Linear and Nonlinear Schemes for Forward and Inverse Problems

Olga Mula

Dauphine | PSL UNIVERSITÉ PARIS

CEREMADE UMR CNRS 7534

Habilitation à Diriger les Recherches

Spécialité: Mathématiques Appliquées

Linear and Nonlinear Schemes for Forward and Inverse Problems

présentée par

Olga MULA

Soutenue publiquement le 16/12/2021 devant le jury composé de:

Prof. Alexandre Ern	École des Ponts	Président
Prof. Anthony Nouy	École Centrale de Nantes	Rapporteur
Prof. Mario Ohlberger	University of Münster	Rapporteur
Prof. Clémentine Prieur	Université Grenoble-Alpes	Examinatrice
Prof. Gianluigi Rozza	SISSA	Rapporteur
Prof. Gabriel Turinici	Université Paris Dauphine	Coordinateur
Prof. Karen Veroy	TU Eindhoven	Examinatrice

Acknowledgments

First of all, I would like to warmly thank Anthony Nouy, Mario Ohlberger, and Gianluigi Rozza for having accepted to review this manuscript, and for providing very valuable feedback. I feel also very grateful to Alexandre Ern, Clémentine Prieur, and Karen Veroy for taking part in the defense as members the jury. Many thanks also to Gabriel Turinici for having coordinated the submission and the defense at Paris Dauphine.

The writing of this document has been yet another occasion for me to measure the immense chance that I have to be able to pursue an academic career, and more particularly, to be able to do it in such a vibrant place such as CEREMADE, the Applied Mathematics Department of Paris Dauphine University. I would like to take the chance to thank all my colleagues for creating such a nice working environment. I have particularly appreciated the discussions generated within the working group STAT-NUM that I have the pleasure to organize with Angelina Roche and Robin Ryder. I have also enjoyed a lot talking to the students from my lectures, from whom I have learnt a lot on many fronts. Many thanks also to the administrative staff from CEREMADE (Isabelle Bellier, César Faivre and Marie Belle), and from MIDO, the teaching department, for making things work as smoothly as possible concerning administration.

I also feel profoundly grateful to the Inria team COMMEDIA led by Miguel Fernandez for having kindly hosted me as a visiting researcher for two years (from September 2019 to September 2021). My stay at Inria has greatly helped me progress in my research, and I have particularly enjoyed working with Damiano Lombardi on many different topics. It has also been a great pleasure to co-supervise Felipe Galarce's PhD thesis together with Damiano and Jean-Frédéric Gerbeau.

I feel particularly indebted to Yvon Maday, my former PhD advisor, for his continuous guidance and encouragements. His unwavering optimism, and energy, together with his incredible scientific knowledge, and inexhaustible mathematical imagination, have always been very inspirational to me. Thanks also for the trust he has placed in me on several occasions such as for the organisation of the CEMRACS 2021.

I would also like to express my tremendous gratitude to Albert Cohen, and Wolfgang Dahmen, two absolutely outstanding people not only on the scientific side but also on the personal front. Their works have greatly influenced my own research and it has been a great pleasure to collaborate with them.

The research that I have done would not exist had I not had the chance to work with all my scientific collaborators. I would like to particularly thank Peter Binev, Ron DeVore, Virginie Ehrlacher, Jalal Fadili, Harsha Hurtridurga, James Nichols, Anthony Patera, Francesco Salvarani, and François-Xavier Vialard for all the exciting, and fruitful exchanges that we have had during these years.

It has been a great pleasure for me to exchange ideas with my PhD students (Felipe Galarce, Helin Gong, Agustín Somacal), and my postdoctoral students (Joubine Aghili, Minh Hieu Do, Walid Kheriji). It has also been a delight to work with other young researchers in the framework of CEMRACS projects.

The terrible health crisis brought by the Covid-19 pandemic in the past 18 months has made me deeply think about the very meaning of my work in academia, and about the role that I could play to help as a researcher with a skillset in numerical applied mathematics but no previous background in epidemiology. I would like to warmly thank Gabriel Turinici for helping me to find my way in this respect, and for organising a very informative working group on the topic of Covid-19. My interest and concerns on the problem eventually led to two very interesting and fruitful collaborations. The first was a collaboration with Yvon Maday, Athmane Bakhta and Thomas Boiveau. In a very nice atmosphere and working dynamic, we built an epidemiological forecasting method which is summarized at the end of this manuscript. The second is a collaboration with computer scientists from PSL University within the "Facing the Virus" initiative led by Jamal Atif. I would like to thank all the members for all the exciting discussions, in particular, to Jamal Atif, Bertrand Cabot, Olivier Cappé, and Raphaël Pinot.

Last, but certainly not least, I would like to thank my family and friends for all the moments that we have shared together, and for their continuous encouragements, unlimited patience, and daily support.

Contents

Chapter 1

Introduction and Overview of the Contributions

1.1 Forward and Inverse Problems

Parametrized partial differential equations are of common use to model complex physical systems, and are routinely involved in design and decision-making processes. Such equations can generally be written in abstract form as

$$\mathcal{P}(u, y) = 0, \tag{1.1.1}$$

where $\mathcal{P}$ is a partial differential operator, and $y = (y_1, \dots, y_p)$ is a vector of scalar parameters ranging in some domain $\mathrm{Y} \subset \mathbb{R}^p$. We assume well-posedness, that is, for any $y \in Y$ the problem admits a unique solution $u = u(y)$ in some Hilbert space V whose elements depend on a physical variable x ranging in a domain $\Omega \subset \mathbb{R}^d$. The variable x usually refers to space but it is not limited to that meaning, and it may also refer to more elaborate sets of variables such as space and time. We may thus regard u as a function $(x, y) \mapsto u(x, y)$ from $\Omega \times \mathrm{Y}$ to $\mathbb{R}$, or we may also consider the *parameter to solution map*

$$y \mapsto u(y), \tag{1.1.2}$$

from Y to V. This map is typically nonlinear, as well as the *solution manifold*

$$\mathcal{M} := \{u(y) \,:\, y \in Y\} \subset V \tag{1.1.3}$$

which describes the collection of all admissible solutions. Throughout this document, we assume that Y is compact in $\mathbb{R}^d$ and that the map (1.1.2) is continuous. Therefore $\mathcal{M}$ is a compact set of V. We sometimes refer to the solution $u(y)$ as the *state* of the system for the given parameter vector y.

The parameters y are used to represent physical quantities such as diffusivity, viscosity, velocity, source terms, or the geometry of the physical domain in which the PDE is posed. In several relevant instances, y may be high or even countably infinite dimensional, that is, $p \gg 1$ or $p = \infty$.

Given this general setting, two families of problems may be considered:

1. *Forward problems* are concerned with the parameter to solution map (1.1.2). For a given parameter $y \in \mathrm{Y}$, the goal is to develop numerical schemes to solve the PDE problem (2.3.1). This is an old topic with a long history in numerical analysis. It can be addressed with classical discretization techniques such as finite element, finite volume spectral methods, or, less

classically, with machine learning techniques such as, for example, Physics-Informed Neural Networks. For general references to these methods, we refer to [42, 76, 12, 83, 45]. Among the main properties that these methods seek to offer stand:

(a) The ability to estimate the error between the computed approximation and the exact solution. This question is usually addressed via the development of *a posteriori error estimators* that connect the residual of the equation with the exact error. Finite elements are particularly well-suited for this task and we could say that the problem is very well-understood for elliptic problems. There are however still numerous open questions, especially in nonlinear and/or non-coercive problems.

(b) The *numerical efficiency,* in the sense of using the minimum number of degrees of freedom to deliver a certain target accuracy. This property is particularly relevant to address the curse of dimensionality, that is, to prevent that the number of degrees of freedom grows exponentially with the dimension d.

(c) The *easiness of implementation*, which is particularly critical in applicative contexts involving complicated PDE models and complicated domains Ω. This point is probably one of the main appealing features of recent approaches involving machine learning techniques such as PINNs despite their current lack of rigorous guarantees concerning the quality of the final approximation.

In numerous design and decision-making processes, one is often confronted to optimization problems defined over the solution manifold $\mathcal{M}$, where the algorithms are usually iterative and require to evaluate solutions $u(y)$ of the PDE on a large set of dynamically updated parameters $y \in \mathrm{Y}$. Computations cannot be addressed rapidly unless the overall complexity has been appropriately reduced, and motivates the search for accurate methods to approximate the family of solutions very quickly at a reduced computational cost. This task, usually known as *reduced modelling* or *model order reduction*, has classically been addressed by approximating $\mathcal{M}$ with well-chosen linear subspaces of V. However, it can be expected to be successful only when the Kolmogorov width of $\mathcal{M}$ decays fast. While this is the case for certain families of parabolic or elliptic problems (see [28]), most transport-dominated problems are expected to present a slow decaying width and require to study nonlinear approximation methods (see, e.g., [18, Chapter 3]).

2. *Inverse Problems* occur when the parameter y is not given, and, instead, we only observe a vector of *linear* measurements

$$z = (z_1, \dots, z_m) \in \mathbb{R}^m, \quad z_i = \ell_i(u), \quad i = 1, \dots, m,$$

where each $\ell_i \in V'$ is a known continuous linear functional on V. We also sometimes use the notation

$$z = \ell(u), \quad \ell = (\ell_1, \dots, \ell_m).$$

In this setting, the goal is to recover the unknown state $u \in \mathcal{M}$ from z or even the underlying parameter vector $y \in Y$ for which $u = u(y)$. Therefore, in an idealized setting, one observes the result of the composition map

$$y \in Y \mapsto u \in \mathcal{M} \mapsto z \in \mathbb{R}^m.$$

for the unknown y. More realistically, the measurements may be affected by additive noise

$$z_i = \ell_i(u) + \eta_i,$$

and the model itself might be biased, meaning that the true state u deviates from the solution manifold $\mathcal{M}$ by some amount. We will come to these important points later on in the manuscript. For the moment, let us simply remark that two main types of inverse problems may be considered:

(a) *State estimation:* recover an approximation u^* of the state u from the observation $z = \ell(u)$ and assuming that u belongs to the manifold $\mathcal{M}$. This inverse problem is linear in nature because the forward map $\ell : u \mapsto \ell(u)$ is linear. It is however challenging because the target u lives in V, which is a space of typically very high or infinite dimension. In addition, the information that $u \in \mathcal{M}$ is difficult to handle given that $\mathcal{M}$ has a complicated geometry, which is only partially known to us by solving forward problems $y \mapsto u(y)$ for different values of $y \in \mathrm{Y}$.

(b) *Parameter estimation:* recover an approximation y^* of the parameter y from the observation $z = \ell(u)$ when $u = u(y)$. This is a nonlinear inverse problem, for which the prior information available on y is given by the domain Y.

These problems become severely ill-posed when Y has dimension $p > m$. For this reason, they are often addressed through Bayesian approaches [57, 23] in which a prior probability distribution P_y is assumed on $y \in Y$. This induces a push forward distribution P_u for $u \in \mathcal{M}$, and the objective is to understand the *posterior* distributions of y or u conditioned by the observations z in order to compute plausible solutions y^* or u^* under such probabilistic priors. The accuracy of these solutions should therefore be assessed in some average sense.

In this work we do not follow this avenue: the only priors made on y and u are their membership to Y and $\mathcal{M}$. We are interested in developping practical estimation methods that offer uniform recovery guarantees under such deterministic priors in the form of upper bounds on the worst case error for the estimators over all $y \in Y$ or $u \in \mathcal{M}$. We also aim to understand whether our error bounds are optimal in some sense. Our primary focus will actually be on state estimation. Nevertheless we present several implications on parameter estimation, which are new to the best of our knowledge.

1.2 Overview of the contributions

In this manuscript, I summarize a series of contributions revolving around the above topics on forward and inverse problems. The common denominator of all the works is the effort to address problems that are challenging for modern computing architectures. The studied problems are mainly connected either to high-dimensional and/or to transport-dominated PDEs and the ability to solve them with guaranteed accuracy, at a reduced computational cost, and ideally with provable optimality properties in a sense that will depend on the specific context. For every topic, I will discuss ongoing works and possible research directions that can be envisaged in future works.

The first part of the manuscript is devoted to forward problems (Section 2):

- Section 2.1 is devoted to *kinetic PDEs*, which is a family of problems that accumulate several intrinsic obstructions to an efficient and accuracy controlled numerical solution:

– The problem is high-dimensional: its solution u depends in general on $2d+1$ variables (space, momentum and time) so "naive" discretization schemes become prohibitive due to the number of degrees of freedom.

– The solutions have low regularity. Standard a priori error estimates involving classical isotropic Sobolev regularity scales are therefore not very useful for controlling accuracy.

– Kinetic problems involve nontrivial scattering kernels which induce a global coupling on momentum variables. This gives rise to very large and dense matrices when using standard methods based on localization only.

– The optical parameters can present high oscillations in space (due to the spatial heterogeneity of the materials) and in energy.

In Section 2.1, I will summarize [A5], a work done in collaboration with Prof. W. Dahmen and F. Gruber, in which we develop accuracy controlled schemes and corresponding stability notions. To the best of my knowledge, this is the first numerical scheme which rigorously connects the exact solution (at the infinite dimensional level) with the discretized one. The work required significant coding efforts, and we have released an open source library called DUNE-DPG (see [A11]). Section 2.1 also summarizes [A7] which is a work done in collaboration with Profs. F. Salvarani and H. Hutridurga on homogenization of the energy variable to study the nature of the PDE in presenc of high oscillations in energy.

- One desirable feature of numerical schemes for high-dimensional problems is the ability to *parallelize* computations. This can be particularly challenging for certain time-dependent problems due to the inherent sequential nature of the time variable. In Section 2.2, I summarize an adaptive parareal method which I have developed in collaboration with Prof. Y. Maday (see [A8]). The main contribution is the improvement of the parallel efficiency of the method.

- Section 2.3 summarizes a recent line of research that I have initiated on nonlinear model reduction in Wasserstein spaces to address transport dominated problems. In [A6], a work in collaboration with V. Ehrlacher, D. Lombardi and Prof. F.X. Vialard, we leveraged the existence of closed forms of the Wasserstein distance in one spatial dimension ($d = 1$) to develop model reduction strategies for conservation laws. In an ongoing collaboration with J. Feydy and H. Do, we are currently extending the procedure to higher dimensions.

The second part of the manuscript is devoted to inverse problems (Section 3). In Sections 3.1 to 3.4, I give an overview in the form of a review of a series of works done in collaboration with Profs. P. Binev, A. Cohen, W. Dahmen, R. DeVore, J. Fadili, and J. Nichols. This corresponds to the publications [A10, A4, S3]. The main contribution is the development of a state and parameter estimation framework which can be seen as a deterministic counterpart to Bayesian inverse problems. We have developed practical estimation methods that involved model order reduction and that offer uniform recovery guarantees in the form of upper bounds on the worst case error for the estimators over all $y \in Y$ or $u \in \mathcal{M}$. In parallel to these theoretical developments, I have devoted considerable efforts to bring the theoretical developments into applications. Interestingly, even though our theoretical framework is formulated in a very general way, each application comes with specific challenging features that have led us to enlarge some aspects of the theory. Section 3.5 summarizes my contributions in applying the methodology to:

- **Neutronics (nuclear engineering):** This work was done in the framework of H. Gong's PhD CIFRE thesis which I co-supervised with Prof. Yvon Maday, and two colleagues from EDF: B. Bouriquet and J.P. Argaud. This has led to the publication of papers [A9, P1, P2].

- **Biomedical applications:** In this case, inverse problems are usually posed on certain organs or portions of the body which inevitably involve morphological variations between individuals. In the framework of F. Galarce's PhD thesis (co-supervised with J.F. Gerbeau and D. Lombardi), we have developed and analyzed an extension of our general methodology in order to allow to take shape variability into account without needing any a priori knowledge on a parametrization of the geometrical variations. The papers connected to this work are [A2, A3, S1].

- **Epidemiology:** Although I was not involved in research on epidemiology before the Covid-19 pandemic, the gravity of the situation has motivated me to try to put my skills at the service of better understanding the dynamic of the pandemic. I have joined forces with colleagues from PSL, Inria and Sorbonne University to help include mobility data in the modeling of the propagation of the epidemy, and to provide more accurate forecasts on the number of infected and hospitalized people. These interactions have led to two publications [Pop1, Pop2] for the greater public which were covered by some national media. I have also published a paper in collaboration with Prof. Y. Maday, A. Bakhta and T. Boiveau from Carnot Smiles on epidemiological forecasting with model reduction of compartmental models (see [A1]).

- **An ongoing work related to pollution:** I am currently involved in using the methodology for the rapid reconstruction of pollutant concentration maps in large urban areas. We will not present the results in the manuscript given that the work is unfinished. It is being done in the framework of the Emergences Project "Models and Measures" in collaboration with Prof. J. Aghili (Maître de Conférences at Strasbourg University who was previously a post-doc funded by Emergences), R. Chakir, A. Cohen and A. Somal (currently doing his PhD with A. Cohen and myself).

1.3 Publication List

I list here my publications. References [A14, A13, A12] and [P5, P6, P4, P7] were produced or substantially initiated during my PhD. The rest are contributions made after my PhD.

Submitted Preprint Articles

[S1] F. Galarce, D. Lombardi, and O. Mula. "State Estimation with Model Reduction and Shape Variability. Application to biomedical problems". 2021. URL: https://arxiv.org/abs/2106.09421.

[S2] J. Aghili and O. Mula. "Depth-Adaptive Neural Networks from the Optimal Control Viewpoint". 2020. URL: https://arxiv.org/abs/2007.02428.

[S3] A. Cohen, W. Dahmen, O. Mula, and J. Nichols. "Nonlinear reduced models for state and parameter estimation". 2020. URL: https://arxiv.org/abs/2009.02687.

Journal Articles

[A1] A. Bakhta, T. Boiveau, Y. Maday, and O. Mula. "Epidemiological Forecasting with Model Reduction of Compartmental Models. Application to the COVID-19 Pandemic". In: *Biology* 10.1 (2021), p. 22. DOI: https://doi.org/10.3390/biology10010022.

[A2] F. Galarce, J.F. Gerbeau, D. Lombardi, and O. Mula. "Fast reconstruction of 3D blood flows from Doppler ultrasound images and reduced models". In: *Computer Methods in Applied Mechanics and Engineering* 375 (2021), p. 113559. DOI: https://doi.org/10.1016/j.cma.2020.113559.

[A3] F. Galarce, D. Lombardi, and O. Mula. "Reconstructing Haemodynamics Quantities of Interest from Doppler Ultrasound Imaging". In: *Int. J. Numer. Meth. Biomedical Eng.* (2021). DOI: https://doi.org/10.1002/cnm.3416.

[A4] A. Cohen, W. Dahmen, R. DeVore, J. Fadili, O. Mula, and J. Nichols. "Optimal reduced model algorithms for data-based state estimation". In: *SIAM Journal on Numerical Analysis* 58.6 (2020), pp. 3355–3381. DOI: https://doi.org/10.1137/19m1255185.

[A5] W. Dahmen, F. Gruber, and O. Mula. "An Adaptive Nested Source Term Iteration for Radiative Transfer Equations". In: *Mathematics of Computation* (2020). DOI: 10.1090/mcom/3505.

[A6] V. Ehrlacher, D. Lombardi, O. Mula, and F.-X. Vialard. "Nonlinear model reduction on metric spaces. Application to one-dimensional conservative PDEs in Wasserstein spaces". In: *ESAIM M2AN* 54.6 (2020), pp. 2159–2197. DOI: 10.1051/m2an/2020013. URL: https://arxiv.org/abs/1909.06626.

[A7] H. Hutridurga, O. Mula, and F. Salvarani. "Homogenization in the energy variable for a neutron transport model". In: *Asymptotic Analysis* 1–2 (2020), pp. 1–25. DOI: https://doi.org/10.3233/ASY-191544.

[A8] Y. Maday and O. Mula. "An Adaptive Parareal Algorithm". In: *Journal of Computational and Applied Mathematics* (2020). DOI: https://doi.org/10.1016/j.cam.2020.112915. URL: https://arxiv.org/abs/1909.08333.

[A9] J.-P. Argaud, B. Bouriquet, F. de Caso, H. Gong, Y. Maday, and O. Mula. "Sensor placement in nuclear reactors based on the generalized empirical interpolation method". In: *Journal of Computational Physics* 363 (2018), pp. 354–370. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2018.02.050. URL: http://www.sciencedirect.com/science/article/pii/S0021999118301414.

[A10] P. Binev, A. Cohen, O. Mula, and J. Nichols. "Greedy Algorithms for Optimal Measurements Selection in State Estimation Using Reduced Models". In: *SIAM/ASA Journal on Uncertainty Quantification* 6.3 (2018), pp. 1101–1126. DOI: 10.1137/17M1157635. URL: https://doi.org/10.1137/17M1157635.

[A11] F. Gruber, A. Klewinghaus, and O. Mula. "The DUNE-DPG library for solving PDEs with Discontinuous Petrov-Galerkin finite elements". In: *Archive of Numerical Software* 5.1 (2017), pp. 111–127. ISSN: 2197-8263. DOI: 10.11588/ans.2017.1.27719. URL: http://journals.ub.uni-heidelberg.de/index.php/ans/article/view/27719.

[A12] Y. Maday, O. Mula, and G. Turinici. "Convergence analysis of the Generalized Empirical Interpolation Method". In: *SIAM Journal on Numerical Analysis* 54.3 (2016), pp. 1713–1731. DOI: 10.1137/140978843. URL: http://dx.doi.org/10.1137/140978843.

[A13] Y. Maday, O. Mula, A. T. Patera, and M. Yano. "The Generalized Empirical Interpolation Method: Stability theory on Hilbert spaces with an application to the Stokes equation". In: *Computer Methods in Applied Mechanics and Engineering* 287.0 (2015), pp. 310–334. ISSN: 0045-7825. DOI: http://dx.doi.org/10.1016/j.cma.2015.01.018. URL: http://www.sciencedirect.com/science/article/pii/S0045782515000389.

[A14] Y. Maday and O. Mula. "A Generalized Empirical Interpolation Method: application of reduced basis techniques to data assimilation". English. In: *Analysis and Numerics of Partial Differential Equations*. Ed. by Franco Brezzi, Piero Colli Franzone, Ugo Gianazza, and Gianni Gilardi. Vol. 4. Springer INdAM Series. Springer Milan, 2013, pp. 221–235. ISBN: 978-88-470-2591-2. DOI: http://dx.doi.org/10.1007/978-88-470-2592-9_13. URL: http://arxiv.org/pdf/1512.00683.pdf.

Proceedings and Book Chapters

[P1] J. P. Argaud, B. Bouriquet, H. Gong, Y. Maday, and O. Mula. "Monitoring flux and power in nuclear reactors with data assimilation and reduced models". In: *International Conference on Mathematics and Computational Methods Applied to Nuclear Science & Engineering*. 2017. URL: https://www.researchgate.net/publication/316620809_Monitoring_flux_and_power_in_nuclear_reactors_with_data_assimilation_and_reduced_models/citations.

[P2] J. P. Argaud, B. Bouriquet, H. Gong, Y. Maday, and O. Mula. "Stabilization of (G)EIM in Presence of Measurement Noise: Application to Nuclear Reactor Physics". In: *Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2016: Selected Papers from the ICOSAHOM conference, June 27-July 1, 2016, Rio de Janeiro, Brazil*. Ed. by Marco L. Bittencourt, Ney A. Dumont, and Jan S. Hesthaven. Cham: Springer International Publishing, 2017, pp. 133–145. ISBN: 978-3-319-65870-4. DOI: 10.1007/978-3-319-65870-4_8. URL: https://doi.org/10.1007/978-3-319-65870-4_8.

[P3] H. Gong, Q. Li, Y.R. Yu, J. P. Argaud, B. Bouriquet, Y. Maday, and O. Mula. "A new data-driven approach for reconstruction with noisy data and physical constraints: application to nuclear reactor physics". In: *Proceedings of the International Congress on Advances in Nuclear Power Plants*. 2017.

[P4] A.-M. Baudron, J.-J. Lautard, Y. Maday, and O. Mula. "MINARET: Towards a parallel 3D time-dependent neutron transport solver". English. In: *SNA + MC 2013 - Joint International Conference on Supercomputing in Nuclear Applications + Monte Carlo*. 2014. DOI: http://dx.doi.org/10.1051/snamc/201404103. URL: http://dx.doi.org/10.1051/snamc/201404103.

[P5] A.-M. Baudron, J.J. Lautard, Y. Maday, and O. Mula. "The parareal in time algorithm applied to the kinetic neutron diffusion equation". English. In: *Domain Decomposition Methods in Science and Engineering XXI*. Vol. 98. Lecture Notes in Computational Science and Engineering. Springer International Publishing, 2014, pp. 437–445. ISBN: 978-3-319-05788-0.

DOI: 10.1007/978-3-319-05789-7_41. URL: http://dx.doi.org/10.1007/978-3-319-05789-7_41.

[P6] Yvon Maday, Olga Mula, and Gabriel Turinici. "A priori convergence of the Generalized Empirical Interpolation Method." In: *10th international conference on Sampling Theory and Applications (SampTA 2013)*. 2013, pp. 168–171. DOI: 10.5281/zenodo.54367. URL: https://doi.org/10.5281/zenodo.54367.

[P7] N.E. Stauff, M. Agard, L. Buiron, B. Fontaine, X. Jeanningros, O. Mula, G. Rimpault, and M. Zabiego. "A new methodology for enhanced natural safety GEN-IV SFR core design: application to a carbide-fueled core". In: *Proceedings of the the International Congress on Advances in Nuclear Power Plants*. Vol. 3-4. paper 11162. 2011. URL: https://inis.iaea.org/search/search.aspx?orig_q=RN:44092836.

Popularization

[Pop1] J. Atif, B. Cabot, O. Cappé, O. Mula, and R. Pinot. *Initiative face au virus. Regards croisés sur l'épidemie de Covid-19 apportés par les données sanitaires et de géolocalisation (mars à octobre 2020)*. 2020. URL: https://hal.archives-ouvertes.fr/hal-03084832/document.

[Pop2] J. Atif, O. Cappé, A. Kazakci, Y. Léo, L. Massoulié, and O. Mula. *Feedback on mobility during the Covid-19 epidemic*. 2020. URL: https://www.psl.eu/sites/default/files/PDF/Scientific_initiative_facing_virus.pdf.

PhD Thesis

[T1] O. Mula. "Some contributions towards the parallel simulation of time dependent neutron transport and the integration of observed data in real time". PhD thesis. Université Pierre et Marie Curie-Paris VI, 2014. URL: http://hal.upmc.fr/tel-01081601/document.

1.4 Advising Activities

I have had the great pleasure of being part of the advising team of several undergraduate, master, PhD and post-doctoral fellows:

Postdoctoral fellows:

- 2020–...: **Minh Hieu Do.** (supervised at 100%)
  Topic: Reduced Modeling based on Computational Optimal Transport.
- 2019–20: **Joubine Aghili.** (supervised at 100%)
  Topic: Data assimilation for Inverse Problems and Statistical Learning.
  Next position: Assistant Professor of Applied Mathematics at Strasbourg University.
- 2016–18: **Walid Kheriji** (co-supervised at 70% with Y. Maday).
  Topic: The parareal algorithm and its application to the neutron transport equation.
  Next position: Data Scientist at VEDECOM.

PhD theses:

- 2020–...: **Agustín Somacal** (co-supervised at 50% with A. Cohen, Sorbonne Université)
  Topic: Nonlinear reduced models and machine learning in forward modeling and inverse problems.

- 2017–2021: **Felipe Galarce** (co-supervised at 50% with J.F. Gerbeau and D. Lombardi, INRIA).
  Topic: Reconstruction of blood flows with reduced models and Doppler ultrasound images.
  Next position: Postdoctoral fellow at the Weierstrass Institute for Applied Analysis and Stochastics (WIAS).

- 2015-18: **Helin Gong** (CIFRE thesis co-supervised at 30% with Y. Maday, Sorbonne Université, and in collaboration with EDF).
  Title: Data assimilation with reduced basis and noisy measurements. Application to nuclear reactor cores.
  Next position: Nuclear engineer at the Nuclear Power Institute of China (NPIC).

Master theses and undergraduate internships:

- April-Sept 2019: Master thesis of **Changqing Fu** (co-supervised at 50% with R. Ryder).
  Title: Classification methods with Approximate Bayesian Computation methods.

- March-Sept 2017: Master thesis of **Felipe Galarce** (co-supervised at 30% with J.F. Gerbeau and D. Lombardi).
  Title: Enhancing Hemodynamics Measurements with Mathematical Modeling.

- June-July 2019: Internship of **Lucas Perrin**, first year master student (co-supervised at 50% with D. Gontier).
  Title: Tensor methods for quantum chemistry.

- June-July 2017: Internship of **Thanh Bao Tran**, second year undergraduate student (supervised at 100%). To my deepest sorrow, Thanh suddenly died before finishing his internship.

CEMRACS Research Projects (=5 week research project + subsequent collaboration):

- July-August 2021: Topic: COVID4CAST: Measuring prediction uncertainty for approximate solutions to PDEs, and application to the Covid-19 pandemic.
  Students: Ludovica Saccaro, Giulia Sambataro.
  Co-supervised with A. Roche and R. Ryder (U. Paris Dauphine).

- July-August 2021: Topic: MORPOR: Model Order Reduction of 1D nonlinear transport PDEs in porous media.
  Students: Beatrice Battisti, Tobias Blickhan.
  Co-supervised with G. Enchéry (IFPEN) and D. Lombardi (Inria Paris).

- July-August 2021: Topic: MOCO: State estimation methods involving physical model corrections. Application to neutronics.
  Students: Yonah Conjungo, David Labeurthre (CEA).
  Co-supervised with F. Madiot (CEA) and T. Taddei (Inria Bordeaux).

- July-August 2021: Topic: GreedyPINNS: Solving high-dimensional PDEs with neural networks and greedy algorithms
  Students: Roberta Flenghi, María Fuente Ruiz, .
  Co-supervised with V. Ehrlacher (École des Ponts ParisTech & Inria-Paris).

- July-August 2021: Topic: Pollution: Inverse Problems on Graphs. Application to Pollution in Urban City Areas.
  Students: Matthieu Dolbeault, Agustín Somacal.
  Co-supervised with A. Cohen (Sorbonne Université).

- July-August 2017: Topic: Quantification of Uncertainties in the Vlasov-Poisson equation.
  Students: Joackim Bernier (ENS Rennes), Pierre Gerhard (Univ. Strasbourg), Anna Yurova (Max-Planck Institute for Plasma-Physik).
  Co-supervised with M. Campos-Pinto (Sorbonne Université) and K. Kormann (MPI).

Chapter 2

Forward Problems: numerical schemes for high-dimensional PDEs and nonlinear model order reduction

In this chapter I summarize a selection of works on forward problems:

- Section 2.1 summarizes [A5, A7], two contributions on the numerical analysis of the linear Boltzmann equation for neutron and radiative transfer.
- Section 2.2 summarizes [A8], a contribution in time domain decomposition.
- Section 2.3 summarizes [A6], a contribution in forward reduced modeling.

At the end of each section I outline future research directions.

2.1 Linear Boltzmann equation for neutron and radiative transfer

The transport of non-charged particles such as neutrons or photons plays a key role in a number of scientific and engineering areas. It is, for example, relevant in understanding certain atmospheric processes and it also plays a major role in the field of nuclear engineering for the safety of nuclear reactors and shielding. The evolution of the particles in all these problems can be modelled by the linear Boltzmann equation that we introduce next.

Let Ω be a bounded domain of $\mathbb{R}^d$ with $\mathcal{C}^1$ boundary. Denoting by $f = f(t, x, v)$ the *population density* of particles which are located at position $x \in \Omega$ at time $t \in \mathbb{R}_+$ and travelling with velocity $v \in \mathbb{V} \subset \mathbb{R}^d$, the dynamics of the gas can be described by the linear Boltzmann equation (sometimes called the *neutron or radiation transport equation*)

$$\partial_t f + v \cdot \nabla_x f + \sigma(x, v) f - \int_{\mathbb{V}} \kappa(x, v \cdot v') f(t, x, v') \, \mathrm{d}v' = q, \tag{2.1.1}$$

where q is a source term, and the non-negative functions σ and κ denote the *total cross-section* of the background material and the *scattering kernel* respectively. In the following, we will sometimes refer to the pair (σ, κ) as the *optical parameters*. The above evolution equation must be supplemented by suitable initial data

$$f(0, x, v) = f_{\text{init}}(x, v),$$

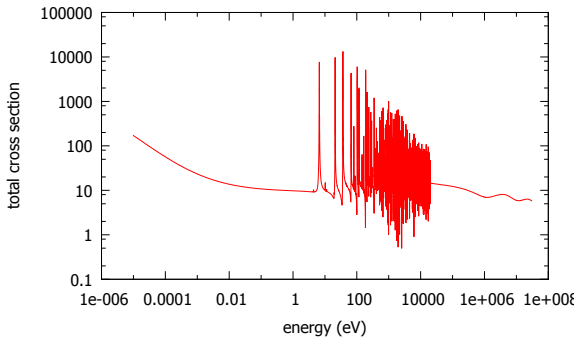


Figure 2.1: Total cross-section σ of Uranium 238 as a function of the energy according to the JEFF 3.1 library [68]. Note the highly oscillatory interval for $E \in [1\text{ eV}; 10^4\text{ eV}]$.

and boundary data on the *incoming phase-space boundary*. For simplicity, we consider *absorption-type* boundary data, i.e.,

$$f(t,x,v) = g \quad \forall t > 0 \text{ and for } (x,v) \in \Gamma_- := \Big\{(x,v) \in \partial\Omega \times \mathbb{R}^d \ : \ n(x)\cdot v < 0\Big\},$$

where $n(x)$ denotes the unit exterior normal to the boundary $\partial\Omega$ at the point x.

The numerical computation of solutions to (2.1.1) is challenging for the following reasons:

- The solution f depends on $(2d+1)$ variables. Hence, the problem is *high-dimensional* enough so that standard schemes become possibly prohibitively inefficient.
- Nontrivial scattering kernels κ give rise to densely populated very large system matrices when using standard discretizations based on localization only.
- These obstructions are aggravated by the fact that solutions exhibit in general only a low degree of regularity, in particular, when dealing with highly concentrated and non-smooth boundary data g. Standard a priori error estimates involving classical isotropic Sobolev regularity scales are therefore not very useful for controlling the accuracy between the exact solution and the computed one.
- On top of the previous, the optical parameters can present high oscillations in space due to the spatial heterogeneity of the materials. They can also oscillate in the energy variable defined as $E = |v|^2/2$, see Figure 2.1 for an illustration.

The above difficulties give rise to very complex discretizations, and it is often tacitly assumed that the numerical output represents the corresponding continuous object reasonably well, without being, however, able to actually quantify output quality in any rigorous sense. Often interest shifts then towards accurately solving the discrete problem which by itself may indeed pose enormous challenges. Instead, the central objective of my collaboration with W. Dahmen and F. Gruber in [A5] was to put forward a new scheme warranting error controlled computation. The main ideas and results are summarized in Section 2.1.1.

Regarding the oscillatory behavior of the optical parameters, this topic has motivated a considerably huge amount of literature in the theory of homogenization (see [79] and references therein).

However, to the best of my knowledge, the existing mathematical theory addresses only high oscillations in the spatial variable and no rigorous results seem to address high oscillations in the energy variable. This point has been treated so far only in the engineering community where the problem is known as energy self-shielding or resonant absorption. In this context, in my collaboration with F. Salvarani and H. Hutridurga we have made a first step to bridge the gap between theory and practice by arriving at some rigorous homogenization results which we have published in [A7].

From the homogenization viewpoint, transport dominated equations such as (2.1.1) are particularly challenging since the structure of the partial differential equation becomes more complex after taking the homogenization limit. This is due to the memory effects induced in the limit that make the dynamics be no longer defined by a semigroup [69, 53, 43]. This, in turn, entails difficulties in the numerical solution of the homogenized equation since the memory effects dramatically increase the computational complexity in terms of the number of degrees of freedom to be used in order to retrieve a certain target accuracy.

2.1.1 An Adaptive Nested Source Term Iteration with a posteriori guarantees

Variational Formulation: For the sake of brevity, we will work with the following simplifications but we emphasize that the development could easily be extended to the general case:

- Stationary: we do not have time dependence.
- Monoenergetic: all the particles have the same velocity modulus $|v|$ so we can work with the direction of propagation of the particles $s = v/|v|$ instead of v, and s takes values in the unit sphere $\mathbb{S}$ of $\mathbb{R}^d$.
- Zero incoming data $g = 0$.

Our starting point equation is thus

$$\begin{aligned} s \cdot \nabla f(x,s) + \sigma(x,s) f(x,s) - \int_{\mathbb{S}} \kappa(x, s' \cdot s) f(x,s') \, \mathrm{d}s' &= q(x,s), \qquad \forall (x,s) \in \Omega \times \mathbb{S}, \\ f &= 0, \qquad \text{on } \Gamma_-. \end{aligned} \tag{2.1.2}$$

with

$$\Gamma_- \coloneqq \{(x,s) \in \partial\Omega \times \mathbb{S} \ : \ n(x) \cdot s < 0\}.$$

In the following, we will also work with the boundary Γ_+ which is defined in the same spirit as Γ_-.

Our approach to derive a numerical scheme with a posteriori error guarantees relies on building stable variational formulations to solve (2.1.2). The trial space U must accommodate solutions of (2.1.1) which are potentially discontinuous. Stability in the present context means that the variational formulation identifies an operator $\mathcal{B}$ as an isomorphism from U onto the dual V' of the test space V. The test and trial Hilbert spaces which we found suitable for our purposes are

$$\begin{aligned} U &\coloneqq L^2(\Omega \times \mathbb{S}) \\ V &\coloneqq H_{0,+}(\Omega \times \mathbb{S}) = \mathrm{clos}_{\|\cdot\|_{H(\Omega\times\mathbb{S})}} \{v \in C^1(\overline{\Omega \times \mathbb{S}}) \ : \ v|_{\Gamma_+} = 0\} \end{aligned} \tag{2.1.3}$$

endowed with norms

$$\begin{aligned} \|u\|_U^2 &\coloneqq \int_{\Omega\times\mathbb{S}} u^2(x,s) \mathrm{d}x\mathrm{d}s, \quad \forall u \in U, \\ \|v\|_V^2 &\coloneqq \|v\|_{L^2(\Omega\times\mathbb{S})}^2 + \int_{\mathbb{S}} \|s \cdot \nabla v\|_{L^2(\Omega)}^2 \mathrm{d}s, \quad \forall v \in V. \end{aligned}$$

On these spaces, we define the bilinear form associated to the transport component of the equation

$$\begin{aligned} a &: U \times V \to \mathbb{R} \\ (u,v) &\mapsto a(u,v) \coloneqq \int_{\Omega\times\mathbb{S}} u(x,s)(\sigma(x,s)v(x,s) - s\cdot\nabla v(x,s))\,\mathrm{d}x\,\mathrm{d}s. \end{aligned}$$

From the bilinear form, we can define the transport operator $\mathcal{T} : U \to V'$ induced by a through

$$(\mathcal{T}u)(v) \coloneqq a(u,v), \quad \forall (u,v) \in U \times V$$

for which we can prove that it is linear and bounded.

We next define the scattering operator

$$\begin{aligned} \mathcal{K} &: L^2(\Omega\times\mathbb{S}) \to L^2(\Omega\times\mathbb{S}) \\ v &\mapsto (\mathcal{K}v)(x,s) \coloneqq \int_{\mathbb{S}} \kappa(x, s\cdot s')v(x,s')\,\mathrm{d}s'. \end{aligned}$$

and define an associated bilinear form

$$\begin{aligned} k &: U \times V \to \mathbb{R} \\ (u,v) &\mapsto k(u,v) \coloneqq \int_{\Omega\times\mathbb{S}} (\mathcal{K}u)(x,s)v(x,s)\,\mathrm{d}x\mathrm{d}s. \end{aligned}$$

The bilinear form associated to our original Boltzmann equation (2.1.2) is the difference between a and k, namely

$$b(u,v) \coloneqq a(u,v) - k(u,v), \quad \forall (u,v) \in U \times V. \tag{2.1.4}$$

The Boltzmann operator $\mathcal{B} : U \to V'$ induced by b is thus

$$(\mathcal{B}u)(v) \coloneqq b(u,v), \quad \forall (u,v) \in U \times V. \tag{2.1.5}$$

A key property to prove the results that follow is *accretivity* of $\mathcal{B}$. In the present context this means that there exists $\alpha > 0$ such that

$$\langle \mathcal{B}u, u\rangle_U \geq \alpha \|u\|_U^2, \quad \forall u \in H_{0,-}(\Omega\times\mathbb{S}) \subset U, \tag{2.1.6}$$

where $H_{0,-}(\Omega\times\mathbb{S})$ is defined similarly as the space $V = H_{0,+}(\Omega\times\mathbb{S})$ but we use Γ_- instead of Γ_+ in (2.1.3). Property (2.1.6) holds true under mild assumptions on the optical parameters which we give in [A5]. We purposefully avoid giving these details in the present document in order to focus on the main ideas without the interference of too many technicalities.

Theorem 2.1.1. *Assume that $\mathcal{B}$ is accretive in the sense that (2.1.6) holds. Given any right-hand side $q \in V'$, the problem*

$$\begin{cases} \text{find } f \in U \text{ such that,} \\ b(f,v) = \langle q, v\rangle_{V',V}, \quad \forall v \in V, \end{cases} \tag{2.1.7}$$

has a unique solution satisfying

$$\|f\|_U \lesssim \|q\|_{V'},$$

with constants depending only on the optical parameters. f is a weak solution to the Boltzmann problem (2.1.2) and the operator $\mathcal{B}$, defined by (2.1.5) is an isomorphism from U onto V'. Furthermore, it holds that

$$\|\mathcal{B}^{-1}\|_{\mathcal{L}(U,U)} \leq \alpha^{-1}. \tag{2.1.8}$$

Iterations in Functional Space: We are now in position to build a numerical scheme based on the above variational formulation of our problem. The main guiding principle underpinning our development is that we *avoid discretization until the last possible moment.* As such, we first formulate a fixed point iteration in the infinite dimensional space U. The idea is to identify an iteration of the form

$$f_{n+1} = f_n + \mathcal{P}(q - \mathcal{B}f_n), \quad n = 0, 1, \dots \tag{2.1.9}$$

where $\mathcal{P} \in \mathcal{L}(V', U)$ is a preconditioner that must be chosen in such a way that

$$\exists\, \rho < 1 \text{ such that } \|f_{n+1} - f\|_U \leq \rho \|f_n - f\|_U, \quad n \in \mathbb{N}. \tag{2.1.10}$$

This holds true if and only if

$$\|\mathrm{id} - \mathcal{P}\mathcal{B}\|_{\mathcal{L}(U,U)} \leq \rho < 1. \tag{2.1.11}$$

Our choice for $\mathcal{P}$ heavily depends on the quantity $\|\mathcal{T}^{-1}\mathcal{K}\|_{\mathcal{L}(U,U)}$, which defines two regimes:

1. Dominating transport: $\|\mathcal{T}^{-1}\mathcal{K}\|_{\mathcal{L}(U,U)} \leq \rho < 1$.
2. Dominating scattering: $\|\mathcal{T}^{-1}\mathcal{K}\|_{\mathcal{L}(U,U)} \geq 1$.

In practice, it is possible to estimate the regime of the problem thanks to the fact that we can find a reasonably tight upper bound ρ for $\|\mathcal{T}^{-1}\mathcal{K}\|_{\mathcal{L}(U,U)}$ with available data as we explain in [A5].

Dominating Transport $\|\mathcal{T}^{-1}\mathcal{K}\|_{\mathcal{L}(U,U)} \leq \rho < 1$: In this case, a suitable choice for the preconditioner is

$$\mathcal{P} \coloneqq \mathcal{T}^{-1}.$$

With this choice, $\|\mathrm{id} - \mathcal{P}\mathcal{B}\|_{\mathcal{L}(U,U)} = \|\mathcal{T}^{-1}\mathcal{K}\|_{\mathcal{L}(U,U)} \leq \rho < 1$ so condition (2.1.11) is satisfied. Iteration (2.1.9) becomes

$$f_{n+1} = f_n + \mathcal{T}^{-1}(q - \mathcal{B}f_n) = \mathcal{T}^{-1}(\mathcal{K}f_n + q), \quad n \in \mathbb{N}_0, \tag{2.1.12}$$

and satisfies (2.1.10), ensuring convergence in U to the solution u of the Boltzmann problem

$$\mathcal{B}f = (\mathcal{T} - \mathcal{K})f = q.$$

In particular, it follows that for any initial guess f_0,

$$\|f - f_n\|_U \leq \rho^n \|f - f_0\|_U.$$

Dominating Scattering $\|\mathcal{T}^{-1}\mathcal{K}\|_{\mathcal{L}(U,U)} \geq 1$: This case is more challenging than the previous. To find a substitute for the preconditioner $\mathcal{P} = \mathcal{T}^{-1}$ of the transport dominated regime, consider for some fixed $r > 0$

$$\mathcal{T}_r \coloneqq \mathcal{T} + r\,\mathrm{id}, \qquad \mathcal{B}_r \coloneqq \mathcal{T}_r - \mathcal{K}$$

and take $\mathcal{P} \coloneqq \mathcal{B}_r^{-1}$. This leads to the iteration

$$f_{n+1} = f_n + (\mathcal{T}_r - \mathcal{K})^{-1}(q - (\mathcal{T} - \mathcal{K})f_n) = r\,\mathcal{B}_r^{-1}\big(f_n + r^{-1}q\big), \quad n \in \mathbb{N}_0, \tag{2.1.13}$$

where we have used that

$$(\mathcal{T}_r - \mathcal{K})^{-1}(\mathcal{T} - \mathcal{K}) = (\mathcal{T}_r - \mathcal{K})^{-1}(\mathcal{T}_r - \mathcal{K} - r\mathrm{id}) = \mathrm{id} - r(\mathcal{T}_r - \mathcal{K})^{-1} = \mathrm{id} - r\mathcal{B}_r^{-1}.$$

Thus, to ensure convergence we need $\|r\mathcal{B}_r^{-1}\|_{\mathcal{L}(U,U)}$ to be a contraction. Note that this is satisfied for any $r > 0$ since we have that $(\mathcal{B}_r v, v) \geq \alpha + r$, which by equation (2.1.8) of Theorem 2.1.1 gives

$$\|r\mathcal{B}_r^{-1}\|_{\mathcal{L}(U,U)} \leq \frac{r}{r+\alpha} < 1.$$

So (2.1.13) converges in $U = L_2(\Omega \times \mathbb{S})$ to the true solution f with the error reduction rate $r/(r+\alpha)$ for any fixed $r > 0$.

At every iteration n, we thus have to solve

$$\mathcal{B}_r \tilde{f}_n = f_n + r^{-1}q.$$

To make this step doable in practice, we choose the parameter r in such a way that the operator $\mathcal{B}_r$ is *transport dominated*. In other words, we fix r to a value r^* so that we have simultaneously

$$\|r^*\mathcal{B}_{r^*}^{-1}\|_{\mathcal{L}(U,U)} \leq \rho^* \quad \text{and} \quad \|\mathcal{T}_{r^*}^{-1}\mathcal{K}\|_{\mathcal{L}(U,U)} \leq \rho^* \quad \text{for some } \rho^* < 1. \tag{2.1.14}$$

As we explain in our paper, it is possible to find such r^* and its value depends on the optical parameters. This way, we can resort to the fixed-point scheme (2.1.12) of the transport dominated case. Therefore, in the case of dominating scattering, we have to solve a nested iterative scheme.

