutions, top and bottom panels respectively.

Part III

Towards more complex problems and emerging methodologies

Chapter 8

A reduced order approach to deflated continuation method

As we understood from the previous investigation, the numerical computation of a bifurcation diagram requires the combination of different strategies. Moreover, these have to be adapted case by case, in particular as concerns the choice of the continuation method and its corresponding guesses.

When complex bifurcating phenomena occur, some of the presented techniques are not always sufficiently stable to discover new branches. Therefore, if the main goal is the complete reconstruction of the bifurcation diagram, a different and more general approach is needed.

Furthermore, we recall that our main interest is an efficient investigation. Despite we addressed this issue with the reduction strategies, it still remains the bottleneck of the initialization of the continuation method. In fact, providing a tailored guess for each sought branch, especially when no physical or analytical considerations can be taken into account, could be an almost impracticable task.

For this reason, we embedded in our continuation strategy the *deflation method* [56, 55], which aims at automatically discovering new branches. In particular, as we will describe later, the deflation prevents the iterative solver from converging to already known solutions and this way, if it converges, it will converge to yet unknown fields.

This methodology have been proved to work well when applied to many different contexts [34, 55], but still it can involve an infeasible computational cost. Hence, we investigated a reduced approach in which an efficient reconstruction of the bifurcation diagram is enabled by the low dimensionality of the RB space.

This part is entirely based on the work done with M. Pintore in collaboration with M. Hess and C. Canuto [122].

8.1 A CFD toy problem with Spectral Element Method

The test case chosen for the development and the analysis of the reduced deflated continuation method is a slight modification of the one studied in Chapter 7. Indeed,

we consider here the Navier-Stokes equations in a straight channel with similar geometry as in Figure 7.1, where only the after inlet length is increased in order to capture different wall-hugging behaviours.

Thus we are interested in the steady and incompressible Navier-Stokes equations for a viscous flow in Ω, given by:

$$\begin{cases} -\mu\Delta v + v\cdot\nabla v + \nabla p = 0 & \text{in } \Omega, \\ \nabla\cdot v = 0 & \text{in } \Omega, \\ v = v_{\text{in}} & \text{on } \Gamma_{\text{in}}, \\ v = 0 & \text{on } \Gamma_{\text{wall}}, \\ -pn + (\mu\nabla v)n = 0 & \text{on } \Gamma_{\text{out}}, \end{cases} \tag{8.1}$$

where $v = (v_{x_1}, v_{x_2})$ is the velocity of the fluid, p is its pressure normalized over a constant density and μ represents the kinematic viscosity.

We chose for the system in (8.1) the same BCs setting as before: a stress free boundary condition on the velocity at the outlet, Γ_{out} with outer normal n, a no-slip (homogeneous) Dirichlet boundary condition on Γ_{wall}, and non-homogeneous Dirichlet boundary conditions v_{in} at the inlet Γ_{in} given by

$$v_{\text{in}}(x_2) = \begin{bmatrix} 20(5-x_2)(x_2-2.5) \\ 0 \end{bmatrix}.$$

We remark that within this physical setting, we selected the parameter range as $\mu \in \mathcal{P} = [0.3, 1]$ since it will contain two bifurcating phenomena. Note that small changes in the geometry can modify the critical points locations, as observed for instance in [73] (we will analyse an example of this behaviour in Section 10.2.2.2).

Contrarily to all the analysis performed throughout this thesis, here we implemented as Full Order Method the Spectral Element Method (SEM) [31, 30]. Thus, starting from the weak formulation in (7.3), we consider the spectral basis $\{\psi_S^i\}_{i=1}^{\mathcal{N}_v}$ and $\{\pi_S^i\}_{i=1}^{\mathcal{N}_p}$ associated to velocity and pressure, respectively. Therefore, the high fidelity fields can be expressed as follows:

$$v_{\mathcal{N}} = \sum_{i=1}^{\mathcal{N}_v} v_{\mathcal{N}}^{(i)} \psi_S^i, \qquad p_{\mathcal{N}} = \sum_{i=1}^{\mathcal{N}_p} p_{\mathcal{N}}^{(i)} \pi_S^i,$$

where $\mathsf{v}_{\mathcal{N}} = \{v_{\mathcal{N}}^{(i)}\}_{i=1}^{\mathcal{N}_v}$ and $\mathsf{p}_{\mathcal{N}} = \{p_{\mathcal{N}}^{(i)}\}_{i=1}^{\mathcal{N}_p}$ are scalar coefficients that characterize the velocity and pressure fields. The main difference with FE method is that in the SEM the basis functions are high-order polynomials inside the associated elements. In particular, in this work we decided to use the stable pair $\mathbb{P}^r/\mathbb{P}^{r-2}$ (see [98, 100] for a more detailed explanation of the methodology).

The high order of the polynomials implies two main consequences. Firstly, since several degrees of freedom are associated with each element, it is possible to obtain accurate solutions even with very coarse meshes. If one is able to reconstruct the

solution field employing only low order polynomials this approach is convenient due to the less degrees of freedom considered. Secondly, even if to compute bifurcation diagrams the mesh generation cost is negligible, generating fine meshes as the ones required by the FE method is an expensive operation. More remarkably, this way it is possible to ensure the exponential convergence of the method comparing with the algebraic one that characterizes the FE method when the solution is smooth enough.

8.2 The reduced deflated continuation method

Let us consider the following nonlinear parametric equation:

$$L(X;\boldsymbol{\mu}) = 0, \tag{8.2}$$

where the unknown X belongs to the functional space $\mathbb{X}$, the parameter $\boldsymbol{\mu}$ belongs to $\mathbb{R}^P$ and L is a nonlinear operator. As usual, we are interested in the computation of the several solutions that can coexist for the same value of the parameter. Until now we presented different bifurcating phenomena, and we summarized their information by means of bifurcation diagrams. We had to implement different strategies in order to tackle the difficulties arising from each specific test case. Therefore, we decided to investigate the properties of a much more general approach based on the computation of the bifurcation diagram through the combination of the continuation method and the deflation method. The main idea behind the coupling of this two techniques is that, while the former allows to properly reconstruct a branch following its behaviour, the latter exploits the continued solution to discover new branches. Such an approach, where the two techniques are alternated in order to discover and follow each branch of the diagram, is called *deflated continuation* [55, 56]. In order to address the huge computational cost required by this technique, we built a reduced version of it, which aims at providing an automatic procedure for the online recovery of the bifurcation diagram.

8.2.1 Pseudo-arclength continuation method

In this section we restrict ourselves, without loss of generalization, to the one parameter case $\mu \in \mathbb{R}$. Indeed, as we have done in the previous chapters we can always vary the bifurcation parameter while keeping fixed the physical/geometrical configuration. We have understood that, in order to ensure its convergence, the nonlinear solver needs an initial guess close enough to the sought solution. Until now, our approach was based on a simple continuation algorithm which exploits the last computed solution and uses it as a guess for the next iteration. Of course this approach is inexpensive and requires a single solution, however as we remarked in Section 3.1.4, properly setting the continuation step $\Delta\mu$ can be a difficult task.

When a waste of the computational time occurs or convergence issues arise, one can choose to rely on the *pseudo-arclength continuation method* [86, 2]. This technique treats the next value of the parameter μ_{i+1}, at which compute the solution, as an

unknown and allows for an alternative parametrization of the branch, characterized by its arclength σ. To derive the system that has to be solved let us consider the following normalization equation

$$N(X, \mu; \Delta\sigma_i) := \dot{X}_i^T (X - X_i) + \dot{\mu}_i(\mu - \mu_i) - \Delta\sigma_i = 0, \tag{8.3}$$

where (X_i, μ_i) is a point on a regular portion of the branch $\mathcal{M}$ and $(\dot{X}_i, \dot{\mu}_i)$ is the unit tangent to the curve in such a point [86]. Equation (8.3) characterizes the plane orthogonal to the vector $(\dot{X}_i, \dot{\mu}_i)$ such that the distance between (X_i, μ_i) and its projection on the plane is $\Delta\sigma_i$. Moreover, if the line described by $(\dot{X}_i, \dot{\mu}_i)$ is a good approximation of $\mathcal{M}$, the orthogonal projection of (X_i, μ_i) on the plane is a good approximation of the sought solution (X_{i+1}, μ_{i+1}).

Consequently, this projection can be used as a close enough initial guess. In order to compute it, one linearizes the following system obtained by the combination of equations (8.2) and (8.3)

$$\begin{cases} L(X; \mu) = 0, \\ \dot{X}_i^T (X - X_i) + \dot{\mu}_i(\mu - \mu_i) - \Delta\sigma_i = 0, \end{cases} \tag{8.4}$$

where the quantities $\dot{X}_i$ and $\dot{\mu}_i$ are approximated as

$$\dot{X}_i \simeq \frac{X_i - X_{i-1}}{\Delta\sigma_{i-1}}, \qquad \dot{\mu}_i \simeq \frac{\mu_i - \mu_{i-1}}{\Delta\sigma_{i-1}}.$$

The main advantage of the pseudo-arclength technique is that the subsequent value of the parameter is automatically chosen. This way, the steps are wider in very smooth regions, while they can be much shorter near the singularities. Finally, we remark that the key ingredient of the accuracy of the pseudo-arclength continuation is that it actually relies on a branch linearization.

8.2.2 Deflating to discover new solutions

Here, we want to discuss the deflation method used for the exploration of the existing solutions for a generic nonlinear parametric PDE. Such a technique has been initially developed to compute multiple roots of a polynomial [56]. Before moving to its description, it is convenient to introduce the concept of *deflation operator.*

Definition 8.2.1. Let us denote by $\mathbb{W}$ and $\mathbb{Z}$ two Banach spaces and by $\mathbb{O}$ an open subset of the Banach space $\mathbb{U}$. Moreover, let $L : \mathbb{O} \subset \mathbb{U} \to \mathbb{W}$ be a Fréchet differentiable operator and $D_X L$ be its Fréchet derivative. Then, let $M(X; W) : \mathbb{W} \to \mathbb{Z}$ be an invertible linear operator for each $W \in \mathbb{O}$ and for each $X \in \mathbb{O}\backslash\{W\}$. If the following holds:

$$L(W) = 0, \qquad D_X L(W) \text{ is nonsingular},$$

and for any arbitrary sequence $\{X_i\}_{i\in\mathbb{N}} \subset \mathbb{O}\backslash\{W\}$ such that $X_i \to \mathbb{W}$, the following inequality holds

$$\liminf_{i\to\infty} \|M(X_i; W)L(X_i)\|_{\mathbb{Z}} > 0, \tag{8.5}$$

then M is a deflation operator.
The deflation operators $M(X;W)$ that are usually considered are of the form

$$M(X;W) = I + \frac{I}{\|X - W\|_{\mathrm{O}}^{p}},$$

where I is the identity operator, and p determines the region of attraction for W. Finally, the new system that one is interested in solving is the following deflated version of (8.2)

$$G(X;\mu) := M(X;W)L(X;\mu) = 0, \tag{8.6}$$

that has the key property of being characterized by the same solutions of $L(X;\mu) = 0$ except for W.

The main advantage of the deflation method consists in the ability to discover unknown branches without any prior knowledge. However, if other branches exist, one cannot be sure that they will be found with this technique. In fact, if a branch $\mathcal{X}$ is too far from any known solution, the solver may diverge before reaching the region of attraction of any of its solutions. Therefore, it is advisable to fix a meaningful maximum number of iterations for the iterative solver when the deflated system is solved. Such a threshold should be high enough in order to leave enough time to the iterative solver, however, if too many iterations are available, a lot of computational resources will be wasted when new branches cannot be found.

It is clear that the deflation method is the bottleneck of the deflated continuation, in fact the system has to be assembled and solved many times. However, as described in [55], it is possible to increase the efficiency of the deflation step by means of the Sherman-Morrison formula [63]. In fact, it allows to express the solution of a generic step of the Newton method for the deflated equation (8.6), in terms of the one for the original problem in (8.2), only exploiting an inexpensive scalar quantity τ. The latter controls somehow the convergence of the deflated problem and one can develop heuristic strategies in order to improve the convergence properties [122].

8.3 Numerical approximation

In this section, we will discuss how the deflation method, applied both to the high fidelity and reduced order discretizations, is able to find new solutions for the Coanda effect. Of course we remark that even if we are considering a different full order method, the only difference lies in the choice of the basis functions. For this reason both discretization phases are defined through the forms already presented in Section 7.2.

Here, we consider the Navier-Stokes equation in (8.1) on the domain Ω depicted in Figure 7.1, discretized with only 19 quadrilateral elements, where Γ_{out} is located at $x_1 = 100$. We remark that despite the longer geometry, the SEM approximation with $r = 12$ only consists in $\mathcal{N} = 7372$ degrees of freedom, in contrast to the $\mathcal{N} = 24301$ of the FE one presented before. While the choice of r influences the accuracy and

the efficiency, we highlight that high order polynomials help to find the bifurcating (asymmetric) configuration and together with the mesh discretization plays a key role for the exact location of critical points, as expected. Contrarily to what we have done in Section 7.2.1, here we considered again a unique global ROM, upon which we based the online phase.

Thus, as explained in [122, 56], for the reconstruction of the bifurcation diagram we performed the following steps. At the beginning, we compute the snapshots with SEM for values of μ selected by the pseudo-arclength algorithm, while the deflation method allows us to find different solutions at each new instance. Then, a unique POD is performed and the continuation and deflation strategies are applied in the manifold spanned by the resulting basis.

To detect the qualitative different solutions which coexist for this model, we considered in Section 7.2.1 a naive output functional, given by the vertical component of the velocity field. Following [56] in this application we chose the symmetry indicator functional given by

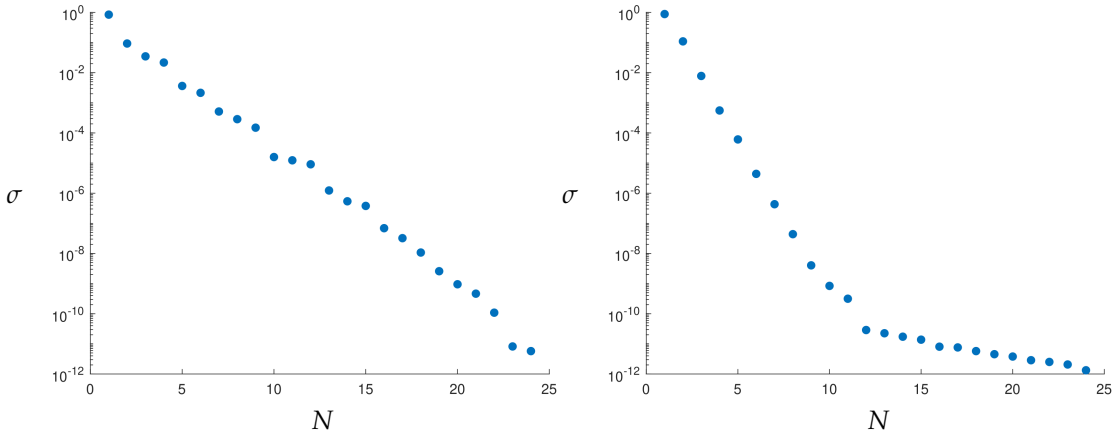
$$s(v) := \pm \int_{\Omega} \|v - \mathcal{R}(v)\|^2 d\Omega,$$

where $\mathcal{R}(\cdot)$ is an operator that reflects a solution over the horizontal symmetry axis, taken with its sign if the jet hugs the upper or lower wall. Therefore, $s(v)$ will be equal to zero if a solution is perfectly symmetrical, while its absolute value will increase with the asymmetry of the velocity field, showing clearly the pitchfork bifurcation.

8.3.1 A deflated bifurcation diagram

Let us consider the one-dimensional parameter range for the viscosity $\mu \in \mathcal{P} = [0.3, 1]$. For the sake of completeness, we present in Figure 8.1 a comparison about the property of the POD technique, between different approaches. The plots represent the decay of the eigenvalues of the correlation matrix (3.23), when $N_{train} = 24$ snapshots are computed in a neighbourhood of the critical point $\mu^* \approx 0.96$. In particular, we compare the result obtained sampling a single (asymmetric) branch with the one coming from the complete sampling of the existing branches. It can be observed that in both cases the decay is exponential [69, 115], even if there is a singular point in the range considered. Moreover, while it can be shown that taking snapshots closer to μ^* has no effect on the decay [122], we can clearly observe that choosing snapshots from a single branch allows for a faster decay, confirming the considerations of the previous sections about global and branch-wise POD. In fact, as expressed in (3.25) the faster is the decay, the better are the RB performances.

In Figure 8.2 we plot the bifurcation diagram with all the solution branches of the system (8.1) found in the viscosity range chosen. In particular, Figure 8.2, shows the bifurcation diagram obtained by the computation of $N_{train} = 224$ snapshots. It can be observed that it includes two bifurcation points and five different branches. Thus, using the deflation, a longer domain and a wider range for the viscosity we found a second pitchfork bifurcation, which occurs at $\mu^* \approx 0.4$.



(a) Sampling three branches (b) Sampling one branch

Figure 8.1: Decays of the eigenvalues of the POD correlation matrix with $N_{train} = 24$.

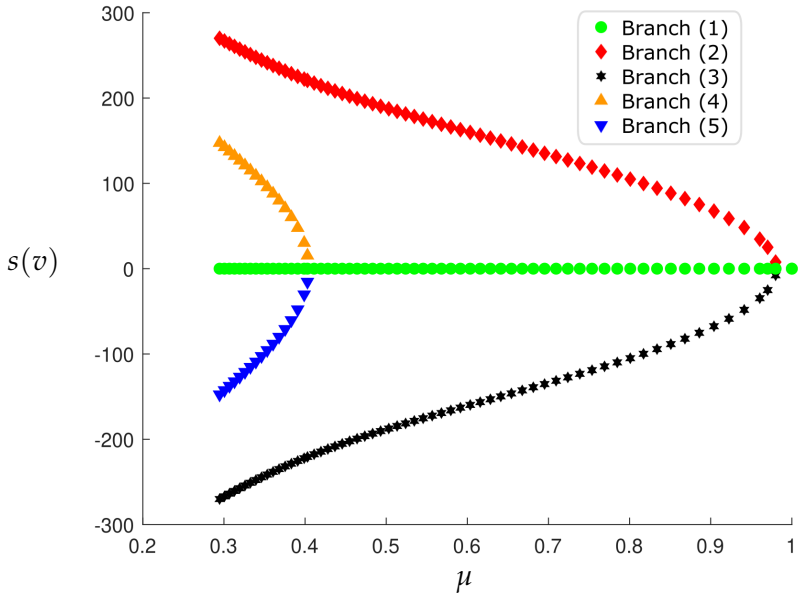
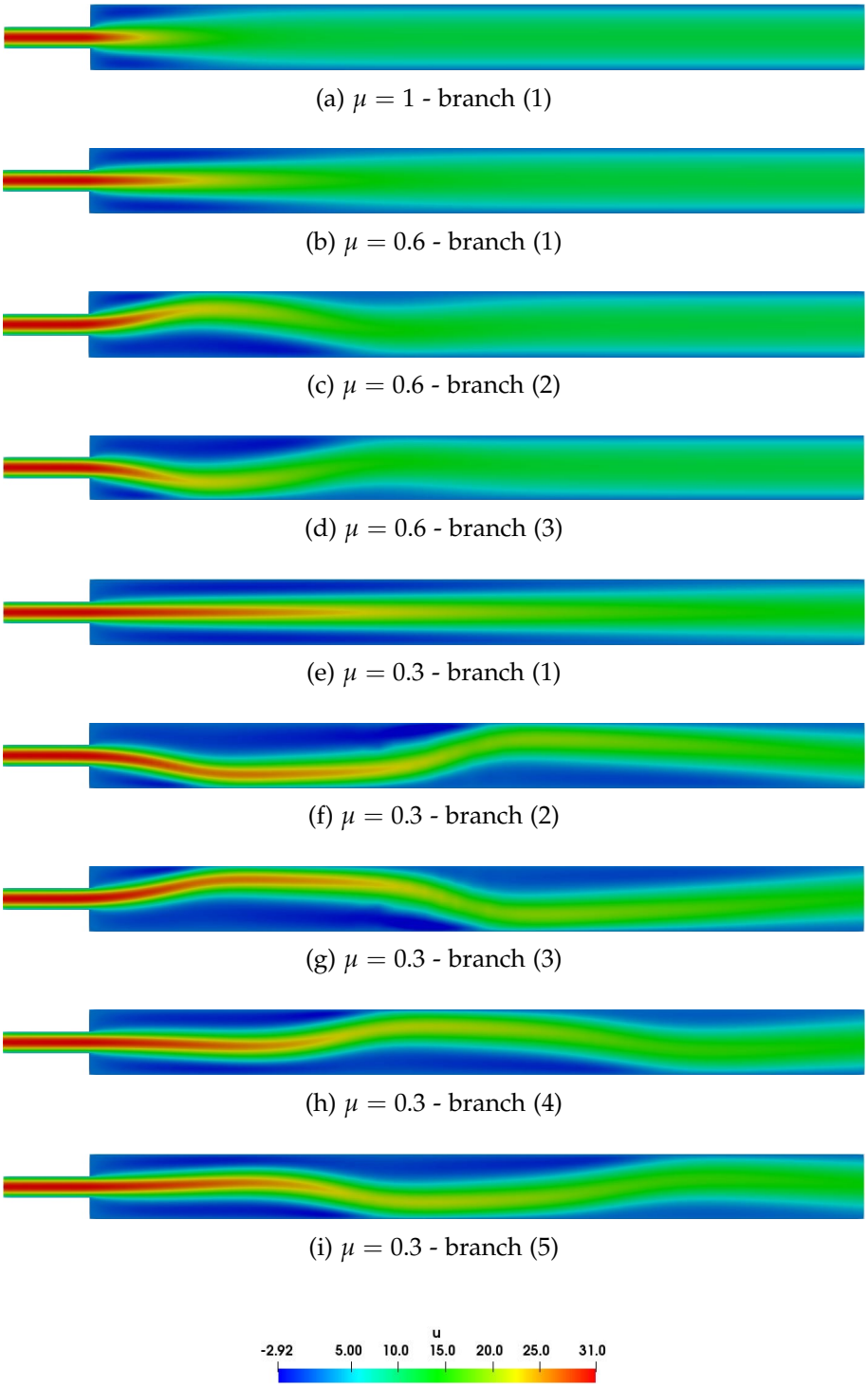


Figure 8.2: High fidelity bifurcation diagram for the Navier-Stokes system.

Let us show in Figure 8.3 some representative solutions for each corresponding branch in Figure 8.2. The symmetric branch is always characterized by the symmetric profile in Figures 8.3a, 8.3b and 8.3e. The first pitchfork bifurcation is represented by the upper/lower wall-hugging profiles in Figures 8.3d, 8.3f and 8.3c, 8.3g respectively. Finally, the second pitchfork bifurcation is represented by the asymmetric profiles, characterized by a delayed wall-hugging phenomenon, in Figures 8.3h and 8.3i.

Obtaining such a diagram in the high fidelity setting is very expensive, also having chosen the spectral element method as FOM. Thus, we applied our reduced strategy



(a) $\mu = 1$ - branch (1)

(b) $\mu = 0.6$ - branch (1)

(c) $\mu = 0.6$ - branch (2)

(d) $\mu = 0.6$ - branch (3)

(e) $\mu = 0.3$ - branch (1)

(f) $\mu = 0.3$ - branch (2)

(g) $\mu = 0.3$ - branch (3)

(h) $\mu = 0.3$ - branch (4)

(i) $\mu = 0.3$ - branch (5)

Figure 8.3: Representative high fidelity horizontal velocities for the five branches varying the viscosity $\mu \in \mathcal{P}$.

with $N_v = 37$ basis function, in order to efficiently obtain its approximated version. The latter is shown in Figure 8.4, where we have been able to reconstruct both pitchfork bifurcations with all corresponding branches.

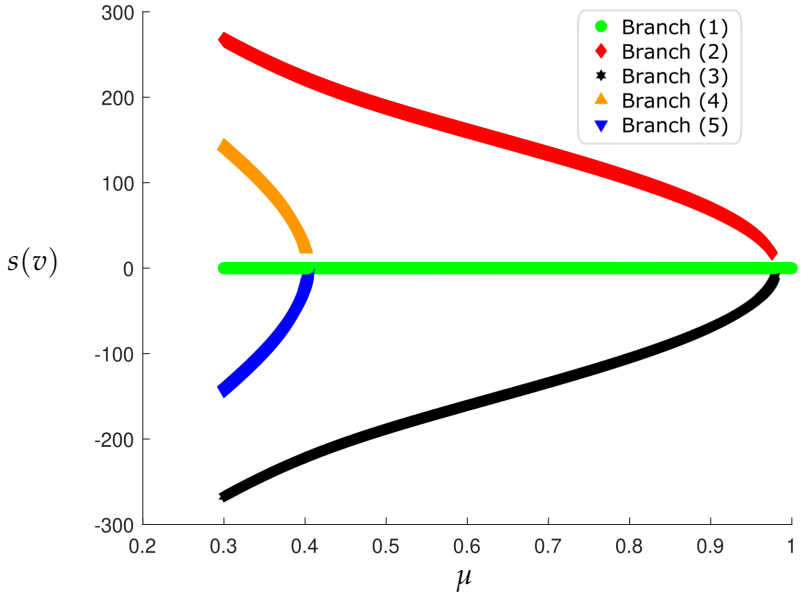


Figure 8.4: Reduced basis bifurcation diagram for the Navier-Stokes system.

Moreover, we highlight that in this case we do not need any empirical interpolation strategies, in fact an efficient assembly of the nonlinear term in a third order tensor was performed. For this reason, we pass from the high fidelity computational time required to build the diagram $t_{HF} = 1292$(s) to $t_{RB} = 293.85$(s) for its reduced version. But we remark that the online bifurcation diagram is approximated by means of $K = 1492$ solutions (vs. $K = 224$ for the offline one), thus the speed-up obtained is almost 290.

To end this analysis, we show in Figure 8.5 the exponential decay of the average and maximum errors with respect to the dimension of the reduced space. Such errors were computed over 300 reduced solution, by means of a ROM built upon 30 snapshots in $\mathcal{P} = [0.825, 1]$. It is also important to observe that the error stops decreasing when it is close to 10^{-10}, indeed this is the tolerance for both the offline and online solver.

Finally, we remark that an extension towards multi-parameter test cases is present in [122], where an additional parameter controls the maximum velocity at the inlet or its width. The latter case will be also investigated through a geometric parametrization by means of new methodologies in Chapter 10.

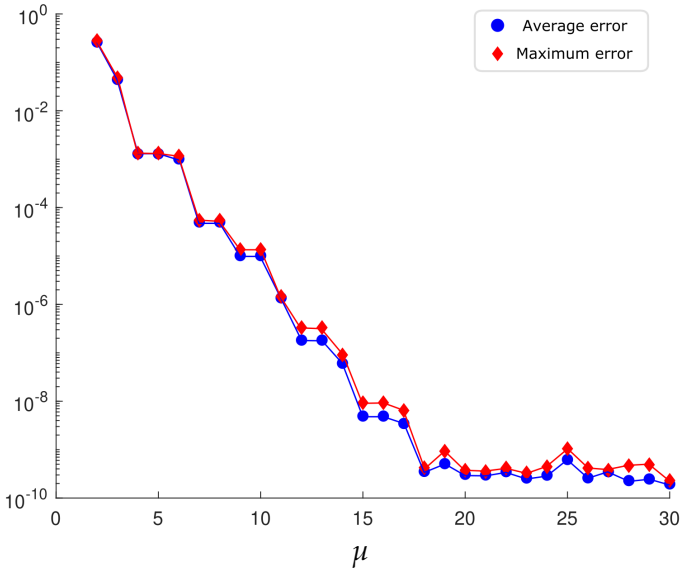


Figure 8.5: Exponential decay of the relative average and maximum errors over the dimension of the reduced basis space.

Chapter 9

An optimal control problem to steer bifurcating systems

Parametrized optimal control problems (OCPs) governed by PDEs are a very powerful tool which aim at driving the solution behaviour towards a convenient desired profile. To this end, OCPs can be interpreted as an input-output system which achieves an observable configuration [20, 66, 77, 152]. They have been exploited in several applications in different scientific fields, see e.g. [95] for an overview.

In this chapter, we thus aim at describing how optimal control theory allows to change the solution profile and stability of state solution branches. In particular, we considered the Navier-Stokes benchmark we have deeply investigated in the previous chapters and we treated it as the state equation of a general control problem. The main goal is to study the effect of several optimal control settings, describing the qualitative and quantitative effect of the control over the pitchfork bifurcation. Moreover, a detailed investigation of the stability eigenvalue analysis of the controlled state is performed. Therefore, optimal control is employed to drive bifurcating state profiles towards a different desired state, which might possibly belong to another state solution branch, playing as an attractor towards a desired configuration.

This part is entirely based on the work done with M. Strazzullo [121].

9.1 Optimal control problem for bifurcating PDEs

In this section, we introduce a generic nonlinear parametrized OCP. We will focus on the minimization of quadratic cost functional under nonlinear parametrized PDE constraint for Hilbert spaces, following the Lagrangian approach [66, 77]. However, the proposed analysis also holds in the more general Banach spaces setting. In the following sections we will provide existence results and optimality conditions for nonlinear parametrized OCPs in their continuous and discretized version, respectively.

9.1.1 Abstract formulation of the model

Optimal control is a mathematical tool which aims at modifying the natural behaviour of a system. Let us suppose to have a *state* PDE

$$G(y;\boldsymbol{\mu}) = f, \tag{9.1}$$

with *state variable* $y := y(\boldsymbol{\mu}) \in \mathbb{Y}$, i.e. $G : \mathbb{Y} \times \mathcal{P} \to \mathbb{Y}^*$ where $\mathbb{Y}$ is a Hilbert space, $f \in \mathbb{Y}^*$ is a forcing term, $\mathcal{P} \subset \mathbb{R}^P$ is a parameter space of dimension $P \geq 1$, while $G(y;\boldsymbol{\mu}) = E_{n\ell}(y;\boldsymbol{\mu}) + E_\ell(y;\boldsymbol{\mu})$ is the *state operator*, with $E_\ell \in \mathcal{L}(\mathbb{Y}, \mathbb{Y}^*)$ and $E_{n\ell}$ representing the linear and nonlinear contributions, respectively. Here, we call $\mathcal{L}(\cdot,\cdot)$ the space of linear continuous functions between two spaces. We now want y to be the most similar to a known solution profile $y_\mathrm{d} := y_\mathrm{d}(\boldsymbol{\mu}) \in \mathbb{Y}_\mathrm{obs} \supseteq \mathbb{Y}$. To this end, a new variable is introduced in the equation, namely the *control variable* $u := u(\boldsymbol{\mu}) \in \mathbb{U}$, with $\mathbb{U}$ another possibly different Hilbert space. Let us define the *controlled equation* $\mathcal{E}(y,u;\boldsymbol{\mu})$, where $\mathcal{E} : \mathbb{Y} \times \mathbb{U} \times \mathcal{P} \to \mathbb{Y}^*$. Then, the controlled equation will be of the following form:

$$\mathcal{E}(y,u;\boldsymbol{\mu}) := G(y;\boldsymbol{\mu}) - C(u) - f = 0,$$

where $C \in \mathcal{L}(\mathbb{U}, \mathbb{Y}^*)$ is the *control operator* describing the action of the variable u on the system. In other words, we are trying to change the behaviour of the state PDE through $C(u)$. The OCP reads: given a $\boldsymbol{\mu} \in \mathcal{P}$, find the pair $(y,u) \in \mathbb{Y} \times \mathbb{U}$ which solves

$$\min_{y\in\mathbb{Y}, u\in\mathbb{U}} J(y,u;y_\mathrm{d}) \text{ subject to } \mathcal{E}(y,u;\boldsymbol{\mu}) = 0, \tag{9.2}$$

where $J : \mathbb{Y} \times \mathbb{U} \times \mathbb{Y}_\mathrm{obs} \to \mathbb{R}$ is the *objective functional* defined by

$$J(y,u;y_\mathrm{d}) := \frac{1}{2}\|y - y_\mathrm{d}\|^2_{\mathbb{Y}_\mathrm{obs}} + \frac{\alpha}{2}\|u\|^2_{\mathbb{U}}, \tag{9.3}$$

and $\alpha \in (0,1]$ is a *penalization parameter*. The role of α is of great interest: indeed, a large value of α translates in a poor capability of the system to be controlled, while $\alpha \ll 1$ allows the functional to be minimized with larger values of the variable u. Problem (9.2) admits a solution if [77, Section 1.5.2]:

(i) $\mathbb{U}$ is convex, bounded and closed;

(ii) $\mathbb{Y}$ is convex and closed;

(iii) for every $\boldsymbol{\mu} \in \mathcal{P}$, the controlled system $\mathcal{E}(y,u;\boldsymbol{\mu}) = 0$ has a bounded solution map $u \in \mathbb{U} \mapsto y(u) \in \mathbb{Y}$;

(iv) for a given $\boldsymbol{\mu} \in \mathcal{P}$, the map $(y,u,\boldsymbol{\mu}) \in \mathbb{Y} \times \mathbb{U} \times \mathcal{P} \to \mathcal{E}(y,u;\boldsymbol{\mu}) \in \mathbb{Y}^*$ is weakly continuous with respect to the first two arguments;

(v) for a given $y_\mathrm{d} \in \mathbb{Y}_\mathrm{obs}$, the objective functional $J(y,u;y_\mathrm{d})$ is weakly lower semicontinous w.r.t. the first two arguments.

We now discuss the Lagrangian structure and the necessary first order optimality conditions. First of all, let $z := z(\boldsymbol{\mu}) \in \mathbb{Y}^{**} = \mathbb{Y}$ be an arbitrary variable called *adjoint variable*. Let us call $X = (y, u, z) \in \mathbb{X} := \mathbb{Y} \times \mathbb{U} \times \mathbb{Y}$ and let us build the *Lagrangian functional* $\mathscr{L} : \mathbb{X} \times \mathbb{Y}_{\text{obs}} \times \mathcal{P} \to \mathbb{R}$ as

$$\mathscr{L}(X; y_{\text{d}}, \boldsymbol{\mu}) := J(y, u; y_{\text{d}}) + \langle z, \mathcal{E}(y, u; \boldsymbol{\mu}) \rangle_{\mathbb{Y}\mathbb{Y}^*}, \tag{9.4}$$

where $\langle \cdot, \cdot \rangle_{\mathbb{Y}\mathbb{Y}^*}$ is the duality pairing of $\mathbb{Y}$ and $\mathbb{Y}^*$. The introduction of the adjoint variable allows to treat problem (9.2) in an unconstrained fashion finding the stationary point of (9.4). We remark that we consider z in the same space of the state variable for a proper definition of the discretized problem: we will clarify the reason in Section 9.1.2. Moreover, the variable X will inherit the parameter dependence by definition, i.e. $X := X(\boldsymbol{\mu})$. Furthermore, let us assume that the following holds:

(vi) $\mathbb{U}$ is nonempty;

(vii) $J : \mathbb{Y} \times \mathbb{U} \times \mathbb{Y}_{\text{obs}} \to \mathbb{R}$ and $\mathcal{E} : \mathbb{Y} \times \mathbb{U} \times \mathcal{P} \to \mathbb{Y}^*$ are continuously Fréchet differentiable w.r.t. the first two arguments;

(viii) given $\boldsymbol{\mu} \in \mathcal{P}$, the controlled system $\mathcal{E}(y, u; \boldsymbol{\mu}) = 0$ has a unique solution $y = y(u) \in \mathbb{Y}$ for all $u \in \mathbb{U}$;

(ix) given $\boldsymbol{\mu} \in \mathcal{P}$, $D_y \mathcal{E}(y, u; \boldsymbol{\mu}) \in \mathcal{L}(\mathbb{Y}, \mathbb{Y}^*)$ has a bounded inverse for all control variables u.

