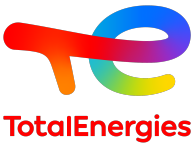
Unstructured Isogeometric Analysis with Applications to Seismic Wave Propagation

Stefano Frambati





Université de Pau et des Pays de l'Adour
Institut National de Recherche en Informatique et Automatique
TotalEnergies SE

Unstructured Isogeometric Analysis with Applications to Seismic Wave Propagation

Thesis by
Stefano Frambati

In partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Mathematics

Scientific advisors

Hélène Barucq, Henri Calandra, Julien Diaz

Reviewers

Rémi Abgrall	Professor, Universität Zürich
Régis Duvigneau	Researcher, Inria Sophia Antipolis

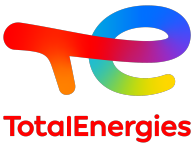
Thesis committee

Rémi Abgrall	Professor, Universität Zürich	Reviewer
Hélène Barucq	Research Director, Inria Bordeaux Sud-Ouest	Advisor
Henri Calandra	Expert, Numerical Methods, TotalEnergies SE	Advisor
Gilles Carbou	Professor, Université de Pau et des Pays de l'Adour	President
Julien Diaz	Research Director, Inria Bordeaux Sud-Ouest	Advisor
Régis Duvigneau	Researcher, Inria Sophia Antipolis	Reviewer
Christian Gout	Professor, INSA Rouen	Examiner
Jeanne Pellerin	R&D Engineer, HDR, TotalEnergies SE	Examiner

13 December 2021







Université de Pau et des Pays de l'Adour
Institut National de Recherche en Informatique et Automatique
TotalEnergies SE

Analyse Isogéométrique Non-Structurée avec Applications à la Propagation des Ondes Sismiques

---

Thèse de doctorat présentée par
Stefano Frambati

Pour l'obtention du grade de
Docteur en Mathématiques

Directeurs de thèse

Hélène Barucq, Henri Calandra, Julien Diaz

Rapporteurs

Rémi Abgrall	Professeur, Universität Zürich
Régis Duvigneau	Chargé de recherche, Inria Sophia Antipolis

Jury

Rémi Abgrall	Professeur, Universität Zürich	Rapporteur
Hélène Barucq	Directrice de recherche, Inria Bordeaux Sud-Ouest	Directrice
Henri Calandra	Ingénieur de recherche expert, TotalEnergies SE	Directeur
Gilles Carbou	Professeur, Université de Pau et des Pays de l'Adour	Président
Julien Diaz	Directeur de recherche, Inria Bordeaux Sud-Ouest	Directeur
Régis Duvigneau	Chargé de recherche, Inria Sophia Antipolis	Rapporteur
Christian Gout	Professeur, INSA Rouen	Examinateur
Jeanne Pellerin	Ingénieur de recherche HDR, TotalEnergies SE	Examinatrice

13 Décembre 2021

A Martine, Chloé, Noah e Iris. Siete la mia vita.

Acknowledgements

First of all, I wish to thank my advisors and mentors, Hélène, Henri and Julien, for providing me with their unwavering help, support, motivation and inspiration throughout this journey. I wish to thank Hélène for her incredible enthusiasm, the unfaltering encouragement I received from her through thick and thin, and the incredible help she gave me, from the most long-term vision down to the gritty details of numerical analysis. She is the engine and heart of the team, and she has created a joyous and superb environment that has helped so many people grow and learn, me included. I wish to thank Julien for the invaluable help and mentoring he provided, always available, quick to answer and very helpful. I have learned so much about numerical methods in the last few years, and it is in large part thanks to your ability to listen and point me towards the right answer. Thank you for sharing all your insights and knowledge, from the most abstract to the most practical, it has helped me so much. Thank you Henri, not only for accepting my weird idea of pursuing a PhD at age 35, but for enthusiastically endorsing it, proposing an incredibly interesting research topic, introducing me to the wonderful family of Magique3D (now Makutu!), helping me convince all the others around us, and supporting me and helping me all the way. Exactly zero part of this work would have been possible without your help. I am really looking forward to keep working with you all.

A huge thank you goes out to my reviewers, Rémi and Régis. You have accepted to review my work, spent an undoubtedly long time reading my verbose and unnecessarily convoluted arguments, and managed to give me invaluable feedback, some very intersting discussions and a huge boost of confidence in my work. I also wish to give a big thank you to the members of my thesis committe, which includes, along with my advisors and reviewers, Gilles, Christian and Jeanne. Thank you for being there, for listening to my presentation and especially for the amazing Q&A session that followed, you challenged me and gave me a lot of interesting feedback, and you allowed me to go into the details of a lot of fascinating points about my work that I did not have the time to include in the main presentation. You have helped me make the conclusion of my adventure as exciting and satisfying as the rest of the journey. I owe you all a great debt of gratitude, and I am really looking forward to being able to work with you again in the future.

I cannot stress enough how this project has been important for me, and it would not have been possible were it not for the help of my colleagues of TotalEnergies, and in particular the Data Science, Artificial Intelligence and Computational Science & Engineering research program. I wish to thank especially Philippe and Alexandre, who have created the ideal conditions for me to work towards my goals, have supported me in practice for such a long time, and have helped me promote our results and our vision within the group. This work would not have been possible without your support. I also would like to thank my colleagues at the CIG department,

and especially Frédérik, for letting me start working on my PhD topic before I officially joined the R&D group, not curbing my enthusiasm but, on the contrary, motivating me even more. A big thank you also goes to the administration of Inria-Bordeaux Sud-Oeust, UPPA and the doctoral school ED211, for helping my somewhat peculiar situation sail smoothly without too many bureaucratic complications.

For a few years, I have been part of this wonderful family of magicians that is Makutu, and I have met so many people who have made this journey an invaluable part of my life. In addition to Hélène and Julien, I had the chance of meeting so many incredible people. Ha–it was fun to go back to quantum physics for a while!, Florian with your kindness and incredible scientific talent, Sébastien and the many smart, piercing remarks you make both at the coffee machine and during Q&A sessions, Victor and your impeccable style, Yder and the many smart ideas you have thrown my way, Marc the king of python, Algiane the queen of meshing, Juliette, Marc and Augustin who helped me appreciate helioseismology and music even more now that I can mentally picture the actual pressure waves inside my guitar, Sylvie who helped an "outsider" like me be a part of the team.

My day-to-day life has been incredibly rich thanks to my fellow past and present Master and PhD students, Mamadou, Aurélien, Elvira, Justine, Chengyi, Pierre, Rose, Nathan, Margot, Vinduja, Arjeta, Julien, Ibrahima, Johan. Even if I couldn't take part in many activities due to my family constraints, you have always made me feel part of this incredible family of magicians, and you have made my days much brighter. A big, big, *big* thank you. I have met incredible talent in this laboratory, and I am sure that you will all make it very far in life! And maybe you will have kids of your own and understand some of the crazy stories that I kept telling you over these years, and that you kindly accepted to listen to ☺. I could write a chapter about each of you, listing all the things I appreciate about your character, your talent, drive, motivation but also and more importantly your kindness, empathy, and the lots of fun that each of you are. I won't write it though, since apparently I already tend to write and talk too much 😉! But I loved being a student again, and the reason I did, is you.

The best part is that, just as an adventure ends, a new one begins, and I will still be able to work with many of you in the joint TE-Inria team Makutu for the net few years. I couldn't have asked for a better outcome, and I am looking forward with happiness and enthusiasm to our work together and the challenges ahead. Go Makutu!

However stimulating the scientific problems and challenges were, there is only one thing I had in mind during my work, and that is my family. The family that brought me into the world and nurtured me, that is, my parents Giorgio and Rosella and my brother Luca (thank you for putting up with me for such a long time ☺). But especially the family that I am so lucky to have found in my wife. Martine, you are not only the love of my life and my motivation, but also my inspiration. You have been helpful, rational, and kept my boat straight even in rocking waters. I am so much indebted to you, and not just for the last few years. I am incredibly lucky to have found you, and even luckier to still have you by my side. And I am so, so grateful every day for the wonderful little family that we have created. Chloé, Noah and Iris, I do (and would do) anything for you, and I hope that some day you might be a little bit proud of your crazy old dad, and maybe even inspired to challenge yourselves in life, and never give up.

Abstract

In this work, we explore the usefulness of spaces of unstructured spline functions in the resolution of hyperbolic problems discretized by explicit-time schemes, and especially the acoustic wave propagation problem with absorbing boundary conditions, and its associated inverse problem known as seismic inversion.

We base our analysis on the known definition of simplex splines through polyhedral projections and Dirichlet averages, and we focus especially on the construction of polynomial-reproducing spaces of simplex splines. We introduce a family of spaces associated to fine zonotopal tilings, which constitute a combinatorial extension of previous results on Delaunay configurations, and which can be constructed on general point configurations including repeated and affinely dependent points. The resulting reduction in the regularity of the function space allows to define external boundary conditions, as well as subdivide the space into subdomains. Furthermore, the combinatorial properties of zonotopal tilings allow us to derive some useful algorithms for spline space construction in all space dimensions, generalizing previously known algorithms, as well as the evaluation of all the spline functions supported at a given point.

We employ these spaces to define an unstructured version of known multi-patch Discontinuous Galerkin (DG) – isogeometric analysis (IGA) numerical schemes, showing that we can recover the usual Bernstein-Bézier DG scheme, as well as a fully unstructured IGA method, as special cases. We also show that the behavior of these spline spaces near the external and internal boundaries is very similar to the behavior of the standard DG bases, thus allowing to derive simple inverse inequalities and reuse the standard results of coercivity and a priori error analysis originally derived for the interior-penalty discontinuous Galerkin (IPGD) method. We illustrate the numerical properties of this discretization scheme with some numerical experiments.

Finally, we explore some possible applications of unstructured spline functions to the seismic inversion problem, in the form of the full waveform inversion (FWI) technique. Specifically, we use the location of the spline knots as degrees of freedom for the inversion, using some known facts about the derivatives of these functions to drive the optimization process. Since the cost function of FWI is not generally differentiable with respect to the geometric degrees of freedom, we introduce this technique using subdifferentials, and we give a derivation of the adjoint state method for the computation of the gradient using a known convex duality theorem, valid for non-differentiable convex functions.

Résumé

Dans ce travail, on explore l'utilisation des espaces de fonctions splines non-structurées dans la résolution des problèmes hyperboliques discrétisés par des schémas en temps explicites, et en particulier le problème de la propagation des ondes acoustiques avec conditions aux limites absorbantes, et son problème inverse associé, l'inversion sismique.

Notre analyse repose sur la définition, connue, de spline simplexe à travers les projections de polyèdres et les moyennes de Dirichlet, et on se concentre en particulier sur la construction d'espaces de splines simplexes capables de reproduire les polynômes. On introduit une famille d'espaces associées aux pavages fins de zonotopes, qui constituent une extension combinatoire de certains résultats connus sur les configurations de Delaunay, et qui peuvent être construits sur configurations générales de points incluant des points répétés et affinement dépendants. La réduction de la régularité de l'espace de fonctions qui en découle permet de définir des conditions aux limites, ainsi que de subdiviser l'espace en sous-domaines. De plus, les propriétés combinatoires des pavages de zonotopes nous permettent de dériver un certain nombre d'algorithmes utiles pour la construction de l'espace de splines dans n'importe quel nombre de dimensions, généralisant un algorithme connu en dimension deux d'espace, ainsi que pour l'évaluation de toutes les fonctions splines de l'espace en un point donné.

On se sert des espaces précédemment construits afin de définir une version non-structurée des schémas multi-patch Galerkine discontinu (DG) – analyse isogéométrique (IGA) connus, puis on montre que le schéma usuel DG Bernstein-Bézier ainsi que la version complètement non-structurée du schéma IGA sont des cas particuliers de cette méthode. On montre aussi que le comportement de ces espaces de fonctions splines à proximité des bords externes et internes est très proche de celui des fonctions standards utilisées dans les schémas DG, ce qui nous permet de dériver des inégalités inverses simples et de réutiliser certains résultats connus concernant la coercivité et l'analyse d'erreur a priori qui avait été développés à l'origine pour la méthode de Galerkine discontinue avec pénalisation symétrique (IPDG). On illustre les propriétés numériques de notre schéma de discrétisation à travers un certain nombre d'expériences numériques.

Pour terminer, on explore certaines applications possibles des fonctions splines non-structurées pour le problème d'inversion sismique, en utilisant la technique de l'inversion des formes d'ondes complètes (FWI). Plus précisément, on interprète la position des nœuds définissants les fonctions splines comme des degrés de liberté d'inversion, en utilisant certaines propriétés connues sur les dérivées de ces fonctions afin d'optimiser le processus. Comme la fonction de coût utilisée en FWI n'est pas en général différentiable par rapport aux degrés de liberté géométriques, on introduit cette technique en utilisant la notion de sous-différentiel, et on montre que la technique de l'état adjoint, utilisée pour le calcul du gradient de la fonction coût, peut être dérivée simplement d'un théorème connu de dualité convexe, qui est aussi valable pour les fonctions convexes non-différentiables.

Publications and patents

Already published

[P1] H. Barucq, H. Calandra, J. Diaz, and S. Frambati*, "Polynomial-reproducing spline spaces from fine zonotopal tilings", *Journal of Computational and Applied Mathematics*, 402 (2022). Also available as a preprint: `https://hal.inria.fr/hal-02865801v3`, `https://arxiv.org/abs/2006.10307`.

[P2] H. Barucq, H. Calandra, J. Diaz, and S. Frambati*, "Unlocking the power of unstructured Isogeometric Analysis: some recent mathematical advances and a more unified framework for the numerical analysis of PDEs", *Mathias Days 2021 – Applied Mathematics, Scientific Computing, Data Science and Artificial Intelligence – TotalEnergies R&D*, Oct 3rd–7th, 2021.

[P3] H. Barucq, H. Calandra, J. Diaz, and S. Frambati*, "Unlocking the power of unstructured Isogeometric Analysis: a unified framework and some applications to wave propagation and elasticity", *IX International Conference on Isogeometric Analysis*, Sept. 27th–29th, 2021.

[P4] S. Frambati,"Performing a deformation-based physics simulation", European Patent Office, patent application no. EP21305691.4, May 26th, 2021.

[P5] H. Barucq, H. Calandra, J. Diaz, and S. Frambati*, "Unstructured multi-patch DG-IGA formulation for wave propagation", *WCCM-ECCOMAS 2020 congress*, Jan 11th–15th, 2021.

In preparation

[P6] H. Barucq, H. Calandra, J. Diaz, and S. Frambati*, "A fully unstructured multi-patch discontinuous Galerkin-Isogeometric Analysis approach for acoustic wave propagation", *in preparation.*

*Presenter/corresponding author.