Perturbed iterations: The practical realization of the scheme boils down to two tasks:

(T1) *Formulate a perturbed version of algorithms* (2.1.12) *and* (2.1.13) *with suitable error tolerances* η_n *that still guarantee convergence to the exact continuous solution.*

For this task, it will be convenient to use the following notational convention: Given an operator $\mathcal{G} \in \mathcal{L}(U, Y)$, we denote for any $\eta > 0$ by $[\mathcal{G}, u; \eta]$ an element in Y satisfying $\|\mathcal{G}u - [\mathcal{G}, u; \eta]\|_Y \leq \eta$. Specifically, for our purposes we require a routine to approximately apply the kernel $\mathcal{K}$, that is,

$$[\mathcal{K}, u; \eta] \to z_\eta \quad \text{such that} \quad \|\mathcal{K}u - z_\eta\|_{V'} \leq \eta. \tag{2.1.15}$$

Likewise the source $q \in V'$ is generally not given exactly and has to be approximated

$$[q; \eta] \to q_\eta \quad \text{such that} \quad \|q - q_\eta\|_{V'} \leq \eta. \tag{2.1.16}$$

The approximation $[q; \eta]$ of q depends on how the data are given. Finally, given a right hand side $g \in V'$, we have to provide a transport solver

$$[\mathcal{T}^{-1}, g; \eta] \to u_\eta \quad \text{such that} \quad \|u_\eta - \mathcal{T}^{-1}g\|_U \leq \eta, \tag{2.1.17}$$

where, as before, $\mathcal{T}$ is viewed as a mapping from U onto V'.

(T2) *Specify how to realize the above routines in* (2.1.15), (2.1.16), *and* (2.1.17).

We first concentrate on task (T1) and assume for the moment that the routines (2.1.15), (2.1.16), and (2.1.17) are available.

- **Transport dominated regime:** An approximate realization of the scheme (2.1.12) is

$$\bar{f}_{n+1} = [\mathcal{T}^{-1}, [\mathcal{K}, \bar{f}_n; \eta_{\mathcal{K}}] + [q; \eta_q]; \eta_{\mathcal{T}}], \quad n \geq 0. \tag{2.1.18}$$

  In the following we take for simplicity $f_0 = 0$. Any other choice for f_0 that exploits additional information would, of course, be possible. We choose the individual tolerances proportional to

$$\eta_n = (1+n)^{-\beta}\rho^n, \tag{2.1.19}$$

  for some fixed $\beta > 1$. Specifically, we set

$$\eta_{\mathcal{K}} \coloneqq c_1\eta_n, \quad \eta_q \coloneqq c_2\eta_n, \quad \eta_{\mathcal{T}} \coloneqq c_3\eta_n,$$

  where the parameters $c_1, c_2, c_3 \geq 0$ satisfy

$$C_{\mathcal{T}}(c_1 + c_2) + c_3 \leq 1,$$

  and where $C_{\mathcal{T}}$ is an upper bound of $\|\mathcal{T}^{-1}\|_{\mathcal{L}(V',U)}$ (a computable upper bound is given in our paper [A5]). We call this algorithm *Adaptive Source Term Iteration* (ASTI). We can prove that it converges to the exact solution f of the infinite dimensional variational formulation (2.1.7) (see Theorem 2.1.2 below). We write ASTI$[\mathcal{T}, \mathcal{K}, q; \varepsilon]$ to denote the routine that computes an approximate solution f_ε such that $\|f - f_\varepsilon\|_U \leq \varepsilon$ using the scheme (2.1.18).

- **Dominating Scattering:** We fix r to a value r^* so that we satisfy (2.1.14). The approximate realization of the scheme (2.1.13) takes the form

$$\bar{f}_{n+1} = [r^*\mathcal{B}_{r^*}^{-1}, \bar{f}_n + [(r^*)^{-1}q; \eta_n^*]; \eta_n^*], \quad n \in \mathbb{N}_0,$$

  where the tolerances η_n^* are chosen as in (2.1.19) but using ρ^*. To compute the above problem, we use ASTI, namely

$$\bar{f}_{n+1} = [r^*\mathcal{B}_{r^*}^{-1}, \bar{f}_n + [(r^*)^{-1}q; \eta_n^*]; \eta_n^*] = r^*\text{ASTI}[\mathcal{T}_{r^*}, \mathcal{K}, \bar{f}_n + [(r^*)^{-1}q; \eta_n^*], \eta_n^*]. \tag{2.1.20}$$

  We write N-ASTI$[\mathcal{T}, \mathcal{K}, q; \varepsilon]$ to denote the routine that computes an approximate solution f_ε such that $\|f - f_\varepsilon\|_U \leq \varepsilon$ using the scheme (2.1.20). The notation N-ASTI stands for Nested ASTI.

The following Theorem guarantees convergence of the nested iterative scheme N-ASTI (as a corollary ASTI converges in the transport dominated regime).

Theorem 2.1.2. *For any target accuracy $\varepsilon > 0$, the iterative scheme (2.1.20) converges to the solution $u \in U$ of the variational problem (2.1.7). The output*

$$f_\varepsilon \coloneqq \text{N-ASTI}[\mathcal{T}, \mathcal{K}, q; \varepsilon]$$

satisfies

$$\|f - f_\varepsilon\|_U \leq \varepsilon,$$

where f is the exact solution of the variational formulation (2.1.7).

Format of the numerical output f_ε: To implement the above the scheme in practice, we use adaptive discretrizations involving piecewise polynomials for the angular variable s, and Discontinuous Petrov–Galerkin (DPG) finite elements for the space variable x. The discretizations are subordinate to partitions $\mathcal{P}_{\mathbb{S}}$ and $\mathcal{P}_\Omega$ of the domains $\mathbb{S}$ and Ω. These partitions are *dynamically refined across the iterations*. In our practical implementation, the partition $\mathcal{P}_\Omega$ also depends on the given direction s. For each cell T of the angular partition, we have used a different spatial partition $\mathcal{P}_\Omega^T$. This degree of flexibility makes the practical implementation a non-trivial task. It is however necessary in order to treat heterogenous transport phenomena with a number of degrees of freedom which is as economic as possible, and also in order to guarantee a final certification for the error.

Let us assume that we use polynomials of degree m and n for space and direction respectively. For every cell T of the angular partition $\mathcal{P}_{\mathbb{S}}$, we denote by $\{\phi_{T,i}\}_{i=1}^n$ an associated polynomial basis of $\mathbb{P}_m(T)$. For a fixed cell $T \in \mathcal{P}_{\mathbb{S}}$ and for every cell K of the spatial partition $\mathcal{P}_\Omega^T$, we denote by $\{\varphi_{K,j}\}_{j=1}^m$ an associated polynomial basis of $\mathbb{P}_m(K)$. With this notation, the numerical approximation of f is of the form

$$\bar{f}(x,s) = \sum_{T\in\mathcal{P}_{\mathbb{S}}} \sum_{K\in\mathcal{P}_\Omega^T} \sum_{i=1}^n \sum_{j=1}^m c_{T,K,i,j}\varphi_{K,j}(x)\phi_{T,i}(s), \quad \forall (x,s) \in \Omega\times\mathbb{S},$$

where the $c_{T,K,i,j} \in \mathbb{R}$ are the coefficients of the polynomial expansion. Note that in the above formula we have implicitly extended the support of the basis functions to the full domain so that $\phi_{T,i}(s) = 0$ if $s \notin T$, and $\varphi_{K,j}(x) = 0$ if $x \notin K$.

Approximation of the scattering kernel $\mathcal{K}\colon U \to U$: For simplicity, let us assume that the scattering coefficient κ does not depend on x. In this case, we need to approximate functions of the form

$$(\mathcal{K}\bar{f})(x,s) = \sum_{T\in\mathcal{P}_{\mathbb{S}}} \sum_{K\in\mathcal{P}_\Omega^T} \sum_{i=1}^n \sum_{j=1}^m c_{T,K,i,j}\varphi_{K,j}(x)\mathcal{K}\phi_{T,i}(s).$$

The simplest realization of $[\mathcal{K},\cdot;\cdot]$ rests on computing η-accurate approximations $w_{T,i} = [\mathcal{K},\phi_{T,i};\eta]$. In practice, to efficiently compute the $w_{T,i}$ we need to have a sparse representation of the scattering operator $\mathcal{K}$. The representation may depend on the nature of the scattering operator as we illustrate for the example of the Henyey–Greenstein kernel for spatial dimension $d = 2$, which reads

$$\kappa(s,s') \propto \frac{1}{\|s-\gamma s'\|_2^2} \tag{2.1.21}$$

for a parameter $\gamma \in]0,1[$ which describes the physical nature of the scattering (if $\gamma \to 1$, the scattering tends to be forward peaked). We can first consider a Hilbert–Schmidt decomposition in which we write

$$k(s,s') = \sum_{i=1}^\infty \sigma_i g_i(s) g_i(s'),$$

where $\sigma_i \geq 0$ are singular values of $\mathcal{K}$ associated to the family of eigenfunctions $\{g_i(s)\}_i$ which forms an orthonormal basis of $L^2(\mathbb{S})$. As Figure 2.2 illustrates, the singular values decay fast only when γ is not close to 1. As a result, we can work with low-rank representations of the kernel (by truncating the above expansion) and the approximation accuracy decays fast with the number of terms. On other hand, when $\gamma \to 1$, we prove in [A5] that $\mathcal{K}$ becomes sparse in a wavelet representation of

the kernel, which allows to leverage wavelet compression techniques. By expressing κ with Alpert wavelets $\{\psi_\lambda\}_{\lambda\in\Lambda}$,

$$\kappa(s,s') = \sum_{(\lambda,\lambda')\in\Lambda\times\Lambda} k_{\lambda,\lambda'}\psi_\lambda(s)\psi_{\lambda'}(s), \quad k_{\lambda,\lambda'} \coloneqq \langle \kappa, \psi_\lambda \otimes \psi_{\lambda'}\rangle_{L^2(\mathbb{S}\times\mathbb{S})},$$

the infinite dimensional matrix $\{k_{\lambda,\lambda'}\}_{(\lambda,\lambda')\in\Lambda\times\Lambda}$ can be truncated to a sparse finite-dimensional matrix which delivers high accuracy. We illustrate this idea in Figure 2.3.

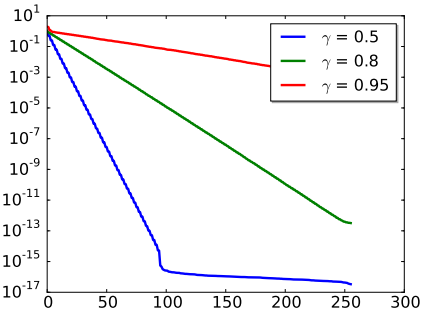


Figure 2.2: Decay of the eigenvalues of the Hilbert–Schmidt decomposition of the Henyey–Greenstein kernel.

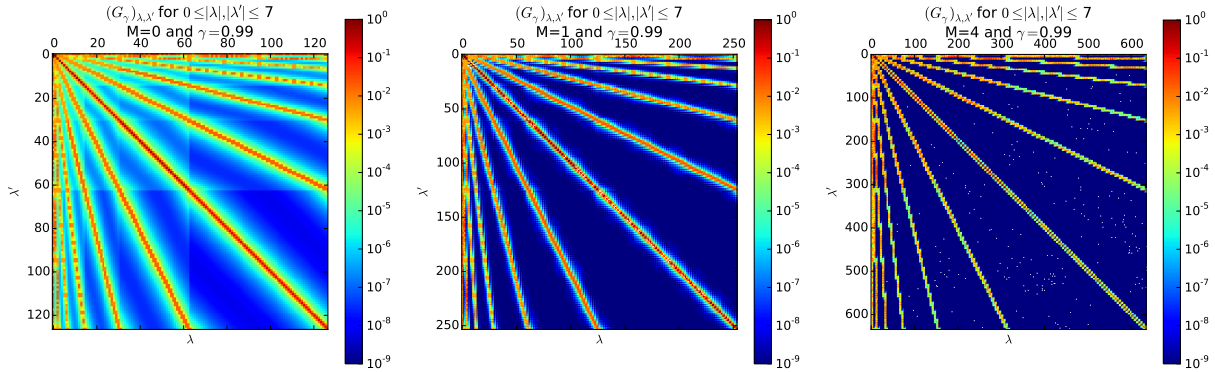


Figure 2.3: Alpert wavelet representation $(k_{\lambda,\lambda'})_{(\lambda,\lambda')\in\Lambda\times\Lambda}$ of the Henyey-Greenstein kernel with degrees $M = 0$, 1 and 4. Here $\gamma = 0.99$. Note the characteristic "finger" structure of a compressible operator.

Approximation of the transport operator $\mathcal{T}\colon U \to V'$: The numerical realization of the routine $[\mathcal{T}^{-1}, \cdot; \cdot]$ is based on solving in the space variable *fiber problems* of the form

$$\mathcal{T}_s f_{n+1} \coloneqq s\cdot\nabla f_{n+1} + \sigma(s) f_{n+1} = \int_{\mathbb{S}} \kappa(\cdot, s, s')\bar{f}_n(\cdot, s')\,\mathrm{d}s' + q, \tag{2.1.22}$$

for properly selected parameters $s \in \mathbb{S}$. Achieving a given target accuracy depends on solving each fiber problem with sufficient accuracy and also on solving sufficiently many of them.

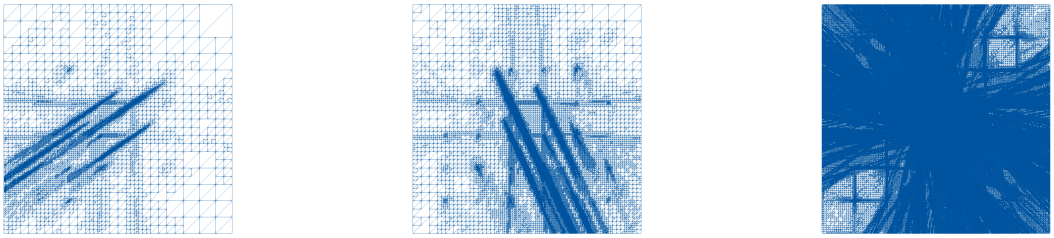


Figure 2.4: Adaptive meshes for fiber transport solutions with respect to two different directions as well as the merged mesh at iteration step 10.

In view of the inherently low regularity of the transport solutions (especially in the presence of rough boundary and source data) we opt for employing an adaptive Discontinuous Petrov–Galerkin (DPG) scheme for each fiber problem (2.1.22). We follow the approach developped and analyzed in [20, 9]. A detailed presentation would exceed the purposes of the current document but the main idea is that this type of discretization makes errors in the U-norm become equivalent to residuals in the V'-norm. The constants of equivalence are uniform in the size of the mesh discretization. This point is crucial to guarantee that the quality of the a posteriori error bounds which we use to estimate the error at each N-ASTI iteration step is not degraded as we dynamically refine.

Since the algorithm is adaptive, the spatial solutions $f_n(\cdot, s)$ are defined in different spatial partitions for each direction $s \in \mathbb{S}$. As a result, one needs to find the union of the meshes when it comes to applying the scattering kernel to f_n. Figure 2.4 shows different meshes for fiber transport solutions with respect to two different directions as well as the merged mesh at iteration step 10 of our numerical example.

The computation of the merged grid is delicate to implement and it leads to a grid which eventually involves a high number of degrees of freedom in final iterations where the accuracy η_n becomes tight. This issue seems inevitable given the nature of the problem but, on the positive side, remark that the merged grid will nevertheless involves much less cells than the underlying uniform mesh which is the one that one would use in a naive approach. In addition to this, note that our approach adresses the most critical issue regarding computational complexity since the bulk of computation lies in the approximate inversions of fiber transport problems (2.1.22), and we handle this operation with an economic number of degrees of freedom. It is therefore of primary importance to keep the size of each fiber transport problem as small as possible, and this is achieved thanks to our adaptive DPG scheme.

Numerical Illustration: We consider the radiative transfer problem (2.1.2) on the unit square domain $\Omega = [0,1]^2$. The spatial structure of the source term q and absorption coefficient σ is depicted in Figure 2.5. More precisely, we take $q = 0$ in the white and gray areas whereas $q = 1$ in the black area. Similarly, we set $\sigma = 10$ in the gray areas and $\sigma = 2$ everywhere else.

We consider a Henyey–Greenstein scatering kernel with $\gamma = 0.5$ (see formula (2.1.21)). Since the singular values of the Hilbert–Schmidt decompositions decay rapidly (see Figure 2.2), we work with adaptive low-rank representations of $\mathcal{K}$ based on it. We present results with Alpert wavelets of degree 2.

We set $\varepsilon = 1.1 \cdot 5.10^{-3}$ as the final target accuracy. The problem is of transport-dominated nature ($\rho \leq 1$) so we can solve it with the ASTI algorithm. The algorithm requires to estimate certain quantities and we refer to [A5] for these details.

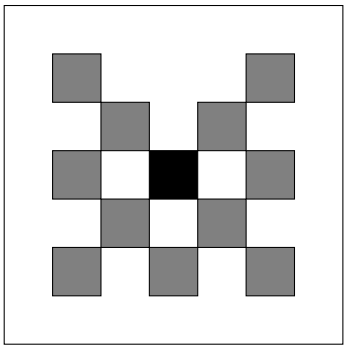


Figure 2.5: Geometry of the checkerboard benchmark.

Figure 2.6, displays the convergence history and degrees of freedom for the above choice of parameters. The left plot gives an approximation error of the scattering application $||\mathcal{K}(\bar{f}_n) - [\mathcal{K}, \bar{f}_n; c_1\eta_n]||_{L_2(\Omega\times\mathbb{S})}$ (dark blue curve), the a posteriori error of the transport solves $||f_n - \bar{f}_n||_{L_2(\Omega\times\mathbb{S})}$ (light blue curve), and a bound for the global error $||f - \bar{f}_n||_{L_2(\Omega\times\mathbb{S})}$ (purple curve). By definition (2.1.19) of the tolerances η_n, the interior solution accuracies need to be somewhat finer which explains the gradual divergence between the global error bound and the interior error tolerances. To avoid this would require total a posteriori bounds based on the full Boltzmann bilinear form b (defined in (2.1.4)) in combination with coarsening strategies, which will be the subject of future work. The shaded blue regions in the right plot indicate statistics about the number of degrees of freedom that are associated for each selected angular direction.

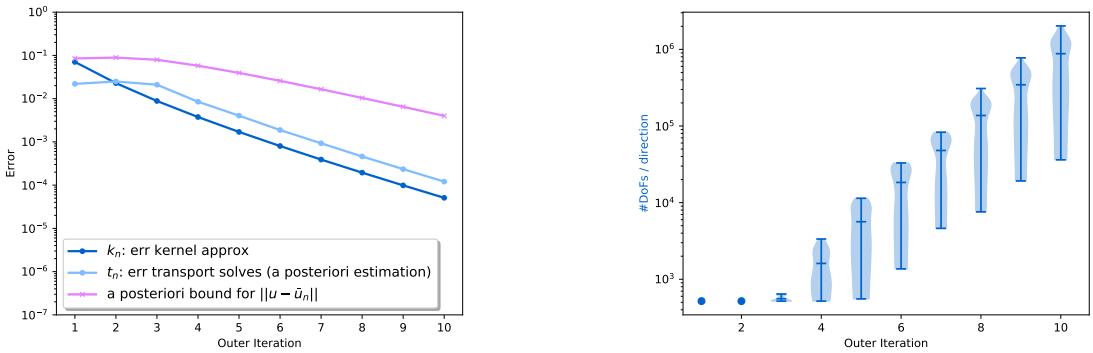
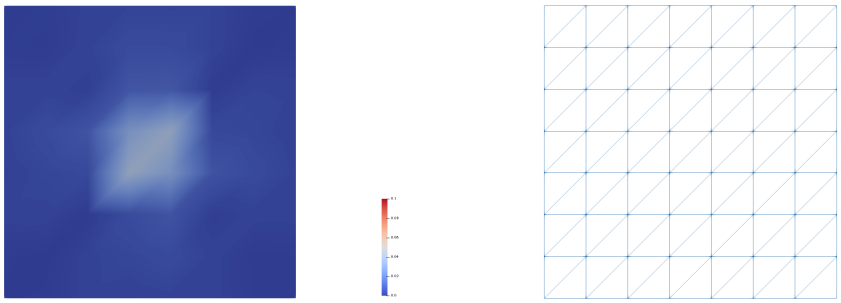


Figure 2.6: Convergence and number of degrees of freedom (DoFs).

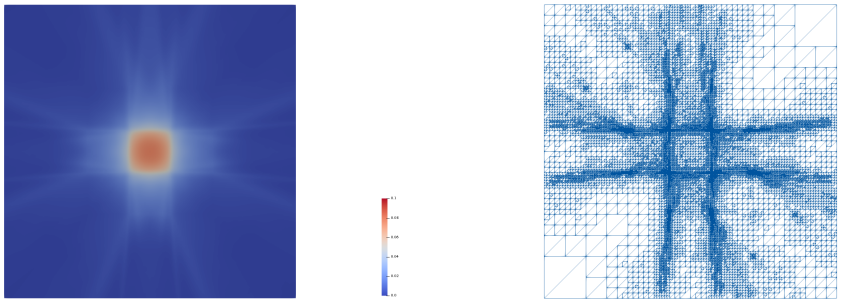
Figure 2.8 shows solutions $\bar{f}_n(\cdot, s)$ with their corresponding grids for the final iterate once the accuracy ε has been reached. Finally, Figure 2.7 shows the final averaged densities $\int_{\mathbb{S}} \bar{f}_n(\cdot, s)\,\mathrm{d}s$. They are computed on the merged grids.

We note that no special *structure preserving* measures had to be imposed on the numerical schemes to produce physically meaningful results.

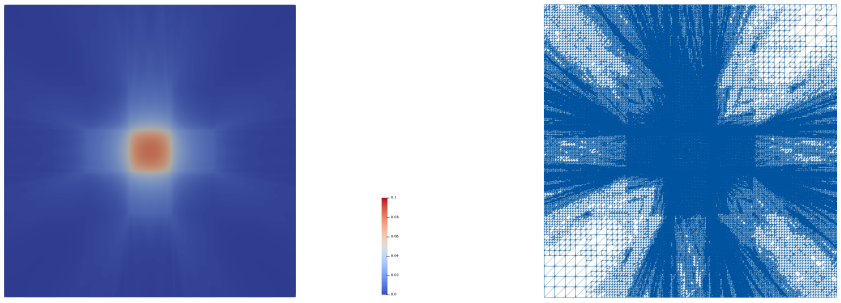
Code: The implementation of the ASTI and N-ASTI algorithms can be done in a very modular fashion thanks to the fact that the main building blocks are very clearly defined. Despite this, the



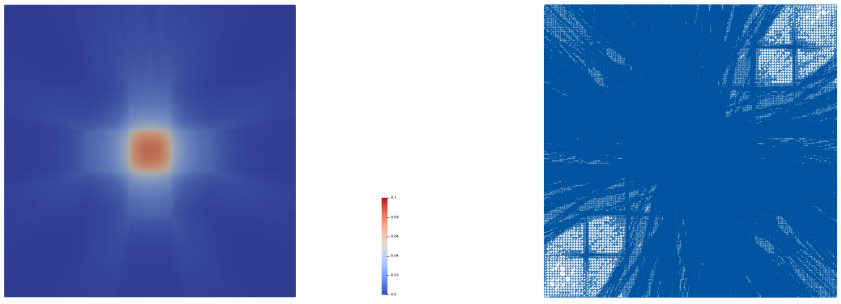
(a) integrated solution and grid for iteration step 2.



(b) integrated solution and grid for iteration step 6.



(c) integrated solution and grid for iteration step 8.



(d) integrated solution and grid for iteration step 10.

Figure 2.7: Integrated solutions $\int_{\mathbb{S}} f_n(\cdot, s)\,\mathrm{d}s$ and corresponding merged grids.

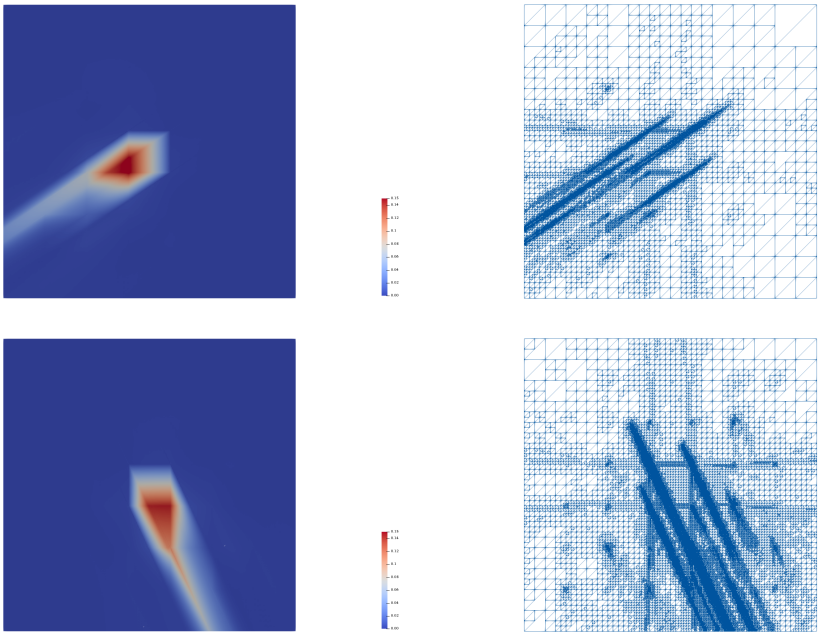


Figure 2.8: Solutions $\bar{f}_n(\cdot, s)$ for different directions s in final outer iterate.

implementation is very challenging due to the numerous different elements to put into place and assemble together: the efficient computation of the kernel, the DPG adaptive solver for the fiber problems, and the management of the grid merging. Felix Gruber was the person who contributed most significantly to this part of the work in the framework of his PhD thesis [16]. On my side, I laid the foundations for the computation of the scattering kernel and the computation of the residual-based a posteriori estimators. My contribution was more focused on the construction of the N-ASTI scheme, and the estimation of technical quantities needed in the algorithm, and upon which I have not insisted a lot in the previous summary.

The implementation makes use of Dune-DPG 0.4.2, a C++ based library which Felix Gruber, Angela Klewinghaus and I built upon the multi-purpose finite element package DUNE. We published an introductive summary of the code in [A11], and details of the Dune-DPG library can be found in [27]. The code to reproduce the numerical part of the present work is available online at:

https://gitlab.dune-project.org/felix.gruber/dune-dpg

2.1.2 Homogenization in the energy variable

In this section, we discuss the oscillatory behavior that the optical parameters often present in the energy variable in realistic scenarios. I summarize the homogenization result on the energy variable that Harsha Hutridurga, Francesco Salvarani and I published in [A7]. To do so, we come back to the general Boltzmann equation (2.1.1), which we recall here:

$$\partial_t f + v \cdot \nabla_x f + \sigma(x, v) f - \int_{\mathbb{V}} \kappa(x, v \cdot v') f(t, x, v') \, \mathrm{d}v' = q.$$

The velocity variable v can be expressed via the couple (ω, E), where $\omega = v/|v|$ is the trajectory angle of the neutron and $E = m|v|^2/2$ is the kinetic energy, m being the mass of the neutron.

Assuming that E ranges in $[E_{\min}, E_{\max}]$, the same equation can equivalently be written in terms of the *neutron flux*

$$\varphi(t,x,\omega,E) = \varphi(t,x,v) := |v| f(t,x,v),$$

which satisfies the equation

$$\sqrt{\frac{m}{2E}}\,\partial_t\varphi + \omega\cdot\nabla_x\varphi + \sigma\left(x,\omega,E\right)\varphi - \int_{E_{\min}}^{E_{\max}}\int_{|\omega'|=1}\kappa\left(x,\omega\cdot\omega',E,E'\right)\varphi(x,\omega',E')\mathrm{d}\omega'\,\mathrm{d}E' = 0,$$

where the optical parameters of the linear Boltzmann equation are appropriately redefined. The above evolution equation is supplemented by suitable initial data, i.e., $\varphi(0,x,\omega,E) = \varphi_{\text{in}}(x,\omega,E)$, and zero boundary data on the *incoming phase-space boundary*.

Real experiments reveal strong oscillations in σ as a function of E when the neutrons interact with relevant materials like, for example, Uranium 238 (see figure 2.1). A similar behaviour is also observed for the scattering kernel κ. These facts motivate us to study the multi-scale linear Boltzmann equation

$$\sqrt{\frac{m}{2E}}\,\partial_t\varphi^\varepsilon + \omega\cdot\nabla_x\varphi^\varepsilon + \sigma^\varepsilon\left(x,\omega,E\right)\varphi^\varepsilon - \int_{E_{\min}}^{E_{\max}}\int_{|\omega'|=1}\kappa^\varepsilon\left(x,\omega\cdot\omega',E,E'\right)\varphi^\varepsilon(x,\omega',E')\mathrm{d}\omega'\,\mathrm{d}E' = 0,$$

where $0<\varepsilon\ll 1$ is a small parameter and

$$\sigma^\varepsilon(x,\omega,E) = \sigma\left(x,\omega,E,\frac{E}{\varepsilon}\right); \qquad \kappa^\varepsilon(x,\omega\cdot\omega',E,E') = \kappa\left(x,\omega\cdot\omega',E,E',\frac{E'}{\varepsilon}\right),$$

where $\sigma(x,\omega,E,y)$ and $\kappa\left(x,\omega\cdot\omega',E,E',y'\right)$ are both assumed to be periodic in the y and y' variables respectively. The equation is complemented with zero incoming flux condition on the phase-space boundary and an initial condition φ_{in} which we assume to be in $\mathrm{L}^2(\Omega\times\mathbb{S}^{d-1}\times(E_{\min},E_{\max}))$.

In addition to the above hypotheses, we also assume that there exists $\alpha>0$ such that for all $\varepsilon>0$,

$$\sigma^\varepsilon(x,\omega,E) - \bar{\kappa}^\varepsilon(x,\omega,E) \geq \alpha \quad \text{and} \quad \sigma^\varepsilon(x,\omega,E) - \tilde{\kappa}^\varepsilon(x,\omega,E) \geq \alpha,$$

where

$$\begin{cases} \bar{\kappa}^\varepsilon(x,\omega,E) & := \displaystyle\int_{E_{\min}}^{E_{\max}}\int_{\mathbb{S}^{d-1}}\kappa^\varepsilon(x,\omega\cdot\omega',E,E')\ \mathrm{d}\omega'\,\mathrm{dE}' \\ \tilde{\kappa}^\varepsilon(x,\omega,E) & := \displaystyle\int_{E_{\min}}^{E_{\max}}\int_{\mathbb{S}^{d-1}}\kappa^\varepsilon(x,\omega\cdot\omega',E',E)\ \mathrm{d}\omega'\,\mathrm{dE}'. \end{cases}$$

From a physical point of view, these assumptions mean that we place ourselves in the so-called subcritical regime where absorption phenomena dominate scattering.

Our main contribution is the homogenization result given in Theorem 2.1.3, where we derive a homogenized equation for the neutron transport problem when the optical parameters oscillate periodically in the energy variable. The result is derived employing the theory of two-scale convergence. For technical reasons, we worked with scattering kernels κ^ε exhibiting separation in the E and E' variables as follows:

$$\kappa^\varepsilon(x,\omega\cdot\omega',E,E') := c_1(x,\omega\cdot\omega',E)c_2\left(x,\omega\cdot\omega',E',\frac{E'}{\varepsilon}\right)$$

with $c_2(x, \omega\cdot\omega', E', y')$ being periodic in the y' variable. A result without this assumption on κ^ε requires further investigation as it is not apparent whether one can derive closed form homogenized equations in the latter case.

The theorem uses the notation $\mathrm{C}_{\mathrm{per}}(\mathrm{Y})$ to denote continuous functions on $\mathbb{R}^d$ which are Y-periodic, and for any given $g \in \mathrm{L}^\infty(\mathrm{Y})$,

$$\mathcal{L}_g\, v \coloneqq gv - \langle gv\rangle \quad \forall v \in \mathrm{L}^2_{\mathrm{per}}(\mathrm{Y}).$$

where $\langle gv\rangle \coloneqq \int_Y g(y)v(y)\,\mathrm{d}y$.

Theorem 2.1.3. *Let $\varphi^\varepsilon = \varphi^\varepsilon(t,x,\omega,E)$ be the solution of the equation*

$$\begin{cases} \partial_t\varphi^\varepsilon + \sqrt{E}\,\omega\cdot\nabla_x\varphi^\varepsilon + \sigma^\varepsilon(\omega,E)\,\varphi^\varepsilon - \displaystyle\int_{E_{\min}}^{E_{\max}}\int_{|\omega'|=1} \kappa^\varepsilon\left(\omega\cdot\omega', E, E'\right)\varphi^\varepsilon(\omega',E')\,\mathrm{d}\omega'\,\mathrm{d}E' = 0 \\ \varphi^\varepsilon(0,x,\omega,E) = \varphi^\varepsilon_{\mathrm{in}}(x,\omega,E), \\ \varphi^\varepsilon(t,x,\omega,E) = 0 \quad \forall t>0 \quad \text{and for } (x,\omega)\in\Gamma_- \coloneqq \left\{(x,\omega)\in\partial\Omega\times\mathbb{S}^{d-1} \;:\; \mathbf{n}(x)\cdot\omega<0\right\}. \end{cases}$$

where the coefficients and the data are of the form

$$\begin{aligned} \sigma^\varepsilon(\omega,E) &\coloneqq \sqrt{E}\,\sigma\left(\omega,E,\frac{E}{\varepsilon}\right) \quad \text{with} \quad \sigma(\omega,E,y)\in\mathrm{L}^\infty(\mathbb{V};\mathrm{C}_{\mathrm{per}}(0,1)) \\ \varphi^\varepsilon_{\mathrm{in}}(x,\omega,E) &\coloneqq \varphi_{\mathrm{in}}\left(x,\omega,E,\frac{E}{\varepsilon}\right) \quad \text{with} \quad \varphi_{\mathrm{in}}(x,\omega,E,y)\in\mathrm{L}^2(\Omega\times\mathbb{V};\mathrm{C}_{\mathrm{per}}(0,1)) \\ \kappa^\varepsilon(\omega\cdot\omega',E,E') &\coloneqq \sqrt{E}\,\kappa_1(\omega\cdot\omega',E)\kappa_2\left(\omega\cdot\omega',E',\frac{E'}{\varepsilon}\right) \text{ with } \kappa_1(\eta,E)\in\mathrm{L}^\infty([-1,1]\times[E_{\min},E_{\max}]) \\ &\qquad \text{and } \kappa_2\left(\eta,E',y'\right)\in\mathrm{L}^\infty([-1,1]\times[E_{\min},E_{\max}];\mathrm{C}_{\mathrm{per}}(0,1)). \end{aligned}$$

Then,

$$\varphi^\varepsilon \rightharpoonup \varphi_{\mathrm{hom}} \quad \text{weakly in } \mathrm{L}^2((0,T)\times\Omega\times\mathbb{V})$$

and φ_{hom} satisfies the following partial integro-differential equation

$$\partial_t\varphi_{\mathrm{hom}} + \sqrt{E}\,\omega\cdot\nabla_x\varphi_{\mathrm{hom}} + \sqrt{E}\,\langle\sigma\rangle\varphi_{\mathrm{hom}} - \int_{E_{\min}}^{E_{\max}}\int_{\mathbb{S}^{d-1}}\sqrt{E}\,\kappa_1(\omega\cdot\omega',E)\int_0^1\kappa_2(\omega\cdot\omega',E',y')\varphi_{\mathrm{hom}}\mathrm{d}y'\mathrm{d}\omega'\mathrm{d}E' =$$

$$\int_{E_{\min}}^{E_{\max}}\int_{\mathbb{S}^{d-1}}\sqrt{E}\,\kappa_1(\omega\cdot\omega',E)\int_0^1\kappa_2(\omega\cdot\omega',E',y')\left[e^{-t\sqrt{E'}\mathcal{L}_\sigma}\mathcal{L}_1\varphi_{\mathrm{in}} - \int_0^t e^{-(t-s)\sqrt{E'}\mathcal{L}_\sigma}\sqrt{E'}\mathcal{L}_1\sigma(\omega',E',y')\varphi_{\mathrm{hom}}\,\mathrm{d}s\right]\mathrm{d}y'\mathrm{d}\omega'\mathrm{d}E'$$

$$-\sqrt{E}\int_0^1\sigma(\omega,E,y)\,\left[e^{-t\sqrt{E}\mathcal{L}_\sigma}\mathcal{L}_1\varphi_{\mathrm{in}} - \int_0^t e^{-(t-s)\sqrt{E}\mathcal{L}_\sigma}\sqrt{E}\mathcal{L}_1\sigma(\omega,E,y)\varphi_{\mathrm{hom}}\,\mathrm{d}s\right]\mathrm{d}y,$$

with initial condition

$$\varphi_{\mathrm{hom}}(0,x,\omega,E) = \langle\varphi_{\mathrm{in}}(x,\omega,E,\cdot)\rangle$$

and zero absorption condition at the in-flux phase-space boundary.

The homogenized equation is integro-differential and presents a memory term which makes the dynamics be no longer defined by a semigroup. This, in turn, entails difficulties in the numerical solution of the homogenized equation since the memory effects dramatically increase the computational complexity in terms of the number of degrees of freedom to be used in order to retrieve a certain target accuracy.

To derive the results, we base our strategy on the method of characteristics, and hence we first derive a homogenization result for an associated ordinary differential equation. In there, we also show that this result is in agreement with previous works on memory effects by Tartar [89], [69, chapter 35]. An interesting result in its own right is that our technique gives an explicit expression of the memory kernel that, in the situation studied by Tartar, is equal to the implicit expression given in [89], [69, chapter 35]. For the sake of brevity, we will not give further details on this results in the present manuscript and refer to [A7].

To conclude this part, let us compare our approach to some standard methods from the nuclear engineering community for treating self-shielding phenomena. To the best of our knowledge, the most widespread technique is a two-stage method originially proposed in [94] by M. Livolant and F. Jeanpierre (we refer to [63, Chapters 8 and 15] for an introductory overview). It consists in finding first the averaged optical parameters which are then plugged into a multigroup version of equation (2.1.1) to compute the behavior of the flux on large and geometrically complex domains such as nuclear reactors. The pre-computation of the averaged parameters is done on a cell problem involving a much simpler spatial geometry and simplified physics. It is nevertheless carefully designed with elaborate physical considerations in a way to keep as much consistency as possible with respect to the original problem. Note that, while this approach implicitly assumes that the homogenized equation is of the same nature as the original Boltzmann problem (2.1.1), our starting point is fundamentally different in the sense that we do not postulate any final form of the limit equation. Our goal is precisely to discover its form from the only assumption that the optical parameters oscillate in energy. As a result, our methodology and conclusions are different from the ones discussed in [94] and do not involve a pre-computation on a cell problem. Another approach, also based on averaging the optical parameters, is the so-called multi-band method (see [90]), where, like in the previous method, the limit equation is assumed to be a Boltzmann equation. Finally, a more recent approach based on averaging arguments taken from results of homogenization of pure transport equations has recently been proposed in [25]. The initial problem there is a time-independent Boltzmann source problem with no oscillations in the scattering kernel.

2.1.3 Research Perspectives

There are several extensions and new research directions which I would like to pursue in the future:

- **Nonlinear problems:**
  - Our contribution [A5] on numerical schemes with a posteriori guarantees for the radiative transfer operator concerns a linear source problem of the form

    $$\mathcal{B}f = (\mathcal{T} - \mathcal{K})f = q,$$

    where the right-hand side is fixed. The obstructions of the source problem are aggravated when going for a generalized eigenpair (λ, u) of the following (nonlinear) problem

    $$(\mathcal{T} - \mathcal{K})u = \lambda \mathcal{F} u \tag{2.1.23}$$

where $\mathcal{T}$ and $\mathcal{K}$ are the transport and scattering operators as introduced in Section 2.1.1, and $\mathcal{F}$ is a fission operator which is of similar structure as $\mathcal{K}$. Under certain assumptions, $\mathcal{B}^{-1}\mathcal{F}$ is compact, and using the Krein-Ruthmann theorm, one can show that there exists a unique smallest positive real eigenvector λ with a nonnegative eigenstate $u \in U$. This problem is routinely solved in the field of nuclear engineering to find a steady state configuration of the nuclear reactor core but so far the numerical schemes do not come with a posteriori guarantees. The main difficulty from the mathematical stand-point is that we are looking for the eigenvalue of a *non-symmetric* problem. We are investigating this topic in an ongoing collaboration with with Prof. Dahmen.

  - Beyond the radiative transfer problem, I would like to develop similar techniques to solve the Vlasov-Poisson problem. The main strategy should rely in building an appropriate infinite dimensional Newton iteration scheme on a suitable linearization of the operator.

- **Model Order Reduction and Uncertainty Quantification:** A second research direction concerns the development of Reduced Order Models for kinetic problems and the study of UQ problems. One interesting problem in this direction would be to develop Model Order Reduction schemes for the non-symmetric eigenvalue problem (2.1.23) where the parameters of the PDE would typically be the optical parameters. Another question would be if we can address the problem of the oscillations in the energy variable as a UQ problem and use techniques from this field.

- **Model Error and Inverse Problems:** In Section 3.5.1, I outline the results of a collaboration with EDF in which we made some contributions in inverse problems for nuclear engineering using a diffusion version of equation (2.1.23). The results were obtained using synthetic measurement observations so we could not study the impact of model error in the reconstructions. In the future it would be very interesting to work on real data in order to study model errors and incorporate techniques that correct or mitigate the model deficiency in inverse problems.

2.2 An Adaptive Parareal Method

2.2.1 Motivation

As already brought up in the general introduction of Section 1, one desirable feature of numerical schemes for high-dimensional problems is the ability to *parallelize* computations. This can be particularly challenging for certain time-dependent problems due to the inherent sequential nature of the time variable. In a collaboration with Prof. Y. Maday, we have published a contribution on this topic in [A8]. The goal was to accelerate the numerical simulation of time dependent problems by time domain decomposition. The available algorithms enabling such decompositions present severe efficiency limitations and are not a competitive option for the solution of large scale and high dimensional problems. Our main contribution is the improvement of the parallel efficiency of the parareal in time method. This method is based on iteratively combining predictions made by a numerically inexpensive solver (with coarse physics and/or coarse resolution) with corrections coming from an expensive solver (with high-fidelity physics and high resolution). At convergence, the algorithm provides a solution that has the fine solver's high-fidelity physics and high resolution. In the classical version, the fine solver has a fixed high accuracy at all iterations (the one of the

expensive solver). This point is the major limitation to achieve a competitive parallel efficiency. In [A8], we develop an *adaptive* variant that overcomes this obstacle by dynamically increasing the accuracy of the fine solver across the parareal iterations. As we will see, the adaptive scheme is built in a similar way as the one developed for the N-ASTI scheme of Section 2.1.1 despite that the context and goals are significantly different.

We theoretically show that with our adaptive strategy the parallel efficiency becomes very competitive in the ideal case where the cost of the coarse solver is small, thus proving that the only remaining factors impeding full scalability become the cost of the coarse solver and communication time. The developed theory has also the merit of setting a general framework to understand the success of several extensions of parareal based on iteratively improving the quality of the fine solver and re-using information from previous parareal steps. We illustrate the actual performance of the method in stiff ODEs, which are a challenging family of problems since the only mechanism for adaptivity is time and efficiency is affected by the cost of the coarse solver.

Roadmap: To develop the adaptive algorithm, we formulate an ideal parareal scheme on an infinite dimensional functional setting. We then present feasible realizations involving a fine solver whose accuracy is adaptively increased across the iterations. We prove that the feasible adaptive algorithm converges at the same rate as the ideal one provided that the tolerances of the fine solver are increased at a certain rate which will be discussed. Finally, we discuss how the new paradigm can be realized thanks to adaptive schemes and/or the re-use of information from previous steps.

2.2.2 Setting and preliminary notations:

Let $\mathbb{U}$ be a Banach space of functions defined over a domain $\Omega \subset \mathbb{R}^d$ ($d \geq 1$). Let

$$\mathcal{E} : [0,T] \times [0,T] \times \mathbb{U} \to \mathbb{U}$$

be a propagator, that is, an operator such that, for any given time $t \in [0,T]$, $s \in [0, T-t]$ and any function $w \in \mathbb{U}$, $\mathcal{E}(t,s,w)$ takes w as an initial value at time t and propagates it at time $t+s$. We assume that $\mathcal{E}$ satisfies the semi group property

$$\mathcal{E}(t_0, t_2 - t_0, w) = \mathcal{E}(t_1, t_2 - t_1, \mathcal{E}(t_0, t_1 - t_0, w)), \quad \forall w \in \mathbb{U}, \forall (t_0, t_1, t_2) \in [0,T]^3,\ t_0 < t_1 < t_2.$$

We further assume that $\mathcal{E}$ is implicitly defined through the solution $u \in \mathcal{C}^1([0,T], \mathbb{U})$ of the time-dependent problem

$$u'(t) + \mathcal{A}\,(t, u(t)) = 0, \quad t \in [0,T], \tag{2.2.1}$$

where $\mathcal{A}$ is an operator from $[0,T] \times \mathbb{U}$ into $\mathbb{U}$ with adequate regularity we shall specify later. Then, given any $w \in \mathbb{U}$, $\mathcal{E}(t,s,w)$ denotes the solution to (2.2.1) at time $t+s$ with initial condition w at time $t \geq 0$. In our problem of interest, we study the evolution given by (2.2.1) when the initial condition is $u(0) \in \mathbb{U}$.

Since, in general, the problem does not have an explicit solution, we seek to approximate it at a given target accuracy. For any initial value $w \in \mathbb{U}$, any $t \in [0,T[$, $s \in [0, T-t]$ and any $\zeta > 0$ we denote by $[\mathcal{E}(t,s,w); \zeta]$ an element of $\mathbb{U}$ that approximates $\mathcal{E}(t,s,w)$ such that we have

$$\|\mathcal{E}(t,s,w) - [\mathcal{E}(t,s,w); \zeta]\| \leq \zeta\, s\, (1 + \|w\|), \tag{2.2.2}$$

where, here and in the following, $\|\cdot\|$ denotes the norm in $\mathbb{U}$. Any realization of $[\mathcal{E}(t,s,w); \zeta]$ involves three main ingredients:

i) a numerical scheme to discretize the time dependent problem (2.2.1) (e.g. an Euler scheme in time),

ii) a certain discretization error (e.g. error associated with the time step size of the Euler scheme),

iii) a numerical implementation to solve the resulting discrete systems (e.g. conjugate gradient, Newton method, SSOR, ...).