Assuming $(y, u) \in \mathbb{Y} \times \mathbb{U}$ to be a solution to (9.2) for a given $\boldsymbol{\mu} \in \mathcal{P}$, thanks to hypotheses (vi) - (ix), there exists an adjoint variable $z \in \mathbb{Y}$ such that the following variational system is satisfied [77]:

$$\begin{cases} D_y \mathscr{L}(X; y_{\text{d}}, \boldsymbol{\mu})[\omega] = 0 & \forall \omega \in \mathbb{Y}, \\ D_u \mathscr{L}(X; y_{\text{d}}, \boldsymbol{\mu})[\kappa] = 0 & \forall \kappa \in \mathbb{U}, \\ D_z \mathscr{L}(X; y_{\text{d}}, \boldsymbol{\mu})[\zeta] = 0 & \forall \zeta \in \mathbb{Y}, \end{cases} \tag{9.5}$$

or equivalently, in strong form

$$\begin{cases} y + D_y \mathcal{E}(y, u; \boldsymbol{\mu})^*(z) = y_{\text{d}}, \\ \alpha u - C^*(z) = 0, \\ \mathcal{E}(y, u; \boldsymbol{\mu}) = 0, \end{cases} \tag{9.6}$$

where $D_y \mathcal{E}(y, u; \boldsymbol{\mu})^* \in \mathcal{L}(\mathbb{Y}, \mathbb{Y}^*)$ is the adjoint operator of the Fréchet linearization of $\mathcal{E}(y, u; \boldsymbol{\mu})$ w.r.t. the state variable, while $C^* \in \mathcal{L}(\mathbb{Y}, \mathbb{U}^*)$ is the adjoint of the control operator. We will equivalently refer to problem (9.5) or (9.6) as *optimality system*. Moreover, writing the latter in compact form, it reads: given $\boldsymbol{\mu} \in \mathcal{P}$, find $X \in \mathbb{X}$ such that

$$\mathcal{G}(X; \boldsymbol{\mu}) = \mathcal{F}, \tag{9.7}$$

with

$$\mathcal{G}(X;\boldsymbol{\mu}) := \begin{bmatrix} y + D_y\mathcal{E}(y,u;\boldsymbol{\mu})^*(z) \\ \alpha u - C^*(z) \\ G(y,\boldsymbol{\mu}) - C(u) \end{bmatrix} \quad \text{and} \quad \mathcal{F} := \begin{bmatrix} y_\mathrm{d} \\ 0 \\ f \end{bmatrix}.$$

In the next section, we will discuss the FE approximation of a solution for a fixed value of the parameter to the nonlinear OCP , restricting ourselves to the well-posed setting presented.

9.1.2 Galerkin FE approximation

Here, we are interested in the numerical approximation of the solution branches $\mathcal{M}_i$ in (3.1) of the nonlinear OCP in (9.7), defined over an open and bounded regular domain $\Omega \subset \mathbb{R}^d$. Our aim is to discretize the problem at hand, in order to investigate its qualitative changes w.r.t. the values of the parameter. As we said, an overall viewpoint can be obtained through the bifurcation diagram, thus, in this section, we will review the building blocks to approximate a specific solution $X(\boldsymbol{\mu})$ for a chosen branch. We also remark that, even considering a single branch, the fulfillment of the well-posedness conditions can fail at some critical point $\boldsymbol{\mu}^*$. Hence, in the following, we assume $\boldsymbol{\mu} \neq \boldsymbol{\mu}^*$ and $X(\boldsymbol{\mu}) \in \mathcal{M}_i$ for some $i \in \{1,\dots,k\}$, thus we call $\mathcal{M}_i$ a *non-singular branch*. Furthermore, we assume the nonlinearity to be at most quadratic in the state variable.

To approximate the system in (9.7), let us consider the FE function space $\mathbb{X}_\mathcal{N} = \mathbb{Y}_{\mathcal{N}_y} \times \mathbb{U}_{\mathcal{N}_u} \times \mathbb{Y}_{\mathcal{N}_y} \subset \mathbb{X}$, of dimension $\mathcal{N} = 2\mathcal{N}_y + \mathcal{N}_u$. The FE approximation of the parametrized problem (9.7) reads: given $\boldsymbol{\mu} \in \mathcal{P}$, find $X_\mathcal{N} := X_\mathcal{N}(\boldsymbol{\mu}) \in \mathbb{X}_\mathcal{N}$ such that

$$\mathcal{G}(X_\mathcal{N};\boldsymbol{\mu}) = \mathcal{F}. \tag{9.8}$$

We now want to make the algebraic structure of the system (9.8) explicit. After the FE discretization, we can define y, u and z as the column vectors which entries are given by the FE coefficients of the state, control and adjoint variables in their approximated spaces, respectively. In the same fashion, we call y_d the column vector of the FE coefficients representing the desired state profile. Let us focus on the structure of the optimality problem. At the FE level, applying our controlled state to FE basis, we can derive the matrices $\mathsf{E}_{n\ell} + \mathsf{E}_\ell - \mathsf{C}$ and the forcing term vector f. Moreover, we define the mass matrices M_y and M_u for state/adjoint variables and control, respectively. We still need to understand the algebraic structure of $D_y\mathcal{E}(y,u;\boldsymbol{\mu})$. Assuming the operator G to be quadratic in the variable y, the Fréchet derivative of the controlled state equation w.r.t. the state y will be $\mathsf{E}'_{n\ell}[\mathsf{y}] + \mathsf{E}_\ell$. In other words, the linear state structure is preserved, the nonlinear operator is linearized in $\mathsf{E}'_{n\ell}[\mathsf{y}]$ and the control contribution

disappears. Then, the global matrix formulation of the optimization system (9.8) is

$$\overbrace{\underbrace{\begin{bmatrix} M_y & 0 & E'_{n\ell}[y]^T + E_\ell^T \\ 0 & \alpha M_u & -C^T \\ E_{n\ell} + E_\ell & -C & 0 \end{bmatrix}}_{G(X;\boldsymbol{\mu})} \underbrace{\begin{bmatrix} y \\ u \\ z \end{bmatrix}}_{X}}^{} = \overbrace{\begin{bmatrix} M_y y_d \\ 0 \\ f \end{bmatrix}}^{F}, \tag{9.9}$$

which in compact form reads:

$$R(X;\boldsymbol{\mu}) := G(X;\boldsymbol{\mu}) - F = 0, \tag{9.10}$$

where $R(X;\boldsymbol{\mu})$ represents the *global residual* of the optimality system. To solve system (9.10), we rely on Netwon method and we solve

$$X_{j+1} = X_j + \mathrm{Jac}(X_j;\boldsymbol{\mu})^{-1}(F - G(X_j;\boldsymbol{\mu})), \qquad j \in \mathbb{N}, \tag{9.11}$$

until a residual based convergence criterion is satisfied. We remark that, since in the controlled context J represents the objective functional, here and in what follows we will denote the Jacobian matrix by Jac

Since the matrix $E'_{n\ell}[y]^T$ still depends on y, the Jacobian matrix will be of the following nature:

$$\mathrm{Jac}(X_j;\boldsymbol{\mu}) = \begin{bmatrix} M_y + D_y(E'_{n\ell}[y_j]^T)[z^j] & 0 & E'_{n\ell}[y_j]^T + E_\ell^T \\ 0 & \alpha M_u & -C^T \\ E'_{n\ell}[y_j] + E_\ell & -C & 0 \end{bmatrix}, \tag{9.12}$$

where the matrix $D_y(E'_{n\ell}[y_j]^T)[z_j]$ does not depend anymore on the state, but only on the $j-$th value of the adjoint variable. This effort is useful to understand the *saddle point* structure of the Jacobian matrix. Indeed, one can write

$$\mathrm{Jac}(X_j;\boldsymbol{\mu}) = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}, \tag{9.13}$$

where

$$A = \begin{bmatrix} M_y + D_y(E'_{n\ell}[y_j]^T)[z_j] & 0 \\ 0 & \alpha M_u \end{bmatrix} \qquad \text{and} \qquad B = \begin{bmatrix} E'_{n\ell}[y_j] + E_\ell & -C \end{bmatrix}. \tag{9.14}$$

We remark that the assumption of at most quadratic nonlinearity in the state variable allows A to be symmetric and, thus, we will always refer to (9.13) as a saddle point structure. To guarantee the solvability of the system we consider A an invertible matrix. Furthermore, we need the following *Brezzi inf-sup condition* to be verified:

$$\beta_{Br,\mathcal{N}}(\boldsymbol{\mu}) := \inf_{0\neq z} \sup_{0\neq x} \frac{z^T B x}{\|x\|_{\mathbb{Y}\times\mathbb{U}} \|z\|_{\mathbb{Y}}} \geq \overline{\beta}_{Br,\mathcal{N}} > 0, \tag{9.15}$$

where $\mathsf{x} = \begin{bmatrix} \mathsf{y} \\ \mathsf{u} \end{bmatrix}$. In the FE context, the inequality (9.15) holds when the function spaces for state and adjoint coincide [109, 110]. The assumption $z \in \mathbb{Y}$, will guarantee the fulfillment of the inf-sup stability condition in the FE approximation. For general nonlinear problems, A may possibly be different form A^T, however the well-posedness results can be extended, the interested reader may refer to [18, 22].

9.1.3 ROMs for a generic nonlinear OCP

In this section we introduce the general setting for a ROM approximation of OCP problem in (9.8). The reduced strategy proposed is independent from the governing state equation. Indeed, we refer to [146, 147, 148, 156] for previous contributions to ROM for nonlinear OCPs, extending them to standard techniques proposed in [72, 120, 119, 122], for bifurcating systems.

As always we relied on POD-Galerkin basis construction [10, 26, 33, 74], combined with aggregated spaces techniques, following the linear OCPs fashion, as already presented in [8, 9, 49, 61, 84, 85, 109, 110, 132].

As in Chapter 7, here we exploited a branch-wise reduction, thus, for every bifurcating solution branch $\mathcal{M}_i$, we build a different ROM. For the moment, we suppose to have already constructed the reduced spaces $\mathbb{Y}_N \subset \mathbb{Y}_{\mathcal{N}_y} \subset \mathbb{Y}$ and $\mathbb{U}_N \subset \mathbb{U}_{\mathcal{N}_u} \subset \mathbb{U}$, the former for state and adjoint variables, the latter for control, respectively. After this reduced spaces building process, one can solve the following low-dimensional problem during the online phase: given $\boldsymbol{\mu} \in \mathcal{P}$, find $X_N(\boldsymbol{\mu}) = (y_N(\boldsymbol{\mu}), u_N(\boldsymbol{\mu}), z_N(\boldsymbol{\mu})) \in \mathbb{X}_N := \mathbb{Y}_N \times \mathbb{U}_N \times \mathbb{Y}_N$ such that it holds

$$\begin{cases} D_y\mathscr{L}(X_N; y_\mathrm{d}, \boldsymbol{\mu})[\omega] = 0 & \forall \omega \in \mathbb{Y}_N, \\ D_u\mathscr{L}(X_N; y_\mathrm{d}, \boldsymbol{\mu})[\kappa] = 0 & \forall \kappa \in \mathbb{U}_N, \\ D_z\mathscr{L}(X_N; y_\mathrm{d}, \boldsymbol{\mu})[\zeta] = 0 & \forall \zeta \in \mathbb{Y}_N. \end{cases} \tag{9.16}$$

As usual, exploiting the ROM is convenient only if one does not have to build from scratch the reduced model for any instance of the parameter. For this reason, the system (9.16) is assumed to be affinely decomposed, that translates in the following finite sums:

$$\begin{aligned} D_y\mathscr{L}(X_N, y_\mathrm{d}, \boldsymbol{\mu})[\omega] &= \sum_{\mathsf{q}=1}^{Q_y} \Theta_y^{\mathsf{q}}(\boldsymbol{\mu}) D_y\mathscr{L}^{\mathsf{q}}(X_N; y_\mathrm{d})[\omega], \\ D_u\mathscr{L}(X_N; y_\mathrm{d}, \boldsymbol{\mu})[\kappa], &= \sum_{\mathsf{q}=1}^{Q_u} \Theta_u^{\mathsf{q}}(\boldsymbol{\mu}) D_u\mathscr{L}^{\mathsf{q}}(X_N; y_\mathrm{d})[\kappa], \\ D_z\mathscr{L}(X_N; y_\mathrm{d}, \boldsymbol{\mu})[\zeta] &= \sum_{\mathsf{q}=1}^{Q_z} \Theta_z^{\mathsf{q}}(\boldsymbol{\mu}) D_z\mathscr{L}^{\mathsf{q}}(X_N; y_\mathrm{d})[\zeta], \end{aligned} \tag{9.17}$$

where $\Theta_y^{\mathfrak{q}}(\boldsymbol{\mu})$, $\Theta_u^{\mathfrak{q}}(\boldsymbol{\mu})$ and $\Theta_z^{\mathfrak{q}}(\boldsymbol{\mu})$ are $\boldsymbol{\mu}$-dependent smooth functions, while the terms $D_y\mathscr{L}^{\mathfrak{q}}(X_N, y_{\mathrm{d}})[\omega]$, $D_u\mathscr{L}^{\mathfrak{q}}(X_N; y_{\mathrm{d}})[\kappa]$ and $D_z\mathscr{L}^{\mathfrak{q}}(X_N; y_{\mathrm{d}})[\zeta]$ represent the $\boldsymbol{\mu}$-independent bilinear forms describing the optimality system.

When the affine dependence assumption is verified, the online phase does not depend on $\mathcal{N}$ and usually is performed in a small amount of time. Conversely, the offline process is performed only once and can take advantage of high performance computing resources. Since in this chapter we will deal with Navier-Stokes governing equations (7.1), then, it has at most quadratically nonlinear terms and the affine decomposition is not fulfilled. Thus, one can employ hyper-reduction techniques to recover it.

Now we focus on the ROM setting for the offline and online phase, showing the strategy employed to build the reduced function spaces. Exploiting the POD, N_{train} snapshots are sampled and then compressed in order to generate function spaces of dimension $N < N_{train}$.

It is well known that optimization governed by PDEs constraints leads to the solution of a saddle point system [18, 20, 77, 145], as already specified in Section 9.1.2. In order to guarantee the well-posedness of such a structure, the matrix B of system (9.13) must verify the inf-sup stability condition (9.15) for every $\boldsymbol{\mu} \in \mathcal{P}$. In the FE approximation, the aforementioned relation holds since state and adjoint spaces are equally discretized. However, the inf-sup stability must hold at the reduced level too, since the relation is provable if the reduced spaces for state and adjoint variables coincide. The standard POD construction process leads to the reduced function spaces for state and adjoint which may be different even under the assumption of the same starting FE spaces. To overcome this issue, the basis are usually manipulated in order to stabilize the system. Indeed, we applied *aggregated spaces* technique, as already did in several papers about ROM for OCPs, see [8, 9, 49, 61, 84, 85, 109, 110, 132] as references. The strategy aims at building a common space for state and adjoint which is able to describe both state and adjoint variables.

Let us suppose to have applied a standard POD for all the involved variables and to have defined the following spaces

$$\mathbb{Y}_N = \text{span } \{\chi_n^y, \chi_n^z,\ n = 1, \dots, N\}, \tag{9.18}$$

$$\mathbb{U}_N = \text{span}\{\chi_n^u,\ n = 1, \dots, N\}, \tag{9.19}$$

and with

$$\mathsf{Z} = \begin{bmatrix} \mathsf{Z}_{\mathsf{x}} \\ \mathsf{Z}_{\mathsf{z}} \end{bmatrix}, \qquad \text{and} \qquad \mathsf{Z}_{\mathsf{x}} = \begin{bmatrix} \mathsf{Z}_{\mathsf{y}} \\ \mathsf{Z}_{\mathsf{u}} \end{bmatrix}$$

where $\mathsf{Z}_{\mathsf{y}} \equiv \mathsf{Z}_{\mathsf{z}} = [\chi_1^y | \cdots | \chi_N^y | \chi_1^z | \cdots | \chi_N^z] \in \mathbb{R}^{\mathcal{N}_y \times 2N}$ and $\mathsf{Z}_{\mathsf{u}} = [\chi_1^u | \cdots | \chi_N^u] \in \mathbb{R}^{\mathcal{N}_u \times N}$ are the reduced basis matrices for each variable and Z spans the global space $\mathbb{X}_N$. We want to solve the optimality system in a low dimensional framework at each parametric instance. To this end, we employ a Galerkin projection into the reduced spaces and the system (9.7) becomes

$$\mathsf{G}_N(\mathsf{X}_N; \boldsymbol{\mu})\mathsf{X}_N = \mathsf{F}_N, \tag{9.20}$$

where

$$\mathsf{G}_N(\mathsf{X}_N;\boldsymbol{\mu}) := \mathsf{Z}^T\mathsf{G}(\mathsf{Z}\mathsf{X}_N;\boldsymbol{\mu}), \qquad \text{and} \qquad \mathsf{F}_N := \mathsf{Z}^T\mathsf{F}.$$

The system (9.20) is nonlinear: thus we can apply Newton method and we iteratively obtain

$$\mathsf{X}_N^{j+1} = \mathsf{X}_N^j + \mathsf{Jac}_N(\mathsf{X}_N^j;\boldsymbol{\mu})^{-1}(\mathsf{F}_N - \mathsf{G}(\mathsf{X}_N^j;\boldsymbol{\mu})\mathsf{X}_N^j), \qquad j \in \mathbb{N}. \tag{9.21}$$

from the FE approximation, the Frechét derivative inherits the saddle point structure, i.e.

$$\mathsf{Jac}_N(\mathsf{X}_N;\boldsymbol{\mu})\mathsf{X}_N = \begin{bmatrix} \mathsf{A}_N & \mathsf{B}_N^T \\ \mathsf{B}_N & 0 \end{bmatrix} \begin{bmatrix} \mathsf{x}_N \\ \mathsf{z}_N \end{bmatrix}, \tag{9.22}$$

with $\mathsf{Jac}_N(\mathsf{X}_N;\boldsymbol{\mu}) = \mathsf{Z}^T\mathsf{Jac}(\mathsf{Z}\mathsf{X}_N;\boldsymbol{\mu})\mathsf{Z}$, $\mathsf{A}_N = \mathsf{Z}_\mathsf{x}^T\mathsf{A}\mathsf{Z}_\mathsf{x}$ and $\mathsf{B}_N = \mathsf{Z}_\mathsf{z}^T\mathsf{B}\mathsf{Z}_\mathsf{x}$

We now have all the ingredients to define a *reduced Brezzi inf-sup condition* as follows

$$\beta_{\mathrm{Br},N}(\boldsymbol{\mu}) := \inf_{0\neq \mathsf{z}_N} \sup_{0\neq \mathsf{x}_N} \frac{\mathsf{z}_N^T\mathsf{B}_N\mathsf{x}_N}{\|\mathsf{x}_N\|_{\mathbb{Y}\times\mathbb{U}}\|\mathsf{z}_N\|_{\mathbb{Y}}} \geq \overline{\beta}_{\mathrm{Br},N} > 0. \tag{9.23}$$

If $\boldsymbol{\mu} \neq \boldsymbol{\mu}^*$, relation (9.23) is verified thanks to the aggregated space definition. We remark that this technique is increasing the dimension of the global reduced system from $3N$ to $5N$. However, it is usually much smaller then $\mathcal{N}$. For the sake of simplicity and for a consistent construction of state and adjoint space, here we always chose N_{train} and N equal for all the involved variables.

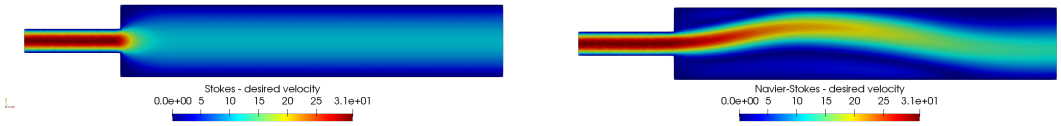
Remark 9.1.1 (Supremizer Stabilization). Dealing with Navier-Stokes governing equations, one has to take care not only of the global inf-sup condition (9.23), but also with the state equation inf-sup condition. Indeed, Navier-Stokes problem can be recast as saddle point itself and in an optimization system it creates a nested framework. The aggregated space techniques has to be accompanied by *supremizer stabilization* for the reduced velocity space. This approach [139] consists in defining a supremizer operator $T^{\boldsymbol{\mu}} : \mathbb{Q}^{\mathcal{N}_p} \to \mathbb{V}^{\mathcal{N}_v}$ as follows:

$$(T^{\boldsymbol{\mu}}s, \varphi)_{\mathbb{V}} = b(\varphi, s;\boldsymbol{\mu}), \qquad \forall \varphi \in \mathbb{V}^{\mathcal{N}_v}, \tag{9.24}$$

where $b(\cdot,\cdot)$ is the bilinear form representing the continuity equation defined in Section 7.1. Then, we enrich the reduced velocity space through the pressure supremizers as follows:

$$\mathbb{V}_N = \mathrm{span}\{\chi_n^v,\ \chi_n^{T_p}, \chi_n^w,\ \chi_n^{T_q},\ n = 1,\dots,N\},$$

where $\chi_n^{T_p}$ and $\chi_n^{T_q}$ are the basis supremizers obtained by state and adjoint pressure snapshots, respectively. Enlarging in this way the reduced space for velocity will guarantee inf-sup stability for the Navier-Stokes equation. This approach, i.e. supremizer stabilization combined with aggregated spaces, is the key for the well-posedness of the whole optimization problem (9.2). This will lead to a reduced system of dimension $13N$, which is still convenient compared to the global FE approximation dimension.



(a) (b)

Figure 9.1: Desired velocity profiles: (a) symmetric profile obtained as Stokes solution for $\mu = 1$; (b) asymmetric profile given by the physically stable Navier-Stokes solution for $\mu = 0.49$.

9.2 Driving Navier-Stokes solutions towards desired branches

In this section we focus on several OCP governed by Navier-Stokes equations (7.1) in the geometrical configuration of a contraction-expansion channel. We aim at understanding how different control problems can affect the solution behaviour discussed in Chapter 7 for the uncontrolled case, especially when bifurcation phenomena are taken into account. Indeed, we know that in such case the Coanda effect occurs and three different solutions coexist for values of μ below μ^*. This leads us to analyse the controlled systems, trying to reach state profiles which are different from the expected uncontrolled solution.

We thus follow the general procedure described in Section 9.1.1, and discuss the specific case of optimal control of Coanda effect as a relevant example of bifurcating state PDE. Nonetheless, the procedure adopted here is general and can be used in wide variety of other different applications.

For all the applications, we will simulate the physical phenomenon over the domain Ω shown in Figure 7.1. Moreover, for the OCPs structure, we will require the velocity solution $v \in \mathbb{V}$ to be the most similar to a desired profile $v_d \in \mathbb{V}_{obs} := (L^2(\Gamma_{obs}))^2$. The *observation domain* $\Gamma_{obs} = \{47\} \times [0, 7.5]$ is a line near the end of the channel. This structure allows the control to change the solution at the outflow following a prescribed convenient configuration. During the rest of this study, we will employ two velocity solution profiles, which are showed in Figure 9.1: we will denote them as the *symmetric desired profile (or target)* for 9.1a and the *asymmetric desired profile (or target)* for 9.1b. The first is the result of a Stokes system over Ω for $\mu = 1$ with the same boundary conditions of the Navier-Stokes uncontrolled equations (7.1). On the contrary, the latter is the physically stable solution of (7.1) for $\mu = 0.49$. While the former choice aims at controlling the system towards a globally symmetric configuration with a weaker outgoing flux, we introduced the latter for completeness, in order to achieve the opposite goal.

The purpose of steering the bifurcating behaviour is summarized in the minimization of the functional

$$J_{NS}(v, u; v_d) = \frac{1}{2}\|v - v_d\|^2_{\mathbb{V}_{obs}} + \frac{\alpha}{2}\|u\|^2_{\mathbb{U}}, \tag{9.25}$$

where $\mathbb{U} := (L^2(\Omega_u))^2$ with $\Omega_u \subseteq \overline{\Omega}$: indeed, the control action can be performed even

over a portion of the boundary $\partial\Omega$. We will refer to Ω_u as the *control domain*. Within this study we will analyse how the choice of Ω_u, combined with different values of the penalization parameter α, affects the solution behaviour of the system, compared to the uncontrolled Navier-Stokes state equation.

Now, we aim at recasting the OCP constrained to Navier-Stokes equations in the algebraic formulation presented in Section 9.1.1. The steady and incompressible controlled Navier-Stokes equations in a given domain Ω are:

$$\begin{cases} -\mu\Delta v + v\cdot\nabla v + \nabla p = C(u) & \text{in } \Omega, \\ \nabla\cdot v = 0 & \text{in } \Omega, \end{cases} \tag{9.26}$$

accompanied by some boundary conditions. The control operator $C : \mathbb{U} \to \mathbb{V}^*$ can represent an external forcing term or a boundary term. If C is defined in the whole domain we will say that the control is *distributed*, while if it is defined in a portion of the internal domain, we will deal with *localized control*. Furthermore, we will refer to *Neumann control* and *Dirichlet control*, if the control acts as Neumann or Dirichlet boundary conditions, respectively. Despite the variety of ways in which the control can act, the optimization system preserves a common structure that we are going to describe in the following.

The weak formulation of (9.26) reads: given $\mu \in \mathcal{P}$, find $v \in \mathbb{V}$, $p \in \mathbb{Q}$ and $u \in \mathbb{U}$ such that

$$\begin{cases} a(v,\psi;\mu) + s(v,v,\psi) + b(\psi,p) = c(u,\psi) & \forall\,\psi \in \mathbb{V}, \\ b(v,\pi) = 0 & \forall\,\pi \in \mathbb{Q}, \end{cases} \tag{9.27}$$

where $a(\cdot,\cdot;\mu)$, $b(\cdot,\cdot)$ and $s(\cdot,\cdot,\cdot)$ have been already defined in (7.4) while $c : \mathbb{U}\times\mathbb{Y} \to \mathbb{R}$ is a bilinear form associated to the operator C. First of all, to derive the optimality conditions, we need the adjoint variables $w \in \mathbb{V}$ and $q \in \mathbb{Q}$ for velocity and pressure, respectively. Let $X = ((v,p),u,(w,q)) \in \mathbb{X} := \mathbb{Y}\times\mathbb{U}\times\mathbb{Y}$ be an optimal solution, where $\mathbb{Y} := \mathbb{V}\times\mathbb{Q}$. The Lagrangian functional for this specific problem is

$$\begin{aligned} \mathscr{L}_{\mathrm{NS}}(X; v_{\mathrm{d}}, \boldsymbol{\mu}) = J_{\mathrm{NS}}(v,u;v_{\mathrm{d}}) \,+\, \mu\int_\Omega \nabla v\cdot\nabla w\,d\Omega \,+\, \int_\Omega (v\cdot\nabla v)\,w\,d\Omega \\ -\int_\Omega p\nabla\cdot w\,d\Omega \,+\, \int_\Omega q\nabla\cdot v\,d\Omega \,-\, c(u,w). \end{aligned} \tag{9.28}$$

The optimality system built through Frechét differentiation is given by:

$$\begin{cases} D_v\mathscr{L}_{\mathrm{NS}}(X; v_{\mathrm{d}}, \boldsymbol{\mu})[\varphi] = 0 & \forall\,\varphi \in \mathbb{V}, \\ D_p\mathscr{L}_{\mathrm{NS}}(X; v_{\mathrm{d}}, \boldsymbol{\mu})[\xi] = 0 & \forall\,\xi \in \mathbb{Q}, \\ D_u\mathscr{L}_{\mathrm{NS}}(X; v_{\mathrm{d}}, \boldsymbol{\mu})[\tau] = 0 & \forall\,\tau \in \mathbb{U}, \\ D_w\mathscr{L}_{\mathrm{NS}}(X; v_{\mathrm{d}}, \boldsymbol{\mu})[\psi] = 0 & \forall\,\psi \in \mathbb{V}, \\ D_q\mathscr{L}_{\mathrm{NS}}(X; v_{\mathrm{d}}, \boldsymbol{\mu})[\pi] = 0 & \forall\,\pi \in \mathbb{Q}, \end{cases} \tag{9.29}$$

where the first two equations form the *adjoint equation*, while differentiating w.r.t. the variable u leads to the *optimality equation*. Finally, the latter two relations coincide with

the equation (9.27). In particular, the adjoint equation has the following form:

$$\begin{cases} m(v,\varphi) + a(w,\varphi;\mu) + s(\varphi,v,w) + s(v,\varphi,w) + b(\varphi,q) = m(v_{\mathrm{d}},\varphi) & \forall\, \varphi \in \mathbb{V}, \\ b(w,\xi) = 0 & \forall\, \xi \in \mathbb{Q}, \end{cases} \tag{9.30}$$

while the optimality equation is given by

$$\alpha r(u,\tau) = c(\tau,w) \quad \forall\, \tau \in \mathbb{U}, \tag{9.31}$$

where $m : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ and $r : \mathbb{U} \times \mathbb{U} \to \mathbb{R}$ terms come from the Frechét derivative of (9.25) w.r.t. the velocity and control, respectively. They represent the L^2 scalar product in Γ_{obs} and Ω_u. Furthermore, we remark that $s(\varphi,v,w) + s(v,\varphi,w)$ is the linearization around v of the trilinear form $s(v,v,\varphi)$, by definition. Therefore, the strong formulation for (9.30) and (9.31) reads:

$$\begin{cases} v\mathbb{I}_{\Omega_{\mathrm{obs}}} - \mu\Delta w - v\cdot\nabla w + (\nabla v)^T w + \nabla q = v_{\mathrm{d}}\mathbb{I}_{\Omega_{\mathrm{obs}}} & \text{in } \Omega, \\ \nabla\cdot w = 0 & \text{in } \Omega, \\ \alpha u \mathbb{I}_{\Omega_u} = C^* w & \text{in } \Omega, \end{cases} \tag{9.32}$$

where $\mathbb{I}_{\Omega_u}$ and $\mathbb{I}_{\Omega_{\mathrm{obs}}}$ are the indicator functions of the control and observation domains, respectively. The global optimization problem reads: given $\mu \in \mathcal{P}$, find $X = ((v,p),u,(w,q)) \in \mathbb{X}$ such that (9.26) and (9.32) are verified.

We remark that, if we call $y := (v,p)$ and $z := (w,q)$, we recover the global algebraic formulation presented in Section 9.1.1 and the saddle point structure is preserved. Indeed, let us suppose to apply the Taylor-Hood approximation $\mathbb{P}^2$-$\mathbb{P}^1$ for state y and adjoint variable z. Furthermore, we discretize the space $\mathbb{U}$ with FE using $\mathbb{P}^2$ polynomials. Recalling the notation of Section 9.1.2, we define the quantities

$$\mathsf{y} = \begin{bmatrix} \mathsf{v} \\ \mathsf{p} \end{bmatrix}, \quad \mathsf{z} = \begin{bmatrix} \mathsf{w} \\ \mathsf{q} \end{bmatrix}, \quad \mathsf{M}_y = \begin{bmatrix} \mathsf{M}_v & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{and} \quad \mathsf{C} = \begin{bmatrix} \mathsf{C}_v \\ 0 \end{bmatrix}, \tag{9.33}$$

where $\mathsf{v}, \mathsf{p}, \mathsf{w}, \mathsf{q}$ are the column vectors of FE coefficients for state and adjoint, velocities and pressures respectively, while M_v is the mass velocity matrix and C_v derives by the bilinear form $c(\cdot,\cdot)$. Furthermore, the linearized state equation structure can be now expressed as

$$\mathsf{E}'_{n\ell}[\mathsf{y}^j] + \mathsf{E}_\ell = \begin{bmatrix} \mathsf{S}[\mathsf{v}^j] & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \mathsf{K} & \mathsf{D}^T \\ \mathsf{D} & 0 \end{bmatrix} = \begin{bmatrix} \mathsf{K} + \mathsf{S}[\mathsf{v}^j] & \mathsf{D}^T \\ \mathsf{D} & 0 \end{bmatrix}, \tag{9.34}$$

where K is the stiffness matrix associated to the bilinear form $a(\cdot,\cdot;\mu)$, D is the continuity equation matrix coming from $b(\cdot,\cdot)$ and $\mathsf{S}[\mathsf{v}^j]$ is the algebraic formulation of $s(v,\cdot,\cdot) + s(\cdot,v,\cdot)$ evaluated at the FE velocity basis functions. It remains to understand the specific structure of $\mathsf{D}_{\mathsf{y}}(\mathsf{E}'_{n\ell}[\mathsf{y}]^T)[\mathsf{z}^j]$ defined in (9.12). To this end, we define $s_{\mathrm{ad}}(v,w,\varphi)$ as the *adjoint operator* of the linearized trilinear form $s(v,v,\cdot)$ around the state velocity v. Applying $s_{\mathrm{ad}}(v,\cdot,\cdot)$ to the basis functions of $\mathbb{V}^{\mathcal{N}_v}$ will result in $\mathsf{S}[\mathsf{v}^j]^T$.

In the Jacobian matrix evaluation, a linearization of $s_{\mathrm{ad}}(w, v, \varphi)$ is performed not only in w, but also w.r.t. the variable v. This process will lead to

$$\mathsf{D}_{\mathsf{y}^j}(\mathsf{E}'_{n\ell}[\mathsf{y}^j]^T)[\mathsf{z}^j] = \begin{bmatrix} \mathsf{D}_{\mathsf{v}}(\mathsf{S}[\mathsf{v}^j]^T)([\mathsf{w}^j]) & 0 \\ 0 & 0 \end{bmatrix}, \tag{9.35}$$

where $\mathsf{D}_{\mathsf{v}}(\mathsf{S}[\mathsf{v}^j]^T)([\mathsf{w}^j])$ is given by the form $s_{\mathrm{ad}}(w, \cdot, \cdot)$ applied to the velocity basis functions. In other words, the whole linearized optimality system reads

$$\mathsf{Jac}_{\mathrm{NS}}(\mathsf{X}^j; \boldsymbol{\mu}) = \begin{bmatrix} \mathsf{M}_v + \mathsf{D}_{\mathsf{v}}(\mathsf{S}[\mathsf{v}^j]^T)[\mathsf{w}^j] & 0 & 0 & \mathsf{K}^T + \mathsf{S}[\mathsf{v}^j]^T & \mathsf{D}^T \\ 0 & 0 & 0 & \mathsf{D} & 0 \\ 0 & 0 & \alpha \mathsf{M}_u & -\mathsf{C}_v^T & 0 \\ \mathsf{K} + \mathsf{S}[\mathsf{v}^j] & \mathsf{D}^T & -\mathsf{C}_v & 0 & 0 \\ \mathsf{D} & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \mathsf{A} & \mathsf{B}^T \\ \mathsf{B} & 0 \end{bmatrix} \tag{9.36}$$

where X is the FE coefficient vector of the optimal solution and

$$\mathsf{A} = \begin{bmatrix} \mathsf{M}_v + \mathsf{D}_{\mathsf{v}}(\mathsf{S}[\mathsf{v}^j]^T)([\mathsf{w}^j]) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \alpha \mathsf{M}_u \end{bmatrix} \quad \text{and} \quad \mathsf{B} = \begin{bmatrix} \mathsf{K} + \mathsf{S}[\mathsf{v}^j] & \mathsf{D}^T & -\mathsf{C}_v \\ \mathsf{D} & 0 & 0 \end{bmatrix}. \tag{9.37}$$

As already specified in Section 9.1.2, we assume that for $\mu \neq \mu^*$ the saddle point (9.36) is well-posed. Moreover, we highlight that we are dealing with a *nested saddle point* structure: indeed, for the state equation (9.34) we require that, for a given $\mu \neq \mu^*$ and fixed v^j, the matrix $\mathsf{K} + \mathsf{S}[\mathsf{v}^j]$ is invertible and that Brezzi inf-sup condition holds, i.e.

$$\beta^{\mathcal{N}}_{\mathrm{Br,NS}} := \inf_{\mathsf{p} \neq 0} \sup_{\mathsf{v} \neq 0} \frac{\mathsf{p}^T \mathsf{D} \mathsf{v}}{\|\mathsf{v}\|_{\mathbb{V}} \|\mathsf{p}\|_{\mathbb{Q}}} \geq \overline{\beta}^{\mathcal{N}}_{\mathrm{Br,NS}} > 0. \tag{9.38}$$

This is indeed the case for the Taylor-Hood discretization.