Contents

General Introduction **1**

1 Waves and inversion **5**
1.1 Seismic imaging principles 5
1.2 Seismic modeling 7
1.2.1 Acoustic wave equation 7
1.2.2 Well-posedness 9
1.3 The inverse problem and Full Waveform Inversion 15
1.3.1 Fenchel-Rockafellar duality and the adjoint state method 17
1.3.2 Simple linear regression via F.-R. duality and the adjoint state method . 22
1.3.3 Full waveform inversion 24
1.3.4 Some worked-out examples 27
1.4 Discussion and further reading 32

2 Discretization, the Galerkin and IPDG methods **35**
2.1 Time integration 35
2.2 Spatial discretization 39
2.2.1 The Galerkin method 40
2.2.2 The interior penalty discontinuous Galerkin method 47
2.3 Discretization for the inverse problem 55
2.3.1 Discretization of the space of physical parameters 55
2.3.2 Minimization of the cost function 57
2.4 Discussion and further reading 58

3 Piecewise polynomial approximation and splines **61**
3.1 Piecewise-polynomial approximations 62
3.1.1 Nodal polynomials 62
3.1.2 Gauss and Gauss-Lobatto quadratures, spectral elements 67
3.1.3 Bernstein polynomials 71
3.1.4 B-splines 74
3.2 CFL condition with cardinal B-splines 79
3.2.1 B-splines and divided differences 80
3.2.2 Spectral properties of the mass and stiffness matrices 86
3.3 Discussion and further reading 91

4 Simplex spline functions **95**
4.1 B-splines as shadows of simplices . . . 96
4.2 Generalization via Dirichlet measures . . . 99
4.2.1 Dirichlet measures and Dirichlet averages . . . 99
4.2.2 Integer parameters and knot multiplicities . . . 101
4.3 Univariate simplex splines and B-splines . . . 105
4.3.1 Recurrence formulas . . . 105
4.3.2 Knot insertion formulas . . . 108
4.3.3 Spatial derivatives . . . 110
4.3.4 Knot dependence of simplex splines . . . 112
4.4 Multivariate simplex splines . . . 118
4.4.1 Recurrence and knot insertion formulas . . . 119
4.4.2 Spatial derivatives . . . 123
4.4.3 Knot dependence . . . 124
4.5 Discussion and further reading . . . 126

5 Simplex spline spaces for numerical analysis **137**
5.1 State of the art . . . 138
5.2 Background . . . 140
5.2.1 Notation . . . 140
5.2.2 Simplex splines . . . 140
5.2.3 Vector configurations and zonotopal tilings . . . 141
5.3 Polynomial-reproducing spline spaces . . . 143
5.3.1 Proof of Theorem 5.3.3 . . . 145
5.3.2 Spline space construction . . . 150
5.3.3 Relationship with centroid triangulations . . . 150
5.3.4 Link regions . . . 151
5.4 Spline spaces from regular fine zonotopal tilings . . . 155
5.4.1 Delaunay triangulations and regular zonotopal tilings . . . 156
5.4.2 Splines supported at a point . . . 157
5.4.3 Spline evaluation . . . 162
5.5 Discussion and further reading . . . 166

6 Fully unstructured multi-patch DG-IGA scheme **169**
6.1 Overview of the numerical scheme . . . 171
6.2 Multi-patch spline space construction . . . 172
6.2.1 Gabriel property . . . 173
6.2.2 Fine-grained height function . . . 174
6.2.3 Multi-patch spline space . . . 175
6.2.4 Some algorithmic aspects . . . 178
6.3 Some numerical properties . . . 181
6.3.1 Splines having a nonzero trace on constraint facets . . . 181
6.3.2 Positivity of the bilinear form . . . 182

6.3.3 *A priori* error analysis 184
6.4 Some numerical results 185
6.4.1 Block-diagonal mass matrix 185
6.4.2 Validation 186
6.4.3 Multi-patch simulation and blending with DG 188
6.4.4 Non-simply-connected domains 188
6.4.5 A simple application to hyperelasticity 195
6.5 Discussion and further reading 195

7 Towards Full Waveform Inversion **203**

Conclusions **213**

List of Figures

1.1 Seismic reflection 6
1.2 Measure condition 13
1.3 Subdifferential of a function 18
1.4 Setup of the worked-out examples 27
1.5 Degeneration of a mesh element in two dimensions 32

2.1 Finite differences, finite volumes, SPH 40
2.2 Representation of the solution in the Galerkin method 43
2.3 Representation of the solution in the discontinuous Galerkin method 51

3.1 Lagrange polynomials 65
3.2 Interpolation points for Lagrange basis 67
3.3 Bernstein polynomials 74
3.4 B-spline basis 78
3.5 Recurrence algorithm for standard B-splines 79
3.6 Cardinal B-spline basis 80
3.7 Comparision of the CFL timestep and precision of SEM, FEM, DG and IGA . . 92

4.1 B-splines as projections of simplices 98
4.2 Dirichlet distribution 100
4.3 Multivariate simplex splines 119

5.1 Schönhardt's polyhedron 141
5.2 Zonotope and zonotopal tiling 143
5.3 Spline spaces associated to zonotopal tilings 149
5.4 Link regions 155
5.5 Oriented dual graph for the determination of supported splines 161
5.6 Auxiliary function construction 164
5.7 Evaluation graph with auxiliary functions 165

6.1 Gabriel facets 174
6.2 Multivariate B-spline 178
6.3 Behavior of spline spaces near constraint facets 183
6.4 Sparsity of multi-patch DG-IGA mass matrices 187

6.5 Two-dimensional homogeneous model simulation 189
6.6 Two-dimensional bi-layered model simulation 190
6.7 Two-dimensional salt model and point configuration 191
6.8 Synthetic salt model simulation via multi-patch IGA 192
6.9 Synthetic salt model simulation via multi-patch hybrid DG-IGA 193
6.10 Detection of numerical artifacts in multi-patch hybrid DG-IGA 194
6.11 Heliosesimology-inspired 2D model simulation via multi-patch IGA 196
6.12 Simulation of a 2D music instrument model via multi-patch IGA 197
6.13 Simulation of a 2D church environment via multi-patch IGA 198
6.14 Simulation of a set of boundary conditions for an internal constraint. 199
6.15 Simulation of a three-dimensional bi-layered model 200
6.16 Simple 2D hyperelasticity simulation 201

7.1 Simple cost function minimization 207
7.2 FWI for a simple one-dimensional two-step model 208
7.3 Optimization of a simple two-dimensional step model 209

Notation, abbreviations and symbols

$\mathrm{aff}(\,\cdot\,)$	Affine span of a set of points.
$f(\,\cdot\,)$	Denotes the function f, without explicitly naming its argument.
$\partial\Omega$	Boundary of the open set Ω, i.e., $\overline{\Omega} \setminus \Omega$.
$\overline{\Omega}$	Closure of the open set Ω, i.e., the smallest closed set containing Ω.
$\mathrm{conv}(\,\cdot\,)$	Convex hull of a set of points.
$\coloneqq$, $\eqqcolon$	Symbol on the left (respectively, right) defined by expression on the right (respectively, left).
$\delta(\,\cdot\,)$	Dirac delta distribution (or measure, depending on context).
δ_{ij}	Kronecker delta symbol, i.e., $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.
$\det(A)$	Determinant of the $(d+1) \times (d+1)$ matrix $(a, 1)_{a \in A}$, ordered so that $\det(A) > 0$.
$\det({}_a^b A)$	Determinant of the points $A \setminus \{a\} \sqcup \{b\}$ ordered so that $\det(A) > 0$, and with the row a replaced by b.
$[\![\,\cdot\,]\!]$	Jump of a variable across an interface.
$\{\!\{\,\cdot\,\}\!\}$	Average of a variable across an interface.
$\widetilde{\cdot}$	Fourier transform.
$A \sqcup B$	Disjoint union of two sets.
i	Imaginary unit.
$\mathbf{1}_A(\,\cdot\,)$	Indicator function of the set A, i.e., $\mathbf{1}_A(x) = 1$ if $x \in A$, 0 otherwise.
$\left(\!\binom{n}{p}\!\right)$	Number of multisets of size p on n symbols (multichoose), equal to $\binom{n+p-1}{p}$.

$\lvert\cdot\rvert$	Cardinality of a set or L^2 norm of a vector, depending on context.
$\lVert\cdot\rVert$	Norm in a Banach space.
$\lVert\!\lvert\cdot\rvert\!\rVert$	Norm in a non-conforming (DG) space.
$\mathcal{Q}_k$	Space of polynomials up to degree k.
$\operatorname{rank}(\cdot)$	Rank of a matrix.
$\langle\cdot,\cdot\rangle$	Scalar product between vectors, or pairing in a Banach space, depending on context.
$\operatorname{sign}(\cdot)$	Sign function: ± 1 if argument positive/negative, zero otherwise.
$H^{1/2}(\partial\Omega)$	Image of the trace operator from $H^1(\Omega)$ to $L^2(\partial\Omega)$.
$H^k(\Omega)$	Sobolev space of order k on Ω, with the L^2 norm.
$L^2(\Omega)$	Space of square-integrable functions over Ω.
$M(\cdot \mid A)$	Normalized simplex spline function defined by the knot vector $A \coloneqq (a_1, \ldots, a_n)$.
∂f	Subdifferential of the convex function f.
$\operatorname{supp}(\cdot)$	Support of a function.
$\operatorname{tr}(\cdot)$	Trace of a matrix.
$\operatorname{vol}^d(\cdot)$	d-dimensional volume of a region.

General Introduction

> Two little mice fell in a bucket of cream. The first mouse quickly gave up and drowned. The second mouse wouldn't quit. He struggled so hard that eventually he churned that cream into butter and crawled out.
>
> *Frank Abagnale Sr., Catch Me If You Can (2002)*

In this short chapter, we introduce the main research challenges that have motivated and guided our investigation, with an eye towards their real-world implications and their industrial value.

Motivation

The exploration of the structure of the Earth's subsoil has long been an important avenue of scientific research, and seismic waves, be they artificially produced or naturally occurring, are one of the key tools at our disposal in this endeavour.

Traditionally, one of the main drivers of geophysical exploration has been the oil and gas industry, where seismic surveying and seismic inversion are used to find and appraise hydrocarbon resources and monitor their production. Recently, as the dramatic consequences of continued, unabated CO_2 emissions on the Earth's climate become more and more apparent, seismic exploration is finding a renewed role in the emerging industrial field of Carbon Capture, Utilization and Storage (CCUS). Here, seismic monitoring is coupled with geomechanical and reservoir flow simulations in order to understand the potential of CO_2 sinks and de-risk CO_2 injection operations, evaluating the mechanical response of the target reservoir and the possible re-activation of natural faults. New challenges, stemming from the chemical and physical properties of the injected CO_2 gas and the different operating conditions, require new computational tools, capable of quickly and accurately simulating wave propagation while efficiently representing the complex geometry of the subsoil.

Similar challenges are posed by geothermal exploration, whose goal is to assess the potential of a given zone for geothermal energy production. Two aspects are extremely important in this case. First, a geothermal exploration campaign is expected to determine the presence of subsurface heat sources and the associated natural heat flows, and their potential for exploitation. These sources tend to be located near heterogeneous geological formations, requiring powerful imaging techniques. Secondly, the presence of risk factors related to high-pressure fluids and

the possible tectonic reactivation of fractures need to be thoroughly taken into account. This aspect is even more relevant than in the case of hydrocarbon exploration and CO_2 storage, since geothermal production activities tend to be located closer to densely populated areas due to the thermal losses associated with transporting heat energy over large distances. A precise seismic imaging technique can be extremely helpful in de-risking geothermal activities, especially when acquisition campaigns are repeated over time for the early detection of tectonic and fluid shifts in a production area.

For all these reasons, there is a strong need in the field of energy production for new numerical schemes dedicated to seismic wave propagation and seismic inversion. This is an area of research that cannot be neglected in order to successfully meet the industrial challenges ahead.

Introduction

Wave propagation problems in geophysics and in engineering often require different tools. In geophysics, one has to contend with heterogeneous and often discontinuous physical properties determined by subsoil structures such as strata and salt domes, often represented via unstructured meshes. Recent works (see e.g. [1]) highlighted the advantages of Discontinuous Galerkin (DG) schemes, able to achieve high-order approximations while relying on block-diagonal matrices, well-suited for parallelization, and especially for time-explicit integration.

Engineering simulations, on the other hand, often involve homogeneous materials with complex, but known, geometries. Isogeometric analysis (IGA) [2], which replaces polynomial bases by B-spline (or Non-Rational Uniform B-Spline, a.k.a. NURBS) functions coming from Computer-Aided Design (CAD) models, has been shown to have higher efficiency per degree of freedom, better convergence in high energy modes and an improved timestep condition for wave propagation.

Recent works have started to bridge the chasm between these two worlds, by formulating a DG scheme over disconnected IGA patches, retaining the numerical advantages of the IGA formulation while allowing for the block-diagonal mass matrix characteristic of DG methods [3]. However, the tensor-product structure of conventional B-spline patches is not well-suited for applications in the natural sciences, where CAD models are not available, discontinuities are often localized and can have arbitrary topology, and inverse problems require a highly flexible geometric description.

Motivated by the need to recover the good numerical properties of the standard multi-patch DG-IGA scheme, while allowing for a more general problem geometry, we set out in this work to explore an innovative multi-patch DG-IGA scheme based on unstructured splines.

The starting point of our investigation lies in some relatively recent work on unstructured spline spaces, i.e., spaces of spline functions that are based on an unstructured set of points. Works by Neamtu [4] and Liu and Snoeyink [5] have started bringing to light some of the theoretical and algorithmic features of these functions. Specifically, Neamtu has given a beautiful geometric description of polynomial-reproducing spline spaces based on higher-order Delaunay configurations, while Liu and Snoeyink have given an explicit construction algorithm for these spaces, although limited to two dimensions and not theoretically well-suited to the most general point sets, which can potentially contain repetitions and affine dependencies. Completed by

Schmitt's proof of the convergence of the two-dimensional construction algorithm at all polynomial degrees [6], this body of work is starting to delineate the contours of a fully-formed scheme for functional approximation and numerical analysis, and the first works exploring this path are starting to appear, see e.g. [7].

However, many desirable and even essential features appear to be still lacking. First and foremost, the construction algorithm itself is only available in two spatial dimensions, and does not naturally take into account point repetitions and affine dependencies, which are essential for the proper treatment of boundary conditions in numerical simulations. Moreover, no quick algorithm for the evaluation of spline functions supported on a given point is available, limiting the computational feasibility of the approach. Finally, the relationship of these unstructured spline spaces with standard bases used in numerical analysis (FEM or DG) is not sufficiently clear. We strive in this work to bring some answers to these needs.

In Chapter 1, we briefly introduce the concept of seismic imaging, the acoustic wave equation, which will be the main application of our work, and the imaging technique known as full waveform inversion. We only tackle the most classical approach to these problems, but we follow a very general and powerful viewpoint, which can be adapted to many different variations and similar settings.

In Chapter 2, we briefly introduce the Galerkin method, the main polynomial-reproducing spaces used in its implementation, and the Interior-Penalty Discontinuous Galerkin scheme whose fluxes and penalty terms are used in our proposed DG-IGA scheme in a later chapter. In the final portion of the section, we give an explicit analytical calculation of the Courant–Friedrichs–Lewy (CFL) condition for the one-dimensional acoustic wave equation under the IGA scheme, showing the theoretical advantage of this approach in a simplified albeit representative configuration.

In Chapter 4, we introduce the mathematical foundations of multivariate (unstructured) spline functions, and derive the relevant properties that make them suitable for our applications.

In Chapter 5, we explore the connections between generalized unstructured spline spaces and some combinatorial structures known as zonotopal tilings. We show how this connection can be used to expand Neamtu, Liu and Snoeyink's work on Delaunay configurations, in order to extend their construction algorithm to all dimensions and to affinely dependent and repeated points. Moreover, we show how the combinatorial properties of these objects can be used to derive an efficient algorithm for the determination of all spline functions supported on a point and their numerical evaluation.

Chapter 6 exploits the results of the previous chapter in order to introduce a fully unstructured multi-patch DG-IGA scheme for the simulation of PDEs. We show how to exploit knot multiplicity to carve out disconnected sub-regions of the simulation domain, that are subsequently coupled via DG fluxes. We demonstrate that our approach can reproduce, for extreme choices of parameters, both a pure DG scheme on unstructured meshes, and a pure IGA scheme, thus allowing to fine-tune the level of domain decomposition and thus the size of the blocks in the mass matrix. Results on two- and three-dimensional wave propagation and two-dimensional elasticity are also presented in Chapter 6, showing that the favorable properties of IGA are retained, along with the parallelization capabilities of DG-like methods.

Chapter 7 is dedicated to the perspective application of our method to seismic full-waveform

inversion (FWI) [8], a proven technique capable of estimating the properties of the subsoil, such as velocity and density, from seismic data. This method has become increasingly attractive in the last decade, as it is capable of creating accurate, high-resolution models of the subsurface even in the presence of complex geological structures. However, the cost-function minimization underlying this technique is inherently non-convex, and can become trapped in local minima due to an unsuitable initial guess, lack of low-frequency content, or noise. Crucially, the number of inversion parameters in the velocity model seems to be an important driver of the stability of the inversion [9]. In this context, an intrinsically mesh-less method such as the one explored in this work seems to be particularly well-suited for the precise determination of the model geometry with a limited amount of model parameters, thus increasing the inversion stability without reducing its precision.

Finally, we draw some conclusions, and discuss some perspectives and avenues of improvement.

Software availability

A piece of software for spline space construction and evaluation, two- and three-dimensional wave propagation and two-dimensional elasticity was written during the course of this project and is available upon request on the Inria git repository (`https://gitlab.inria.fr/sframbat/iga-dg`).

1 | Waves and inversion

> The earth keeps some vibration going
> There in your heart, and that is you.
>
> *Edgar Lee Masters, Spoon River Anthology, Fiddler Jones (1915)*

We introduce in this chapter the main features of seismic surveying, seismic wave propagation and seismic inversion that underlie and motivate a substantial part of our work.

In the first part, we take a look at seismic surveying, its importance in industrial applications for energy and oil & gas, and how it is used to gain an understanding of the subsoil and its structure. We then focus on the acoustic wave equation, with the boundary conditions commonly used in seismic applications, we introduce its adjoint problem and we use it to show that the wave simulation problem (i.e., the *forward problem*) is well-posed. Finally, we introduce the concept of seismic inversion, the inverse problem associated with wave propagation. We introduce the associated cost function, and we adopt a slightly less common but more general point of view, showing how convex duality can be used to introduce the *adjoint state method*, in a very general form and even when the cost function is not differentiable. We exemplify these concepts using a simplified model based on simple linear regression, before moving to the full-fledged seismic inversion formulation. Lastly, we explicitly work out the simple case of one-dimensional seismic inversion on a few simple models, illustrating its main features with some concrete examples.

1.1 Seismic imaging principles

Seismic depth imaging is an indirect process that allows to recover the geometric structure and distribution of the physical parameters of the subsoil (called in geophysics the *subsurface*) from measured seismic data. In most practical cases, this goal is achieved through *seismic reflection*, a process whereby an oscillation (a shock wave) is imparted to the surface of the earth, and the reflected waves are detected and measured at a set of locations on the surface [10]. The features of the subsurface are then deduced from the measured data through various computational techniques, an example of a (rather complex) *inverse problem*. The physics underlying this process is the same as in ultrasound techniques used in the medical field, although the scale, range of wave frequencies and source and receiver types are all quite different.