In the following, we will use the term *solver* to denote a particular choice for i), ii) and iii). Given a solver $\mathcal{S}$, we will use the same notation as for the exact propagator $\mathcal{E}$ to express that $\mathcal{S}(t,s,w)$ is an approximation of $\mathcal{E}(t,s,w)$ with a certain accuracy ζ. In other words, we can write $\mathcal{S}(t,s,w) = [\mathcal{E}(t,s,w);\zeta]$.

2.2.3 An idealized version of the parareal algorithm

We introduce a decomposition of the time interval $[0,T]$ into $\underline{N}$ subintervals $[T_N, T_{N+1}]$, $N = 0,\dots,\underline{N}-1$. Without loss of generality, we will take them of uniform size $\Delta T = T/\underline{N}$ which means that $T_N = N\Delta T$ for $N = 0,\dots,\underline{N}$. For a given target accuracy $\eta > 0$, the primary goal of the parareal in time algorithm is to build an approximation $\tilde{u}(T_N)$ of $u(T_N)$ such that

$$\max_{1\le N\le \underline{N}} \|u(T_N) - \tilde{u}(T_N)\| \le \eta.$$

The classical way to achieve this is to set

$$\tilde{u}(T_N) = \mathcal{S}_{\text{seq}}(0, T_N, u(0)) = [\mathcal{E}(0, T_N, u(0)); \zeta], \quad 1 \le N \le \underline{N},$$

where $\mathcal{S}_{\text{seq}}$ is some sequential solver in $[0,T]$ with $\zeta = \eta/(T(1+\|u(0)\|))$ in (2.2.2). Since this comes at the cost of solving the evolution over the whole time interval $[0,T]$, the main goal of the parareal in time algorithm is to speed up the computing time, while maintaining the same target accuracy η. This is made possible by first decomposing the computations over the time domain. Instead of solving over $[0,T]$, we perform $\underline{N}$ parallel solves over each interval $(T_N, T_{N+1}]$ of size ΔT. We next introduce an idealized version of it which will not be feasible in practice but which will be the starting point of subsequent implementable versions. The algorithm relies on the use of a solver $\mathcal{G}$ (known as the coarse solver) with the following properties involving the operator

$$\delta\mathcal{G} \coloneqq \mathcal{E} - \mathcal{G}.$$

Hypotheses (H): There exists constants $\varepsilon_{\mathcal{G}},\ C_c,\ C_d > 0$ such that for any function $x,\ y \in \mathbb{U}$ and for any $t \in [0, T[$ and $s \in [0, T-t]$,

$$\mathcal{G}(t,s,x) = [\mathcal{E}(t,s,x), \varepsilon_{\mathcal{G}}] \quad \Leftrightarrow \quad \|\delta\mathcal{G}(t,s,x)\| \le s(1+\|x\|)\varepsilon_{\mathcal{G}} \tag{H1}$$

$$\|\mathcal{G}(t,s,x) - \mathcal{G}(t,s,y)\| \le (1 + C_c s)\|x - y\|, \tag{H2}$$

$$\|\delta\mathcal{G}(t,s,x) - \delta\mathcal{G}(t,s,y)\| \le C_d s\varepsilon_{\mathcal{G}}\|x - y\| \tag{H3}$$

Note that hypothesis (H1) to (H3) are the classical abstract formulations of the properties of numerical schemes related to stability and accuracy. Hypothesis (H2) is a Lipschitz condition and the quantity $\varepsilon_{\mathcal{G}}$ is a small constant which, in the case of a Euler scheme, would be equal to the time step size.

The idealized version of the algorithm consists in building iteratively a series $(y_k^N)_k$ of approximations of $u(T_N)$ for $0 \leq N \leq \underline{N}$ following the recursive formula

$$\begin{cases} y_0^{N+1} = \mathcal{G}(T_N, \Delta T, y_0^N), & 0 \leq N \leq \underline{N} - 1 \\ y_{k+1}^{N+1} = \mathcal{G}(T_N, \Delta T, y_{k+1}^N) + \mathcal{E}(T_N, \Delta T, y_k^N) \\ \qquad\qquad - \mathcal{G}(T_N, \Delta T, y_k^N), & 0 \leq N \leq \underline{N} - 1, \quad k \geq 0, \\ y_0^0 = u(0). \end{cases} \tag{2.2.3}$$

At this point, several comments are in order:

1. The computation of y_k^N only requires propagations with $\mathcal{E}$ over intervals of size ΔT. As follows from (2.2.3), for a given iteration k, $\underline{N}$ propagations of this size are required, each of them over distinct intervals $[T_N, T_{N+1}]$ of size ΔT, each of them with independent initial conditions. Since they are independent from each other, they can be computed over $\underline{N}$ parallel processors and the original computation over $[0, T]$ is decomposed into parallel computations over $\underline{N}$ subintervals of size ΔT.

2. The algorithm may not be implementable in practice because it involves the exact propagator $\mathcal{E}$. Feasible instantiations consist of replacing $\mathcal{E}(T_N, \Delta T, y_k^N)$ by some approximation $[\mathcal{E}(T_N, \Delta T, y_k^N), \zeta_k^N]$ with a certain accuracy ζ_k^N which has to be carefully chosen. We will come to this point later on.

3. Note that, in the current version of the algorithm, for all $N = 0, \ldots, \underline{N}$, the exact solution $u(T_N)$ is obtained after exactly $k = N$ parareal iterations. This number can be reduced when we only look for an approximate solution with accuracy η. Depending on the problem, the final number of iterations $K(\eta)$ can actually be much smaller than $\underline{N}$.

The convergence result of Theorem 2.2.1 is helpful to understand the main mechanisms driving the convergence of the algorithm and explaining its behavior. To present it, we introduce the shorthand notation for the error norm

$$E_k^N := \|u(T_N) - y_k^N\|, \quad k \geq 0, \ 0 \leq N \leq \underline{N},$$

and the quantities

$$\mu = \frac{e^{C_c T}}{C_d} \max_{0 \leq N \leq \underline{N}} (1 + \|u(T_N)\|), \quad \text{and} \quad \tau := C_d T e^{-C_c \Delta T} \varepsilon_{\mathcal{G}}.$$

Theorem 2.2.1. *If $\mathcal{G}$ and $\delta\mathcal{G}$ satisfy Hypothesis (H1) to (H3), then,*

$$\max_{0 \leq N \leq \underline{N}} \|u(T_N) - y_k^N\| \leq \mu \frac{\tau^{k+1}}{(k+1)!}, \quad \forall k \geq 0. \tag{2.2.4}$$

Note that τ is the quantity driving convergence and its speed. Introducing the quantity

$$\bar{\varepsilon}_{\mathcal{G}} := \frac{e^{C_c \Delta T}}{C_d T},$$

we can write

$$\tau = \frac{\varepsilon_{\mathcal{G}}}{\bar{\varepsilon}_{\mathcal{G}}}$$

and we note that a sufficient condition to converge is that

$$\tau < 1 \quad \Leftrightarrow \quad \varepsilon_{\mathcal{G}} < \bar{\varepsilon}_{\mathcal{G}}. \tag{2.2.5}$$

In other words, $\bar{\varepsilon}_{\mathcal{G}}$ is the minimal accuracy that the coarse solver has to satisfy in order to guarantee convergence of the ideal parareal algorithm. In the following, we will work under the assumption that $\varepsilon_{\mathcal{G}}$ satisfies (2.2.5).

As we will see next, $\bar{\varepsilon}_{\mathcal{G}}$ plays also a critical role in certain convergence properties of the perturbed algorithm so we finish this section by discussing the behavior of $\bar{\varepsilon}_{\mathcal{G}}$ depending on several scenarios. First, C_c and C_d are Lipschitz constants (fixed by the properties of the evolution problem) so they could be potentially large numbers. As a result, $\bar{\varepsilon}_{\mathcal{G}}$ could be a large number and condition (2.2.5) would not be very stringent. The value of $\bar{\varepsilon}_{\mathcal{G}}$ can be small for very long time simulations where T becomes large or if ΔT becomes small compared to C_c (that is, if the number $\underline{N}$ of processors becomes large).

2.2.4 Feasible realizations of the parareal algorithm

Feasible versions of algorithm (2.2.3) involve approximations of $\mathcal{E}(T_N, \Delta T, y_k^N)$ with a certain accuracy ζ_k^N. This leads to consider algorithms of the form

$$\begin{cases} y_0^{N+1} = \mathcal{G}(T_N, \Delta T, y_0^N), & 0 \le N \le \underline{N} - 1 \\ y_{k+1}^{N+1} = \mathcal{G}(T_N, \Delta T, y_{k+1}^N) + [\mathcal{E}(T_N, \Delta T, y_k^N); \zeta_k^N] \\ \qquad\qquad - \mathcal{G}(T_N, \Delta T, y_k^N), & 0 \le N \le \underline{N} - 1,\ k \ge 0, \\ y_0^0 = u(0). \end{cases} \tag{2.2.6}$$

Since no feasible version will converge at a better rate than (2.2.4), we need to analyze what is the minimal accuracy ζ_k^N that preserves it. A result in this direction is given in the following theorem. It requires to introduce the quantity

$$\nu_p := \frac{\max_{0 \le N \le \underline{N}}(1 + \|y_p^N\|)}{\max_{0 \le N \le \underline{N}}(1 + \|u(T_N)\|)}, \quad \forall p \ge 0.$$

which tends to 1 as $p \to \infty$.

Theorem 2.2.2. *Let $\mathcal{G}$ and $\delta\mathcal{G}$ satisfy Hypothesis (H1) to (H3). Let $k \ge 0$ be any given positive integer. If for all $0 \le p < k$ and all $0 \le N < \underline{N}$, the approximation $[\mathcal{E}(T_N, \Delta T, \zeta_p^N)]$ has accuracy*

$$\zeta_p^N \le \zeta_p := \frac{\varepsilon_{\mathcal{G}}^{p+2}}{(p+1)!\nu_p}, \tag{2.2.7}$$

then the $(y_k^N)_N$ of the feasible parareal scheme (2.2.6) satisfy

$$\max_{0 \le N \le \underline{N}} \|u(T_N) - y_k^N\| \le \mu \frac{\tilde{\tau}^{k+1}}{(k+1)!}, \tag{2.2.8}$$

with

$$\tilde{\tau} := \tau + \varepsilon_{\mathcal{G}}.$$

Let us make a couple of remarks:

1. The sufficient condition to converge is now

$$\tilde{\tau} < 1 \quad \Leftrightarrow \quad \varepsilon_{\mathcal{G}} < \frac{\bar{\varepsilon}_{\mathcal{G}}}{1 + \bar{\varepsilon}_{\mathcal{G}}}$$

   so the minimal accuracy required for the coarse solver is stronger than the one for the ideal case (see (2.2.5)). Note however that when $\bar{\varepsilon}_{\mathcal{G}}$ is small (roughly, $\bar{\varepsilon}_{\mathcal{G}} \leq 1$), the condition on $\varepsilon_{\mathcal{G}}$ is similar in the ideal and perturbed case.

2. Comparing (2.2.4) and (2.2.8), the rate of convergence $\tilde{\tau}$ of the feasible parareal algorithm deviates from τ, the ideal one, by a factor

$$\frac{\tilde{\tau}}{\tau} = \frac{\tau + \varepsilon_{\mathcal{G}}}{\tau} = 1 + \frac{e^{C_c \Delta T}}{C_d T} = 1 + \bar{\varepsilon}_{\mathcal{G}}.$$

   The parameter $\bar{\varepsilon}_{\mathcal{G}}$ plays again a critical role in the convergence properties and determines whether convergence is close to the ideal rate τ, or deviates from it by a potentially important factor.

Practical realization of $[\mathcal{E}(T_N, \Delta T, y_k^N), \zeta_k^N]$: Since the accuracy ζ_k^N needs to improve with k, the most natural way to build the approximations $[\mathcal{E}(T_N, \Delta T, y_k^N), \zeta_k^N]$ is with adaptive techniques and with adaptive refinements at every step k. The implementation ultimately rests on the use of a posteriori error estimators. It opens the door to local time step adaptation in the parareal algorithm as well as spatial coarsening or refinement if the problem involves additional spatial variables.

In principle, as ζ_k^N decreases with k, the numerical cost increases in terms of degrees of freedom and also in terms of computing time. This actually reveals the key idea of this new approach which is that we would like that only the last fine solver is expensive and the cost of the previous ones is a small fraction of the cost of the last one (we refer to the next section for a more precise statement). By re-using information from previous iterations, we can limit the cost of internal solvers required in $[\mathcal{E}(T_N, \Delta T, y_k^N), \zeta_k^N]$ and enhance the speed-up. This depends of course on the nature of the specific problem.

The idea about re-using information from previous steps is actually the main point of contact between our work and previous contributions from the literature which have incorporated it with encouraging results in a variety of contexts. Among the most relevant ones stand the coupling of the parareal algorithm with spatial domain decomposition (see [72, 46, 17]), the combination of the parareal algorithm with iterative high order methods in time like spectral deferred corrections (see [61, 56, 34]) and, in a similar spirit, applications of the parareal algorithm to solve optimal control problems (see [72, 65]). Another relevant scenario where efficiency could be enhanced by reusing information from previous iterations is when internal iterative schemes are involved to solve the equation at every time step. This idea was first identified and proposed in my PhD thesis [T1], where I provided an analysis on a restricted setting. It has later been applied more extensively in the framework of the MGRIT algorithm that couples parareal with multigrid iterative schemes (see [24]).

2.2.5 Parallel efficiency

It is difficult to give accurate a priori estimations for the speed-up and efficiency of the method due to its adaptive nature so the actual performance can only be established through relevant examples but we make here some general remarks to highlighting the relevance of the cost of the coarse solver in the parallel speed-up. The speed-up is defined as the ratio

$$\mathbf{speed\text{-}up}_{\mathrm{AP/seq}}(\eta,[0,T]) \coloneqq \frac{\mathbf{cost}_{\mathrm{seq}}(\eta,[0,T])}{\mathbf{cost}_{\mathrm{AP}}(\eta,[0,T])} \tag{2.2.9}$$

between the cost to run a sequential fine solver achieving a target accuracy η with the cost to run an adaptive parareal algorithm providing at the end the same target accuracy η. The parallel efficiency of the method is then defined as the ratio of the above speed up with the number of processor which gives a target of 1 to any parallel solver:

$$\mathbf{eff}_{\mathrm{AP/seq}}(\eta,[0,T]) \coloneqq \frac{\mathbf{speed\text{-}up}_{\mathrm{AP/seq}}(\eta,[0,T])}{\underline{N}}.$$

The next Proposition gives an estimate of the parallel efficiency in the ideal case in which the cost of the coarse solve is negligible, and that there is no communication delay.

Proposition 2.2.3. *Suppose that the cost to realize* $[\mathcal{E}(T_N,\Delta T,y_k^N),\zeta_k^N]$ *is* $f_k^N=\Delta T(\zeta_k^N)^{-1/\alpha}$ *for some* $\alpha>0$*. Then, if the cost of the coarse solver is negligible with respect to* f_k^N *for any* $k\geq 0$*, we have*

$$\textit{eff}_{AP/seq}(\eta,[0,T]) = \frac{1-\tau^{1/\alpha}}{1-\tau^{K(\eta)/\alpha}} \sim \frac{1}{(1+\varepsilon_{\mathcal{G}}^{1/\alpha})}.$$

Therefore

$$\textit{speed-up}_{AP/seq}(\eta,[0,T]) \sim \underline{N}\frac{1}{(1+\varepsilon_{\mathcal{G}}^{1/\alpha})}.$$

In the ideal setting of Proposition 2.2.3:

- The parallel efficiency of the adaptive parareal algorithm does not depend on *the final number of iterations*. This is in contrast to the classical version whose efficiency decreases with the final number of iterations $K(\eta)$ as $1/K(\eta)$.
- The efficiency behaves like $1-o(\varepsilon_{\mathcal{G}})$ in the adaptive version, and $o(\varepsilon_{\mathcal{G}})$ rapidly goes to zero with $\varepsilon_{\mathcal{G}}$. As soon as $\varepsilon_{\mathcal{G}}$ becomes negligible with respect to 1, we will be in the range of full scalability.

We emphasize that, obviously, the above idealized setting will never hold in practice, but the result is interesting since it highlights that the cost of the fine solver is no longer the main obstacle for full scalability in the adaptive setting: the cost of the coarse solver becomes now the major obstruction towards full efficiency.

2.2.6 Guidelines for a practical implementation

- **Practical choice of** ζ_k^N Formula (2.2.7) of the convergence analysis of Theorem 2.2.2 gives an estimate for ζ_k^N that one could in principle use for the implementation. However, these

tolerances may not be optimal because they are derived from a theoretical convergence analysis based on abstract conditions for the coarse and fine solvers. This was confirmed during our numerical tests where we observed that using estimates (2.2.7) for ζ_k^N did not deliver satisfactory enough results. This is the reason why it is necessary to build a practical rule to set ζ_k^N. We have explored the following choice: if η is the final target accuracy, the classical parareal algorithm is usually run with a solver that delivers a slightly higher accuracy, say $\eta/2$. Assume that the classical algorithm converges in $K_{\mathrm{CP}}(\eta) = K$ iterations. We propose to build the tolerances of ζ_k^N in such a way to target that $K_{\mathrm{AP}}(\eta) = K_{\mathrm{CP}}(\eta)$ and such that the cost of the last fine propagation is of the order of the sum of the previous ones. This motivates to set

$$\zeta_k^N = \begin{cases} \varepsilon_{\mathcal{G}}^{1-\frac{k+1}{K}} \left(\frac{\eta}{2}\right)^{\frac{k+1}{K}}, & \text{if } k < K \\ \eta/2, & \text{if } k \geq K. \end{cases}$$

The numerical example of the next section uses these tolerances.

- **Load balancing:** For simplicity of exposition, the algorithm has so far been discussed for $\underline{N}$ subintervals of uniform size ΔT. However, this decomposition may lead to a task imbalance because some time intervals may have more complex dynamics than others, requiring more degrees of freedom, thus more computational time. In order to balance tasks as efficiently as possible, we dynamically adapt the size of the $\underline{N}$ subintervals in a way to have the fine solver propagations as balanced as possible among processors.

2.2.7 Numerical tests for several stiff ODEs

We apply our adaptive algorithm to several stiff ODEs where the only mecanism for adaptivity is time. Our results illustrate that our approach improves the speed-up and efficiency with respect to the classical non-adaptive parareal method. We also show that the main element affecting performance is no longer the cost of the fine solver but the cost of the coarse solver. In extreme cases, this cost may even prevent any speed-up at all and puts this obstruction at the forefront for future research. The code to reproduce the numerical results is available online at:

https://plmlab.math.cnrs.fr/mulahernandez/parareal-adaptive

Other ODEs can easily be tested as indicated in the instructions. Note that the algorithm could also be applied to PDEs but we defer the presentation of numerical examples to future works since this requires full space-time adaptive techniques which are a topic in itself since they are challenging to formulate and deploy, and there are also very specific to each type of problem.

The Brusselator system: We consider the brusselator system

$$\begin{cases} x' = A + x^2 y - (B+1)x \\ y' = Bx - x^2 y, \end{cases}$$

with initial condition $x(0) = 0$ and $y(0) = 1$. This is a stiff ODE that models a chain of chemical reactions. It was already studied in a previous work on the parareal algorithm (see [60]). The system has a fixed point at $x = A$ and $y = B/A$ which becomes unstable when $B > 1 + A^2$ and leads to oscillations. We place ourselves in this oscillatory regime by setting $A = 1$ and $B = 3$. The

dynamics present large velocity variations in some time subintervals, making the use of adaptive time-stepping schemes particularly desirable for an appropriate treatment of the transient.

For the coarse solver, we set

$$\varepsilon_{\mathcal{G}} = 0.1,$$

and use an explicit Runge Kutta method of order 5 with an adaptive time-stepping (see [91]). For the fine solver, we use the implicit Runge-Kutta method of the Radau IIA family of order 5 with adaptive time-stepping (see [85, 81]). Both integrators are available in the ODE integration library of Scipy[1] which we have used in our library.

As already discussed, the target accuracies ζ_k^N should be ensured by rigorous a posteriori error estimators. However, these type of estimators are unfortunately not available in the Scipy library and we are not aware of any mainstream library with this capability. As a surrogate, we have used the above mentioned classical ODE integrators that only guarantee *local* accuracy between time-steps $t_n \to t_{n+1}$, but not *global* accuracy between macro intervals $[T_N, T_{N+1}]$ (composed of several time-steps). The local accuracy can be specified in the library routine via the parameters `atol` and `rtol` of the function `scipy.integrate.solve_ivp`. To relate this local accuracy control to the global one, we have built a priori a "chart" mapping accuracies of the solver on macro-intervals against the tolerance parameters `atol` and `rtol` of the library. To simplify, these two parameters have been set to be equal (`atol` = `rtol`) and their value is fixed according to the chart. As an example, we provide a chart for $T = 20$ for the scheme of the fine solver in Figure 2.9. The dots are computed values: for a given value of the parameter `atol`, we examine the accuracy ε of the solver. We then interpolate the points with a cubic spline interpolation. This way, for a given intermediate accuracy ζ_k^N in the parareal algorithm, we can easily adapt the parameter value `atol` that is required.

We use formula (2.2.9) to compare the speed-up of the classical and adaptive parareal algorithm in terms of the number of operations involved in the numerical solution (communication delays have not been taken into account). For the costs g_k^N and f_k^N, we take into account:

- the number of time steps (which is adaptively increased as we tightened the accuracy),
- the number of right-hand side evaluations,
- for the fine solver, we additionally count the number of evaluations of the Jacobian matrix and of the number of linear system inversions.

In Figure 2.10, we plot the obtained speed-up for different configurations:

- the final time T varies from 100 to 900,
- the final target accuracy is $\eta = 10^{-6}$ or $\eta = 10^{-8}$,
- the number of processors $\underline{N}$ varies from 10 to 100.

The speed-up of the adaptive parareal is always superior to the one of the classical parareal. We observe that the gain is marginal for a moderate accuracy ($\eta = 10^{-6}$) but it is about 2.5 times larger for $\eta = 10^{-8}$. Note that sometimes the speed-up does not increase monotonically as the number of processors $\underline{N}$ increases. Also, the speed-up generally increases with $\underline{N}$ but the increase is rather moderate.

[1] https://docs.scipy.org/doc/scipy/reference/generated/scipy.integrate.solve_ivp.html

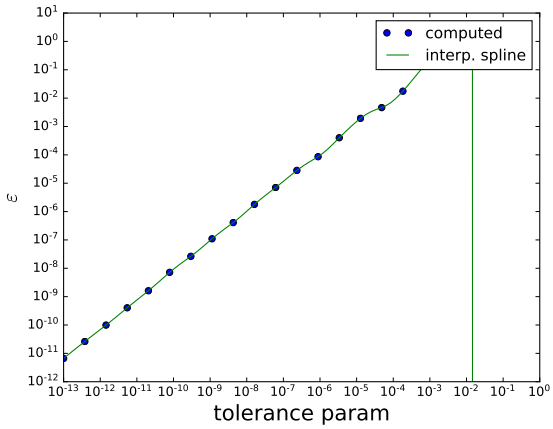


Figure 2.9: Mapping of the accuracies ε against the tolerance parameters (`atol` = `rtol`) of the library. The dots are computed values: for a given value of the parameter, we examine the accuracy ε of the solver. We then interpolate the points with a cubic spline interpolation. This way, for a given intermediate accuracy ζ_k^N in the algorithm, we can infer the parameter value `atol` and `rtol`. Case $T = 20$, integrator of the fine solver.

The values significantly differ from the range of full scalability and we next explain why this is mainly due to the cost of the coarse solver. Since the problem is stiff and we consider relatively long time intervals, it has been necessary to use a sufficiently accurate coarse solver. This explains our choice of an explicit Runge-Kutta scheme of order 5. To illustrate the impact of its cost, let us fix $T = 500$, $\eta = 10^{-8}$ and $\underline{N} = 50$ (other parameters would yield similar conclusions). We compare the speed-up and efficiency when we count or do not count the cost of the coarse solver in Table 2.1. Obviously, when we do not count the cost of the coarse solver, the performance of both algorithms improves but it is particularly increased in the case of the adaptive version. If the cost of $\mathcal{G}$ was negligible, it would deliver a very satisfactory efficiency of 75.52%. This is five times larger than what the classical parareal would yield. This analysis illustrates that the major obstacle to achieve competitive scalabilities is no longer the cost of the fine solver like in the classical version, but the cost of the coarse propagator.

We next give some insight on the differences in the convergence behavior of both algorithms. We fix $T = 20$, $\eta = 10^{-8}$ and $\underline{N} = 20$ and plot in Figure 2.11 the convergence history of the parareal solution in terms of:

- the errors of the fine solver at every fine time-step
- the maximum error of the parareal solution at the macro-intervals

$$\max_N \|u(T_N) - y_k^N\|$$

Note that the maximum error in the adaptive scheme steadily decreases to the desired accuracy whereas the error in the classical scheme degrades at iteration $k = 1$ before converging. This type of behavior has been observed for all other configurations and we conjecture that an important difference in accuracy between the coarse and the fine solver at early stages of the algorithm may

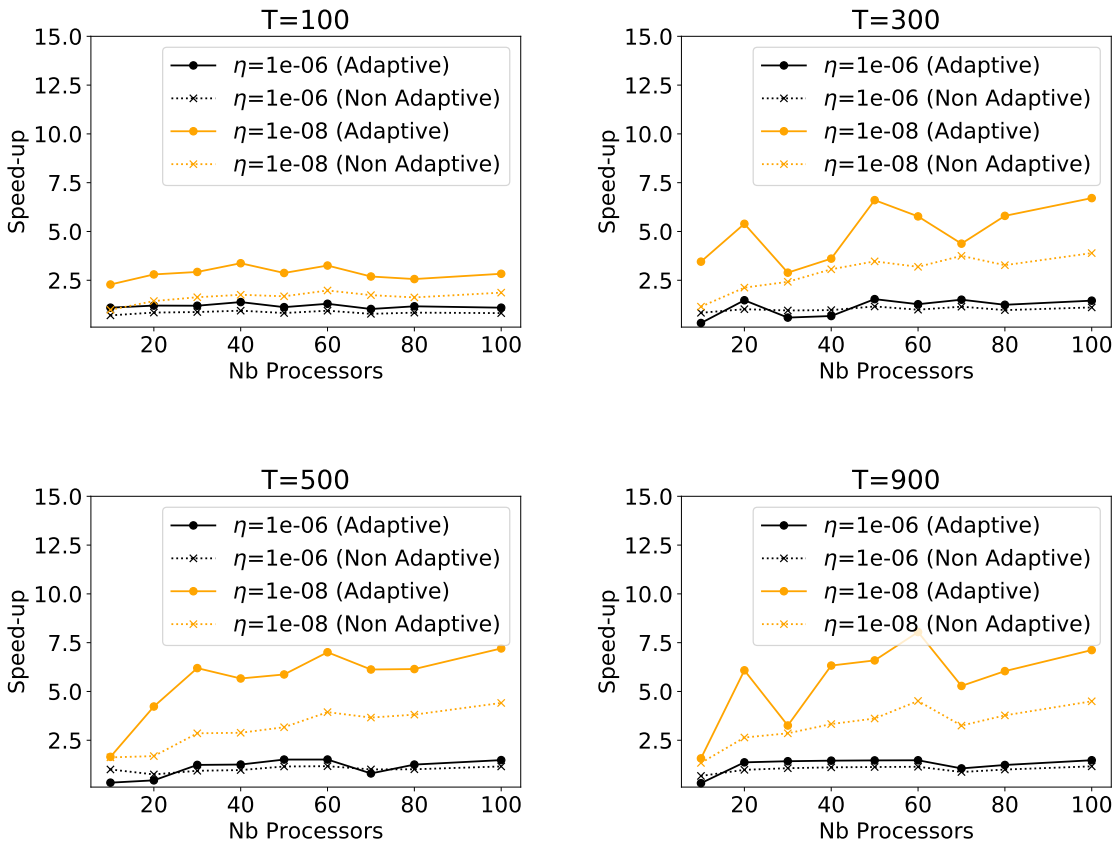


Figure 2.10: Speed-up in comparison to running a sequential fine solver as a function of the number of processors $\underline{N}$. Dashed lines: classical parareal. Continuous lines: Adaptive parareal.

Speed-up	Classical parareal	Adaptive Parareal
With cost $\mathcal{G}$	4.06	7.38
Without cost $\mathcal{G}$	7.38	37.76

Efficiency	Classical parareal	Adaptive Parareal
With cost $\mathcal{G}$	8%	14.76%
Without cost $\mathcal{G}$	14.76%	75.52%

Table 2.1: **Brusselator:** Impact of the cost of the coarse solver. Speed-up and efficiency with $T = 500$, $\eta = 10^{-8}$ and $\underline{N} = 50$.

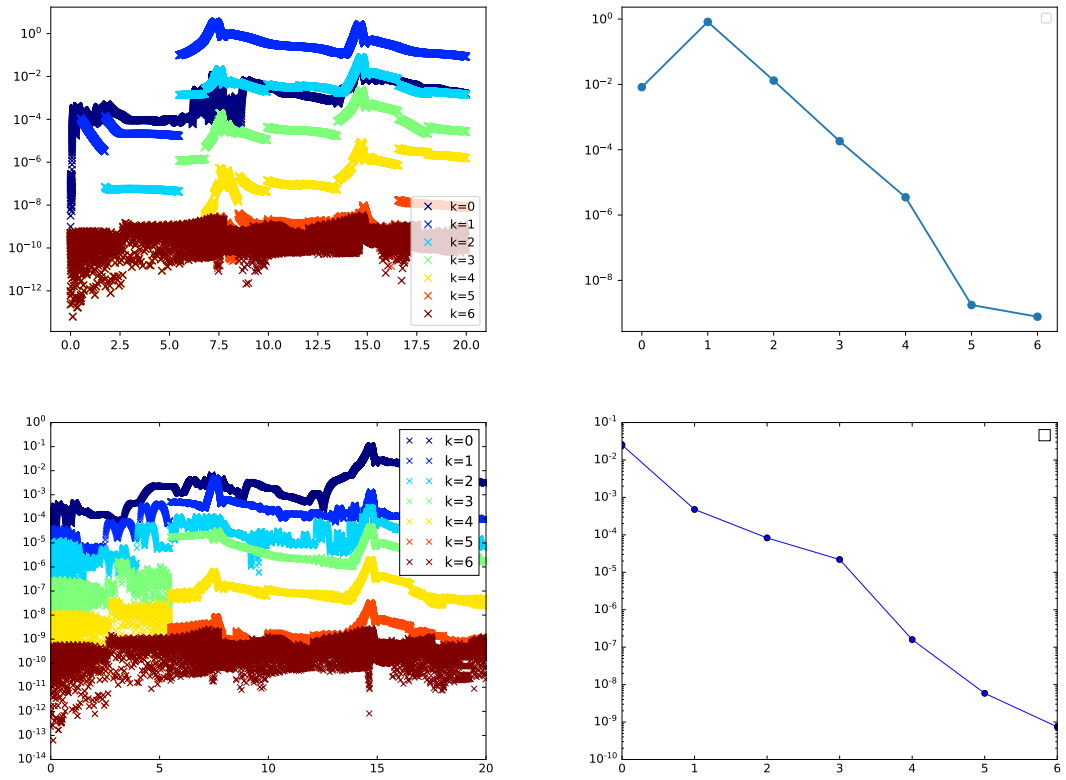


Figure 2.11: **Brusselator:** Convergence history of the errors for $T = 20$, $\eta = 10^{-8}$ and $\underline{N} = 20$. Top: classical parareal. Bottom: adaptive parareal. Left: errors of the fine solver at every fine time-step. Right: maximum parareal error at each iteration k.

be the cause. Finally, an inspection of the error of the fine solver shows that the adaptive algorithm succeeds to reduce the error at every time t in a much more uniform way than the classical algorithm.

The Van der Pol oscillator We next consider the Van der Pol oscillator

$$\begin{cases} x' = y \\ y' = \mu(1 - x^2)y - x, \end{cases}$$

with initial condition $x(0) = 2$ and $y(0) = 0$. When $\mu = 0$, this equation is a simple nonstiff harmonic oscillator. When $\mu > 0$, the system has a limit cycle and becomes stiffer and stiffer as its value is increased. For our tests, we set $\mu = 4$ which is already a relatively stiff case.

Like in the example of the Brusselator system, we set $\varepsilon_{\mathcal{G}} = 0.1$ for the coarse solver and use an explicit Runge Kutta method of order 5 with an adaptive time-stepping (see [91]). For the fine solver, we use the implicit Runge-Kutta method of the Radau IIA family of order 5 with adaptive time-stepping.

In Figure 2.12, we plot the obtained speed-up for different configurations:

- the final time T is 1000 or 2000,
- the final target accuracy is $\eta = 10^{-6}$ or $\eta = 10^{-8}$,
- the number of processors $\underline{N}$ varies from 10 to 100.

Like in the previous example, the adaptive algorithm outperforms the nonadaptive version in terms of speed-up. However, the gain is marginal for moderate accuracies $\eta = 10^{-6}$. For high accuracy $\eta = 10^{-8}$, the adaptive algorithm improves the speed-up by a factor of about 2 to 3 times with respect to the classical one. The improvement is more significant for large T.

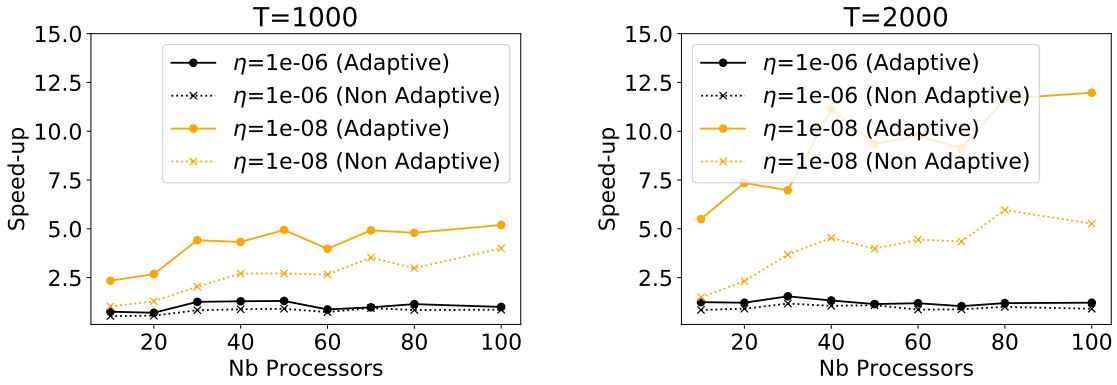


Figure 2.12: **Van der Pol:** speed-up in comparison to running a sequential fine solver as a function of the number of processors $\underline{N}$. Dashed lines: classical parareal. Continuous lines: Adaptive parareal.

In Table 2.2, we illustrate that the coarse solver is again the main bottleneck to reach high parallel efficiency in the adaptive algorithm: we examine the speed-up and efficiency for $T = 2000$, $\eta = 10^{-8}$ and $\underline{N} = 40$ when we take and do not take into account the cost of the coarse solver.

Speed-up	Classical parareal	Adaptive Parareal
With cost $\mathcal{G}$	4.54	11.14
Without cost $\mathcal{G}$	6.61	32.63

Efficiency	Classical parareal	Adaptive Parareal
With cost $\mathcal{G}$	11.35%	27.8%
Without cost $\mathcal{G}$	16.5%	81.56%

Table 2.2: **Van der Pol:** Impact of the cost of the coarse solver. Speed-up and efficiency with $T = 2000$, $\eta = 10^{-8}$ and $\underline{N} = 40$.

2.2.8 Research Perspectives

Further investigations which could be pursued on this topic are:

- **Further enhancements and applications of the adaptive parareal algorithm:**
  - The main bottleneck of the parareal algorithm after our contribution [A8] relies in the cost of the coarse solver. One natural idea to explore is that if we need to repeat the simulations for different parameter values, one could build a coarse solver based on reduced modelling or machine learning techniques. An interesting element could be to try to exploit the coarse solver propagations computed during the parareal iterations to enrich the training set on the fly.
  - It would be interesting to consider PDE problems in which the adaptive parareal algorithm is coupled with adaptive spatial domain decomposition.
  - The adaptive parareal algorithm could be applied to speed up optimal control problems involving the solution of the Pontryagin Maximum Principle. In particular, this idea could be leveraged for the training of ODE-Nets as we outline in our paper [S2].
- **Alternative time domain decomposition techniques without a coarse solver:** The question of developing perfectly scalable algorithms based on domain decomposition is an old problem which is still nowadays open because there does not seem to exist a general recipe to succeed. This is particularly the case in transport dominated PDEs. To make progress in this challenging task, it may be necessary to develop schemes that are not based on a prediction-correction strategy like in the case of the parareal algorithm, but rather on "one-shot" straightforward strategies. A direction which, in my view, may be promising would be to search for coordinate transformations, possibly time-dependent, for which the nature of the PDE in the transformed coordinates makes the problem easier to parallelize.

2.3 Model-Order Reduction for Transport Dominated Problems

In this Section I summarize [A6], a work in collaboration with V. Ehrlacher, D. Lombardi and F.X. Vialard in which we extend the notion of of reduced order modeling to metric spaces, and leverage the new point of view to address model reduction of transport dominated problems.

2.3.1 Motivation

In modern applications of science, industry and numerous other fields, the available time for design and decision-making is often short, and some tasks are even required to be performed in real time. The process usually involves predictions of the state of complex systems which, in order to be reliable, need to be described by sophisticated models. The predictions are generally the output of inverse or optimal control problems that are formulated on these models and which cannot be solved in real time unless the overall complexity has been appropriately reduced. Our focus lies in the case where the model is given by a PDE that depends on certain parameters. In this setting, the routines for prediction require to evaluate solutions of the PDE on a large set of dynamically updated parameters. This motivates the search for accurate and online methods to approximate the solutions at a reduced computational cost. Depending on the different scientific communities and on the nature of the strategy to reduce the cost, this task is known under the name of *reduced modelling*, *model order reduction*, *surrogate modeling* or *metamodeling*.

Following the notation of the main introduction (Section 1.1), we write the parametric PDE as

$$\mathcal{P}(u, y) = 0, \tag{2.3.1}$$

where $\mathcal{P}$ is a partial differential operator, and $y = (y_1, \dots, y_p)$ is a vector of scalar parameters ranging in some domain $\mathrm{Y} \subset \mathbb{R}^p$. We assume well-posedness, that is, for any $y \in Y$ the problem admits a unique solution $u = u(y)$ in some Hilbert, Banach or metric space (V, d) with metric d. The set of solutions when the parameters y range in Y defines the so-called *solution manifold*

$$\mathcal{M} := \{u(y) \,:\, y \in Y\} \subset V$$

which we already introduced in equation (1.1.3) of Section 1.1.

The bottom line of most model reduction strategies has so far been based on posing the problem on a Hilbert or Banach space V and determining a "good" n-dimensional subspace $V_n = \operatorname{span}\{v_1, \dots, v_n\} \subset V$ that yields efficient approximations of $u(z)$ in V_n of the form

$$u_n(y) := \sum_{i=1}^{n} c_i(y) v_i \tag{2.3.2}$$

for some coefficients $c_1(y), \cdots, c_n(y) \in \mathbb{R}$. This approach is the backbone of most existing methods among which stand the reduced basis method ([32, 30]), the empirical interpolation method and its generalized version (G-EIM, [73, 64, A14, A12]), Principal Component Analysis (PCA, see [18, Chapter 1]), polynomial-based methods like [54, 50] or low-rank methods ([51]).

The approximation quality of the obtained subspace V_n is either measured through the worst case error

$$e^V_{\mathrm{wc}}(\mathcal{M}, V_n) := \sup_{y \in \mathrm{Y}} \inf_{v \in V_n} d\left(u(y), v\right),$$

or the average error

$$e^V_{\mathrm{av}}(\mathcal{M}, V_n) := \left(\int_{y \in \mathrm{Y}} \inf_{v \in V_n} d^2\left(u(y), v\right) \, \mathrm{d}\mu(y) \right)^{1/2},$$

where μ is a probability measure on Y, given a priori and from which the parameters are sampled.

The reduction method is considered efficient if $e^V_{\mathrm{wc}}(\mathcal{M}, V_n)$ or $e^V_{\mathrm{av}}(\mathcal{M}, V_n)$ decays rapidly to 0 as n goes to ∞. There is sound evidence of efficiency only in the case of certain elliptic and parabolic

PDEs. More precisely, it has been shown in [28] that for this type of equations, under suitable assumptions, the L^∞-Kolmogorov width defined as

$$d_n^V(\mathcal{M}) := \inf_{\substack{V_n \subset V \\ \dim V_n = n}} e_{\mathrm{wc}}^V(\mathcal{M}, V_n) \tag{2.3.3}$$

and the L^2-Kolmogorov width

$$\delta_n^V(\mathcal{M}) := \inf_{\substack{V_n \subset V \\ \dim V_n = n}} e_{\mathrm{av}}^V(\mathcal{M}, V_n) \tag{2.3.4}$$

decay exponentially or polynomially with high exponent as n grows. In the context of model reduction, this quantity gives the best possible performance that one can achieve when approximating $\mathcal{M}$ with n-dimensional linear spaces.

Optimal linear subspaces $V_n \subset V$ of dimension n which realize the infimum of (2.3.3) cannot be computed in practice in general. However, it has been shown that greedy algorithms can be used to build sequences of linear spaces $(V_n)_{n\geq 1}$ whose approximation error $e_{\mathrm{wc}}^V(\mathcal{M}, V_n)$ decay at a comparable rate as the Kolmogorov n-width $d_n^V(\mathcal{M})$. These algorithms are the backbone of the so-called Reduced Basis method [48]. In the case of (2.3.4), the optimal subspaces for which the minimum is attained are obtained using the PCA or Proper Orthogonal Decomposition (POD) method.

In [A6], our goal has been to extend the above notion of model reduction to more general metric spaces in view of the following facts:

- First of all, in the context of Banach or Hilbert spaces, linear methods are unfortunately not well suited for hyperbolic problems. Among others, this is due to the transport of shock discontinuities whose locations may vary together with the parameters. It was proved in [18, Chapter 3, see equation (3.76)] that the L^∞-Kolmogorov width of simple pure transport problems decays very slowly, at a rate $n^{-1/2}$ if $V = L^2$ (similar examples can be found in [29, 26, A10]). The same type of result has recently been derived for wave propagation problems in [11]. These results highlight that linear methods of the type (2.3.2) are not expected to provide a fast decay in numerous transport dominated problems, and may be highly suboptimal in terms of the trade off between accuracy and numerical complexity. For these classes of problems, an efficient strategy for model reduction requires to look for *nonlinear methods* that capture the geometry of $\mathcal{M}$ in a finer manner than linear spaces.

- In addition to the idea of searching for nonlinear methods, it may be beneficial to move from the classical Banach/Hilbert metric framework to more general metric spaces in order to better quantify the ability to capture specific important features like translations or shifts.

- Finally, this broader setting enlarges the scope of problems that can be treated. Relevant classes of problems could be posed either on Banach spaces or on metric spaces and the latter characterization *may be* more convenient for model reduction in some situations. To name a few examples involving gradient flows, we cite [77] for Hele-Shaw flows, [58] for quantum problems, [78] for porous media flows, [82] for Fokker-Planck equations and [35, 39] for Keller-Segel models in chemotaxis. Other examples involving metric spaces that are not necessarily related to gradient flows are [70] for the Camassa-Holm equation, [8] for the Hunter-Saxton equation. Such examples can often be interpreted as a geodesic flow on a group of diffeomorphisms and can thus be encoded as Hamiltonian flows. In addition to this, there are

other problems which cannot be defined on Banach vector spaces and can only be defined over metric spaces. Consider for instance the case of a pure transport equation with constant velocity where the initial data is a Dirac measure concentrated in one point. The solution of this PDE remains at all times a (translated) Dirac mass. More generally, it has been proven that solutions to certain nonlinear dissipative evolution equations with measure-valued initial data are measure-valued and do not belong to some standard Lebesgue or Sobolev spaces. They can however be formulated in the form of Wasserstein gradient flows.

2.3.2 Model Order Reduction in the L^2-Wasserstein space

Setting: In [A6], we have developped a model order reduction method for conservation laws based on the following remark: for any given parameter $y \in \mathrm{Y}$, if the solution $u(y)$ of a conservative PDE is nonnegative, it induces a measure $\rho(y)$ defined by

$$\rho(y)(A) \coloneqq \int_A u(y)(x)\mathrm{d}x, \quad \text{for any borel subset } A \text{ of } \Omega,$$

Thus we assume in the following that $u(y) \geq 0$ and that $\int_\Omega u(y)(x)\mathrm{d}x = 1$. We also assume that the regularity of $u(y)$ is such that the measure $\rho(y)$ belongs to $\mathcal{P}_2(\Omega)$, the set of probability measures on Ω with finite second-order moments. The space $\mathcal{P}_2(\Omega)$ is a metric space when endowed with the L^2-Wasserstein distance defined as

$$W_2(u,v) \coloneqq \inf_{\pi \in \Pi(u,v)} \left(\int_{\Omega\times\Omega} (x-y)^2 \, d\pi(x,y) \right)^{1/2}, \quad \forall (u,v) \in \mathcal{P}_2(\Omega) \times \mathcal{P}_2(\Omega),$$

where $\Pi(u,v)$ is the set of probability measures on $\Omega \times \Omega$ with marginals u and v. This space is usually called the L^2-Wasserstein space (see [75] for more details).

In the particular case of one dimensional marginal domains, the L^2-Wasserstein distance can be expressed using the inverse cumulative distribution functions as

$$W_2(u,v) = \| \operatorname{icdf}(u) - \operatorname{icdf}(v) \|_{L^2([0,1])}. \tag{2.3.5}$$

Wasserstein Barycenters: The proposed model reduction strategy is based on using Wasserstein barycenters. We next recall the main definitions. Let $n \in \mathbb{N}^*$ and let

$$\Sigma_n \coloneqq \left\{ (\lambda_1, \cdots, \lambda_n) \in [0,1]^n, \quad \sum_{i=1}^n \lambda_i = 1 \right\}$$

be the set of barycentric weights. For any $\mathrm{U}_n = (u_i)_{1\leq i\leq n} \in \mathcal{P}_2(\Omega)^n$ and barycentric weights $\Lambda_n = (\lambda_i)_{1\leq i\leq n} \in \Sigma_n$, an associated barycenter is an element of $\mathcal{P}_2(\Omega)$ which minimizes

$$\inf_{v\in\mathcal{P}_2(\Omega)} \sum_{i=1}^n \lambda_i W_2^2(v,u_i). \tag{2.3.6}$$

The existence of a minimizer is guaranteed under mild conditions. For instance, it suffices that at least one barycentric function u_i vanishes on small sets in the sense defined in [47]. In full generality, minimizers to (2.3.6) may not be unique. In the following, we call $\mathrm{Bar}(\mathrm{U}_n, \Lambda_n)$ the set of minimizers to (2.3.6), which is the set of barycenters of U_n with barycentric weights Λ_n.