In the next subsections we will analyse how the controlled problem behaves, comparing its properties with the ones of the uncontrolled system presented in Chapter 7. For the sake of notation, we will reserve the words symmetric/asymmetric to describe the profile of the desired velocity target or during the description of a visual representation of the obtained optimal solution. In contrast, we will use the word *natural optimal branch* to describe the branch that is obtained by running Algorithm 1 with a trivial initial guess. This branch may consist of either symmetric or asymmetric configurations, depending on the test case. Numerical experience shows that the natural optimal branch is the simplest branch to achieve by the optimal control system using the Algorithm 1, even for a wider range of initial guesses (obtained e.g. by random perturbation of the trivial guess). Further branches may exist, but are much harder to compute in practice and require very tailored initial guesses in a neighborhood of μ^*, and will be named *non-natural optimal branches*. For OCPs it makes no sense to talk about the physical stability of the global optimal solution. Indeed, the system itself is "artificially" built by adding non-physical adjoint variables, with the aim of changing the system behaviour.

In the next sections, we will provide the analysis of different optimal control systems, as follows.

Sec. 9.2.1. A *weak control* is built by controlling a Neumann boundary and the optimality system slightly affects the usual bifurcating nature of the uncontrolled Navier-Stokes equations.

Sec. 9.2.2. A *strong control* effect can be observed over the classical bifurcating behaviour of the uncontrolled solution by acting on the forcing term.

Sec. 9.2.3. The *penalization parameter* α is analysed while acting at the end of the inlet channel, and we discuss how changing α results in different orders of magnitude for the optimal control.

Sec. 9.2.4. We show how imposing different *boundary flux* conditions completely changes the known behaviour of the starting system.

Finally, in Section 9.2.5, final remarks and comparisons concerning the spectral analysis of the four test cases are presented.

9.2.1 Neumann control: weak steering

The first test case we present is a Neumann control over the boundary Γ_{out}, where homogeneous Dirichlet conditions are applied to $\Gamma_{\text{wall}} := \Gamma_0 \cup \Gamma_{\text{D}}$. More specifically, in this case, the optimality conditions read: given $\mu \in \mathcal{P}$ find $X \in \mathbb{X}$ such that

$$\begin{cases} v\mathbb{I}_{\Gamma_{\text{obs}}} - \mu\Delta w - v\cdot\nabla w + (\nabla v)^T w + \nabla q = v_{\text{d}}\mathbb{I}_{\Gamma_{\text{obs}}} & \text{in } \Omega, \\ \nabla\cdot w = 0 & \text{in } \Omega, \\ w = 0 & \text{on } \Gamma_{\text{in}} \cup \Gamma_{\text{wall}}, \\ -qn + (\mu\nabla w)n = 0 & \text{on } \Gamma_{\text{out}}, \\ \alpha u\mathbb{I}_{\Gamma_{\text{out}}} = w\mathbb{I}_{\Gamma_{\text{out}}} & \text{in } \Omega, \\ -\mu\Delta v + v\cdot\nabla v + \nabla p = 0 & \text{in } \Omega, \\ \nabla\cdot v = 0 & \text{in } \Omega, \\ v = v_{\text{in}} & \text{on } \Gamma_{\text{in}}, \\ v = 0 & \text{on } \Gamma_{\text{wall}}, \\ -pn + (\mu\nabla v)n = u & \text{on } \Gamma_{\text{out}}. \end{cases} \tag{9.39}$$

The desired velocity v_{d} will always be of the symmetric type for this specific example. In other words, we are studying which is the best choice for Neumann boundary condition, to reach the exiting symmetric profile shown in Figure 9.1a. We study the behaviour of the controlled solution varying $\alpha = 1, 0.1, 0.001, 0.0001$, where the greater is the value of α the lower is the strength of the control. In Figure 9.2 we show some representative solutions for $\alpha = 0.01$ and $\mu = 0.5$, for state velocity and pressure variables. In this case, the natural optimal branch is composed by asymmetric solutions (Figure 9.2, top), while there is a further non-natural optimal branch made up

by symmetric solutions (Figure 9.2, bottom). Results obtained following the natural optimal and non-natural optimal branches are shown in Figures 9.3 and 9.4, respectively. Therefore, we conclude that the Neumann control affects *weakly* the system, as it is not able to steer it towards the desired symmetric configuration after the bifurcation has occured, thus not changing drastically the features already observed for the uncontrolled state equations (see Figure 7.2).

The left plot of Figure 9.3 depicts the velocity profile magnitude over Γ_{obs} for the highest value of the Reynolds number when following the natural optimal branch. Even though the obtained velocity (marked by an orange line) is indeed different from the desired profile (denoted by a blue line), especially for what concerns peak values, we observe that the Neumann control straightens the flux near the end of the channel (compare the orange line to the green line, which represents the uncontrolled asymmetric profile), even when high Reynolds numbers are considered. The resulting profile is similar to the uncontrolled symmetric velocity (red line), even though full symmetry is not achieved. The action of the control variable is shown in the right plot of Figure 9.3 when changing the parameter μ following the natural optimal branch: the control is stronger for $\mu < \mu^*$ (i.e., when the wall-hugging phenomenon occurs and straightening in necessary), while it remains low in magnitude for $\mu > \mu^*$.

Similarly, the left plot of Figure 9.4 shows the velocity profile magnitude over Γ_{obs} for the highest value of the Reynolds number when following the non-natural optimal branch. In this case, the controlled symmetric profile (orange line) coincides with the uncontrolled symmetric profile (red line). Furthermore, the right plot of Figure 9.4 shows that the control variable around the critical μ^* (e.g., $\mu = 1$ and $\mu = 0.95$) is asymmetric to counteract the stable wall-hugging physically driven behaviour of the uncontrolled system. We further remark that, compared to the natural optimal branch, the control variable of the non-natural optimal branch is much lower in magnitude.

Table 9.1 shows the value of the cost functional (9.25) for several values of μ (rows) and α (columns), following either the natural optimal or non-natural optimal branches. The first column also shows the value of the *uncontrolled functional*, i.e. (9.25) evaluated for the uncontrolled velocity v of the equation (7.1) and zero control.

The main observation is that decreasing the value of α results in lower cost functional values, since a lower value of α allows stronger control to take place and drive the velocity to the desired configuration. In all cases, the non-natural branch presents lower values of the functional compared to the natural branch; this has to be expected, as the cost functional measures deviation from a symmetric target, and the non-natural branch is clearly closer to the target being made of symmetric solution (compare e.g. for $\mu = 0.05$ and $\alpha = 0.001$ the left panels of Figures 9.3–9.4). However, the natural branch is the one for which the control procedure is influencing the most the cost functional values: for instance, for $\mu = 0.05$ and $\alpha = 0.001$, the cost functional is decreased by 6% on the non-natural branch and by 55% on the natural one w.r.t. the corresponding uncontrolled configuration. Again, this has to be expected from the previous discussion of Figures 9.3–9.4, which shows larger impact of the control procedure to straighten the solution on the natural branch. Finally, large values of μ have negligible

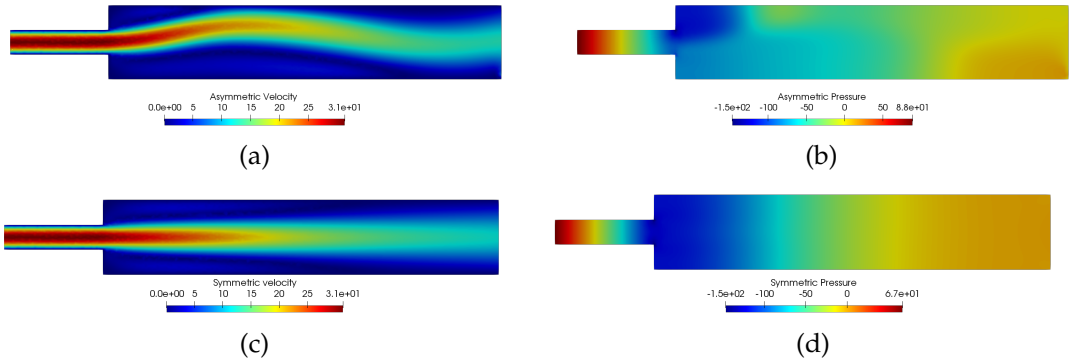


Figure 9.2: Optimal solutions with $\alpha = 0.01$ and $\mu = 0.5$, belonging to the natural optimal (panels (a) and (b) for state velocity and pressure, respectively) and the non-natural optimal (panels (c) and (d)) branches.

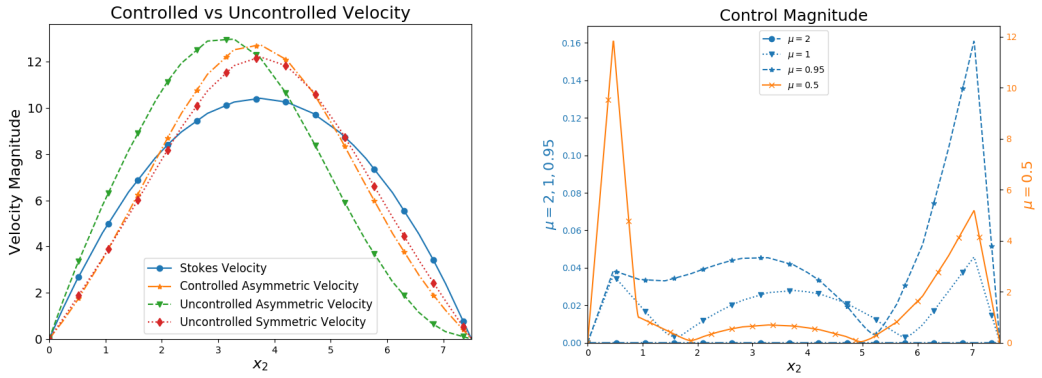


Figure 9.3: *Left*: comparison of velocity profiles in the controlled and uncontrolled cases for $\alpha = 0.01$, $\mu = 0.5$ on Γ_{obs} w.r.t. the desired profile when following the natural optimal branch. *Right*: representation of control variable evolution for $\alpha = 0.01$, $\mu = 2, 1, 0.95, 0.5$ over Γ_{out} when following the natural optimal branch.

cost functionals, as the target velocity almost coincides with the uncontrolled velocity. From such an analysis we deduce that, when bifurcating phenomena occur, a configuration can perform better than another one, and finding all the solution branches can be of great importance to understand the solution that best recover the desired profile.

Concerning the stability of the solution, we performed the eigenvalues analysis described in Algorithm 1. We can derive several informations from the Figure 9.5, which represents the global eigenvalue problem for the natural branch, against the parameter μ such that $\Re(\sigma_\mu) = [-0.01, 0.01]$.

We plot the first $N_{eig} = 100$ eigenvalues of the linearised system (3.12) around the global optimal solution, using a Krylov-Schur algorithm. From the plot, we observe

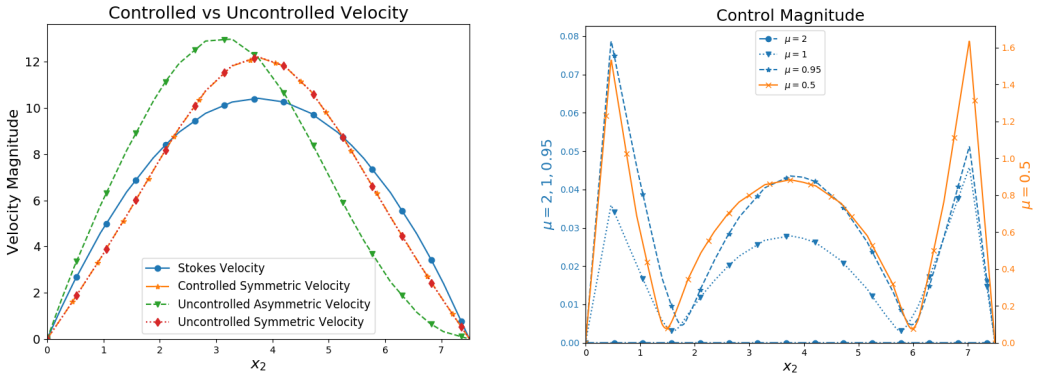


Figure 9.4: *Left*: comparison of velocity profiles in the controlled and uncontrolled cases for $\alpha = 0.01$, $\mu = 0.5$ on Γ_{obs} w.r.t. the desired profile when following the non-natural optimal branch. The lines marked by "Controlled Symmetric Velocity" and "Uncontrolled Symmetric Velocity" overlap. *Right*: representation of control variable evolution for $\alpha = 0.01$, $\mu = 2, 1, 0.95, 0.5$ over Γ_{out} when following the non-natural optimal branch.

two eigenvalues (highlighted with blue markers) approaching $\Re(\sigma_\mu) = 0$: we will refer to this behaviour as *shears phenomenon*. Moreover, the number of positive eigenvalues grows inversely with the value of the penalization parameter, and the negative eigenvalues are lowering except for the negative shear eigenvalue. Furthermore, the positive real eigenvalues accumulate in the value of α: this is very clear in subplots 9.5c and 9.5d. From the plot, a single eigenvalue (denoted by red markers) approaching zero is clearly visible.

One of the conclusion we can obtain from the global eigenvalue analysis is how the concentration of negative eigenvalues is affected by the greater action of the control variable obtained by decreasing α: for a fixed range of $\Re(\sigma_\mu)$, decreasing α (i.e., a more controlled system) results in larger number of positive eigenvalues in $\Re(\sigma_\mu)$.

Unfortunately, we can not derive information about the physical stability of the global solution from the performed global eigenvalue analysis, since similar eigenvalue structures are observed for both the natural and non-natural branches (only the former being shown here for the sake of brevity). Therefore, our considerations throughout the study are limited to the numerical stability represented by natural optimality, as discussed above.

Regardless, we thus conclude that the Neumann control is not able to fully steer uncontrolled solutions towards the desired symmetric configuration. However, this will be achieved in the next section, where a stronger control action will be presented.

Table 9.1: Comparison of the functional value for Neumann control w.r.t. stable and unstable uncontrolled solutions. (Nat.) Natural optimal branch. (n-Nat.) Non-natural optimal branch.

μ	Stable	Unstable	Nat.	n-Nat.	Nat.	n-Nat.	Nat.	n-Nat.	Nat.	n-Nat.
	Uncontrolled		$\alpha = 1$		$\alpha = 0.1$		$\alpha = 0.01$		$\alpha = 0.001$	
2	5.14e–9	5.14e–9	5.13e–9	5.13e–9	5.13e–9	5.13e–9	5.13e–9	5.13e–9	5.07e–9	5.07e–9
1.5	4.38e–6	4.38e–6	4.38e–6	4.38e–6	4.38e–6	4.38e–6	4.37e–6	4.37e–6	4.28e–6	4.28e–6
1	4.10e–3	4.10e–3	4.10e–3	4.10e–3	4.10e–3	4.10e–3	4.08e–3	4.10e–3	3.92e–3	3.92e–3
0.9	3.33e–2	1.63e–2	3.33e–2	1.63e–2	3.30e–2	1.63e–2	3.15e–2	1.63e–2	2.93e–2	1.55e–2
0.8	2.08e–1	6.52e–2	2.07e–1	6.52e–2	2.04e–1	6.51e–2	1.88e–1	6.51e–2	1.70e–1	6.15e–2
0.7	1.01e+0	2.59e–1	1.01e+0	2.59e–1	9.80e–1	2.59e–1	8.63e–1	2.59e–1	7.67e–1	2.43e–1
0.6	4.48e+0	1.70e+0	4.44e+0	1.02e+0	4.15e+0	1.02e+0	3.33e+0	1.02e+0	2.91e+0	9.57e–1
0.5	1.88e+1	3.92e+0	1.83e+1	3.92e+0	1.50e+1	3.92e+0	9.61e+0	3.92e+0	8.54e+0	3.68e+0

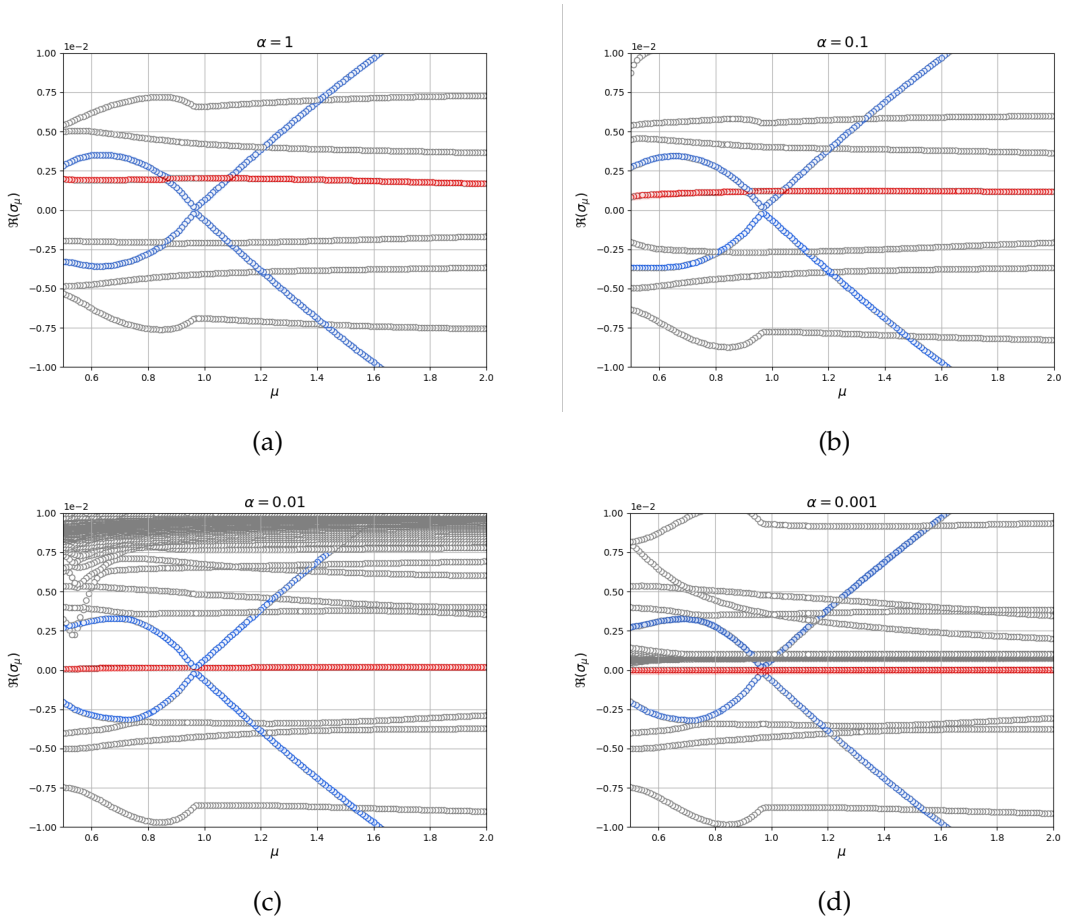


Figure 9.5: Spectral analysis of Neumann control with $\alpha = 1, 0.1, 0.01, 0.001$

.

9.2.2 Distributed control: strong steering

This section deals with a distributed control in $\Omega_u \equiv \Omega$, thus the control variable u acts as an external forcing term on the whole domain. Here we consider again $\Gamma_{\text{wall}} = \Gamma_0 \cup \Gamma_{\text{D}}$. Given $\mu \in \mathcal{P}$, the optimal solution $X \in \mathbb{X}$ satisfies the following system:

$$\begin{cases} v\mathbb{I}_{\Gamma_{\text{obs}}} - \mu\Delta w - v \cdot \nabla w + (\nabla v)^T w + \nabla q = v_{\text{d}}\mathbb{I}_{\Gamma_{\text{obs}}} & \text{in } \Omega, \\ \nabla \cdot w = 0 & \text{in } \Omega, \\ w = 0 & \text{on } \Gamma_{\text{in}} \cup \Gamma_{\text{wall}}, \\ -qn + (\mu\nabla w)n = 0 & \text{on } \Gamma_{\text{out}}, \\ \alpha u = w & \text{in } \Omega, \\ -\mu\Delta v + v \cdot \nabla v + \nabla p = u & \text{in } \Omega, \\ \nabla \cdot v = 0 & \text{in } \Omega, \\ v = v_{\text{in}} & \text{on } \Gamma_{\text{wall}}, \\ v = 0 & \text{on } \Gamma_0, \\ -pn + (\mu\nabla v)n = 0 & \text{on } \Gamma_{\text{out}}, \end{cases} \tag{9.40}$$

First of all, we underline that in distributed OCPs the action of the control is usually stronger and it affects deeply the original system.

To show this, we will steer the system towards either symmetric or asymmetric desired profiles v_{d}:

- *Symmetric target*: the aim of this setting is to steer the solution of (9.40) to a symmetric profile. We plot two representative control solutions in Figures 9.6a and 9.6b, obtained for $\mu = 2$ and $\mu = 0.5$ when following the natural optimal branch, which is composed of symmetric solutions. The stronger action of the control allows the controlled velocity profile to be more diffusive compared to the uncontrolled symmetric profile, as represented in the left plot of Figure 9.7, corresponding to the observed slice of the velocity solution for $\mu = 0.5$: in this case the controlled velocity (orange line) and the symmetric target (blue line) almost coincide. The right plot of Figure 9.7 shows that a slightly asymmetric control is required only near the critical value μ^* (also compare to Figures 9.6a and 9.6b for the cases $\mu = 2$ and $\mu = 0.5$). Furthermore, the control action is clearly higher when the Re value increases. Indeed, for $\mu = 2$ the control exclusively acts in the proximity of Γ_{obs} with a maximum magnitude of $1.8 \cdot 10^{-4}$, while for $\mu = 0.5$ it reaches a value of 1.6 of magnitude.

  A further non-natural optimal branch exists, and is made of symmetric solutions, but is hardly reachable by numerical continuation methods unless tailored guesses are provided to Algorithm 1.

- *Asymmetric target*: in this case, we desire to recover the asymmetric target for all $\mu \in \mathcal{P}$. We plot two representative control solutions in Figures 9.6c and 9.6d,

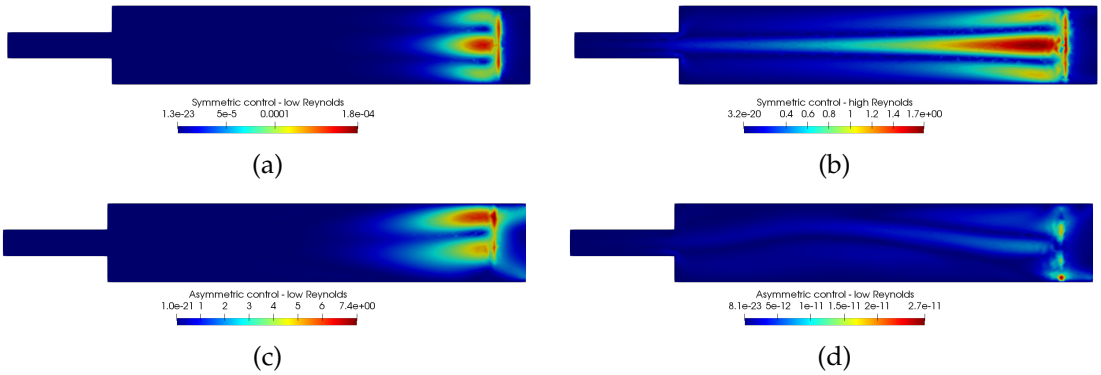


Figure 9.6: Optimal control profiles for $\alpha = 0.01$. Left: $\mu = 2$ in (a) and (c); right: $\mu = 0.5$ in (b) and (d). Top: symmetric target in (a) and (b); bottom: asymmetric target in (c) and (d).

obtained for $\mu = 2$ and $\mu = 0.5$ when following the natural optimal branch, which is made of asymmetric solutions. The action of the control is also visible in the left plot of Figure 9.8, obtained for $\mu = 2$: indeed, we see how the flux over Γ_{obs} is pushed towards the domain wall (orange line), in contrast to the symmetric profile of the uncontrolled velocity (green line). Namely, also in this case, the distributed control is able to drive the solution towards the desired state. In order to do so, the control variable has to be large when $\mu > \mu^*$, i.e. when the uncontrolled configuration on Γ_{obs} would lead to a symmetric profile. Indeed, in Figure 9.6c the maximum control value reaches 7 for $\mu = 2$ in the upper part of the domain. In contrast Figure 9.6d it lowers to 10^{-11} for $\mu = 0.5$ when the stable asymmetric velocity solution does not need to be controlled by an external forcing term. This is confirmed in the right plot of Figure 9.8 for several values of μ. Also in this case a non-natural optimal branch (featuring symmetric solutions) continues to exist, but it is numerically difficult to reach.

We show the comparison of the values of the cost functional (9.25) in Table 9.2 for the reached natural branch for both symmetric and asymmetric targets. Several values of μ (rows) and α (columns) have been analysed and compared to the uncontrolled functional, computed as in the Neumann test case. As expected, we notice that the functional is lower for smaller α. For the symmetric target, the action of the distributed control is indeed able to steer the solution towards the desired symmetric profile. Indeed, for $\mu = 0.05$, the choice $\alpha = 0.01$ shows that the functional is decreased by a 90% w.r.t. its uncontrolled counterpart, while for $\alpha = 0.001$, the cost functional is almost decreased by 99%. Similarly, for the asymmetric target, the maximum action of the control variable is given for low Reynolds and, for $\mu = 2$ we can observe a decrease of the functional of the 77.5% for $\alpha = 0.01$, up to a 97% for $\alpha = 0.001$. We remark that no control action is needed for $\mu = 0.5 \approx 0.49$, which is the parameter value for which the asymmetric v_{d} was computed: this was underlined by very low values of (9.25),

Table 9.2: Comparison of the functional value for distributed control. (Sym.) Natural optimal branch for symmetric target. (Asym.) Natural optimal branch for asymmetric target. (Sym.-U.) Unstable uncontrolled solution with symmetric target. (Asym.-S.) Stable uncontrolled solution with asymmetric target. (B.M.E.) Below machine epsilon.

μ	Sym.-U.	Asym.-S.	Sym.	Asym.	Sym.	Asym.	Sym.	Asym.	Sym.	Asym.
	Uncontrolled		$\alpha = 1$		$\alpha = 0.1$		$\alpha = 0.01$		$\alpha = 0.001$	
2	5.14e–9	1.88e+1	5.06e–9	1.81e+1	4.51e–9	1.36e+1	2.22e–9	4.23e+0	4.04e–10	5.66e–1
1.5	4.38e–6	1.88e+1	4.29e–6	1.77e+1	3.61e–6	1.20e+1	1.46e–6	3.09e+0	2.28e–7	3.87e–1
1	4.10e–3	1.86e+1	3.95e–3	1.67e+1	2.99e–3	9.15e+0	9.14e–4	1.86e+0	1.23e–4	2.17e–1
0.9	1.63e–2	1.84e+1	1.56e–2	1.54e+1	1.14e–2	7.88e+0	3.26e–3	1.50e+0	4.26e–4	1.73e–1
0.8	6.52e–2	1.54e+1	6.21e–2	1.31e+1	4.36e–2	6.06e+0	1.14e–2	1.08e+0	1.45e–3	1.22e–1
0.7	2.59e–1	1.15e+1	2.45e–1	9.28e+0	1.63e–1	3.68e+0	3.93e–2	6.16e–1	4.81e–3	6.90e–2
0.6	1.70e+0	5.34e+0	9.54e–1	3.76e+0	5.94e–1	1.24e+0	1.28e–1	2.00e–1	1.70e–2	2.22e–2
0.5	3.92e+0	B.M.E.	3.59e+0	B.M.E.	2.04e+0	B.M.E.	3.92e–1	B.M.E.	4.47e–2	B.M.E.

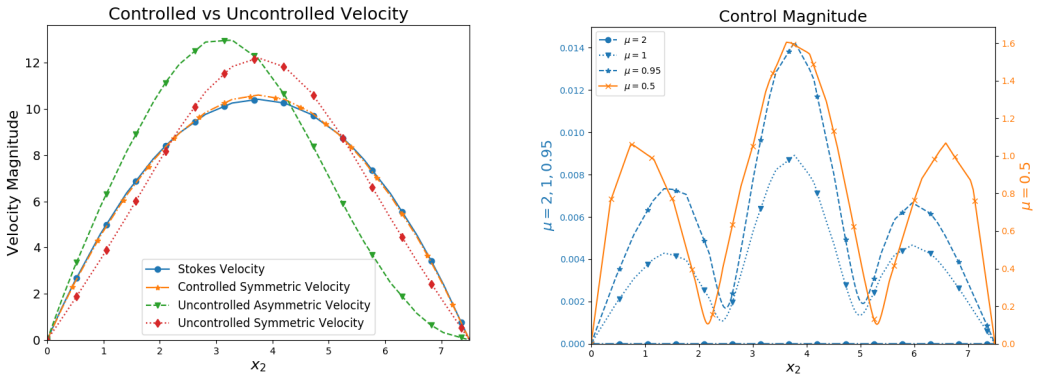


Figure 9.7: *Left*: comparison of velocity profiles in the controlled and uncontrolled cases for $\alpha = 0.01$, $\mu = 0.5$ on Γ_{obs} w.r.t. the symmetric desired profile when following the natural optimal branch. *Right*: representation of control variable evolution for $\alpha = 0.01$, $\mu = 2, 1, 0.95, 0.5$ for $x_1 = 45$ when following the natural optimal branch.

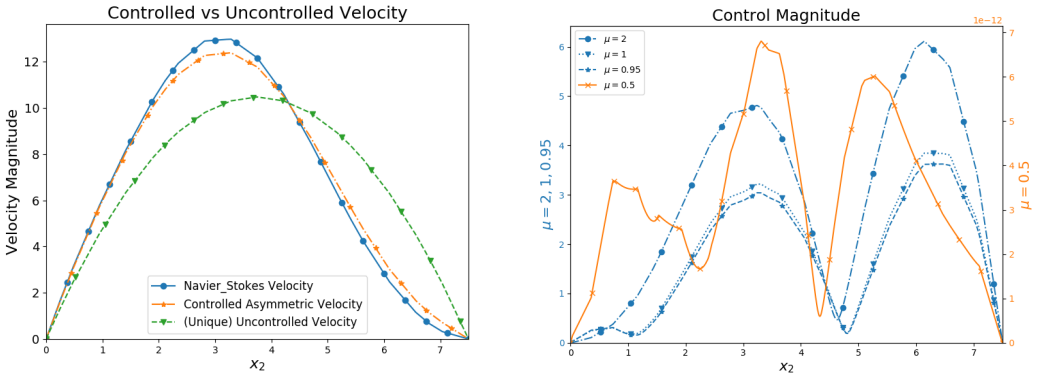


Figure 9.8: *Left*: comparison of velocity profiles in the controlled and uncontrolled cases for $\alpha = 0.01$, $\mu = 2.$ on Γ_{obs} w.r.t. the asymmetric desired profile when following the natural optimal branch. *Right*: representation of control variable evolution for $\alpha = 0.01$, $\mu = 2, 1, 0.95, 0.5$ for $x_1 = 45$ when following the natural optimal branch.

which were below the machine precision.

The plots of Figure 9.9 represent the spectral analysis for this optimal control problem: in particular, Figures 9.9a ($\alpha = 1$) and 9.9b ($\alpha = 0.1$) are associated with the symmetric target when following the corresponding natural optimal branch, while Figures 9.9c ($\alpha = 1$) and 9.9d ($\alpha = 0.1$) consider the asymmetric target when following its natural optimal branch. As the behavior between the top and bottom panels of Figure 9.9 is comparable, we will only comment in the following on the role of α. Furthermore, as we discussed in the previous test case, we cannot recover stability information about the configuration obtained by the optimal solution by means of the global eigenvalue problem, because the remaining (i.e., non-natural optimal) branches show very similar patterns to the ones in Figure 9.9.

We plot the eigenvalues for $\alpha = 1$ in $\Re(\sigma_\mu) = [-0.01, 0.01]$ and for $\alpha = 0.1$ in $\Re(\sigma_\mu) = [-0.005, 0.005]$. For this test case, the predominance of positive eigenvalues is visible also for large values of the penalization parameter. The smaller is α, the more negative eigenvalues are lowered. For all the α taken into account, the shears phenomenon does not appear: for $\alpha = 1$ a small trace of the shears structure is still visible (highlighted in blue) in Figures 9.9a and 9.9c where the bottom part of the shear is pushed away from $\Re(\sigma_\mu) = 0$. Instead, for $\alpha = 0.1$ the shears structure is completely lost: Figure 9.9c and 9.9d shows that only one eigenvalue (representing the top of the shears, and marked in blue) approaches $\Re(\sigma_\mu) = 0$ without crossing it.

We finally notice that the point μ^{**} where the upper shears curve is the closest to the axis $\Re(\sigma_\mu) = 0$ allows to obtain further information on the bifurcating phenomenon. From Figure 9.9, $\mu^{**} \approx 0.96$ for the symmetric target, regardless of α, while requiring an asymmetric target leads to $\mu^{**} \in [1.0, 1.2]$ with a mild dependence on α. Thanks to the aid of Figure 9.10, which shows the bifurcation diagram for the controlled solution with asymmetric target (a similar plot can be obtained for the

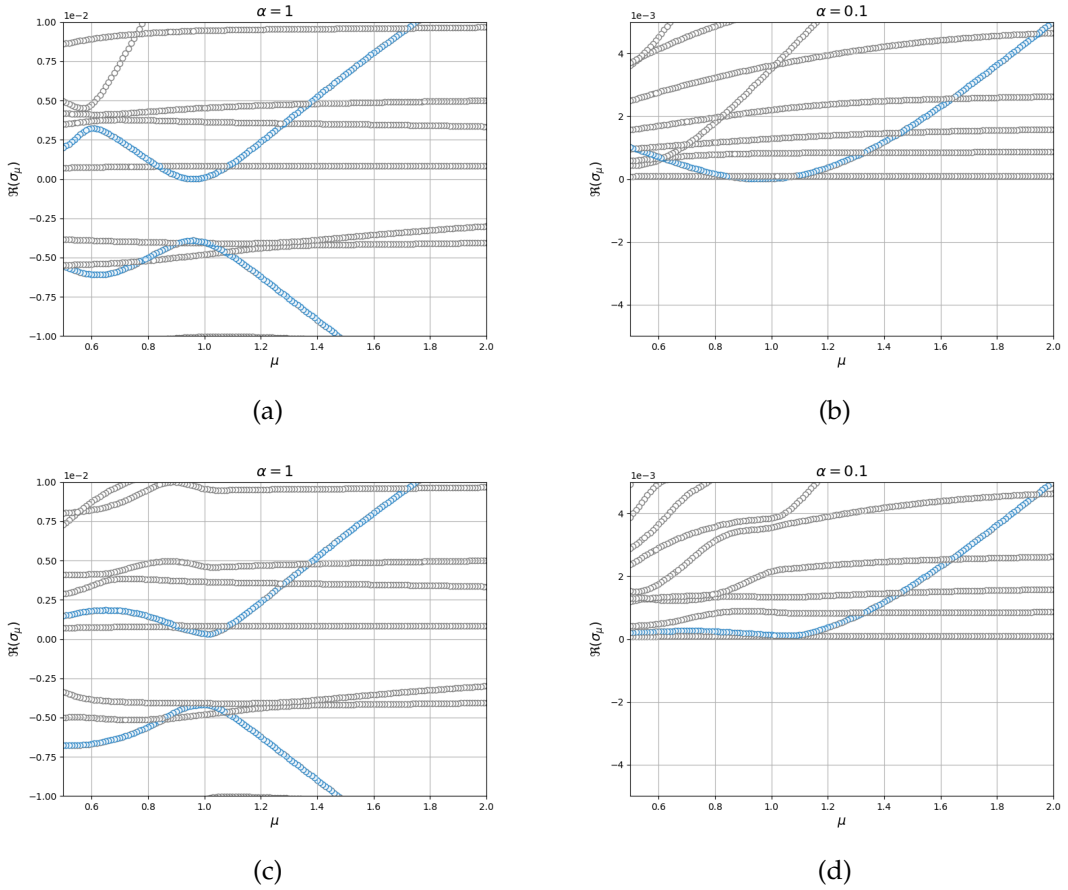


Figure 9.9: Spectral analysis with $\alpha = 1, 0.1$ (left to right) for the natural optimal branch with symmetric (top) and asymmetric (bottom) targets.

symmetric target as well, but is here omitted because the lines almost overlap), we can state that μ^{**} provides an indication on the location where bifurcation of the controlled system occurs.

We further note that for the symmetric target $\mu^{**} \approx \mu^*$, being μ^* the bifurcation point of the uncontrolled case, while $\mu^{**} \neq \mu^*$ for the asymmetric target, with a mild dependence on α. Thus, optimal control is not only able to steer the state solution towards a desired branch, but may also affect the location of the bifurcation point.