The physical measurements take place during seismic acquisition campaigns over a given area, named *seismic surveys*. These measurements can be performed either on land or at sea; in the latter case, the source of oscillations and the receivers are both placed at the surface

of the water, usually on a boat with trailing cables called *streamers*. Physically, the source of oscillations is a boat-mounted air gun for marine campaigns or a vibrator truck for terrestrial ones. Each source activation (or *shot*) has a very limited time duration, and a main frequency in the range of a few Hertz (i.e., oscillations per second) to a few tens of Hertz. Many shots are performed, and for each shot, many receivers (*geophones*) are present, usually distributed along a line. Using this setup, many different locations are probed, usually forming a set of lines in

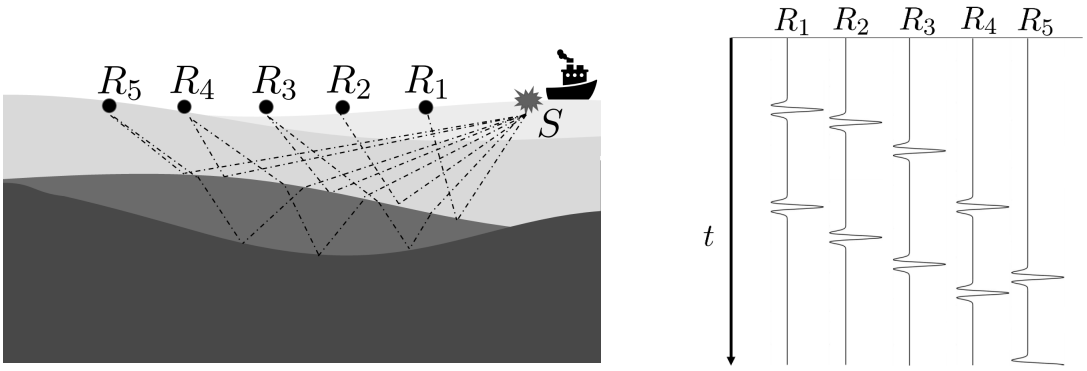


Figure 1.1: (Left) Seismic reflection in a marine environment. The source (air gun) produces waves that propagate into the subsoil, get reflected by the main sedimentary interfaces, and are recorded at a series of trailing receivers. Many shots are performed at different locations. (Right) Seismic traces $d_r(t)$ recorded by the receivers.

land campaigns (2D acquisition), or a regular grid for marine campaigns (3D acquisition).

In a seismic campaign, the oscillations induced in the subsurface by the artificial source can be subdivided into *P*- (pressure or primary) waves and *S*- (shear or secondary) waves, according to whether the oscillation occurs in a plane parallel or perpendicular to the direction of propagation of the wave, respectively [11]. Notice that, although a liquid such as water is not able to carry S-waves, these oscillations are nonetheless generated when a pure pressure wave reaches the water bottom and interacts with the solid surface. In addition to these waves, surface waves are also produced, that travel directly from the source to the geophones. All these waves have different speeds in different materials, and all must be taken into account in order to correctly explain the measured data. Surface waves can carry information about the shallow and medium-depth structures of the subsurface, for which they find ample use especially in geophysics and civil engineering (see, e.g., [12–14]), but do not contain information about the deepest part of the subsurface, and are often filtered out from seismic data in a pre-processing step based on time-of-arrival.

The literature on seismic imaging is very large, and even a brief overview of the main techniques would take us out of the scope of this work. Instead, we refer the interested reader to some of the many very complete texts on the subject, as for example [15, 16], and all the references contained there. Here, we will focus mainly on seismic imaging via P-waves, i.e., *acoustic seismic imaging*, through a technique known as *full waveform inversion* (FWI). We introduce these concepts in the next sections.

1.2 Seismic modeling

We discuss here the origin of the acoustic wave propagation problem and the mathematical properties that are important for the development of our work. We introduce the wave equation based on the theory of linear elasticity. Although nonlinear effects have been proven to be important for the full description of the physics of seismic waves, we will not explore this aspect in the present work, and we will only focus on linear waves. For a simple, beautiful and intuitive introduction to these topics, we refer the reader to Chapters 31, 38 and 39 of the timeless classic [17].

1.2.1 Acoustic wave equation

When a piece of material is subject to a force (stress), it deforms (strains). If the force is small enough, the deformation, i.e., the relative displacement of the various positions in the material, is linearly proportional to the force, an *elastic* behaviour captured by Hooke's law. Consider a piece of material occupying a region $\Omega \subset \mathbb{R}^d$, and suppose that it is deformed by a (small) displacement field $u : \Omega \mapsto \mathbb{R}^d$ such that the point at position $x \in \Omega$ finds itself at position $x + u(x)$. Since a constant displacement is simply a translation, the deformation must be given by the first derivatives of u, i.e.,

$$\epsilon_{ij} \coloneqq \frac{\partial u_i}{\partial x_j}. \tag{1.1}$$

However, the antisymmetric part of ϵ merely describes rotations, which are also rigid motions. Therefore, the deformation induced by u is described by the symmetric part of (1.1), i.e., the *strain tensor*

$$\varepsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right). \tag{1.2}$$

By Hooke's law, if we consider a test area A^j normal to the direction j and measure the i-th component of the force σ^j through it, i.e., $(\sigma_i^j) =: \sigma_{ij}$, we find a linear relationship with ε,

$$\sigma_{ij} = \sum_{k,l=1}^{d} C_{ijkl}\varepsilon_{kl}, \tag{1.3}$$

The linear map C is called the *stiffness* or *elasticity* tensor and is in general a function of x. Notice that the term σ of (1.3) is a force per unit area. To obtain the total net force per unit volume on any test volume, one may simply take the (signed) average of the contributions on opposite areas, i.e., the divergence

$$f_i = \sum_{j=1}^{d} \frac{\partial \sigma_{ij}}{\partial x_j}.$$

By Newton's law, f_i is proportional to the second time derivative of the displacement through the inertia given by the density ρ. If we also add an external forcing $g = g(x,t)$, we obtain the

general *elastic wave equation*

$$\rho \frac{\partial^2 u}{\partial t^2} = \sum_{j=1}^{d} \frac{\partial}{\partial x_j} \left(\sum_{k,l=1}^{d} C_{ijkl} \varepsilon_{kl} \right) + g. \tag{1.4}$$

In the present work, we wish to focus on the foundational aspects of our numerical scheme. To avoid cluttering our presentation and clouding our results, we simplify the physical systems of interest by removing all anisotropy and focusing only on pressure waves (equivalently, setting the Poisson's ratio of the material to zero). This restriction does not hide the main characteristics of the wave equation that are important in view of our results, i.e., its hyperbolic character, the spatial dependence of the physical parameters, and the convergence characteristics of the associated inverse problem. Henceforth, we will only deal with the *acoustic wave equation*. For an isotropic material, (1.3) can be simplified to

$$\sigma_{ij} = \lambda \delta_{ij} \sum_{k=1}^{d} \varepsilon_{kk} + 2\mu \varepsilon_{ij},$$

where λ and μ are known as *Lamé parameters*. If the *shear modulus* μ is equal to zero, then σ is proportional to the identity. Taking the (over)pressure $p \coloneqq \frac{1}{3} \operatorname{tr} \sigma$ as our main variable, we have from (1.2) and (1.3)

$$p = \lambda \nabla \cdot u,$$

and, taking the second time derivative and using (1.4), we arrive to the final form of the acoustic wave equation,

$$\frac{1}{\lambda} \frac{\partial^2 p}{\partial t^2} - \nabla \cdot \left(\frac{1}{\rho} \nabla p \right) = s, \tag{1.5}$$

where $s = s(x,t)$ is the source term. The P-wave velocity can be recovered via $c^2 = \lambda/\rho$. Notice that in usual applications, these physical parameters do not depend on time. Their possible time dependence, due to fluid flow in reservoirs and/or the movement of faults and other tectonic features, are much too slow to be relevant in a single seismic campaign, and are only picked up by repeating the acquisition campaign over many periods of time spread over multiple years.

The physical parameters are usually taken from a well-definite space of (positive) functions, such as continuous, piecewise constant or piecewise polynomial functions, but for some applications such as scattering from small obstacles or lattices (see, e.g., [18–21] and references therein), more general choices can be required. Thus, in all generality, the physical parameters ρ and λ are assumed to be positive *measures* on Ω.

The source term $s(x,t)$ used in seismic surveys is usually taken to be a point-like *Ricker wavelet* [22], whose time dependence is given by the second derivative of a Gaussian, i.e.,

$$s(x,t) = A\delta(x - x_s) 2\pi \nu_s^2 \left(1 - 2\pi^2 \nu_s^2 (t - t_s)^2\right) e^{-\pi^2 \nu_s^2 (t - t_s)^2}, \tag{1.6}$$

where A is an amplitude parameter with units of volume, $\delta(x)$ is the Dirac delta distribution, x_s is the spatial location of the source (usually at the surface of the earth or of the water), t_s is the source delay, and ν_s is the source peak frequency, usually in the range $1 \div 100$Hz.

The usual setup for the solution (1.5) is an *initial value problem* with a given value of $u(x,t)$ and its time derivative at $t=0$, usually imposing no pressure and no inertia,

$$p(x,0) = \frac{\partial p}{\partial t}(x,t)\bigg|_{t=0} = 0. \tag{1.7}$$

Additionally, a set of boundary conditions on $\partial\Omega$ is applied throughout the simulation time. For seismic surveys, the top surface of the domain $\partial\Omega_F$, corresponding to the surface of the water or the earth, is usually taken to be a *free surface*, i.e., in the acoustic case, a simple Neumann condition is imposed,

$$\frac{\partial p}{\partial n}\bigg|_{\partial\Omega_F} = 0, \tag{1.8}$$

where $\partial/\partial n$ denotes the normal derivative. The remaining portion of the boundary, $\partial\Omega_A \coloneqq \partial\Omega \setminus \partial\Omega_F$, is used to approximate the infiniteness of the domain using various techniques. Among the most widely used approaches are *perfectly matched layers* (PMLs, see, e.g., [23]) and *absorbing boundary conditions* (ABCs, see, e.g., [24]). In the present work, we use the lowest-order absorbing boundary conditions, which simply reads [25]

$$\frac{\partial p}{\partial t}\bigg|_{\partial\Omega_A} + c\,\frac{\partial p}{\partial n}\bigg|_{\partial\Omega_A} = 0. \tag{1.9}$$

These conditions are exact for $d=1$ and reasonably accurate in $d\geq 2$ for incidence angles less than about $\pi/6$.

1.2.2 Well-posedness

We wish to show that the acoustic wave problem determined by (1.5) with initial conditions (1.7) and boundary conditions (1.8) and (1.9) is *well-posed* in Hadamard's sense, i.e., that under mild technical assumptions on Ω and the regularity of the physical parameters ρ, λ, c, and the source term $s(x,t)$, this problem has exactly one solution $p(x,t)$ for $x\in\Omega$ and $t\in\mathbb{R}_{\geq 0}$ and this solution depends continuously on the problem parameters. This result is proven in [26], but we give here the main argument as it can be rather instructive, and it will help us introduce the notion of adjoint operator.

Some elements of Hille-Yosida theory on Hilbert spaces

Consider a first-order time-evolution problem expressed as

$$\frac{\partial u}{\partial t} = Au + s, \tag{1.10}$$

where u belongs to a given real Hilbert space $\mathcal{H}$ where the domain $D(A)$ of A is dense. The well-posedness of (1.10) is often tackled by proving that there exists a unique one-parameter semigroup $\mathcal{T}$ of *time evolution operators* $T(t)$, $t\in\mathbb{R}_{\geq 0}$, with $T(0)=I$ and $T(r)T(s)=T(r+s)$, such that the solution at time t can be simply expressed via Duhamel's formula (see, e.g., [27,

28])

$$u(x,t) = T(t)u(x,0) + \int_0^t T(t-\tau)s(x,\tau)\,\mathrm{d}\tau, \tag{1.11}$$

and that the map $t \mapsto T(t)u(x,0)$ is strongly continuous (i.e., continuous in the norm of $\mathcal{H}$). The expression given for $u(x,t)$ by (1.11) then represents the unique solution of (1.10). If $\mathcal{H}$ were finite-dimensional, the solution to (1.10) would be trivially obtained by exponentiating A, i.e.,

$$u(t) = e^{tA}u(0) + \int_0^t e^{(t-\tau)A}s(\tau)\,\mathrm{d}\tau.$$

In the infinite-dimensional case, however, things are not quite that simple, and even for bounded operators, one must be careful with the definition of the operator domain. There are however a few theorems, part of the Hille-Yosida theory, that allow the construction of $\mathcal{T}$ if A is *dissipative*, i.e., if

$$\langle Au, u\rangle_{\mathcal{H}} \leq 0 \text{ for all } u \in D(A).$$

Intuitively, A is dissipative if its spectrum is negative, and there is a nonnegative quantity, encoded by the norm on $\mathcal{H}$, that can be interpreted as the "energy" of the system and is never increased by the time evolution. The other conditions that are required of A rely on its *adjoint* operator.

Recall that the *adjoint* $A^\star$ of A in the Hilbert space $\mathcal{H}$, endowed with the scalar product $\langle\,\cdot\,,\,\cdot\,\rangle_{\mathcal{H}}$, is defined as follows. First the domain $D(A^\star) \subseteq \mathcal{H}$ is defined as the set of all elements $u' \in \mathcal{H}$ such that the linear functional $\varphi(u) \coloneqq \langle Au, u'\rangle_{\rho,\lambda}$ is continuous for all $u \in D(A)$. Since $D(A)$ is dense in $\mathcal{H}$, by the Hahn-Banach theorem, this linear functional can be extended to the whole $\mathcal{H}$, i.e., to an element of the dual $\mathcal{H}^\star$. By the Riesz representation theorem, there is a unique element $v \in \mathcal{H}$ such that

$$\left\langle Au, u'\right\rangle_{\mathcal{H}} = \langle u, v\rangle_{\mathcal{H}} \text{ for all } u \in \mathcal{H}.$$

We then simply declare

$$A^\star u' \coloneqq v.$$

Then, it is always true that, for $u \in D(A)$ and $u' \in D(A^\star)$,

$$\left\langle Au, u'\right\rangle_{\mathcal{H}} = \left\langle u, A^\star u'\right\rangle_{\mathcal{H}}, \tag{1.12}$$

which is the defining property of the Hilbert adjoint.

We are now ready to state two fundamental theorems in (Hilbert space) Hille-Yosida theory. Let the domain $D(A)$ be dense in $\mathcal{H}$, and let the operator A be *closed*, i.e., let the graph of A be a closed set in $\mathcal{H} \times \mathcal{H}$. Then the two following theorems hold.

Theorem 1.2.1 (Stone's theorem for Hilbert spaces, [29])**.** *There exists a unique semigroup of strongly continuous time evolution operators* (1.11) *that are unitary, i.e., they preserve the norm on $\mathcal{H}$, if and only if A is skew-symmetric, i.e., $A = -A^\star$.*

Theorem 1.2.2 (Lumer-Phillips' theorem for Hilbert spaces, [30])**.** *There exists a semigroup of strongly continuous time evolution operators* (1.11) *that are contractions, i.e., they do not*

increase the norm on $\mathcal{H}$, if and only if A is closed and both A and $A^\star$ are dissipative.

Notice that in the statement of Stone's theorem, dissipativity is not required, since by (1.12), a skew-self-adjoint operator satisfies $\langle Au, u\rangle_{\mathcal{H}} = 0$ and is thus automatically dissipative. In the following subsections, we show how the Hille-Yosida theory can be applied to the acoustic wave problem to construct a semigroup of contraction operators and thus prove its well-posedness.

The acoustic wave operator and the simulation domain

Let us now return to the acoustic wave equation (1.5). If we consider a (generic) displacement field $u(x,t)$ propagating through space according to the wave equation

$$\alpha \frac{\partial^2 u}{\partial t^2} = \nabla \cdot (\beta \nabla u)\,, \tag{1.13}$$

it is well-known that the total *energy* of this system, expressed in terms of the usual sum of a kinetic term and a potential term,

$$E_u(t) \coloneqq \frac{1}{2}\int_\Omega \alpha \left(\frac{\partial u}{\partial t}\right)^2 + \beta\,|\nabla u|^2 \,\mathrm{d}\Omega,$$

is conserved. Comparing (1.13) with (1.5), one can see that the quantity

$$E_p(t) \coloneqq \frac{1}{2}\int_\Omega \frac{1}{\lambda} \left(\frac{\partial p}{\partial t}\right)^2 + \frac{1}{\rho}\,|\nabla p|^2 \,\mathrm{d}\Omega,$$

even if it does not have the physical unit of an energy, is a good candidate to establish the dissipativity of the acoustic wave problem.