We next introduce the notion of *optimal barycenter* of an element $u \in \mathcal{P}_2(\Omega)$ for a given family $\mathrm{U}_n \in \mathcal{P}_2(\Omega)^n$. The set of barycenters with respect to U_n is

$$\mathcal{B}_n(\mathrm{U}_n) \coloneqq \{\mathrm{Bar}(\mathrm{U}_n, \Lambda_n) \ : \ \Lambda_n \in \Sigma_n\}$$

and an optimal barycenter of the function $u \in V$ with respect to the set U_n is a minimizer of

$$\min_{b \in \mathcal{B}_n(\mathrm{U}_n)} W_2^2(u, b). \tag{2.3.7}$$

This problem can equivalently be written as the search for optimal barycentric weights

$$\Lambda_n^*(u) \in \arg\min_{\Lambda_n \in \Sigma_n} W_2^2(u, \mathrm{Bar}(\mathrm{U}_n, \Lambda_n))$$

In other words, a minimizer to (2.3.7) has the form $\mathrm{Bar}(\mathrm{U}_n, \Lambda_n^*(u))$ and it is the projection of u on the set of barycenters $\mathcal{B}_n(\mathrm{U}_n)$.

A Barycentric Greedy Algorithm: In [A6] we explored two model reduction strategies for conservation laws. The first was the so-called Tangent PCA method (tPCA), which consists in mapping the manifold to a tangent space and performing a standard PCA on this linearization. This method has drawn significant attention in numerous fields like pattern recognition, shape analysis, medical imaging, computer vision (see [74, 37]). It is well-documented that this approach suffers from certain stability issues. In [A6] we have developed an alternative strategy based on a barycentric greedy algorithm which we call gBar in the following. To the best of our knowledge, this strategy is novel, and it could be used as an alternative to tPCA in other applications apart from model reduction.

The advantages of our appoach with respect to tPCA are the following. Our strategy can be defined for general metric spaces (V, d). Contrary to tPCA, the space does not need to be embedded with a Riemannian manifold structure. The method is also guaranteed to be stable in the sense that all the steps of the algorithm are well-defined. The stability comes at the price of difficulties in connecting theoretically its approximation quality with optimal performance quantities. Thus its quality has been evaluated through numerical examples.

The offline phase of the gBar method is an iterative algorithm which can be written as follows:

- Compute a training set of functions $\mathcal{M}_{\mathrm{tr}} \subset \mathcal{M}$ associated to training parameters Y_{tr}.
- **Initialization:** Find $(u_1, u_2) \in \mathcal{M}_{\mathrm{tr}} \times \mathcal{M}_{\mathrm{tr}}$ such that

  $$(u_1, u_2) \in \operatorname*{argmax}_{(v_1, v_2) \in \mathcal{M}_{\mathrm{tr}} \times \mathcal{M}_{\mathrm{tr}}} d(v_1, v_2)^2,$$

  and define $\mathrm{U}_2 \coloneqq \{u_1, u_2\}$. Then compute and store

  $$\Lambda_2(y) \in \operatorname*{argmin}_{\Lambda_2 \in \Sigma_2} d\left(u(y), \mathrm{Bar}(\mathrm{U}_2, \Lambda_2)\right)^2, \quad \forall y \in \mathrm{Y}_{\mathrm{tr}}.$$

- **Iteration** $n \geq 3$**:** Find $u_n \in \mathcal{M}_{\mathrm{tr}}$ such that

  $$u_n \in \operatorname*{argmax}_{v \in \mathcal{M}_{\mathrm{tr}}} \min_{b \in \mathcal{B}_{n-1}(\mathrm{U}_{n-1})} d(v, b)^2.$$

  and set $\mathrm{U}_n \coloneqq \mathrm{U}_{n-1} \cup \{u_n\}$. Then compute and store

  $$\Lambda_n(y) \in \operatorname*{argmin}_{\Lambda_n \in \Sigma_n} d\left(u(y), \mathrm{Bar}(\mathrm{U}_n, \Lambda_n)\right)^2, \quad \forall y \in \mathrm{Y}_{\mathrm{tr}}.$$

- **Terminal Step:** For a given target accuracy $\varepsilon > 0$, the algorithm terminates when

$$\max_{\tilde{y}\in \mathrm{Y}_{\mathrm{tr}}} \min_{b\in\mathcal{B}_{n-1}(\mathrm{U}_{n-1})} d(u(\tilde{y}), b)^2 = \min_{b\in\mathcal{B}_{n-1}(\mathrm{U}_{n-1})} d(u(y_n), b)^2 < \varepsilon^2.$$

Note that the gBar algorithm selects via a greedy procedure particular snapshots

$$\mathrm{U}_n = \{u(y_1), \cdots, u(y_n)\}$$

in order to approximate as well as possible each element $u(y) \in \mathcal{M}_{\mathrm{tr}}$ with its optimal barycenter associated to the family U_n. The barycentric weights have to be determined via an optimization procedure. We momentarily postpone the discussion on the feasibility of all the steps in order to present the online procedure.

In principle, we can consider two different versions of the online phase of the gBar algorithm:

- **Projection:** Let $y \in \mathrm{Y}$. Compute $\Lambda_n(y) \in \Sigma_n$ a minimizer of

$$\Lambda_n(y) \in \operatorname*{argmin}_{\Lambda_n\in\Sigma_n} d(u(y), \mathrm{Bar}(\mathrm{U}_n, \Lambda_n))^2,$$

  and approximate $u(y)$ with $u_n^{\mathrm{gBar,proj}}(y) \in \mathrm{Bar}(\mathrm{U}_n, \Lambda_n(y))$.

- **Interpolation:** From the values $(\Lambda_n(y))_{y\in\mathrm{Y}_{\mathrm{tr}}}$ which are known from the offline stage, compute an interpolant $\overline{\Lambda}_n : \mathrm{Y} \to \Sigma_n$ such that

$$\overline{\Lambda}_n(y) = \Lambda_n(y), \quad \forall y \in \mathrm{Y}_{\mathrm{tr}}.$$

  For a given $y \in \mathrm{Y}$, we approximate $u(y)$ with $u_n^{\mathrm{gBar,interp}}(y) \in \mathrm{Bar}(\mathrm{U}_n, \overline{\Lambda}_n(y))$.

In practice, the only viable online strategy is the one based on the interpolation of the barycentric coefficients since the projection method requires the computation of the full solution $u(y)$ for $y \in \mathrm{Y}$. Both approaches are purely data-driven and do not involve solving the original PDE in a reduced space or manifold in the online phase. We compare the quality of both strategies in our numerical tests.

Challenges for practical implementation in the L^2-Wasserstein space: We apply the gBar algorithm to conservation laws using the L^2-Wasserstein distance. In this case, the implementation of the above steps crucially relies in the ability to accurately compute Wasserstein distances and barycenters. It is also required that the computation of barycenters is fast enough in order to guarantee that the online step is competitive with respect to directly computing the solution with the high fidelity numerical solver. In the case of one spatial dimension, computations are greatly facilitated and sped up thanks to the existence of the closed form (2.3.5) for the Wasserstein distance. As a consequence, this entails that the computation of the optimal barycentric weights becomes a simple convex quadratic optimization problem. In [A6], we leverage these facts to show in numerical examples that the strategy is online efficient. As we illustrate in our numerical examples below, the obtained speed-ups factors were of the order of 5.10^2 compared to a direct computation with the high-fidelity solver.

The extension to higher spatial dimensions is far from trivial from a practical computational standpoint. One viable possibility is to resort to entropic regularized version of the Wasserstein

distance, leading to Sinkhorn-type algorithms amenable for computation (see [15]). In a current collaboration with Hieu Do, a post-doctoral student under my supervision, and Jean Feydy (postdoc at Imperial College), we are currently exploring the potential of the same algorithm by replacing W_2 with the regularized version $W_2^{(\eta)}$ where $\eta > 0$ is a regularization parameter (see, e.g., [15, Chapter 4] for the definition). In this setting, it is possible to compute barycenters with a Sinkhorn-type algorithm. With a highly optimized implementation that resorts to the latest developments in numerical optimal transport, if the functions are discretized with N degrees of freedom, the incompressible cost of the barycentric computation is of the order $\mathcal{O}(nN \log N)$ with a proportionality factor that depends mildly on the dimension d. Although this cost may seem large because it involves the degrees of freedom of the high-fidelity discretization, note that the cost of computing one single time step with a classical solver is at least of order $\mathcal{O}(N^{(1+\alpha)})$ where $\alpha > 0$ depends on the specific matrix system to invert and the solver used to perform the inversion. As a result, when $N \gg 1$ and the number of calls is sufficiently large, one can expect significant gains in computing times.

2.3.3 Numerical Experiments

As a support for our tests, we consider in [A6] four different conservative PDEs in one space dimension:

- An inviscid and Burgers' equation for which we have explicit expressions of the solutions and some theoretical estimates of the performance of the method (see [A6]).
- A version with viscosity of the previous Burgers' equation.
- A Camassa Holm equation.
- A Korteveg de Vries equation.

For each PDE, we compare the performance of the four following model reduction methods:

- The classical linear PCA method in L^2,
- The tangent PCA method in W_2,
- The gBar method (with interpolation and projection) in W_2.

The performance is measured in terms of the average and worst case approximation error of a set on a discrete test set of 500 functions. Each test set is different from the training set $\mathcal{M}_{\text{tr}}$. The training set is composed of randomly generated snapshots. For every example, the number of training snapshots is $\#\mathcal{M}_{\text{tr}} = 5.10^3$ (see [A6] for a discussion on the impact of the number of training snapshots).

In addition to the error study, we also provide run time statistics but only for the case of the viscous Burgers' equation since it is the only example that involves a high-fidelity solver. In the case of the inviscid Burgers' equation and KdV, the exact solutions can be explicitly written down with formulas so we did not use a solver to which we can compare ourselves to. In the case of the Camassa Holm equation, the solution was nearly analytic too and we could not consider its numerical solution as a representative example which involves a solver.

The code to reproduce the numerical results is available online at:

https://github.com/olga-mula/2019-RBM-metric-spaces

For each PDE example, we also provide reconstruction videos of a full propagation on the same link. In this manuscript, we only report results for the viscous Burgers' equation for the sake of brevity.

Viscous Burgers' Equation: The considered equation is

$$u_t + \frac{1}{2}(u^2)_x - \nu \partial_x^2 u = 0, \quad \rho(0,x,y) = \begin{cases} 0, & -3 \leq x < 0 \\ h, & 0 \leq x < \frac{1}{h} \\ 0, & \frac{1}{h} \leq x \leq 5, \end{cases}$$

where the parameters are $h \in [0.5, 3]$, the viscosity term $\nu \in [5.10^{-5}, 0.1]$, and the time $t \in [0, 3]$. The parameter domain is thus

$$\mathrm{Y} = \{(t, y, \nu) \in [0, 3] \times [0.5, 3] \times [5.10^{-5}, 0.1]\}.$$

The spatial coordinate is x and it ranges in $\Omega = [-3, 5]$ in our computations.

Figure 2.13 gives the errors in average and worst case sense for the test set $\mathcal{M}_{\text{test}}$. If we first examine the errors in the natural norms (plots on the left), it appears that the errors in tPCA do not seem to decay significantly faster than in PCA. Also, the approach with barycenters does not seem to give a very good performance and seems to perform worse than PCA. In order to make a "fair" comparison, it is necessary to mesure errors in a common metric which is "fair" for all approaches and which also quantifies the potential numerical instabilities of tPCA. Since we are looking for metrics that quantify the quality of transport rather than spatial averages, we discard the L^2 metric in favor of the H^{-1} metric which can be seen as a relaxation of the W_2 distance. The space H^{-1} is taken here as the dual of H_0^1 and its norm computed accordingly. When we examine the errors in the unified H^{-1} metric, we see that all the nonlinear methods are clearly outperforming PCA. This is more in accordance with what we visually observe when we examine the reconstructed functions given by each method (see Figure 2.16). Note that in this particular example the tPCA presents a sharp unnatural spike at the propagation front, due to the above discussed stability issues of this method. This is in contrast to the approach with barycenters which does not suffer from this issue at the cost of slightly degrading the final approximation quality. Like for the other examples, the reader may watch videos of the reconstruction on the link above.

We next provide run time statistics for this test case. For any $u \in \mathcal{M}_{\text{tr}}$, let $r_{\text{HF}}(u)$, $r_{\text{PCA}}(u)$, $r_{\text{tPCA}}(u)$, $r_{\text{gBar}}(u)$ and $r_{\text{gBar}}^{\text{interp}}(u)$ be the respective run times of the high-fidelity solver, and of the PCA, tPCA, and gBar with projection and interpolation methods. The high-fidelity solver uses an explicit piecewise linear finite-volume method to evaluate the advective flux and then discretize the diffusion part implicitly (Crank-Nicolson) with the advective piece as a source to update in time. The resulting discretization is second-order in space and time. For each dynamic, the time step δt is fixed to be sufficiently small in order to satisfy a CFL condition. Figure 2.14 shows, as a function of the reduced dimension n, the average and the median of the ratios between the run time of a given method and the run time of the high-fidelity computation,

$$R_{\text{av}}^* = \frac{1}{\#\mathcal{M}_{\text{tr}}} \sum_{u \in \mathcal{M}_{\text{tr}}} \frac{r_*(u)}{r_{\text{HF}}(u)} \quad \text{and} \quad R_{\text{median}}^* = \text{median} \left\{ \frac{r_*(u)}{r_{\text{HF}}(u)} \ : \ u \in \mathcal{M}_{\text{tr}} \right\}.$$

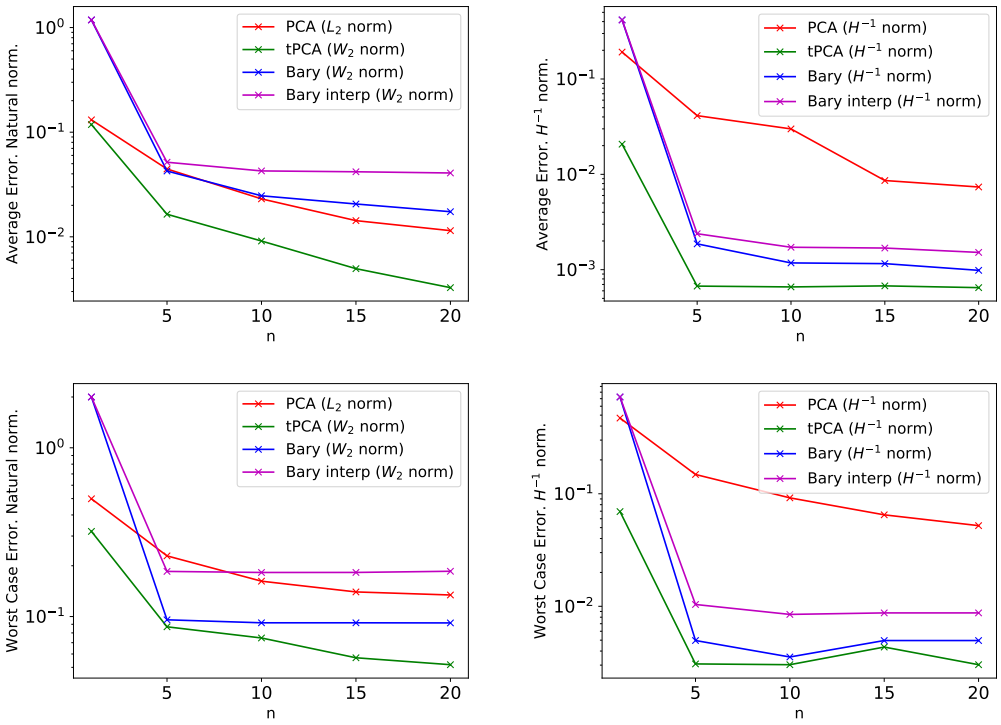


Figure 2.13: Errors on $\mathcal{M}_{\text{test}}$ for the ViscousBurgers equation. Top figures: average error. Bottom figures: worst case error. Left: natural norms. Right: H^{-1} norm.

The $*$ symbol denotes all the previous methods. The figure shows that the run time is reduced by a factor of about 100 in average and of about 500 in the median for all the methods. We observe that the classical PCA is slightly faster than tPCA and the gBar algorithm. This is essentially related to the fact that we need to compute exponential maps for the latter methods. We may also note that the run time remains constant with n: we think that this is due to the fact that n is pretty small and expect a mild increase for larger values of n. One last final observation for these plots is to remark that the gBar method with interpolation performs almost identically than the one with interpolation.

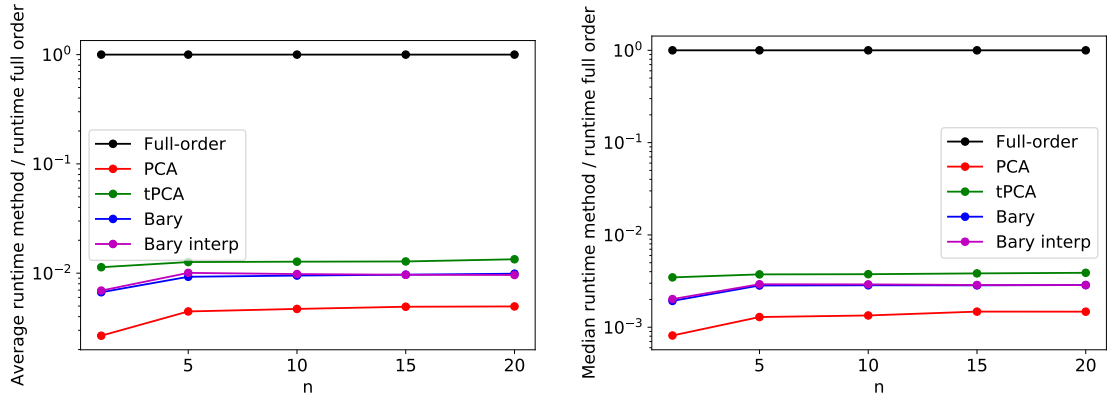


Figure 2.14: Run times as a function of the dimension n: Average R^*_{av} (left plot) and median R^*_{median} (right plot).

Since the time variable is treated as a parameter in our approach, it is interesting to compare run times with respect to t since we expect that the high-fidelity method will be faster for smaller times. Figure 2.15 shows the run times to compute each snapshot of $\mathcal{M}_{\text{tr}}$ as a function of its corresponding parameter t (we take $n = 20$). We observe that the reduced models are significantly faster, even for small values of t. As expected, the difference grows as t increases.

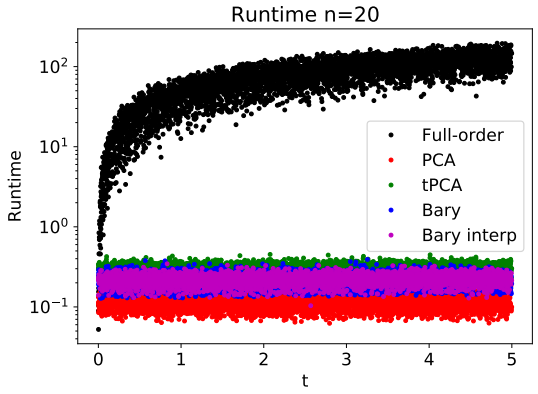


Figure 2.15: Run times to compute each snapshot of $\mathcal{M}_{\text{tr}}$ as a function of the parameter t ($n = 20$).

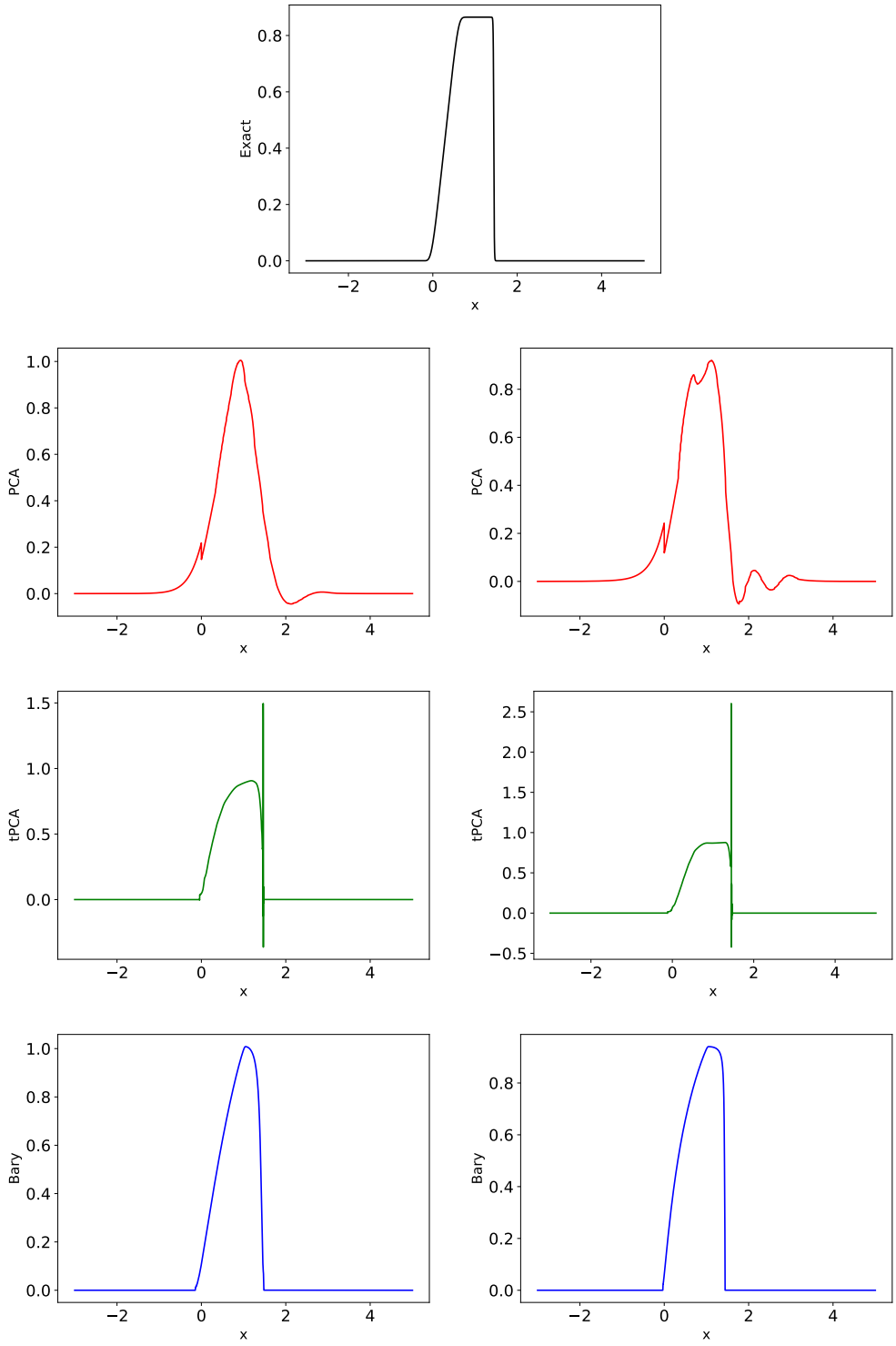


Figure 2.16: ViscousBurgers equation: Reconstruction of a function with $n = 5$ (left) and $n = 10$ (right). Black: exact function. Red: PCA. Green: tPCA. Blue: gBar.

2.3.4 Research Perspectives

There are numerous research perspectives connected to the question of model reduction for parametrized, transport dominated PDEs which I would like to explore in the future. We start by discussing some perspectives directly related to [A6] and then enlarge the scope of the discussion:

- **Enhancing the theoretical foundations:** In [A6] we were able to develop some theoretical statements in order to show the potential of the approach. However, the results were very pessimistic compared to what we observed in the numerical tests. Can we find better strategies of proof to give sharper bounds? And more importantly, in [48, 41] it was proven that greedy algorithms for Hilbert and Banach spaces had a similar rate of convergence as the Kolmogorov width. Can we relate the decay rate of the gBar algorithm to some adapted version of the Kolmogorov width for metric spaces?

- **Extension to higher dimensions:** As already mentionned above, I am working with H. Do and J. Feydy in the extension to dimension 2 and 3 of the same type of algorithm. We resort to the entropic regularized version of the Wasserstein distance in order to leverage the latest developments in numerical optimal transport to build online efficient strategies. The development requires a highly optimized GPU implementation which is currently in progress.

- **Nonlinear dimensionality reduction:** One can further refine the above strategy by adaptively choosing the barycentric functions U_n for each given y. This could be done by a parameter splitting strategy in the spirit of our approach in [S3] for nonlinear linear inverse state estimation, or by nonlinear dimensionality learning strategies such as Isomap (see e.g. [71]), which aims at finding locally low-dimensional isometric embeddings. This type of ideas are being tested with H. Do, J. Feydy. I am also engaged in another collaboration with Prof. D. Guignard on these topics.

- **Metric Learning, beyond conservation laws:** Since each instance of the entropic regularization parameter leads to a different metric, the optimization over these parameters will naturally lead to metric learning strategies. In addition, it will be possible to address certain families of non-conservative problems thanks to the unbalanced formulation of optimal transport (see e.g. [13]). This will also require learning the additional parameters required in this approach.

- **Foreseen limitations, thoughts on possible remedies, and further directions:**

  - **Classes of problems:** Although we expect that the above approach will significantly broaden the scope of problems that can efficiently be addressed in forward reduced modelling, it seems clear that the success of the above strategy is linked to the question as to whether the problem presents a somewhat "additive transport structure", where regions with high density are stretched or contracted. Problems that we think can possibly be well described by this property are Burgers' equations, KdV, Gross–Pitaevskii, Vlasov–Poisson equations (and possibly others). However, PDEs such as the wave equation are expected to pose problems and in general any PDE whose solution is not positive poses a challenge. Although the naive approach would be to add a sufficiently large constant to make the solutions positive, we do not expect that this would work because it may produce concentrations on regions with depressions. A possible strategy in this case would

be to search for a lifting in a higher dimensional space by appropriately transporting the graphs $(x, y, u(y)) \in \Omega \times \mathrm{Y} \times \mathcal{M}$.

- **Very high dimension:** Despite the development of efficient numerical approaches for the computation of optimal transport distances and barycenters, these computations still remain intractable when the dimension d of the problem becomes very high. One possible angle of attack in this context is to approximate the solutions with mixtures of Gaussians or mixtures of densities supported on polygons with simple shapes where we can ressort to closed forms for the Wasserstein distances. Since optimal transport plans between mixture models are usually not mixture models themselves, we could consider variants of the Wasserstein distance by restricting the set of possible coupling measures to certain mixture models as has recently been studied in [4].
- **How can nonlinear approximation methods coming from Machine Learning help to address the problem?** Given the tremendous impact that learning with Deep Neural Networks is having in a variety of applicative problems, it is natural to wonder how to leverage this type of approach to address the present topic of nonlinear model reduction of transport dominated PDEs. In the framework of Agustín Somacal's PhD thesis which I am co-supervising with Prof. A. Cohen, we are exploring several ideas in this direction.

Chapter 3

Inverse state and parameter estimation using reduced models

In this chapter I give an overview of my works on inverse problems:

- Sections 3.1 to 3.4 summarize the theory and algorithms of state and parameter estimation framework that my collaborators and I have built in [A10, A4, S3].
- Section 3.5 summarizes my contributions in applying the methodology to applications (neutronics [A9, P1, P2, 10], and biomedicine [A2, A3, S1]). I also include another application related to epidemiology forecasting in which the method is not directly built from the previous theoretical developments (see [A1]).
- Section 3.6 discusses future research directions in the field of inverse problems.

3.1 Optimality Benchmarks for State Estimation

Let V be a Hilbert space defined over a domain $\Omega \subseteq \mathbb{R}^d$ and equipped with some norm $\|\cdot\|$ and inner product $\langle \cdot, \cdot \rangle$. We want to recover an approximation to an unknown function $u \in V$ from data given by m linear measurements

$$\ell_i(u), \quad i = 1, \dots, m,$$

where the ℓ_i are m linearly independent linear functionals over V. This problem appears in many different settings. The particular one that motivates our work is the case where $u = u(y)$ represents the *state* of a physical system described as a solution to a parametric PDE

$$\mathcal{P}(u, y) = 0 \tag{3.1.1}$$

for some unknown parameter vector $y = (y_1, \dots, y_p)$ that belongs to some admissible set $Y \subseteq \mathbb{R}^p$. The dimension p of the parameter space can be finite or infinite. The ℓ_i are a mathematical model for sensors that capture some partial information on the unknown solution $u(y) \in V$.

Denoting by $\omega_i \in V$ the Riesz representers of the ℓ_i, such that $\ell_i(v) = \langle \omega_i, v \rangle$ for all $v \in V$, and defining

$$W := \operatorname{span}\{\omega_1, \dots, \omega_m\},$$

the measurements are equivalently represented by

$$w = P_W u,$$

where P_W is the orthogonal projection from V onto W. A *recovery algorithm* is a computable map

$$A : W \to V$$

and the approximation to u obtained by this algorithm is

$$u^* = A(w) = A(P_W u).$$

The construction of A should be based on the available prior information that describes the properties of the unknown u, and the evaluation of its performance needs to be defined in some precise sense. Two distinct approaches are usually followed:

- In the *deterministic setting*, the sole prior information is that u belongs to the set

$$\mathcal{M} := \{u(y) \ : \ y \in Y\}, \tag{3.1.2}$$

  of all possible solutions. The set $\mathcal{M}$ is sometimes called the *solution manifold*. The performance of an algorithm A over the class $\mathcal{M}$ is usually measured by the "worst case" reconstruction error

$$E_{\rm wc}(A, \mathcal{M}) = \sup\{\|u - A(P_W u)\| \ : \ u \in \mathcal{M}\}. \tag{3.1.3}$$

  The problem of finding an algorithm that minimizes $E_{\rm wc}(A)$ is called *optimal recovery*. It has been extensively studied for convex sets $\mathcal{M}$ that are balls of smoothness classes [87, 92, 62], which is not the case for (3.1.2).

- In the *stochastic setting*, the prior information on u is described by a probability distribution p on V, which is supported on $\mathcal{M}$, typically induced by a probability distribution on Y that is assumed to be known. It is then natural to measure the performance of an algorithm in an averaged sense, for example through the mean-square error

$$E_{\rm ms}(A, p) = \mathbb{E}(\|u - A(P_W u)\|^2) = \int_V \|u - A(P_W u)\|^2 dp(u). \tag{3.1.4}$$

  This stochastic setting is the starting point for *Bayesian estimation* methods [23]. Let us observe that for any algorithm A one has $E_{\rm ms}(A, p) \leq E_{\rm wc}(A, \mathcal{M})^2$.

Note that, of course, in full generality, the measurement observations are noisy and the target function u may not belong to $\mathcal{M}$ due to model error. For the sake of clarity, we will carry the discussion by placing ourselves in the ideal case of no noise and no model error, and we will discuss how to analyze noise and model error later on in the manuscript.

My contributions concentrate mostly on the deterministic setting according to the above distinction, although we gave some remarks on the analogies with the stochastic setting in [A4]. In the deterministic setting, the performance benchmark of recovery algorithms is given by

$$E^*_{\rm wc}(\mathcal{M}) = \inf_{A: W \to V} E_{\rm wc}(A, \mathcal{M}),$$

where the infimum is taken over all possible maps $A : W \to V$.

There is a simple mathematical description of an optimal map that meets this benchmark. To define it, we note that in the absence of model bias and when a noiseless measurement $w = P_W u$ is given, our knowledge on u is that it belongs to the set

$$\mathcal{M}_w := \mathcal{M} \cap (\omega + W^\perp). \tag{3.1.5}$$

The best possible recovery map can be described through the following general notion.

Definition 1. The Chebychev ball of a bounded set $S \in V$ is the closed ball $B(v, r)$ of minimal radius that contains S. One denotes by $v = \mathrm{cen}(S)$ the Chebychev center of S and $r = \mathrm{rad}(S)$ its Chebychev radius.

In particular one has

$$\frac{1}{2}\mathrm{diam}(S) \leq \mathrm{rad}(S) \leq \mathrm{diam}(S), \tag{3.1.6}$$

where $\mathrm{diam}(S) := \sup\{\|u - v\| \ : \ u, v \in S\}$ is the diameter of S. Therefore, the recovery map that minimizes the worst case error over $\mathcal{M}_w$ for any given w, and therefore over $\mathcal{M}$ is defined by

$$A^*_{\mathrm{wc}}(w) = \mathrm{cen}(\mathcal{M}_w). \tag{3.1.7}$$

Its associated worst case error is

$$E^*_{\mathrm{wc}}(\mathcal{M}) = \sup\{\mathrm{rad}(\mathcal{M}_w) \ : \ w \in W\}.$$

Note that the map A^*_{wc} is also optimal among all algorithms for each $\mathcal{M}_w$, where $w \in P_W(\mathcal{M})$, since

$$E_{\mathrm{wc}}(A^*_{\mathrm{wc}}, \mathcal{M}_w) = \min_A E_{\mathrm{wc}}(A, \mathcal{M}_w) = \mathrm{rad}(\mathcal{M}_w), \quad \forall w \in P_W(\mathcal{M}).$$

However, there may exist other maps A such that $E_{\mathrm{wc}}(A, \mathcal{M}) = E^*_{\mathrm{wc}}(\mathcal{M})$, since we also supremize over $w \in P_W(\mathcal{M})$.

In view of the equivalence (3.1.6), we can relate $E^*_{\mathrm{wc}}(\mathcal{M})$ to the quantity

$$\delta_0 = \delta_0(\mathcal{M}, W) := \sup\{\mathrm{diam}(\mathcal{M}_w) \ : \ w \in W\} = \sup\{\|u - v\| \ : \ u, v \in \mathcal{M}, \ u - v \in W^\perp\}, \tag{3.1.8}$$

by the equivalence

$$\frac{1}{2}\delta_0 \leq E^*_{\mathrm{wc}}(\mathcal{M}) \leq \delta_0. \tag{3.1.9}$$

Note that injectivity of the measurement map P_W over $\mathcal{M}$ is equivalent to $\delta_0 = 0$. We provide in Figure (3.6a) an illustration of the mapping A^*_{wc} and of the above benchmark concepts.

If $w = P_W u$ for some $u \in \mathcal{M}$, then any $u^* \in \mathcal{M}$ such that $P_W u^* = w$, meets the ideal benchmark $\|u - u^*\| \leq \delta_0$. Therefore, one way of finding such a u^* would be to minimize the distance to the manifold over all functions $v \in V$ such that $P_W v = w$, that is, to solve

$$\min_{v \in \omega + W^\perp} \mathrm{dist}(v, \mathcal{M}) = \min_{v \in \omega + W^\perp} \min_{y \in Y} \|u(y) - v\|.$$

This problem is computationally out of reach since it amounts to the nested minimization of two non-convex functions in high dimension, and motivates the search for suboptimal, yet computationally feasible algorithms. We discuss our contributions on this front in Sections 3.2, to 3.4. Section 3.5 will be devoted to applications to which we have brought the developed methodology.

3.2 Optimal Affine Algorithms

3.2.1 Definition and preliminary remarks

In practice the above map A^*_{wc} cannot be easily constructed due to the fact that the solution manifold $\mathcal{M}$ is a high-dimensional and geometrically complex object. One is therefore interested in designing "sub-optimal yet good" recovery algorithms and analyze their performance. One possibility in this direction is to restrict the search to linear recovery mappings $A \in \mathcal{L}(W, V)$. One vehicle for constructing them is to use *reduced modeling*.

Generally speaking, reduced models consist of linear spaces $(V_n)_{n\geq 0}$ with increasing dimension $\dim(V_n) = n$ which uniformly approximate the solution manifold in the sense that

$$\mathrm{dist}(\mathcal{M}, V_n) \coloneqq \max_{u\in\mathcal{M}} \|u - P_{V_n} u\| \leq \varepsilon_n, \tag{3.2.1}$$

where

$$\varepsilon_0 \geq \varepsilon_1 \geq \cdots \geq \varepsilon_n \geq \cdots \geq 0,$$

are known tolerances. Instances of reduced models for parametrized families of PDEs with provable accuracy are provided by polynomial approximations in the y variable [31, 50] or reduced bases [44, 67]. The construction of a reduced model is typically done offline, using a large training set of instances of $u \in \mathcal{M}$ called *snapshots*. The offline stage potentially has a high computational cost. Once this is done, the online cost of recovering $u^* = A(w)$ from any data w using this reduced model should in contrast be moderate.

In [33], a simple reduced-model based recovery algorithm was proposed. Assuming that we have a reduced model V_n, the algorithm, called Parametrized Background Data-Weak (PBDW), is defined in terms of the map

$$A_n(w) := \mathrm{argmin}\{\mathrm{dist}(v, V_n) \ : \ v \in \omega + W^\perp\}, \tag{3.2.2}$$

which is well defined provided that $V_n \cap W^\perp = \{0\}$. A necessary (but not sufficient) condition to guarantee this is to have $n \leq m$, which we will assume in the following. As a side-remark, it is interesting to note that the reconstruction algorithm (3.2.2) was proposed simultaneously in the field of model order reduction and by researchers seeking to build infinite dimensional generalizations of compressed sensing (see [38]). In the applications of this community, V_n is usually chosen to be a "multi-purpose" basis such as the Fourier basis, as opposed to our current envisaged applications in which V_n is a subspace specifically tailored to approximate $\mathcal{M}$. The analysis that follows remains however valid also for these types of "multi-purpose" spaces.

We can easily prove that A_n is a linear mapping and it was shown in [19] that A_n has a simple interpretation in terms of the cylinder

$$\mathcal{K}_n := \{v \in V \ : \ \mathrm{dist}(v, V_n) \leq \varepsilon_n\}, \tag{3.2.3}$$

that contains the solution manifold $\mathcal{M}$. Namely, the algorithm A_n is also given by

$$A_n(w) = \mathrm{cen}(\mathcal{K}_{n,w}), \quad \mathcal{K}_{n,w} := \mathcal{K}_n \cap (\omega + W^\perp),$$

and the map is shown to be the optimal when $\mathcal{M}$ is replaced by the simpler containment set $\mathcal{K}_n$, that is

$$A_n = \operatorname*{arg\,min}_{A:W\to V} E_{\mathrm{wc}}(A, \mathcal{K}_n).$$

The substantial advantage of this approach is that, in contrast to A^*_{wc}, the map A_n can be easily computed by solving a simple least-squares minimization problem of size $n \times m$. Note that A_n depends on V_n and W, but not on ε_n in view of (3.2.2). This is important because ε_n is only known approximately in practice.

This algorithm satisfies the performance bound

$$\|u - A_n(P_W u)\| \leq \mu_n \mathrm{dist}(u, V_n \oplus (V_n^\perp \cap W)) \leq \mu_n \mathrm{dist}(u, V_n) \leq \mu_n \varepsilon_n, \tag{3.2.4}$$

where the last inequality holds when $u \in \mathcal{M}$. Here

$$\mu_n = \mu(V_n, W) := \max_{v \in V_n} \frac{\|v\|}{\|P_W v\|}$$

is the inverse of the inf-sup constant

$$\beta_n := \min_{v \in V_n} \max_{w \in W} \frac{\langle v, w \rangle}{\|v\| \, \|w\|},$$

which describes the angle between V_n and W. In particular $\mu_n = \infty$ in the event where $V_n \cap W^\perp$ is non-trivial.

An important observation is that the PBDW algorithm (3.2.2) has a simple extension to the setting where V_n is an affine space rather than a linear space, namely, when

$$V_n^{(\mathrm{aff})} = \bar{u} + V_n, \tag{3.2.5}$$

with V_n a linear subspace of dimension n and $\bar{u}$ a given offset that is known to us. In this case, denoting

$$\bar{\omega} := P_W \bar{u},$$

the affine version of (3.2.2) reads

$$A_n^{(\mathrm{aff})}(w) := \arg\min\{\mathrm{dist}(v, \bar{u} + V_n) \ : \ v \in \omega + W^\perp\}, \tag{3.2.6}$$

which can also be written as

$$A_n^{(\mathrm{aff})}(w) = \bar{u} + A_n(\omega - \bar{\omega}).$$

At first sight, affine spaces do not bring any significant improvement in terms of approximating the solution manifold, due to the following elementary observation: if $\mathcal{M}$ is approximated with accuracy ε by an n-dimensional affine space V_n given by (3.2.5), it is also approximated with accuracy $\widetilde{\varepsilon} \leq \varepsilon$ by the $n+1$-dimensional linear space

$$\widetilde{V}_{n+1} := V_n \oplus \mathbb{R}\bar{u}.$$

However, the choice of an affine subspace may significantly improve the performance of the algorithm (3.2.2) in the case where the parametric solution $u(y)$ is a "small perturbation" of a nominal solution $\bar{u} = u(\bar{y})$ for some $\bar{y} \in Y$, in the sense that

$$\mathrm{diam}(\mathcal{M}) \ll \|u\|.$$

Indeed, suppose in addition that $\bar{u}$ is badly aligned with respect to the measurement space W in the sense that

$$\|P_W \bar{u}\| \ll \|u\|.$$

In such a case, any linear space V_n that is well tailored to approximating the solution manifold (for example a reduced basis space) will contain a direction close to that of $\bar{u}$ and thus, we will have that $\mu_n \gg 1$, rendering the reconstruction by the linear PBDW method much less accurate than the approximation error by V_n. The use of the affine mapping (3.2.5) has the advantage of elimitating the bad direction $\bar{u}$ since μ_n will now be computed with respect to the linear part $\widetilde{V}_n$.

The standard constructions of reduced models are targeted at making the spaces V_n as efficient as possible for approximating $\mathcal{M}$, that is, making ε_n as small as possible for each given n. For example, for the reduced basis spaces, it is known [48, 41] that a certain greedy selection of snapshots generates spaces V_n such that $\operatorname{dist}(\mathcal{M}, V_n)$ decays at the same rate (polynomial or exponential) as the Kolmogorov n-width

$$d_n(\mathcal{M}) := \inf\{ \operatorname{dist}(\mathcal{M}, E) \ : \ \dim(E) = n\}. \tag{3.2.7}$$

However these constructions do not ensure the control of μ_n and therefore these reduced spaces may be much less efficient when using the PBDW algorithm for the recovery problem.

In view of the above observations, two main strategies are possible. First, we can build affine spaces V_n that are better targeted towards the recovery task. In other words, we want to build the spaces V_n to make the recovery algorithm A_n as efficient as possible, given the measurement space W. This was the topic of our work [A4] which I summarize in Sections 3.2.2 and 3.2.3, and where we develop an implementable method to find the optimal affine subspace for the reconstruction task. A second strategy can be considered if we are allowed to select the measurement functionals ℓ_i from some admissible dictionary. This amounts to optimizing the space W and was the topic of our work [A10] which I summarize in Section 3.3. In Section 3.4 we present a strategy that goes beyond linear and affine algorithms.

3.2.2 Characterization of Affine Algorithms

The goal of our contribution [A4] was to characterize the best affine subspace V_n to apply the PBDW algorithm (3.2.6), and to develop an implementable strategy to find it. Here, we consider our measurement system to be imposed on us, and therefore W is fixed.

It turns out that searching for the best affine subspace V_n for the PBDW algorithm (3.2.6) is equivalent to searching for the best affine reconstruction map $A^*_{\text{aff}} : W \to V$ defined as

$$A^*_{\text{aff}} \ \in \ \operatorname*{arg\,min}_{\substack{A:W\to V \\ A \text{ affine}}} E_{\text{wc}}(A, \mathcal{M}), \tag{3.2.8}$$

where the existence of the minimum is guaranteed under very mild assumptions as we outline next. The algorithm A^*_{aff} reaches the performance among all affine algorithms, namely,

$$E_{\text{wc}}(A^*_{\text{aff}}, \mathcal{M}) = E^*_{\text{wca}}(\mathcal{M}) := \min_{\substack{A:W\to V \\ A \text{ affine}}} E_{\text{wc}}(A, \mathcal{M}). \tag{3.2.9}$$

Note that we wrote $E^*_{\text{wca}}(\mathcal{M})$ with the subindex "wca" to indicate that it is the optimal performance in the worst case sense among all affine maps. Obviously, $E^*_{\text{wca}}(\mathcal{M}) \geq E^*_{\text{wc}}(\mathcal{M})$ since $E^*_{\text{wc}}(\mathcal{M})$ is the optimal performance in the worst case among all maps (affine and nonlinear) as defined in (3.2.9).

We next characterize A^*_{aff}. In order to do this, as a first observation, note that since we are given the measurement observation w, any algorithm A which is a candidate to optimality must satisfy

$P_W(A(w)) = w$ (otherwise the reconstruction error would not be minimized). Thus a necessary condition for optimality is that A should have the form

$$A(w) = w + B(w), \tag{3.2.10}$$

where $B : W \to W^\perp$ with $W^\perp$ the orthogonal complement of W in V. Therefore, in going further, we always require that A has the above form (3.2.10) and concentrate on the construction of good affine maps B.

Our next observation is that any affine algorithm A of the form (3.2.10) can always be interpreted as a PBDW algorithm A_n for a certain space V_n with $n \leq m$.

Lemma 3.2.1 (See [A4])**.** *A is an affine map of the form (3.2.10) if and only if there there exists $\bar{u} \in V$ and a linear subspace V_n of dimension $n \leq m$ such that A coincides with the affine PBDW algorithm (3.2.6) for $V_n^{(aff)} = \bar{u} + V_n$.*

In view of this result, the search for an affine reduced model $\bar{u} + V_n$ that is best tailored to the recovery problem is equivalent to the search of an optimal affine map. Our next result is that such an optimal map always exist when $\mathcal{M}$ is a bounded set.

Theorem 3.2.2. *Let $\mathcal{M}$ be a bounded set. Then there exists a map A^*_{wca} that minimizes $E_{\text{wc}}(A, \mathcal{M})$ among all affine maps A.*

3.2.3 A practical algorithm for optimal affine recovery and some numerical tests

Discretization and truncation: Since we are searching among algorithms of the form (3.2.10), we have that

$$\begin{aligned} E_{\text{wc}}(A^*_{\text{aff}}, \mathcal{M}) &= \min_{A: W \to V \text{ affine}} \max_{u \in \mathcal{M}} ||u - A(\omega)|| \\ &= \min_{B: W \to W^\perp \text{ affine}} \max_{u \in \mathcal{M}} ||u - \omega - B(\omega)|| \\ &= \min_{c \in W^\perp,\, B: W \to W^\perp \text{ linear}} \max_{u \in \mathcal{M}} ||P_{W^\perp} u - c - B(\omega)||. \end{aligned}$$

This means that the optimal affine recovery map is obtained by minimizing the convex function

$$F(c, B) = \max_{u \in \mathcal{M}} \|P_{W^\perp} u - c - B(P_W u)\|,$$

over $W^\perp \times \mathcal{L}(W, W^\perp)$. This optimization problem cannot be solved exactly for two reasons:

1. The sets $W^\perp$ as well as $\mathcal{L}(W, W^\perp)$ are infinite dimensional when V is infinite dimensional.
2. One single evaluation of $F(c, B)$ requires in principle to explore the entire manifold $\mathcal{M}$.

The first difficulty is solved by replacing V by a subspace Z_N of finite dimension $\dim(Z_N) = N$ that approximates the solution manifold $\mathcal{M}$ with an accuracy of smaller order than that expected for the recovery error. One possibility is to use a finite element space $Z_N = V_h$ of sufficiently small mesh size h. However its resulting dimension $N = N(h)$ needed to reach the accuracy could still be quite large. An alternative is to use reduced model spaces Z_N which are more efficient for the approximation of $\mathcal{M}$, as we discuss further.