The role of the penalization parameter α will be clarified in the next section and it will result into a completely new optimal solution behaviour in section 9.2.4.

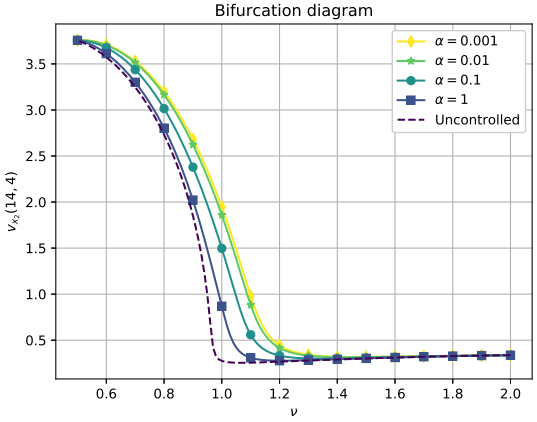


Figure 9.10: Bifurcation diagram (upper branch only) for controlled state velocity obtained with $\alpha = 1, 0.1, 0.01, 0.001$ and asymmetric target, compared to the uncontrolled velocity.

9.2.3 Channel control: the α effect

This section aims at describing how the value of the penalization parameter α can affect the natural convergence towards a symmetric target over Γ_{obs}. Towards this goal, we analysed the action of a control variable defined at the end of inlet channel, i.e. $\Omega_u = \Gamma_{\text{ch}}$, as depicted in Figure 7.1. The boundary Γ_{wall} is, once again, $\Gamma_0 \cup \Gamma_{\text{D}}$. Within this setting, the problem reads: given $\mu \in \mathcal{P}$ find the optimal solution $X \in \mathbb{X}$ such that the following holds:

$$
\begin{cases}
v\mathbb{I}_{\Gamma_{\text{obs}}} - \mu\Delta w - v \cdot \nabla w + (\nabla v)^T w + \nabla q = v_{\text{d}}\mathbb{I}_{\Gamma_{\text{obs}}} & \text{in } \Omega, \\
\nabla \cdot w = 0 & \text{in } \Omega, \\
w = 0 & \text{on } \Gamma_{\text{in}} \cup \Gamma_{\text{wall}}, \\
-qn + (\mu\nabla w)n = 0 & \text{on } \Gamma_{\text{out}}, \\
\alpha u\mathbb{I}_{\Gamma_{\text{ch}}} = w\mathbb{I}_{\Gamma_{\text{ch}}} & \text{in } \Omega, \\
-\mu\Delta v + v \cdot \nabla v + \nabla p = u\mathbb{I}_{\Gamma_{\text{ch}}} & \text{in } \Omega, \\
\nabla \cdot v = 0 & \text{in } \Omega, \\
v = v_{\text{in}} & \text{on } \Gamma_{\text{in}}, \\
v = 0 & \text{on } \Gamma_{\text{wall}}, \\
-pn + (\mu\nabla v)n = 0 & \text{on } \Gamma_{\text{out}}.
\end{cases}
\tag{9.41}
$$

The optimal control acts as a forcing term capable to change the way the flow enters in the expansion channel. In Figure 9.11 we show the adjoint velocity and pressure profiles obtained for $\mu = 0.5$ and two different penalization values, namely $\alpha = 1$ and $\alpha = 0.01$. In the first case, following Algorithm 1, the natural optimal branch presents a wall-hugging behaviour, while for smaller values of α the control variable

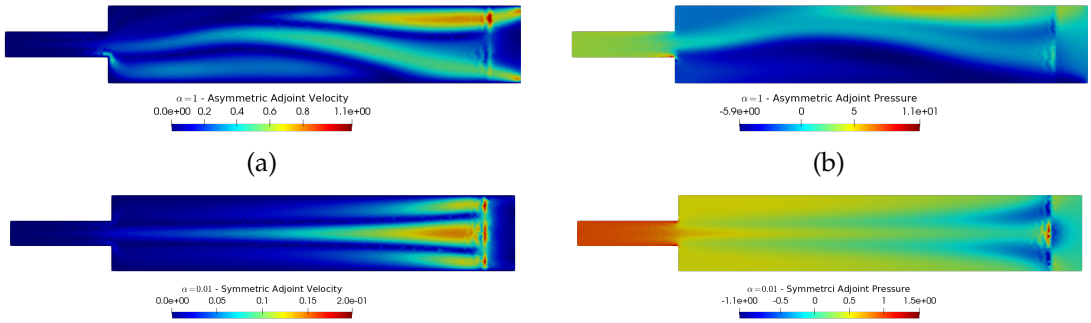


Figure 9.11: Two optimal solutions for adjoint velocity and pressure for $\mu = 0.5$: $\alpha = 1$ in (a) and (b), and $\alpha = 0.01$ in (c) and (d), respectively.

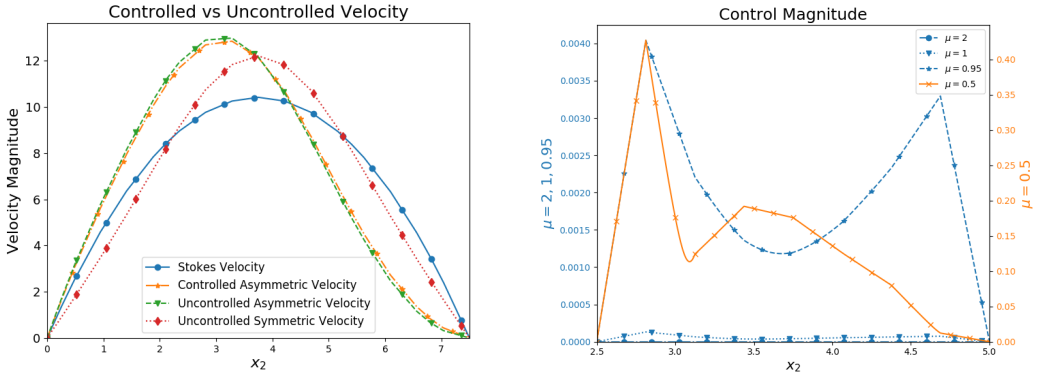


Figure 9.12: *Left*: comparison of velocity profiles in the controlled and uncontrolled cases for $\alpha = 1$, $\mu = 0.5$ on Γ_{obs} w.r.t. the symmetric desired profile when following the natural optimal branch. *Right*: representation of control variable evolution for $\alpha = 1$, $\mu = 2, 1, 0.95, 0.5$ at $x_1 = 10$ when following the natural optimal branch.

is able to drive the velocity towards a straight flux (see the left panels of Figures 9.12 and 9.13). Therefore, for large values of α the natural optimal branch is composed by asymmetric solutions (i.e., far away from the target), while for smaller values of α the natural optimal branch is made of symmetric solutions.

From the right plots of Figures 9.12 and 9.13, the control is very sensitive close to μ^* and this is shown by its asymmetric configuration both for the wall-hugging solution and the straight one. For $\alpha = 1, 0.1, 0.01$, we were able to detect two solutions using different initial guesses in the continuation method, showing symmetric and asymmetric features coexisting for some values of $\mu < \mu^*$. The smaller was α, the more difficult was to recover the non-natural branch. For example, when $\alpha = 0.001$, the action of the control variable drives the wall-hugging phenomenon towards a straight flux so strongly that we were not able to really reconstruct the whole optimal non-

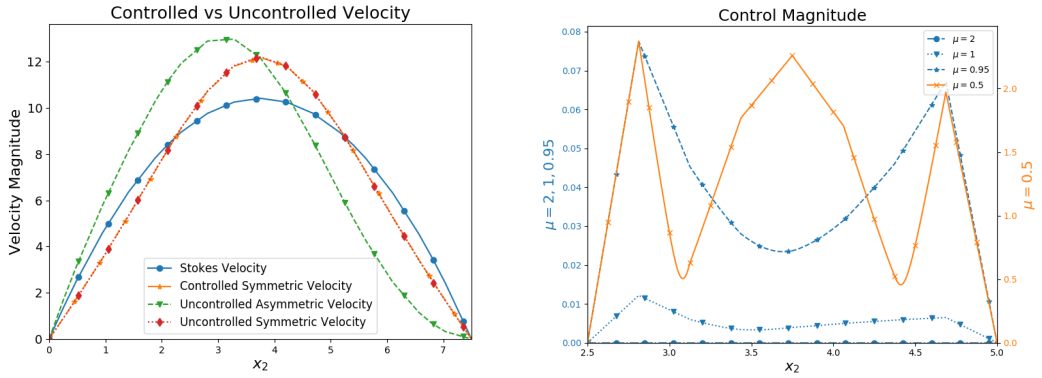


Figure 9.13: *Left*: comparison of velocity profiles in the controlled and uncontrolled cases for $\alpha = 0.01$, $\mu = 0.5$ on Γ_{obs} w.r.t. the symmetric desired profile when following the natural optimal branch. *Right*: representation of control variable evolution for $\alpha = 1$, $\mu = 2, 1, 0.95, 0.5$ for $x_1 = 10$ when following the natural optimal branch.

natural branch. Indeed, either the Newton solver did not converge (this happens also for $\alpha = 0.1$ and $\mu = 0.5$, compare Table 9.3) or converged to the natural branch consisting in symmetric features.

As usual, the role of α is highlighted in reducing objective functional, as Table 9.3 shows. As already specified in Sections 9.2.1 and 9.2.2, the straight configuration is lowering much more the functional than the asymmetric solution, due to its similarity with the symmetric v_{d}, which is the fixed target for this test case. In this case, the role of α is crucial in order to reach a solution which represents better the desired state. Indeed, the control was able to steer the solution to the symmetric profile for $\alpha = 0.1, 0.01, 0.001$. From the functional point of view, we do not have a notable decrease, as can be observed in Table 9.3, where the value of (9.25) is presented for different values of μ and α w.r.t. the uncontrolled problem solution. Yet, acting at the end of the inlet channel still allows to drive the optimal solution towards a natural convergence to the symmetric v_{d}, but the parabolic profile on Γ_{obs} is not reached (see e.g. that the functional decreases only of a 10% for $\mu = 0.5$ and $\alpha = 0.001$ w.r.t. the uncontrolled symmetric solution).

Figure 9.14 shows the eigenvalues of the global eigenproblem in the range $\Re(\sigma_\mu) = [-0.01, 0.01]$ when following the natural optimal branch. We remark that the eigenvalues behaviour is preserved also for the other non-natural optimal branch. For $\alpha = 1$, we can observe the shears phenomenon, which disappears for other values of the penalization parameter. Lowering the value of α, leads to a positive-dominated eigenvalues ensemble. Furthermore, a clustering around the value of α can be observed in plots 9.14c and 9.14d. In the next section, very peculiar features have been observed as well, while changing the value of the penalization parameter α in a Dirichlet control.

Table 9.3: Comparison of the functional value for channel control w.r.t. stable and unstable uncontrolled solutions. *Headers:* (Nat.) Natural optimal branch. (n-Nat.) Non-natural optimal branch. *Trailing cell characters*: (s) the solution has symmetric profile. (a) The solution has asymmetric profile. (nat-C.) Converging to natural branch despite tailored guess. (non-C.) Non-converging Newton's solver for tailored guess.

μ	Stable	Unstable	Nat.	n-Nat.	Nat.	n-Nat.	Nat.	n-Nat.	Nat.	n-Nat.
	Uncontrolled		$\alpha = 1$		$\alpha = 0.1$		$\alpha = 0.01$		$\alpha = 0.001$	
2	5.14e–9	5.14e–9	5.14e–9s	5.14e–9s	5.14e–9s	5.14e–9s	5.14e–9s	5.14e–9s	5.07e–9s	5.14e–9s
1.5	4.38e–6	4.38e–6	4.38e–6s	4.38e–6s	4.38e–6s	4.38e–6s	4.38e–6s	4.38e–6s	4.28e–6s	4.38e–6s
1	4.10e–3	4.10e–3	4.10e–3s	4.10e–3s	4.10e–3s	4.10e–3s	4.08e–3s	4.10e–3s	3.92e–3s	4.10e–3s
0.9	3.33e–2	1.63e–2	3.33e–2a	1.63e–2s	1.63e–1s	3.33e–2a	1.63e–1s	nat-C.	2.93e–2s	non-C.
0.8	2.08e–1	6.52e–2	2.08e–1a	6.52e–2s	6.52e–2s	2.07e–1a	6.52e–2s	2.04e–1a	6.51e–2s	nat-C.
0.7	1.01e+0	2.59e–1	1.01e+0a	2.59e–1s	2.59e–1s	1.01e+0a	2.59e–1s	9.76e–1a	2.24e–1s	nat-C.
0.6	4.48e+0	1.70e+0	4.48e+0a	1.02e+0s	1.02e+0s	4.43e+0a	1.02e+0s	4.03e+0a	9.90e–1s	nat-C.
0.5	1.88e+1	3.92e+0	1.87e+1a	3.92e+0s	3.92e+1s	non-C.	3.87e+0s	non-C.	3.50e+0s	nat-C.

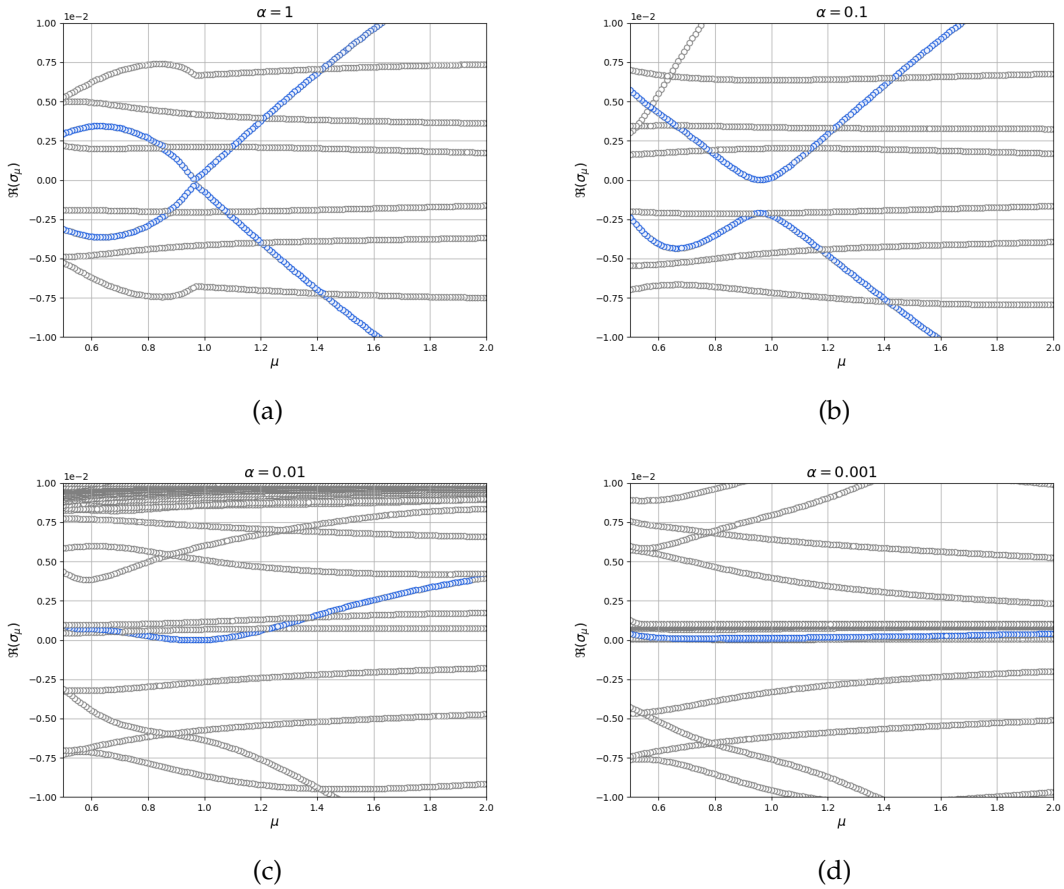


Figure 9.14: Spectral analysis with $\alpha = 1, 0.1, 0.01, 0.001$ when following the natural optimal branch.

9.2.4 Dirichlet control: flux action

In this final example, we propose a Dirichlet control over the boundary $\Omega_u \equiv \Gamma_D$. We fix the symmetric configuration as desired state v_d over the line Γ_{obs}, while we set $\Gamma_{wall} = \Gamma_0$. In other words, we are trying to control a Dirichlet boundary condition in order to lead the controlled solution towards the symmetric profile. The problem to be solved reads: given $\mu \in \mathcal{P}$ find $X \in \mathbb{X}$ such that

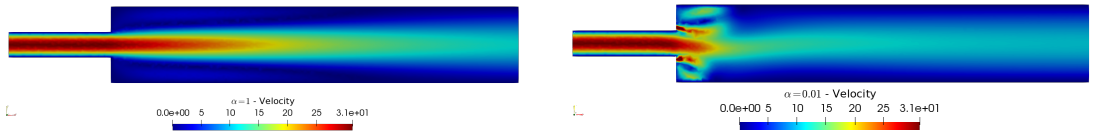
$$\begin{cases} v\mathbb{I}_{\Gamma_{obs}} - \mu\Delta w - v \cdot \nabla w + (\nabla v)^T w + \nabla q = v_d \mathbb{I}_{\Gamma_{obs}} & \text{in } \Omega, \\ \nabla \cdot w = 0 & \text{in } \Omega, \\ w = 0 & \text{on } \Gamma_{in} \cup \Gamma_D \cup \Gamma_{wall}, \\ -qn + (\mu\nabla w)n = 0 & \text{on } \Gamma_{out}, \\ \alpha u \mathbb{I}_{\Gamma_D} = w\mathbb{I}_{\Gamma_D} & \text{in } \Omega, \\ -\mu\Delta v + v \cdot \nabla v + \nabla p = 0 & \text{in } \Omega, \\ \nabla \cdot v = 0 & \text{in } \Omega, \\ v = v_{in} & \text{on } \Gamma_{in}, \\ v = u & \text{on } \Gamma_D, \\ v = 0 & \text{on } \Gamma_{wall}, \\ -pn + (\mu\nabla v)n = 0 & \text{on } \Gamma_{out}. \end{cases} \tag{9.42}$$

Allowing the flux to freely enter or exit from the boundary Γ_D drastically changes the optimal solution behaviour. Since we are asking for a symmetric desired profile, the main action of the control is to straighten the flow: this behaviour can be observed from Figure 9.15a and the left plot of Figure 9.16. Indeed, even for large values of α, the velocity profile reaches the symmetric configuration, while for lower values of the penalization parameter, the velocity in Γ_{obs} is parabolic.

This feature is highlighted also from the functional values in Table 9.4, where the functional (9.25) is shown for several μ (rows) and α (columns) w.r.t. uncontrolled stable and unstable solutions. The cost functional largely decreases for smaller values of α, e.g. $\alpha = 0.001$: for example, focusing on $\mu = 0.5$, the functional only lowers of 18% for $\alpha = 0.01$, in contrast to almost 82% for $\alpha = 0.001$.

Within the setting of $\alpha = 0.001$, the system manifests an interesting and unexpected profile, shown in Figure 9.15b. The flux presents an asymmetric configuration for low value of μ. Namely, for low values of α a bifurcating solution appears as depicted in Figure 9.15b. The asymmetric behaviour is due to the control variable which not only allows the flow to exit from Γ_D (in order to avoid the asymmetric recirculation of the wall-hugging solution), but it is adding flux near the channel in order to achieve the straight configuration and the parabolic velocity profile given by the symmetric target velocity over the observation domain, as it is represented in the right plot of Figure 9.16.

The eigenvalue analysis is provided in Figure 9.17, where we show some close-ups starting with $\Re(\sigma_\mu) = [-0.001, 0.001]$ for $\alpha = 1$ in the top-right image, and restricting



(a) (b)

Figure 9.15: Two optimal velocity solutions for $\mu = 0.5$, with $\alpha = 1$ and $\alpha = 0.001$, left and right respectively.

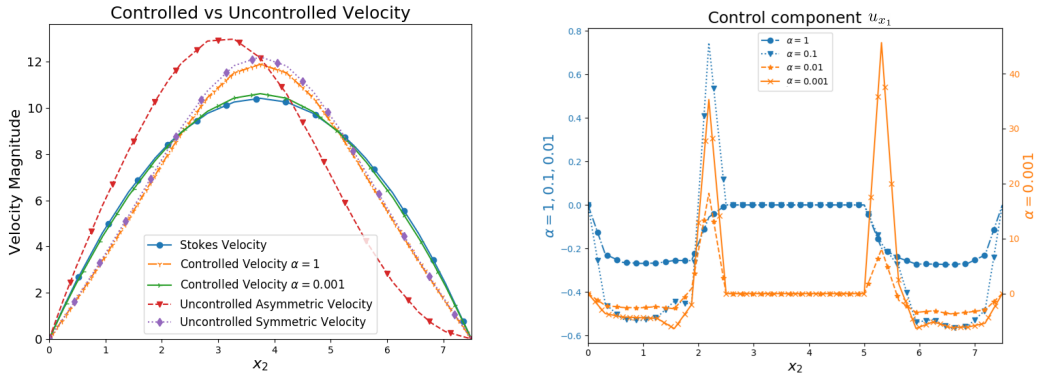


Figure 9.16: *Left*: comparison of velocity profiles in the controlled and uncontrolled cases for $\alpha = 1, 0.01$, $\mu = 0.5$ on Γ_{obs} w.r.t. the symmetric desired profile when following the natural optimal branch. *Right*: representation of control variable evolution for $\alpha = 0.1, 0.01, 0.001, 0.001$, at $x_1 = 10$ when following the natural optimal branch.

Table 9.4: Comparison of the functional value for Dirichlet control w.r.t. the stable and unstable uncontrolled solutions.

μ	Stable	Unstable	Controlled Solution			
	Uncontrolled		$\alpha = 1$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.001$
2	5.14e–9	5.14e–9	4.98e–9	4.83e–9	4.79e–9	4.79e–9
1.5	4.38e–6	4.38e–6	4.24e–6	4.10e–6	4.07e–6	4.06e–6
1	4.10e–3	4.10e–3	3.94e–3	3.78e–3	3.74e–3	3.72e–3
0.9	3.33e–2	1.63e–2	1.56e–2	1.49e–2	1.47e–2	1.45e–2
0.8	2.08e–1	6.52e–2	6.20e–2	5.88e–2	5.78e–2	5.46e–2
0.7	1.01e+0	2.69e–1	2.44e–1	2.29e–1	2.21e–1	1.82e–1
0.6	4.48e+0	1.70e+0	9.49e–1	8.73e–1	8.09e–1	3.57e–1
0.5	1.88e+1	3.92e+0	3.58e+0	3.21e+0	2.41e+0	4.73e–1

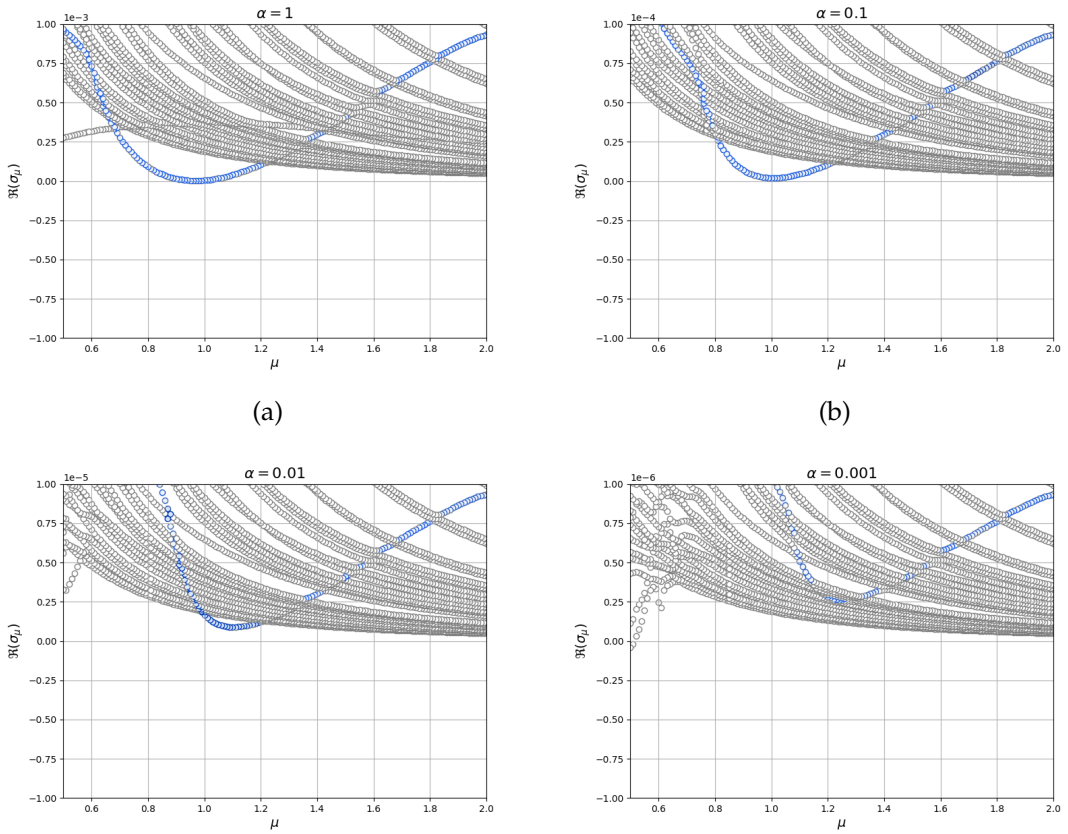


Figure 9.17: Spectral analysis with $\alpha = 1, 0.1, 0.01, 0.001$

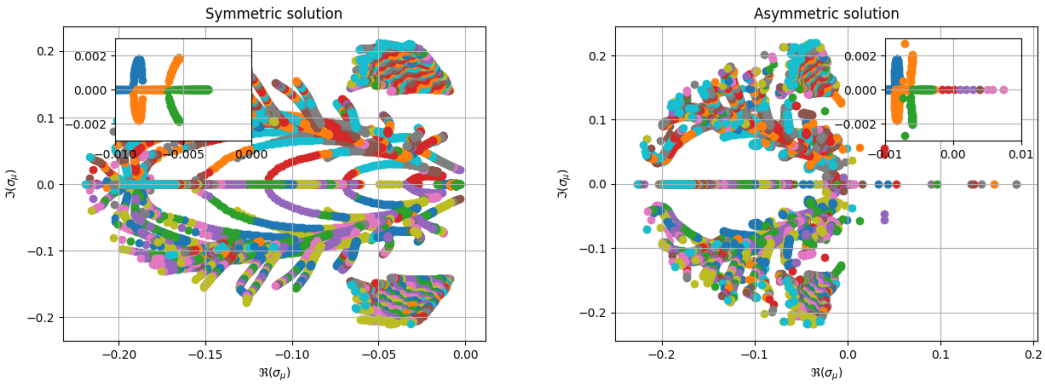


Figure 9.18: Eigenvalues of the state eigenproblem in the complex plane for the Dirichlet optimal control: asymmetric and symmetric solutions, left and right respectively.

the vertical interval due to the order of the lowered α in the remaining images. As already noticed in Section 9.2.2, the strongest is the control, the more the eigenvalues are positive. Furthermore, as it already happened in the distributed control case with asymmetric target in Section 9.2.2, the value μ^* seems to be not relevant anymore as an indication of the bifurcation point. Recalling the definition of μ^{**} in Section 9.2.2 as the value of the parameter for which the top curve of the shears (marked in blue) is approaching $\Re(\sigma_\mu) = 0$ from above, Figure 9.17 shows that such a curve is moving away from $\Re(\sigma_\mu) = 0$ as α decreases, and thus no such point μ^{**} exists.

The results of the previous sections have shown that the top shear structure is typically associated to the wall-hugging bifurcation, and that μ^{**} provides an indication of the bifurcation: we are thus lead to believe that the standard bifurcating configuration observed in the uncontrolled case, consisting of a branch of symmetric solutions and a branch of wall-hugging ones, is not present here, with the latter branch disappearing. However, the system seems to be featuring a *different bifurcation*, presented in Figure 9.15b. Indeed, we can see an eigenvalue crossing the line $\Re(\sigma_\mu) = 0$ for the global eigenproblem in Figure 9.17d for $\alpha = 0.001$.

Furthermore, it is the first time that we can see an eigenvalue crossing the line $\Re(\sigma_\mu) = 0$ for the global eigenproblem, as depicted in Figure 9.17d for $\alpha = 0.001$. We believe that this structure is strictly connected to the bifurcating solution presented in Figure 9.15b. Therefore, if we plot the eigenvalues of the state eigenproblem of Algorithm 1 (Figure 9.18), we see how the symmetric profile does never cross the origin, while the asymmetric solution in Figure 9.15b for $\alpha = 0.001$ does. In the setting with the modified BCs, the physical stable solution behaviour is a feature of straight profile.

Moreover, from Figure 9.18, we can clearly observe a couple of complex and conjugate eigenvalue crossing the imaginary axis. This is, in fact, a paradigm for Hopf bifurcation [128, 142] and represents another evidence of how deeply the system changed its inner features.

Remark 9.2.1 (Lagrange multipliers). From a numerical point of view, we employed Lagrange multipliers to solve the optimality system (9.42). The condition $v = u$ on Γ_D has been weakly imposed in integral form

$$\int_{\Gamma_D} v\lambda \, ds = \int_{\Gamma_D} u\lambda \, ds \quad \forall \lambda \in \mathbb{V}. \tag{9.43}$$

This equation reflects in system (9.27), since the term

$$\int_{\Gamma_D} \varphi(\nabla \cdot vn + pn) \, ds \quad \forall \varphi \in \mathbb{V}, \tag{9.44}$$

appears. The two equations will result in extra terms in the adjoint equations. Furthermore, we weakly impose also the boundary condition $w = 0$ on Γ_D with another multiplier.

9.2.5 A comparative eigenvalue analysis

In this section, we sum up all the observations and results derived from the global eigenvalue analysis over the four test cases. Therefore, we now list the similarities between these, especially for what concerns the variation against different values of the penalization parameter α:

- the *eigenvalues cluster* around the value of α. This behaviour is well represented in Figures 9.5c, 9.5d, 9.14c and 9.14d. These eigenvalues come from the optimality equation;
- the *predominance of positive eigenvalues* over the negative ones. In all the performed spectral analysis we have observed that the control action lowers the negative eigenvalues. The stronger is the control, the greater is the number of positive eigenvalues, as it is represented in Figures 9.9b and 9.17;
- the *shears effect* for low controlled system. The shears configuration is characteristic of control problems which do not highly change the uncontrolled system solution. It is the case of Neumann control in Figure 9.5a and of the channel control for $\alpha = 1$ as shown in Figure 9.14a. For the other cases, the smaller is α the least visible are these shearing eigenvalues: in some cases, the structure is completely broken;
- the μ^* *identification*. For the shears, it is clear that they approach to the real line in the point for which the bifurcating phenomenon occurs. This is the same situation we found for the uncontrolled problem, in which the path of the eigenvalue identifies the value of bifurcation parameter. Moreover, regardless the power of the control imposed this situation is preserved. Indeed, the positive shears eigenvalue is still present in Figures 9.9b, 9.14b, 9.14c and even in the Dirichlet optimal control, as shown in Figure 9.17. In some cases, a shift of the μ^* has been observed.

Since the structure of the spectral analysis is highly influenced by the control strength, we tried to perform an eigenvalue analysis dealing with only state and adjoint equations. For all the test cases, shears occurs. The shears structure is symmetric when the solution shows the wall-hugging property, while it is slightly asymmetric when the state flow is straight. We guess that this behaviour is due to the different reaction of state and adjoint blocks to the bifurcating phenomena. Indeed, for the symmetric flux, the behaviour of the state equation has to be preserved for all μ, while the adjoint problem, which is strictly linked to the control variable, puts more effort in rebalancing the flux, resulting in an asymmetric contribution that causes the shears to be slightly asymmetric.

The spectral analysis of a nonlinear system is a indispensable tool to understand bifurcation phenomena, eventually. Under the point of view of computational costs, it is a very tough task, most of all for nonlinear OCP. Indeed, as always FE discretization leads to huge systems to be solved for a wide sample of parameter $\mu \in \mathcal{P}$. Thus, in

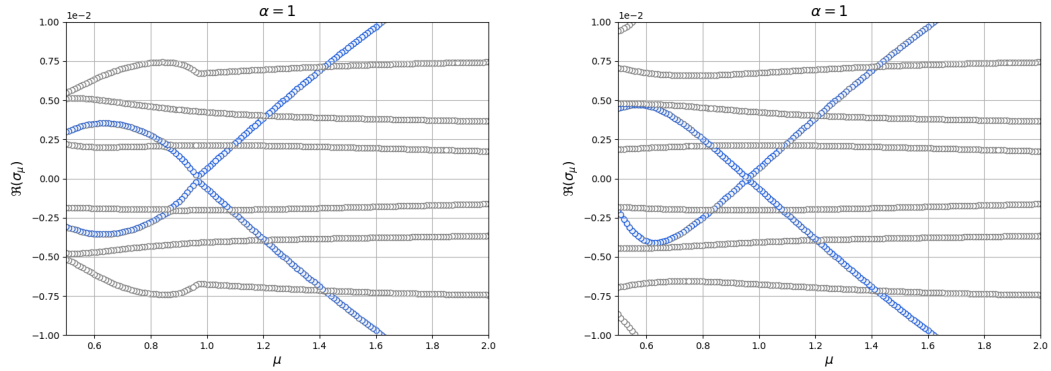


Figure 9.19: *Left*: asymmetric velocity with Neumann control for $\alpha = 1$. *Right*: symmetric velocity with distributed control for $\alpha = 1$.

the next section, we investigate the ROM approach as a suitable way to overcome this issue.

9.2.6 A ROM approach

In this section, we are presenting the numerical results deriving from the reduction of the four controlled test cases described previously. For each numerical test case, the offline setting is given by $N_{train} = 51$ snapshots evaluated for equidistant parameters in the range of $\mathcal{P} = [0.5, 2]$ and the POD algorithm is chosen for the ROM construction. Let us define the *basis number* $\overline{N}$ as the maximum dimension of the reduced spaces, which is the same for each variable. For Dirichlet test case we chose $\overline{N} = 12$ basis functions, while for the other test cases the basis number is $\overline{N} = 20$. The choice was taken due to the presence of the two multipliers variables in the Dirichlet test case, which increases the global dimension of the system from $13\overline{N}$ to $15\overline{N}$.

Then, we perform an online phase solving (9.16) for $K = 151$ equidistant value of μ in the same parameter space $\mathcal{P}$. The performance has been tested through separate error analysis for each variable. The effectiveness of the ROM approach have been evaluated through

- an average error over the parameter space against an increasing value of the reduced spaces dimension N up to $\overline{N}$;
- a μ-dependent error computed for the value $\overline{N}$.

The two error analysis specify different features of the system that we are going to highlight in the following. Indeed, the average error gives us information about how the performance changes as the behaviour of the solution changes. The straight profile appears to be always the best approximated, due to its Stokes-like (symmetric) nature for all the value $\mu \in \mathcal{P}$. This is the case of Neumann and Channel control, which

average error is depicted in Figures 9.20a and 9.22a. Their asymmetric counterparts, Figures 9.20b and 9.22b, show how representing the two different features of the solution, a Stokes-like one for $\mu > \mu^*$ and a wall-hugging for the lower values of μ, using the same value of $\overline{N}$ is more difficult than the symmetric one. Nonetheless, the provided accuracy for basis size $\overline{N}$ is satisfactory for many practical applications in both target cases.

This argument applies to the control and adjoint variables, yet the state ones are the best described by ROM for all the test cases. Because of the optimality equation, the adjoint variables feel the direct influence of the control, which is the most challenging one to be approximated by the reduced model due to its high variability in μ. Indeed, the control variable presents a sort of *on-off* behaviour which drastically affects the efficiency of reduced representation. For example, if we deal with Stokes target v_{d}, the control is *off* for high values of viscosity, but, when $\mu \approx \mu^*$, it starts to grow in magnitude and to change drastically its features. This is represented in Figures 9.20d and 9.23d, where the higher values for the control error and, thus, for the adjoint variable, is for higher values of μ.