In this spirit, let us identify $(p(x,t), 1/\lambda\,\partial_t p(x,t))$ with a vector $P(x,t) = (p_1(x,t), p_2(x,t))$ in the Hilbert space $H^1_\rho(\Omega) \oplus L^2_\lambda(\Omega)$, analogous to the usual space $H^1(\Omega) \oplus L^2(\Omega)$, but where the finite positive measures ρ, λ on Ω are used to define the modified scalar product $\langle\,\cdot\,,\,\cdot\,\rangle_{\rho,\lambda}$ as

$$\langle (p_1,p_2),(q_1,q_2)\rangle_{\rho,\lambda} \coloneqq \int_\Omega \nabla p_1 \cdot \nabla q_1 \frac{\mathrm{d}\Omega}{\rho} + \int_\Omega p_2 q_2 \lambda \,\mathrm{d}\Omega. \tag{1.14}$$

The distributional derivatives are taken in the weak sense using this scalar product. We can write (1.5) in the form (1.10) using the vector variable $P(x,t)$ as

$$\frac{\partial P}{\partial t} = AP + S, \tag{1.15}$$

where

$$A \coloneqq \begin{pmatrix} 0 & \lambda \\ \nabla \cdot \left(\frac{1}{\rho}\nabla\right) & 0 \end{pmatrix} \text{ and } S \coloneqq \begin{pmatrix} 0 \\ s(x,t) \end{pmatrix}. \tag{1.16}$$

Notice that we can recover the original pressure field by identifying $p(x,t) \coloneqq p_1(x,t)$ and $\partial_t p(x,t) \coloneqq \lambda p_2(x,t)$.

We want to solve (1.15) distributionally, i.e., by projecting it on test functions $\Phi \coloneqq (\varphi_1, \varphi_2) \in C_0^\infty(\Omega) \times C_0^\infty(\Omega)$. We will assume hereafter that Ω is a bounded open subdomain of $\mathbb{R}^d$, and

thus we can write

$$\int_\Omega (AP)\cdot\Phi\,\mathrm{d}\Omega \coloneqq \int_\Omega p_2\varphi_1\lambda\,\mathrm{d}\Omega - \int_\Omega \nabla p_1\cdot\nabla\varphi_2\frac{\mathrm{d}\Omega}{\rho}.$$

In order to define the domain $D(A)\subset H^1_\rho(\Omega)\oplus L^2_\lambda(\Omega)$, we need to incorporate the boundary conditions (1.8) and (1.9), and therefore we need to place some constraints on Ω and define a suitable trace operator.

Following similar arguments that exist in the literature for comparable differential operators (see, e.g., [31] for the Schrödinger equation), assume that the boundary $\partial\Omega$ is piecewise C^1 and *locally Lipschitz*, i.e., locally the graph of a Lipschitz function. A boundary that is piecewise smooth and whose pieces connect through edges and corners forming positive angles is indeed locally Lipschitz, see, e.g., [32]. This includes all polytopal (e.g., polygonal or polyhedral) domains, and is clearly more than enough for most numerical simulations. If Ω satisfies this regularity property, then the normal vector to $\partial\Omega$ can be defined almost everywhere, and one can apply Stein's extension theorem and Sobolev's embedding theorem (see, e.g., [32, Chapter 5] or the Raviart-Thomas' approach [33]) to show that there is a linear, continuous and surjective trace operator $\mathrm{Tr}: H^1(\Omega)\mapsto H^{1/2}(\partial\Omega)$, with a linear, continuous right inverse. By using this classical embedding, one can evaluate the trace of every function in $H^1(\Omega)$ on $\partial\Omega$, and, if the function sits in $H^2(\Omega)$, also its normal derivatives.

However, there is still a small caveat: in our case, the Sobolev space depends on a non-standard metric induced by ρ and λ. Even so, if a measure μ satisfies the following *measure density condition*,

$$\begin{aligned}&\text{there exists a constant } C>0, \text{ s.t. } \mu(B_{x,r}\cap\Omega)\geq Cr^d\\ &\text{for every ball } B_{x,r} \text{ centered in } x\in\Omega \text{ with radius } 0<r\leq 1,\end{aligned}\tag{1.17}$$

then it can be proven that the extension theorem also applies to the modification of the Sobolev space $H^1(\Omega)$, where the scalar product is taken with respect to the measure μ (see [34] and [35, Theorem 5]). Condition (1.17) requires that the region near the boundary is well behaved, i.e., that (Ω,μ) does not become too thin near the boundary, or more precisely, that it does not go to zero faster than r^d. This implies in particular that $\mu(B_{x,r}\cap\Omega)/\left|B_{x,r}\right|$ goes to a finite positive constant as $r\to 0$ for all $x\in\partial\Omega$, cf. [34, Lemma 2.1] and Figure 1.2, and thus that the Lebesgue measure of $\partial\Omega$ must be zero and μ must be well-behaved near the boundary. The measure density property is certainly satisfied by every positive, bounded and well behaved set of physical parameters ρ, λ and regular domains that are interesting for numerical analysis. For example, it is satisfied by smooth or piecewise smooth functions, and even if a set of discrete Dirac masses is added to the interior of Ω, as long as the measure stays positive.

As a consequence of this construction, by choosing P in $H^2_{\rho,\lambda}\oplus H^1_{\rho,\lambda}$, p_1, p_2 and the first derivatives of p_1 all possess a well-defined trace on $\partial\Omega$. If we furthermore assume that the measures λ, ρ also possess a well-defined positive trace such that $c\coloneqq\sqrt{\lambda/\rho}$ is well-defined a.e. on $\partial\Omega$, then (1.8) and (1.9) become meaningful linear conditions on $\mathcal{H}_{\rho,\lambda,\Omega}$, the extension of $H^2_{\rho,\lambda}\oplus H^1_{\rho,\lambda}$ through the Sobolev embedding theorem discussed above. We can now define the

Figure 1.2: A well-behaved domain (equivalently, measure), satisfying the measure condition (1.2) (left), and a domain not satisfying the condition (right).

domain of the operator A as

$$D(A) \coloneqq \{(p_1, p_2) \in \mathcal{H}_{\rho,\lambda,\Omega} : A(p_1, p_2) \in \mathcal{H}_{\rho,\lambda,\Omega},\ (\partial_n p_1)|_{\partial\Omega_F} = 0,\ (\partial_n p_1)|_{\partial\Omega_A} + \lambda\, p_2|_{\partial\Omega_A} = 0\}.$$

It is easy to see that the domain $D(A)$ is dense in $\mathcal{H}_{\rho,\lambda,\Omega}$, and in fact in $L^2(\Omega) \oplus L^2(\Omega)$, since it contains $C_0^\infty(\Omega) \times C_0^\infty(\Omega)$. Moreover, A is closed in $\mathcal{H}_{\rho,\lambda,\Omega}$. In fact, its domain is complete as it is obtained from a complete space via a simple linear condition, and since the image of A is contained in $\mathcal{H}_{\rho,\lambda,\Omega}$, it follows by standard arguments that A is bounded. Thus, A satisfies the prerequisites required to apply the machinery of Hille-Yosida theory introduced above.

Before moving to the next subsection, we need to point out that, in order for the absorbing boundary condition (1.9) (and the adjoint condition (1.21)) to make physical sense, the domain Ω must also chosen to be *convex*. In fact, perfectly absorbing boundary conditions (of which (1.9) are an approximation) rely on the hypothesis that waves that leave Ω do not make a contribution to the signal at the receivers, and can therefore be ignored. This is usually the case if the propagation time is small enough (or the domain large enough) so that waves reflected by heterogeneities outside Ω do not have the time to travel back to the receivers and be measured. However, if Ω is not convex, some waves could exit Ω only to re-enter after a very short time, and therefore cannot be safely ignored. Nevertheless, this requirement, however important for real-world applications, is not necessary for the well-posedness of the acoustic problem with the lowest-order absorbing boundary conditions (1.9).

Dissipativity and the adjoint operator

Once the base Hilbert space $\mathcal{H}_{\rho,\lambda,\Omega}$ has been defined, we wish to show that the operator A is dissipative. We can easily compute

$$\begin{aligned}
\langle A(p_1, p_2), (q_1, q_2)\rangle_{\rho,\lambda} &= \int_\Omega \nabla(\lambda p_2) \cdot \nabla q_1 \frac{\mathrm{d}\Omega}{\rho} + \int_\Omega q_2 \nabla \cdot \left(\frac{1}{\rho}\nabla p_1\right) \lambda\, \mathrm{d}\Omega, \qquad (1.18)\\
&= \int_\Omega \nabla(\lambda p_2) \cdot \nabla q_1 \frac{\mathrm{d}\Omega}{\rho} - \int_\Omega \nabla(\lambda q_2) \cdot \nabla p_1 \frac{\mathrm{d}\Omega}{\rho} + \int_{\partial\Omega} c^2 q_2 \frac{\partial p_1}{\partial n} \cdot\, \mathrm{d}S,\\
&= \int_\Omega \nabla(\lambda p_2) \cdot \nabla q_1 \frac{\mathrm{d}\Omega}{\rho} - \int_\Omega \nabla(\lambda q_2) \cdot \nabla p_1 \frac{\mathrm{d}\Omega}{\rho} - \int_{\partial\Omega_A} \lambda c\, p_2 q_2\, \mathrm{d}S.
\end{aligned}$$

If there were no absorbing boundary conditions, i.e., if $\partial\Omega_A = \varnothing$, the last term in (1.18) would vanish and the operator A would in fact be skew-symmetric. Stone's theorem would then apply,

and A would generate a unitary time evolution semi-group, i.e., it would conserve the energy associated to the scalar product $\langle \cdot, \cdot \rangle_{\rho,\lambda}$. The absorbing boundary condition is therefore solely responsible for all the energy change inside Ω, as expected. In any case, A is *dissipative*, since

$$\langle A(p_1, p_2), (p_1, p_2) \rangle_{\rho,\lambda} = -\int_{\partial\Omega_A} \lambda c\, p_2^2 \, \mathrm{d}\Omega \leq 0.$$

One can use (1.18) to compute the Hilbert adjoint $A^\star$. In fact, imposing the adjoint condition (1.12) yields

$$\begin{aligned} \langle (p_1, p_2), A^\star(q_1, q_2) \rangle_{\rho,\lambda} &= \langle A(p_1, p_2), (q_1, q_2) \rangle_{\rho,\lambda}\,, \\ &= \int_\Omega \nabla(\lambda p_2) \cdot \nabla q_1 \frac{\mathrm{d}\Omega}{\rho} - \int_\Omega \nabla(\lambda q_2) \cdot \nabla p_1 \frac{\mathrm{d}\Omega}{\rho} - \int_{\partial\Omega_A} \lambda c\, p_2 q_2 \, \mathrm{d}S, \\ &= -\int_\Omega p_2 \nabla \cdot \left(\frac{1}{\rho} \nabla q_1 \right) \lambda \, \mathrm{d}\Omega + \int_{\partial\Omega_A} c^2 p_2 \frac{\partial q_1}{\partial n} \cdot \mathrm{d}S - \int_\Omega \nabla(\lambda q_2) \cdot \nabla p_1 \frac{\mathrm{d}\Omega}{\rho} \\ &\qquad - \int_{\partial\Omega_A} \lambda c\, p_2 q_2 \, \mathrm{d}S, \\ &=: \left\langle (p_1, p_2), (q_1', q_2') \right\rangle_{\rho,\lambda}\,, \end{aligned} \tag{1.19}$$

where the last step gives the definition of $(q_1', q_2') = A^\star(q_1, q_2)$. Comparing (1.14) and (1.19), by equating the integrals on Ω and canceling the boundary integrals in (1.19), we deduce that

$$A^\star = \begin{pmatrix} 0 & -\lambda \\ -\nabla \cdot \left(\frac{1}{\rho} \nabla \right) & 0 \end{pmatrix} \tag{1.20}$$

and

$$D(A^\star) \coloneqq \{ (p_1, p_2) \in \mathcal{H}_{\rho,\lambda,\Omega} : A^\star(p_1, p_2) \in \mathcal{H}_{\rho,\lambda,\Omega},\ (\partial_n p_1)|_{\partial\Omega_F} = 0,\ c\,(\partial_n p_1)|_{\partial\Omega_A} - \lambda\, p_2|_{\partial\Omega_A} = 0 \}.$$

In other words, the operator A is *almost* skew-symmetric, in the sense that its adjoint can be obtained by taking minus the differential operator in (1.5), and replacing the absorbing boundary conditions (1.9) by the *adjoint absorbing conditions*

$$\left.\frac{\partial p}{\partial t}\right|_{\partial\Omega_A} - c \left.\frac{\partial p}{\partial n}\right|_{\partial\Omega_A} = 0, \tag{1.21}$$

where the sign of the normal derivative has been changed. It can be then easily verified that $A^\star$ is also dissipative, by noticing that, compared to the derivation (1.18), the minus sign in (1.20) and the opposite sign of the boundary conditions (1.21) conspire to keep the sign of the last term in (1.18) unchanged. Consequently,

$$\langle A^\star(p_1, p_2), (p_1, p_2) \rangle_{\rho,\lambda} = -\int_{\partial\Omega_A} \lambda c\, p_2^2 \, \mathrm{d}\Omega \leq 0$$

for all $(p_1, p_2) \in D(A^\star)$.

We can therefore invoke Lumer-Phillips' theorem to conclude that the acoustic problem is well posed. We restate this result through the following theorem, which summarizes all the major hypotheses made so far on Ω, ρ and λ.

Theorem 1.2.3. *Let $\Omega \subseteq \mathbb{R}^d$ be an open set whose boundary $\partial\Omega$ is piecewise C^1 and locally Lipschitz, and let $\partial\Omega_F$ and $\partial\Omega_A$ be two disjoint, locally Lipschitz and piecewise C^1 subsets of $\partial\Omega$. If ρ and λ are positive finite measures on Ω satisfying* (1.17) *and c is a well-defined positive measure on $\partial\Omega$, then the problem described by* (1.5) *with boundary conditions* (1.7), (1.8) *and* (1.9) *and its adjoint are both well-posed. Moreover, if the source $S(x,t) \in C^0(L^2(\Omega); [0, +\infty))$, then $P \in C^0(\mathcal{H}_{\rho,\lambda,\Omega}; [0, +\infty))$.*

Proof. The first part was proven in the text of the last two subsections. The regularity of the solution is a simple consequence of the strong continuity of the contraction semigroup and the expression of P via Duhamel's formula (1.11) with $P(x, 0) = 0$, see, e.g., [36]. □

As we will see in the next section, the adjoint operator introduced here will come in handy when discussing the seismic inverse problem.

1.3 The inverse problem and Full Waveform Inversion

In real applications, one is rarely interested in the solution of (1.5) *per se*. Rather, one is given the result of some pressure (or other) measurements $D_r(t)$, $r \in R_i$, $i = 1, \ldots, n_s$, where n_s is the number of shots and R_i is the set of receiver indices for the i-th shot, with receiver positions $\{x_r\}_{r\in R}$. One is then tasked with finding the physical parameters $\lambda = \lambda(x)$, $\rho = \rho(x)$ of (1.5) that best fit the experimental data. The resulting *inverse problem* is called *seismic inversion*. Notice that, using (1.11), the wave propagation problem can be recast as a *convolution filter*, parameterized by the physical properties of the sub-soil, that takes the source data and transforms it into receiver data. Seismic inversion is then the corresponding deconvolution operation [15].

Most inversion methods start from a tentative guess of the physical model, and proceed to iteratively update the model to match the receiver data. For each shot point, we denote by $s_i(x,t)$, $i = 1, \ldots, n_s$ the corresponding source term, and by $P_i(x,t)$ the corresponding solution to (1.15). The misfit between the current model and the measurements can be quantified by the *cost function*, which is to be (locally) minimized through modifications to the model $m \coloneqq (\rho, \lambda)$. The most common form for the cost function is a simple L^2 distance, expressed as

$$J(m) \coloneqq \sum_{i=1}^{n_s} \sum_{r\in R_i} \int_0^{t_f} \|P_i(x_r, t) - D_r(t)\|^2 , \tag{1.22}$$

where each P_i is a solution of (1.15) with physical model m and source term s_i, see, e.g., [37]. The time interval $[0, t_f]$ is chosen to include the whole duration of the source and receiver data, typically a few seconds. In the literature, many other misfit functions have been used instead of (1.22), mainly with the goal of improving robustness to noise or computational performance, see, e.g., [38–44] for a non-exhaustive list. For the sake of simplicity, however, we will not explore

these possibilities here, but we will nonetheless choose an approach general enough to handle most of the possible variations of the cost function. Thus, our goal is the minimization of (1.22) over all allowable λ, ρ. This minimization, where the pressure values are computed as solutions to the wave equation, is called *full waveform inversion* (FWI).

Computing the cost function for a given choice of physical parameters requires the solution of a full acoustic wave problem, and its minimization needs many such evaluations, which can be very computationally expensive. On the other hand, using the full wave equation without significant approximation allows to extract the maximum amount of information from the measured data. Consequently, FWI is a very powerful technique, capable of reconstructing very complex subsurface structures, but which needs a significant amount of computational power in order to be applied to real-world problems.

The first, very fundamental problem that we face in the minimization of (1.22) is the fact that the cost function J as defined is generally not convex. In fact, if P and P' are two solutions to the acoustic wave problem with the same source term and different physical parameters, there does not necessarily exist a set of physical parameters such that the convex combination $\gamma P + (1-\gamma)P$, $0 \leq \gamma \leq 1$ is a solution with the same source. The non-convexity of the seismic inversion problem is a very well known fact, see, e.g., [45, 46], and it means that the best we can hope for is to obtain a *local* minimization of J, i.e., to find the local minimum around a given starting point $\overline{\rho}$, $\overline{\lambda}$, the *initial guess*. Therefore, our search is limited to a subspace of the space of finite positive measures, the local *attraction basin* in which the initial model lies.