We therefore minimize $F(c,B)$ over $\widetilde{W}^\perp \times \mathcal{L}(W,\widetilde{W}^\perp)$, where $\widetilde{W}^\perp$ is the orthogonal complement of W in the space $W+Z_N$, and obtain an affine map $\widetilde{A}_{\rm wca}$ defined by

$$\widetilde{A}_{\rm wca}(w) = w + \bar{c} + \bar{B}w, \quad (\bar{c},\bar{B}) := \arg\min\{\widetilde{F}(c,B) : \ c \in \widetilde{W}^\perp, B \in \mathcal{L}(W,\widetilde{W}^\perp)\}.$$

with

$$\widetilde{F}(c,B) = \max_{u\in\widetilde{\mathcal{M}}} \|P_{W^\perp}u - c - B(P_W u)\|.$$

In order to compare the performance of $\widetilde{A}_{\rm wca}(w)$ with that of $A^*_{\rm wca}$, we first observe that

$$\|P_{W^\perp}u - P_{\widetilde{W}^\perp}u\| \le \varepsilon_N := \sup_{u\in\mathcal{M}} \operatorname{dist}(u,Z_N).$$

For any $(c,B) \in W^\perp \times \mathcal{L}(W,W^\perp)$, we define $(\widetilde{c},\widetilde{B}) \in \widetilde{W}^\perp \times \mathcal{L}(W,\widetilde{W}^\perp)$ by $\widetilde{c} = P_{\widetilde{W}^\perp}c$ and $\widetilde{B} = P_{\widetilde{W}^\perp} \circ B$. Then, for any $u \in \mathcal{M}$,

$$\begin{aligned}\|P_{W^\perp}u - \widetilde{c} - \widetilde{B}(P_W u)\| &\le \|P_{\widetilde{W}^\perp}(P_{W^\perp}u - c - B(P_W u))\| + \|P_{W^\perp}u - P_{\widetilde{W}^\perp}u\| \\ &\le \|P_{W^\perp}u - c - BP_W u\| + \varepsilon_N.\end{aligned}$$

It follows that we have the framing

$$E(A^*_{\rm wca},\mathcal{M}) \le E(\widetilde{A}_{\rm wca},\mathcal{M}) \le E(A^*_{\rm wca},\mathcal{M}) + \varepsilon_N, \tag{3.2.11}$$

which shows that the loss in the recovery error is at most of the order ε_N.

To understand how large N should be, let us observe that a recovery map A of the form (3.2.10) takes it value in the linear space

$$F_{m+1} = \mathbb{R}c + \operatorname{range}(B),$$

which has dimension $m+1$. It follows that the recovery error is always larger than the approximation error by such a space. Therefore

$$E_{\rm wc}(A^*_{\rm wca},\mathcal{M}) \ge d_{m+1}(\mathcal{M}),$$

where $d_{m+1}(\mathcal{M})$ is the Kolmogorov n-width defined by (3.2.7) for $n = m+1$. Therefore, if we could use the space Z_n that exactly achieves the infimum in (3.2.7), we would be ensured that, with $N = m+1$, the additional error $\varepsilon_N = \delta_{m+1}(\mathcal{M})$ in (3.2.11) is of smaller order than $E_{\rm wc}(A^*_{\rm wca},\mathcal{M})$. As a result we would obtain the framing

$$E(A^*_{\rm wca},\mathcal{M}) \le E(\widetilde{A}_{\rm wca},\mathcal{M}) \le 2E(A^*_{\rm wca},\mathcal{M}).$$

In practice, since we do not have access to the n-width spaces, we use instead the reduced basis spaces $Z_n := V_n$ which are expected to have comparable approximation performances in view of the results from [48, 41]. We take N larger than m but of comparable order.

The second difficulty is solved by replacing the set $\mathcal{M}$ in the supremum that defines $F(c,B)$ by a discrete training set $\widetilde{\mathcal{M}}$, which corresponds to a discretization $\widetilde{Y}$ of the parameter domain Y, that is

$$\widetilde{\mathcal{M}} := \{u(y) \ : \ y \in \widetilde{Y}\},$$

with finite cardinality.

We therefore minimize over $\widetilde{W}^\perp \times \mathcal{L}(W, \widetilde{W}^\perp)$ the function

$$\widetilde{F}(c, B) = \sup_{u \in \widetilde{\mathcal{M}}} \|P_{W^\perp} u - c - B P_W u\|,$$

which is computable. The additional error resulting from this discretization can be controlled from the resolution of the discretization. Namely, let $\varepsilon > 0$ be the smallest value such that $\widetilde{\mathcal{M}}$ is an ε-approximation net of $\mathcal{M}$, that is, $\mathcal{M}$ is covered by the balls $B(u, \varepsilon)$ for $u \in \widetilde{\mathcal{M}}$. Then, we find that

$$\widetilde{F}(c, B) \le F(c, B) \le \widetilde{F}(c, B) + \varepsilon \|B\|_{\mathcal{L}(W, \widetilde{W}^\perp)},$$

which shows that the additional recovery error will be of the order of ε amplified by the norm of the linear part of the optimal recovery map.

One difficulty is that the cardinality of ε-approximation nets become potentially untractable for small ε as the parameter dimension becomes large, due to the curse of dimensionality. This difficulty also occurs when constructing reduced basis by a greedy selection process which also needs to be performed in a sufficiently dense discretized sets. Recent results obtained in [14] show that, in certain relevant instances, ε-approximation nets can be replaced by random training sets of smaller cardinality. One interesting direction for further research is to apply similar ideas in the context of the present work.

Optimization algorithms: As already brought up, the practical computation of $\widetilde{A}_{\mathrm{wc}}$ consists in solving

$$\min_{(c,B) \in \widetilde{W}^\perp \times \mathcal{L}(W, \widetilde{W}^\perp)} \underbrace{\max_{u \in \widetilde{\mathcal{M}}} \|P_{W^\perp} u - c - B P_W u\|^2}_{= \widetilde{F}(c,B)}, \tag{3.2.12}$$

The numerical solution of this problem is challenging due to its lack of smoothness (the objective function $\widetilde{F}$ is convex but non differentiable) and its high dimensionality (for a given target accuracy ε_N, the cardinality of $\widetilde{\mathcal{M}}$ might be large). One could use classical subgradient methods, which are simple to implement. However these schemes only guarantee a very slow $O(k^{-1/2})$ convergence rate of the objective function, where k is the number of iterations. This approach did not give satisfactory results in our case: due to the slow convergence, the solution update of one iteration falls below machine precision before approaching the minimum close enough, see Figure 3.1. This has motivated the use of a primal-dual splitting method which is known to ensure a $O(1/k)$ convergence rate on the partial duality gap. We next briefly describe this method.

We assume without loss of generality that $\dim(W + V_N) = m + N$ and that $\dim \widetilde{W}^\perp = N$. Let $\{\psi_i\}_{i=1}^{m+N}$ be an orthonormal basis of $W + V_N$ such that $W = \operatorname{span}\{\psi_1, \dots, \psi_m\}$. Since for any $u \in V$,

$$P_{W+V_N} u = \sum_{i=1}^{m+N} u_i \psi_i,$$

the components of u in W can be given in terms of the vector $\mathbf{w} = (u_i)_{i=1}^m$ and the ones in $\widetilde{W}^\perp$ with $\mathbf{u} = (u_{i+m})_{i=1}^N$.

We now consider the finite training set

$$\widetilde{\mathcal{M}} := \{u^1, \dots, u^J\}, \quad J := \#(\widetilde{\mathcal{M}}) < \infty,$$

and denote by $\mathbf{w}^j$ and $\mathbf{u}^j$ the vectors associated to the snapshot functions u^j for $j = 1, \dots, J$. One may express the problem (3.2.12) as the search for

$$\min_{\substack{(\mathbf{R},\mathbf{b})\in \\ \mathbb{R}^{N\times m}\times\mathbb{R}^N}} \max_{j=1,\dots,J} \|\mathbf{u}^j - \mathbb{R}\mathbf{w}^j - \mathbf{b}\|_2^2.$$

Concatenating the matrix and vector variables $(\mathbf{R}, \mathbf{b})$ into a single $\mathbf{x} \in \mathbb{R}^{m(N+1)}$, we rewrite the above problem as

$$\min_{\mathbf{x}\in\mathbb{R}^{m(N+1)}} \max_{j=1,\dots,J} f_j(\mathbf{Q}_j\mathbf{x}), \tag{3.2.13}$$

where $\mathbf{Q}_j \in \mathbb{R}^{N\times m(N+1)}$ is a sparse matrix built using the coefficients of $\mathbf{w}^j$ and $f_j(\mathbf{y}) := \|\mathbf{u}^j - \mathbf{y}\|_2^2$.

The key observation to build our algorithm is that problem (3.2.13) can be equivalently written as a minimization problem on the epigraphs, i.e.,

$$\begin{aligned} &\min_{(\mathbf{x},t)\in\mathbb{R}^{m(N+1)}\times\mathbb{R}^+} t \quad \text{subject to} \quad f_j(\mathbf{Q}_j\mathbf{x}) \le t, \quad j = 1, \dots, J \\ \iff &\min_{(\mathbf{x},t)\in\mathbb{R}^{m(N+1)}\times\mathbb{R}^+} t \quad \text{subject to} \quad (\mathbf{Q}_j\mathbf{x}, t) \in \mathrm{epi}_{f_j}, \quad j = 1, \dots, J, \end{aligned}$$

or, in a more compact (and implicit) form,

$$\min_{(\mathbf{x},t)\in\mathbb{R}^{m(N+1)}\times\mathbb{R}^+} t + \sum_{j=1}^{J} \iota_{\mathrm{epi}_{f_j}}(\mathbf{Q}_j\mathbf{x}, t).$$

where, for any non-empty set S the indicator function ι_S has value 0 on S and $+\infty$ on S^c.

This problem takes the following canonical expression, which is amenable to a primal-dual proximal splitting algorithm

$$\min_{(\mathbf{x},t)\in\mathbb{R}^{m(N+1)}\times\mathbb{R}} G(\mathbf{x}, t) + F \circ L(\mathbf{x}, t).$$

Here, G is the projection map for the second variable

$$G(\mathbf{x}, t) = t,$$

the linear operator L is defined by

$$L(\mathbf{x}, t) := ((\mathbf{Q}_1\mathbf{x}, t), (\mathbf{Q}_2\mathbf{x}, t), \cdots, (\mathbf{Q}_J\mathbf{x}, t))$$

and acts from $\mathbb{R}^{m(N+1)} \times \mathbb{R}$ to $\times_{j=1}^{J}(\mathbb{R}^N \times \mathbb{R})$ and the function F acting from $\times_{j=1}^{J}(\mathbb{R}^N \times \mathbb{R})$ to $\mathbb{R}$ is defined by

$$F((\mathbf{v}_1, t_1), \cdots, (\mathbf{v}_J, t_J)) := \sum_{j=1}^{J} \iota_{\mathrm{epi}_{f_j}}(\mathbf{v}_j, t_j).$$

Note that F is the indicator function of the cartesian product of epigraphs.

Before introducing the primal-dual algorithm, some remarks are in order:

1. We recall that if ϕ is a proper closed convex function on $\mathbb{R}^d$, its proximal mapping prox_ϕ is defined by
$$\mathrm{prox}_\phi(y) = \mathrm{argmin}_{\mathbb{R}^d}\left(\phi(x) + \frac{1}{2}\|x - y\|_2^2\right).$$

2. The adjoint operator L^* is given by

$$L^*\left((\mathbf{v}_1,t_1),\cdots,(\mathbf{v}_J,t_J)\right) := \left(\sum_{j=1}^{J}\mathbf{Q}_j^T\mathbf{v}_j, \sum_{j=1}^{J}t_j\right). \tag{3.2.14}$$

It can be easily shown that the operator norm of L satisfies $\|L\|^2 \leq J + \sum_{j=1}^{J}\|\mathbf{Q}_j\|^2$.

3. Both G and F are simple functions in the sense that their proximal mappings, prox_G and prox_F, can be computed in closed form.

The iterations of our primal-dual splitting method read for $k \geq 0$,

$$\begin{aligned}
(\mathbf{x},t)^{k+1} &= \text{prox}_{\gamma_G G}\left((\mathbf{x},t)^k - \gamma_G L^*\left(((\mathbf{v}_1,\xi_1),\ldots,(\mathbf{v}_J,\xi_J))^k\right)\right),\\
(\bar{\mathbf{x}},\bar{t})^{k+1} &= (\mathbf{x},t)^{k+1} + \theta\left((\mathbf{x},t)^{k+1} - (\mathbf{x},t)^k\right),\\
((\mathbf{v}_1,\xi_1),\ldots,(\mathbf{v}_J,\xi_J))^{k+1} &= \text{prox}_{\gamma_F \hat{F}}\left(((\mathbf{v}_1,\xi_1),\ldots,(\mathbf{v}_J,\xi_J))^k + \gamma_F L(\bar{\mathbf{x}},\bar{t})^{k+1}\right),
\end{aligned}$$

where $\hat{F}$ is the Fenchel-Legendre transform of F, $\gamma_G > 0$ and $\gamma_F > 0$ are such that $\gamma_G\gamma_F < 1/\|L\|^2$, and $\theta \in [-1,+\infty[$ (it is generally set to $\theta = 1$ as in [49]).

To illustrate the relevance of this algorithm for our purposes, we compare its performance with a standard subgradient method. Figure 3.1 plots the convergence history of the objective function across the iterations of both optimization methods in the example described in the next page ($m = 40$, $N = 110$ and $J = 10^3$). Two different reconstruction maps have been considered as starting guesses: $A(w) = w = P_W u$, and the PBDW algorithm A_{n^*} based on reduced basis spaces V_n with an optimal choice n^* for n. The convergence plot shows the superiority of the primal-dual method which converges to the same minimal value of the objective function after 10^5 iterations regardless of the intialization, while the subgradient method fails to reach it since its increments fall below machine precision.

For the same numerical example described next, we vary m and consider $m = 10$, 20, 30, 40, 50. Figure 3.2 gives the convergence of the reconstruction error over the training set $\widetilde{\mathcal{M}}$ across the primal-dual iterations (for simplicity, we took P_{W_m} as the starting guess for $A^{(m)}_{\text{wca}}$). To make sure that we reach convergence, we performed 10^6 iterations for each case. As expected, we observe in this figure that the final value of the objective function decreases as we increase the value of m (the reconstruction error decreases as we increase the number of measurements).

Numerical example: The example under consideration is the elliptic problem

$$\begin{cases} -\text{div}\left(a(y)\nabla u\right) &= f, \quad x \in D\\ u(x) &= 0, \quad x \in \partial D \end{cases} \tag{3.2.15}$$

on the unit square $D =]0,1[^2$, with a certain parameter dependence in the field a. More precisely, for a given $p \geq 1$, we consider "checkerboard" random fields where $a(y)$ is piecewise constant on a $p \times p$ subdivision of the unit-square.

$$D = \bigcup_{i,j=0}^{p-1} S_{i,j},$$

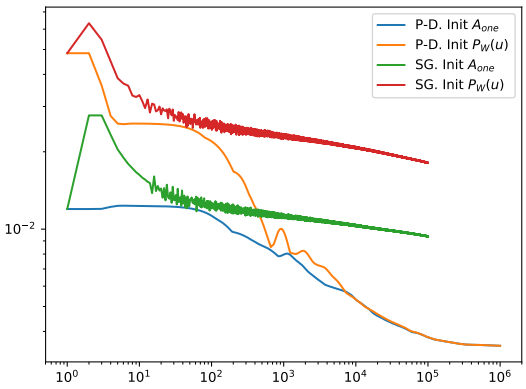


Figure 3.1: Convergence of the objective function for two different optimization algorithms and starting guesses. P.D. = Our Primal-Dual splitting method (worked well). S.G.=Subgradient (struggled to converge). Here, $m = 40$.

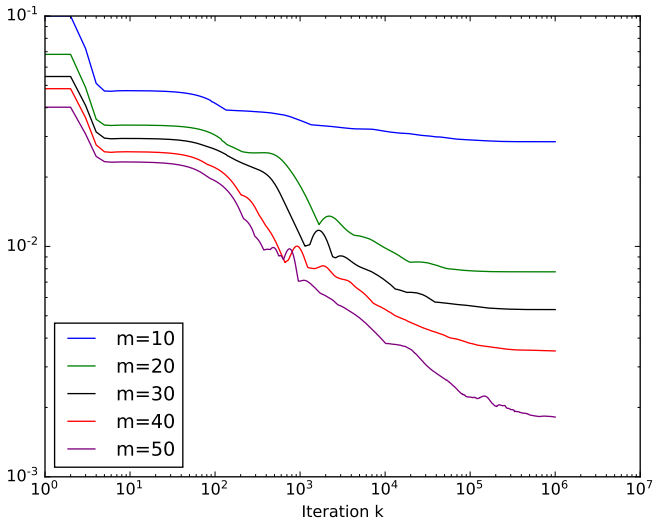


Figure 3.2: Convergence of the objective function in the primal-dual iterations for $m = 10$, 20, 30, 40, 50.

with

$$S_{i,j} := \left[\frac{i}{p}, \frac{i+1}{p}\right[\times \left[\frac{j}{p}, \frac{j+1}{p}\right[, \qquad i,j \in 0,\dots,p-1.$$

The random field is defined as

$$a(y) = 1 + \frac{1}{2}\sum_{i,j=0}^{p-1} \chi_{S_{i,j}} y_{i,j},$$

where χ_S denotes the characteristic function of a set S, and the $y_{i,j}$ are random coefficients that are independent, each with identical uniform distribution on $[-1,1]$. Thus, our vector of parameters is

$$\mathbf{y} = (y_{i,j})_{i,j=0}^{p-1} \in \mathbb{R}^{p\times p}.$$

In our numerical tests, we take $p = 4$, that is 16 parameters, and work in the ambient space $V = H_0^1(\Omega)$. The sensor measurements are modelled with linear functionals that are local averages of the form

$$\ell_{\mathbf{x},\tau}(u) = \int_D u(\mathbf{r})\varphi_\tau(\mathbf{r}-\mathbf{x})\,\mathrm{d}\mathbf{r},$$

where

$$\varphi_\tau(\mathbf{r}) \propto \exp(-|\mathbf{r}|/2\tau^2)$$

is a radial function such that $\int \varphi_\tau = 1$. The parameter $\tau > 0$ represents the spread around the center $\mathbf{x}$. For the observation space W of our example, we randomly select $m = 50$ centers $\mathbf{x}_i \in [0.1, 0.9]^2$ and spreads $\tau_i \in [0.05, 0.1]$, and compute the Riesz representers $\omega_{\mathbf{x}_i,\tau}$ of $\ell_{\mathbf{x}_i,\tau}$ in $H_0^1(\Omega)$. We then set

$$W \coloneqq \{\omega_{\mathbf{x}_i,\tau}\}_{i=1}^M$$

which is a space of dimension $m = 50$. Figure 3.3 shows the m centers $\mathbf{x}_i$. As an example, the figure also plots the function $\omega_{\mathbf{x}_i,\tau}$ for $i = 10$, which has center $\mathbf{x}_i = (0.23, 0.75)$ and spread $\tau_i = 0.06$.

In our numerical experiments, we aimed primarily at comparing in terms of the worst case reconstruction error our approximation to the optimal affine algorithm A^*_{wca} with three other affine algorithms. The performance results are given in Figure 3.4, and they highlight the superiority of the best affine algorithm (green curves in the figure). This comes however at the cost of a computationally intensive training phase to run our primal-dual algorithm as we are going to discuss. The affine algorithms to which we do the comparison are:

- $A_{\mathrm{mvn}}(\omega) = \omega$ (blue color in Figure 3.4)
- the affine PBDW algorithm $A^{(\mathrm{aff})}_{n^*}$ for an optimized value of $n \le m$ (denoted A_{one} in Figure 3.4, and in black color)
- an affine algorithm A^*_{msa}, which corresponds to the best algorithm in the mean square sense when we assume a normal distribution on $\mathcal{M}$ (see [A4] for the details on this algorithm, and red color in Figure 3.4).

In order to illustrate the impact of the number of measurements that are used, in Figure 3.4 we consider the nested subspaces

$$W_m = \mathrm{span}\{\omega_{\mathbf{x}_i,\tau_i}\}_{i=1}^m \subset W$$

for $m = 10$, 20, 30, 40, 50 so that $W_{50} = W$.

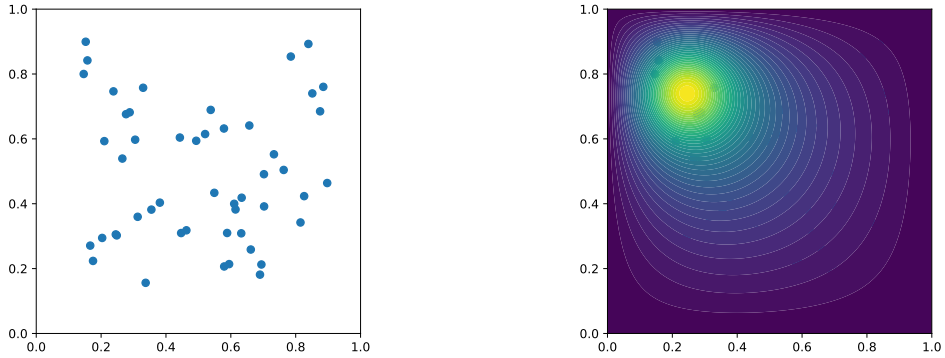


Figure 3.3: Sensor locations and the function $\omega_{\mathbf{x}_i,\tau_i}$ for $i = 10$ ($\mathbf{x}_i = (0.23, 0.75)$ and $\tau_i = 0.06$).

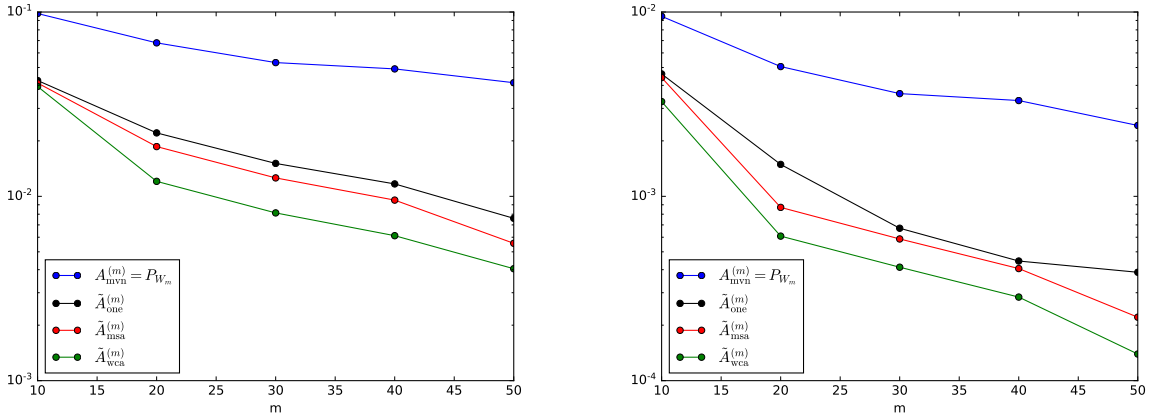


Figure 3.4: Comparison of the reconstruction errors (left: $H_0^1(\Omega)$ norm; right: $L^2(\Omega)$ norm).

Brief discussion on the complexity of the primal-dual algorithm: At each iteration of the algorithm, the main bottleneck is the computation of L^* (see equation (3.2.14)). It requires to do J matrix-vector products with the matrices $Q_j \in \mathbb{R}^{N\times m(N+1)}$ and then do a summation of the resulting vectors. The cost of these operations thus increases linearly with J in terms of computational time and memory ressources. In fact, the limitation in memory was the main reason to fix $J = 10^3$ in our case and not work with a larger number of training snapshots. Let us make a quick count on the cost in terms of the number of elements to store at each iteration. The matrices Q_j are sparse. For each row, there are $m+1$ nonnegative coefficients. Therefore we need to store $N(m+1)$ coefficients for each matrix, therefore a total number of $JN(m+1)$ coefficients. In our case, $N = 110$ was carefully fixed to guarantee that

$$\max_{u\in\mathcal{M}_{\text{greedy}}\cup\mathcal{M}_{\text{test}}} \|u - P_{V_N}u\| \leq \varepsilon_N = 5.10^{-5}.$$

We have m ranging between 10 and 50. Thus the number of nonnegative elements that we have to store for each Q_j ranges between 1210 and 5610. Therefore, taking $J = 10^3$ as in our computation, we need to handle a total number of coefficients ranging between $1.21.10^6$ and $5.61.10^6$.

3.3 Sensor placement

In section 3.2 we have summarized a strategy to find an optimal affine reconstruction algorithm A^*_{aff} for a given observation space W. This algorithm is connected to an optimal subspace V_n^{opt} to use in the PBDW method although we note that our procedure does not yield an explicit characterization of V_n^{opt} and a further post-processing may have been necessary to find it in practice. In [A10], we have considered the "reciprocal" problem, namely, for a given reduced model space V_n with a good accuracy ε_n, the question is how to guarantee a good reconstruction accuracy with a number of measurements $m \geq n$ as small possible. In view of the error bound (3.2.4), one natural objective is to guarantee that $\mu(V_n, W_m)$ is maintained of moderate size. Note that taking $W_m = V_n$ would automatically give the minimal value $\mu(V_n, W_m) = 1$ with $m = n$. However, in a typical data acquisition scenario, the measurements that span the basis of W_m are chosen from within a limited class. This is the case for example when placing m pointwise sensors at various locations within the physical domain Ω.

We model this restriction by asking that the ℓ_i are picked within a *dictionary* $\mathcal{D}$ of V', that is a set of linear functionals normalized according to

$$\|\ell\|_{V'} = 1, \quad \ell \in \mathcal{D},$$

which is *complete* in the sense that $\ell(v) = 0$ for all $\ell \in \mathcal{D}$ implies that $v = 0$. With an abuse of notation, we identify $\mathcal{D}$ with the subset of V that consists of all Riesz representers ω of the above linear functionals ℓ. With such an identification, $\mathcal{D}$ is a set of functions normalized according to

$$\|\omega\| = 1, \quad \omega \in \mathcal{D},$$

such that the finite linear combinations of elements of $\mathcal{D}$ are dense in V. Our task is therefore to pick $\{\omega_1, \ldots, \omega_m\} \in \mathcal{D}$ in such a way that

$$\beta(V_n, W_m) \geq \beta^* > 0, \tag{3.3.1}$$

for some prescribed $0 < \beta^* < 1$, with m larger than n but as small as possible. In particular, we may introduce

$$m^* = m^*(\beta^*, \mathcal{D}, V_n),$$

the minimal value of m such that there exists $\{\omega_1, \dots, \omega_m\} \in \mathcal{D}$ satisfying (3.3.1).

In [A10], we show two "extreme" results:

- For any V_n and $\mathcal{D}$, there exists $\beta^* > 0$ such that $m^* = n$, that is, the inf-sup condition (3.3.1) holds with the minimal possible number of measurements. However this β^* could be arbitrarily close to 0.
- For any prescribed $\beta^* > 0$ and any model space V_n, there are instances of dictionaries $\mathcal{D}$ such that m^* is arbitrarily large.

We then discuss particular cases of relevant dictionaries for the particular space $V = H_0^1(\Omega)$, with inner product and norms

$$\langle u, v \rangle \coloneqq \int_\Omega \nabla u(x) \cdot \nabla v(x) \, \mathrm{d}x \quad \text{and} \quad \|u\| \coloneqq \|\nabla u\|_{L^2(\Omega)}.$$

The considered dictionaries model local sensors, either as point evaluations or as local averages. In the first case,

$$\mathcal{D} = \{\ell_x = \delta_x \ : \ \forall x \in \Omega\},$$

which requires that V is a reproducing kernel Hilbert space (RKHS) of functions defined on Ω, that is a Hilbert space that continuously embeds in $\mathcal{C}(\Omega)$. Examples of such spaces are the Sobolev spaces $H^s(\Omega)$ for $s > d/2$, possibly with additional boundary conditions. In the case of local averages, the linear functionals are of the form

$$\ell_{x,\tau}(u) = \int_\Omega u(y) \varphi_\tau(y - x) dy,$$

where

$$\varphi_\tau(y) \coloneqq \tau^{-d} \varphi\left(\frac{y}{\tau}\right),$$

for some fixed radial function φ compactly supported in the unit ball $B = \{|x| \leq 1\}$ of $\mathbb{R}^d$ and such that $\int \varphi = 1$, and $\tau > 0$ representing the point spread. The dictionary in this case is

$$\mathcal{D} = \{\ell_{x,\tau} \ : \ \forall x \in \Omega\}.$$

We could even consider an interval of values for τ in $[\tau_{\min}, \tau_{\max}]$ with $0 < \tau_{\min} \leq \tau_{\max}$,

$$\mathcal{D} = \{\ell_{x,\tau} \ : \ \forall (x, \tau) \in \Omega \times [\tau_{\min}, \tau_{\max}]\}.$$

For the above cases of dictionaries, we provide upper estimates of m^* in the case of spaces V_n that satisfy some inverse estimates, such as finite element or trigonometric polynomial spaces. The optimal value m^* is proved to be of the same order as n when the sensors are uniformly spaced.

This a-priori analysis is not possible for more general spaces V. It is not possible either for subspaces V_n such as reduced basis spaces, which are preferred to finite element spaces for model order reduction because the approximation error ε_n of the manifold $\mathcal{M}$ defined in (3.2.1) is expected to decay much faster in elliptic and parabolic problems). For such general spaces, we need a strategy

to select the measurements. In practice, V is of finite but very large dimension and $\mathcal{D}$ is of finite but very large cardinality

$$M := \#(\mathcal{D}) \gg 1.$$

For this reason, the exhaustive search of the set $\{\omega_1, \dots, \omega_m\} \subset \mathcal{D}$ maximizing $\beta(V_n, W_m)$ for a given $m > 1$ is out of reach. One natural alternative is to rely on greedy algorithms where the ω_j are picked incrementally.

Our starting point to the design of such algorithms is the observation that (3.3.1) is equivalent to having

$$\sigma_m = \sigma(V_n, W_m) := \sup_{v \in V_n, \|v\|=1} \|v - P_{W_m} v\| \leq \sigma^*, \quad \sigma^* := \sqrt{1 - (\beta^*)^2} < 1. \tag{3.3.2}$$

Therefore, our objective is to construct a space W_m spanned by m elements from $\mathcal{D}$ that captures all unit norm vectors of V_n with the prescribed accuracy $\sigma^* < 1$. This leads us to study and analyze algorithms which may be thought as generalization to the well-studied orthogonal matching pursuit algorithm (OMP), equivalent to the algorithms we study here when applied to the case $n = 1$ with a unit norm vector ϕ_1 that generates V_1. In [A10], we proposed and analyzed two algorithms and I briefly summarize the main results in Sections 3.3.1 and 3.3.2. In Section 3.3.3 I discuss the case of pointwise evaluations. In Section 3.3.4 I report some of our numerical tests. The main theoretical result is that we show that both algorithms always converge, ensuring that (3.3.1) holds for m sufficiently large, and we also give conditions on $\mathcal{D}$ that allow us to a-priori estimate the minimal value of m where this happens. The main observation from our numerical experiments is the ability of our greedy algorithms to pick good points. In particular, in the case of dictionaries of point evaluations or local averages, we observe that the selection performed by the greedy algorithms is near optimal in simple 1D cases in the sense that it achieves (3.3.1) after a number of iterations which is proportional to n and which can be predicted in theory.

Before finishing this section, let us outline the main differences and points of contact between our approach and existing works in the literature. The problem of optimal placement of sensors, which corresponds to the particular setting where the linear functionals are point evaluations or local averages, has been extensively studied since the 1970's in control and systems theory. In this context, the state function to be estimated is the realization of a Gaussian stochastic process, typically obtained as the solution of a linear PDE with a white noise forcing term. The error is then measured in the mean square sense (3.1.4), rather than in the worst case performance sense (3.1.3) which is the point of view adopted in our work. The function to be minimized by the sensors locations is then the trace of the error covariance, while we target at minimizing the inverse inf-sup constant $\mu(V_n, W)$. See in particular [96] where the existence and characterization of the optimal sensor location is established in this stochastic setting. Continuous optimization algorithms have been proposed for computing the optimal sensor location, see e.g. [93, 97, 95]. One common feature with our approach is that the criterion to be minimized by the optimal location is non-convex, which leads to potential difficulties when the number of sensors is large. This is our main motivation for introducing a greedy selection algorithm, which in addition allows us to consider more general dictionaries.

3.3.1 A collective OMP algorithm

In this section we discuss a first numerical algorithm for the incremental selection of the spaces W_m, inspired by the orthonormal matching pursuit (OMP) algorithm which is recalled below. More

precisely, our algorithm may be viewed as applying the OMP algorithm for the collective approximation of the elements of an orthonormal basis of V_n by linear combinations of m members of the dictionary.

Our objective is to reach a bound (3.3.2) for the quantity σ_m. Note that this quantity can also be written as

$$\sigma_m = \|(I - P_{W_m})|_{V_n}\|_{\mathcal{L}(V_n,V)},$$

that is, σ_m is the spectral norm of $I - P_{W_m}$ restricted to V_n.

Description of the algorithm: When $n = 1$, there is only one unit vector $\phi_1 \in V_1$ up to a sign change. A commonly used strategy for approximating ϕ_1 by a small combination of elements from $\mathcal{D}$ is to apply a greedy algorithm, the most prominent one being the orthogonal matching pursuit (OMP): we iteratively select

$$\omega_k = \arg\max_{\omega\in\mathcal{D}} |\langle \omega, \phi_1 - P_{W_{k-1}}\phi_1\rangle|,$$

where $W_{k-1} \coloneqq \operatorname{span}\{\omega_1, \dots, \omega_{k-1}\}$ and $W_0 \coloneqq \{0\}$. In practice, one often relaxes the above maximization, by taking ω_k such that

$$|\langle \omega_k, \phi_1 - P_{W_{k-1}}\phi_1\rangle| \geq \kappa \max_{\omega\in\mathcal{D}} |\langle \omega, \phi_1 - P_{W_{k-1}}\phi_1\rangle|,$$

for some fixed $0 < \kappa < 1$, for example $\kappa = \frac{1}{2}$. This is known as the weak OMP algorithm, but we refer to it as OMP, as well. It has been studied in [59, 84], see also [52] for a complete survey on greedy approximation.

For a general value of n, one natural strategy is to define our greedy algorithm as follows: we iteratively select

$$\omega_k = \arg\max_{\omega\in\mathcal{D}} \max_{v\in V_n, \|v\|=1} |\langle \omega, v - P_{W_{k-1}}v\rangle| = \arg\max_{\omega\in\mathcal{D}} \|P_{V_n}(\omega - P_{W_{k-1}}\omega)\|. \tag{3.3.3}$$

Note that in the case $n = 1$, we obtain the original OMP algorithm applied to ϕ_1.

As to the implementation of this algorithm, we take $(\phi_1, \dots, \phi_n)$ to be any orthonormal basis of V_n. Then

$$\|P_{V_n}(\omega - P_{W_{k-1}}\omega)\|^2 = \sum_{i=1}^{n} |\langle \omega - P_{W_{k-1}}\omega, \phi_i\rangle|^2 = \sum_{i=1}^{n} |\langle \phi_i - P_{W_{k-1}}\phi_i, \omega\rangle|^2$$

Therefore, at every step k, we have

$$\omega_k = \arg\max_{\omega\in\mathcal{D}} \sum_{i=1}^{n} |\langle \phi_i - P_{W_{k-1}}\phi_i, \omega\rangle|^2,$$

which amounts to a stepwise optimization of a similar nature as in the standard OMP. Note that, while the basis $(\phi_1, \dots, \phi_n)$ is used for the implementation, the actual definition of the greedy selection algorithm is independent of the choice of this basis in view of (3.3.3). It only involves V_n and the dictionary $\mathcal{D}$. Similar to OMP, we may weaken the algorithm by taking ω_k such that

$$\sum_{i=1}^{n} |\langle \phi_i - P_{W_{k-1}}\phi_i, \omega_k\rangle|^2 \geq \kappa^2 \max_{\omega\in\mathcal{D}} \sum_{i=1}^{n} |\langle \phi_i - P_{W_{k-1}}\phi_i, \omega\rangle|^2,$$

for some fixed $0 < \kappa < 1$.

For such a basis, we introduce the residual quantity

$$r_m \coloneqq \sum_{i=1}^{n} \|\phi_i - P_{W_m}\phi_i\|^2.$$

This quantity allows us to control the validity of (3.3.1) since we have

$$\sigma_m = \sup_{v\in V_n, \|v\|=1} \|v - P_{W_m}v\| = \sup_{\sum_{i=1}^n c_i^2=1} \Big\| \sum_{i=1}^{n} c_i(\phi_i - P_{W_m}\phi_i)\Big\| \leq r_m^{1/2},$$

and therefore (3.3.1) holds provided that $r_m \leq \sigma^2 = 1 - \gamma^2$.

Remark 3.3.1. *The quantity $r_m^{1/2}$ is the Hilbert-Schmidt norm of the operator $I - P_{W_m}$ restricted to V_n. The inequality $\sigma_m \leq r_m^{1/2}$ simply expresses the fact that the Hilbert-Schmidt norm controls the spectral norm. On the other hand, in dimension n, the Hilbert-Schmidt norm can be up to $n^{1/2}$ times the spectral norm. This lack of sharpness is one principle limitation in our convergence analysis which uses the fact that we can estimate the decay of r_m, but not directly that of σ_m.*

Convergence analysis: By analogy to the analysis of OMP provided in [84], we introduce for any $\Psi = (\psi_1, \ldots, \psi_n) \in V^n$ the quantity

$$\|\Psi\|_{\ell^1(\mathcal{D})} \coloneqq \inf_{c_{\omega,i}} \Big\{ \sum_{\omega\in\mathcal{D}} \Big(\sum_{i=1}^{n} |c_{\omega,i}|^2 \Big)^{1/2} \ : \ \psi_i = \sum_{\omega\in\mathcal{D}} c_{\omega,i}\omega, \quad i = 1, \ldots, n \Big\},$$

or equivalently, denoting $c_\omega \coloneqq \{c_{\omega,i}\}_{i=1}^n$,

$$\|\Psi\|_{\ell^1(\mathcal{D})} \coloneqq \inf_{c_\omega} \Big\{ \sum_{\omega\in\mathcal{D}} \|c_\omega\|_2 \ : \ \Psi = \sum_{\omega\in\mathcal{D}} c_\omega\omega \Big\}.$$

This quantity is a norm on the subspace of V^n on which it is finite.

Given that $\Phi = (\phi_1, \ldots, \phi_n)$ is any orthonormal basis of V_n, we write

$$J(V_n) \coloneqq \|\Phi\|_{\ell^1(\mathcal{D})}.$$

This quantity is indeed independent on the orthonormal basis Φ: if $\tilde{\Phi} = (\tilde{\phi}_1, \ldots, \tilde{\phi}_n)$ is another orthonormal basis, we have $\tilde{\Phi} = U\Phi$ where U is unitary. Therefore any representation $\Phi = \sum_{\omega\in\mathcal{D}} c_\omega\omega$ induces the representation

$$\tilde{\Phi} = \sum_{\omega\in\mathcal{D}} \tilde{c}_\omega\omega, \quad \tilde{c}_\omega = Uc_\omega,$$

with the equality

$$\sum_{\omega\in\mathcal{D}} \|\tilde{c}_\omega\|_2 = \sum_{\omega\in\mathcal{D}} \|c_\omega\|_2,$$

so that $\|\Phi\|_{\ell^1(\mathcal{D})} = \|\tilde{\Phi}\|_{\ell^1(\mathcal{D})}$.

One important observation is that if $\Phi = (\phi_1, \dots, \phi_n)$ is an orthonormal basis of V_n and if $\Phi = \sum_{\omega \in \mathcal{D}} c_\omega \omega$, one has

$$n = \sum_{i=1}^{n} \|\phi_i\| \leq \sum_{i=1}^{n} \sum_{\omega \in \mathcal{D}} |c_{\omega,i}| = \sum_{\omega \in \mathcal{D}} \|c_\omega\|_1 \leq \sum_{\omega \in \mathcal{D}} n^{1/2} \|c_\omega\|_2.$$

Therefore, we always have

$$J(V_n) \geq n^{1/2}.$$

Using the quantity $J(V_n)$, we can generalize the result of [84] on the OMP algorithm in the following way.

Theorem 3.3.2. *Assuming that $J(V_n) < \infty$, the collective OMP algorithm satisfies*

$$r_m \leq \frac{J(V_n)^2}{\kappa^2} (m+1)^{-1}, \quad m \geq 0. \tag{3.3.4}$$

Remark 3.3.3. *Note that the right side of (3.3.4), is always larger than $n(m+1)^{-1}$, which is consistent with the fact that $\beta(V_n, W_m) = 0$ if $m < n$.*

One natural strategy for selecting the measurement space W_m is therefore to apply the above described greedy algorithm, until the first value $\tilde{m} = \tilde{m}(n)$ is met such that $\beta(V_n, W_m) \geq \gamma$. According to (3.3.4), this value satisfies

$$m(n) \leq \frac{J(V_n)^2}{\kappa^2 \sigma^2}. \tag{3.3.5}$$

For a general dictionary $\mathcal{D}$ and space V_n we have no control on the quantity $J(V_n)$ which could even be infinite, and therefore the above result does not guarantee that the above selection strategy eventually meets the target bound $\beta(V_n, W_m) \geq \gamma$. In order to treat this case, we establish a perturbation result similar to that obtained in [59] for the standard OMP algorithm.

Theorem 3.3.4. *Let $\Phi = (\phi_1, \dots, \phi_n)$ be an orthonormal basis of V_n and $\Psi = (\psi_1, \dots, \psi_n) \in V^n$ be arbitrary. Then the application of the collective OMP algorithm on the space V_n gives*

$$r_m \leq 4 \frac{\|\Psi\|^2_{\ell^1(\mathcal{D})}}{\kappa^2} (m+1)^{-1} + \|\Phi - \Psi\|^2, \quad m \geq 1.$$

where $\|\Phi - \Psi\|^2 := \|\Phi - \Psi\|^2_{V^n} = \sum_{i=1}^n \|\phi_i - \psi_i\|^2$.

As an immediate consequence of the above result, we obtain that the collective OMP converges for any space V_n, even when $J(V_n)$ is not finite.

The next corollary shows that if $\gamma > 0$, one has $\beta(V_n, W_m) \geq \gamma$ for m large enough.

Corollary 3.3.5. *For any n dimensional space V_n, the application of the collective OMP algorithm on the space V_n gives that $\lim_{m \to +\infty} r_m = 0$.*

3.3.2 A worst case OMP algorithm

We present in this section a variant of the previous collective OMP algorithm. This algorithm was first tested in [33] but without any convergence analysis. In our numerical experiments this variant performs better than the collective OMP algorithm, however its analysis is more delicate. In particular we do not obtain convergence bounds that are as good.

Description of the algorithm: We first take

$$v_k \coloneqq \operatorname{argmax}\Big\{\|v - P_{W_{k-1}} v\| \,:\, v \in V_n,\, \|v\| = 1\Big\},$$

the vector in the unit ball of V_n that is less well captured by W_{k-1} and then define ω_k by applying one step of OMP to this vector, that is

$$|\langle v_k - P_{W_{k-1}} v_k, \omega_k\rangle| \geq \kappa \max\Big\{|\langle v_k - P_{W_{k-1}} v_k, \omega\rangle| \,:\, \omega \in \mathcal{D}\Big\},$$

for some fixed $0 < \kappa < 1$.

Convergence analysis: The first result gives a convergence rate of r_m under the assumption that $J(V_n) < \infty$, similar to Theorem 3.3.2, however with a multiplicative constant that is inflated by n^2.

Theorem 3.3.6. *Assuming that $J(V_n) < \infty$, the worst case OMP algorithm satisfies*

$$r_m \leq \frac{n^2 J(V_n)^2}{\kappa^2}(m+1)^{-1}, \quad m \geq 0. \tag{3.3.6}$$

For the general case, we establish a perturbation result similar to Theorem 3.3.4, with again a multiplicative constant that depends on the dimension of V_n.

Theorem 3.3.7. *Let $\Phi = (\phi_1, \ldots, \phi_n)$ be an orthonormal basis of V_n and $\Psi = (\psi_1, \ldots, \psi_n) \in V^n$ be arbitrary. Then the application of the worst case OMP algorithm on the space V_n gives*

$$r_m \leq 4\frac{n^2 \|\Psi\|^2_{\ell^1(\mathcal{D})}}{\kappa^2}(m+1)^{-1} + n^2\|\Phi - \Psi\|^2, \quad m \geq 1.$$

where $\|\Phi - \Psi\|^2 \coloneqq \|\Phi - \Psi\|^2_{V^n} = \sum_{i=1}^n \|\phi_i - \psi_i\|^2$.

By the exact same argument as in the proof of Corollary 3.3.5 we find that that the worst case OMP converges for any space V_n, even when $J(V_n)$ is not finite.

Corollary 3.3.8. *For any n dimensional space V_n, the application of the worst case OMP algorithm on the space V_n gives that $\lim_{m\to+\infty} r_m = 0$.*

3.3.3 Application to point evaluation

As a simple example, we consider a bounded univariate interval $\Omega = I$ and take $V = H^1_0(I)$ which is continuously embedded in $\mathcal{C}(I)$. Without loss of generality we take $I =]0,1[$. For every $x \in]0,1[$,

the Riesz representer of δ_x is given by the solution of $\omega'' = \delta_x$ with zero boundary condition. Normalising this solution ω it with respect to the V norm, we obtain

$$\omega_x(t) = \begin{cases} \frac{t(1-x)}{\sqrt{x(1-x)}}, & \text{for } t \leq x \\ \frac{(1-t)x}{\sqrt{x(1-x)}}, & \text{for } t > x. \end{cases}$$

For any set of m distinct points $0 < x_1 < \cdots < x_m < 1$, the associated measurement space $W_m = \operatorname{span}\{\omega_{x_1}, \ldots, \omega_{x_m}\}$ coincides with the space of piecewise affine polynomials with nodes at $x_1, \ldots, x_m$ that vanish at the boundary. Denoting $x_0 \coloneqq 0$ and $x_{m+1} \coloneqq 1$, we have

$$W_m = \{\omega \in \mathcal{C}^0([0,1]),\ \omega|_{[x_k, x_{k+1}]} \in \mathbb{P}_1,\ 0 \leq k \leq m,\ \text{and } \omega(0) = \omega(1) = 0\}.$$

As an example for the space V_n, let us consider the span of the Fourier basis (here orthonormalized in V),

$$\phi_k \coloneqq \frac{\sqrt{2}}{\pi k} \sin(k\pi x), \quad 1 \leq k \leq n. \tag{3.3.7}$$

Let us now estimate $m(n)$ in this example if we choose the points with the greedy algorithms that we have introduced. This boils down to estimate for $J(V_n)$. In this simple case,

$$J(V_n) \coloneqq \|\Phi\|_{\ell^1(\mathcal{D})} = \inf \left\{ \int_{x \in [0,1]} \|c_x\|_2 \, \mathrm{d}x \ : \ \Phi = \int_{x \in [0,1]} c_x \omega_x \, \mathrm{d}x \right\}$$

and we can derive c_x for every $x \in [0,1]$ by differentiating twice the components of Φ since

$$\Phi''(x) = \int_{y \in [0,1]} c_y \omega_y''(x) \, \mathrm{d}y = -\int_{y \in [0,1]} c_y \delta_y(x) \, \mathrm{d}x = -c_x.$$

Thus, using the basis functions ϕ_k defined by (3.3.7), we have

$$J(V_n) = \int_{x \in [0,1]} \left(\sum_{k=1}^{n} |\phi_k''(x)^2| \right)^{1/2} \mathrm{d}x = \int_{x \in [0,1]} \left(\sum_{k=1}^{n} 2k\pi |\sin(k\pi x)|^2 \right)^{1/2} \mathrm{d}x \sim n^{3/2}.$$

Estimate (3.3.5) for the convergence of the collective OMP approach yields

$$m(n) \gtrsim \frac{n^3}{\kappa^2 \sigma^2},$$

while for the worst case OMP, estimate (3.3.6) gives

$$m(n) \gtrsim \frac{n^5}{\kappa^2 \sigma^2}.$$

These bounds deviate from the optimal estimation due to the use of the Hilbert-Schmidt norm in the analysis. Our numerical results reported in [A10] revealed that the greedy algorithms actually behave much better in this case. I do not summarize them here for the sake of brevity. Instead, I recall in the next section only one numerical test when V_n is a reduced model space.