Since the control magnitude is essentially zero, for Channel and Dirichlet test cases with low Reynold number (high viscosity), we chose to plot the absolute errors instead of the relative ones, in order to prevent division by zero. This is not the case of Distributed control, see for example Figure 9.21b, which presents good errors decay for all the variables, since its strong action causes the control magnitude to be always meaningful.

The most challenging case to be approximated is the Dirichlet for $\alpha = 0.001$. It is the most complex dynamic, where a new bifurcation appears. Indeed, it can reach an error of almost 10^{-3} for the controlled state. Even though this result is worse than the ones obtained for the other test cases, where average error are ranging between 10^{-5} and 10^{-8}, the accuracy provided by the ROM is still acceptable for many practical purposes.

We remark that this performance is strictly correlated to the more complex features of the Dirichlet problem. This is evident also from the μ-dependent errors in Figure 9.23c and 9.23d. In both the pictures, we see a great increment of the error for high Re. Although the phenomenon appears also for the other test cases, see Figures 9.21d, 9.22c and 9.22d, it is not as strong as the Dirichlet case.

Furthermore, the μ-dependent error gives an a posteriori information about the bifurcation point. Indeed, we remark that in order to have good accuracy property, ROM approach requires regularity on the parametric dependence of the solution instead of the spatial one. This means that reduced errors will generally exhibits a peak at μ^*.

In fact, an increasing value of the error can be seen around $\mu^* \sim 0.96$ for Neumann, Distributed and Channel controls. This feature can be very useful when there is no previous knowledge about bifurcating behaviours. In this sense, ROM is not only an useful approach to fast solve very complicated time consuming systems, but also to detect parameters which can be related to the bifurcating nature of the problem at hand, since their instances will be the worst approximated.

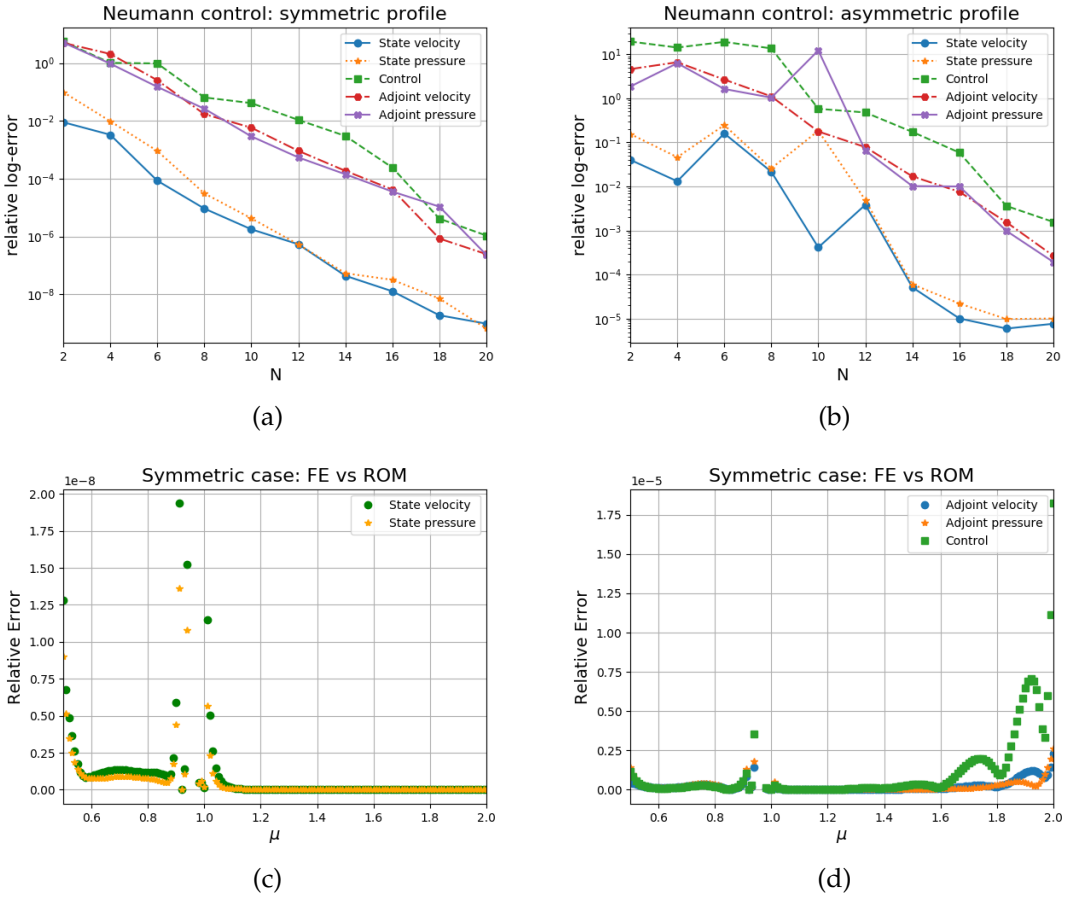


Figure 9.20: Error Analysis of Neumann control for $\overline{N} = 20$ and $\alpha = 0.01$. Average error over μ for symmetric and asymmetric profile in (a) and (b), respectively. The μ-dependent error for symmetric profile for state variable in (c) and adjoint and control variables in (d).

We conclude this analysis by noticing that the same considerations hold for the Dirichlet control, but this time at the left end of the parametric domain $\mathcal{P}$ where such evidence is clearly influenced by the new configuration observed in Figure 9.15b.

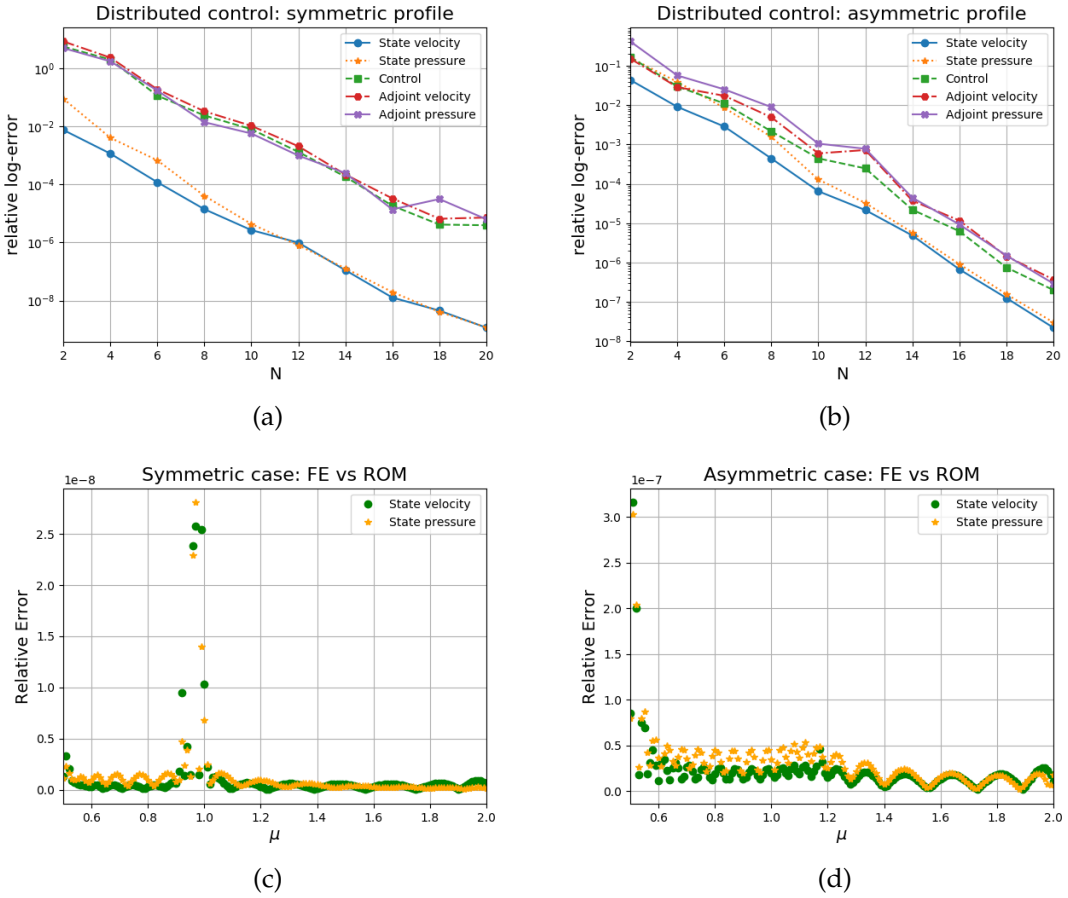


Figure 9.21: Error Analysis of Distributed control for $\overline{N} = 20$ and $\alpha = 0.01$. Average error over μ for symmetric and asymmetric profile in (a) and (b), respectively. The μ-dependent error for symmetric and asymmetric profile of the state variable in (c) and (d).

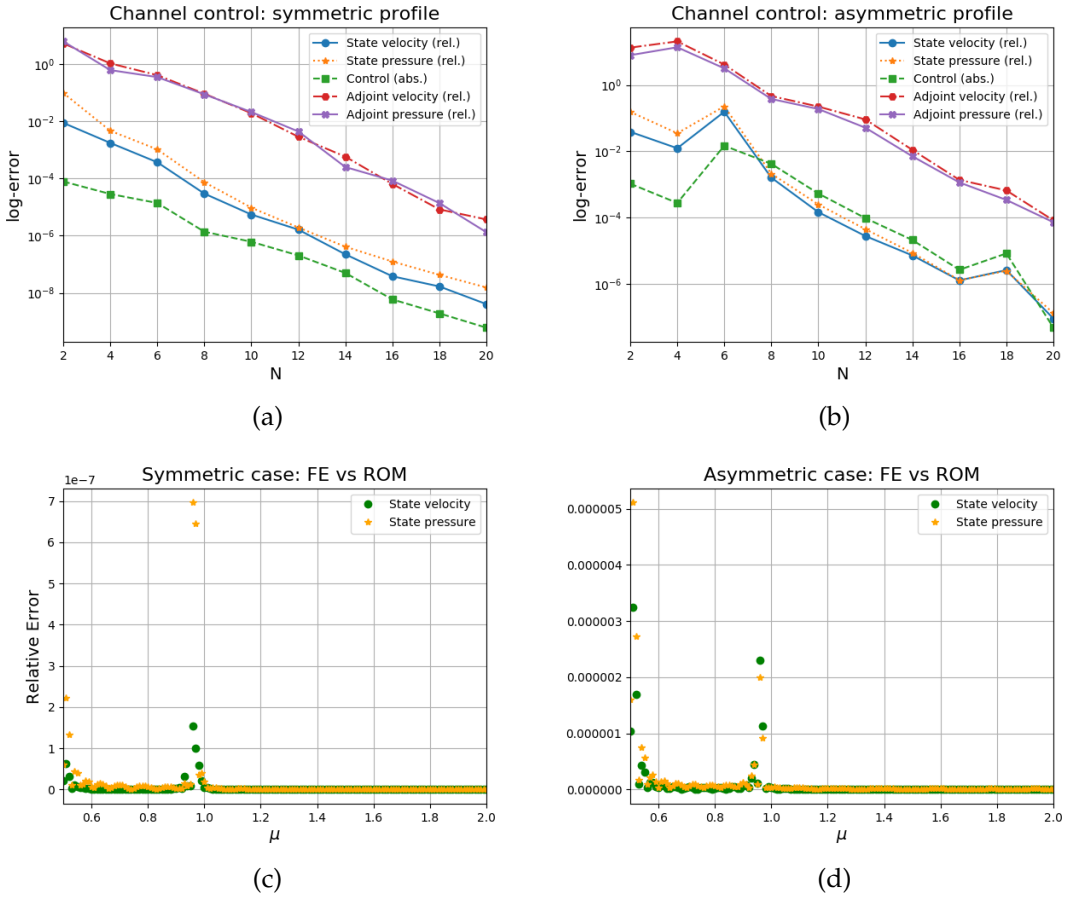


Figure 9.22: Error Analysis of Channel control for $\overline{N} = 20$. Average error over μ for symmetric ($\alpha = 1$) and asymmetric ($\alpha = 0.01$) profile in (a) and (b), respectively. The μ-dependent error for symmetric and asymmetric profile of the state variable in (c) and (d).

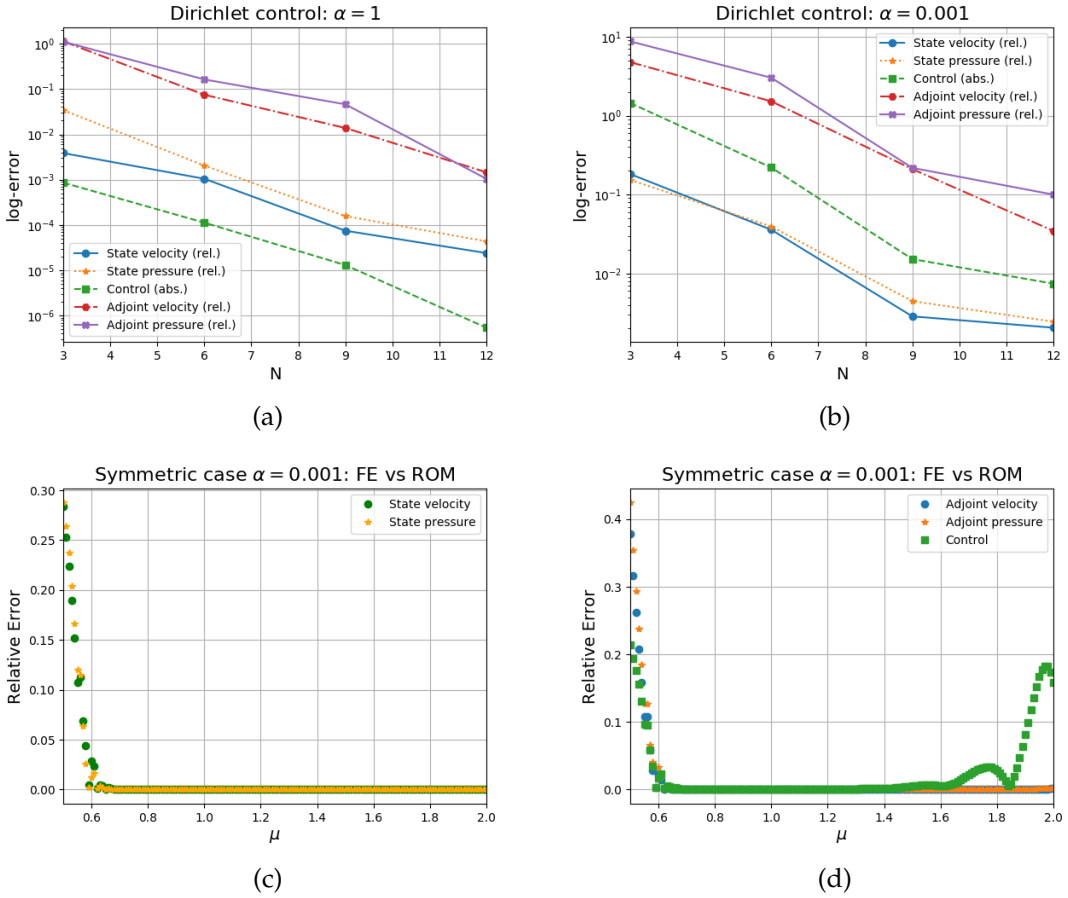


Figure 9.23: Error Analysis of Dirichlet control for $\overline{N} = 12$. Average error over μ for $\alpha = 1$ and $\alpha = 0.001$ in (a) and (b), respectively. The μ-dependent error for $\alpha = 0.001$ for state variable in (c) and adjoint and control variables in (d).

Chapter 10

An artificial neural network investigation of bifurcations

In this last chapter, we want to address the issue of the efficiency, that we have encountered many times throughout this thesis, by means of a *non-intrusive* technique. Dealing with the reduction of nonlinear parametrized PDEs led us to the implementation of the (intrusive) hyper-reduction strategies, in the sense that a further level of approximation have to be introduced, with derived quantities entering in the online phase, to recover the efficiency.

In order to obtain the full decoupling between offline and online phase, we applied here a regression-based approach in the context of *Machine Learning* (ML). In particular, among all the different regression methods in the supervised learning context, we chose the POD-NN approach presented in [76], where the *Artificial Neural Networks* (ANNs) are coupled with the POD technique.

Here, we present the results concerning its application to bifurcating phenomena in CFD, showing that the network is able to learn the location of the critical points for the model, and the corresponding branching behaviour. Moreover, we used this as a starting point to develop a buckling detector tool which can represent an inexpensive way to obtain preliminary information about the system.

This part is entirely based on the work in preparation with J. S. Hesthaven [117].

10.1 The POD-NN approach

Nowadays, many authors are deeply investigating the benefits that a machine learning approach could bring to different topics in numerical analysis [134, 67, 116, 92, 94]. From one side, the higher computational resources and the increasing data available makes the application of these methodologies easier and faster, but it is clear that not relying on a solid mathematical analysis they could led to inaccurate or not reliable results. An example of this is given by the duality between e.g. projection-based and data-driven methods. While the former are built upon the mathematical formulation of the problems (the weak formulation in our nonlinear parametrized PDE context),

the latter exploit the relation between input data and teaching outputs (in the supervised learning paradigm) to learn the dynamics.

The scientific community is making a great effort focusing the attention on a better mathematical characterization of the approximation properties of these techniques [48, 143, 107]. Despite this, practicable results are still limited.

Moreover, as we have seen in the previous sections, the reduction of nonlinear parametrized PDEs involving bifurcating phenomena usually still suffer from the course of dimensionality, since hyper-reduction techniques as EIM/DEIM are in these cases too much costly and of difficult application.

For these reasons, we chose the POD-NN method presented in [76] for a non-intrusive investigation of the models, in particular in connection to bifurcating phenomena.

This approach is the combination between the POD technique based on the Galerkin FE as full order model and the feedforward neural networks, also called (multilayer) *perceptrons*. It enables a complete offline-online decoupling recovering the efficiency by means of a non-intrusive methodology, while relying on reliable high fidelity approximation.

In the following, we want to describe briefly how the neural network can be used for the reduced approximation of a generic nonlinear parameter dependent PDE. For a much detailed presentation we recommend [76, 137, 58]. Thus, let us consider the nonlinear function $f : \mathbb{R}^{M_I} \to \mathbb{R}^{M_O}$, where M_I and M_O are the input and output dimensions. We aim at approximating such function starting from a training set given by the pairs $(\boldsymbol{\mu}_i, f(\boldsymbol{\mu}_i))_{1 \leq i \leq N_{train}}$, respectively the input pattern in $\mathbb{R}^{M_I}$ and teaching inputs in $\mathbb{R}^{M_O}$.

Therefore, we consider a feedforward neural networks with an input and output layer, respectively of M_I and M_O neurons, and $L_K - 2$ inner layers each one with H_K computing neurons, this way we will obtain a network with L_K layers and $L_K H_K$ neurons.

Coming back to the notation in Section 3.2.1, we remark that we have denoted with $\{E^i\}_{i=1}^{\mathcal{N}}$ a basis for $\mathbb{X}_{\mathcal{N}}$ and with $\{\Sigma^i\}_{i=1}^{N}$ the reduced basis for $\mathbb{X}_N$. Finally, we recall that the $||\cdot||_{\mathbb{X}_{\mathcal{N}}}$ - closest element of $\mathbb{X}_N$ to the high fidelity solution $X_{\mathcal{N}}$ can be expressed as

$$X_{\mathcal{N}}^{\mathsf{V}}(\boldsymbol{\mu}) = \sum_{j=1}^{\mathcal{N}} (\mathsf{V}\mathsf{V}^T X_{\mathcal{N}}(\boldsymbol{\mu}))_j E^j = \sum_{i=1}^{N} (\mathsf{V}^T X_{\mathcal{N}}(\boldsymbol{\mu}))_i \Sigma^i.$$

In this context, the aim of the neural network is to approximate the function $\pi : \mathcal{P} \subset \mathbb{R}^P \to \mathbb{R}^N$, which maps each input parameter $\boldsymbol{\mu} \in \mathcal{P}$ to the coefficient $\mathsf{V}^T X_{\mathcal{N}}(\boldsymbol{\mu})$ for the expansion of $X_{\mathcal{N}}^{\mathsf{V}}(\boldsymbol{\mu})$ in the reduced basis $\{\Sigma^i\}_{i=1}^{N}$.

As we said, the learning step is performed through a supervised learning approach, based on the training set given by the pairs $\{(\boldsymbol{\mu}^{(i)}, \mathsf{V}^T X_{\mathcal{N}}(\boldsymbol{\mu}^{(i)}))\}_{1 \leq i \leq N_{train}}$. Actually, in order to be consistent, instead of computing the training output as $\mathsf{V}^T X_{\mathcal{N}}(\boldsymbol{\mu}^{(i)})$ we project the snapshots through the normal equations, solving the linear system given by $\mathsf{V}^T \mathsf{M} \mathsf{V} X_N(\boldsymbol{\mu}) = \mathsf{V}^T \mathsf{M} X_{\mathcal{N}}(\boldsymbol{\mu})$, where M is the inner product matrix of the discretization.

The main advantage of this training process is that it does not affect the offline phase for the reduction strategies. Indeed, one only have to rely on the computed snapshots to perform the POD compression step.

Moreover, once computed the dataset given by the parameters and the corresponding target outputs, for cross-validation purposes we split it as follows: a training set $\Xi_{tr} = \{\boldsymbol{\mu}^{(1)}, \dots, \boldsymbol{\mu}^{(N_{tr})}\}$ with $N_{tr} = 4N_{train}/6$, a validation set $\Xi_{va} = \{\boldsymbol{\mu}^{(1)}, \dots, \boldsymbol{\mu}^{(N_{va})}\}$ with $N_{va} = N_{train}/6$ and a testing set $\Xi_{te} = \{\boldsymbol{\mu}^{(1)}, \dots, \boldsymbol{\mu}^{(N_{te})}\}$ with $N_{te} = N_{train}/6$.

Here, we remark the fact that, while testing the ability of the network to learn with different size of N_{train}, the dimension of the testing set Ξ_{te} increases proportionally. Thus we cannot expect a monotone decrease of the error improving the training set, also in a non-overfitting regime.

Despite the huge variety of neural network structures that can be implemented, we chose here a very standard setting, i.e. a *learning rate* η, the weighted sum as a propagation function, the hyperbolic tangent $\tanh(x)$ or the rectified linear unit ReLU $\max(0, x)$ as activation functions and finally the identity as output function.

As concerns the training procedure, we took into account both mini-batches and learning through epochs paradigms. Within this setting we adjust the weights of the network by means of the ADAM optimizer [90], guided by the mean squared error (MSE) function as an indicator of the performance. The ADAM method is a first order technique that consider only the gradient of the cost function, while for example the optimizer used in [76] is the Levenberg-Marquardt algorithm, which is of the second order and thus requires the computation of the hessian. This makes out approach faster but less accurate.

In particular, starting from the consideration that a network has to be trained with a sufficiently large enough number of sample to obtain good accuracy properties, which is in contrast with the ROM philosophy, we implemented some basic tools to speed up the learning process and to avoid the overfitting issue [64]. Thus, we will both consider an increase of the number of training points and specific procedure for epochs. We initialize the weights with a uniform distribution (without using a multiple restarts approach), while for the learning rate decay we tested both following hyperbolic relations

$$\eta = \frac{\eta_0}{1 + \theta \cdot \text{epoch}} \qquad \text{ans} \qquad \eta = \frac{\eta_0}{1 + \sqrt{\text{epoch}}}, \tag{10.1}$$

weighted respectively by the hyper-parameter $\theta \in [0, 1]$ and the square root function.

In this way we could reach a better accuracy, since the learning parameter gives a weighted importance to the informations at any instance of the training loop on epochs. As we said, the latter is controlled by an early-stop procedure, which checks whether or not the training validation accuracy has decreased within the last N_{it}^{ep} iterations, or the loop has reached a maximum of instances $N_{\max}^{ep}$. A smarter way to break the learning phase could be implemented, for example allowing only to the most significant improvements to keep the loop running, but we remark that here we want to focus on bifurcating applications rather then ad-hoc settings for the network. Moreover, we remark that a critical task dealing with neural networks is the setting

of the so called *hyper-parameters*, some of them we have already introduced (number of layers, number of neurons, learning rate, activation function). Nonetheless, a more robust investigation of the hyper-parameter space can be done through automated machine learning (AutoML) and Bayesian optimization [79].

At the end of the procedure, we consider the network identified by the weights configuration associated to the best validation accuracy and we use it for the non-intrusive step.

Moreover, we introduced a normalization on the dataset in order rescale it in the range $[0, 1]$. This was a necessary step to let the network learn from the different order of magnitudes of the training dataset and to improve the results.

Furthermore, we want to remark that although many different sampling procedures can be adopted, we chose an equispaced or log-equispaced distribution, since they better guarantee that after the splitting of the dataset, each portion of the parameter domain is well represented for training, testing and validating phases also when N_{train} is small. Another observation is that we used an initialization seed in order to be sure of the reproducibility of these results, indeed a different initial random splitting, weights' setting and order of the batches would bring to different results. As we said before, we fixed a priori this the random seed without considering a multiple restarts approach.

Of course, all the ad-hoc settings for the network that one can implement, increase its dimensionality and this way the computational time for the learning procedure could become too costly.

As concerns the errors, for the analysis of the POD-NN methodology we will consider for a given parameter $\boldsymbol{\mu} \in \mathcal{P}$ the quantities

$$\varepsilon_{RB}(\boldsymbol{\mu}) = \frac{\|X_{\mathcal{N}}(\boldsymbol{\mu}) - X_N(\boldsymbol{\mu})\|}{\|X_{\mathcal{N}}(\boldsymbol{\mu})\|} \qquad \text{and} \qquad \varepsilon_{NN}(\boldsymbol{\mu}) = \frac{\|X_{\mathcal{N}}(\boldsymbol{\mu}) - X_{NN}(\boldsymbol{\mu})\|}{\|X_{\mathcal{N}}(\boldsymbol{\mu})\|}$$

respectively the POD-Galerkin error and the neural network one, being $X_{NN}(\boldsymbol{\mu}) = \mathsf{V}\pi(\boldsymbol{\mu})$.

We evaluated these errors on the test parameter set $\Xi_{te} \subset \mathcal{P}$ of dimension N_{te}, and as statistics of the performance we consider the average of these quantities, which we denoted by

$$\bar{\varepsilon}_{RB} = \frac{\sum_{\boldsymbol{\mu}\in\Xi_{te}} \varepsilon_{RB}(\boldsymbol{\mu})}{N_{te}} \qquad \text{and} \qquad \bar{\varepsilon}_{NN} = \frac{\sum_{\boldsymbol{\mu}\in\Xi_{te}} \varepsilon_{NN}(\boldsymbol{\mu})}{N_{te}}$$

and their maximum values

$$\varepsilon_{RB}^{\max} = \max_{\boldsymbol{\mu}\in\Xi_{te}} \varepsilon_{RB}(\boldsymbol{\mu}) \qquad \text{and} \qquad \varepsilon_{NN}^{\max} = \max_{\boldsymbol{\mu}\in\Xi_{te}} \varepsilon_{NN}(\boldsymbol{\mu}),$$

over the test data set Ξ_{te}.

Finally, we remark that we we will use the same notation also when the unknown X will contain multiple variables, thus specifying the corresponding field.

10.2 Numerical approximation

Here, we want to analyse the approximation properties of the POD-NN strategy in connection to bifurcating models held by Navier-Stokes equations. We will start with a benchmark multi-parameter problem based on a simple thermal block application based on the Poisson equation. Then, we will consider how the network is able to learn the bifurcating phenomena in a varying geometry setting in the following cases: the channel flow model investigated in Chapter 8 and a cavity flow inside a triangular geometry.

10.2.1 Thermal block toy problem

As a first benchmark, we now consider a steady heat conduction problem [74] in a two-dimensional domain. In particular, we define $\Omega = \Omega_1 \cup \Omega_2 = [-1,1] \times [-1,1]$ divided in two sub-domains Ω_1 and Ω_2 such that the former is a disk centred at the origin with radius $r_0 = 0.5$ and the latter is defined as $\Omega_2 := \Omega / \overline{\Omega}_1$. The conductivity κ of this phenomena is given by

$$\kappa(\mu_1) = \begin{cases} \mu_1 & \text{in } \Omega_1 \\ 1 & \text{in } \Omega_2 \end{cases}$$

where we imposed through the parameter μ_1 the diffusion on the sub-domain Ω_1. Moreover, we took into account a second parameter which acts as a constant heat flux of magnitude μ_2 over $\Gamma_{\text{base}} = [-1,1] \times \{-1\}$.

Thus, we have a parameter $\boldsymbol{\mu} = (\mu_1, \mu_2)$ in the space $\mathcal{P} = [0.1, 10] \times [0.1, 1]$ and we model the heat transfer process due to the heat flux over the bottom boundary Γ_{base}. The remaining boundary conditions prescribed for this problem are homogeneous Neumann for $\Gamma_{\text{side}} = \{-1,1\} \times [-1,1]$ and homogeneous Dirichlet on $\Gamma_{\text{top}} = [-1,1] \times$

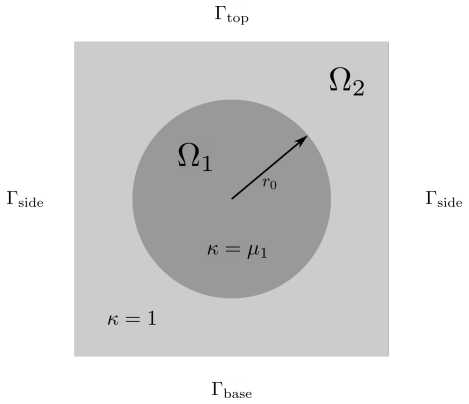


Figure 10.1: Domain Ω for the thermal block test case, with the two sub-domains Ω_1 and Ω_2.

$\{1\}$. From point of view of the physics, we are imposing insulated side edges while fixing a reference temperature on the top one.

Let $u(\boldsymbol{\mu})$ be the temperature unknown in the domain Ω, the strong formulation of the parametrized problem is given by:

$$\begin{cases} \operatorname{div}(\kappa(\mu_1)\nabla u(\boldsymbol{\mu})) = 0 & \text{in } \Omega \\ u(\boldsymbol{\mu}) = 0 & \text{in } \Gamma_{\text{top}} \\ \kappa(\mu_1)\nabla u(\boldsymbol{\mu}) \cdot n = 0 & \text{in } \Gamma_{\text{side}} \\ \kappa(\mu_1)\nabla u(\boldsymbol{\mu}) \cdot n = \mu_2 & \text{in } \Gamma_{\text{base}} \end{cases} \tag{10.2}$$

where n is the outer normal to the boundaries Γ_{side} and Γ_{base}. Finally, if we define the function space as $\mathbb{X} = \{v \in H^1(\Omega) : v|_{\Gamma_{\text{top}}} = 0\}$ the weak formulation of the problem reads: given $\boldsymbol{\mu} \in \mathcal{P}$, find $u(\boldsymbol{\mu}) \in \mathbb{X}$ such that

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

where we introduced the following linear and bilinear forms as

$$a(u, v; \boldsymbol{\mu}) = \int_\Omega \kappa(\mu_1)\nabla u \cdot \nabla v \, dx \,,$$
$$f(v; \boldsymbol{\mu}) = \mu_2 \int_{\Gamma_{\text{base}}} v \, ds \,.$$

Even if this is just a simple linear problem, which serves us to test the POD-NN strategy, it will already provide good results for the multi-parameter context.

Let us now review the network structure. The input dimension corresponds to the dimension of the parameter space $P = 2$ while the following choices for the hyper-parameters were made. In order to study the learning properties of the network w.r.t. the size of the dataset, we considered $N_{train} = n_b^2$ equispaced points in the parameter set $\mathcal{P}$, where $n_b = 6(i+1)$ with $i = 1, 2, 3$.

For the mini-batch technique we selected n_b pairs depending on the size N_{train} of the dataset. An example of a split dataset for $i = 2$ is given in Figure 10.2. We start from an initial learning parameter $\eta_0 = 5.\text{e-}2$ and we update it by means of the hyperbolic decay (10.1) with $\theta = 0.01$, considering $N_{it}^{ep} = 5.\text{e+}2$ and $N_{\max}^{ep} = 1.\text{e+}4$.

As concerns the reduction strategy, after processing the $\mathcal{N}_{train}$ snapshots we selected the first $N = 4$ basis functions from the POD technique (thus the output layer will have 4 neurons). The obtained dataset is then divided as explained before between training, validation and testing sets. This will serve us to learn/validate and test the network during the offline and online phase, respectively.

Figure 10.3 represents the decay of the loss function for the training set. Together with the validation one, this exhibits a clear sign that the network is learning from the training data Ξ_{tr} already for the first 100 epochs and it is able to generalize for Ξ_{va}, which means that the network can actually predict with good accuracy unseen values during the learning procedure. We can observe that the validation curve is under the training one, this implies a possible underfitting of the model. In such situation

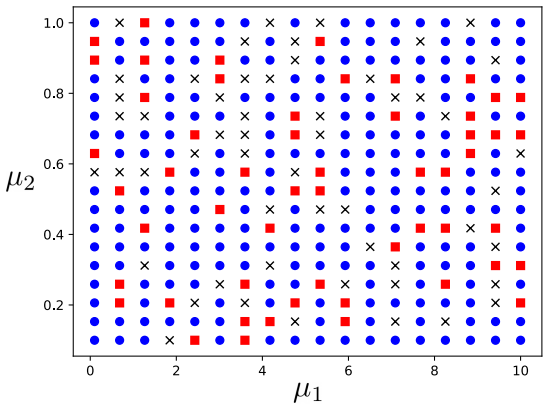


Figure 10.2: Dataset for ANN: (1) red squares are the training points in Ξ_{tr} (2) blue circles are the validation points in Ξ_{va} (3) black crosses are the testing points in Ξ_{te}.

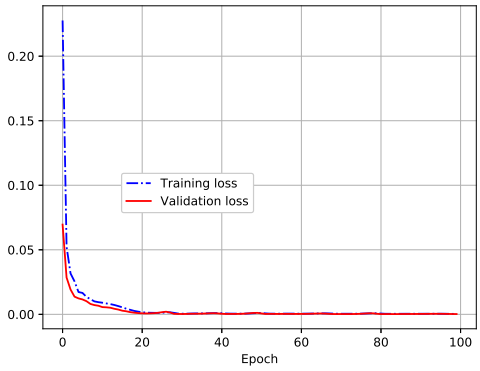


Figure 10.3: Training and validation loss function with respect to epoch.

one can either work with a more complex network adding layers or increasing its dimension adjusting H_K.

For this reason, we tested the network with different number of data inputs.

Now we report some experiments on the network properties, in order to find the best configuration for its weights. As we already said, the following hyper-parameters tuning is usually time consuming and automated procedure should be implemented in realistic cases.

We have changed the number $\mathcal{N}_{train}$ of points sampled from $\mathcal{P}$, so that after the splitting we have a training set Ξ_{tr} of dimension 96, 216 and 384, respectively for $i = 1, 2, 3$. We have also considered an increasing number of hidden layers, from one to three, together with a different dimension of the layers. Indeed, we let vary H_K from 5 to 50. Finally, we have also changed the activation function, comparing the hyperbolic tangent with the ReLU nonlinearity for different configurations of the

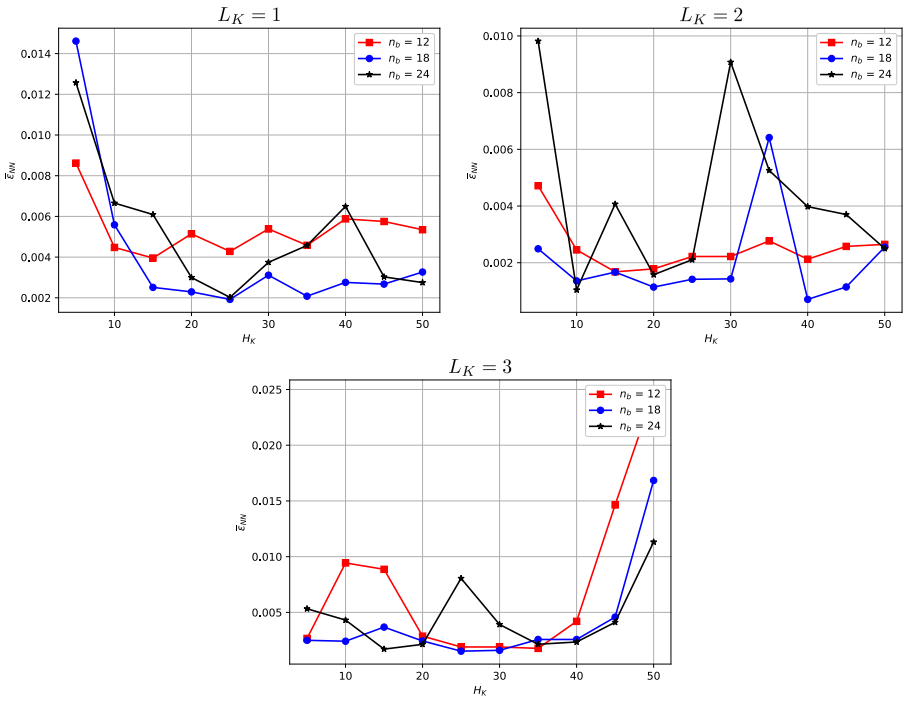


Figure 10.4: Mean NN errors $\bar{\varepsilon}_{NN}$ for different number of layers: $L_K = 1$ (top-left), $L_K = 2$ (top-right) and $L_K = 3$ (bottom).

network.