In the literature, the local minimization of (1.22) is usually performed via a simple gradient descent, although full Newton-type algorithms involving the computation of the Hessian have been proposed and their efficacy weighted against their accrued cost, see, e.g., [40, 47, 48]. As we will see, the cost of computing the Hessian can indeed be prohibitive even in the case of the spline spaces studied in this work. For this reason, we will mostly limit ourselves to simple gradient descent or quasi-Newton methods such as BFGS and L-BFGS [49, 50], which are capable of iteratively constructing an approximate Hessian based only on first-order information, and only require the computation of the gradient of the cost function.

Finally, it has been shown that the inversion problem is in general very ill-conditioned, even more so as the number of degrees of freedom increases (see, e.g., [9]). Incorporating *a priori* information over the kind of physical model that one seeks can therefore drastically improve the convergence rate of the inverse problem, and guide it towards an acceptable minimum. For example, one could promote sparsity in the gradient of m (i.e., promote piecewise-constant models) by adding to $J(m)$ a regularization term in the form of the *total variation* norm [51]

$$\|m\|_{L^1} \coloneqq \int_\Omega |\nabla m(x)| \, \mathrm{d}x,$$

see, e.g., [52]. One could also look for solution in the form of sparse (Dirac-delta) scatterers. For this goal, one might rely on a penalization term based on the *measure norm* [53]

$$\|m\|_{\mathcal{M}} \coloneqq \sup_{\varphi \in C^0(\Omega)} \int_\Omega \varphi \, \mathrm{d}m : |\varphi(x)| < 1 \text{ for all } x,$$

see, e.g., [54] for an application in this sense. In the literature, many different kinds of penalization and regularization terms have been introduced in order to make the minimization quicker or more resilient to noise (see, e.g., [55, 56]). All these norms are convex, but often non-differentiable, and sufficiently different from one another to require dedicated tools for their theoretical and practical manipulation.

Great variety is also seen in the choice of the degrees of freedom over which the minimization is performed. In practical applications, one often discretizes the model space in some way, often by projecting it over some finite space of functions defined over a mesh or a discrete set of measures (which can include Dirac deltas to simulate pointlike scatterers). The degrees of freedom allowed during seismic inversion are not always limited to the coefficients of this expansion, but can include the mesh shape (see, e.g., [57, 58]) and other aspects of the basis functions. This will also be the case in the approach that we suggest at the end of this manuscript. The common feature of these choices is that the physical model can be assumed to live in an appropriate Banach space, as is the case for example of the general measures ρ, λ, that we have used in the previous section.

For all these reasons, we deem it necessary to employ general tools, that do not require differentiability and are capable to handle general convex optimization problems, in our approach to FWI. We therefore introduce in the next section some general tools from convex analysis, showing how they can be successfully employed in seismic inversion. We also refer the reader to, e.g., [54, 59] for a few examples of application of this powerful duality outside the context of seismic inversion.

1.3.1 Fenchel-Rockafellar duality and the adjoint state method

Let us now introduce the powerful Fenchel-Rockafellar duality (see, e.g., [60, Chapter 15]), a fundamental duality theorem in convex theory. Let $\mathcal{X}$ be a Banach space with dual $\mathcal{X}^\star$, and let $\langle u, x\rangle$ denote the pairing between $u \in \mathcal{X}^\star$ and $x \in \mathcal{X}$ (i.e., the linear action of u on x). For every function $f : \mathcal{X} \mapsto \mathbb{R} \sqcup \{+\infty\}$, define its *convex dual* $f^* : \mathcal{X}^\star \mapsto \mathbb{R} \sqcup \{+\infty\}$ (see, e.g., [61]) as

$$f^*(u) \coloneqq \sup_{x\in\mathcal{X}} \left(\langle u, x\rangle - f(x)\right), \tag{1.23}$$

which is always convex even when f is not. Notice that in general $(f+g)^* \neq f^* + g^*$. By the definition of convex dual (1.23), for every $x \in \mathcal{X}$ and $u \in \mathcal{X}^\star$

$$f(x) + f^*(u) \geq \langle u, x\rangle, \tag{1.24}$$

which is known as the Fenchel-Young inequality. Suppose that, for a given $x \in \mathcal{X}$, (1.24) holds with an equality, i.e., that we can find a linear functional $\overline{u}$ such that

$$f(x) + f^*(\overline{u}) = \langle \overline{u}, x\rangle. \tag{1.25}$$

Then, for all $x' \in \mathcal{X}$,

$$\left\langle \overline{u}, x'\right\rangle - \langle \overline{u}, x\rangle \leq f(x') - f(x), \tag{1.26}$$

i.e., the affine functional $x' \mapsto (f(x) - \langle \overline{u}, x \rangle) + \langle \overline{u}, x' \rangle$ is always lower than f. The set of such linear functionals is called the *subdifferential* of f at x, and it is denoted by $\partial f(x)$,

$$\partial f(x) \coloneqq \{u \in \mathcal{X}^\star : \langle u, x' - x \rangle \leq f(x') - f(x) \text{ for all } x' \in \mathcal{X}\}.$$

For example, if $f(x) = 1/2x^2 + |x - 1|$ is a real function of $x \in \mathbb{R}$, then $\partial f(0)$ contains only one element, $u(x) = -x$, but $\partial f(1)$ contains all the functions $u(x) = \gamma x$ for $\gamma \in [0, 2]$. We show this definition in Figure 1.3.

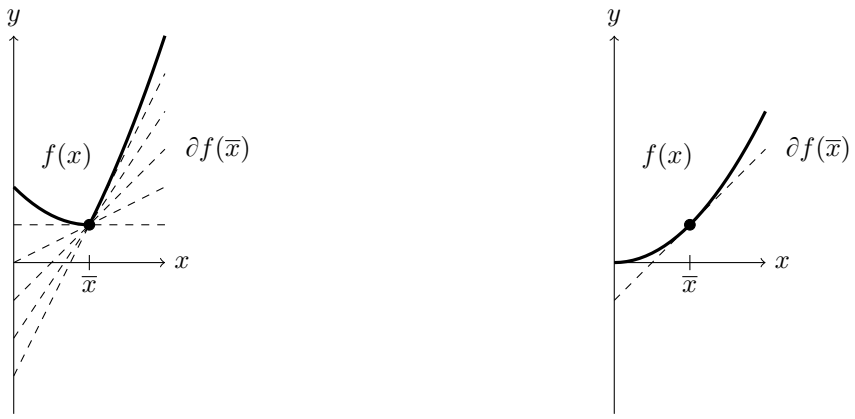


Figure 1.3: (Left) some linear functionals in the subdifferential $\partial f(\overline{x})$ of a non-differentiable function f. (Right) when the function is differentiable, the subdifferential contains only one linear functional.

Thus, we can reformulate the Fenchel-Young relation (1.25) equivalently as

$$\overline{u} \in \partial f(x).$$

Recall that a convex function f is Fréchet-differentiable at x if and only if $\overline{u}$, as defined in (1.26), is unique. Therefore, in this case, $\overline{u}$ corresponds to the gradient of f,

$$\overline{u} = \nabla f(x). \tag{1.27}$$

In applications, one often discretizes the space $\mathcal{X}$ by introducing a basis, thus reducing it to a problem on $\mathbb{R}^N$ for some N. For example, $\mathcal{X}$ could be the set of piecewise-constant or piecewise-linear functions over a mesh, and therefore N is the size of the corresponding basis of functions. In this case, if f is differentiable, then (1.27) is simply a manifestation of the usual Lagrange duality. However, the notion of subdifferential is more general and valid for any Banach space, even large ones such as the space of measures on which we formulate our seismic inversion problem, and even when f is not differentiable. For this reason, the concept of subdifferential is very relevant in optimization, and can be used to define a local descent direction that is guaranteed to reduce f even when f is not smooth. In particular, equation (1.27) is very useful for two distinct reasons: when read from left to right, it allows to recover the dual variable from the primal one, and when read from right to left, it allows to compute the gradient of the primal problem if the dual variable is known.

We are ready to state the important Fenchel-Rockafellar duality theorem [61, 62]. Let $\mathcal{X}$ and $\mathcal{Y}$ be two Banach spaces with respective duals $\mathcal{X}^\star$ and $\mathcal{Y}^\star$, and let $A : \mathcal{X} \mapsto \mathcal{Y}$ be a bounded linear operator with adjoint $A^\star : \mathcal{Y}^\star \mapsto \mathcal{X}^\star$. Then, if f and g are two convex functions with domain $\mathcal{X}$ and $\mathcal{Y}$, respectively, and image in $\mathbb{R} \sqcup \{+\infty\}$, the convex duality can be expressed as follows.

Theorem 1.3.1 (Fenchel-Rockafellar)**.** *The following weak duality holds:*

$$\inf_{x\in\mathcal{X}} \left(f(x) + g(Ax)\right) \geq \sup_{u\in\mathcal{Y}^\star} \left(-f^*(A^\star u) - g^*(-u)\right). \tag{1.28}$$

Furthermore, if there are two elements $\overline{x} \in \mathcal{X}$ and $\overline{u} \in \mathcal{Y}^\star$ satisfying the optimality conditions

$$A^\star \overline{u} \in \partial f(\overline{x}), \tag{1.29a}$$
$$-\overline{u} \in \partial g(A\overline{x}), \tag{1.29b}$$

then (1.28) *holds with an equality (strong duality), and $\overline{x}$, $\overline{u}$ are optimal arguments of the primal and dual problem respectively.*

Proof. From the definition of inf, sup, convex dual (1.23) and the Fenchel-Young inequality (1.24),

$$\begin{aligned}
\inf_{x\in\mathcal{X}} \left(f(x) + g(Ax)\right) &\geq \inf_{x\in\mathcal{X}} \left(f(x) + \sup_{u\in\mathcal{Y}^\star} \left(\langle -u, Ax\rangle - g^*(-u)\right)\right), \\
&\geq \sup_{u\in\mathcal{Y}^\star} \left(-g^*(-u) + \inf_{x\in\mathcal{X}} \left(f(x) - \langle u, Ax\rangle\right)\right), \\
&= \sup_{u\in\mathcal{Y}^\star} \left(-g^*(-u) + \inf_{x\in\mathcal{X}} \left(f(x) - \langle A^\star u, x\rangle\right)\right), \\
&= \sup_{u\in\mathcal{Y}^\star} \left(-g^*(-u) - \sup_{x\in\mathcal{X}} \left(\langle A^\star u, x\rangle - f(x)\right)\right), \\
&= \sup_{u\in\mathcal{Y}^\star} \left(-g^*(-u) - f^*(A^\star u)\right).
\end{aligned}$$

Finally, if there exist $\overline{x} \in \mathcal{X}$ and $\overline{u} \in \mathcal{Y}^\star$ satisfying (1.29), then by (1.25) we have

$$\begin{aligned}
f(\overline{x}) + f^*(A^\star \overline{u}) &= \langle A^\star \overline{u}, \overline{x}\rangle, \\
g(A\overline{x}) + g^*(-\overline{u}) &= -\langle \overline{u}, A\overline{x}\rangle,
\end{aligned}$$

and summing these two expressions yields $f(\overline{x}) + g(A\overline{x}) + f^*(A^\star\overline{u}) + g^*(-\overline{u}) = 0$, as required. □

Formulating the dual of an optimization problem has many advantages. First of all, the dual functions f^* and g^* and their respective dual spaces can be easier to describe. For example, the dual of the L^p norm is the L^q norm with $\frac{1}{p} + \frac{1}{q} = 1$, with the interesting special case $p = 1$, $q = \infty$, and the dual of a measure space is simply a space of continuous functions. Thus, regularization terms often become simpler to incorporate. Moreover, the dual problem is sometimes better conditioned, which is an interesting feature in applications, especially since the measured data is noisy (and thus potentially contradictory) and generally insufficient to recover the physical

model uniquely. We illustrate this point later in this chapter with a simple problem. Finally, the Fenchel-Rockafellar duality is relevant to seismic inversion since it allows to naturally derive the *adjoint state method* for the calculation of the gradient of a cost function, even when this function is not differentiable, as we do presently.

Suppose that we are given a Banach space $\mathcal{M}$ of *physical models*, a Banach space $\mathcal{X}$ of *solutions*, and that we wish to minimize some cost function $f(x)$ over $x \in \mathcal{X}$ subject to the constraint $A(m)x = s$ for some bounded linear operator $A(m)$ dependent on $m \in \mathcal{M}$. Here, $s \in \mathcal{Y}$, where $\mathcal{Y}$ is another Banach space, namely the space of *constraints* or, in the case of seismic inversion, the space of *sources*. Suppose that we want to compute the gradient of the constrained cost function $f(m) \coloneqq f(x(m))$, where $x(m)$ is chosen to be a solution to the constraint. The adjoint state technique can be employed to compute this gradient efficiently, and it is a very important tool in practical applications, virtually omnipresent in the seismic inversion literature. Intuitively, one can express, for some notion of differentiability δ (e.g., Fréchet differentiability),

$$\frac{\delta f}{\delta m} = \left\langle \frac{\delta f}{\delta x}, \frac{\delta x}{\delta m} \right\rangle,$$

and after finding a vector u such that $A^\star(m)u = \delta f/\delta x$, one can use the product rule of differentiation

$$\frac{\delta}{\delta m} A(m)x(m) = \frac{\delta A(m)}{\delta m} x(m) + A(m)\frac{\delta x(m)}{\delta m}$$

to transform the gradient into the form

$$\frac{\delta f}{\delta m} = \left\langle -u, \frac{\delta A(m)}{\delta m} x \right\rangle.$$

The element u is called the *adjoint state*. The main computational advantage of this formulation is that the derivatives $\delta A(m)/\delta m$ are much easier to compute than $\delta x/\delta m$, which requires the knowledge of the whole forward propagation kernel. However, this derivation relies on the Fréchet differentiablity of $f(x(m))$ and the other quantities in play, as well as a model space that is well-behaved enough. We show presently that the adjoint state method is a simple consequence of the Fenchel-Rockafellar duality, and therefore it has in fact a wider applicability.

Let us express the constraint as a minimization of the form

$$x \in \operatorname*{arg\,min}_{w \in \mathcal{X}} g(A(m)w) \tag{1.30}$$

where $A(m) : \mathcal{X} \to \mathcal{Y}$ is a bounded linear operator for every $m \in M \subset \mathcal{M}$ and g is the convex function defined as

$$g(u) = \begin{cases} 0 & \text{if } u = s, \\ +\infty & \text{otherwise.} \end{cases} \tag{1.31}$$

Then, (1.30) is equivalent to the condition $A(m)x = s$. One can then formulate a constrained optimization problem by minimizing the cost function $f(x) + g(A(m)x)$ over both $m \in \mathcal{M}$ and $x \in \mathcal{X}$. The Fenchel-Rockafellar duality then yields directly the adjoint state method.

First, we will need the following product rule for subdifferentials.

Lemma 1.3.2. *Let $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{M}$ be Banach spaces, $A(m)$ be a family of bounded linear operators from $\mathcal{X}$ to $\mathcal{Y}$ and $x(m)$ be a family of elements in $\mathcal{X}$, both parameterized by $m \in \mathcal{M}$ and both Lipschitz-continuous in their argument at $\overline{m}$. Let $f : \mathcal{Y} \mapsto \mathbb{R} \sqcup \{+\infty\}$ be a strictly differentiable function. Then*

$$\partial(f(A(m)x(m)))(\overline{m}) = \partial(f(A(\overline{m})x(m)) + f(A(m)x(\overline{m})))(\overline{m}).$$

Proof. Very similar to Theorem 5.3 in [63]. Specifically, it follows from [63, Theorem 4.2] by taking $X \coloneqq \mathcal{M}$, $Y \coloneqq B(\mathcal{X}, \mathcal{Y}) \times \mathcal{X}$, $\Phi(m) \coloneqq (A(m), x(m))$ and $\varphi(A(m), x(m)) \coloneqq f(A(m)x(m))$. □

We are now ready to introduce the ajoint state method.