3.3.4 Some numerical illustrations

We consider the same checkerboard problem as the one described in (3.2.15). We work here with 16 parameters (that is, a 4×4 checkerboard diffusion field). The PDE solution snapshots are computed using $\mathbb{P}_1$ finite elements on a uniform trianuglar mesh $\mathcal{T}_h$ of size $h = 2^{-7}$.

We consider a dictionary $\mathcal{D}$ composed of local averages by the nodal basis functions of the finite element space. We run our greedy procedure to select the observation space W_m for two different spaces V_n. A space of trigonometric polynomials

$$V_n^{\sin} = \operatorname{span}\{\phi_{k,\ell} : 1 \leq k \times \ell \leq n\},$$

and a reduced basis space

$$V_n = \operatorname{span}\{u_h(y^{(1)}), \ldots, u_h(y^{(n)})\}$$

spanned by solutions $u_h(y^{(i)})$ to (3.2.15) for a given parameter $y^{(i)}$

We recall that the worst case performance of the state estimation algorithm as defined in (3.2.4) is given by the product of the inverse inf-sup constant $\mu(V_n, W_m)$ by the approximation error $\varepsilon_n = \operatorname{dist}(\mathcal{M}, V_n)$. Since the exact computation of ε_n is out of reach, we instead study the average projection error for a collection of solutions $u_h(a(y))$ to $V_n = V_n^{\text{red}}$ or $V_n^{\sin}$. The left side of Figure 3.5 shows that the reduced bases outperform the trigonometric polynomial spaces by several order of magnitude, as to the decay of this approximation error. On the other hand, the right side of Figure 3.5 shows (here in the case $n = 20$) that when applying the greedy algorithm, the inf-sup constant $\beta(V_n, W_m)$ is better behaved for the trigonometric polynomial spaces, however only by a moderate factor of around 1.1. Therefore the final trade-off is clearly in favor of reduced basis spaces.

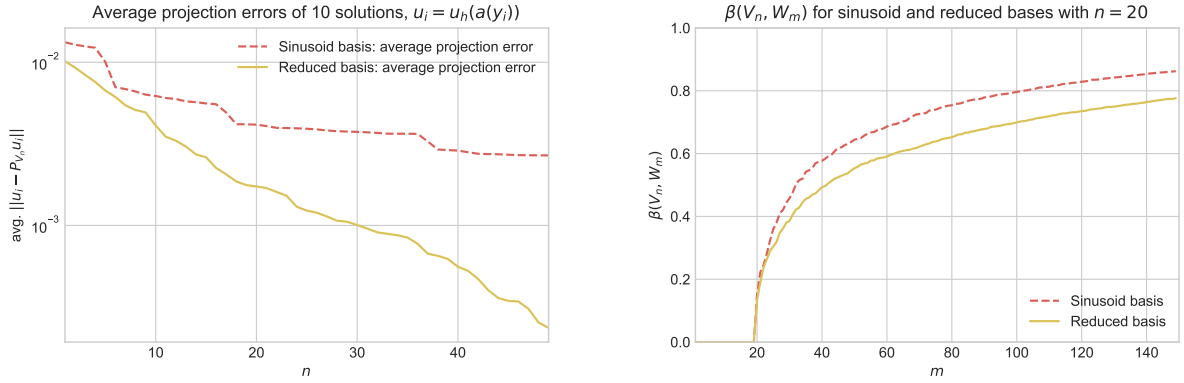


Figure 3.5: Results on unit square.

3.4 A Piece-Wise Affine Algorithm to reach the Benchmark Optimality

The simplicity of the plain PBDW method (3.2.2) and its above variants come together with a fundamental limitation of performance: since the map $w \mapsto A_n(w)$ is linear or affine, the reconstruction

necessarily belongs to an m or $m+1$ dimensional space, and therefore the worst case performance is necessarily bounded from below by the Kolmogorov width $d_m(\mathcal{M})$ or $d_{m+1}(\mathcal{M})$. In other words, if we restrict ourselves to affine algorithms, we have

$$\min_{A:W\to V} E_{\mathrm{wc}}(A,\mathcal{M}) \leq d_{m+1}(\mathcal{M}) \leq \min_{\substack{A:W\to V \\ A \text{ affine}}} E_{\mathrm{wc}}(A,\mathcal{M}).$$

and affine algorithms will miss optimality especially in cases where

$$\min_{A:W\to V} E_{\mathrm{wc}}(A,\mathcal{M}) \ll d_{m+1}(\mathcal{M}).$$

This is expected to happen in elliptic problems with weak coercivity or in hyperbolic problems.

In view of this limitation, the principal objective of our contribution [S3] is to develop *nonlinear* state estimation techniques which *provably* overcome the bottleneck of the Kolmogorov width $d_m(\mathcal{M})$. In the next pages, I summarize the main ideas from this contribution. I will focus particularly on summarizing a nonlinear recovery method based on a family of affine reduced models $(V_k)_{k=1,\dots,K}$. Each V_k has dimension $n_k \leq m$ and serves as a local approximation to a portion $\mathcal{M}_k$ of the solution manifold. Applying the PBDW method with each such space, results in a collection of state estimators u_k^*. The value k for which the true state u belongs to $\mathcal{M}_k$ being unknown, we introduce a *model selection* procedure in order to pick a value k^*, and define the resulting estimator $u^* = u_{k^*}^*$. We show that this estimator has performance comparable to optimal in a sense which we make precise later on, and which cannot be achieved by the standard linear/affine PBDW method due to the above described limitations.

Model selection is a classical topic of mathematical statistics [66], with representative techniques such as complexity penalization or cross-validation in which the data are used to select a proper model. Our approach differs from these techniques in that it exploits (in the spirit of *data assimilation*) the PDE model which is available to us, by evaluating the distance to the manifold

$$\mathrm{dist}(v,\mathcal{M}) = \min_{y\in Y} \|v - u(y)\|, \tag{3.4.1}$$

of the different estimators $v = u_k^*$ for $k = 1,\dots,K$, and picking the value k^* that minimizes it. In practice, the quantity (3.4.1) cannot be exactly computed and we instead rely on a computable surrogate quantity $\mathcal{S}(v,\mathcal{M})$ expressed in terms of the residual to the PDE. One typical instance where such a surrogate is available and easily computable is when the parametric PDE (3.1.1) has the form of a linear operator equation

$$\mathcal{B}(y)u = f(y),$$

where $\mathcal{B}(y)$ is boundedly invertible from V to V', or more generally, from $V \to Z'$ for a test space Z different from V, uniformly over $y \in Y$. Then $\mathcal{S}(v,\mathcal{M})$ is obtained by minimizing the residual

$$\mathcal{R}(v,y) = \|\mathcal{B}(y)v - f(y)\|_{Z'},$$

over $y \in Y$. In other words,

$$\mathcal{S}(v,\mathcal{M}) = \min_{y\in \mathrm{Y}} \mathcal{R}(v,y).$$

This task itself is greatly facilitated in the case where the operators $A(y)$ and source terms $f(y)$ have affine dependence in Y. One relevant example is the second order elliptic diffusion equation

with affine diffusion coefficient,

$$-\mathrm{div}(a\nabla u) = f(y), \quad a = a(x;y) = \bar{a}(x) + \sum_{j=1}^{d} y_j \psi_j(x).$$

Optimality benchmark under perturbations

In order to present our piece-wise affine strategy and its performance, we need to enrich the notions of benchmark optimality introduced in section 3.1. In that section, we introduced in (3.1.8) the quantity δ_0 which was defined as

$$\delta_0 = \delta_0(\mathcal{M}, W) \coloneqq \sup\{\mathrm{diam}(\mathcal{M}_w) \ : \ w \in W\} = \sup\{\|u - v\| \ : \ u, v \in \mathcal{M}, \ u - v \in W^{\perp}\}.$$

We saw in (3.1.9) that δ_0 can be related to the worst-case optimal performance $E^*_{\mathrm{wc}}(\mathcal{M})$ by the equivalence

$$\frac{1}{2}\delta_0 \leq E^*_{\mathrm{wc}}(\mathcal{M}) \leq \delta_0.$$

We next introduce a somewhat relaxed benchmark quantity to take into account the fact that computationally feasible algorithms usually introduce simplifications of the geometry of the manifold. In the case of the plain PBDW, the simplification is that the manifold is "replaced" by a linear or an affine subspace V_n, which makes that for most practical and theoretical purposes, $\mathcal{M}$ could be replaced by the cylinder $\mathcal{K}_n$ introduced in (3.2.3). As we will see later on, the relaxed benchmark will also allow us to take into account model error and measurement noise in the analysis.

In order to account for manifold simplification as well as model bias, for any given accucary $\sigma > 0$, we introduce the σ-offset of $\mathcal{M}$,

$$\mathcal{M}_\sigma \coloneqq \{v \in V \ : \ \mathrm{dist}(v, \mathcal{M}) \leq \sigma\} = \bigcup_{u \in \mathcal{M}} B(u, \sigma),$$

where $B(u, \sigma)$ is the ball of center u and radius σ. Likewise, we introduce the set

$$\mathcal{M}_{\sigma,w} = \mathcal{M}_\sigma \cap (\omega + W^{\perp}),$$

which is a perturbed set of $\mathcal{M}_w$ introduced in (3.1.5) (note that this set still excludes uncertainties in w but we will come to this in a moment).

Our benchmark for the worst case error is now defined as

$$\delta_\sigma \coloneqq \max_{w \in W} \mathrm{diam}(\mathcal{M}_{\sigma,w}) = \max\{\|u - v\| \ : \ u, v \in \mathcal{M}_\sigma, \ u - v \in W^{\perp}\}. \tag{3.4.2}$$

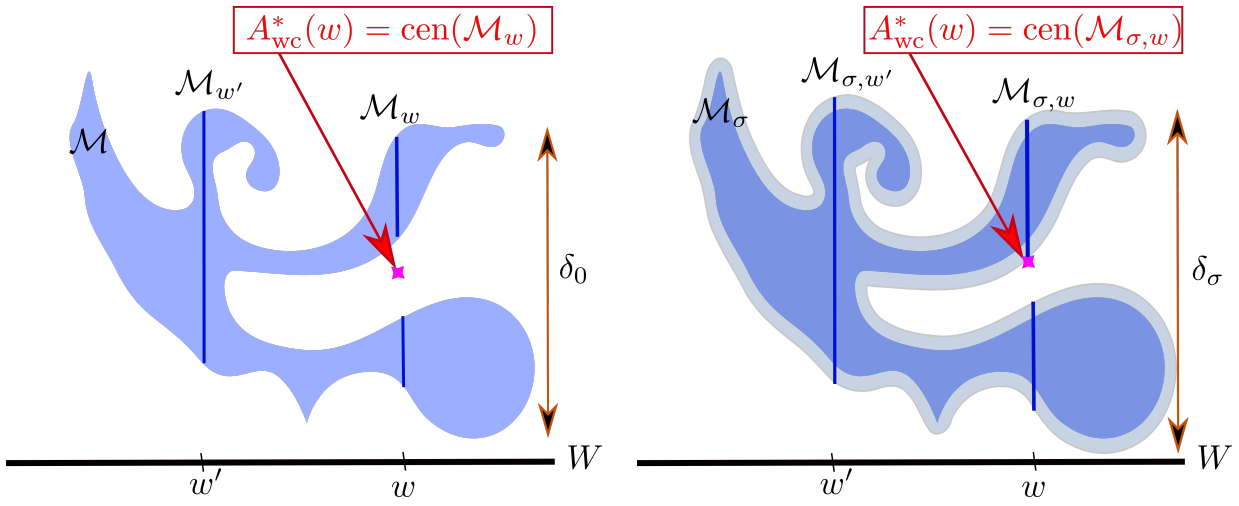
Figures 3.6a and 3.6b give an illustration of δ_0, δ_σ and the optimal scheme A^*_{wc} based on Chebyshev centers which was introduced in (3.1.7).

To account for measurement noise, we introduce the quantity

$$\tilde{\delta}_\sigma \coloneqq \max\{\|u - v\| \ : \ u, v \in \mathcal{M}, \ \|P_W u - P_W v\| \leq \sigma\}.$$

The two quantities δ_σ and $\tilde{\delta}_\sigma$ are not equivalent, however one has the framing

$$\delta_\sigma - 2\sigma \leq \tilde{\delta}_{2\sigma} \leq \delta_\sigma + 2\sigma.$$



(a) Perfect Model. (b) Perturbed Model.

Figure 3.6: Illustration of the optimal recovery benchmark on a manifold in the two dimensional Euclidean space.

In the following analysis of reconstruction methods, we use the quantity δ_σ as a benchmark which, in view of this last observation, also accounts for the lack of accuracy in the measurement of $P_W u$. Our objective is therefore to design an algorithm that, for a given tolerance $\sigma > 0$, recovers from the measurement $w = P_W u$ an approximation to u with accuracy comparable to δ_σ. Such an algorithm requires that we are able to capture the solution manifold up to some tolerance $\varepsilon \leq \sigma$ by some reduced model.

Piecewise affine reduced models

Linear or affine reduced models, as used in the PBDW algorithm, are not suitable for approximating the solution manifold when the required tolerance ε is too small. In particular, when $\varepsilon < d_m(\mathcal{M})$ one would then need to use a linear space V_n of dimension $n > m$, therefore making $\mu(V_n, W)$ infinite.

One way out is to replace the single space V_n by a *family* of affine spaces

$$V_k = \bar{u}_k + \bar{V}_k, \quad k = 1, \ldots, K,$$

each of them having dimension

$$\dim(V_k) = n_k \leq m,$$

such that the manifold is well captured by the union of these spaces, in the sense that

$$\mathrm{dist}\left(\mathcal{M}, \bigcup_{k=1}^{K} V_k\right) \leq \varepsilon$$

for some prescribed tolerance $\varepsilon > 0$. This is equivalent to saying that there exists a partition of the solution manifold

$$\mathcal{M} = \bigcup_{k=1}^{K} \mathcal{M}_k,$$

such that we have local certified bounds

$$\operatorname{dist}(\mathcal{M}_k, V_k) \leq \varepsilon_k \leq \varepsilon, \quad k = 1, \ldots, K. \tag{3.4.3}$$

We may thus think of the family $(V_k)_{k=1,\ldots,K}$ as a piecewise affine approximation to $\mathcal{M}$. We stress that, in contrast to the hierarchies $(V_n)_{n=0,\ldots,m}$ produced by reduced modeling algorithms, the spaces V_k do not have dimension k and are not nested. Most importantly, K is not limited by m while each n_k is.

The objective of using a piecewise reduced model in the context of state estimation is to have a joint control on the local accuracy ε_k as expressed by (3.4.3) and on the stability of the PBDW when using any individual V_k. This means that, for some prescribed $\mu > 1$, we ask that

$$\mu_k = \mu(\bar{V}_k, W) \leq \mu, \quad k = 1, \ldots, K. \tag{3.4.4}$$

According to (3.2.4), the worst case error bound over $\mathcal{M}_k$ when using the PBDW method with a space V_k is given by the product $\mu_k \varepsilon_k$. This suggests to alternatively require from the collection $(V_k)_{k=1,\ldots,K}$, that for some prescribed $\sigma > 0$, one has

$$\sigma_k \coloneqq \mu_k \varepsilon_k \leq \sigma, \quad k = 1, \ldots, K. \tag{3.4.5}$$

This leads us to the following definitions.

Definition 2. The family $(V_k)_{k=1,\ldots,K}$ is *σ-admissible* if (3.4.5) holds. It is *(ε, μ)-admissible* if (3.4.3) and (3.4.4) are jointly satisfied.

Obviously, any (ε, μ)-admissible family is σ-admissible with $\sigma \coloneqq \mu\varepsilon$. In this sense the notion of (ε, μ)-admissibility is thus more restrictive than that of σ-admissibility. The benefit of the first notion is in the uniform control on the size of μ which is critical in the presence of noise.

If $u \in \mathcal{M}$ is our unknown state and $w = P_W u$ is its observation, we may apply the PBDW method for the different V_k in the given family, which yields a corresponding family of estimators

$$u_k^* = u_k^*(w) = \operatorname{argmin}\{\operatorname{dist}(v, V_k) \,:\, v \in \omega + W^\perp\}, \quad k = 1, \ldots, K. \tag{3.4.6}$$

If $(V_k)_{k=1,\ldots,K}$ is σ-admissible, we find that the accuracy bound

$$\|u - u_k^*\| \leq \mu_k \operatorname{dist}(u, V_k) \leq \mu_k \varepsilon_k = \sigma_k \leq \sigma,$$

holds whenever $u \in \mathcal{M}_k$.

Therefore, if in addition to the observed data w one had an oracle giving the information on which portion $\mathcal{M}_k$ of the manifold the unknown state sits, we could derive an estimator with worst case error

$$E_{\mathrm{wc}} \leq \sigma.$$

This information is, however, not available and such a worst case error estimate cannot be hoped for, even with an additional multiplicative constant. Indeed, as we shall see below, σ can be fixed arbitrarily small by the user when building the family $(V_k)_{k=1,\ldots,K}$, while we know from (3.1.9) that the worst case error is bounded from below by $E_{\mathrm{wc}}^*(\mathcal{M}) \geq \frac{1}{2}\delta_0$ which could be non-zero. We will thus need to replace the ideal choice of k by a model selection procedure only based on the data w, that is, a map

$$w \mapsto k^*(w),$$

leading to a choice of estimator $u^* = u^*_{k^*} = A_{k^*}$. We shall prove further that such an estimator is able to achieve the accuracy

$$E_{\mathrm{wc}}(A_{k^*}, \mathcal{M}) \leq \delta_\sigma,$$

that is, the benchmark introduced in §2.2. Before discussing this model selection, we discuss the existence and construction of σ-admissible or (ε, μ)-admissible families.

Constructing admissible reduced model families

For any arbitrary choice of $\varepsilon > 0$ and $\mu \geq 1$, the existence of an (ε, μ)-admissible family results from the following observation: since the manifold $\mathcal{M}$ is a compact set of V, there exists a finite ε-cover of $\mathcal{M}$, that is, a family $\bar{u}_1, \ldots, \bar{u}_K \in V$ such that

$$\mathcal{M} \subset \bigcup_{k=1}^{K} B(\bar{u}_k, \varepsilon),$$

or equivalently, for all $v \in \mathcal{M}$, there exists a k such that $\|v - \bar{u}_k\| \leq \varepsilon$. With such an ε cover, we consider the family of trivial affine spaces defined by

$$V_k = \{\bar{u}_k\} = \bar{u}_k + \bar{V}_k, \quad \bar{V}_k = \{0\},$$

thus with $n_k = 0$ for all k. The covering property implies that (3.4.3) holds. On the other hand, for the 0 dimensional space, one has

$$\mu(\{0\}, W) = 1,$$

and therefore (3.4.4) also holds. The family $(V_k)_{k=1,\ldots,K}$ is therefore (ε, μ)-admissible, and also σ-admissible with $\sigma = \varepsilon$.

This family is however not satisfactory for algorithmic purposes for two main reasons. First, the manifold is not explicitly given to us and the construction of the centers $\bar{u}_k$ is by no means trivial. Second, asking for an ε-cover, would typically require that K becomes extremely large as ε goes to 0. For example, assuming that the parameter to solution $y \mapsto u(y)$ has Lipschitz constant L,

$$\|u(y) - u(\tilde{y})\| \leq L|y - \tilde{y}|, \quad y, \tilde{y} \in Y,$$

for some norm $|\cdot|$ of $\mathbb{R}^d$, then an ε cover for $\mathcal{M}$ would be induced by an $L^{-1}\varepsilon$ cover for Y which has cardinality K growing like ε^{-d} as $\varepsilon \to 0$. Having a family of moderate size K is important for the estimation procedure since we intend to apply the PBDW method for all $k = 1, \ldots, K$.

In order to construct (ε, μ)-admissible or σ-admissible families of better controlled size, we need to split the manifold in a more economical manner than through an ε-cover, and use spaces V_k of general dimensions $n_k \in \{0, \ldots, m\}$ for the various manifold portions $\mathcal{M}_k$. To this end, we combine standard constructions of linear reduced model spaces with an iterative splitting procedure operating on the parameter domain Y. Let us mention that various ways of splitting the parameter domain have already been considered in order to produce local reduced bases having both controlled cardinality and prescribed accuracy [55, 43, 1]. Here our goal is different since we want to control both the accuracy ε and the stability μ with respect to the measurement space W.

We describe the greedy algorithm for constructing σ-admissible families, and explain how it should be modified for (ε, μ)-admissible families. For simplicity we consider the case where Y is a rectangular domain with sides parallel to the main axes, the extension to a more general bounded

domain Y being done by embedding it in such a hyper-rectangle. We are given a prescribed target value $\sigma > 0$ and the splitting procedure starts from Y.

At step j, a disjoint partition of Y into rectangles $(\mathrm{Y}_k)_{k=1,\dots,K_j}$ with sides parallel to the main axes has been generated. It induces a partition of $\mathcal{M}$ given by

$$\mathcal{M}_k \coloneqq \{u(y) \,:\, y \in \mathrm{Y}_k\}, \quad k = 1, \dots, K_j.$$

To each $k \in \{1, \dots, K_j\}$ we associate a hierarchy of affine reduced basis spaces

$$V_{n,k} = \bar{u}_k + \bar{V}_{n,k}, \quad n = 0, \dots, m.$$

where $\bar{u}_k = u(\bar{y}_k)$ with $\bar{y}_k$ the vector defined as the center of the rectangle Y_k. The nested linear spaces

$$\bar{V}_{0,k} \subset \bar{V}_{1,k} \subset \cdots \subset \bar{V}_{m,k}, \quad \dim(\bar{V}_{n,k}) = n,$$

are meant to approximate the translated portion of the manifold $\mathcal{M}_k - \bar{u}_k$. For example, they could be reduced basis spaces obtained by applying the greedy algorithm to $\mathcal{M}_k - \bar{u}_k$, or spaces resulting from local n-term polynomial approximations of $u(y)$ on the rectangle Y_k. Each space $V_{n,k}$ has a given accuracy bound and stability constant

$$\mathrm{dist}(\mathcal{M}_k, V_{n,k}) \leq \varepsilon_{n,k} \quad \text{and} \quad \mu_{n,k} \coloneqq \mu(\bar{V}_{n,k}, W).$$

We define the test quantity

$$\tau_k = \min_{n=0,\dots,m} \mu_{n,k}\varepsilon_{n,k}. \tag{3.4.7}$$

If $\tau_k \leq \sigma$, the rectangle Y_k is not split and becomes a member of the final partition. The affine space associated to $\mathcal{M}_k$ is

$$V_k = \bar{u}_k + \bar{V}_k,$$

where $V_k = V_{n,k}$ for the value of n that minimizes $\mu_{n,k}\varepsilon_{n,k}$. The rectangles Y_k with $\tau_k > \sigma$ are, on the other hand, split into a finite number of sub-rectangles in a way that we discuss below. This results in the new larger partition $(\mathrm{Y}_k)_{k=1,\dots,K_{j+1}}$ after relabelling the Y_k. The algorithm terminates at the step j as soon as $\tau_k \leq \sigma$ for all $k = 1, \dots, K_j = K$, and the family $(V_k)_{k=1,\dots,K}$ is σ-admissible. In order to obtain an (ε, μ)-admissible family, we simply modify the test quantity τ_k by defining it instead as

$$\tau_k \coloneqq \min_{n=0,\dots,m} \max\left\{\frac{\mu_{n,k}}{\mu}, \frac{\varepsilon_{n,k}}{\varepsilon}\right\}$$

and splitting the cells for which $\tau_k > 1$.

The splitting of one single rectangle Y_k can be performed in various ways. When the parameter dimension d is moderate, we may subdivide each side-length at the mid-point, resulting into 2^d sub-rectangles of equal size. This splitting becomes too costly as d gets large, in which case it is preferable to make a choice of $i \in \{1, \dots, d\}$ and subdivide Y_k at the mid-point of the side-length in the i-coordinate, resulting in only 2 sub-rectangles. In order to decide which coordinate to pick, we consider the d possibilities and take the value of i that minimizes the quantity

$$\tau_{k,i} = \max\{\tau_{k,i}^-, \tau_{k,i}^+\},$$

where $(\tau_{k,i}^-, \tau_{k,i}^+)$ are the values of τ_k for the two subrectangles obtained by splitting along the i-coordinate. In other words, we split in the direction that decreases τ_k most effectively. In order

to be certain that all sidelength are eventually split, we can mitigate the greedy choice of i in the following way: if Y_k has been generated by l consecutive refinements, and therefore has volume $|\mathrm{Y}_k| = 2^{-l}|Y|$, and if l is even, we choose $i = (l/2 \bmod d)$. This means that at each even level we split in a cyclic manner in the coordinates $i \in \{1, \dots, d\}$.

Using such elementary splitting rules, we are ensured that the algorithm must terminate. Indeed, we are guaranteed that for any $\eta > 0$, there exists a level $l = l(\eta)$ such that any rectangle Y_k generated by l consecutive refinements has side-length smaller than 2η in each direction. Since the parameter-to-solution map is continuous, for any $\varepsilon > 0$, we can pick $\eta > 0$ such that

$$\|y - \tilde{y}\|_{\ell^\infty} \leq \eta \implies \|u(y) - u(\tilde{y})\| \leq \varepsilon, \quad y, \tilde{y} \in Y.$$

Applying this to $y \in \mathrm{Y}_k$ and $\tilde{y} = \bar{y}_k$, we find that for $\bar{u}_k = u(\bar{y}_k)$

$$\|u - \bar{u}_k\| \leq \varepsilon, \quad u \in \mathcal{M}_k.$$

Therefore, for any rectangle Y_k of generation l, we find that the trivial affine space $V_k = \bar{u}_k$ has local accuracy $\varepsilon_k \leq \varepsilon$ and $\mu_k = \mu(\{0\}, W) = 1 \leq \mu$, which implies that such a rectangle would not anymore be refined by the algorithm.

Reduced model selection and recovery bounds

We return to the problem of selecting an estimator within the family $(u_k^*)_{k=1,\dots,K}$ defined by (3.4.6). In an idealized version, the selection procedure picks the value k^* that minimizes the distance of u_k^* to the solution manifold, that is,

$$k^* = \operatorname{argmin}\{\operatorname{dist}(u_k^*, \mathcal{M}) \,:\, k = 1, \dots, K\} \tag{3.4.8}$$

and takes for the final estimator

$$u^* = u^*(w) \coloneqq A_{k^*}(w) = u_{k^*}^*(w). \tag{3.4.9}$$

Note that k^* also depends on the observed data w. This estimation procedure is not realistic since the computation of the distance of a known function v to the manifold

$$\operatorname{dist}(v, \mathcal{M}) = \min_{y \in Y} \|u(y) - v\|,$$

is a high-dimensional non-convex problem which necessitates to explore the whole solution manifold. A more realistic procedure is based on replacing this distance by a surrogate quantity $\mathcal{S}(v, \mathcal{M})$ that is easily computable and satisfies a uniform equivalence

$$r \operatorname{dist}(v, \mathcal{M}) \leq \mathcal{S}(v, \mathcal{M}) \leq R \operatorname{dist}(v, \mathcal{M}), \quad v \in V,$$

for some constants $0 < r \leq R$. We then instead take for k^* the value that minimizes this surrogate, that is,

$$k^* = \operatorname{argmin}\{\mathcal{S}(u_k^*, \mathcal{M}) \,:\, k = 1, \dots, K\}. \tag{3.4.10}$$

Before discussing the derivation of $\mathcal{S}(v, \mathcal{M})$ in concrete cases, we establish a recovery bound in the absence of model bias and noise.

Theorem 3.4.1. *Assume that the family $(V_k)_{k=1,\dots,K}$ is σ-admissible for some $\sigma > 0$. Then, the idealized estimator based on (3.4.8), (3.4.9), satisfies the worst case error estimate*

$$E_{\mathrm{wc}}(A_{k^*}, \mathcal{M}) = \max_{u \in \mathcal{M}} \|u - u^*(P_W u)\| \leq \delta_\sigma,$$

where δ_σ is the benchmark quantity defined in (3.4.2). When using the estimator based on (3.4.10), the worst case error estimate is modified into

$$E_{\mathrm{wc}}(A_{k^*}, \mathcal{M}) \leq \delta_{\kappa\sigma}, \quad \kappa = \frac{R}{r} > 1.$$

In the above result, we do not obtain the best possible accuracy satisfied by the different u_k^*, since we do not have an oracle providing the information on the best choice of k. We can show that this order of accuracy is attained in the particular case where the measurement map P_W is injective on $\mathcal{M}$ (which implies $\delta_0 = 0$).

Theorem 3.4.2. *Assume that $\delta_0 = 0$ and that*

$$\mu(\mathcal{M}, W) = \frac{1}{2} \sup_{\sigma > 0} \frac{\delta_\sigma}{\sigma} < \infty.$$

Then, for any given state $u \in \mathcal{M}$ with observation $w = P_W u$, the estimator u^* obtained by the model selection procedure (3.4.10) satisfies the oracle bound

$$\|u - u^*\| \leq C \min_{k=1,\dots,K} \|u - u_k^*\|, \quad C \coloneqq 2\mu(\mathcal{M}, W)\kappa.$$

In particular, if $(V_k)_{k=1,\dots,K}$ is σ-admissible, it satisfies

$$\|u - u^*\| \leq C\sigma.$$

We next discuss how to incorporate model bias and noise in the recovery bound, provided that we have a control on the stability of the PBDW method, through a uniform bound on μ_k, which holds when we use (ε, μ)-admissible families.

Theorem 3.4.3. *Assume that the family $(V_k)_{k=1,\dots,K}$ is (ε, μ)-admissible for some $\varepsilon > 0$ and $\mu \geq 1$. If the observation is $w = P_W u + \eta$ with $\|\eta\| \leq \varepsilon_{noise}$, and if the true state does not lie in $\mathcal{M}$ but satisfies $\mathrm{dist}(u, \mathcal{M}) \leq \varepsilon_{model}$, then, the estimator based on (3.4.10) satisfies the estimate*

$$\|u - u^*(w)\| \leq \delta_{\kappa\rho} + \varepsilon_{noise}, \quad \rho \coloneqq \mu(\varepsilon + \varepsilon_{noise}) + (\mu + 1)\varepsilon_{model}, \quad \kappa = \frac{R}{r},$$

and the idealized estimator based on (3.4.8) satifies a similar estimate with $\kappa = 1$.

A numerical example: constructing σ-admissible families:

In this example we examine the behavior of the splitting scheme to construct σ-admissible families in the example of the elliptic PDE with a checkerboard diffusion field. The manifold $\mathcal{M}$ is thus given by the solutions to equation (3.2.15) associated to the diffusivity field

$$a(y) = \overline{a} + \sum_{\ell=1}^{d} c_\ell \chi_{D_\ell} y_\ell, \quad y \in Y,$$

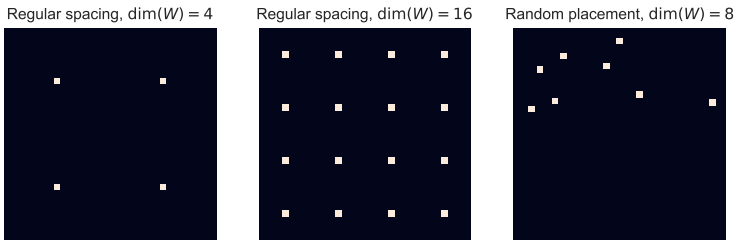


Figure 3.7: Measurement locations.

where χ_{D_ℓ} is the indicator function on the set D_ℓ, and parameters ranging uniformly in $Y = [-1, 1]^p$. We study the impact of the intrinsic dimensionality of the manifold by considering two cases for the partition of the unit square D, a 2×2 uniform grid partition resulting in $p = 4$ parameters, and a 4×4 grid partition of D resulting in $p = 16$ parameters. We also study the impact of coercivity and anisotropy on our reconstruction algorithm by examining the different manifolds generated by taking $c_\ell = c_1 \ell^{-r}$ with $c_1 = 0.9$ or 0.99 and $r = 1$ or 2. The value $c_1 = 0.99$ corresponds to a severe degeneration of coercivity, and the rate $r = 2$ corresponds to a more pronounced anisotropy.

We use two different measurement spaces, one with $m = \dim(W) = 4$ evenly spaced local averages and the other with $m = 16$ evenly spaced local averages. The measurement locations are shown diagrammatically in Figure 3.7. The local averages are taken as squares of side-length 2^{-6}. Note that the two values $m = 4$ and $m = 16$ which we consider for the dimension of the measurement space are the same as the parameter dimensions $p = 4$ and $p = 16$ of the manifolds. This allows us to study different regimes:

- When $m < p$, we have a highly ill-posed problem since the intrinsic dimension of the manifold is larger than the dimension of the measurement space. In particular, we expect that the fundamental barrier $\delta_0(\mathcal{M})$ is strictly positive. Thus we cannot expect very accurate reconstructions even with the splitting strategy.

- When $m \geq p$, the situation is more favorable and we can expect that the reconstruction involving manifold splitting brings significant accuracy gains.

As in the previous case, the training set $\widetilde{\mathcal{M}}$ is generated by a subset $\widetilde{Y}_{\text{tr}} = \{y_j^{\text{tr}}\}_{j=1,\dots,N_{\text{tr}}}$ of $N_{\text{tr}} = 5000$ samples taken uniformly on Y. We build the σ-admissible families outlined in §3.4 using a dyadic splitting and the splitting rule is given by (3.4.7). For example, our first split of Y results in two rectangular cells Y_1 and Y_2, and the corresponding collections of parameter points $\widetilde{Y}_1 \subset Y_1$ and $\widetilde{Y}_2 \subset Y_2$, as well as split collections of solutions $\widetilde{\mathcal{M}}_1$ and $\widetilde{\mathcal{M}}_2$. On each $\widetilde{\mathcal{M}}_k$ we apply the greedy selection procedure, resulting in V_k, with computable values μ_k and ε_k. The coordinate direction in which we split Y is precisely the direction that gives us the smallest resulting $\sigma = \max_{k=1,2} \mu_k \varepsilon_k$, so we need to compute greedy reduced bases for each possible splitting direction before deciding which results in the lowest σ. Subsequent splittings are performed in the same manner, but at each step we first chose cell $k_{\text{split}} = \arg\max_{k=1,\dots,K} \mu_k \varepsilon_k$ to be split.

After $K - 1$ splits, the parameter domain is divided into $Y = \bigcup_{k=1}^K Y_k$ disjoint subsets Y_k and we have computed a family of K affine reduced spaces $(V_k)_{k=1,\dots,K}$. For a given $w \in W$, we have K possible reconstructions $u_1^*(w), \dots u_K^*(w)$ and we select a value k^* with the surrogate based model selection given in equation (3.4.10). The test is done on a test set of $N_{\text{te}} = 1000$ snapshots which are different from the ones used for the training set $\widetilde{\mathcal{M}}$.

In Figure 3.8 we plot the reconstruction error, averaged over the test set, as a function of the number of splits K for all the different configurations: we consider the 2 different diffusivity fields $a(y)$ with $p = 4$ and $p = 16$ parameters, the two measurement spaces of dimension $m = 4$ and $m = 16$, and the 4 different ellipticity/coercivity regimes of c_ℓ in $a(y)$. We also plot the error when taking for k^* the *oracle value* that corresponds to the value of k that contains the parameter y which gave rise to the snapshot and measurement.

Our main findings can be summarized as follows:

1. The error decreases with the number of splits. As anticipated, the splitting strategy is more effective in the overdetermined regime $m \geq p$.

2. Degrading coercivity has a negative effect on the estimation error, while anisotropy has a positive effect. In our computations, a larger r in c_ℓ corresponds to a higher degree of anisotropy, and in turn to a reduced width of the solution manifold $\mathcal{M}$ in dimensions associated with the less active coordinates. Hence it is no surprise that the approximation errors from our algorithm are lower for these higher anisotropy examples.

3. Choosing k^* by the surrogate based model selection yields error curves that are above yet close to those obtained with the oracle choice. The largest discrepancy is observed when coercivity degrades.

Figure 3.9 presents the error bounds $\sigma_K := \max_{k=1,\dots,K} \mu_k \varepsilon_k$ which are known to be upper bounds for the estimation error when choosing the oracle value for k^* at the given step K of the splitting procedure. We observe that these worst upper bounds have similar behaviour as the averaged error curves depicted on Figure 3.8. In Figure 3.10, for the particular configuration $\dim(Y) = \dim(W) = 16$, we demonstrate that σ_K indeed acts as an upper bound for the worst case error of the oracle estimator.

3.5 Applications

In the previous sections, I have summarized the main theoretical aspects of a deterministic theory to address inverse state and parameter estimation problems. The algorithms that we have developed come with certain optimality guarantees regarding the quality of approximation. This feature, and the fact that our theory is formulated in very general terms, is a great opportunity to contribute to different applications since many of them require quality certificates. That is why, in parallel to the above theoretical developments, I have devoted considerable efforts to bring the above methodology into concrete applications, and this section summarizes my works on this front. Interestingly, even though our theoretical framework is formulated in a very general way, each application comes with specific challenging features that have led us to enlarge some aspects of our approach. One salient example are biomedical applications where we must take into account the morphological variations that inevitably arise between individuals. Therefore we must enlarge the theory and consider families of domains instead of a fixed domain Ω (see Section 3.5.2).

3.5.1 Neutronics and collaboration with EDF

In the context of Helin Gong's CIFRE PhD thesis, the company EDF was interested in addressing inverse problems arising in the field of neutronics with our methodology involving reduced model

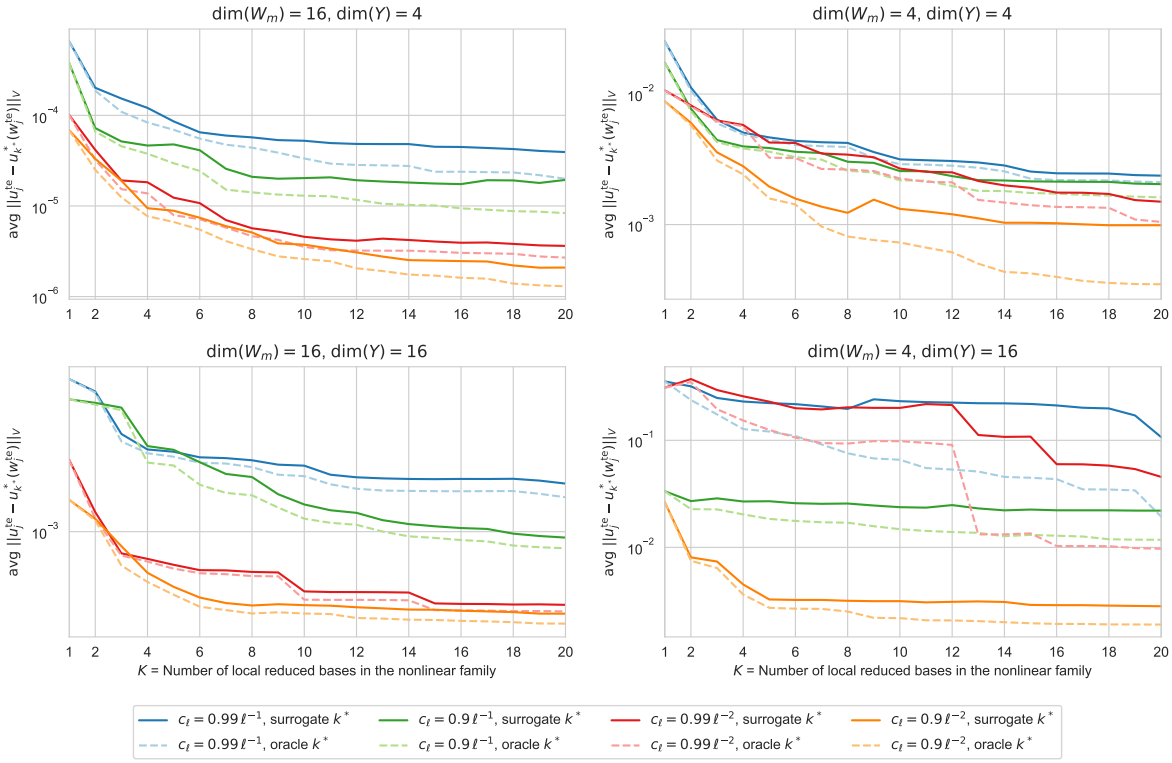


Figure 3.8: Average of errors $\|u_j^{\mathrm{te}} - u_{k^*}^*(w_j^{\mathrm{te}})\|$ for different choices of k^* .

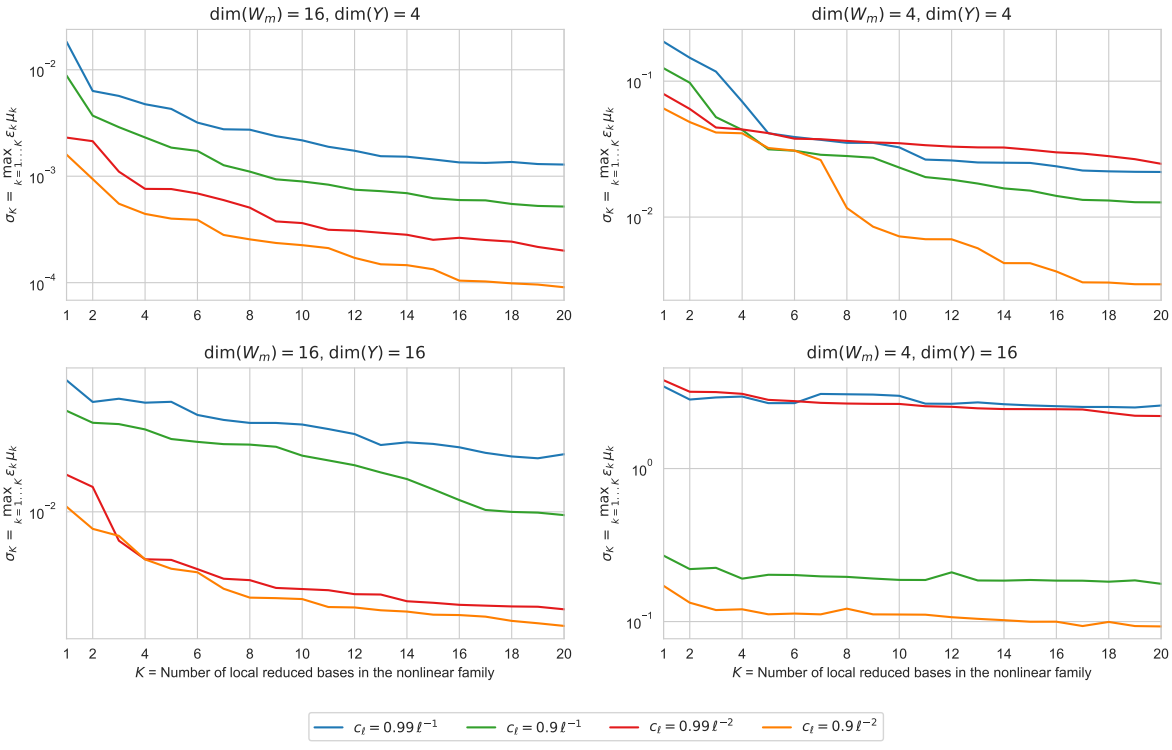


Figure 3.9: Error bounds of local linear families, given by $\sigma_K = \max_{k=1\ldots K} \mu_k \varepsilon_k$.