In Figure 10.4, for each plot which refers to the number of layers L_K, we show a comparison of the mean NN errors $\bar{\varepsilon}_{NN}$ with respect to the number of neurons H_K, for different sizes of the dataset (and corresponding mini-batches) $n_b = 12, 18, 24$, fixing the hyperbolic tangent as activation function.

For the sake of clarity, we show in Figure 10.5 the same comparison as before, but fixing in each plot the number of training points through n_b and letting vary the number of layers L_K. This way, we can clearly see the overfitting issue which appears when the dimension of the network is too big with respect to the features to be learned.

Moreover, in Figure 10.6 we show a comparison between the activation functions of the mean NN errors $\bar{\varepsilon}_{NN}$ with respect to the number H_K of neurons in each hidden layer $L_K = 1, 2$ for $n_b = 12, 18$. Since from these plots we can observe a general better behaviour of the tanh activation function, from now on we will only consider networks with this kind of nonlinearity. We could expect such result, in fact the potentiality of the ReLU activation function usually assume a greater importance when considering *deep neural networks*, i.e. one deals with very complex and huge structures.

In Figures 10.7, 10.8 we show the errors $\varepsilon_{RB}(\boldsymbol{\mu})$ and $\varepsilon_{NN}(\boldsymbol{\mu})$ computed for the parameters $\boldsymbol{\mu}$ belonging to Ξ_{te} with respect to the best case scenario for the POD-NN. In such case, we obtained $\varepsilon_{NN}^{\max} = 0.00473$ as the maximum error (near the origin where the solution changes more its properties) and $\bar{\varepsilon}_{NN} = 0.000706$ as the mean error over

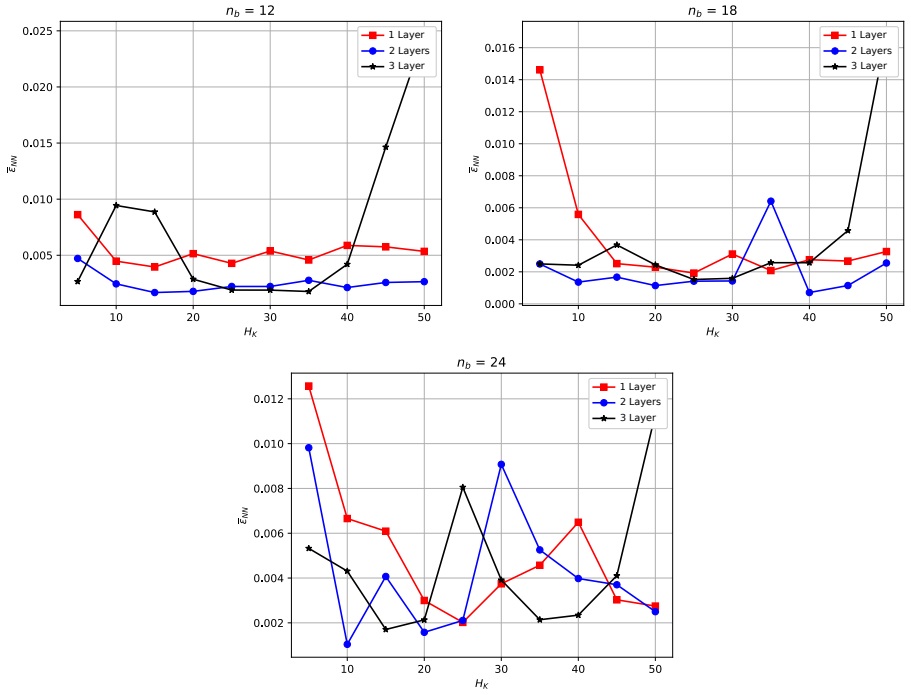


Figure 10.5: Mean NN errors $\bar{\varepsilon}_{NN}$ for different number of training points: $n_b = 12$ (top-left), $n_b = 18$ (top-right) and $n_b = 24$ (bottom).

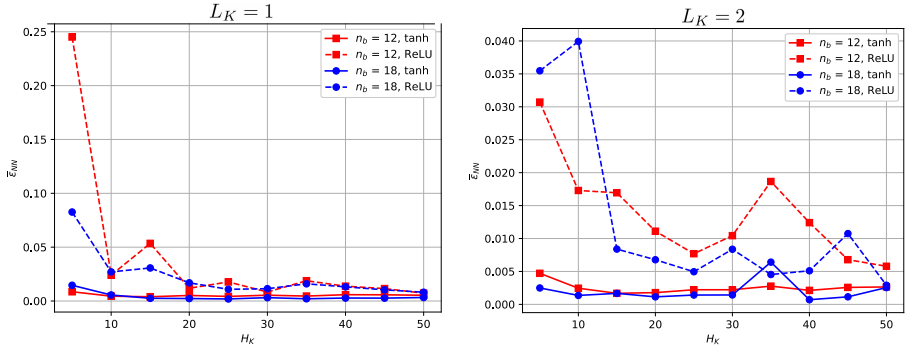


Figure 10.6: Mean NN errors $\bar{\varepsilon}_{NN}$ for $L_K = 1$ and $L_K = 2$ layers for different activation functions and number of points in Ξ_{tr}, left and right respectively.

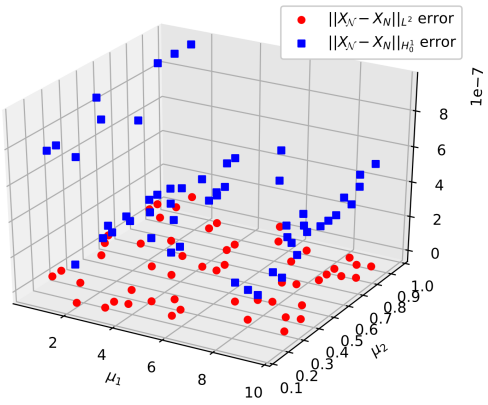


Figure 10.7: Error $\varepsilon_{RB}(\boldsymbol{\mu})$ computed on Ξ_{te}.

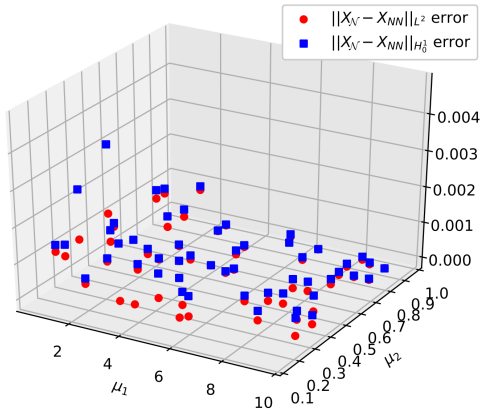


Figure 10.8: Error $\varepsilon_{NN}(\boldsymbol{\mu})$ computed on Ξ_{te}.

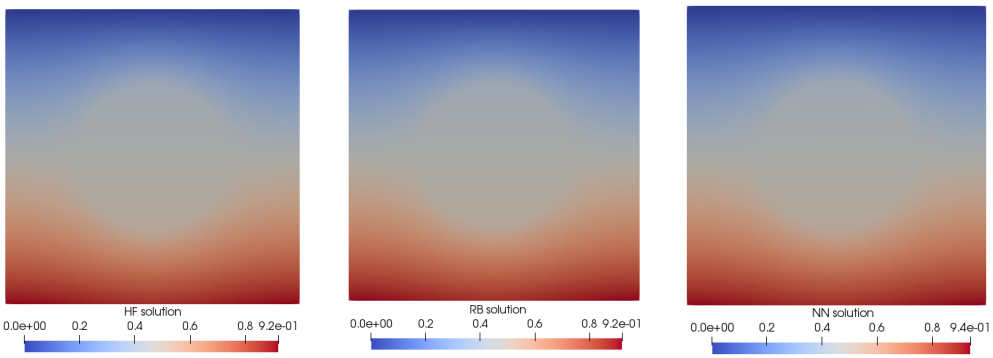


Figure 10.9: High fidelity, reduced basis and neural network solutions for the thermal block problem with $\boldsymbol{\mu} = (8.6, 0.6)$.

the testing dataset Ξ_{te}.

Moreover, we show in Figure 10.9 a representative solution obtained with the FE method, RB method and POD-NN technique. The corresponding errors with respect to the FE solution are shown in Figure 10.10 for both the RB and the POD-NN strategies.

Finally, the speed-up of the evaluation of the network on Ξ_{te} is almost 800 with respect to the FE approximation. Moreover, since the problem is linear and involves only two parameters, the corresponding RB speed-up remains quite low, almost 3. Thus, within the POD-NN setting we obtained a good accuracy while reaching a speed-up that is two order greater than the RB one.

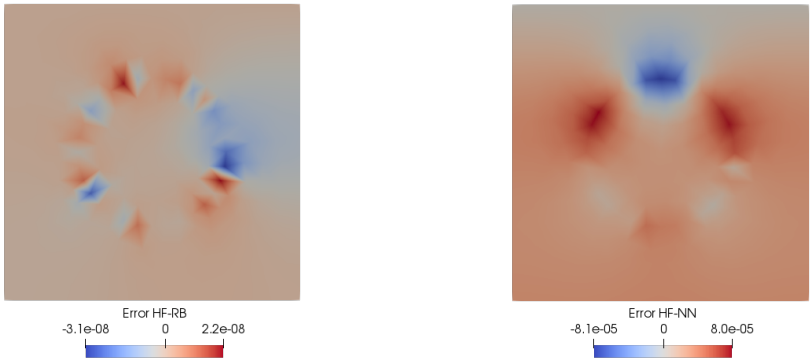


Figure 10.10: Relative errors HF-RB and HF-NN on the solution for the thermal block problem with $\boldsymbol{\mu} = (5.9, 0.2)$.

10.2.2 Navier-Stokes system in a channel

Let us now move to the more complex nonlinear Navier-Stokes equations. Here we consider again the viscous, steady and incompressible flow in the long planar straight channel with a narrow inlet, analysed in Chapter 8.

We recall that in this first preliminary study, we consider the Navier-Stokes equations

$$\begin{cases} u \cdot \nabla u - \mu \Delta u + \nabla p = 0 & \text{in } \Omega, \\ \nabla \cdot u = 0 & \text{in } \Omega, \end{cases} \tag{10.3}$$

where u is the velocity and p the pressure normalized over a constant density, in the non-bifurcating regime. This means that for the moment we let vary the viscosity μ in the range $\mathcal{P} = [2, 30]$, where no bifurcations occur. In fact, as studied before, the first critical value appears at $\mu^* \approx 0.96$. The same boundary conditions as in Section 8.1 have been supplemented to the system (10.3).

We remark that instead of considering a network that aims at approximating all the coefficients monolithically, i.e. trying to recover simultaneously the reduced coefficients for velocity and pressure with the same network, we constructed two different artificial neural networks with their own weights each field. This is crucial to obtain good approximation results, indeed even if the networks share the same inputs, the corresponding outputs can in principle have very different behaviours.

In practice, for this test case we have a one dimensional parameter space for the viscosity, namely $\mu \in \mathcal{P}$ corresponding to $P = 1$ and we chose a maximum number of basis functions for both fields u and p corresponding to $N_u = N_p = 10$. We remark that actually the dimension of the reduced basis space for the velocity is increased by one, due to the lifting coming from the inhomogeneous Dirichlet BC.

As concerns the learning procedure through epochs, we started from $\eta_0 = 3.\text{e-}1$ for both networks and we updated it by means of the hyperbolic decay weighted by the square root in (10.1), considering $N_{it}^{ep} = 5.\text{e+}2$ and $N_{\max}^{ep} = 1.\text{e+}4$.

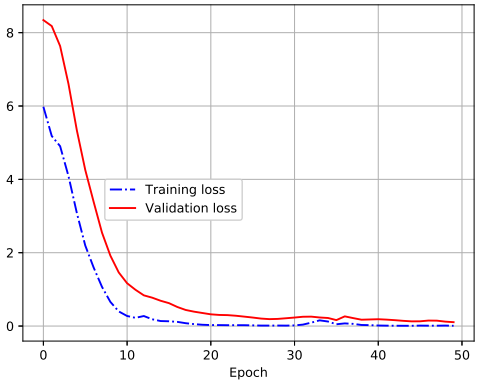


Figure 10.11: Training and validation loss function for the velocity field with respect to each epoch.

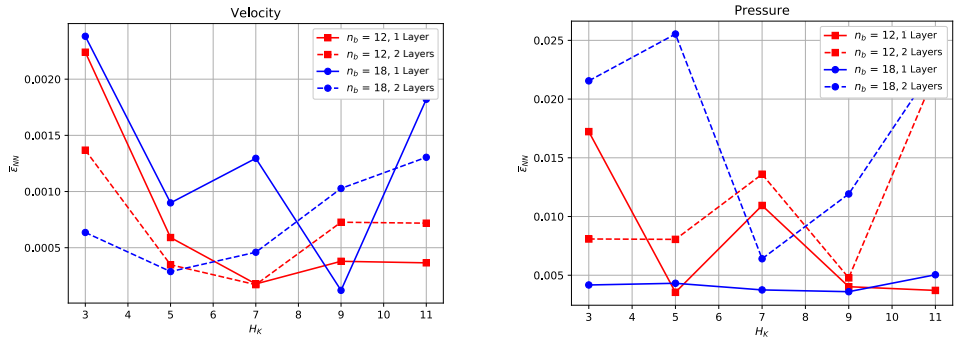


Figure 10.12: Mean NN errors $\bar{\varepsilon}_{NN}$ for velocity and pressure fields, left and right respectively.

Once again, the offline phase is initialized with $N_{train} = n_b^2$ (and corresponding size n_b for the mini-batches) equispaced points in the parameter set $\mathcal{P}$, using the same ratio as before to split the dataset in train, test and validation sets.

We present the decay of the velocity loss function for training and validating sets in Figure 10.11. We can see that the nonlinearity clearly complicates the setting, with an initial high value for both the loss functions. Despite this, we can see that already at the first 50 epochs the network learned the map from parameter to reduced coefficients and it is able to generalize for different values of the input.

Once again, we tested the network accuracy properties with respect to the dimension of the dataset (96, 216 and 384 training points in Ξ_{te}) while increasing the number of hidden layers from $L_K = 1$ to $L_K = 2$. Thus, in Figure 10.12 we show a comparison of the mean NN errors $\bar{\varepsilon}_{NN}$ for the velocity and pressure fields with respect to the number H_K of neurons in each hidden layer.

In Figures 10.13, 10.14 we show the error $\varepsilon_{NN}(\boldsymbol{\mu})$ computed for the parameters $\boldsymbol{\mu}$ belonging to Ξ_{te} for the velocity and the pressure. Also for this nonlinear problem, in

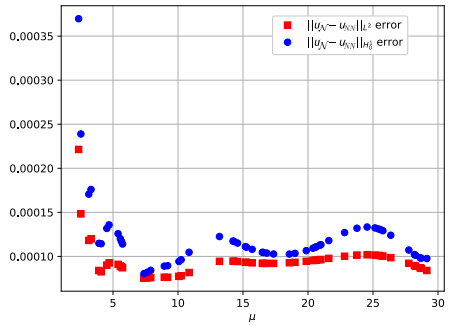


Figure 10.13: Error $\varepsilon_{NN}(\mu)$ for the velocity field u computed on Ξ_{te}.

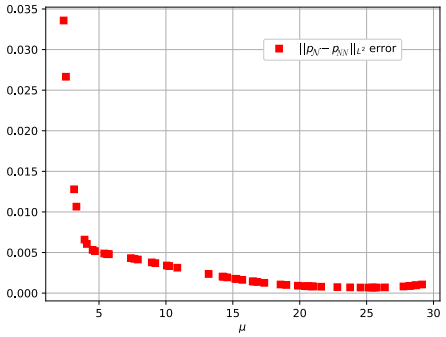


Figure 10.14: Error $\varepsilon_{NN}(\mu)$ for the pressure field p computed on Ξ_{te}.

the best case scenario, we were able to gain good approximation properties. In fact, we obtained for the velocity field $\varepsilon_{NN}^{\max} = 0.000369$ as the maximum error (near the left extrema of the parameter set $\mathcal{P}$, where we approach to the bifurcation point), and $\bar{\varepsilon}_{NN} = 0.00012$ as the mean error over the testing dataset. Similar considerations hold for the pressure field, where an increased error is observed mostly while approaching to lower values of the viscosity. We remark that such region is characterized by flow with increasing flux and of course it has to be properly represented inside the training set Ξ_{tr} after the splitting.

We show in Figures 10.15a, 10.15b and 10.15 a representative solution obtained with the FE method, RB method and POD-NN technique. Moreover, we show the corresponding errors with respect to the FE solution in Figures 10.16 and 10.17, respectively for velocity and pressure fields, using both the RB and the POD-NN strategies.

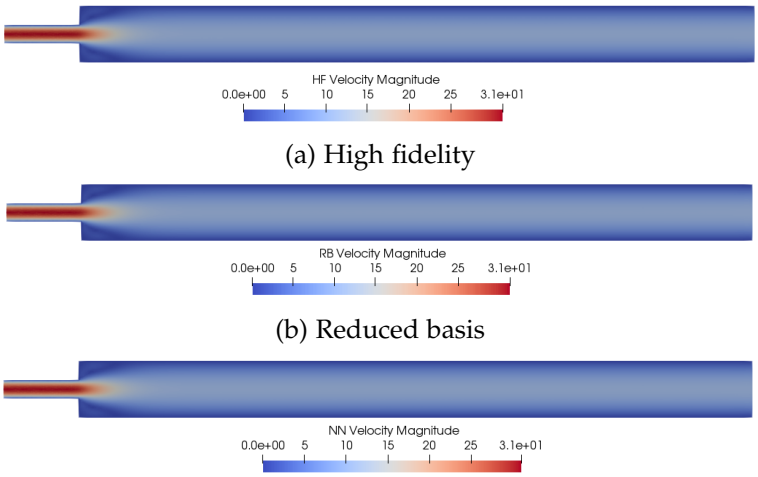
Finally, the speed-up for the evaluation of the network on Ξ_{te}, with respect to the FE method is almost 1.2e+6, while the RB one is equal to 1.25. This is consistent with the results obtained in the previous section, but we notice also that the gap between the two speed-ups increased. Indeed, the nonlinearity of the problem together with the lack of an efficient hyper-reduction/tensor assembly for the weak form, causes the standard RB approach to be characterized by a $\mathcal{N}$-dependent online phase.

In particular, the (actually unfair regarding what we just said) comparison between intrusive and non-intrusive strategy, show a difference of five orders of magnitude.

10.2.2.1 The bifurcating regime

In the previous section, we have discussed the POD-NN approximation of the nonlinear Navier-Stokes test case when critical points are not taken into consideration. Now, we can finally discuss the approximation of the bifurcating regime.

We remark that in this case, following the analysis done in Section 7 we are only interested in the approximation of the asymmetric (e.g. downwards wall-hugging behaviour) branch. Indeed, the reconstruction of the full bifurcation diagram as in Chapter 8 would require the network to assign different outputs to the same input pa-



(a) High fidelity

(b) Reduced basis

(c) POD-NN

Figure 10.15: Comparison between velocity field for the Navier-Stokes system with $\mu = 2.66$

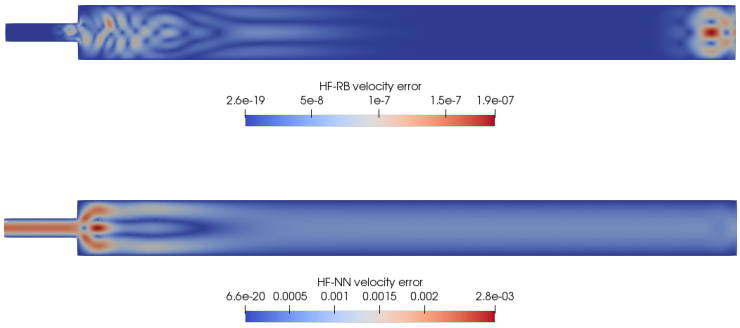


Figure 10.16: Relative errors HF-RB (top) and HF-NN (bottom) on the velocity field for the Navier-Stokes problem with $\mu = 4.09$.

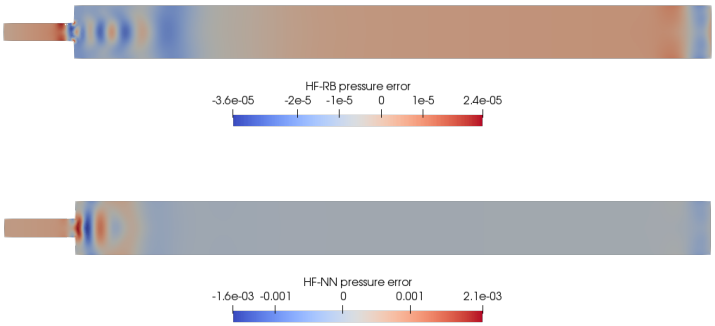


Figure 10.17: Relative errors HF-RB (top) and HF-NN (bottom) on the pressure field for the Navier-Stokes problem with $\mu = 4.09$.

rameter. To solve this issue, one can exploit a branch-wise strategy, building a specific network for each branch or consider more involved machine learning strategies. Therefore, it is clear that a key aspect for the reconstruction of bifurcating behaviour, within an artificial neural network context, is a complete identification of the singularities of the model. On this point we will come back in Section 10.3.

Here we consider the range for the viscosity as $\mathcal{P} = [0.5, 2]$ in such a way that it includes the critical value $\mu^* \approx 0.96$.

We remark that in order to be consistent with the analysis performed before, we still want to apply the simple continuation method to recover the solution at the high fidelity and reduced basis levels. Of course, this is not required by the network that simply evaluates the input parameters returning the value of the reduced coefficients. Therefore, in order to ensure the convergence of the FE and RB solutions, we applied a simple continuation method with $K = 151$ equispaced points in $\mathcal{P}$. Finally, as concerns the reduction strategies, we chose $N_u = N_p = 10$ basis functions for the reduced basis approximation.

Moreover, we highlight that we are fixing a priori the testing set Ξ_{te} through the continuation method, thus the usual splitting of the N_{train} snapshots is adapted incorporating the its previously assigned points to the training set Ξ_{tr}

Since we are now dealing with a more complex phenomenon, we expect the training process to be more difficult.

As concerns the accuracy, we tested again the properties of the network changing its hyper-parameters. Thus, in Figure 10.18, we show a comparison of the mean NN errors $\bar{\varepsilon}_{NN}$ for the velocity and pressure fields for the bifurcation test case, with respect to the number H_K of neurons in each hidden layer, $L_K = 1, 2$, for $n_b = 12, 18$.

We show in Figure 10.19 the error $\varepsilon_{NN}(\mu)$ for both velocity and pressure fields in the best case scenario which corresponds to the case with $L_K = 2$ layers, $n_b = 18$ and $H_K = 7$. In this case the maximum and mean errors over the testing dataset are respectively $\varepsilon_{NN}^{\max} = 0.02104$ and $\bar{\varepsilon}_{NN} = 0.00506$. Similar result holds for the pressure

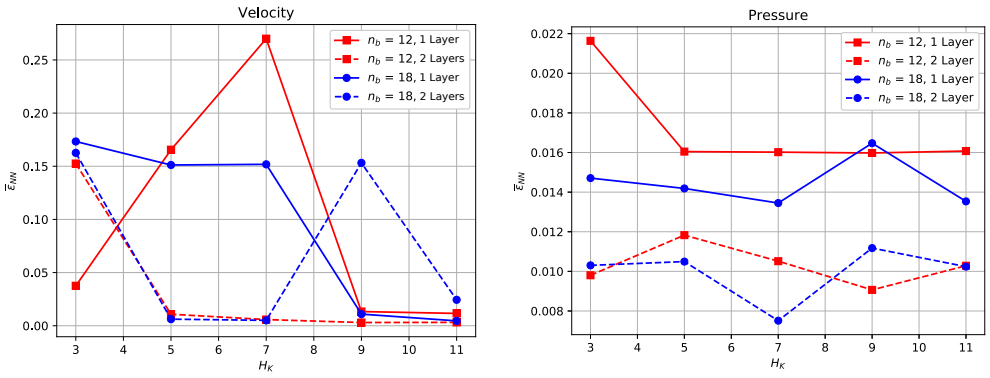


Figure 10.18: Mean NN errors $\bar{\varepsilon}_{NN}$ for velocity and pressure fields in the bifurcating regime, left and right respectively.

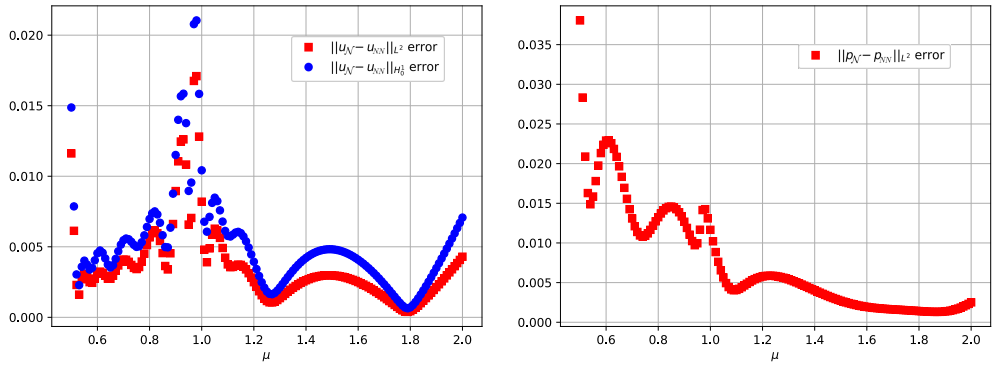


Figure 10.19: Error $\varepsilon_{NN}(\boldsymbol{\mu})$ for velocity and pressure fields in the bifurcating regime computed on Ξ_{te}, left and right respectively.

field, where we have $\varepsilon_{NN}^{\max} = 0.03805$ and $\bar{\varepsilon}_{NN} = 0.00751$.

As a comparison, we can show in Figure 10.20 the reduced basis error, on the same testing set Ξ_{te}, when employing the same number of basis functions. Employing the RB method (without hyper-reduction) we obtained the maximum and mean errors over Ξ_{te} as $\varepsilon_{RB}^{\max} = 0.02579$ and $\bar{\varepsilon}_{RB} = 0.00115$. As concerns the pressure field, we have $\varepsilon_{RB}^{\max} = 0.01726$ and $\bar{\varepsilon}_{RB} = 0.00086$.

As we can see from the errors, in the proposed setting, the POD-NN approach was able to properly recover the bifurcating behaviour of the model. Moreover, it reached an impressive order of accuracy also in comparison with the RB strategy. In fact, as we understood from the whole analysis of bifurcating problems, when the solution does not depend smoothly from the parameter, we can not expect the classic reduced basis error behaviour.

In Figure 10.21 we show a representative solution of the bifurcated state for $\mu = 0.5$, in which the velocity flow attaches to the bottom of the channel. In Figure 10.22

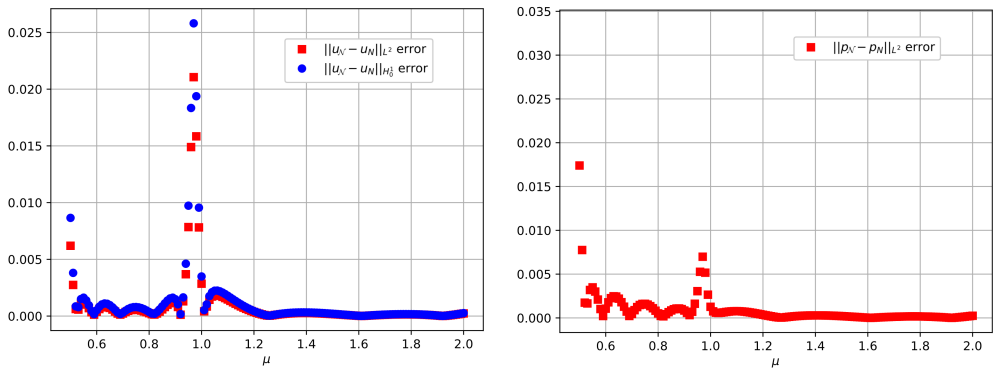


Figure 10.20: Error $\varepsilon_{RB}(\boldsymbol{\mu})$ for velocity and pressure fields in the bifurcating regime computed on Ξ_{te}, left and right respectively.

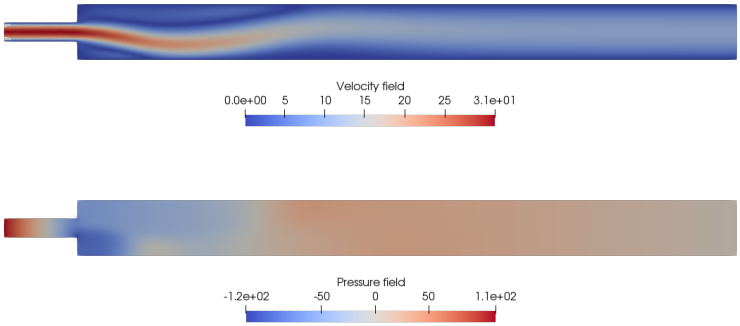


Figure 10.21: High fidelity bifurcated solution for the Navier-Stokes system for $\boldsymbol{\mu} = 0.5$, velocity and pressure fields.

we show the bifurcation diagram obtained through the POD-NN approach for the Navier-Stokes test case, where for each value of the viscosity $\mu \in \Xi_{te}$ we consider the horizontal velocity in the point $s(u) = u_y(20, 2.5)$ as output.

Of course, the results for the reduced approximation of the bifurcation test case are always worse with respect to the ones where the critical point is not in the parameter range. In general it could be a limitation, but we remark that, on the contrary, since the network is only built from the snapshots, it does not see the singularities of the model and it is only slightly affected by it. This results in the similar order of accuracy between mean and max errors for the POD-ANN, while as we have seen the same quantities for the RB are of different orders of magnitude.

Finally, we conclude noticing that the computational speed-up provides qualitatively similar results as the one in the non-bifurcating regime, since the convergence cost is dominated by the intrusive strategy.

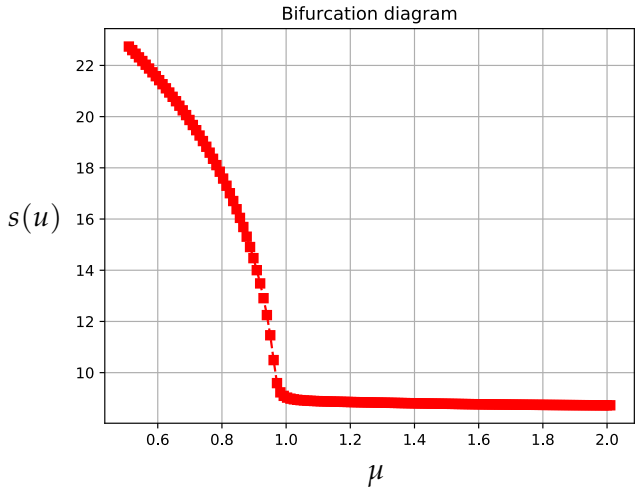


Figure 10.22: Bifurcation diagram for the Navier-Stokes system reconstructed through POD-NN.

10.2.2.2 Multi-parameter case with varying geometry

As we said many times, the recovery of a single bifurcating branch is interesting, but the more complex multi-parameter case is worth to be studied. Indeed, in such case, one can investigate how the bifurcating properties of the system change. Thus, we generalize the model we discussed above with a geometrical parametrization setting. In fact, we consider the usual channel geometry and we introduce a parametrization on the semi-height of the narrow inlet $w \in \mathcal{P}_g = [0.5, 2.0]$.

The physics context is clear, by reducing or increasing the height of the inlet channel, we are actually changing the intrinsic value of the Reynolds number, balancing differently the inertial and viscous forces. This means that we expect the critical point of the model to vary accordingly to the different geometry considered. More specifically, we aim at observing that the smaller is the parameter w, and thus the corresponding inlet, the sooner the bifurcation can occur.

As an example we show in Figure 10.23 four solutions at $\mu = 0.5$ for different values of the parameter w.

Therefore, as we can see already from these velocity profiles, varying the geometry we obtained different bifurcation points, one for each fixed value of w, since we restrict ourselves to the first critical point. So, if we correctly explore the parameter space offline, then we will be able to efficiently reconstruct and investigate the 3-D bifurcation diagram during the online phase, with both RB and POD-NN strategies.

As we observed empirically from the velocity profiles in Figure 10.23, the bifurcation phenomena happen for values of w in $[0.5, 1.5]$, thus, we restricted the study to this range for the geometrical parameter.

We present the bifurcation diagrams in Figures 10.24 and 10.25, respectively a 3-D version with respect to the viscosity μ and the 2-D one with respect to the Reynolds

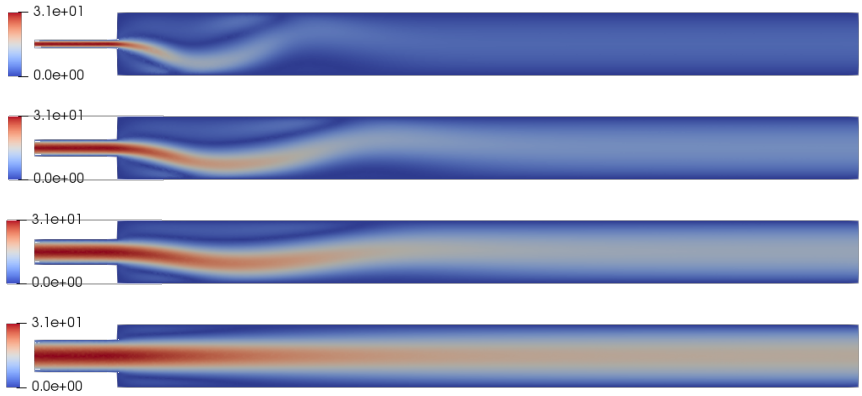


Figure 10.23: High fidelity velocity profiles for the Navier-Stokes system with $w = 0.5, 1.0, 1.5, 2.0$ and $\mu = 0.5$, from top to bottom.

number linked to the geometry chosen. For the output, in this case we chose a measure of the symmetry of the flow w.r.t. the mid-line of the channel. From the aforementioned figures, it is evident the effect of the geometrical parametrization on the location of the critical point.

We remark that we are still considering an output which depends on the $\mathcal{N}$ degrees of freedom of the high fidelity solution. Again, this is only for visual purposes, in fact different affine decomposable functionals can be considered. But, in this context, we will develop a strategy in order to fully exploit the artificial neural network capability, efficiently recovering the bifurcation diagram during the online phase, without involving any $\mathcal{N}$-dependent quantity.

For the offline sampling, we chose 6 equispaced points on $\mathcal{P}_g = [0.5, 1.5]$ and then we proceeded as always, with $N_{train} = 216$ equispaced values of μ for each branch.

Since this generalized problem is much more difficult then the previous one, we selected as a basis for the reduced manifold $N_u = 50$ velocity basis functions and $N_p = 24$ pressure basis functions corresponding to a POD tolerance equal to $\varepsilon_{POD} = 10^{-8}$.