Theorem 1.3.3 (Adjoint state method for subdifferentials)**.** *Let f be convex, g be defined by* (1.31)*, and let $A(m)$ be a family of bounded linear operators from $\mathcal{X}$ to $\mathcal{Y}$. Define the cost function*

$$J(m) \coloneqq \inf_{x \in \mathcal{X}} \left(f(x) + g(A(m)x) \right),$$

and let $\overline{m} \in \mathcal{M}$, $\overline{x} \in \mathcal{X}$ and $\overline{u} \in \mathcal{Y}$ be optimal values in the sense that the conditions (1.29) *are satisfied for $A \coloneqq A(\overline{m})$. Assume that $A(m)$ is Lipschitz-continuous at $\overline{m}$ with a Lipschitz-continuous right inverse. Then, the function*

$$h(m) \coloneqq \langle -\overline{u}, A(m)\overline{x} \rangle . \tag{1.32}$$

satisfies

$$\partial J(\overline{m}) \supseteq \partial h(\overline{m}).$$

Proof. Applying the Fenchel-Rockafellar duality, we have, for all $x \in \mathcal{X}$ and $m \in \mathcal{M}$,

$$\begin{aligned} f(x) + g(A(m)x) &\geq f(x) + g(A(\overline{m})\overline{x}) && \text{by (1.31),} \\ &\geq \langle A(\overline{m})^\star \overline{u}, x \rangle - f^*(A(\overline{m})^\star \overline{u}) + \langle -\overline{u}, A(\overline{m})\overline{x} \rangle - g^*(-\overline{u}) && \text{by (1.29) and (1.25),} \\ &= \langle \overline{u}, A(\overline{m})(x - \overline{x}) \rangle + J(\overline{m}) && \text{by (1.28).} \end{aligned}$$

Let us take the infimum on x on the left side, yielding $J(m)$. The infimum must be taken over all $x(m)$ satisfying the condition $g(A(m)x(m)) < +\infty$, i.e., $A(m)x(m) = s$. This can be done Lipschitz-continuously in m by hypothesis. We obtain

$$\langle \overline{u}, A(\overline{m})(x(m) - \overline{x}) \rangle \leq J(m) - J(\overline{m}). \tag{1.33}$$

We can now apply Lemma 1.3.2 to transform the left hand side. Specifically,

$$\partial \langle \overline{u}, A(\overline{m})x(m) + A(m)\overline{x} \rangle = \partial \langle \overline{u}, A(m)x(m) \rangle = \{0\},$$

since $A(m)x(m)$ is constant. This implies

$$0 \leq \langle \overline{u}, A(\overline{m})x(m)\rangle + \langle \overline{u}, A(m)\overline{x}\rangle - 2\langle \overline{u}, A(\overline{m})\overline{x}\rangle . \tag{1.34}$$

Summing (1.33) and (1.34) finally yields

$$J(m) - J(\overline{m}) \geq \langle \overline{u}, (A(\overline{m}) - A(m))\overline{x}\rangle = \langle -\overline{u}, (A(m) - A(\overline{m}))\overline{x}\rangle .$$

Thus, $\delta \in \partial \langle -\overline{u}, A(m)\overline{x}\rangle$ implies $\delta \in \partial J(m)$, completing the proof. □

Theorem 1.3.3 illustrates the computational advantages of the adjoint state method. Specifically, it states that it is possible to find a subgradient of the constrained cost function at $\overline{m}$ (and thus a viable descent direction) by computing a subgradient of the much simpler function $h(m) \coloneqq \langle -\overline{u}, A(m)\overline{x}\rangle$, after solving the primal and dual problems,

$$\overline{x} \in \operatorname*{arg\,min}_{A(m)x=s} f(x), \tag{1.35a}$$

$$A^\star \overline{u} \in \partial f(\overline{x}). \tag{1.35b}$$

These equations correspond to the *forward* and *adjoint* problems in the seismic inversion literature. The function $h(m)$ depends on m only through the operator $A(m)$, and thus its subdifferential is much easier to compute. Notice that no assumption is made on the differentiability of the constrained cost function, only that $A(m)$ is Lipschitz-continuous with Lipschitz-continuous right inverse.

We exemplify this optimization scheme in the next section by reformulating the simple linear regression problem similarly to the classical adjoint-state approach, before moving to seismic inversion in the following section.

1.3.2 Simple linear regression via F.-R. duality and the adjoint state method

In this section, we delve into the details of the Fenchel-Rockafellar duality and the adjoint state method in a somewhat simplified setting, by tackling a well-known and intuitive problem, the simple linear regression problem. We will however reshape this problem into a form that is very similar to seismic inversion, and we will draw the necessary parallels between the two, so that most of the results can be transported to the seismic context without too much effort.

Suppose that we wish to find the best affine function $y(x) = \alpha x + \beta$ that fits a set of data points $(x_i, d_i) \in \mathbb{R}^2$, $i = 1, \ldots, n$ in the least-squares sense, i.e., that minimizes the function

$$f(y) \coloneqq \sum_{i=1}^{n} \frac{1}{2} |y(x_i) - d_i|^2 . \tag{1.36}$$

Let $I \subset \mathbb{R}$ be a finite interval containing all the coordinates x_i. Then this problem is equivalent to finding a real parameter α and a differentiable function $y : I \mapsto \mathbb{R}$ minimizing (1.36) and satisfying the differential equation $y' = \alpha$.

Let us use the adjoint state method, in the form of Theorem 1.3.3, to minimize J by (sub)gradient descent. Specifically, let us reformulate the constraint on y by defining, for $\alpha \neq 0$,

$$A(\alpha) \coloneqq \frac{1}{\alpha}\frac{\mathrm{d}}{\mathrm{d}x},$$

and using the convex function

$$g(w) \coloneqq \begin{cases} 0 & \text{if } w = 1, \\ +\infty & \text{otherwise.} \end{cases}$$

Suppose now that we are given an initial guess $\overline{\alpha}$ for the physical parameter α, and that we wish to find the derivative of the cost function $J(\alpha) \coloneqq \inf_{x\in\mathcal{X}} (f(x) + g(A(\alpha)x))$ with respect to α at the point $\alpha = \overline{\alpha}$. The solution to the primal problem (1.35a)

$$\overline{y} \in \operatorname*{arg\,min}_{C\in\mathbb{R}} f(\overline{\alpha}x + C)$$

is, as can be easily derived,

$$\overline{y}(x) = \overline{\alpha}(x - \langle x\rangle) + \langle d\rangle, \tag{1.37}$$

where we have used the average of a variable $\langle z\rangle \coloneqq \sum_{i=1}^n z_i/n$. The source of the adjoint problem $\overline{v} \coloneqq A(\overline{\alpha})^\star \overline{u}$ (1.35b) can be obtained using Fenchel-Young's relation (1.25), i.e.,

$$\overline{v} \in \partial f(\overline{y}) \Leftrightarrow f(\overline{y}) + f^*(\overline{v}) = \langle \overline{v}, \overline{y}\rangle\,, \tag{1.38}$$

We can compute $f^*(v)$ using (1.23). We have

$$f^*(v) \coloneqq \sup_{y\in C^1(I)} \left(\langle v, y\rangle - \sum_{i=1}^n \frac{1}{2}\left|y(x_i) - d_i\right|^2 \right).$$

This expression is clearly unbounded in y unless v is supported only on the points x_i, i.e., $\langle v, y\rangle = \sum_{i=1}^n \varepsilon_i y(x_i)$ for some values ε_i. One obtains then

$$f^*(v) = \begin{cases} \displaystyle\sum_{i=1}^n \frac{1}{2}\left(|\varepsilon_i + d_i|^2 - |d_i|^2\right) & \text{if } \langle v, y\rangle = \displaystyle\sum_{i=1}^n \varepsilon_i y(x_i) \\ +\infty \text{ otherwise.} \end{cases} \tag{1.39}$$

Using now the definitions of f (1.36) and f^* (1.39), we conclude that the functional $\overline{v}$ must have the form

$$\langle \overline{v}, y\rangle = \sum_{i=1}^n \overline{\varepsilon}_i y(x_i),$$

with, due to (1.38),

$$\overline{\varepsilon}_i = \overline{y}(x_i) - d_i.$$

Notice that the source of the adjoint problem $\overline{v}$ is supported only at the receiver locations x_i, and its amplitude is the difference between the forward solution $\overline{y}(x_i)$ and the measured data

d_i. This characterizes also the seismic inversion problem, as we shall see in the next section.

Since $J(\alpha)$ is differentiable, its subdifferential $\partial J(\overline{\alpha})$ only contains its gradient $\partial J/\partial\alpha|_{\alpha=\overline{\alpha}}$, and it can be computed by finding $\partial h(\overline{\alpha})$, where $h(\alpha)$ is given using (1.32) as

$$h(\alpha) \coloneqq \langle -\overline{u}, A(\alpha)\overline{y}\rangle = \left\langle -\overline{u}, \frac{\overline{\alpha}}{\alpha}A(\overline{\alpha})(\overline{\alpha}x + C)\right\rangle = -\frac{\overline{\alpha}}{\alpha}\langle \overline{v}, \overline{\alpha}x + C\rangle = -\frac{\overline{\alpha}^2}{\alpha}\sum_{i=1}^{n}(\overline{y}(x_i) - d_i)\, x_i.$$

Notice that the choice of $C \in \mathbb{R}$ is irrelevant, since $\sum_{i=1}^{n}(\overline{y}(x_i) - d_i) = 0$. Thus,

$$\left.\frac{\partial J}{\partial \alpha}\right|_{\alpha=\overline{\alpha}} = \left.\frac{\partial h}{\partial \alpha}\right|_{\alpha=\overline{\alpha}} = \sum_{i=1}^{n}(\overline{y}(x_i) - d_i)\, x_i,$$

with $\overline{y}(x)$ given by (1.37). One can easily check that this is indeed the gradient of the cost function with respect to α by plugging directly the best-fitting line with slope α, namely $y = \alpha(x - \langle x\rangle) + \langle d\rangle$, into (1.36) and differentiating.

Although very convoluted, this formulation of the optimal solution of the simple linear regression problem has the advantage of being very similar to the seismic inversion problem, sharing most of its major features, while being set in a simpler and much more intuitive context. After going through all the details in its derivation, we are ready to transfer these results to full waveform inversion.

1.3.3 Full waveform inversion

We consider first, for simplicity, the case of full waveform inversion with a single shot and a single receiver. In this problem, we are given a source term $s(x,t)$, for example a point-like Ricker wavelet as in (1.6), and a receiver position x_r with measured data $D_r(t)$, either only pressure data, i.e., $D_r(t) = (d_r(t), 0)$, or pressure and speed data. We will focus here on pressure data only, for the sake of simplicity. In order to avoid confusion, we will make the dependence of the acoustic wave operator (1.16) on the physical parameters explicit, by writing $A_{\rho,\lambda}$ instead of A.

The FWI problem is formally very similar to the simple linear regression problem seen in the previous section, since the solution can be expressed as the minimization of the function

$$f(P) \coloneqq \frac{1}{2}\int_0^{t_f} |p(x_r,t) - d_r(t)|^2 \,\mathrm{d}t, \tag{1.40}$$

where $p \coloneqq P_1$, under the constraint that $P(x,t) \coloneqq (P_1, P_2)$ is a solution to the acoustic problem (1.15), with its associated initial and boundary conditions, for some physical parameters ρ, $\lambda > 0$.

Suppose that we seek to minimize (1.40) using the adjoint state method, starting from an initial guess for the physical model $\overline{\rho}$, $\overline{\lambda}$. This problem can be recast as the minimization of

$$J(\rho,\lambda) \coloneqq \inf_{P\in\mathcal{X}}\left(f(P) + g((\partial_t - A_{\rho\lambda})P)\right),$$

where $g(W) = 0$ if $W = S$ and $+\infty$ otherwise.

According to (1.35), we need to compute the subdifferential of f and the adjoint operator to $\partial/\partial t - A_{\rho,\lambda}$.

Let us first compute $f^*(V)$ for a linear operator $U \in \mathcal{X}^\star$ from its definition (1.23), as done in the previous section. We have

$$f^*(V) \coloneqq \sup_{P\in\mathcal{X}} \left(\langle V, P\rangle - \frac{1}{2}\int_0^{t_f} |p(x_r,t) - d_r(t)|^2 \,\mathrm{d}t \right).$$

Once again, f^* is infinite unless V satisfies $V_2 = 0$ and V_1 is supported only at x_r. Thus,

$$f^*(V) = \begin{cases} \dfrac{1}{2}\displaystyle\int_0^{t_f} |\varepsilon(t) + d_r(t)|^2 - |d_r(t)|^2 \,\mathrm{d}t & \text{if } V_2 = 0 \text{ and } \langle V_1, p\rangle = \int_0^{t_f} \varepsilon(t)p(x_r,t)\,\mathrm{d}t \\ & \text{for some function } \varepsilon(t), \\ +\infty & \text{otherwise.} \end{cases} \tag{1.41}$$

By the Fenchel-Young relation (1.25), the dual source U is in the subdifferential of f if and only if $f(P) + f^*(V) = \langle V, P\rangle$. Using (1.40) and (1.41), one sees right away that V must satisfy $V_2 = 0$ and

$$v \coloneqq V_1 = \delta(x - x_r)(p(x_r,t) - d_r(t)).$$

Moving to the adjoint operator, let $\sigma \coloneqq \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, and let $U(x,t)$ be a solution to the adjoint problem (1.20) with source $-\sigma V$, i.e., since $\sigma = \sigma^{-1}$,

$$-\sigma\left(\frac{\partial}{\partial t} - A^\star_{\rho,\lambda}\right) U = V,$$

with the initial conditions replaced by the *final conditions*

$$U(x,t_f) = 0, \text{ that is, } u_f(x,t_f) = \partial_t u(x,t_f) = 0. \tag{1.42}$$

Notice that, as a differential operator, $A^\star_{\rho,\lambda} = -A_{\rho,\lambda}$, and therefore U represents a wavefield that propagates *backwards* in time from t_f to 0. Notice also that (1.42) implies

$$0 = \int_0^{t_f} \frac{\partial}{\partial t}(U_i P_j)\,\mathrm{d}t = \int_0^{t_f} \frac{\partial P_i}{\partial t} U_j \,\mathrm{d}t + \int_0^{t_f} P_i \frac{\partial U_j}{\partial t}\,\mathrm{d}t \tag{1.43}$$

for $i, j = 1, 2$. Therefore,

$$\begin{aligned}
\left\langle -\sigma A^\star_{\rho,\lambda} U, P\right\rangle &= \int_0^{t_f}\!\!\int_\Omega \lambda U_2 P_2 \,\mathrm{d}\Omega\,\mathrm{d}t + \int_0^{t_f}\!\!\int_\Omega \nabla\cdot\left(\frac{1}{\rho}\nabla U_1\right) P_1 \,\mathrm{d}\Omega\,\mathrm{d}t, \\
&= \int_0^{t_f}\!\!\int_\Omega \lambda U_2 P_2 \,\mathrm{d}\Omega\,\mathrm{d}t - \int_0^{t_f}\!\!\int_\Omega \frac{1}{\rho}\nabla U_1\cdot\nabla P_1 \,\mathrm{d}\Omega\,\mathrm{d}t + \int_0^{t_f}\!\!\int_{\Omega_A} \frac{1}{\rho} P_1 \frac{\partial U_1}{\partial n}\,\mathrm{d}S\,\mathrm{d}t, \\
&= \int_0^{t_f}\!\!\int_\Omega \lambda U_2 P_2 \,\mathrm{d}\Omega\,\mathrm{d}t - \int_0^{t_f}\!\!\int_\Omega \frac{1}{\rho}\nabla U_1\cdot\nabla P_1 \,\mathrm{d}\Omega\,\mathrm{d}t + \int_0^{t_f}\!\!\int_{\Omega_A} c P_1 U_2 \,\mathrm{d}S\,\mathrm{d}t, \\
&= \int_0^{t_f}\!\!\int_\Omega \lambda U_2 P_2 \,\mathrm{d}\Omega\,\mathrm{d}t - \int_0^{t_f}\!\!\int_\Omega \frac{1}{\rho}\nabla U_1\cdot\nabla P_1 \,\mathrm{d}\Omega\,\mathrm{d}t - \int_0^{t_f}\!\!\int_{\Omega_A} c P_2 U_1 \,\mathrm{d}S\,\mathrm{d}t,
\end{aligned}$$

$$
\begin{aligned}
&= \int_0^{t_f}\!\!\int_\Omega \lambda U_2 P_2 \,\mathrm{d}\Omega\,\mathrm{d}t - \int_0^{t_f}\!\!\int_\Omega \frac{1}{\rho}\nabla U_1 \cdot \nabla P_1 \,\mathrm{d}\Omega\,\mathrm{d}t + \int_0^{t_f}\!\!\int_{\Omega_A} \frac{1}{\rho}\frac{\partial P_1}{\partial n} U_2 \,\mathrm{d}S\,\mathrm{d}t,\\
&= \int_0^{t_f}\!\!\int_\Omega \lambda U_2 P_2 \,\mathrm{d}\Omega\,\mathrm{d}t + \int_0^{t_f}\!\!\int_\Omega U_1 \nabla\cdot\left(\frac{1}{\rho}\nabla P_1\right)\,\mathrm{d}\Omega\,\mathrm{d}t,\\
&= \langle U, \sigma A_{\rho,\lambda} P\rangle\,,
\end{aligned}
$$

where we have used the adjoint absorbing boundary conditions (1.21) on U and the absorbing boundary conditions (1.9) on P in the second and fourth step, respectively, and (1.43) in the third step. Similarly,

$$
\left\langle \sigma\frac{\partial}{\partial t}U, P\right\rangle = \left\langle U, -\sigma\frac{\partial}{\partial t}P\right\rangle,
$$

thanks to (1.43). Thus,

$$
\left[\sigma\left(\frac{\partial}{\partial t} - A_{\rho,\lambda}\right)\right]^\star = -\sigma\left(\frac{\partial}{\partial t} - A^\star_{\rho,\lambda}\right).
$$

We now have all the ingredients that go into the implementation of the adjoint state via Theorem 1.3.3, and therefore we only have to explicitly state here the algorithm that needs to be followed:

(i) Compute $\overline{P}$ solution to the forward problem

$$
\left(\frac{\partial}{\partial t} - A_{\rho,\lambda}\right)\overline{P} = S,
$$

from $t = 0$ to $t = t_f$, with initial conditions $P(x, 0) = 0$;

(ii) Compute the source term for the dual problem, namely

$$
\overline{V} = (\overline{p}(x_r, t) - d_r(t))\,\delta(x - x_r); \tag{1.44}
$$

(iii) Solve the adjoint problem

$$
-\sigma\left(\frac{\partial}{\partial t} - A^\star_{\rho,\lambda}\right)\overline{U} = \overline{V},
$$

propagating backwards from $t = t_f$ to $t = 0$, starting with the final conditions $V(x, t_f) = 0$;

(iv) Compute the function

$$
h(\rho, \lambda) \coloneqq \left\langle -\overline{U}, -\sigma A_{\rho,\lambda}\overline{P}\right\rangle. \tag{1.45}
$$

Any element of the subdifferential $\partial h\left(\overline{\rho}, \overline{\lambda}\right)$ is in $\partial J\left(\overline{\rho}, \overline{\lambda}\right)$.