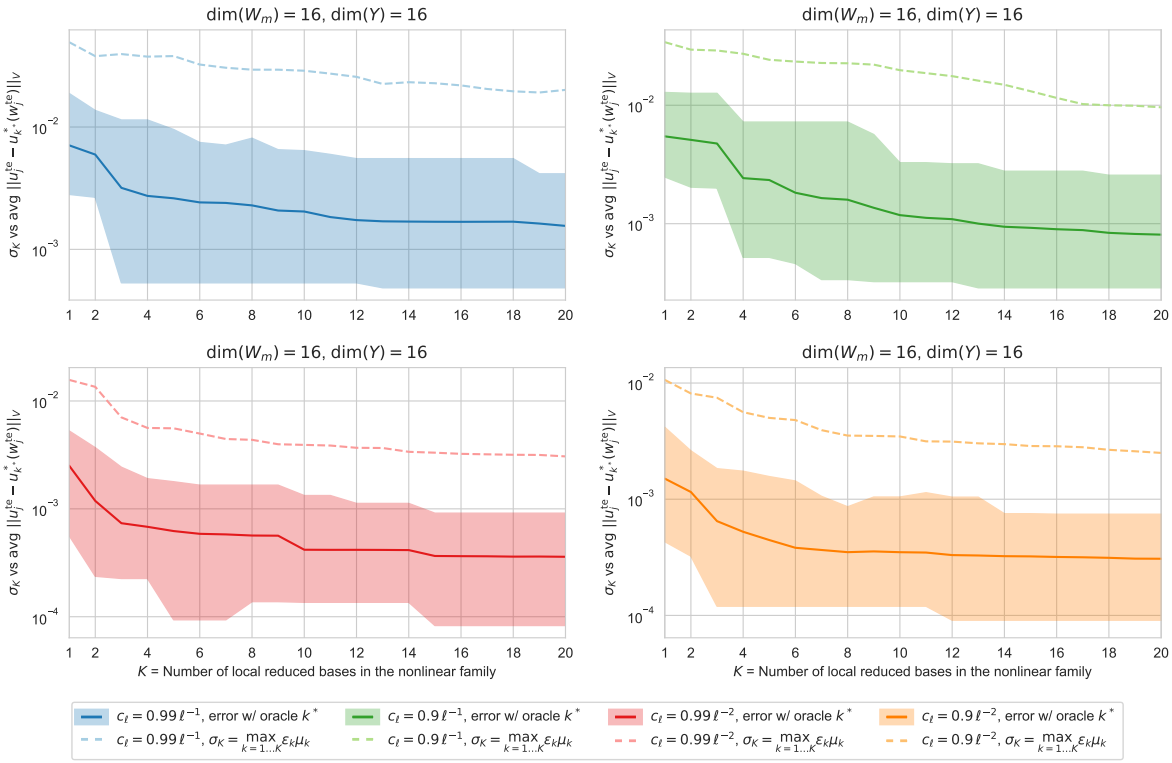


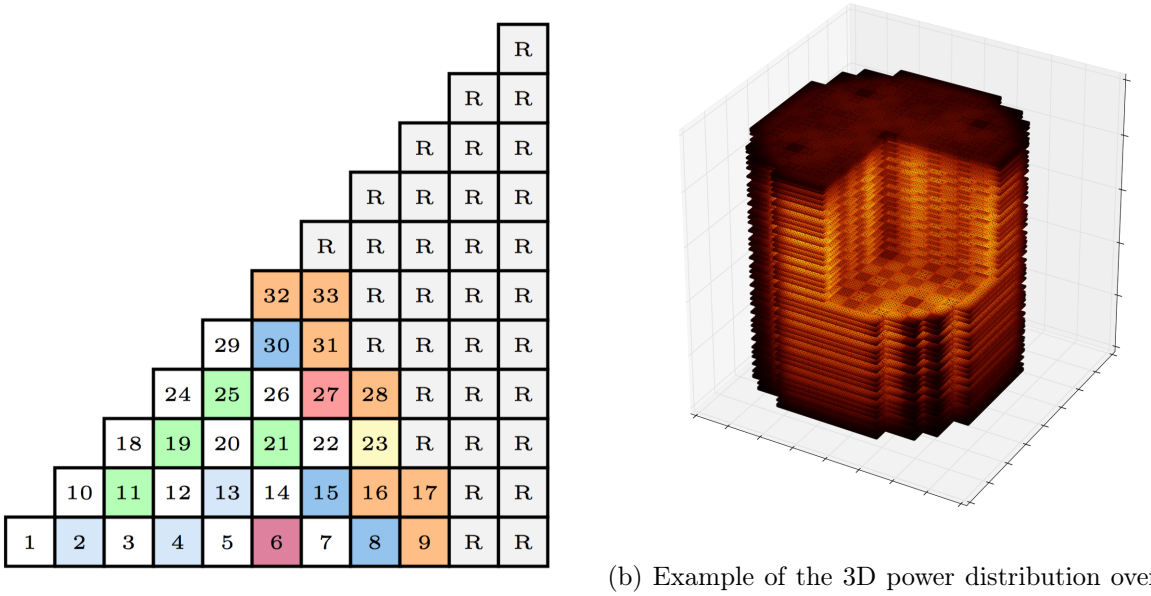
Figure 3.10: Comparison between σ_K (dashed curve), the averaged oracle error (full curve) and the the range from maximum to minimum oracle error (shaded region).

spaces. The thesis was co-supervised by Yvon Maday (Sorbonne Université) and myself on the academic side, and by Jean-Philippe Argaud and Bertrand Bouriquet, research engineers at EDF.

In order to position better our contribution, we need to recall some elements of context regarding the main scientific challenges in the field of nuclear engineering. An essential element to bear in mind is that the production of nuclear energy is done under very high safety standards. Understandably, this leads to a certain type of extremely conservative manner of carrying design, safety studies in which the experience of engineers plays a crucial role in order to find acceptable configurations. Due to the complexity of the physics, it is usually necessary to combine the expertise of engineers from different fields and the modelling/optimization process often require several iterations between different experts before satisfying all the desired criteria. These iterations take a lot of effort, and are time-consuming.

In this context, our collaboration with EDF can be seen as a transfer of knowledge to the nuclear industry, in which the main goal was to help make some of the above tasks become more *agile and systematic* by blending in a more natural manner multiphysics phenomena and information of different nature (PDE models and data observations). The fact that computations take place in reduced spaces is also a crucial point to accelerate calculations.

In [A9, P1], we applied some of the above techniques to the field of neutronics. The goal was to give a proof of concept that our methods were able to do fast state estimation of the population of neutrons (and some related quantities of interest such as the power distribution). Our numrical tests were done with synthetic measurements but we worked on realistic reactor core configurations. As a parametrized PDE model, we used a two-group neutron diffusion equation (obviously, more



(a) Eighth of core (1-33: fuel assembly; R: reflector).

(b) Example of the 3D power distribution over the core. Computed with the COCAGNE code [21].

Figure 3.11: The fuel assembly loading scheme and an example of 3D power distribution over the core in a realistic 1450 MWe reactor at EDF.

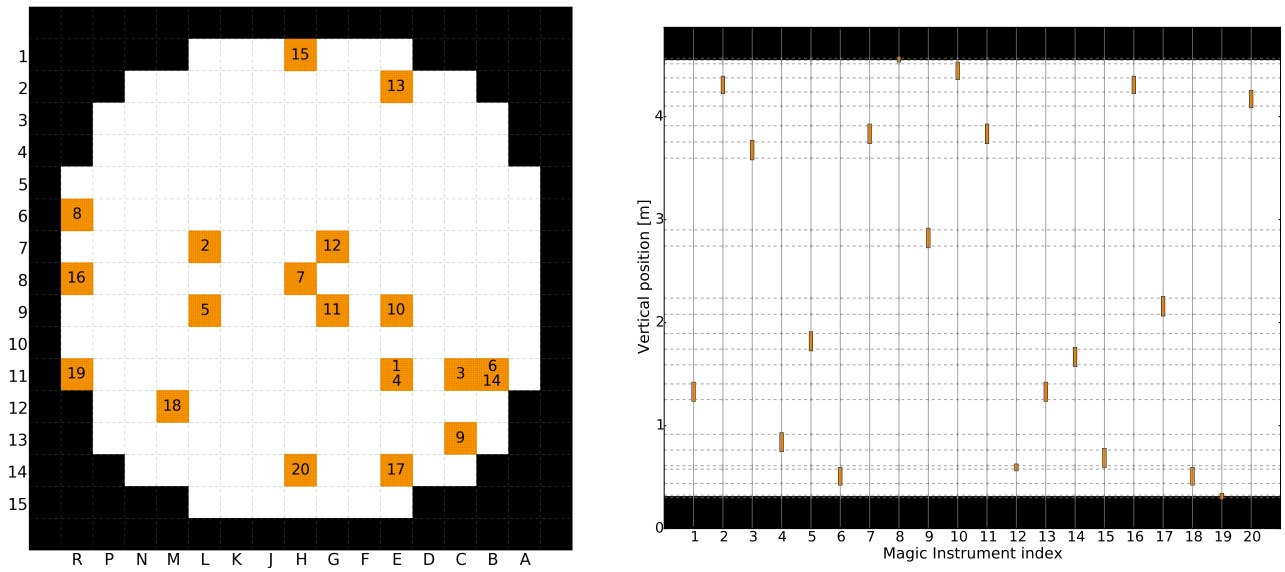
complicated models could be considered). This model is actually the diffusion limit of the Boltzmann equation that we discussed in Section 2.1. In [A9], we gave results revolving around the search for optimal sensor locations to measure certain quantities of interest during the operation of the core. For this, we used the Generalized Empirical Interpolation Method (GEIM, [T1, A14, A13, A12]). This method is a particular version of the linear PBDW method in which $m = n$. The method was originally introduced in my PhD thesis (see [T1, A14, A13, A12]). Another possibility could have been to use the greedy algorithm that we developped in Section 3.3. Our numerical experiments were done on several reactor geometries. Figure 3.11 illustrates the geometry of a realistic Pressurized Water Reactor of 1450 MWe which is studied by EDF, and Figure 3.12 shows the 20 first locations for sensor positioning that our algorithms gave.

Our collaboration also led to works on how to do state estimation in presence of measurement noise. In [P2, 10], we developed a reconstruction strategy involving least-squares projections with constrains using some a priori knowledge of the geometry of the manifold formed by all the possible physical states of the system.

3.5.2 Biomedical Applications and Problems with shape variability

In this section I summarize a series of three works [A2, A3, S1] in the field of biomedical applications. The results have been obtained in the framework of Felipe Galarce's PhD thesis at the Inria Commedia group, which I have co-supervised with D. Lombardi, and J. F. Gerbeau.

The overarching topic to which we have contributed is related to the challenge of developing numerical tools to assist medical doctors in their decisions and diagnoses. This requires to solve both quickly and in a reliable manner data assimilation and inverse problems, which share the



(a) Radial (b) Axial

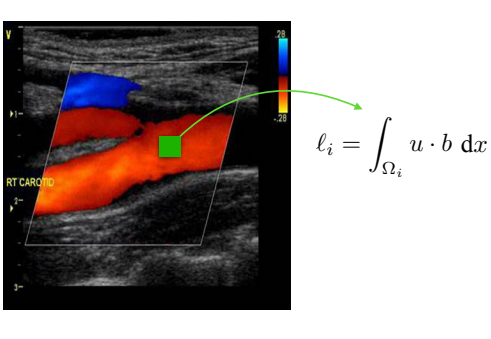
Figure 3.12: The first 20 sensors placements suggested by our algorithms on a realistic PWR reactor of 1450 MWe.

following general common features:

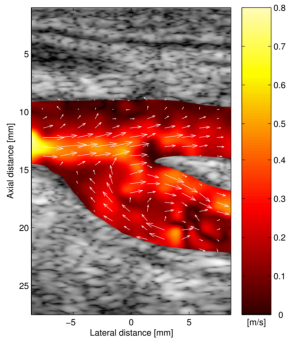
- The problems are typically posed on certain organs or portions of the body which inevitably involve morphological variations.
- The available data are often corrupted by noise and obtained with medical imaging techniques, which have the advantage of being *non-invasive*.
- There may be morphological constraints that prevent from measuring at specific locations.
- In some cases, the device may not be able to measure directly the desired quantity of interest (QoI), and a post-processing may be required to obtain an estimate of it.
- In order to be of practical use, the problems need to be solved in close to real time in order that their outputs can be taken into account by the doctors in the process of diagnosing.

State estimation of blood velocity and pressure on carotid arteries using Doppler ultrasound velocity images: As a guiding example, we have focused on a specific problem which is the one of reconstructing 3D blood velocity, and pressure fields as well as related quantities by using doppler ultrasound velocity images. This type of imaging is one of the most used, clinically available technologies to monitor blood flows in the heart cavity and in several segments of the vascular tree. Its main advantages are that it is fast, non-invasive, and cheap. Its main drawback lies in the space resolution: observations are noisy averages over some voxels of the projection of the velocity field over a given direction or over a given plane. Figure 3.13 gives examples of velocity ultrasound images. In this framework, providing full 3D reconstructions of blood velocity, and pressure fields is of interest since it could enrich the available information used in the diagnosis of

certain pathologies (e.g. stenosis in patients with sickle cell disease) which is currently solely relying on the ultrasound images, and the crucial experience and intuition of medical doctors.



(a) Color flow image (CFI)



(b) Vector flow image (VFI)

Figure 3.13: Velocity image of the common carotid bifurcation of two ultrasound devices.

In [A2, A3], we illustrate how the general methodology which we have explained in Sections 3.1 to 3.4 can be applied to this specific inverse problem. Due to our lack of real ultrasound images, our experiments present certain limitations: we have worked with synthetic images and have used an admittedly simple Gaussian modelling of the ultrasound noise (Doppler ultrasound images present a very involved space-time structure which is not the main topic of our work). The manifold $\mathcal{M}$ considered to describe the haemodynamics is a family of parametrized incompressible Navier-Stokes equations. This model is generally acknowledged to be accurate for large vessels such as the carotid artery, which is why we assume in the following that there is no model error and that the true system is governed by these equations. This is admittedly a rather strong assumption but note that we are also led to it because it is not possible to study the impact of the model error without real measurements.

The Navier-Stokes equations model the dynamics of a fluid with density $\rho \in \mathbb{R}^+$ and dynamic viscosity $\mu \in \mathbb{R}^+$. In our case, the fluid under consideration is blood. For all times $t \in [0, T]$, the blodd velocity and pressure fields $(u(t), p(t))$ are the solution to

$$\begin{cases} \rho \dfrac{\partial u}{\partial t}(t) + \rho\, u(t) \cdot \nabla u(t) - \mu \Delta u(t) + \nabla p(t) = 0, & \text{in } \Omega \\ \nabla \cdot u(t) = 0, & \text{in } \Omega. \end{cases}$$

The equations are closed by prescribing an initial condition and boundary conditions. A weak formulation of this equation makes the problem be well posed when we seek $(u(t), p(t))$ in the space $H^1(\mathrm{div}, \Omega) \times L^2(\Omega)$, where

$$H^1(\mathrm{div}, \Omega) \coloneqq \{f \in L^2(\Omega, \mathbb{R}^3) \ : \ \mathrm{div}(f) = 0\}$$

is the space of divergence free fields.

Our Navier-Stokes modeling of the blood dynamics involves quantities such as the heart rate which we take as parameters $y \in Y \in \mathbb{R}^p$. Since the dynamics behaves with a period given by the heart beats, time will be considered as a parameter so t will be one of the coordinates of y. A manifold $\mathcal{M}$ is generated by the variations of the parameters

$$\mathcal{M} \coloneqq \{(u(y), p(y)) \in V \ : \ y \in \mathrm{Y}\}. \tag{3.5.1}$$

In the following, instead of working with $H^1(\mathrm{div},\Omega)$ for the velocity space, we work with $H^1(\Omega,\mathbb{R}^3)$. That is, we take

$$V = U \times V \coloneqq H^1(\Omega,\mathbb{R}^3) \times L^2(\Omega) \tag{3.5.2}$$

as the ambient space so that we view $\mathcal{M}$ as a subset of V. The reason for choosing to work in $H^1(\Omega,\mathbb{R}^3)$ instead of $H^1(\mathrm{div},\Omega)$ is because the computation of the Riez representers ω_i is facilitated.

In [A2], we focus on the reconstruction of the 3D velocity field using PBDW. The manifold is therefore slighlty simplified to

$$\mathcal{M}_{(\mathrm{vel})} \coloneqq \{u(y) \,:\, y \in \mathrm{Y}\} \subset U.$$

We study the impact of working with different types of reduced models:

- A plain linear (or affine) version in which V_n is given by the Singular Value Decomposition of $\mathcal{M}_{(\mathrm{vel})}$.
- A piece-wise affine version where we avoid the step (3.4.10) of model selection thanks to the fact that, in our construction, it is possible to know in the online phase to which partition of the manifold the solution belongs to.
- A non-linear data-driven version in which V_n is built online with a greedy OMP algorithm that uses the observation ω. The algorithm goes as follows. Let

  $$\mathcal{D} \coloneqq \{v = u/\|u\| \;:\; u \in \mathcal{M}_{(\mathrm{vel})},\ u \neq 0\}.$$

  be the set of normalized functions from $\mathcal{M}$. The first element φ_1 is chosen as

  $$\varphi_1 = \frac{1}{\#\mathcal{D}} \sum_{v \in \mathcal{D}} v,$$

  and we set $V_1 \coloneqq \mathrm{span}\{\varphi_1\}$. For $n > 1$, given $V_n = \mathrm{span}\{\varphi_1,\dots,\varphi_n\}$, we select

  $$\varphi_{n+1} \in \arg\max_{v \in \mathcal{D}} \left| \left\langle w - P_{P_{W_m} V_n} w, \frac{P_{W_m} v}{\|P_{W_m} v\|} \right\rangle \right|$$

  where $P_{W_m} V_n = \mathrm{span}\{P_{W_m}\varphi_1,\dots,P_{W_m}\varphi_n\}$. We set $V_{n+1} = \mathrm{span}\{V_n, \varphi_{n+1}\}$. The algorithm can easily be extended to build an affine space $\bar{u} + V_n$ for a given $\bar{u}$. For this, we introduce $\bar{\omega} = P_{W_m}\bar{u}$ and the shifted set

  $$\delta_{\bar{u}}\mathcal{D} = \left\{ v = \frac{u - \bar{u}}{\|u - \bar{u}\|} \;:\; u \in \mathcal{M}_{(\mathrm{vel})},\ u \neq \bar{u} \right\},$$

  and it suffices to apply the previous greedy algorithm to the target function $\omega - \bar{\omega}$ instead of ω and do the search over $\delta_{\bar{u}}\mathcal{D}$ instead of $\mathcal{D}$.

Our main conclusion is that, for the example under consideration, all methods deliver a good accuracy but the piece-wise affine approach is the most accurate. We have thus worked with this type of approach in [A3] in order to show how to reconstruct unobserved fields such as the pressure from ultrasound velocity observations. This can easily be accomplished with our approach by performing a joint velocity and pressure reconstruction. For this, we consider the full velocity-pressure manifold (3.5.1) and we endow the space (3.5.2) with the external direct sum and product

structure to build a Hilbert space. That is, for any two elements (u_1, p_1) and (u_2, p_2) of $V = U \times P$ and any scalar $\alpha \in \mathbb{R}$,

$$(u_1, p_1) + (u_2, p_2) = (u_1 + u_2, p_1 + p_2), \quad \alpha(u_1, p_1) = (\alpha u_1, \alpha p_1).$$

The inner product is defined as the sum of component-wise inner products

$$\langle (u_1, p_1), (u_2, p_2) \rangle_V = \langle u_1, u_2 \rangle_U + \langle p_1, p_2 \rangle_P ,$$

and it induces a norm on V,

$$\|(u, p)\| \coloneqq (\langle (u, p), (u, p) \rangle_V)^{1/2}, \quad \forall (u, p) \in V.$$

When we are given partial information on (u, p) from Doppler velocity measures, we are given the projection

$$\omega = P_{W_m}(u, p),$$

where W_m is the observation space

$$W_m \coloneqq W_m^{(u)} \times \{0\} = \operatorname{span}\{\omega_1, \ldots, \omega_m\} \times \{0\} \subset V,$$

and the ω_i are the Riesz representers in U of each voxel $\ell_i \in U'$ of the imaging device,

$$\langle \omega_i, v \rangle_U = \ell_i(v) = \int_{\Omega_i} v \cdot b \, \mathrm{d}x, \quad \forall v \in U.$$

We are now in position to apply directly the reconstruction algorithms from the previous sections to do the joint reconstruction of (u, p) with the current particular choice of Hilbert space V and observation space W_m. Note that, in general, stability is degraded in the joint reconstruction compared to the single velocity reconstruction. In fact, if the reduced model is taken as a product of two reduced spaces, namely, if $V_n = V_{n_u}^{(u)} \times V_{n_p}^{(p)}$ with $V_{n_u}^{(u)} \subset U$ and $V_{n_p}^{(p)} \subset P$, we can easily prove that if the inf-sup constant in the single velocity space satisfies

$$\beta(V_{n_u}^{(u)}, W_m^{(u)}) > 0,$$

with $V_{n_u}^{(u)}$ and $W_m^{(u)} \subset U$, then the inf-sup of the joint reconstruction satisfies

$$0 < \beta(V_{n_u}^{(u)} \times V_{n_p}^{(p)}, W_m^{(u)} \times \{0\}) \leq \beta(V_{n_u}^{(u)}, W_m^{(u)}).$$

On the one hand, the left-hand side of the bound guarantees that the joint reconstruction is well-posed. On the other hand, from the right-hand side it follows that stability cannot be better than the one of the single velocity reconstruction.

Figures 3.14 to 3.18 give some illustrations of our reconstructions on an example of a carotid bifurcation. The imaged velocity field ω is the projection of the velocity on a 2D plane on the region before the carotid bifurcation as shown in Figure 3.14. We refer to [A3] for the presentation of the error reconstruction study associated to the numerical example.

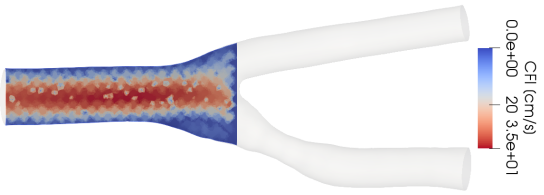
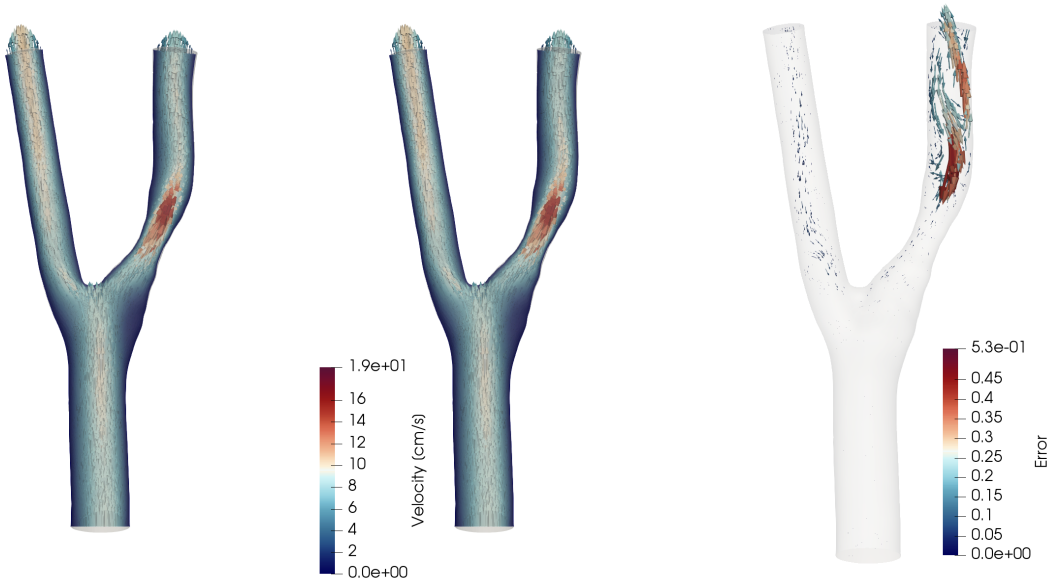
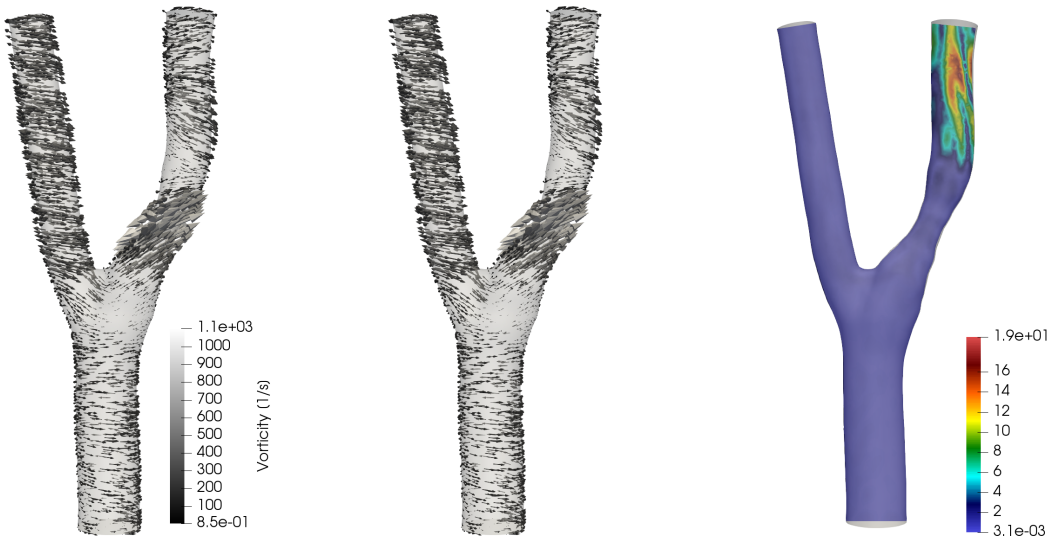


Figure 3.14: Example of observation $P_W u$: synthetic CFI of the common carotid branch with $m = 233$ voxels of 0.15 [cms] each (image from the systole period). Note that we only receive information before the bifurcation.



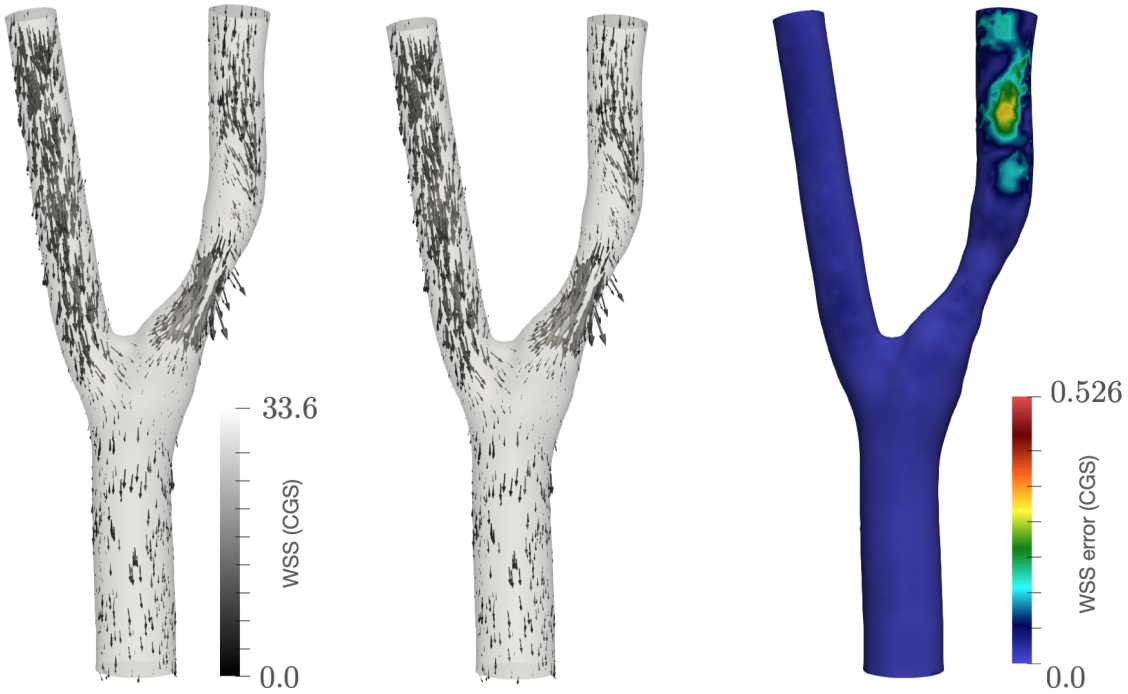
(a) u. (b) u^* (c) $u - u^*$

Figure 3.15: Example of reconstruction of the velocity. We observe a region with higher errors close to the stenosis.



(a) Θ. (b) Θ^* (c) $\Theta - \Theta^*$

Figure 3.16: Example of reconstruction of the vorticity $\Theta := \nabla \times u$.



(a) $\mathcal{S}(u)$. (b) $\mathcal{S}^*(u)$ (c) $\mathcal{S}(u) - \mathcal{S}^*(u)$

Figure 3.17: Example of reconstruction of the wall shear stress. This quantity is a mapping $\mathcal{S} : U \to [H^{-1/2}(\partial\Omega_{\text{wall}})]^3$ that returns the tangential component of the force that the blood applies on the vessel wall $\mathcal{S}(u) := 2\mu \{I - n \otimes n\} \left(\frac{\nabla u + \nabla u^T}{2} n \right)$, on $\partial\Omega_{\text{wall}}$.

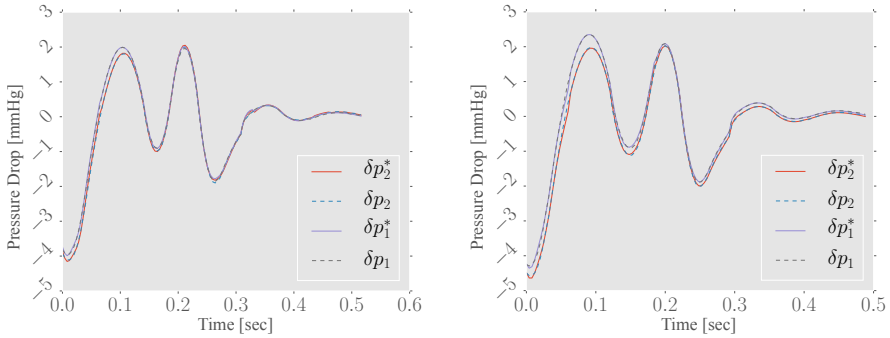


Figure 3.18: Pressure drops at the two outlets in two different simulations using the joint velocity and pressure reconstruction method with the piecewise linear algorithm. The dashed lines show the ground truth δp_1 and δp_2. The continuous lines show the reconstruction. The vertical axis shows the pressure drop in [mmHg] and the horizontal axis the time in seconds.

Multidomain State Estimation: The speed of the affine PBDW reconstruction algorithm crucially relies on the fact that we have assumed that the spatial domain Ω is given to us a priori. Thanks to this we can precompute the reduced models $V_n(\Omega)$ before the reconstruction takes place, and we only need to solve (3.2.6) during the reconstruction, which is a computation that can be done in near real-time. The offline computation of the reduced model should be seen as a training phase, and it can be computationally intensive and time-consuming for complex physical systems.

There are however cases in which we cannot assume that Ω is given a priori. This situation typically arises in biomedical applications where state estimation needs to be performed on a certain part of the body for different patients which inevitably present morphological variations. One example is the application discussed above. In this case, given a new target geometry Ω, one could of course generate $\mathcal{M}(\Omega)$ and derive a reduced model $V_n(\Omega)$ but this task would not be feasible in real-time, and the method would no be useful for real time decisions. To avoid this computational bottleneck, we propose in [S1] a method to quickly build a space $V_n(\Omega)$ by using reduced models which have been pre-computed on a database of template geometries which we suppose to be available offline. The idea consists in finding the best reduced model from the template geometries, and then to transport it to the target geometry Ω. Once this is done, we reconstruct with PBDW on the target geometry. We next outline the details of our proposed strategy.

We consider a set G of spatial domains in $\mathbb{R}^d$. The set can potentially be infinite. An example for G is the set of human carotid arteries or, more generally, the set of shapes of a certain organ. Our goal is to build a state estimation procedure that is fast for every geometry $\Omega \in \mathrm{G}$. For this, our approach is based on a learning phase that involves computations on a dataset of available template geometries. We next summarize the main steps. Some of them involve certain routines which are introduced at an abstract level in the current presentation. We refer to [S1] for further details on possible choices and practical implementations.

Training/Learning phase (offline)

- **Database of Template Geometries:** Gather a family of K template domains

$$\mathrm{G}_{\text{templates}} = \{\Omega_1, \ldots, \Omega_K\} \subseteq \mathrm{G}.$$

This family will serve as a database for our subsequent developments.

- **Database of Template Reduced Models:** For every $\Omega \in \mathrm{G}_{\text{templates}}$, we consider a parameter-dependent PDE
$$\mathcal{P}(u, y) = 0,$$
where the parameters y take values in Y and the solution $u(y)$ belongs to a Hilbert space $V(\Omega)$. Note that the differential operator $\mathcal{P}$ and the parameter domain Y could vary with the geometry Ω. However, to simplify the presentation, we assume that $\mathcal{P}$ and Y are taken identical for all $\Omega \in \mathrm{G}_{\text{templates}}$. The set of solutions yields the solution manifold $\mathcal{M}(\Omega)$ and it describes all the possible physical states of the system under consideration for the given geometry. We summarize the physics by precomputing a template reduced model $V_n(\Omega)$,
$$\mathcal{M}(\Omega) \approx V_n(\Omega), \quad \forall \Omega \in \mathrm{G}_{\text{templates}}.$$

- **Transport snapshots and reduced-models between geometries:**
  - We need to define a map to transport functions between different geometries
  $$\tau_{\Omega\to\Omega'} : V(\Omega) \to V(\Omega'), \quad \forall (\Omega, \Omega') \in \mathrm{G} \times \mathrm{G}.$$
  For some applications, it will be important that τ satisfies some physical properties such as mass conservation.
  - We also need to define a map to transport subspaces into subspaces. For this we introduce the mapping
  $$\widehat{\tau}_{\Omega\to\Omega'} : V_n(\Omega) \subseteq V(\Omega) \to V_{n'}(\Omega') \subseteq V(\Omega'), \quad \forall (\Omega, \Omega') \in \mathrm{G} \times \mathrm{G}.$$
  If $V_n(\Omega)$ is spanned by the family of functions $\{v_i\}_{i=1}^n$, one possibility to define $\widehat{\tau}_{\Omega\to\Omega'}(V_n(\Omega))$ is
  $$\widehat{\tau}_{\Omega\to\Omega'}(V_n(\Omega)) = \mathrm{span}\{\tau_{\Omega\to\Omega'}(v_i)\}_{i=1}^n$$
  which is a space of dimension $n' \leq n$.
  - We refer to [S1] for details on how we have built τ and $\widehat{\tau}$ in practice.

- `Best-Template`: For the reconstruction task, we need to identify for each new target geometry $\Omega \in \mathrm{G}$, which template geometry $\Omega_t \in \mathrm{G}_{\text{templates}}$ has the most appropriate reduced model $V_n(\Omega_t)$ that we have to transport to Ω. For this, we need to build a best template map
$$\begin{aligned} \mathtt{BT} : \mathrm{G} &\to \mathrm{G}_{\text{templates}} \\ \Omega &\mapsto \Omega_t^*. \end{aligned}$$
In our case, the selection strategy is based on defining and estimating distances between reduced models $V_n(\Omega)$ from different geometries $\Omega \in \mathrm{G}$. We use a dimensionality reduction technique called Multi-Dimensional Scaling for this task (MDS, see e.g. [86, 80, 71, 6]). In our work, the distance between two reduced models $V_n(\Omega)$ and $V_n(\Omega')$ is defined as a symetrized version of formula (3.5.3), which is a quantity involved in our numerical analysis of the main underlying mechanisms that drive the reconstruction quality with transported subspaces (see below). We refer to [S1] for details concerning the practical implementation.

Reconstruction phase (online)

We are given a target domain $\Omega \in \mathrm{G}$, and our goal is to give a fast reconstruction of an unknown function $u \in V(\Omega)$ given m measurement observations $\ell(u) = (\ell_i(u))_{i=1}^m$. Note that since $\ell_i \in V'(\Omega)$, the observation space depends on the geometry and $W = W(\Omega)$.

- If $\Omega \in \mathrm{G}_{\text{templates}}$ (the target geometry is in our template dataset), then we simply reconstruct with $A_{n,m}^{(\text{pbdw})}(P_{W(\Omega)}u)$ with the pre-computed reduced model $V_n(\Omega)$.
- If $\Omega \notin \mathrm{G}_{\text{templates}}$:
  - We need to find an appropriate reduced model for the reconstruction. For this, we apply the best-template mapping $\mathtt{BT}$ and we set
  $$\Omega_t^* = \mathtt{BT}(\Omega) \ \in \mathrm{G}_{\text{templates}}.$$
  - We transport the template reduced model $V_n(\Omega_t^*)$ to Ω by applying $\widehat{\tau}_{\Omega_t^* \to \Omega}$, namely
  $$\widehat{V}_n(\Omega) = \widehat{\tau}_{\Omega_t^* \to \Omega}(V_n(\Omega_t^*))$$
  - In Ω, we reconstruct with PBDW using $W_m(\Omega)$ and $\widehat{V}_n(\Omega)$.

Error Analysis of the reconstruction quality using a transported reduced model: Suppose we are given a target geometry $\Omega_1 \in \mathrm{G}$ and that we want to reconstruct an unknown function $u \in \mathcal{M}(\Omega_1)$ from its observations $\ell_i(u)$, $i = 1, \dots, m$. Suppose further that there exists a reduced space $V_n(\Omega_1)$ that accurately approximates $\mathcal{M}(\Omega_1)$ but computing it would would prevent the reconstruction to be in real time. We can instead transport a pre-computed reduced model from a template geometry. Suppose we fix the template geometry to be $\Omega_0 \in \mathrm{G}_{\text{templates}}$ and we transport the reduced model space $V_n(\Omega_0)$ to the target geometry by applying $\widehat{\tau}_{0\to 1}$. This yields

$$\widehat{V}_n(\Omega_1) \coloneqq \widehat{\tau}_{0\to 1}(V_n(\Omega_0)).$$

In Theorem 3.5.1 we analyse the reconstruction error by involving the Hausdorff distance between $\mathbb{S}(V_n(\Omega_1))$ and $\mathbb{S}(\widehat{V}_n(\Omega_1))$, the unit spheres of $V_n(\Omega_1)$ and $\widehat{V}_n(\Omega_1)$. The distance is defined as

$$\begin{aligned} d_H^2(\mathbb{S}(\widehat{V}_n(\Omega_1)), \mathbb{S}(V_n(\Omega_1))) &\coloneqq \max\left(\max_{\hat{v}\in\widehat{V}_n(\Omega_1)} \frac{\|\hat{v} - P_{V_n(\Omega_1)}\hat{v}\|^2}{\|\hat{v}\|^2}; \max_{v\in V_n(\Omega_1)} \frac{\|v - P_{\widehat{V}_n(\Omega_1)}v\|^2}{\|v\|^2} \right) \\ &= \max\left(1 - \beta^2(\widehat{V}_n, V_n); 1 - \beta^2(V_n, \widehat{V}_n)\right) \\ &= 1 - \min\left(\beta^2(\widehat{V}_n, V_n); \beta^2(V_n, \widehat{V}_n)\right) \end{aligned} \tag{3.5.3}$$

Theorem 3.5.1. *Suppose $V_n(\Omega_1)$ is a reduced model space such that*

$$\max_{u\in\mathcal{M}(\Omega_1)} \|u - P_{V_n(\Omega_1)}u\| \le \varepsilon,$$
$$\beta(V_n(\Omega_1), W(\Omega_1)) \ge \underline{\beta} > 0.$$

Let $\widehat{V}_n(\Omega_1) = \widehat{\tau}_{0\to 1}(V(\Omega_0))$ *be a transported subspace from* Ω_0 *to* Ω_1 *such that*

$$d_H(\mathbb{S}(\widehat{V}_n(\Omega_1)), \mathbb{S}(V_n(\Omega_1))) \leq \delta_H.$$

Then the reconstruction of $\mathcal{M}(\Omega_1)$ *with PBDW using* $V_n(\Omega_1)$ *is well-posed and the error is bounded by*

$$\max_{u\in\mathcal{M}(\Omega_1)} \|u - A_{V_n(\Omega_1)}(P_W u)\| \leq \frac{\varepsilon}{\underline{\beta}}.$$

If we use $\widehat{V}_n(\Omega_1)$*, the reconstruction is well posed if and only if*

$$\delta_H < \underline{\beta}$$

and the reconstruction error is bounded by

$$\max_{u\in\mathcal{M}(\Omega_1)} \|u - A_{\widehat{V}_n(\Omega_1)}(P_W u)\| \leq \frac{\varepsilon + 2\delta_H \max_{u\in\mathcal{M}(\Omega_1)} \|P_{V_n+\widehat{V}_n} u\|}{\underline{\beta}(1-\delta_H/\underline{\beta})^{1/2}((2+\delta_H)/\underline{\beta}-1)^{1/2}}. \tag{3.5.4}$$

From the error bound (3.5.4) from Theorem 3.5.1, it follows that the Hausdorff distance between subspaces plays a crucial role in the final reconstruction quality. This motivates to use this distance in our MDS approach to build the routine `BT` to select the best template. If the transported subspace $\widehat{V}_n(\Omega_1)$ deviates from $V_n(\Omega_1)$ by a quantity of the order $\delta_H \leq \varepsilon/\max_{u\in\mathcal{M}(\Omega_1)} \|u\|$, then

$$\max_{u\in\mathcal{M}(\Omega_1)} \|u - A_{\widehat{V}_n(\Omega_1)}(P_W u)\| \leq C\frac{\varepsilon}{\underline{\beta}},$$

for a relatively moderate constant $C \geq 1$. In this scenario, the reconstruction with the transported subspace is of the same quality as the one with the reduced model $V_n(\Omega_1)$ (which we are avoiding to compute in order to speed-up the state estimation procedure).

Numerical results on a simple example:

Geometry: In our example, the family G of geometries is a set of 3D Venturi tubes with variations on three geometrical parameters concerning the tube coarctation (see Figure 3.19). The parameters are the coarctation length S_l, its radius S_r, and its position along the $y-$axis S_x. The ranges of the geometrical parameters are $S_r \in [1.4, 2.6]$ mm, $S_l \in [0.8L, 1.2L]$ and $S_x \in [5, 11]$ mm. The length of the tube is fixed to $L = 5$ cm, and its diameter to $D = 0.4$ cm.

Physical Model: We work with $K = 64$ template geometries for the database $\mathrm{G}_{\mathrm{templates}}$. They are computed using a uniform grid sample on the three geometrical parameters.

We assume that the fluid is governed by the Stokes equations defined, for a given $\Omega \in \mathrm{G}$, as the

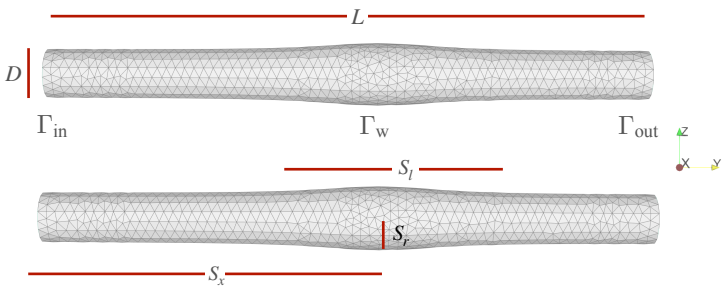


Figure 3.19: Scheme for the generation of the set G.

problem of finding the velocity $u \in \left[H^1(\Omega;[0,T])\right]^3$ and the pressure $p \in L^2(\Omega;[0,T])$ such that

$$\begin{cases} \partial_t u - \mu\Delta u + \nabla p = 0 \text{ in } \Omega \\ \nabla \cdot u = 0 \text{ in } \Omega \\ u = (0,0,0) \text{ on } \Gamma_{\mathrm{w}} \\ u = u_0 \left(0, 1 - \dfrac{x^2+z^2}{(D/2)^2}, 0\right) \sin(2\pi t) \text{ on } \Gamma_{\mathrm{in}} \\ \left(\dfrac{\nabla^T u + \nabla u}{2} - p\mathbf{I}\right) \cdot n = (0,0,0) \text{ on } \Gamma_{\mathrm{out}}, \end{cases}$$

where $\mathbf{I}$ is an identity matrix of size three, n is a unitary vector pointing outwards the working domain, and $u_0 \in \mathbb{R}_+$. The boundary $\partial\Omega$ is decomposed into 3 disjoint subdomains,

$$\partial\Omega = \Gamma_{\mathrm{in}} \cup \Gamma_{\mathrm{out}} \cup \Gamma_{\mathrm{w}},$$

where Γ_{in} is the inflow part, Γ_{out} the outflow, and Γ_{w} corresponds to the walls (see Figure 3.19).

In our example, we reconstruct velocities taking $V(\Omega) = L^2(\Omega, \mathbb{R}^3)$ as the ambient reconstruction space. This choice was made in order to target the reconstruction of the field and not its derivatives. For each $\Omega \in \mathrm{G}$, we work with the manifold

$$\mathcal{M}(\Omega) \coloneqq \{u(y) \ : \ y \in \mathrm{Y}\},$$

with

$$\mathrm{Y} \coloneqq \{y = (t, u_0, \mu) \in [0, 0.5 \text{ s.}] \times [0.01, 1 \text{ cm/s}] \times [0.01, 0.1 \text{ P}]\}.$$

Training/Learning Phase:

- For each $\Omega \in G_{\text{templates}}$, we compute a finite training subset of $\mathcal{M}(\Omega)$ with $N_s = 12\,800$ snapshots, and we compute its Proper Orthogonal Decomposition (POD).
- We build the best template routine `BT` by applying MDS.

Measurement space $W(\Omega)$: For a given $\Omega \in \mathrm{G}$, we consider a partition of $\Omega = \cup_{i=1}^{m} \Omega_i^{\text{voxel}}$ into m disjoint subdomains (voxels) Ω_i^{voxel}. We mimic getting ultrasound images by defining the

linear functionals $\ell_i \in L^2(\Omega)$ as

$$\ell_i(u) = \int_{\Omega_i^{\text{voxel}}} u \cdot b \, \mathrm{d}x, \quad 1 \leq i \leq m,$$

where b is a unitary vector giving the direction of the ultrasound beam. In our case, the plane is chosen to be $z = 0$, the ultrasound direction is $b = [0, \sqrt{2}/2, \sqrt{2}/2]$ and the size of voxels is 2.5 mm^3. The dimension m of the total number of observations changes slightly between geometries. The geometry with the smallest amount of voxels, i.e., the geometry corresponding to the smaller parameter S_r and maximal S_l, is $m = 59$.

Study of the reconstruction quality: Figure 3.20 shows the relative average reconstruction L^2-errors for four different target geometries. Each curve depicts the error made when we use the different template geometries $\Omega_t \in G_{\text{templates}}$ from our database. The role of the routine `BT` which we have built in the learning stage is to quickly select the template which will be the most appropriate so that we obtain the most accurate reconstruction results. The selection with our proposed construction yields the error curve in blue which is labeled MDS. We tested several possibilities for the definition of the metric for MDS but the one based on (3.5.3) produced systematically the best results, so, for the sake of clarity, we only present the results for this choice, which, in addition, is also involved in our numerical analysis above. We observe in Figure 3.20 that the selection method is near-optimal in the sense that it chooses either a good or the best available template among the 64 template domains. Figure 3.21 gives an illustration of the reconstruction of one snapshot with our pipeline.

3.5.3 Epidemiology

In this section I summarize [A1], a contribution in epidemiological forecasting made in collaboration with Prof. Yvon Maday, and two engineers from Summit, Athmane Bakhta and Thomas Boiveau. The work was made in the course of the year 2020, during the first waves of the Covid-19 pandemic in France which led to several lockdown periods. Our goal was to develop a method for forecasting the series of infected and recovered people on a two-week horizon at the regional and interregional resolution. The main challenge in this task is related to the paradox that, on the one hand, the mechanisms of an epidemic spread can be very well understood through epidemiological models, and all the details of its evolution can be very accurately modelled in theory. On the other hand, it is very difficult to benefit from the power of epidemiological models in actual practice because they involve numerous parameters, and many of them are very difficult to accurately estimate in practice. This difficulty legitimately raises the question as to whether one should not resort to purely data-driven strategies but, in this approach, we may quickly be limited by the data: their nature, their quality, and our ability to access it or not. In addition, outputs from purely data-driven strategies may lack interpretability. In our approach, we decided not to dismiss the high potential and interpretability of epidemiological models. We work with a limited amount of health data, and address the difficulty of handling the potential numerous parameters with the angle of attack which I explain next.