In Figure 10.26 we show the error $\varepsilon_{NN}(\boldsymbol{\mu})$ for velocity and pressure fields. For both networks, velocity and pressure, we chose the hyper-parameters in the following way: $L_K = 2$ layers, $n_b = 18$ and $H_K = 15$. This way, we were able to reconstruct 11 equispaced branches for $w \in [0.5, 1.5]$ where the maximum and the mean errors for the velocity field, over the testing dataset represented again by the continuation sample, are respectively $\varepsilon_{NN}^{\max} = 0.06250$ and $\bar{\varepsilon}_{NN} = 0.01181$. Similar result holds for the pressure field, where we have $\varepsilon_{NN}^{\max} = 0.11417$ and $\bar{\varepsilon}_{NN} = 0.01150$.

Instead, as concerns the RB approximation results, we have $\varepsilon_{RB}^{\max} = 0.75538$ and $\bar{\varepsilon}_{RB} = 0.01297$ for the velocity field and $\varepsilon_{RB}^{\max} = 0.69706$ and $\bar{\varepsilon}_{RB} = 0.00984$ for the pressure field. Thus, we can see that in some cases the RB technique fails the convergence to the solution, giving a bad approximation results with an unacceptable maximum error $\varepsilon_{RB}^{\max}$. On the contrary, since the POD-NN strategy is not based on the explicit for-

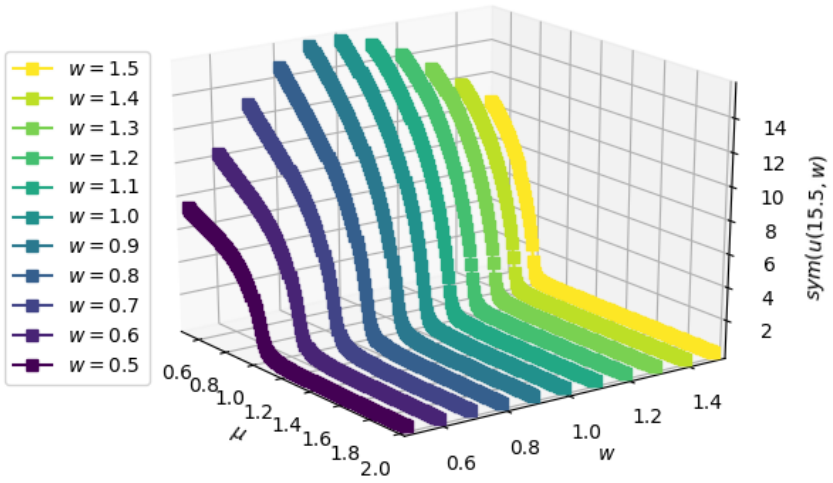


Figure 10.24: 3D bifurcation diagram for the Navier-Stokes system with geometrical parametrization of the narrow inlet, with respect to the viscosity.

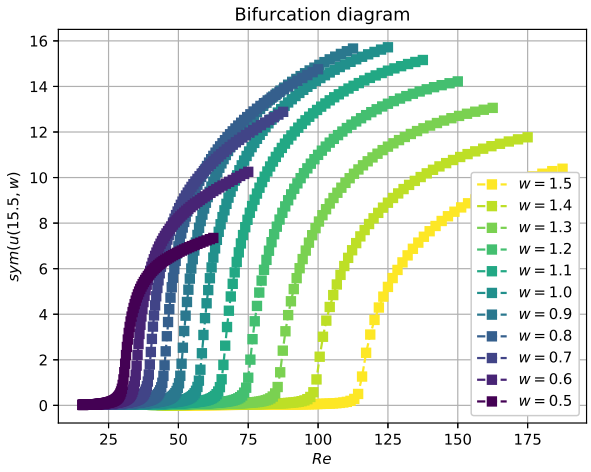


Figure 10.25: 2-D bifurcation diagram for the Navier-Stokes system with geometrical parametrization of the narrow inlet with respect to the Reynolds number.

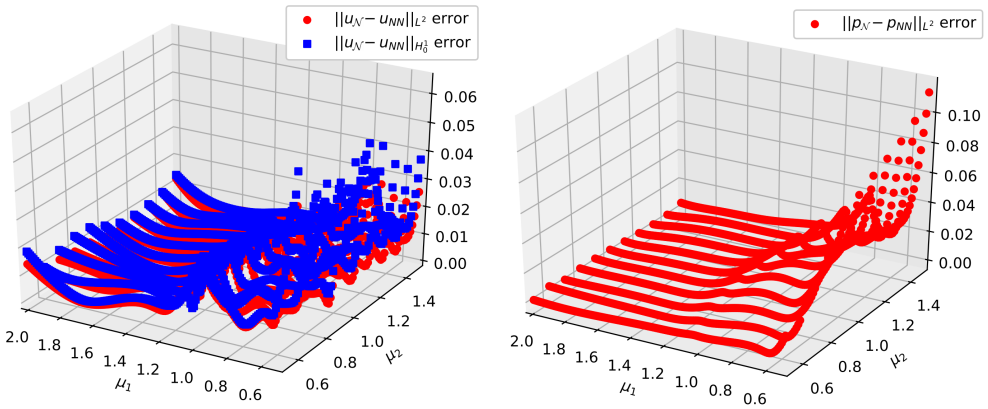


Figure 10.26: Error $\varepsilon_{NN}(\boldsymbol{\mu})$ for velocity and pressure fields, left and right respectively

mulation of the model, it does not involve any convergence issue but only a network evaluation.

As regard the computational time, since we are now considering a greater number of basis functions (in order to compress the bifurcating phenomena linked with different critical points in the parametric space) all the efficiency of the reduced model is lost (here we have not implemented any hyper-reduction technique or efficient tensor assembly). In fact, each RB or HF solution requires almost 10(s). This means that if we want a complete bifurcation diagram with e.g. $n = 151$ values for μ and $m = 51$ points for w, we would need 7701 solution, thus almost 1 day of computational time using either RB or HF solver. Contrarily, a single evaluation of the network, which provides the coefficients for the reduced basis approximation, takes only $t_{NN} = 1.2 \cdot 10^{-5}$(s), enabling for a speed-up of almost $8 \cdot 10^5$.

10.2.3 A flow inside a triangular cavity

In this section, we present the preliminary results for the triangular cavity benchmark. This test case is again modelled by the Navier-Stokes equations. In order to have a coherent structure, we chose here to remain within the velocity-pressure variables pair, instead of modelling the flow by means of the streamfunction-vorticity formulation that usually appears in this context.

We consider the triangular domain Ω in Figure 10.27 identified by the vertexes $A = [a_x, a_y]$, $B = [b_x, b_y]$ and $C = [c_x, c_y]$. The Navier-Stokes system in (10.3) is supplemented here with the following boundary conditions: $u = 0$ on Γ_{wall} and $u = [1, 0]^T$ on Γ_{lid}. Moreover, since there are no Neumann conditions, we have to impose e.g. on the bottom vertex C a zero value for the pressure field. In particular, the geometry represented in Figure 10.27 is the equilateral reference configuration, where $A = [-\sqrt{3}, 1]$, $B = [\sqrt{3}, 1]$ and $C = [0, -2]$.

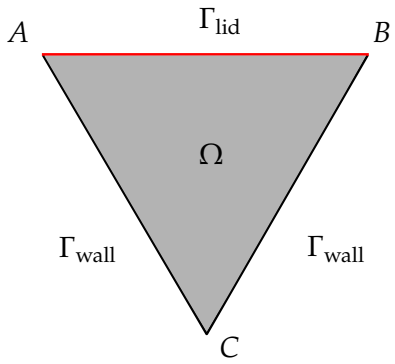


Figure 10.27: Domain Ω which represents a triangular cavity.

10.2.3.1 Equilateral geometry toy problem

As a first benchmark, we investigated the one dimensional parameter space, in which we study the behaviour of both the RB methodology and the POD-NN method for the approximation of the solution varying the viscosity parameter μ. Here we present only the best case scenario for the POD-NN, corresponding to the following setting of the hyper-parameters. The viscosity range is chosen as $\mathcal{P} = [0.001, 1]$, thus corresponding to a maximum value of the Reynolds number equal to 1000.

Having a one dimensional parameter space, corresponds to an input dimension for the network $P = 1$, while the output layer has dimension equal to the number of reduced basis considered, thus $N_u = N_p = 14$ basis functions which were selected by the POD with tolerance $\varepsilon_{POD} = 10^{-9}$. We remark that for the velocity field we have to add again the lifting basis function, to keep track of the non-homogeneous Dirichlet condition on the moving wall Γ_{lid}.

The network is constructed using $H_K = 15$ hidden neurons per layer, the initial learning parameter is fixed to $\eta_0 = 5.\text{e-}2$ with square root learning decay in (10.1). We train and validate the network with a batch size $n_b = 24$, which corresponds to $i = 3$ and $N = 432$ training points. Such points are selected through a logarithmically equispaced distribution on $\mathcal{P}$ and then divided using the 75% of the snapshots as the training points, while the remaining ones for the cross-validation. This choice is motivated by the fact that slight changes for μ have a greater importance when the viscosity is small. For the online simulation Ξ_{te} is chosen as a logarithmically equispaced sample of 250 points in $[0.001, 1]$. Here we remark that a much lower number of training points, e.g. $N_{train} = 36$, is indeed sufficient to reach a good accuracy of the neural network approximation.

In the best case scenario, reported in Figure 10.28, the maximum error and the mean error over the testing dataset for the velocity field are respectively $\varepsilon_{NN}^{\max} = 0.00265$ and $\bar{\varepsilon}_{NN} = 0.00037$. Similar result holds for the pressure field in Figure 10.29, where we have $\varepsilon_{NN}^{\max} = 0.06637$ and $\bar{\varepsilon}_{NN} = 0.01091$. As comparison we show the reduced basis error behaviour for the velocity and pressure field in Figure 10.30, 10.31.

As we can see, in this case the neural network approximation of the velocity field

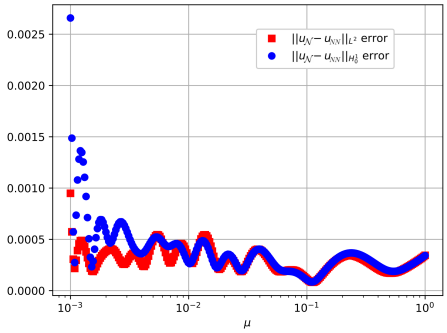


Figure 10.28: Error $\varepsilon_{NN}(\boldsymbol{\mu})$ for the velocity field computed on Ξ_{te}.

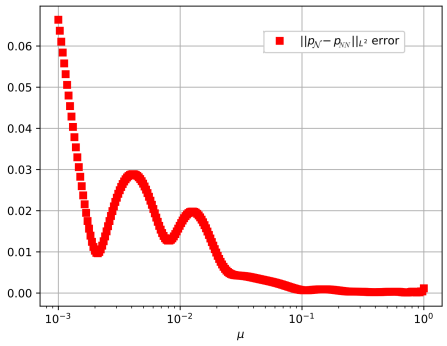


Figure 10.29: Error $\varepsilon_{NN}(\boldsymbol{\mu})$ for the pressure field computed on Ξ_{te} .

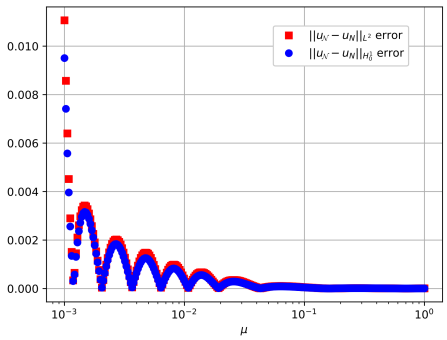


Figure 10.30: Error $\varepsilon_{RB}(\boldsymbol{\mu})$ for the velocity field computed on Ξ_{te}.

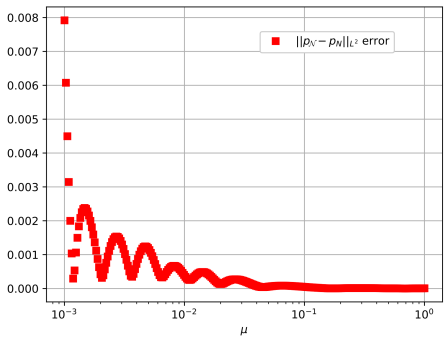


Figure 10.31: Error $\varepsilon_{RB}(\boldsymbol{\mu})$ for the pressure field computed on Ξ_{te} .

outperforms the reduced basis approximation. The same is not true for the recovery of the pressure field that seems more sensitive for small values of the viscosity.

Moreover, we show in Figures 10.32, 10.33 some representative high fidelity velocity and pressure fields solutions obtained for different values of the viscosity $\mu = \{1, 0.1, 0.01, 0.001\}$.

In Figure 10.34 we show the relative errors between the high fidelity approximation and the neural network one for $\mu = 0.001$ for the velocity and pressure fields.

We remark that the streamfunction ψ and the vorticity ω fields can be easily recovered from the velocity field by solving linear Poisson equations. The speed-up of the evaluation of the network on Ξ_{te} with respect to the high fidelity analysis is of order $6.5 \cdot 10^4$.

Moreover, we tried to understand the behaviour of the system for a much higher value of the Reynolds number, choosing a value for the viscosity up to the order 10^{-4}. Unfortunately, if the mesh is not sufficiently refined at a certain point the high fidelity solver starts an oscillating behaviour that prevents the convergence. This happens for a Reynolds number Re ≈ 2702. The last computed high fidelity solution is depicted

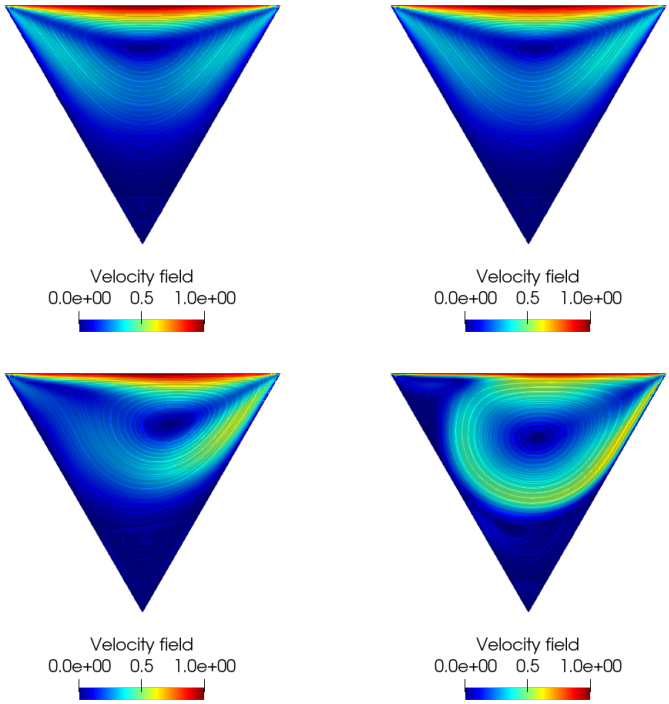


Figure 10.32: High fidelity velocity field for the Navier-Stokes system in the triangular cavity for $\mu = \{1, 0.1, 0.01, 0.001\}$, from left to right, top to bottom.

in Figure 10.35. What we have understood from the previous chapters is that this non-convergence behaviour could have been caused by a bifurcating phenomenon. Indeed, in a recent paper the authors found another configuration for Re $\approx$ 5000 [5]. For this reason in the next section will show some preliminary results of the critical points investigation for the cavity flow when a geometrical parametrization is taken into consideration.

Finally, when attempting to recover an higher value of Reynolds number within the reduced context, also the nonlinear iteration at the reduced level fails to converge. Indeed when Ξ_{te} is chosen as a logarithmically equispaced sample of $K = 250$ points in $[0.00037, 1]$, the maximum and the mean neural network error for the velocity field are respectively given by $\varepsilon_{NN}^{\max} = 0.01449$ and $\bar{\varepsilon}_{NN} = 0.00271$. The corresponding quantities for the reduced basis approximation, which confirm what we have already seen for the channel flow, are $\varepsilon_{RB}^{\max} = 0.98907$ and $\bar{\varepsilon}_{RB} = 0.05078$.

10.2.3.2 Bifurcating regime with parametrized geometry

In this section, we consider the same benchmark analysed previously, but in a multi-parameter context. During the previous analysis we observed that increasing the Re number, a vortex originates inside the cavity, thus our aim is to investigate if this phys-

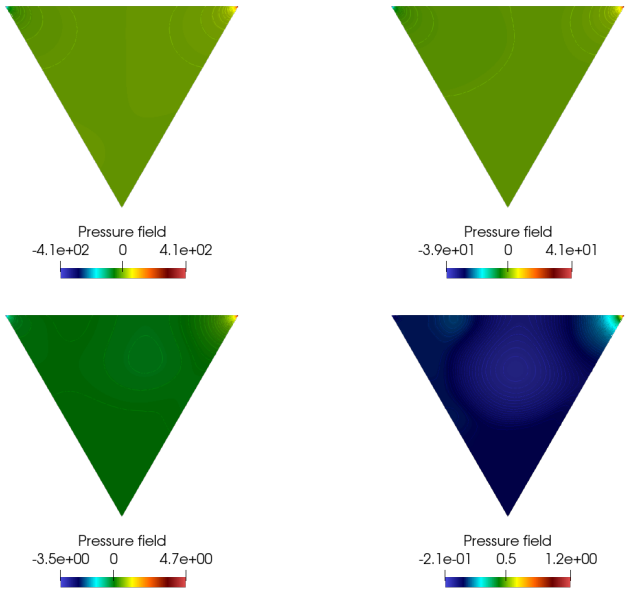


Figure 10.33: High fidelity pressure field for the Navier-Stokes system in the triangular cavity for $\mu = \{1, 0.1, 0.01, 0.001\}$, from left to right, top to bottom.

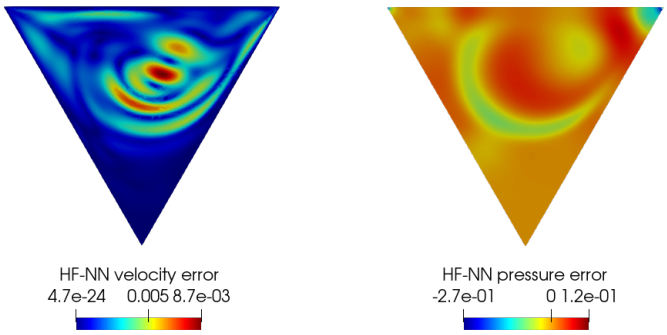


Figure 10.34: Relative errors HF-NN on the velocity and pressure fields for the Navier-Stokes system in the triangular cavity with $\mu = 0.001$.

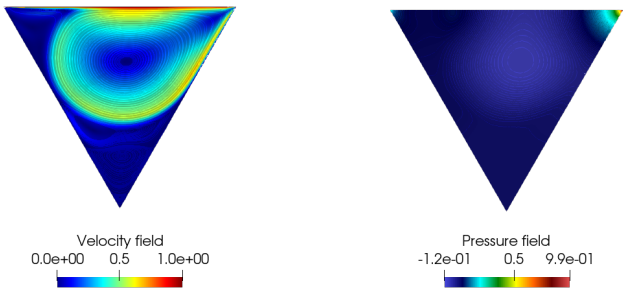


Figure 10.35: Velocity and pressure fields for the Navier-Stokes system in the triangular cavity for $\mu = 0.00037$.

ical structure looses its stability also in connection to different geometries. Indeed, we added a geometrical parametrization of the triangular cavity, where the bottom vertex C has now coordinate $[c_x, c_y] = [\mu_1, \mu_2]$ that we let vary in parameter domain $\mathcal{P}_g = [-1.5, 1.5] \times [-3, -1]$. Of course, since we are changing a lot the characteristic of the triangular domain we expect that this will deeply influence the flow behaviour. As a preliminary test case, we considered an offline phase with $N_{train} = 64$ logarithmically equispaced points, for each of the 5×5 geometrical configurations of the cavity obtained by an equispaced sampling in $\mathcal{P}_g$. As concerns the online simulation, Ξ_{te} is chosen as a logarithmically equispaced sample of $K = 120$ points in $[0.005, 1]$, for a random pair in $\mathcal{P}_g$.

Moreover, we show in Figure 10.36 some representative high fidelity velocity field solutions obtained for $\mu = 0.005$ and different geometry configurations. It is clear that, even if we have considered a lower value for the largest Re number, varying the geometry already produced a change in the behaviour of the velocity field. Indeed from the top row in Figure 10.36 we can observe that if the angle corresponding to the vertex B is small enough, there exists a value for the viscosity μ such that the vortex starts moving towards the top-right vertex.

For this benchmark, we have $\boldsymbol{\mu} = (\mu, \mu_1, \mu_2)$ thus $P = 3$, while we need a much larger set of basis functions for the reduced model, which translates in $N_u = 100$ and $N_p = 44$. The network is constructed using $H_K = 20$ hidden neurons for each layer $L_K = 3$, the initial learning parameter is fixed to $\eta_0 = 5.e\text{-}2$ with square root learning decay in (10.1). We train and validate the network with a batch size $n_b = 8$. We divide again the dataset, depicted in Figure 10.37 using the 75% of the snapshots as the training points.

For this test case, we did not investigate the property of the network varying its hyper-parameters, indeed this could entail a too heavy computational cost also for the offline step. Therefore, we show in Figure 10.38 a representative high fidelity velocity and pressure fields solution obtained for the pair $[\mu_1, \mu_2] = [1.19, -1.48]$ and viscosity $\mu = 0.005$.

The maximum error and the mean error over the testing dataset for the velocity

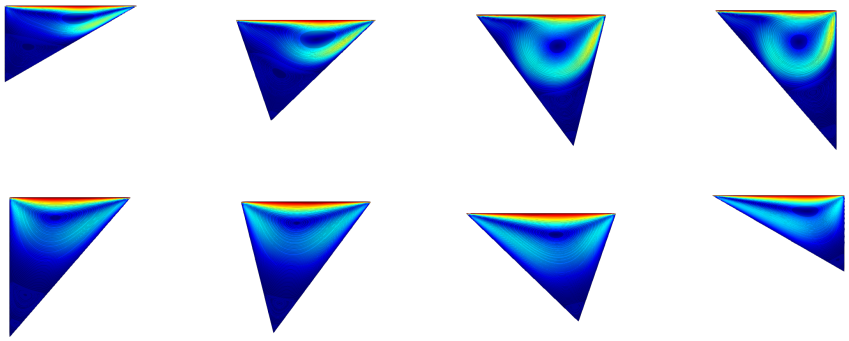


Figure 10.36: High fidelity velocity fields for the Navier-Stokes system in the triangular cavity for $\mu = 0.005$

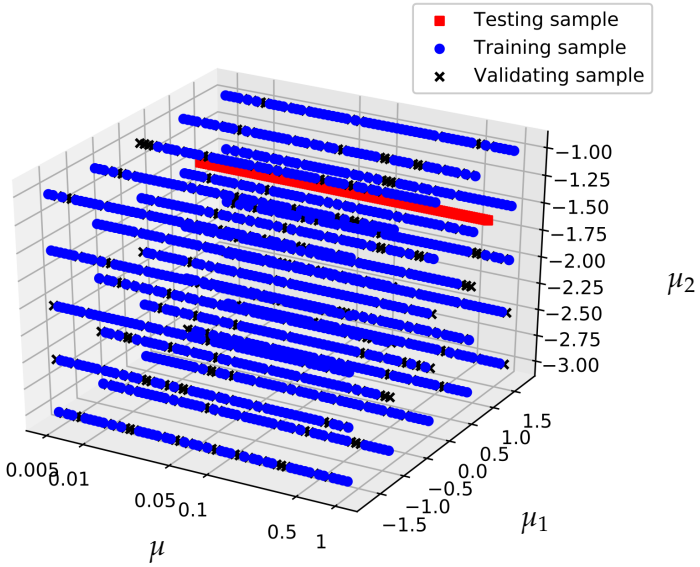


Figure 10.37: Dataset for tricavity benchmark.

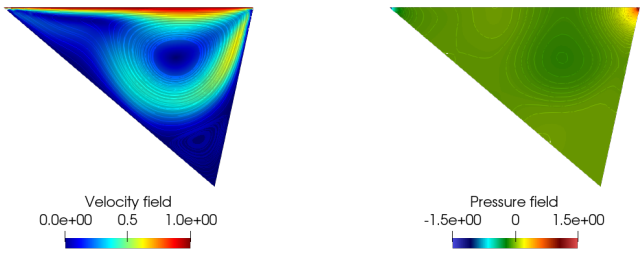


Figure 10.38: Velocity and pressure fields for the Navier-Stokes system in the triangular cavity with $[\mu_1, \mu_2] = [1.19, -1.48]$ for $\mu = 0.005$.

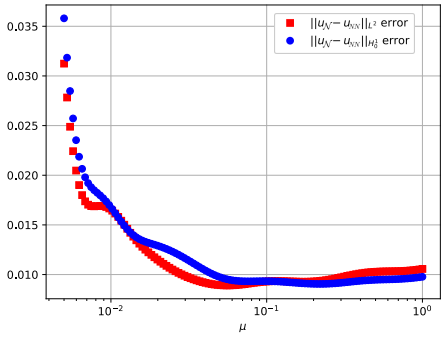


Figure 10.39: Error $\varepsilon_{NN}(\boldsymbol{\mu})$ for the velocity field computed on Ξ_{te}.

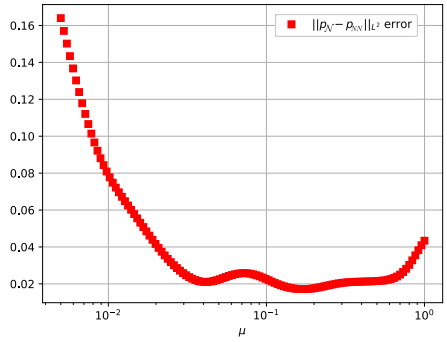


Figure 10.40: Error $\varepsilon_{NN}(\boldsymbol{\mu})$ for the pressure field computed on Ξ_{te}.

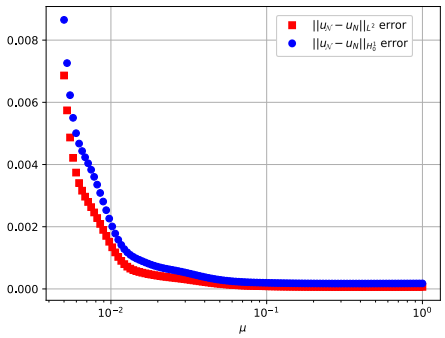


Figure 10.41: Error $\varepsilon_{RB}(\boldsymbol{\mu})$ for the velocity field computed on Ξ_{te}.

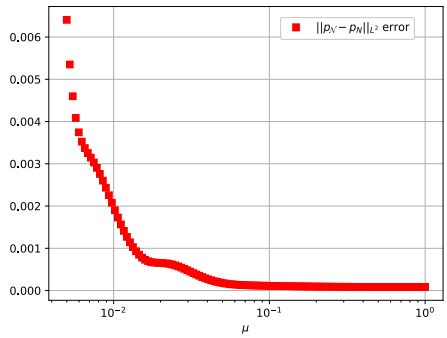


Figure 10.42: Error $\varepsilon_{RB}(\boldsymbol{\mu})$ for the pressure field computed on Ξ_{te} .

field are respectively $\varepsilon_{NN}^{\max} = 0.03661$ and $\bar{\varepsilon}_{NN} = 0.01267$. For the pressure field we have a not satisfactory $\varepsilon_{NN}^{\max} = 0.16154$ and $\bar{\varepsilon}_{NN} = 0.04086$. We report in Figure 10.39 and 10.40 the neural network errors for the velocity and pressure fields. As comparison we show the reduced basis error behaviour for the velocity and pressure field in Figure 10.41, 10.42. Finally, the speed-up of the POD-NN technique with respect to the HF method is again of order $1.2 \cdot 10^5$.

Motivated by the investigation of the flow properties when dealing with high Reynolds number, we considered an isosceles triangular reference configuration with $A = [-0.5, 0]$, $B = [0.5, 0]$ and $C = [\mu_1, -0.5]$, varying $\mu_1 \in \mathcal{P}_g = [-0.5, 0.5]$. Refining properly the mesh and the continuation method, we were able to overcome the non-converging issue, reaching the viscosity value $\mu = 0.0002$ corresponding to Re $= 5000$. From this setting, as we can see in Figure 10.43, is still more evident the vertex-attaching behaviour that we observed previously. We remark that, as can be seen from Figure 10.43, it seems that it exists a critical value for the angle on the vertex B, in this case equal to $\pi/4$, after that such phenomena does no longer occur. This will serve as a starting point for an eigenvalue analysis, as the one performed in the previous

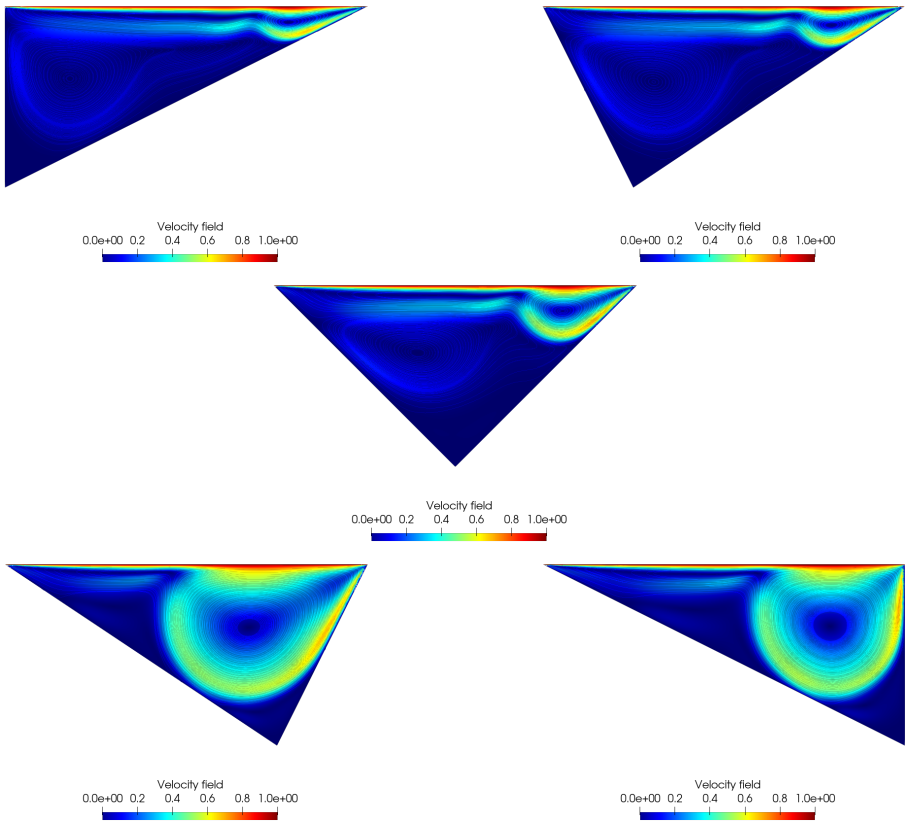


Figure 10.43: High fidelity velocity fields for the Navier-Stokes system in the triangular cavity for $\mu = 0.0002$

chapter, to discover exactly how the bifurcation point is affected by the geometry.

10.3 A reduced manifold based bifurcation diagram

In this section, we aim at developing a non-intrusive strategy, based on theoretical considerations, in order to reproduce the bifurcation diagram of a potentially unknown bifurcating phenomenon. As we have seen in the previous section, the investigation of the critical points, e.g. for parametrized models, can be too expansive considering the full order models. Thus, with these motivations, we want to find an efficient way to predict the critical points of a parametrized model. The starting point will be the POD-NN technique, which allows us to recover information on the dynamics with a non-intrusive approach.

This can be crucial in the design of a nonlinear solver when a parametrized problem shows non-uniqueness behaviour. In fact in order to catch the bifurcating phenomena we usually have an information on the position of the critical point and we refine the parameter grid until it is able to reproduce the symmetry breaking.

We thus aim at efficiently reconstruct a bifurcation diagram where the output is entirely based on the reduced coefficients that appears in the reduced basis expansion. We remark that the approach we are describing can be equivalently be pursued with the reduced coefficients obtained by POD-NN or RB, i.e. from the output of the neural network or the solution of the RB projection. As we will see the main issue with the pure RB approach is that we need a reasonable rich set of reduced vector, that in absence of hyper-reduction technique is infeasible. For this reason, from now on, all the simulations we are showing are obtained from the evaluation of the network.

The paradigm of the reduced order models is the approximation of the truth manifold of the solution for a parametrized PDE, by means of a reduced manifold spanned by some suitable basis functions. In this context, from the POD technique, we already have the basis and the reduced coefficients that express the solution in this basis.

As we understood, dealing with bifurcating models is a difficult task since the solutions near to a critical point change suddenly their behaviour. Thus, the truth manifold, represented by some chosen output, exhibits a $\mathcal{C}^1$ discontinuity, as we observed also in this chapter in Figures 10.24 and 10.25.

The idea is to take advantage of the non-smoothness of the manifold, constructing a detection tool that is able to track the critical points. Indeed, if we are able to reconstruct at the reduced level the bifurcation diagram with good accuracy (as proved by the many different applications considered), while using the same basis for the reduced manifold, then it means that also the reduced coefficients will show a sudden variation, since they encode the parametrized information.

Motivated by this consideration we decided to choose as detection tool, an approximation of the curvature of the *reduced coefficients manifold*. Indeed, if we assume that the basis we have constructed for the reduced space contains enough information, the reduced coefficients have to reflect the bifurcation phenomenon, showing the maximum curvature w.r.t. the parameter space near the critical points.

Moreover, we remark now the connection between the curvature and the Laplace operator. The latter being the divergence of the gradient, measures how much the value of the field differs from its average value taken over the surrounding points.

Let us describe more in detail the approach we developed. We will apply this methodology to the channel flow test case with geometrical parametrization we have analysed in Section 10.2.2.2. As before, we consider the parametrization $\boldsymbol{\mu} = (\mu, w) \in \mathcal{P} = [0.5, 2] \times [0.5, 2]$ and we sample $n = 301$ points for the viscosity μ and $m = 151$ points for the semi-height w of the narrow inlet. In the following, we will only discuss the approach based on the velocity field. From the POD strategy we obtain N velocity basis functions $\{\Sigma_u^k\}_{k=1}^N$ such that we can write

$$u^{NN}(\boldsymbol{\mu}) = \sum_{k=1}^{N} u_{NN}^{(k)}(\boldsymbol{\mu})\Sigma_u^k,$$

where $\mathsf{u}_{NN}(\boldsymbol{\mu}) = \{u_{NN}^{(k)}(\boldsymbol{\mu})\}_{k=1}^N$ is the vector of the reduced coefficients, obtained through the POD-NN technique, for each $\boldsymbol{\mu} \in \mathcal{P}$.

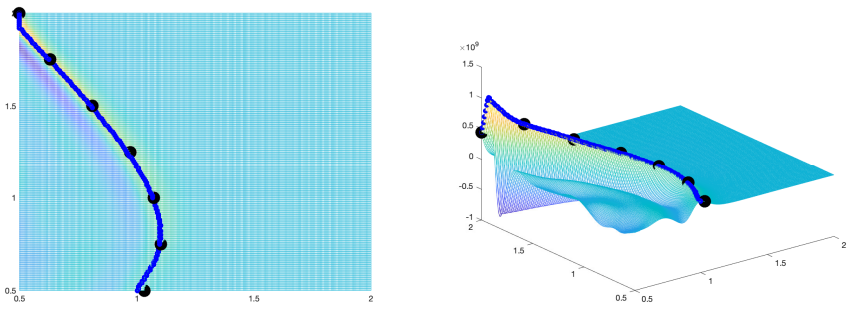


Figure 10.44: Reduced manifold based 3D bifurcation diagram for the Navier-Stokes system with geometrical parametrization.

Thus, we can efficiently and non-intrusively compute $u_{NN}(\boldsymbol{\mu})$ for each pair of the grid $G = \{(\mu_i, w_j)\}_{(i,j)=(1,1)}^{(n,m)}$ we defined above. In this way, we obtain a 3D tensor given by $U_{i,j}^k$ for each $i = 1, \dots, n$, $j = 1, \dots, m$ and $k = 1, \dots, N$.