If J is differentiable, this corresponds exactly to the usual adjoint state method as encountered in standard seismic imaging, see, e.g., [45, 64, 65]. For example, if ρ and λ are piecewise-constant functions, it is straightforward to compute directly the derivatives of (1.45) and obtain, after a few simple manipulations, the following well-known expressions:

$$
\begin{aligned}
\frac{\partial J}{\partial \lambda^{-1}} &= \int_0^{t_f}\!\!\int_\Omega \overline{u}\frac{\partial^2 \overline{p}}{\partial t^2}\,\mathrm{d}\Omega\,\mathrm{d}t,\\
\frac{\partial J}{\partial \rho^{-1}} &= \int_0^{t_f}\!\!\int_\Omega \overline{u}\Delta\overline{p}\,\mathrm{d}\Omega\,\mathrm{d}t.
\end{aligned} \tag{1.46}
$$

Notice that the primal and adjoint wavefields propagate in opposite time directions, and thus in order to compute their superposition integral, one would need in principle to store in memory the whole time-dependent wavefields $\overline{P}(x,t)$ and $\overline{U}(x,t)$. In practice, some trade-off between memory use and computational time is usually done, for example subsampling or storing only boundary data, see, e.g., [66–69].

Finally, we return to the full-scale seismic inversion problem with $n_s > 1$ shots and multiple receiver positions per shot. In this case, by looking at the similarities with the linear regression problem of the last section, it should come as no surprise that the adjoint state method for the computation of the subdifferential of $J(m)$ takes a very similar form, except that there have to be n_s forward and n_s adjoint solutions, with source s_i for the i-th forward problem, and adjoint source v_i given by

$$v_i(x,t) \coloneqq \sum_{r \in R_i} \varepsilon_r(t)\delta(x - x_r)$$

for the i-th adjoint problem, leading to n_s scalar products which define the n_s functions $h_i(\rho,\lambda) \coloneqq \langle -\sigma\overline{U}_i, A_{\rho,\lambda}\overline{P}_i \rangle$. The function h is then simply given by the sum

$$h(\rho,\lambda) \coloneqq \sum_{i=1}^{n_s} h_i(\rho,\lambda),$$

and everything else proceeds as discussed above.

1.3.4 Some worked-out examples

We end this chapter by illustrating some of the features of the seismic inversion problem (namely, the computation of the gradient via the adjoint method, non-convexity, the lack of well-posedness, and the lack of differentiablity) with a few simple examples. The setup for these examples is shown in Figure 1.4.

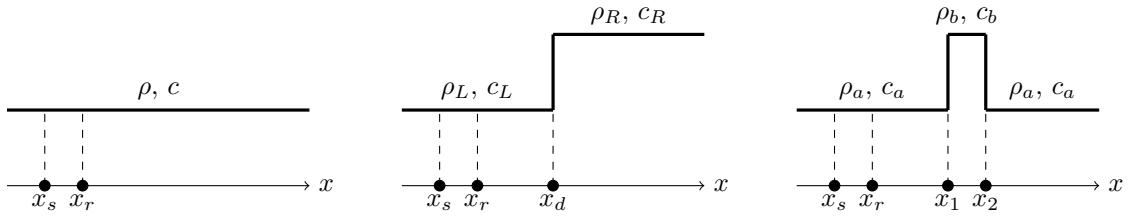


Figure 1.4: Position of the receiver, source and discontinuities for the one-dimensional homogeneous (left), step (middle) and barrier (right) examples.

One-dimensional infinite homogeneous medium

Let us consider first a one-dimensional homogeneous domain, represented by an interval $\Omega \subset \mathbb{R}$, on which we place a source and a receiver at x_s, $x_r \in \Omega$ respectively, with a pressure measurement $d_r(t)$ at x_r. In this case, one can rely on the well known Green's function

$$G(x,t) = \frac{c\rho}{2}\theta(ct - |x|),$$

where as usual $c \coloneqq \sqrt{\lambda/\rho}$, and θ represents the Heaviside theta function $\theta(t) \coloneqq 1$ if $t \geq 0$, and zero otherwise. This expression directly leads to an explicit solution when $s(x,t) \coloneqq s_0\delta(x - x_s)s(t)$ is a point-like source with amplitude s_0 and $s(t \leq 0) = 0$, namely

$$\overline{p}(x,t) = \frac{s_0\rho c}{2}\int_0^t \theta\left(c(t-\tau) - |x - x_s|\right) s(\tau)\,\mathrm{d}\tau.$$

One can verify this statement by explicitly computing

$$\frac{1}{\lambda}\frac{\partial^2 p}{\partial t^2}(x,t) = \frac{s_0}{2c}s'\left(t - \frac{|x_r - x_s|}{c}\right)$$

and

$$\frac{1}{\rho}\frac{\partial^2 p}{\partial x^2}(x,t) = \frac{s_0}{2c}s'\left(t - \frac{|x_r - x_s|}{c}\right) + s_0\delta(x - x_r)s(t).$$

The forward solution $p(x,t)$, evaluated at $x = x_r$, is given simply by

$$\overline{p}(x_r,t) = \frac{s_0 c\rho}{2}S\left(t - \frac{|x_r - x_s|}{c}\right), \tag{1.47}$$

where $S(t)$ is a primitive of $s(t)$ with $S(t \leq 0) = 0$. This is the expected result for a non-dispersive medium. Our problem is therefore analogous to finding a shift and a multiplicative coefficient such that the correlation with $d_r(t)$ is maximized.

Let us use the adjoint state method to perform this optimization. As per (1.44), the adjoint problem must be solved with the adjoint source

$$\overline{v}(x,t) = \delta(x - x_r)\left(\frac{s_0 c\rho}{2}S\left(t - \frac{|x_r - x_s|}{c}\right) - d_r(t)\right).$$

Recall that the solution must be found using the adjoint operator, and starting from the final conditions $u(x,t_f) = \partial_t u(x,t_f) = 0$ and propagating backwards. Thus,

$$\overline{u}(x,t) = \frac{s_0\rho c}{2}\left(\frac{s_0\rho c}{2}S_2\left(t - \frac{|x_r - x_s|}{c} + \frac{|x - x_r|}{c}\right) - D_r\left(t + \frac{|x - x_r|}{c}\right)\right),$$

where $D_r(t)$ is the primitive $\int_{t_f}^t (d_r(\tau))\,\mathrm{d}\tau$ and S_2 is a primitive of S. Using (1.46), integrating by parts to transfer the derivatives from a function to another, and remembering that s, d and all their derivatives and primitives are zero outside $[0, t_f]$,

$$\frac{\partial J}{\partial \lambda^{-1}} = \frac{s_0^2\rho\lambda}{4}\left(-\frac{s_0\rho c}{2}\int_0^{t_f}\int_{x_s}^{x_r} S\left(t - \frac{|x_r - x_s|}{c} + \frac{|x - x_r|}{c}\right) s\left(t - \frac{|x - x_s|}{c}\right)\mathrm{d}t \right.$$
$$\left. + \int_0^{t_f}\int_{x_s}^{x_r} s\left(t - \frac{|x - x_s|}{c}\right) d_r\left(t + \frac{|x - x_r|}{c}\right)\mathrm{d}t\right).$$

Let us choose, without loss of generality, $x_r > x_s$. Then $-|x_r - x_s| + |x - x_r| = -(x - x_s)$ and $|x - x_r| = x_r - x = -(x - x_s) + (x_r - x_s)$. Making these substitutions and shifting the

integration variable $t \to t - (x_r - x_s)/c$ in the second integral, we find

$$\frac{\partial J}{\partial \lambda^{-1}} = \frac{s_0^2 \rho \lambda}{4} \left(-\frac{s_0 \rho c}{2} \int_0^{t_f} \int_{x_s}^{x_r} S\left(t - \frac{x - x_s}{c}\right) s\left(t - \frac{x - x_s}{c}\right) \mathrm{d}t \right.$$
$$\left. + \int_0^{t_f} \int_{x_s}^{x_r} s\left(t - \frac{x - x_s}{c} - \frac{x_r - x_s}{c}\right) d_r\left(t - \frac{x - x_s}{c}\right) \mathrm{d}t \right),$$

and, after realizing that the variable x is irrelevant and can be integrated away to obtain a simple factor $(x_r - x_s)$, substituting $c = \sqrt{\lambda/c}$, we obtain the final result

$$\frac{\partial J}{\partial \lambda^{-1}} = \frac{s_0^2 \rho \lambda}{4} (x_r - x_s) \left(-\frac{s_0 \sqrt{\rho \lambda}}{2} \int_0^{t_f} S(t) s(t) \, \mathrm{d}t + \int_0^{t_f} s\left(t - \frac{x_r - x_s}{c}\right) d_r(t) \, \mathrm{d}t \right). \tag{1.48}$$

One can check that (1.48) is the correct form by plugging (1.47) directly into J and differentiating. A similar calculation, that we omit, can be performed for $\partial J / \partial \rho^{-1}$.

Even in this elementary case, the seismic inversion problem is nonconvex. For example, if the functions appearing in (1.48) are peaked around a given frequency ω_0, like the Ricker wavelet (1.6), and the receiver datum is perfect, i.e. $d_r(t) = S(t - (x_r - x_s)/c_0)$ for some given velocity c_0, then

$$\int_{\mathbb{R}} s\left(t - \frac{x_r - x_s}{c}\right) d_r\left(t - \frac{x_r - x_s}{c_0}\right) \mathrm{d}t = \int_{\mathbb{R}} \overline{\widetilde{s}}(\omega) \widetilde{d}_r(\omega) e^{\mathrm{i}\omega(\alpha - \alpha_0)} \, \mathrm{d}\omega \approx \overline{\widetilde{s}}(\omega_0) \widetilde{d}_r(\omega_0) e^{\mathrm{i}\omega_0(\alpha - \alpha_0)},$$

with $\alpha = (x_r - x_s)/c$, $\alpha_0 = (x_r - x_s)/c_0$ ad $\widetilde{\cdot}$ denotes the Fourier transform. Thus, the gradient vanishes for velocities c such that $\omega_0(\alpha - \alpha_0) \in 2\pi\mathbb{Z}$, which represent local minima in the cost function. The local minimization will then possibly mismatch the cycles of the wave compared with the real underlying model, a phenomenon known as *cycle skipping*.

This issue is much more evident in real-world seismic problems, where wave packets reflected by nearby layers can have similar shapes, and thus lead in ambiguities in their attribution. For example in Figure 1.1, one might be tempted to assign the first pulse of the receiver R_5 to the wave reflected by the second layer instead of the first. The local minima that appear in this case are much harder to treat, and are at the core of the problems arising in seismic imaging, and inverse problems more generally.

One-dimensional single step

Suppose now that that the physical model is instead an infinite bi-layered medium, identified with $\mathbb{R}$, with the discontinuity placed at x_d, $\rho = \rho_L + \theta(x - x_d)(\rho_R - \rho_L)$, and similarly for λ, with θ being the Heaviside step function. The source point and the receiver are both located on the left half of the model at x_s and x_r respectively, with $x_s < x_r < x_d$, and we want to recover all the physical parameters of both media using full waveform inversion.

In this case, the right-moving part of the wave propagates by translation from x_s to x_d and splits into two waves, a reflected wave (with a possibly inverted phase) that travels back to x_r, and a transmitted wave that reaches $+\infty$. The ratio between the reflected and transmitted waves

and the incoming wave is given, respectively, by the *reflection* and *transmission* coefficients

$$\mathcal{R} \coloneqq \frac{Z_R - Z_L}{Z_L + Z_R},\ \mathcal{T} \coloneqq \frac{2Z_R}{Z_L + Z_R}, \tag{1.49}$$

where the impedances are given by $Z_i = \sqrt{\rho_i \lambda_i}$. Only two independent quantities can be measured by the receiver: the time delay between shot and detection, and the ratio between the amplitude of the reflected wave and that of the source. Since there are four physical parameters to be recovered, with no *a priori* relation between them, the problem is under-determined.

For our simple bi-layered model, the pressure field at the receiver is then given by

$$p(x_r, t) = S\left(t - \frac{x_r - x_s}{c_L}\right) + \mathcal{R}S\left(t - \frac{2x_d - x_s - x_r}{c_L}\right) \tag{1.50}$$

with $c_L^2 = \lambda_L/\rho_L$.

If we are given a measurement $d_r(t)$ at x_r, and we try to minimize the cost function, we see right away that the problem is ill-posed. Let us concentrate on the reflected wave, arriving between t_1 and t_2, which is the only (interesting) wave that can be measured at x_r. Using the explicit solution (1.50),

$$J(c, \rho) = \frac{1}{2}\int_{t_1}^{t_2} |p(x_r, t) - d_r(t)|^2 = \frac{1}{2}\int_{t_1}^{t_2} \left|\mathcal{R}\frac{s_0 c_L \rho_L}{2} S\left(t - \frac{2x_d - x_s - x_r}{c_L}\right) - d_r(t)\right|^2$$

where again $S' = s$. There are therefore only two adjustable parameters, the amplitude $\mathcal{R}s_0 c_L \rho_L/2$, and the delay $(2x_d - x_r - x_s)/c_L$. Thus, $c_L = \sqrt{\lambda_L/\rho_L}$ can be determined uniquely, but only $\mathcal{R}$, and not the individual parameters ρ_L, ρ_R and c_R, can be recovered. Measuring also the time derivative of the pressure would yield no new information, but the spatial derivative $\partial_x p$ at x_r would allow to determine ρ_L and thus $\mathcal{R}$. Nevertheless, this dataset would still be insufficient to recover λ_R and c_R individually. This problem is greatly magnified when the number of degrees of freedom to be reconstructed is very large, as is the case, e.g., when the modeling takes place over a mesh with hundreds of thousands of elements, and at least a couple of values per mesh element need to be determined. Thus, multiple different seismic models are compatible with the minimization of the cost function. This illustrates another aspect under which the seismic problem is ill-posed: there are many, potentially very different models that achieve comparable minima, and the seismic data are not able to discriminate them even in the best conditions. We also refer the reader to [70] for important considerations about the stability of the seismic inverse problem.

One-dimensional barrier

Finally, consider the case of a barrier, i.e., a finite homogeneous layer starting at x_1 and ending at x_2, with physical parameters ρ_b, c_b, surrounded by an infinite homogeneous medium with parameters ρ_a, c_a. Suppose that all the physical and geometrical values are fixed except for x_2, and let us perform an inversion over this parameter. This extremely simple problem mimics the inversion methods that contain geometric degrees of freedom (such as the position of the mesh vertices) as inversion parameters. We allow x_2 to take values less than x_1, which corresponds

to an inversion of the barrier.

For our one-dimensional problem, suppose that the source and receiver are placed at x_s and x_r respectively, with $x_s < x_r < x_1$. Multiple waves reach the receiver in this case. The first three waves to reach x_r correspond to the direct wave from the source, the wave reflected at $\min(x_1, x_2)$, and a wave that is transmitted at $\min(x_1, x_2)$, reflected back at $\max(x_1, x_2)$, and transmitted again at $\min(x_1, x_2)$. Another (in principle infinite) series of waves with decreasing amplitudes, known as *multiples*, reaches x_r after the first three, corresponding to multiple reflections inside the barrier. Let us ignore the direct wave from the source, which only carries redundant information (and would be discarded anyway in real seismic data). The solution at x_r can therefore be written as

$$\begin{aligned} p(x_r, t) =& \mathcal{R}_{ab} S\left(t - \frac{2\min(x_1, x_2) - x_r - x_s}{c_a}\right) \\ &+ \sum_{r=0}^{\infty} \mathcal{T}_{ab}\mathcal{R}_{ba}\left(\mathcal{R}_{ba}\right)^{2r}\mathcal{T}_{ba} S\left(t - \frac{2\min(x_1, x_2) - x_r - x_s}{c_a} - (2r+1)\frac{|x_2 - x_1|}{c_b}\right) \end{aligned}$$

where $\mathcal{R}_{ij}$ and $\mathcal{T}_{ij}$ indicate respectively the reflection and transmission coefficients (1.49) from the infinite medium to the barrier (if $ij = ab$) or vice-versa (if $ij = ba$). One can check that the solution is continuous as expected in the variable x_2, even when $x_2 = x_1$, since

$$\mathcal{R}_{ab} + \mathcal{T}_{ab}\mathcal{T}_{ba}\mathcal{R}_{ba}\sum_{r=0}^{\infty}\left(\mathcal{R}_{ab}\right)^{2r} = \mathcal{R}_{ab} + \mathcal{T}_{ab}\mathcal{T}_{ba}\mathcal{R}_{ba}\frac{(Z_a + Z_b)^2}{4Z_aZ_b} = 0,$$

i.e., no reflection, which is the expected result when the barrier becomes infinitely thin. However, if one computes the derivatives of $p(x_r, t)$ with respect to x_2, taken from the left and right sides (recall that $\sum_{r=0}^{\infty}(2r+1)x^{2r} = (x^2+1)/(x^2-1)^2$ for $|x| < 1$), one obtains

$$\begin{aligned} \left.\frac{\partial p(x_r, t)}{\partial x_2}\right|_{x_2 = x_1^+} &= -\frac{1}{c_b}\frac{(Z_a - Z_b)(Z_b^2 + Z_a^2)}{2Z_aZ_b(Z_a + Z_b)} s\left(t - \frac{2x_1 - x_r - x_s}{c_a}\right), \\ \left.\frac{\partial p(x_r, t)}{\partial x_2}\right|_{x_2 = x_1^-} &= \frac{1}{c_b}\frac{(Z_a - Z_b)(Z_b^2 + Z_a^2)}{2Z_aZ_b(Z_a + Z_b)} s\left(t - \frac{2x_1 - x_r - x_s}{c_a}\right). \end{aligned}$$

Therefore, the cost function is continuous, but not differentiable, at $x_2 = x_1$.