We assume that we are given health data in a time window $[0, T]$, where $T > 0$ is taken to be the present time. The observed data is the series of infected[1] people, denoted I_{obs}, and removed people

[1] In fact, the observed series is the of the hospitalized people H_{obs}. In [A1], we apply a correction factor $\alpha = 15$

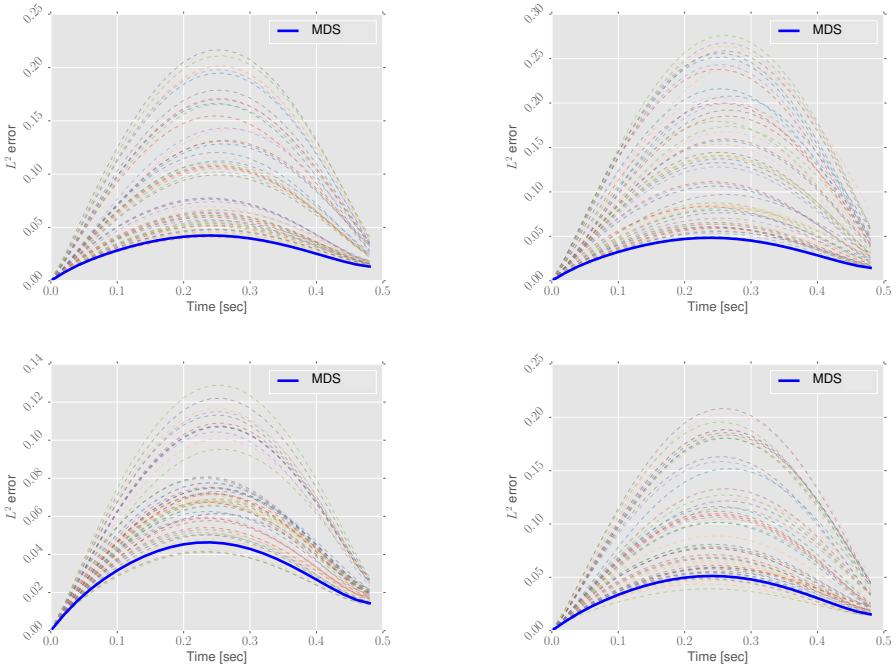
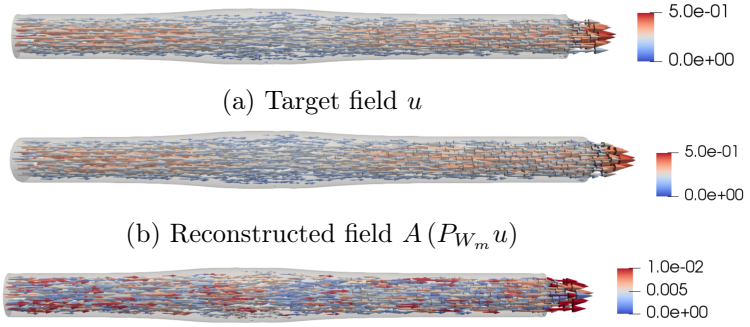


Figure 3.20: Relative reconstruction error for 4 different target geometries $\Omega \in G_{\text{test}}$. Each individual curve depicts the error when we transport the reduced model from a given template geometry. The `Best-Template` methods is able to identify a good or the best template.



(a) Target field u

(b) Reconstructed field $A\left(P_{W_m} u\right)$

(c) Error field $u - A\left(P_{W_m} u\right)$

Figure 3.21: Example of field reconstruction for one target snapshot

denoted R_{obs}. They are usually given at a national or a regional scale and on a daily basis. For our discussion, it will be useful to work with time-continuous functions and $t \to I_{\text{obs}}(t)$ will denote the piecewise constant approximation in $[0,T]$ from the given data (and similarly for $R_{\text{obs}}(t)$). Our goal is to give short-term forecasts of the series in a time window $\tau > 0$ whose size will be about two weeks. We denote by $I(t)$ and $R(t)$ the approximations to the series $I_{\text{obs}}(t)$ and $R_{\text{obs}}(t)$ at any time $t \in [0,T]$.

We propose to fit the data for $t \in [0,T]$ and provide forecasts for $t > T$ with SIR models with time-dependent parameters [98]. It is based on a partition of the population into:

- Uninfected people, called susceptible (S),
- Infected and contagious people (I), with more or less marked symptoms,
- People removed (R) from the infectious process, either because they are cured or unfortunately died after being infected.

If N denotes the total population size that we assume to be constant in time, we have

$$N = S(t) + I(t) + R(t), \forall t \in [0,T],$$

and the evolution from S to I and from I to R is given for all $t \in [0,T]$ by

$$\begin{aligned}
\frac{dS}{dt}(t) &= -\frac{\beta(t) I(t) S(t)}{N} \\
\frac{dI}{dt}(t) &= \frac{\beta(t) I(t) S(t)}{N} - \gamma(t) I(t) \\
\frac{dR}{dt}(t) &= \gamma(t) I(t).
\end{aligned}$$

In the following, we use bold-faced letters for the graph of time-dependent functions. For example, $\mathbf{f} \coloneqq \{f(t) : 0 \le t \le T\}$ for any function $\mathbf{f} \in L^\infty([0,T])$. Using this notation, for any given $\boldsymbol{\beta}$ and $\boldsymbol{\gamma} \in L^\infty([0,T])$ we denote by

$$(\mathbf{S}, \mathbf{I}, \mathbf{R}) = \texttt{SIR}(\boldsymbol{\beta}, \boldsymbol{\gamma}, [0,T])$$

the solution of the associated SIR dynamics in $[0,T]$.

The SIR model is one of the simplest epidemiological models. It has only two parameters:

- $\gamma > 0$ represents the recovery rate. In other words, its inverse γ^{-1} can be interpreted as the length (in days) of the contagious period.
- $\beta > 0$ is the transmission rate of the disease. It essentially depends on two factors: the contagiousness of the disease and the contact rate within the population. The larger this second parameter is, the faster the transition from susceptible to infectious will be.

In the most simple SIR model, β and γ are constant in time. The main motivation for considering them time-dependent is because the family of SIR models with time-dependent coefficients in $L^\infty([0,T])$ has optimal fitting and forecasting properties for our purposes in the sense that we explain next. In addition, the variations of β and γ are reasonable from the epidemiological point

to infer the series of infected people $I_{\text{obs}} = \alpha H_{\text{obs}}$. Obviously, this factor is uncertain and could be improved in the light of further retrospective studies of the outbreak.

of view given that social distancing measures affect the value of β, and the improvement of medical treatments at the hospital have an impact on γ.

Defining the cost function

$$\mathcal{J}(\boldsymbol{\beta}, \boldsymbol{\gamma} \,|\, I_{\text{obs}}(t), R_{\text{obs}}(t), [0,T]) \coloneqq \int_0^T \left(|I(t) - I_{\text{obs}}(t)|^2 + |R(t) - R_{\text{obs}}(t)|^2 \right) \mathrm{d}t$$

such that

$$(\mathbf{S}, \mathbf{I}, \mathbf{R}) = \texttt{SIR}(\boldsymbol{\beta}, \boldsymbol{\gamma}, [0,T]),$$

we define the fitting problem of approximating the observed health series $\mathbf{I}_{\text{obs}}$, and $\mathbf{R}_{\text{obs}}$ with a SIR evolution as the optimal control problem of finding

$$J^* = \inf_{(\boldsymbol{\beta},\boldsymbol{\gamma}) \in L^\infty([0,T]) \times L^\infty([0,T])} \mathcal{J}(\boldsymbol{\beta}, \boldsymbol{\gamma} \,|\, \mathbf{I}_{\text{obs}}, \mathbf{R}_{\text{obs}}, [0,T]). \tag{3.5.5}$$

It is straightforward to observe that if $S_{\text{obs}}, I_{\text{obs}}, R_{\text{obs}}$ are of class $\mathcal{C}^1([0,T])$, then setting

$$\begin{cases} \beta^*_{\text{obs}}(t) & \coloneqq -\dfrac{N}{I_{\text{obs}}(t) S_{\text{obs}}(t)} \dfrac{dS_{\text{obs}}}{dt}(t) \\ \gamma^*_{\text{obs}}(t) & \coloneqq \dfrac{1}{I_{\text{obs}}(t)} \dfrac{dR_{\text{obs}}}{dt}(t), \end{cases}$$

we have that

$$(\mathbf{S}_{\text{obs}}, \mathbf{I}_{\text{obs}}, \mathbf{R}_{\text{obs}}) = \texttt{SIR}(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, [0,T])$$

is a minimizer of the optimal control problem since

$$\mathcal{J}(\boldsymbol{\beta}^*_{\text{obs}}, \boldsymbol{\gamma}^*_{\text{obs}}, [0,T]) = 0,$$

which obviously implies that $J^* = 0$.

This simple observation means that there exists a time-dependent SIR model which can perfectly fit the data of any epidemiological evolution. In particular, we can perfectly fit the series of sick people with a time-dependent SIR model (modulo a smoothing of the local oscillations due to noise). This great approximation power comes however at the cost of defining the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in $L^\infty([0,T])$, which is a space that is too large in order to be able to define any feasible prediction strategy.

In order to pin down a smaller manifold where these parameters may vary without sacrificing much on the fitting and forecasting power, our strategy is the following:

1. **Learning phase:** The fundamental hypothesis of our approach is the confidence that the specialists of epidemiology have well understood the mechanisms of infection transmission. We thus propose to generate a large number of virtual epidemics with highly detailed compartmental models involving constant parameters $\mu \subset \mathbb{R}^p$. Since the models are assumed to be detailed, they involve a potentially large number of parameters ($p \gg 1$) for which the optimal value is unknown. The main steps of our the learning step are thus as follows:

   (a) **Generate virtual scenarios using detailed models with constant coefficients:**

   - Define the notion of `Detailed_Model` which is most appropriate for the epidemiological study. Note that several models could be considered simultaneously. In the case of the Covid-19 pandemic, the detailed models that we considered are the `SEI5CHRD` model from [5] involving 11 compartments or the `SE2IUR` proposed in [7].

- Define an interval range $\mathcal{P} \subset \mathbb{R}^p$ where the parameters μ of `Detailed_Model` will vary. We call the solution manifold $\mathcal{U}$ the set of virtual dynamics over $[0, T+\tau]$, namely

$$\mathcal{U} \coloneqq \{\mathbf{u}(\mu) = \texttt{Detailed\_Model}(\mu, [0, T+\tau]) \ : \ \mu \in \mathcal{P}\}.$$

  The fact that the functions $\mathbf{u}(\mu)$ are defined in the fitting and forecasting window $[0, T+\tau]$ is crucial to provide our forecasts.
- Draw a finite training set

$$\mathcal{P}_{\text{tr}} = \{\mu_1, \dots, \mu_K\} \subseteq \mathcal{P}$$

  of $K \gg 1$ parameter instances and we compute $\mathbf{u}(\mu) = \texttt{Detailed\_Model}(\mu, [0, T+\tau])$ for $\mu \in \mathcal{P}_{\text{tr}}$. Each $\mathbf{u}(\mu)$ is a virtual epidemiological scenario. An important detail for our prediction purposes is that the simulations are done in $[0, T+\tau]$, that is, we simulate not only in the fitting time interval but also in the prediction time interval. We call

$$\mathcal{U}_{\text{tr}} = \{\mathbf{u}(\mu) \ : \ \mu \in \mathcal{P}_{\text{tr}}\}$$

  the training set of all virtual scenarios.

(b) **Collapse:** Collapse every detailed model $\mathbf{u}(\mu) \in \mathcal{U}_{\text{tr}}$ into a SIR model following the procedure outlined in our paper [A1]. For every $\mathbf{u}(\mu)$, the procedure gives time-dependent parameters $\boldsymbol{\beta}(\mu)$ and $\boldsymbol{\gamma}(\mu)$ and associated SIR solutions $(\mathbf{S}, \mathbf{I}, \mathbf{R})(\mu)$ in $[0, T+\tau]$. This yields the sets

$$\mathcal{B}_{\text{tr}} \coloneqq \{\boldsymbol{\beta}(\mu) \ : \ \mu \in \mathcal{P}_{\text{tr}}\} \quad \text{and} \quad \mathcal{G}_{\text{tr}} \coloneqq \{\boldsymbol{\gamma}(\mu) \ : \ \mu \in \mathcal{P}_{\text{tr}}\}.$$

(c) **Compute reduced models:**
We apply model reduction techniques using $\mathcal{B}_{\text{tr}}$ and $\mathcal{G}_{\text{tr}}$ as training sets in order to build two basis

$$\mathrm{B}_n = \operatorname{span}\{b_1, \dots, b_n\}, \quad \mathrm{G}_n = \operatorname{span}\{g_1, \dots, g_n\} \subset L^\infty([0, T+\tau], \mathbb{R}),$$

which are defined over $[0, T+\tau]$. The space B_n is such that it approximates well all functions $\boldsymbol{\beta}(\mu) \in \mathcal{B}_{\text{tr}}$ (resp. all $\boldsymbol{\gamma}(\mu) \in \mathcal{G}_{\text{tr}}$ can be well approximated by elements of G_n). For this step, it is interesting to note that classical model reduction strategies such as Singular Value Decomposition or a classical greedy algorithm did not work because they do not preserve positivity of the basis functions. Due to this, reduction with Nonnegative Matrix Factorization (NMF, see [88, 36]), a variant of SVD involving nonnegative modes and expansion coefficients, delivered better results. We additionally developed a new greedy algorithm, called Enlarged Nonnegative Greedy (ENG), which not only allows to preserve positivity but also other types of bounds. In our numerical results, ENG systematically outperformed NMF.

2. **Fitting on the reduced spaces:** We next solve the fitting problem (3.5.5) in the interval $[0, T]$ by searching $\boldsymbol{\beta}$ (resp. $\boldsymbol{\gamma}$) in B_n (resp. in G_n) instead of in $L^\infty([0, T])$, that is,

$$J^*_{(\mathrm{B}_n, \mathrm{G}_n)} = \min_{(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathrm{B}_n \times \mathrm{G}_n} \mathcal{J}(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{I}_{\text{obs}}, \mathbf{R}_{\text{obs}}, [0, T]). \tag{3.5.6}$$

Note that the respective dimensions of B_n and G_n can be different, for simplicity we take them equal in the following. Obviously, since B_n and $\mathrm{G}_n \subset L^\infty([0,T])$, we have that

$$J^* \leq J^*_{(\mathrm{B}_n,\mathrm{G}_n)},$$

but we numerically observe that the function $n \mapsto J^*_{(\mathrm{B}_n,\mathrm{G}_n)}$ decreases very rapidly as n increases, which indirectly confirms the fact that the manifold generated by the two above models accommodates well the COVID-19 epidemics.

The solution of problem (3.5.6) gives us coefficients $(c_i^*)_{i=1}^n$ and $(\tilde{c}_i^*)_{i=1}^n \in \mathbb{R}^n$ such that the time-dependent parameters

$$\beta_n^*(t) = \sum_{i=1}^n c_i^* b_i(t), \quad \forall t \in [0, T+\tau],$$
$$\gamma_n^*(t) = \sum_{i=1}^n \tilde{c}_i^* g_i(t).$$

achieve the minimum (3.5.6).

3. **Forecast:** For a given dimension n of the reduced spaces, propagate in $[0, T+\tau]$ the associated SIR model

$$(\mathbf{S}_n^*, \mathbf{I}_n^*, \mathbf{R}_n^*) = \mathtt{SIR}(\boldsymbol{\beta}_n^*, \boldsymbol{\gamma}_n^*, [0, T+\tau])$$

The values $I_n^*(t)$ and $R_n^*(t)$ for $t \in [0, T[$ are by construction close to the observed data $\mathbf{I}_{\text{obs}}, \mathbf{R}_{\text{obs}}$ (up to some numerical optimization error). The values $I_n^*(t)$ and $R_n^*(t)$ for $t \in [T, T+\tau]$ are then used for prediction.

4. **Forecast Combination/Aggregation of Experts (optional step):** By varying the dimension n and using different model reduction approaches, we can easily produce a collection of different forecasts and the question of how to select the best predictive model arises. Alternatively, we can also resort to Forecast Combination techniques: denoting $(I_1, R_1), \ldots, (I_P, R_P)$ the different forecasts, the idea is to search for an appropriate linear combination

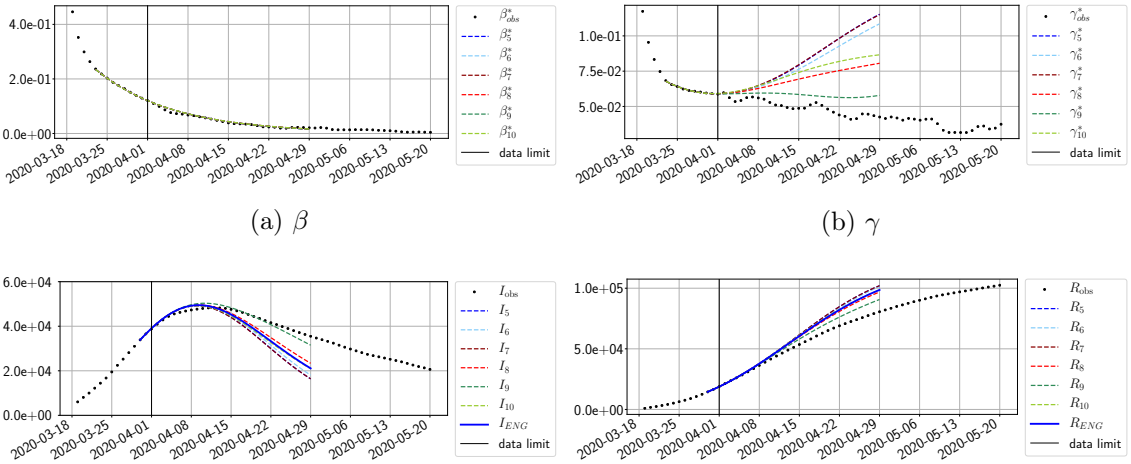
$$I^{\mathrm{FC}}(t) = \sum_{p=1}^P w_p I_p(t)$$

and similarly for R. Note that these combinations do not need to involve forecasts from our methodology only. Other approaches like time series forecasts could also be included. One simple forecast combination is the average, in which all alternative forecasts are given the same weight $w_p = 1/P$, $p = 1, \ldots P$. More elaborate approaches consist in estimating the weights that minimize a loss function involving the forecast error.

Some numerical results: We consider the forecasting of I and R for the first two epidemic waves in the Paris region which took place around March-May 2020 and November 2020. We use public observed data from Santé Publique France[2] to get the number $I_{\text{obs}}(t)$ of infected, and $R_{\text{obs}}(t)$ of removed people. Figures 3.22 to 3.24 show forecasts with our approach on a 28-day ahead window

[2] https://www.data.gouv.fr/en/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/

for different times T. The plots depict the evolution of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ and the resulting evolution of the infected I and removed R. We only show results using the ENG algorithm to find the reduced model spaces B_n and G_n since this was the approach that delivered the best predictions. Note that the method has difficulties in forecasting $\boldsymbol{\gamma}$ due to the oscillatory behavior of the series. However, the obtained forecasts for I and R are in general very satisfactory over the whole time window. We accurately predict the peak of both waves at least 10 days in advance.



(a) β (b) γ

(c) Infected (d) Removed

Figure 3.22: ENG forecast from $T = 01/04$.

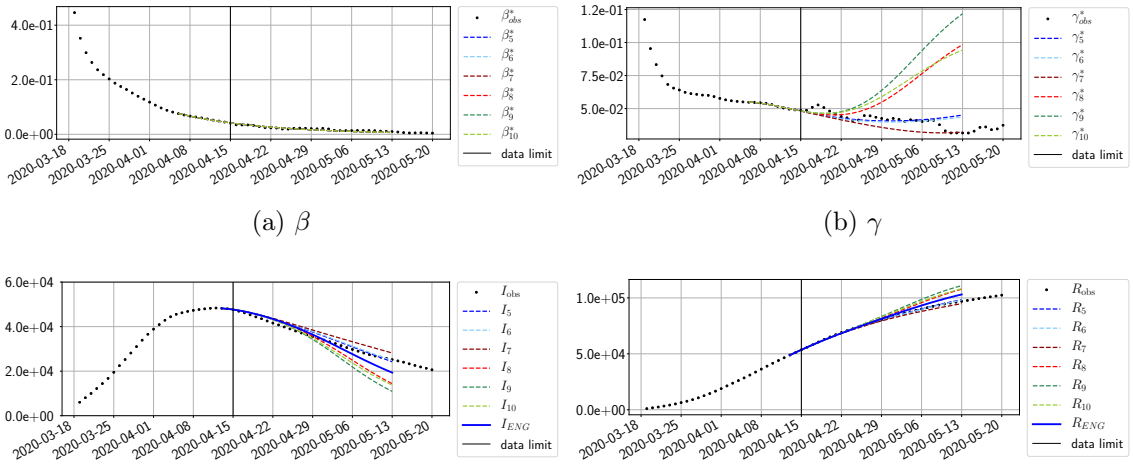
3.6 Research perspectives on inverse problems

In this section I outline some research perspectives on inverse problems that I am either currently investigating, or that I think would be worth considering for future works:

Future works on Theory and Algorithms:

- **Efficient algorithms for advection dominated problems:** The theory and algorithms of Sections 3.1 to 3.4 is general in the sense that it holds for solution sets $\mathcal{M}$ of any nature (elliptic, or hyperbolic problems). However, since the developed algorithms are based on a piecewise affine reconstruction strategy, we expect that advection dominated problems will require a very large number of partitions, making our model selection approach too computationally demanding to be acceptable. The fundamental obstruction is connected to the fact that partitioning a solution set $\mathcal{M}$ from a hyperbolic problem will not improve the slow convergence rate of the Kolmogorov n-width. A relevant topic for future works is therefore the one of searching for alternative nonlinear reconstruction algorithms that are more efficient than our piecewise affine approach in this setting. Can nonlinear schemes from machine learning help us in this task? With what theoretical guarantees and at which training cost?

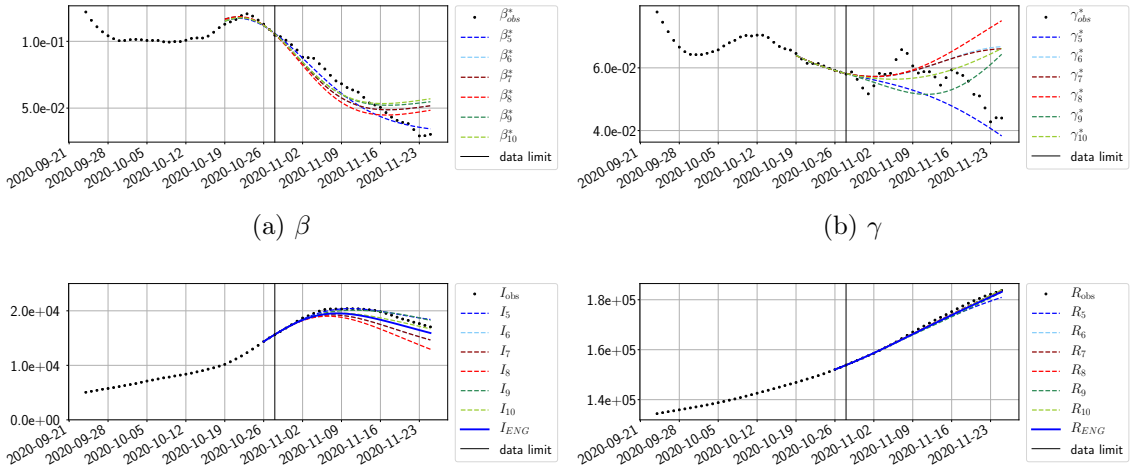
- **Extension of the theory and algorithms for Banach and metric spaces:** Many advection dominated problems are naturally posed on Banach or even metric spaces. Can we extend



(a) β (b) γ

(c) Infected (d) Removed

Figure 3.23: ENG forecast from $T = 15/04$



(a) β (b) γ

(c) Infected (d) Removed

Figure 3.24: ENG forecast from $T = 28/10$

the theory to these spaces and can we develop relatively simple reconstruction algorithms? In the case of Wasserstein spaces, can we involve the model reduction ideas from Chapter 2, Section 2.3?

- **Sampling:** The construction of the local reduced models is usually made with algorithms that involve a finite training set of snapshots obtained through a discretization of the parameter domain. In forward problems, the influence of sampling has been investigated in, e.g., [3]. Its impact on inverse problems has not been explored to the best of my knowledge and it would be worth studying. Even more: if the sampling is not enough, how can we dynamically enrich it?

- **Sensor placement for the piecewise affine reconstruction:** It would be interesting to develop extensions of the work [A10] on sensor placement that I have summarized in Section 3.3. Another very interesting contribution which has connections on sensor placement is [40] and its subsequent extensions/improvements such as [22, 2]. For any of these approaches, the selection of the sensors is done assuming that we work with one reduced model. One line of research could consist in seeing how to perform a joint selection of the optimal observation space and a family of piecewise reduced models.

- **Model error:** The current methodology does not inform on how to proceed if the model error is large. Can we build a systematic strategy to estimate and compensate for the model bias?

Applications:

- **Beyond synthetic observations:** I personally regret that most applications of this manuscript have been done with synthetic measurements. This is of course due to the difficulty of getting access to real observations for some applications such as neutronics. Despite this, I am optimistic that we will be able to work with real data in the near future at last for certain biomedical problems. This will allow us to extend the works [A2, A3, S1] presented in Section 3.6 and to bring the approach closer to a real production use.

- **Variable geometries:** Several extensions of our work [S1] could be envisaged. It would be interesting to extend the strategy to the piecewise reconstruction approach, and for parameter estimation. Another question is related to the database of available template geometries: how can we detect that we do not have enough template geometries and how to enrich the database? For the generation of new meaningful geometries, we may consider using machine learning techniques such as variational autoencoders. For the detection, we may consider using tools from topological data analysis in order to reduce the comparisons of shapes and physical regimes to comparisons of algebraic invariants.

- **Epidemiology:** One key issue in epidemiological forecasting is how to leverage mobility data in the modeling and forecasting of the outbreak dynamics. It would therefore be interesting to develop a multiregional version of the forecasting approach proposed in [A1] and summarized in Section 3.6. This requires the use of interregional population mobility data, which is in general difficult to have access to, but this data is available in our case thanks to an agreement between Paris Sciences Lettres University (PSL) and Facebook. Together with colleagues from computer science at PSL, we have already studied this data in order to understand the

connections between the different epidemic waves and population mobility. Two reports for the greater public have been published on this matter (see [Pop1, Pop2]). It would now be worth using the data for the multiregional extension of [A1]. Beyond this development, the topic of optimal vaccination policies also arises a relevant question to explore, with interesting mathematical challenges. Taking inspiration from the current pandemic situation in which a vaccine is available, but not everybody has the possibility of being vaccinated immediately, the question that we would like to address is the following: given an insufficient and fixed number of vaccines, what is the optimal vaccination strategy? Should we invest all vaccines in vulnerable people, or should we vaccinate some people with risky social behavior to mitigate the spread? We expect that the rigorous formulation of this problem will lead to the study of a mean field equilibrium for which we would have to study its theoretical properties and develop numerical algorithms for solving it.

- **Inverse problems posed on graphs:** Numerous applications are posed on graphs. Among the many examples which we could think of stand the connections between trees from a forest, the traffic on the graph of roads, interactions on social networks... Applying and further extending our methodology to these types of problems seems interesting and original to me. In the framework of Agustín Somacal's thesis, we are currently studying state and parameter estimation problems related to the transport of pollutants in urban areas.

3.7 References

Bibliography

[1] Bonito A., Cohen A., R. DeVore, D. Guignard, P. Jantsch, and G. Petrova. "Nonlinear methods for model reduction". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 55.2 (2021), pp. 507–531.

[2] A. Cohen and M. Dolbeault. "Optimal pointwise sampling for L^2 approximation". In: *arXiv preprint arXiv:2105.05545* (2021).

[3] A. Cohen, W. Dahmen, R. DeVore, and J. Nichols. "Reduced basis greedy selection using random training sets". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 54.5 (2020), pp. 1509–1524.

[4] Julie Delon and Agnès Desolneux. "A Wasserstein-type distance in the space of gaussian mixture models". In: *SIAM Journal on Imaging Sciences* 13.2 (2020), pp. 936–970.

[5] Laura Di Domenico, Giulia Pullano, Chiara E Sabbatini, Pierre-Yves Boëlle, and Vittoria Colizza. "Expected impact of lockdown in Île-de-France and possible exit strategies". In: *medRxiv* (2020).

[6] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley. "Multidimensional scaling, Sammon mapping, and Isomap: Tutorial and survey". In: *arXiv preprint arXiv:2009.08136* (2020).

[7] P. Magal and G. Webb. "Predicting the number of reported and unreported cases for the COVID-19 epidemic in South Korea, Italy, France and Germany". In: *Medrxiv* (2020). URL: https://doi.org/10.1101/2020.03.21.20040154.

[8] J. A. Carrillo, K. Grunert, and H. Holden. "A Lipschitz metric for the Hunter–Saxton equation". In: *Communications in Partial Differential Equations* 44.4 (2019), pp. 309–334. DOI: 10.1080/03605302.2018.1547744. URL: https://doi.org/10.1080/03605302.2018.1547744.

[9] W. Dahmen and R. P. Stevenson. "Adaptive strategies for transport equations". In: *Comp. Meth. Appl. Math.* 19.3 (2019), pp. 431–464.

[10] H. Gong, Y. Maday, O. Mula, and T. Taddei. "PBDW method for state estimation: error analysis for noisy data and nonlinear formulation". In: *arXiv e-prints*, arXiv:1906.00810 (June 2019), arXiv:1906.00810.

[11] C. Greif and K. Urban. "Decay of the Kolmogorov N-width for wave problems". In: *Applied Mathematics Letters* 96 (2019), pp. 216–222.

[12] 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". In: *Journal of Computational Physics* 378 (2019), pp. 686–707.

[13] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. "Scaling algorithms for unbalanced optimal transport problems". In: *Mathematics of Computation* 87.314 (2018), pp. 2563–2609.

[14] A. Cohen, W. Dahmen, and R. DeVore. "Reduced basis greedy selection using random training sets". In: *arXiv preprint arXiv:1810.09344* (2018).

[15] M. Cuturi and G. Peyré. *Computational Optimal Transport*. 2018.

[16] Felix Gruber. "Adaptive Source Term Iteration: A Stable Formulation for Radiative Transfer". PhD thesis. RWTH Aachen University, 2018. DOI: 10.18154/RWTH-2018-230893.

[17] S. Aouadi, D. Q. Bui, R. Guetat, and Y. Maday. *Convergence analysis of the coupled parareal-Schwarz waveform relaxation method*. in preparation. 2017.

[18] P. Benner, A. Cohen, M. Ohlberger, and K. Willcox. *Model Reduction and Approximation: Theory and Algorithms*. Vol. 15. SIAM, 2017.

[19] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk. "Data assimilation in reduced modeling". In: *SIAM/ASA Journal on Uncertainty Quantification* 5.1 (2017), pp. 1–29. DOI: 10.1137/15M1025384. URL: http://dx.doi.org/10.1137/15M1025384.

[20] D. Broersen, W. Dahmen, and R. Stevenson. "On the stability of DPG formulations of transport equations". In: *Math. Comp.* (2017). DOI: https://doi.org/10.1090/mcom/3242. URL: http://www.ams.org/journals/mcom/0000-000-00/S0025-5718-2017-03242-7/#Additional.

[21] A. Calloo, D. Couyras, F. Fécotte, and M. Guillo. "Cocagne: EDF new neutronic core code for ANDROMEDE calculation chain". In: *Proceedings of International Conference on Mathematics & Computational Methods Applied to Nuclear Science & Engineering (M&C), Jeju, Korea*. 2017.

[22] A. Cohen and G. Migliorati. "Optimal weighted least-squares methods". In: *The SMAI journal of computational mathematics* 3 (2017), pp. 181–203.

[23] M. Dashti and A. M. Stuart. "The Bayesian Approach to Inverse Problems". In: *Handbook of Uncertainty Quantification*. Ed. by R. Ghanem, D. Higdon, and H. Owhadi. Cham: Springer International Publishing, 2017, pp. 311–428. ISBN: 978-3-319-12385-1. DOI: 10.1007/978-3-319-12385-1_7. URL: https://doi.org/10.1007/978-3-319-12385-1_7.

[24] R. D. Falgout, T. A. Manteuffel, B. O'Neill, and J. B. Schroder. "Multigrid reduction in time for nonlinear parabolic problems: A case study". In: *SIAM Journal on Scientific Computing* 39.5 (2017), S298–S322.

[25] T.S. Haut, C. Ahrens, A. Jonko, R. Lowrie, and A. Till. "A new multigroup method for cross-sections that vary rapidly in energy". In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 187 (2017), pp. 461–471. ISSN: 0022-4073. DOI: 0.1016/j.jqsrt.2016.10.019.

[26] Gerrit Welper. "Interpolation of functions with parameter dependent jumps by transformed snapshots". In: *SIAM Journal on Scientific Computing* 39.4 (2017), A1225–A1250.

[27] M. Blatt, A. Burchardt, A. Dedner, C. Engwer, J. Fahlke, B. Flemisch, C. Gersbacher, C Gräser, F. Gruber, C. Grüninger, D. Kempf, R. Klöfkorn, T. Malkmus, S. Müthing, M. Nolte, M. Piatkowski, and O. Sander. "The Distributed and Unified Numerics Environment, Version 2.4". In: *Archive of Numerical Software* 4.100 (2016), pp. 13–29. DOI: 10.11588/ans.2016.100.26526.

[28] A. Cohen and R. DeVore. "Kolmogorov widths under holomorphic mappings". In: *IMA Journal of Numerical Analysis* 36.1 (2016), pp. 1–12. DOI: 10.1093/imanum/dru066. URL: http://dx.doi.org/10.1093/imanum/dru066.

[29] M. Ohlberger and S. Rave. "Reduced Basis Methods: Success, Limitations and Future Challenges". In: *Proceedings of the Conference Algoritmy*. 2016, pp. 1–12. URL: http://www.iam.fmph.uniba.sk/amuc/ojs/index.php/algoritmy/article/view/389.

[30] A. Quarteroni, A. Manzoni, and F. Negri. "Reduced Basis Methods for Partial Differential Equations. An Introduction". In: La Matematica per il 3+2. 92 (2016). DOI: 10.1007/978-3-319-15431-2. URL: http://infoscience.epfl.ch/record/218966.

[31] A. Cohen and R. DeVore. "Approximation of high-dimensional parametric PDEs". In: *Acta Numerica* 24 (2015), pp. 1–159. DOI: 10.1017/S0962492915000033.

[32] J. S. Hesthaven, G. Rozza, and B. Stamm. "Certified Reduced Basis Methods for Parametrized Partial Differential Equations". In: *SpringerBriefs in Mathematics* (2015).

[33] Y. Maday, A. T. Patera, J. D. Penn, and M. Yano. "A parameterized-background data-weak approach to variational data assimilation: formulation, analysis, and application to acoustics". In: *International Journal for Numerical Methods in Engineering* 102.5 (2015), pp. 933–965. DOI: 10.1002/nme.4747. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/nme.4747.

[34] M. L. Minion, R. Speck, M. Bolten, M. Emmett, and D. Ruprecht. "Interweaving PFASST and parallel multigrid". In: *SIAM Journal on Scientific Computing* 37.5 (2015), S244–S263.

[35] J. Zinsl and D. Matthes. "Exponential convergence to equilibrium in a coupled gradient flow system modeling chemotaxis". In: *Analysis & PDE* 8.2 (2015), pp. 425–466.

[36] N. Gillis. "The why and how of nonnegative matrix factorization". In: *Regularization, optimization, kernels, and support vector machines* 12.257 (2014), pp. 257–291.

[37] S. Sommer, F. Lauze, and M. Nielsen. "Optimization over geodesics for exact principal geodesic analysis". In: *Advances in Computational Mathematics* 40.2 (Apr. 2014), pp. 283–313. ISSN: 1572-9044. DOI: 10.1007/s10444-013-9308-1. URL: https://doi.org/10.1007/s10444-013-9308-1.

[38] B. Adcock, A. C. Hansen, and C. Poon. "Beyond consistent reconstructions: optimality and sharp bounds for generalized sampling, and application to the uniform resampling problem". In: *SIAM Journal on Mathematical Analysis* 45.5 (2013), pp. 3132–3167.

[39] A. Blanchet and P. Laurençot. "The parabolic-parabolic Keller-Segel system with critical diffusion as a gradient flow in $\mathbb{R}^d$, $d \geq 3$". In: *Communications in Partial Differential Equations* 38.4 (2013), pp. 658–686.

[40] A. Cohen, M. A. Davenport, and D. Leviatan. "On the stability and accuracy of least squares approximations". In: *Foundations of computational mathematics* 13.5 (2013), pp. 819–834.

[41] R. DeVore, G. Petrova, and P. Wojtaszczyk. "Greedy algorithms for reduced bases in Banach spaces". In: *Constructive Approximation* 37.3 (2013), pp. 455–466.

[42] A. Ern and J.L. Guermond. *Theory and practice of finite elements*. Vol. 159. Springer Science & Business Media, 2013.

[43] Y. Maday and B. Stamm. "Locally adaptive greedy approximations for anisotropic parameter reduced basis spaces". In: *SIAM J. Scientific Computing* 35.6 (2013), A2417–A2441.

[44] A. Buffa, Y. Maday, A. T. Patera, C. Prud'homme, and G. Turinici. "A priori convergence of the greedy algorithm for the parametrized reduced basis method". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 46.3 (2012), pp. 595–603.

[45] C. Canuto, M. Y. Hussaini, and A. Quarteroni. *Spectral methods in fluid dynamics*. Springer Science & Business Media, 2012.

[46] R. Guetat. "Méthode de parallélisation en temps: Application aux méthodes de décomposition de domaine". PhD thesis. Paris VI, 2012.

[47] Martial Agueh and Guillaume Carlier. "Barycenters in the Wasserstein space". In: *SIAM Journal on Mathematical Analysis* 43.2 (2011), pp. 904–924.

[48] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk. "Convergence Rates for Greedy Algorithms in Reduced Basis Methods". In: *SIAM Journal on Mathematical Analysis* 43.3 (2011), pp. 1457–1472. DOI: 10.1137/100795772. URL: https://doi.org/10.1137/100795772.

[49] A. Chambolle and T. Pock. "A first-order primal-dual algorithm for convex problems with applications to imaging". In: *Journal of Mathematical Imaging and Vision* 40.1 (2011), pp. 120–145.

[50] A. Cohen, R. DeVore, and C. Schwab. "Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's". In: *Analysis and Applications* 09.01 (2011), pp. 11–47. DOI: 10.1142/S0219530511001728. URL: https://doi.org/10.1142/S0219530511001728.

[51] B. N Khoromskij and C. Schwab. "Tensor-structured Galerkin approximation of parametric and stochastic elliptic PDEs". In: *SIAM Journal on Scientific Computing* 33.1 (2011), pp. 364–385.

[52] V. Temlyakov. *Greedy Approximation*. Vol. 20. Cambridge University Press, 2011.

[53] E. Bernard, F. Golse, and F. Salvarani. "Homogenization of transport problems and semigroups". In: *Math Method Appl Sci* 33.10 (2010), pp. 1228–1234. DOI: 10.1002/mma.1319. URL: https://doi.org/10.1002/mma.1319.

[54] A. Cohen, R. DeVore, and C. Schwab. "Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs". In: *Foundations of Computational Mathematics* 10.6 (2010), pp. 615–646.

[55] J. L. Eftang, A. T. Patera, and E. M. Rønquist. "An" hp" certified reduced basis method for parametrized elliptic partial differential equations". In: *SIAM Journal on Scientific Computing* 32.6 (2010), pp. 3170–3200.

[56] M. Minion. "A hybrid parareal spectral deferred corrections method". In: *Comm. App. Math. and Comp. Sci.* 5.2 (2010).

[57] A. M. Stuart. "Inverse problems: a Bayesian perspective". In: *Acta Numerica* 19 (2010), pp. 451–559.

[58] U. Gianazza, G. Savaré, and G. Toscani. "The Wasserstein gradient flow of the Fisher information and the quantum drift-diffusion equation". In: *Archive for rational mechanics and analysis* 194.1 (2009), pp. 133–220.

[59] A. R. Barron, A. Cohen, W. Dahmen, and R. A. DeVore. "Approximation and learning by greedy algorithms". In: *The annals of statistics* 36.1 (2008), pp. 64–94.

[60] M. J. Gander and E. Hairer. "Nonlinear convergence analysis for the parareal algorithm". In: *Domain decomposition methods in science and engineering XVII*. Springer, 2008, pp. 45–56.

[61] M. L. Minion, A. Williams, T. E. Simos, G. Psihoyios, and C. Tsitouras. "Parareal and spectral deferred corrections". In: *AIP Conference Proceedings*. Vol. 1048. 1. 2008, p. 388.

[62] E. Novak and H. Wozniakowski. "Tractability of Multivariate Problems, Volume I: Linear Information, European Math". In: *Soc., Zürich* 2.3 (2008).

[63] P. Reuss. *Neutron Physics*. EDP Sciences, 2008.

[64] M.A. Grepl, Y. Maday, N.C. Nguyen, and A.T. Patera. "Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations." In: *ESAIM, Math. Model. Numer. Anal.* 41(3) (2007), pp. 575–605.

[65] Y. Maday, J. Salomon, and G. Turinici. "Monotonic parareal control for quantum systems". In: *SIAM Journal on Numerical Analysis* 45.6 (2007), pp. 2468–2482.

[66] P. Massart. "Concentration inequalities and model selection". In: (2007). DOI: 10.1007/978-3-540-48503-2.

[67] G. Rozza, D. B. P. Huynh, and A. T. Patera. "Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations". In: *Archives of Computational Methods in Engineering* 15.3 (Sept. 2007), p. 1. ISSN: 1886-1784. DOI: 10.1007/BF03024948. URL: https://doi.org/10.1007/BF03024948.

[68] A. et al. Koning. *The JEFF-3.1 Nuclear Data Library-JEFF Report 21*. Tech. rep. Organisation for Economic Co-operation and Development, 2006. URL: https://www.oecd-nea.org/dbdata/nds_jefreports/jefreport-21/jeff21.pdf.

[69] L. Tartar. *An introduction to Navier-Stokes equation and Oceanography*. Vol. 1. Lecture Notes of the Unione Matematica Italiana. Springer-Verlag Berlin Heidelberg, 2006. DOI: 10.1007/3-540-36545-1. URL: https://doi.org/10.1007/3-540-36545-1.

[70] A. Bressan and M. Fonte. "An Optimal Transportation Metric for Solutions of the Camassa-Holm Equation". In: *Methods Appl Anal* 12 (May 2005). DOI: 10.4310/MAA.2005.v12.n2.a7.

[71] D. L. Donoho and C. Grimes. "Image manifolds which are isometric to Euclidean space". In: *Journal of mathematical imaging and vision* 23.1 (2005), pp. 5–24.

[72] Y. Maday and G. Turinici. "The Parareal in Time Iterative Solver: a Further Direction to Parallel Implementation". In: *Domain Decomposition Methods in Science and Engineering*. Springer Berlin Heidelberg, 2005, pp. 441–448.

[73] M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera. "An Empirical Interpolation Method: application to efficient reduced-basis discretization of partial differential equations." In: *C. R. Acad. Sci. Paris, Série I.* 339 (2004), pp. 667–672. DOI: https://doi.org/10.1016/j.crma.2004.08.006.

[74] P. T. Fletcher, C. Lu, S. M. Pizer, and S.C. Joshi. "Principal geodesic analysis for the study of nonlinear statistics of shape". In: *IEEE Transactions on Medical Imaging* 23 (2004), pp. 995–1005.

[75] Cédric Villani. *Topics in optimal transportation*. 58. American Mathematical Soc., 2003.

[76] R. J. LeVeque. *Finite volume methods for hyperbolic problems*. Vol. 31. Cambridge university press, 2002.

[77] L. Giacomelli and F. Otto. "Variatonal formulation for the lubrication approximation of the Hele-Shaw flow". In: *Calculus of Variations and Partial Differential Equations* 13.3 (2001), pp. 377–403.

[78] F. Otto. "The geometry of dissipative evolution equations: the porous medium equation". In: (2001).

[79] L. Dumas and F. Golse. "Homogenization of transport equations". In: *SIAM Journal on Applied Mathematics* 60.4 (2000), pp. 1447–1470.

[80] J. B. Tenenbaum, V. De Silva, and J. C. Langford. "A global geometric framework for nonlinear dimensionality reduction". In: *science* 290.5500 (2000), pp. 2319–2323.

[81] Ernst Hairer and Gerhard Wanner. "Stiff differential equations solved by Radau methods". In: *Journal of Computational and Applied Mathematics* 111.1 (1999), pp. 93–111. ISSN: 0377-0427. DOI: https://doi.org/10.1016/S0377-0427(99)00134-X. URL: http://www.sciencedirect.com/science/article/pii/S037704279900134X.

[82] R. Jordan, D. Kinderlehrer, and F. Otto. "The variational formulation of the Fokker–Planck equation". In: *SIAM journal on mathematical analysis* 29.1 (1998), pp. 1–17.

[83] C. Bernardi and Y. Maday. "Spectral methods". In: *Handbook of numerical analysis* 5 (1997), pp. 209–485.

[84] R. A. DeVore and V. N. Temlyakov. "Some remarks on greedy algorithms". In: *Advances in Computational Mathematics* 5.1 (1996), pp. 173–187.

[85] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*. Vol. 14. Jan. 1996.

[86] H. Murase and S. K. Nayar. "Visual learning and recognition of 3-D objects from appearance". In: *International journal of computer vision* 14.1 (1995), pp. 5–24.

[87] B. Bojanov. "Optimal recovery of functions and integrals". In: *First European Congress of Mathematics*. Springer. 1994, pp. 371–390.

[88] P. Paatero and U. Tapper. "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values". In: *Environmetrics* 5.2 (1994), pp. 111–126. URL: https://doi.org/10.1002/env.3170050203.

[89] L. Tartar. "Nonlocal effects induced by homogenization". In: *Partial differential equations and the calculus of variations*. Springer, 1989, pp. 925–938.

[90] D. E. Cullen and G. C. Pomraning. "The multiband method in radiative transfer calculations". In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 24.2 (1980), pp. 97–117. ISSN: 0022-4073. DOI: 10.1016/0022-4073(80)90024-2. URL: http://www.sciencedirect.com/science/article/pii/0022407380900242.

[91] J. R. Dormand and P. J. Prince. "A family of embedded Runge-Kutta formulae". In: *Journal of computational and applied mathematics* 6.1 (1980), pp. 19–26.

[92] C. A. Micchelli and Th. J. Rivlin. *A survey of optimal recovery*. Springer, 1977.

[93] S.E. Aidarous, M.R. Gevers, and M.J. Installe. "Optimal sensors' allocation strategies for a class of stochastic distributed systems". In: *International Journal of Control* 22.2 (1975), pp. 197–213.

[94] M. Livolant and F. Jeanpierre. *Autoprotection des résonnances dans les réacteurs nucléaires*. Tech. rep. 4533. CEA, 1974. URL: https://inis.iaea.org/collection/NCLCollectionStore/_Public/05/153/5153104.pdf.

[95] T. K. Yu and J. H. Seinfeld. "Observability and optimal measurement location in linear distributed parameter systems". In: *Int. J. Control* 18.4 (1973), pp. 785–799.

[96] A. Bensoussan. "Optimization of sensors' location in a distributed filtering problem". In: *Stability of stochastic dynamical systems*. Springer, 1972, pp. 62–84.

[97] J.R. Cannon and R.E. Klein. "Optimal selection of measurement locations in a conductor for approximate determination of temperature distributions". In: *J. Dyn. Sys. Meas. Control* 93.3 (1971), pp. 193–199. DOI: 10.1115/1.3426496.

[98] W. O. Kermack, A. G. McKendrick, and G. T. Walker. "A contribution to the mathematical theory of epidemics". In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 115.772 (1927), pp. 700–721. DOI: 10.1098/rspa.1927.0118. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1927.0118.