The algorithm reported in Algorithm 3, illustrates the procedure to find the critical points. For each value of the inlet's semi-height w, we define the curvature by means of the Matlab function *del2* of the sum over all the N coefficients vector U_i^k. Now, we have for each point in the grid G the value of the corresponding curvature, thus we can simply compute the maximum value of this quantity for each fixed w, to obtain a curve in G that describe the critical points location.

Algorithm 3 Reduced manifold based bifurcation diagram

1: **for** $j = 1 : m$ **do** ▷ Loop on each fixed geometry
2: $\quad u_{i,k} = U_{i,j}^k \in \mathbb{R}^{n \times N}$ ▷ Extract the coefficients for the basis for all μ
3: $\quad C_j = \sum_{k=1}^{N} \Delta u_{i,k} \in \mathbb{R}^n$ ▷ Compute the global curvature
4: **end for**
5: $M = \arg\max_i \Delta C_{i,j} \in \mathbb{R}^m$ ▷ Find indexes of local maxima for the curvature
6: $X_c = \{\mu_i\}_{i \in M}$ ▷ Extract the vector of critical points for each w

We can see a preliminary result in Figure 10.44, where the black dots are the exact (high fidelity) critical points for seven equispaced values of w in $\mathcal{P}_g$, and the blue line is the approximation for each value w_j in G.

We remark that in this setting the L^2 relative error for the vector of the critical points is of the order 10^{-2}. Finally, we remark that none of the reconstructed branches, to which correspond the black dots, have been sampled during the offline strategy.

Conclusions and perspectives

In this thesis we have developed and analysed a reduced order methodology to investigate complex bifurcating behaviours held by nonlinear parametrized PDEs. In particular, a branch-wise algorithm for the reconstruction of the bifurcation diagram was developed both at the high fidelity and the reduced order level. This was implemented by combining Newton method, Galerkin FE method and a continuation method, in order to linearize, solve and follow the solution branch. Such approach allowed us to depict the multiple coexisting solutions for many different models, investigating their stability properties together with the singularities of the equations which define them. Buckling phenomena, bifurcations from linear limit and wall-hugging effect were studied in an efficient framework, by means of the application of empirical interpolation strategies. Many different test cases were discussed, taking into consideration both geometrical and physical parametrization, understanding how the critical points of the model behave with respect to multi-parameter settings. A great effort was spent to analyse the stability of the solutions, linking both theoretically and numerically this property to the study of the eigenvalues. Let us recap some conclusions.

- The plate buckling phenomenon was studied by means of the Von Kármán equations finding the different modes which correspond to the successive bifurcation of the model. We observed that changing sign eigenvalues are linked to the bifurcation points of the model and we confirmed this by means of a non-parametrized generalized eigenvalue problem. Moreover, we showed that compressing non-uniformly the plate one obtains different critical points, and we developed an ad-hoc strategy based on the eigenvectors to investigate the post buckling behaviour. We were able to detect multiple bifurcations, which occur in connection to multiple eigenvalue, and secondary bifurcations, utilizing a more complicated continuation method. Finally, we apply the Empirical Interpolation Method to recover the efficiency, obtaining a good speed-up.

- The Gross-Pitaevskii model and its rich bifurcating behaviour was investigated, for an efficient recovery of its first six bifurcating branches. A multi-parameter analysis were performed varying the trap strength to confine the solution in

a smaller region, and at the same time the expected evolution of the critical points was observed. We studied the comparison between the two empirical interpolation technique presented, remarking the issues that such methodologies could exhibit within the bifurcating context.

- We performed an extensive study of the buckling problem for beams. In particular, the Saint Venant-Kirchhoff and neo-Hookean constitutive relations were analysed and discussed in connection to different test cases. We studied the forced and the unforced, two-dimensional and three-dimensional rectangular beam subjected to a compression imposed by means of Neumann and Dirichlet boundary conditions. Two multi-parameter cases were studied, varying the physical properties of the beam, such as the Young modulus and the Poisson ratio, and its length through a geometrical parameter. As expected we found that also in these cases the second parameter change the position of the bifurcation point, as an example we observed that longer beams buckles at lower compression. These investigations allowed us to deal with a more complex benchmark coming from industrial application. A three-dimensional element with annular cross section was studied, finding unexpected behaviours, investigating the buckling with respect to its length and to the type of compression imposed.

- We studied a CFD application governed by the Navier-Stokes equations in a channel, which exhibits a bifurcating phenomena of relevant interest also from the medical viewpoint. The Coanda effect, which translates into a wall hugging behaviour of the velocity profile, was discussed and its stability analysed. Indeed, we observed that the straight unstable profile is characterized by the change of sign of the rightmost eigenvalue. On the contrary, the coexisting wall-hugging profile originates as the stable configuration and all its corresponding eigenvalues have negative real parts.

- We exploited the latter application to investigate more complex models employing emerging methodologies. We developed a reduced order approach to deflated continuation method, which allowed us to discover automatically and efficiently the full bifurcation diagram, which present in this case two kind of wall-hugging behaviour.

- We modelled and investigated an optimal control approach to bifurcating systems. We showed that we were able to act on the state system, steering the solution towards desired profiles. In particular, we analyzed a weak Neumann boundary control on the outlet, which was only capable of changing quantitatively the solutions. We discussed a strong action obtained by imposing a distributed control, which allowed us to define the concept of natural solution. We investigated the effect of the control parameter, and we found that the smaller it was the more difficult was to recover the non-natural branch. Finally, we imposed the control by means of Dirichlet boundary conditions and we discovered a change in the stability properties of the studied profiles.

- To overcome the issues and the computational complexity deriving from the implementation of hyper-reduction strategies, we investigated a POD-NN non-intrusive approach to bifurcating phenomena. We were able to recover efficiently the bifurcation diagram for a multi-parameter geometrical variant of the channel flow. We also showed preliminary results for the bifurcating behaviour of a cavity flow in a triangular domain, for which we discussed the existence of critical quantities which can determine the vortex structure. Finally, we developed a reduced manifold based approach for the detection of bifurcation points which allowed us to discover non-intrusively the singularities of the model for each geometrical configuration.

Perspectives

During the investigation of the different methodologies and models we have presented in this thesis, we have identified several connected research topics that should be further studied and developed. As concerns the methodological approach one could investigate the Greedy sampling procedure, by theoretically develop a posteriori error estimates based upon the Brezzi-Rappaz-Raviart theory [24]. Such analysis could also be fundamental for multi-parameter applications in view of a POD-Greedy approach. The properties of the Empirical Interpolation Method could be theoretically investigated, in particular when a global approach for the snapshot sampling is taken into consideration.

From the viewpoint of the applications, many different bifurcating models can be considered within the same context we developed. Indeed, it could be used to discover efficiently the critical points of a physical phenomenon which depends on multiple parameters. As an example, one could apply this methodology to structural stability problems in a real-time setting. Indeed, the efficiency enabled by the ROMs allows us to perform such complex, parameter dependent, computations even on a webserver, as we did for the buckling plate benchmark in ARGOS [126].

The reduced order approach to deflated continuation method could be investigated further, understanding how it behaves when a branch is not sampled during the offline procedure. Moreover, this could be applied to recover information for time-dependent problems, which are usually characterized by Hopf bifurcations.

Optimal control problems proved to be a fundamental tool to drive the behaviour of bifurcating models. We studied different settings, but the complex action of the control can be discovered for many different applications, depending on the goal one wants to achieve. Indeed, as we showed for the Dirichlet control, such models could lead to unexpected solutions.

Many different paths could be pursued by applying machine learning techniques to the bifurcating context. We restricted ourselves to the POD-NN technique, but more advanced strategies can be investigated, as the physics informed neural networks (PINNs) [134]. Indeed, these could be used to obtain better accuracy while reducing the amount of offline data required [37]. Moreover, in this context, one could search

for the best configuration for the network by means of Bayesian optimization. Finally, a deeper analysis of the POD-NN based bifurcation detection tool can be made, since it provided several promising results already at this stage.

Bibliography

[1] G. Alfimov and D. Zezyulin. Nonlinear modes for the Gross–Pitaevskii equation - a demonstrative computation approach. *Nonlinearity*, 20(9):2075, 2007.

[2] E. Allgower and K. Georg. *Introduction to Numerical Continuation Methods*. Society for Industrial and Applied Mathematics, 2003.

[3] A. Ambrosetti and A. Malchiodi. *Nonlinear Analysis and Semilinear Elliptic Problems*. Number v. 10 in Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2007.

[4] A. Ambrosetti and G. Prodi. *A Primer of Nonlinear Analysis*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995.

[5] B. An, J. M. Bergada, and F. Mellibovsky. The lid-driven right-angled isosceles triangular cavity flow. *Journal of Fluid Mechanics*, 875:476–519, 2019.

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

[7] I. Babuška and J. Osborn. Eigenvalue problems. *Handbook of numerical analysis*, 2:641–787, 1991.

[8] E. Bader, M. Kärcher, M. A. Grepl, and K. Veroy. Certified reduced basis methods for parametrized distributed elliptic optimal control problems with control constraints. *SIAM Journal on Scientific Computing*, 38(6):A3921–A3946, 2016.

[9] E. Bader, M. Kärcher, M. A. Grepl, and K. Veroy-Grepl. A certified reduced basis approach for parametrized linear-quadratic optimal control problems with control constraints. *IFAC-PapersOnLine*, 48(1):719–720, 2015.

[10] F. Ballarin, A. Manzoni, A. Quarteroni, and G. Rozza. Supremizer stabilization of POD–Galerkin approximation of parametrized steady incompressible Navier–Stokes equations. *International Journal for Numerical Methods in Engineering*, 102(5):1136–1161, 2015.

[11] F. Ballarin, F. Pichi, and G. Rozza. A successive partition method for the efficient evaluation of parametrized stability factors. Preprint, 2020.

[12] M. Barrault, N. C. Nguyen, Y. Maday, and A. T. Patera. An "empirical interpolation" method: Application to efficient reduced-basis discretization of partial differential equations. *C. R. Acad. Sci. Paris, Série I.*, 339:667–672, 2004.

[13] L. Bauer, H. B. Keller, and E. L. Reiss. Multiple eigenvalues lead to secondary bifurcation. *SIAM Review*, 17(1):101–122, 1975.

[14] L. Bauer and E. L. Reiss. Nonlinear buckling of rectangular plates. *Journal of the Society for Industrial and Applied Mathematics*, 13(3):603–626, 1965.

[15] P. Benner, A. Cohen, M. Ohlberger, and K. Willcox. *Model Reduction and Approximation: Theory and Algorithms*. Computational Science and Engineering Series. SIAM, Society for Industrial and Applied Mathematics, 2017.

[16] P. Benner, S. Gugercin, and K. Willcox. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Review*, 57(4):483–531, 2015.

[17] P. Benner, M. Ohlberger, A. T. Patera, G. Rozza, and K. Urban, editors. *Model Reduction of Parametrized Systems*. MS&A Series Vol. 17. Springer International Publishing, 2017.

[18] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 2005.

[19] M. S. Berger. On Von Kármán's equations and the buckling of a thin elastic plate, I the clamped plate. *Communications on Pure and Applied Mathematics*, 20(4):687–719, 1967.

[20] P. B. Bochev and M. D. Gunzburger. *Least-squares finite element methods*, volume 166. Springer-Verlag, New York, 2009.

[21] N. Boullé and P. E. Farrell. Bifurcation analysis with symmetry groups. *ENS Rennes and University of Oxford. Summer internship - Report*, 2017.

[22] J. Bramble, J. Pasciak, and A. Vassilev. Uzawa type algorithms for nonsymmetric saddle point problems. *Mathematics of Computation*, 69(230):667–689, 2000.

[23] F. Brezzi. Finite element approximations of the Von Kármán equations. *RAIRO. Anal. numér.*, 12(4):303–312, 1978.

[24] F. Brezzi, J. Rappaz, and P. A. Raviart. Finite dimensional approximation of nonlinear problems. *Numerische Mathematik*, 36(1):1–25, 1980.

[25] A. Buffa, Y. Maday, A. Patera, and G. Turinici. A priori convergence of the greedy algorithm for the parameterized reduced basis. *ESAIM Mathematical Modelling and Numerical Analysis*, 46(3):595–603, 2011.

[26] J. Burkardt, M. Gunzburger, and H. C. Lee. POD and CVT-based reduced-order modeling of Navier–Stokes flows. *Computer Methods in Applied Mechanics and Engineering*, 196(1-3):337–355, 2006.

[27] G. Caloz and J. Rappaz. Numerical analysis for nonlinear and bifurcation problems. *Handbook of numerical analysis*, 5:487–637, 1997.

[28] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík. Guaranteed a posteriori bounds for eigenvalues and eigenvectors: multiplicities and clusters. *Mathematics of Computation*, March 2020.

[29] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík. Post-processing of the planewave approximation of Schrödinger equations. Part I: linear operators. Preprint, January 2020.

[30] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods: fundamentals in single domains*, volume XXII, 581. Scientific Computation. Springer-Verlag Berlin Heidelberg, Berlin, 2006.

[31] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods: evolution to complex geometries and applications to fluid dynamics*, volume XXX, 596. Springer Science & Business Media, Berlin, 2007.

[32] C. Canuto, T. Tonn, and K. Urban. A posteriori error analysis of the reduced basis method for nonaffine parametrized nonlinear PDEs. *SIAM Journal on Numerical Analysis*, 47(3):2001–2022, 2009.

[33] D. Chapelle, A. Gariah, P. Moireau, and J. Sainte-Marie. A Galerkin strategy with proper orthogonal decomposition for parameter-dependent problems: Analysis, assessments and applications to parameter estimation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 47(6):1821–1843, 2013.

[34] E. Charalampidis, N. Boullé, P. E. Farrell, and P. Kevrekidis. Bifurcation analysis of stationary solutions of two-dimensional coupled Gross–Pitaevskii equations using deflated continuation. *arXiv:1912.00023*, 2019.

[35] E. Charalampidis, P. Kevrekidis, and P. E. Farrell. Computing stationary solutions of the two-dimensional Gross–Pitaevskii equation with deflated continuation. *Communications in Nonlinear Science and Numerical Simulation*, 54:482 – 499, 2018.

[36] S. Chaturantabut and D. C. Sorensen. Nonlinear model reduction via discrete empirical interpolation. *SIAM Journal on Scientific Computing*, 32(5):2737–2764, 2010.

[37] W. Chen, Q. Wang, J. S. Hesthaven, and C. Zhang. Physics-informed machine learning for reduced-order modeling of nonlinear problems. Preprint, 2020.

[38] Y. Chen, J. S. Hesthaven, Y. Maday, and J. Rodríguez. Certified reduced basis methods and output bounds for the harmonic Maxwell's equations. *SIAM Journal on Scientific Computing*, 32(2):970–996, 2010.

[39] C. S. Chien and M. S. Chen. Multiple bifurcation in the Von Kármán equations. *SIAM Journal on Scientific Computing*, 18, 1997.

[40] C. S. Chien, S. Y. Gong, and Z. Mei. Mode jumping in the Von Kármán equations. *SIAM Journal on Scientific Computing*, 22(4):1354–1385, 2000.

[41] P. G. Ciarlet. *Mathematical Elasticity, Volume I: Three-Dimensional Elasticity*. Elsevier Science, 1988.

[42] P. G. Ciarlet. *Mathematical Elasticity: Volume II: Theory of Plates*. Studies in Mathematics and its Applications. Elsevier Science, 1997.

[43] P. G. Ciarlet. *Linear and Nonlinear Functional Analysis with Applications*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 2013.

[44] P. G. Ciarlet and P. A. Raviart. A mixed finite element method for the biharmonic equation. In *Proceedings of Symposium on Mathematical Aspects of Finite Elements in PDE*, pages 125–145, 1974.

[45] A. Cohen and R. DeVore. Kolmogorov widths under holomorphic mappings. *IMA Journal of Numerical Analysis*, 36(1):1–12, 03 2015.

[46] M. Coles, D. Pelinovsky, and P. Kevrekidis. Excited states in the large density limit: a variational approach. *Nonlinearity*, 23(8):1753, 2010.

[47] A. Contreras and C. Garcia-Azpeitia. Global bifurcation of vortex and dipole solutions in Bose-Einstein condensates. *Comptes Rendus Mathematique*, 354(3):265 – 269, 2016.

[48] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[49] L. Dedè. Reduced basis method and a posteriori error estimation for parametrized linear-quadratic optimal control problems. *SIAM Journal on Scientific Computing*, 32(2):997–1019, 2010.

[50] S. Deparis. Reduced basis error bound computation of parameter-dependent Navier–Stokes equations by the natural norm approach. *SIAM journal on numerical analysis*, 46(4):2039–2067, 2008.

[51] E. J. Doedel. *Lecture Notes on Numerical Analysis of Nonlinear Equations*, pages 1–49. Springer Netherlands, Dordrecht, 2007.

[52] E. J. Doedel, H. B. Keller, and J. P. Kernevez. Numerical analysis and control of bifurcation problems (i): Bifurcation in finite dimensions. *International Journal of Bifurcation and Chaos*, 01(03):493–520, 1991.

[53] G. Dusson and Y. Maday. A posteriori analysis of a nonlinear Gross—Pitaevskii-type eigenvalue problem. *IMA Journal of Numerical Analysis*, 37:94–137, 2017.

[54] L. C. Evans. *Partial Differential Equations*. Graduate studies in mathematics. American Mathematical Society, 2010.

[55] P. E. Farrell, C. H. L. Beentjes, and A. Birkisson. The computation of disconnected bifurcation diagrams. *arXiv preprint arXiv:1603.00809*, 2016.

[56] P. E. Farrell, A. Birkisson, and S. W. Funke. Deflation techniques for finding distinct solutions of nonlinear partial differential equations. *SIAM Journal on Scientific Computing*, 37(4):A2026–A2045, 2015.

[57] FEniCS. https://fenicsproject.org.

[58] S. Fresca, L. Dede, and A. Manzoni. A comprehensive deep learning-based approach to reduced order modeling of nonlinear time-dependent parametrized PDEs. https://arxiv.org/abs/2001.04001, 2020.

[59] I. Fumagalli, A. Manzoni, N. Parolini, and M. Verani. Reduced basis approximation and a posteriori error estimates for parametrized elliptic eigenvalue problems. *ESAIM: M2AN*, 50(6):1857–1885, 2016.

[60] C. García-Azpeitia and D. Pelinovsky. Bifurcations of Multi-Vortex Configurations in Rotating Bose–Einstein Condensates. *Milan Journal of Mathematics*, 85(2):331–367, Dec 2017.

[61] A. L. Gerner and K. Veroy. Certified reduced basis methods for parametrized saddle point problems. *SIAM Journal on Scientific Computing*, 34(5):A2812–A2836, 2012.

[62] DNV GL. *Class guideline: Buckling DNVGL-CG-0128*, October 2015.

[63] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 2012.

[64] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[65] M. A. Grepl, Y. Maday, N. C. Nguyen, and A. T. Patera. Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 41(3):575–605, 8 2007.

[66] M. D. Gunzburger. *Perspectives in flow control and optimization*, volume 5. SIAM, Philadelphia, 2003.

[67] M. Guo and J. S. Hesthaven. Data-driven reduced order modeling for time-dependent problems. *Computer methods in applied mechanics and engineering*, 345:75–99, 2019.

[68] M. E. Gurtin. *An Introduction to Continuum Mechanics*. ISSN. Elsevier Science, 1982.

[69] B. Haasdonk. Convergence rates of the POD-greedy method. *ESAIM: Mathematical Modelling and Numerical Analysis*, 47(3):859–873, 2013.

[70] B. Haasdonk and M. Ohlberger. Reduced basis method for finite volume approximations of parametrized linear evolution equations. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 42(2):277–302, 2008.

[71] H. Herrero, Y. Maday, and F. Plá. RB (Reduced basis) for RB (Rayleigh–Bénard). *Computer Methods in Applied Mechanics and Engineering*, pages 132–141, 2013.

[72] M. Hess, A. Alla, A. Quaini, G. Rozza, and M. Gunzburger. A localized reduced-order modeling approach for PDEs with bifurcating solutions. *Computer Methods in Applied Mechanics and Engineering*, 351:379 – 403, 2019.

[73] M. Hess, A. Quaini, and G. Rozza. Reduced basis model order reduction for Navier–-Stokes equations in domains with walls of varying curvature. *International Journal of Computational Fluid Dynamics*, pages 1–8, 2019.

[74] J. S. Hesthaven, G. Rozza, and B. Stamm. *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*. SpringerBriefs in Mathematics. Springer International Publishing, 2015.

[75] J. S. Hesthaven, B. Stamm, and S. Zhang. Efficient greedy algorithms for high-dimensional parameter spaces with applications to empirical interpolation and reduced basis methods. *ESAIM: Mathematical Modelling and Numerical Analysis*, 48(1):259–283, 2014.

[76] J. S. Hesthaven and S. Ubbiali. Non-intrusive reduced order modeling of nonlinear problems using neural networks. *Journal of Computational Physics*, 363:55–78, 2018.

[77] M. L. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23. Springer Science & Business Media, Antwerp, 2008.

[78] T. Horger, B. Wohlmuth, and T. Dickopf. Simultaneous reduced basis approximation of parameterized elliptic eigenvalue problems. *ESAIM: M2AN*, 51(2):443–465, 2017.

[79] F. Hutter, L. Kotthoff, and J. Vanschoren, editors. *Automatic machine learning: methods, systems, challenges*. Challenges in Machine Learning. Springer, Germany, 2019.

[80] D. B. P. Huynh, A. T. Patera, and G. Rozza. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations: Application to transport and continuum mechanics. *Archives of Computational Methods in Engineering*, 15:229–275, 2008.

[81] D. B. P. Huynh, F. Pichi, and G. Rozza. *Reduced Basis Approximation and A Posteriori Error Estimation: Applications to Elasticity Problems in Several Parametric Settings*, pages 203–247. Springer International Publishing, Cham, 2018.

[82] M. Kahlbacher and S. Volkwein. Galerkin proper orthogonal decomposition methods for parameter dependent elliptic systems. *Discussiones Mathematicae, Differential Inclusions, Control and Optimization*, 27(1):95–117, 2007.

[83] T. Kapitula and K. Promislow. *Spectral and dynamical stability of nonlinear waves*, volume 185. Springer, 2013.

[84] M. Kärcher and M. A. Grepl. A certified reduced basis method for parametrized elliptic optimal control problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 20(2):416–441, 2014.

[85] M. Kärcher, Z. Tokoutsi, M. A. Grepl, and K. Veroy. Certified reduced basis methods for parametrized elliptic optimal control problems with distributed controls. *Journal of Scientific Computing*, 75(1):276–307, 2018.

[86] H. B. Keller. Lectures on numerical methods in bifurcation problems. *Applied Mathematics*, 217, 1987.

[87] P. Kevrekidis, D. Frantzeskakis, and R. Carretero-González. *The Defocusing Nonlinear Schrödinger Equation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2015.

[88] P. Kevrekidis, V. Konotop, A. Rodrigues, and D. Frantzeskakis. Dynamic generation of matter solitons from linear states via time-dependent scattering lengths. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 38(8):1173, 2005.

[89] H. Kielhöfer. *Bifurcation Theory: An Introduction with Applications to PDEs*. Applied Mathematical Sciences. Springer New York, 2006.

[90] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[91] Y. Kivshar, T. Alexander, and S. Turitsyn. Nonlinear modes of a macroscopic quantum oscillator. *Physics Letters A*, 278(4):225 – 230, 2001.

[92] G. Kutyniok, P. Petersen, M. Raslan, and R. Schneider. A theoretical analysis of deep neural networks and parametric PDEs. *arXiv preprint arXiv:1904.00377*, 2019.

[93] Y. Kuznetsov. *Elements of Applied Bifurcation Theory*. Applied Mathematical Sciences. Springer New York, 2004.

[94] K. Lee and K. T. Carlberg. Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders. *Journal of Computational Physics*, 404:108973, 2020.

[95] G. Leugering, P. Benner, S. Engell, A. Griewank, H. Harbrecht, M. Hinze, R. Rannacher, and S. Ulbrich. *Trends in PDE constrained optimization*. Springer, New York, 2014.

[96] L. Machiels, Y. Maday, I. B. Oliveira, A. T. Patera, and D. V. Rovas. Output bounds for reduced-basis approximations of symmetric positive definite eigenvalue problems. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 331(2):153–158, 2000.

[97] Y. Maday, N. C. Nguyen, A. T. Patera, and G. S. H. Pau. A general multipurpose interpolation procedure: the magic points. *Communications on Pure & Applied Analysis*, 8(1):383, 2009.

[98] Y. Maday and A. T. Patera. Spectral element methods for the Navier-Stokes equations. *State-of-the-art surveys on computational mechanics (A90-47176 21-64). New York*, pages 71–143, 1989.

[99] Y. Maday, A. T. Patera, and J. Peraire. A general formulation for a posteriori bounds for output functionals of partial differential equations; application to the eigenvalue problem. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 328(9):823–828, 1999.

[100] Y. Maday, A. T. Patera, and E. M. Rønquist. A well-posed optimal spectral element approximation for the Stokes problem. *Final Report National Aeronautics and Space Administration. Langley Research Center, Hampton, VA.*, 1987.

[101] A. Manzoni. An efficient computational framework for reduced basis approximation and a posteriori error estimation of parametrized Navier–Stokes flows. *ESAIM: Mathematical Modelling and Numerical Analysis*, 48(4):1199–1226, 2014.

[102] B. J. Matkowsky, L. J. Putnick, and E. L. Reiss. Secondary states of rectangular plates. *SIAM Journal on Applied Mathematics*, 38(1):38–51, January 1980.

[103] J. M. Melenk. On n-widths for elliptic problems. *Journal of Mathematical Analysis and Applications*, 247:272–289, 2000.

[104] S. Middelkamp, P. Kevrekidis, D. Frantzeskakis, R. Carretero-González, and P. Schmelcher. Bifurcations, stability, and dynamics of multiple matter-wave vortex states. *Phys. Rev. A*, 82:013646, Jul 2010.

[105] S. Middelkamp, P. Kevrekidis, D. Frantzeskakis, R. Carretero-González, and P. Schmelcher. Emergence and stability of vortex clusters in Bose-Einstein condensates: A bifurcation approach near the linear limit. *Physica D: Nonlinear Phenomena*, 240(18):1449 – 1459, 2011.

[106] F. Millar and D. Mora. A finite element method for the buckling problem of simply supported Kirchhoff plates. *Journal of Computational and Applied Mathematics*, 286:68 – 78, 2015.

[107] S. Mishra and R. Molinaro. Estimates on the generalization error of Physics Informed Neural Networks (PINNs) for approximating PDEs. *ArXiv*, abs/2006.16144, 2020.

[108] H. K. Moffatt. Viscous and resistive eddies near a sharp corner. *Journal of Fluid Mechanics*, 18(1):1–18, 1964.

[109] F. Negri, A. Manzoni, and G. Rozza. Reduced basis approximation of parametrized optimal flow control problems for the Stokes equations. *Computers & Mathematics with Applications*, 69(4):319–336, 2015.

[110] F. Negri, G. Rozza, A. Manzoni, and A. Quarteroni. Reduced basis method for parametrized elliptic optimal control problems. *SIAM Journal on Scientific Computing*, 35(5):A2316–A2340, 2013.

[111] N. C. Nguyen, G. Rozza, and A. T. Patera. Reduced basis approximation and a posteriori error estimation for the time-dependent viscous Burgers' equation. *Calcolo*, 46(3):157–185, 2009.

[112] A. Noor. On making large nonlinear problems small. *Computer Methods in Applied Mechanics and Engineering*, 34(1):955 – 985, 1982.

[113] A. Noor and J. Peters. Multiple-parameter reduced basis technique for bifurcation and post-buckling analyses of composite materiale. *International Journal for Numerical Methods in Engineering*, 19:1783–1803, 1983.

[114] Norsok. *Norsok standard: Design of steel structures*, October 2004.

[115] A.T. Patera and G. Rozza. *Reduced basis approximation and A posteriori error estimation for Parametrized Partial Differential Equation*. MIT Pappalardo Monographs in Mechanical Engineering, Copyright MIT (2007-2010).

[116] B. Peherstorfer and K. Willcox. Dynamic data-driven reduced-order models. *Computer Methods in Applied Mechanics and Engineering*, 291:21–41, 2015.

[117] F. Pichi, F. Ballarin, G. Rozza, and J. S. Hesthaven. Artificial neural network for bifurcating phenomena modelled by nonlinear parametrized PDEs. Preprint, 2020.

[118] F. Pichi, J. Eftang, G. Rozza, and A. T. Patera. Reduced order models for the buckling of hyperelastic beams. MIT-FVG "ROM2S" report, 2020.

[119] F. Pichi, A. Quaini, and G. Rozza. A reduced order modeling technique to study bifurcating phenomena: Application to the Gross–Pitaevskii equation. *SIAM Journal on Scientific Computing*, 42(5):B1115–B1135, 2020.

[120] F. Pichi and G. Rozza. Reduced basis approaches for parametrized bifurcation problems held by non-linear Von Kármán equations. *Journal of Scientific Computing*, 81(1):112–135, 2019.

[121] F. Pichi, M. Strazzullo, F. Ballarin, and G. Rozza. Driving bifurcating parametrized nonlinear PDEs by optimal control strategies: application to Navier-Stokes equations and model reduction. Preprint, 2020.

[122] M. Pintore, F. Pichi, M. Hess, G. Rozza, and C. Canuto. Efficient computation of bifurcation diagrams with a deflated approach to reduced basis spectral element method. arXiv:1912.06089, 2019.

[123] L. P. Pitaevskii and S. Stringari. *Bose-Einstein Condensation*. International Series of Monographs on Physics. Oxford University Press, 2003.

[124] G. Pitton, A. Quaini, and G. Rozza. Computational reduction strategies for the detection of steady bifurcations in incompressible fluid-dynamics: Applications to Coanda effect in cardiology. *Journal of Computational Physics*, 344:534 – 557, 2017.

[125] G. Pitton and G. Rozza. On the application of reduced basis methods to bifurcation problems in incompressible fluid dynamics. *Journal of Scientific Computing*, 73(1):157–177, 2017.

[126] ARGOS: Advanced Reduced Groupware Online Simulation Platform. Developed by Prof. Gianluigi Rozza's group within SISSA mathLab, Trieste, Italy.

[127] C. Prud'homme, D. Rovas, K. Veroy, Y. Maday, A. T. Patera, and G. Turinici. Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods. *Journal of Fluids Engineering*, 124(1):70–80, March 2002.

[128] A. Quaini, R. Glowinski, and S. Canic. Symmetry breaking and preliminary results about a Hopf bifurcation for incompressible viscous flow in an expansion channel. *International Journal of Computational Fluid Dynamics*, 30(1):7–19, 2016.

[129] A. Quarteroni. *Numerical Models for Differential Problems*. MS&A Series Vol. 16. Springer International Publishing, 2017.

[130] A. Quarteroni, A. Manzoni, and F. Negri. *Reduced Basis Methods for Partial Differential Equations: An Introduction*. UNITEXT. Springer International Publishing, 2015.

[131] A. Quarteroni, G. Rozza, and A. Manzoni. Certified reduced basis approximation for parametrized partial differential equations and applications. *Journal of Mathematics in Industry*, 1(1):3, 2011.

[132] A. Quarteroni, G. Rozza, and A. Quaini. Reduced basis methods for optimal control of advection-diffusion problems. In *Advances in Numerical Mathematics*, pages 193–216. RAS and University of Houston, 2007.

[133] A. Quarteroni and A. Valli. *Numerical approximation of partial differential equations*, volume 23. Springer Science & Business Media, Berlin and Heidelberg, 2008.

[134] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, February 2019.

[135] RBniCS. http://mathlab.sissa.it/rbnics.

[136] T. C. Rebollo, E. D. Ávila, M. G. Mármol, F. Ballarin, and G. Rozza. On a certified Smagorinsky reduced basis turbulence model. *SIAM Journal on Numerical Analysis*, 55(6):3047–3067, 2017.

[137] F. Regazzoni, L. Dedè, and A. Quarteroni. Machine learning for fast and reliable solution of time-dependent differential equations. *Journal of Computational Physics*, 397:108852, 2019.

[138] L. Reinhart. On the numerical analysis of the Von Kármán equations: Mixed finite element approximation and continuation techniques. *Numerische Mathematik*, 39(3):371–404, Oct 1982.

[139] G. Rozza and K. Veroy. On the stability of the reduced basis method for Stokes equations in parametrized domains. *Computer Methods in Applied Mechanics and Engineering*, 196(7):1244–1260, 2007.

[140] S. Salsa. *Partial Differential Equations in Action: From Modelling to Theory*. UNITEXT. Springer International Publishing, 2016.

[141] D. Schaeffer and M. Golubitsky. Boundary conditions and mode jumping in the buckling of a rectangular plate. *Communications in Mathematical Physics*, 69(3):209–236, 1979.

[142] R. Seydel. *Practical Bifurcation and Stability Analysis*. Interdisciplinary Applied Mathematics. Springer New York, 2009.

[143] Y. Shin, J. Darbon, and G. E. Karniadakis. On the convergence and generalization of Physics Informed Neural Networks. *ArXiv preprint arXiv:2004.01806*, 2020.

[144] M. Stein. *Loads and Deformations of Buckled Rectangular Plates*. NASA Technical Report. U.S. Government Printing Office, 1959.

[145] M. Stoll and A. Wathen. All-at-once solution of time-dependent Stokes control. *J. Comput. Phys.*, 232(1):498–515, January 2013.

[146] M. Strazzullo, F. Ballarin, R. Mosetti, and G. Rozza. Model reduction for parametrized optimal control problems in environmental marine sciences and engineering. *SIAM Journal on Scientific Computing*, 40(4):B1055–B1079, 2018.

[147] M. Strazzullo, F. Ballarin, and G. Rozza. POD-Galerkin model order reduction for parametrized nonlinear time dependent optimal flow control: an application to Shallow Water Equations. Submitted, 2020.

[148] M. Strazzullo, Z. Zainib, F. Ballarin, and G. Rozza. Reduced order methods for parametrized nonlinear and time dependent optimal flow control problems: towards applications in biomedical and environmental sciences. Submitted, https://arxiv.org/abs/1912.07886, 2019.

[149] F. Terragni and J. Vega. On the use of POD-based ROMs to analyze bifurcations in some dissipative systems. *Physica D: Nonlinear Phenomena*, 241.17:1393–1405, 2012.

[150] S. P. Timoshenko and J. M. Gere. *Theory of Elastic Stability*. Dover Civil and Mechanical Engineering. Dover Publications, 2009.

[151] D. J. Tritton. *Physical fluid dynamics*. Springer Science & Business Media, 2012.

[152] F. Tröltzsch. Optimal control of partial differential equations. *Graduate studies in mathematics*, 112, Verlag, Wiesbad, 2010.

[153] K. Veroy. *Reduced-Basis Methods Applied to Problems in Elasticity: Analysis and Applications*. PhD thesis, Massachusetts Institute of Technology, 2003.

[154] K. Veroy and A. T. Patera. Certified real-time solution of the parametrized steady incompressible navier stokes equations: rigorous reduced-basis a posteriori error bounds. *International Journal for Numerical Methods in Fluids*, 47(8-9):773–788, 2005.

[155] T. Von Kármán. Festigkeitsprobleme im Maschinenbau. *Encyclopädie der Mathematischen Wissenschaften*, 4, 1910.

[156] Z. Zainib, F. Ballarin, S. Fremes, P. Triverio, L. Jiménez-Juan, and G. Rozza. Reduced order methods for parametric optimal flow control in coronary bypass grafts, toward patient-specific data assimilation. *International Journal for Numerical Methods in Biomedical Engineering*, page e3367, 2020.

[157] L. Zanon. *Model Order Reduction for Nonlinear Elasticity: Applications of the Reduced Basis Method to Geometrical Nonlinearity and Finite Deformation*. PhD thesis, RWTH Aachen University, 2017.

[158] E. Zeidler. *Nonlinear Functional Analysis and its Applications*, volume I: Fixed-Point Theorems. Springer-Verlag New York, 1985.

[159] S. Zhang and Z. Zhang. Invalidity of decoupling a biharmonic equation to two Poisson equations on non-convex polygons. *Int. J. Numer. Anal. Model*, 5(1):73–76, 2008.