When x_2 crosses x_1, the barrier (seen as a geometrical object) undergoes an inversion. In more than one-dimension, this kind of inversion can happen even without the need for points to superpose. For example, a triangular obstacle can become degenerate and perform a similar inversion when one vertex is displaced, without the need for the vertex to cross the opposite segment (see Figure 1.5). In general, degenerate configurations can happen much more easily in $d > 1$.

Notice that, in addition to degenerate elements, the non-differentiability of the cost function can also be due to a change in shape of a mesh element causing the source location or one of the receiver locations to change the element in which it lies. More generally, if the geometry of the problem discretization is included as an explicit degree of freedom of the inversion, one might be forced to work with nonsmooth cost functions, justifying the need for subdifferentials (and

Figure 1.5: Example of how a mesh element in two dimensions can be reversed without any vertex superposing an edge. The cost function is generally not differentiable at the inversion location.

subgradient descent) in the minimization.

1.4 Discussion and further reading

We have presented in this chapter the main motivations of this work, namely the forward and inverse problems related to seismic surveying and subsurface exploration. Although other applications can be targeted by our method, this subject remains the main recipient of the techniques that we have developed.

For the sake of simplicity, and not to overcrowd this section with non-essential information, we have limited our account to the basic introduction of acoustic wave propagation in its second-order formulation, and to the classic inverse problem formulated through FWI via steepest descent and the adjoint-state method, using the classic least-squares cost function.

Many generalizations are possible. For the forward problem, the full elastic problem and its elasto-acoustic special case, with both isotropic and non-isotropic materials, are often studied [10, 71, 72]. Among non-isotropic materials, particular attention is given to TTI (Tilted Transverse Isotropic) media [73–75], which are general enough to matter in real-world applications, but symmetric enough to benefit from specific optimizations. However, the nature of the wave propagation problem as it pertains to our work is not fundamentally different in these cases: they are all time evolution problems guided by hyperbolic partial differential equations, with null initial conditions and (usually) absorbing boundary conditions, which are well-posed and depend on non-homogeneous physical parameters. As we shall see in the remainder, these are the core features of the kind of problems that we target in the present work.

Regarding the inverse problem, different imaging techniques are sometimes used, both pre-stack and post-stack, including Kirchhoff migration [16, 76] and RTM (Reverse-Time Migration) [77–79]. However, FWI seems to be today the most powerful inversion technique, whose application to real-world phenomena has only been delayed by its relatively sizeable computational requirements. Full-waveform inversion can be formulated both in a time-explicit fashion, as we have done here, or in the frequency domain, in which case the corresponding forward problem is the Helmholtz equation. Although different in many respects, both cases share the same basic optimization technology, and both are usually approached using steepest descent with the gradient computed via the adjoint state method. Thus, our presentation only requires some minor changes in order to be applied to the frequency domain. For a more complete overview of frequency-domain seismic inversion, see, e.g., [40, 47, 80] and the references therein, as well as, e.g., [81, 82] for some hybrid approaches.

As discussed in the main text, many different cost functions and regularization terms have been introduced to increase the robustness, accuracy and computational performance of FWI, see, e.g., [55, 56]. Once again, by adopting a very general point of view rooted in the some basic convex dualities, our presentation can accommodate most of these variations, although we do not discuss this here explicitly as it is not the main focus of this work.

All in all, we believe that our way of introducing these subjects is somewhat original and interesting, compared to the standard approach, for two reasons.

First, in the case of the forward problem, we have swapped the classic Hille-Yosida conditions on the range of the operator A with Lumer-Phillips' theorem, which relies on a condition on its adjoint $A^\star$. Aside from not requiring the introduction of more sophisticated concepts such as the resolvent set of a linear operator, this approach provides an excuse to bring the adjoint acoustic operator into the game early on, thus anticipating its importance in the formulation of the inverse problem.

Second, formulating the optimization problem underlying FWI through the powerful Fenchel-Rockafellar duality allows to derive the usual adjoint state method in a more abstract setting, as a direct consequence of a fundamental convex duality on Banach spaces. This illustrates that the adjoint state method is not a unique approach to a specific class of problems, but a more broadly applicable technique. As an instance of this, we have applied it, somewhat clumsily, to the simple linear regression problem. Finally, another reason why convex duality is relevant for seismic applications is that dual problems are often better conditioned and more well-posed that their primal counterparts. Thus, convex duality might perhaps be useful in shedding some light on the controllability of the seismic inversion problem, whose well-posedness has long remained an elusive goal.

2 | Discretization, the Galerkin and IPDG methods

> [...] Non domandarci la formula che mondi possa aprirti,
> sì qualche storta sillaba e secca come un ramo.
> Codesto solo oggi possiamo dirti,
> ciò che non siamo, ciò che non vogliamo.
>
> *Eugenio Montale, Non chiederci la parola che squadri da ogni lato, Ossi di seppia (1925)*

We have explored in the previous chapter the acoustic wave problem and its inverse. We have kept a physically-grounded, theoretical approach. However, the computational requirements necessary for the solution of the forward and especially inverse problem for real-world applications can only be met via computer simulations. Especially in the case of FWI, dedicated HPC platforms are often needed, see, e.g., [83, 84]. This is mainly due to the fact that, as we saw in Chapter 1, the evaluation of a viable direction for minimization entails the solution of two acoustic problems, a forward one and a backward one, and must be repeated for each shot, and for each iteration in the minimization process. Consequently, the computational demands of FWI are indeed extremely taxing. Thus, it is no surprise that the introduction of highly efficient computational models for wave propagation are at the forefront of FWI technology.

We present very briefly in this section the main state-of-the-art strategies for the simulation of PDEs, with special focus on hyperbolic problems of the second order. We introduce the main time discretization techniques used in these problems, before moving to the Galerkin method for space discretization, the main theorems on which it relies, and the main variants used in applications. We also introduce discontinuous Galerkin methods, and in particular an interior penalty version, which has interesting performance characteristics and has been used as a basis for a part for our work. From there, we briefly discuss the main issues in the discretization of the physical parameters, and the importance that they have in the inversion problem. Finally, we present a few aspects of the iterative methods employed in the minimization of the cost function.

2.1 Time integration

We begin this chapter by discussing the main strategies employed in the discretization of the time variable of the acoustic equation (1.5). Notice that space and time play a symmetric

role in the wave equation, and there exist some simulation techniques that put them on the same footing, discretizing the whole $(d+1)$-dimensional space, using causality to define the shape of the space-time domains. The reader may refer to, e.g., [85, 86] for a recent approach in this direction. However, adding one dimension often increases significantly the complexity of the discretization. Therefore, most simulation methods for time-dependent PDEs today discretize the time coordinate separately from the spatial coordinates, exploiting the fact that hyperbolic equations are formulated as Cauchy initial-value problems, and therefore can be computed iteratively starting from an initial known configuration.

In most cases, the time variable is simply treated by selecting a sequence of equispaced discrete times $t_1, \ldots, t_n$, starting with $t_1 = 0$, at which the solution $p(x, t_i)$ is computed from the knowledge of $p(x, t_j)$, $j < i$.

One of the simplest techniques, which is also used in the present work, consists in exchanging the continuous differential operator ∂_t^2 of (1.5) for a discretized version that converges to it in the limit $\Delta t \coloneqq t_{i+1} - t_i \to 0$. The simplest choice is perhaps the second-order approximation obtained via the three-point stencil of finite differences, i.e.,

$$\frac{\partial^2 p}{\partial t^2}(x, t_i) = \frac{p(x, t_{i+1}) - 2p(x, t_i) + p(x, t_{i-1})}{\Delta t^2} + O(\Delta t^2). \tag{2.1}$$

This scheme is also known as the second order *leapfrog* scheme, since it can be obtained by reformulating the wave equation as a couple of first order equations, whose variables are then evaluated at staggered intervals, i.e., $p_i, p_{i+1}, \ldots$ and $v_{i+\frac{1}{2}}, v_{i+\frac{3}{2}}, \ldots$. Thus, the two variables "leapfrog" over one another as the simulation progresses. The expression (2.1) computes $\partial_t^2 p(x, t_i)$ from the knowledge of p at times t_{i+1}, t_i and t_{i-1}. For example, if the differential equation reads

$$\frac{\partial^2 p}{\partial t^2}(x, t) = F(p(x, t)), \tag{2.2}$$

one may substitute $p_i \coloneqq p(x, t_i)$ for $p(x, t)$ inside F on the right hand side of the equation, using (2.1) to obtain

$$p_{i+1}(x) = 2p_i(x) - p_{t-1}(x) + \Delta t^2 F(p_i(x)),$$

which yields directly, if $p(x, t_{i-1})$ and $p(x, t_i)$ have been already computed, the pressure at the next step, $p_{i+1}(x)$. This value approximates $p(x, t_{i+1})$ within an error of order $O(\Delta t^2)$, and is used in place of $p_{i+1}(x)$ for the successive steps. This integration scheme is called *explicit*, as it does not involve the solution to a linear system. In contrast, one could evaluate F on a combination of the values of p at different times. For example, if F is linear in p and is evaluated at a combination of $p_{i+1}(x)$, $p_i(x)$ and $p_{i-1}(x)$, one can rewrite (2.2) in the form

$$p_{i+1}(x) - \Delta t^2 F_1(t, p_{i+1}(x)) = 2p_i(x) - p_{i-1}(x) + \Delta t^2 F_2(t, p_i(x), p_{i+1}(x)),$$

which can be solved to obtain $p_{i+1}(x)$. This scheme is called *implicit*, and involves a solution to a (usually linear) system at every timestep.

One of the main differences between implicit and explicit time integration schemes lies in their numerical stability. Specifically, implicit schemes are more computationally intensive but are often more stable or even *unconditionally stable*, i.e., the solution error does not grow without

bounds whatever the choice of timestep Δt. Notice that this does not mean that the solution is always accurate, only that the integration scheme yields *some* solution. In contrast, explicit schemes impose limits on the maximum allowable timestep that can be chosen in order to avoid numerical instability. These limits are usually obtained through von Neumann stability analysis [87]. In the case of hyperbolic PDEs, the condition on the timestep takes the name of Courant–Friedrichs–Lewy (CFL) condition [88]. For the second-order leapfrog (LF2) scheme (2.1), if the operator F appearing on the right hand side of (2.2) is a source term plus a bounded linear operator A, the condition reads (see, e.g., [89])

$$\Delta t_{\mathrm{CFL}} \leq \frac{2}{\sqrt{\lambda_{\max}(A)}}, \tag{2.3}$$

where $\lambda_{\max}(A)$ is the spectral radius of A, e.g., the norm of the largest eigenvector of the system matrix in a discretized problem. Intuitively, one must avoid timesteps Δt large enough that the wave could travel more than the space between two adjacent degrees of freedom during Δt, as this would lead to an exponential amplification of numerical errors (as well as an incorrect depiction of the underlying physics). Notice that this is merely a necessary condition, and not a sufficient one, for numerical stability. We will see that this constraint poses some requirements on the discretization of the operator F, as one must seek to maximize Δt_{CFL} in order to minimize the number of iterations and thus the computational cost of a simulation. Specifically, it will be interesting to study its behaviour as the space discretization becomes finer, and the order of spatial approximation k increases.

Notice that (2.3) is not the only condition on the timestep. We will see in a later section that the size of the spatial discretization also imposes a limit on the maximum allowable timestep, above which we may incur aliasing effects at the frequency(ies) of the source. In any case, one is free to choose a timestep below the limit (2.3), and the numerical result will generally have an order of convergence of $O(\Delta t^2)$ as Δt becomes smaller and smaller.

Higher-order variations of (2.1) exist, based on wider stencils. For example, the second derivative based on the five-point stencil reads

$$\frac{\partial^2 p(x,t_i)}{\partial t^2} \coloneqq \frac{-p(x,t_{i-2}) + 16p(x,t_{i-1}) - 30p(x,t_i) + 16p(x,t_{i+1}) - p(x,t_{i+2})}{12\Delta t^2} + O(\Delta t^4),$$

and is exact at order Δt^4. In fact, the $(2k+1)$-point stencil can be found by interpolating $p(x,t)$ on the points p_{i+j}, $j = -k, \ldots, k$ with a *Lagrange polynomial* $\ell(t)$ (see Chapter 3), and computing the corresponding derivatives of ℓ. The corresponding CFL conditions are also less stringent as the size of the stencil increases, by a factor $\sqrt{3}$ in the example of the five-point stencil [90]. Notice that larger stencils require more memory space, as they require to store solutions for more timesteps. Non-central approximations, that use indices $i+j$ not distributed symmetrically around i, are also available. As a final consideration on leapfrog schemes, we mention that some authors have found ways to achieve a higher convergence rate without requiring larger stencils, simply by virtue of adding more terms to the Taylor expansion (2.1), and replacing the higher time derivatives of p with time derivatives of $F(p(x,t))$, using the wave equation [91]. This shows that higher orders of convergence are achievable even with very (time-)local information on p, for the price of a few additional matrix multiplications.

Finally, another widely-used category of time integration schemes is the family of *Runge-Kutta* methods [92]. These schemes use intermediate points between t_i and t_{i+1} to improve the accuracy of the integration, by iteratively correcting the slope based on the "peeked values". Runge-Kutta methods are available at any order, but by far the most used is the fourth-order one (RK4). These methods are usually applied to first-order equations, so one needs to transform (1.5) into a set of coupled first-order equations of the form

$$\frac{\partial p}{\partial t} = G(t, p(x,t)).$$

One then proceeds via Algorithm 1. The last step in this scheme contains a weighted average of

Algorithm 1 Time integration via the RK4 scheme.

1: $(\bar{t}_1, \bar{p}_1) \leftarrow (t_i, p_i(x))$ ▷ starting values
2: $k_1 \leftarrow G(\bar{t}_1, \bar{p}_1)$ ▷ slope at starting values
3: $(\bar{t}_2, \bar{p}_2) \leftarrow (t_i + \frac{\Delta t}{2}, p_i(x) + k_1 \frac{\Delta t}{2})$ ▷ midpoint, according to slope k_1
4: $k_2 \leftarrow G(\bar{t}_2, \bar{p}_2)$ ▷ slope at $(\bar{t}_2, \bar{p}_2)$
5: $(\bar{t}_3, \bar{p}_3) \leftarrow (t_i + \frac{\Delta t}{2}, p_i(x) + k_2 \frac{\Delta t}{2})$ ▷ midpoint, according to slope k_2
6: $k_3 \leftarrow G(\bar{t}_3, \bar{p}_3)$ ▷ slope at $(\bar{t}_3, \bar{p}_3)$
7: $(\bar{t}_4, \bar{p}_4) \leftarrow (t_i + \Delta t, p_i(x) + k_3 \Delta t)$ ▷ endpoint, according to slope k_3
8: $k_4 \leftarrow G(\bar{t}_4, \bar{p}_4)$ ▷ slope at $(\bar{t}_4, \bar{p}_4)$
9: $p_{i+1}(x) \leftarrow p_i(x) + \frac{k_1 + 2k_2 + 2k_3 + k_4}{6} \Delta$

the slopes at the four computed positions, whose weights have been computed beforehand with the goal of cancelling lower order terms, leaving a final convergence rate of $O(\Delta t^4)$. Other time discretization schemes, such as the *Adams-Bashforth* and *Adams-Moulton* methods, are based on the same principle, but use the information gathered at previous steps to compute the next one (they are known as *multi-step* methods).

This family of methods is generally very accurate and thus widely used. However, they are not symplectic, i.e., they do not preserve the total energy of the system, which can lead to some instability for longer simulation times. This is usually not an insurmountable problem for seismic wave propagation, since energy conservation is not a pressing issue, and energy dissipation occurs anyway due to the boundary conditions or to penalization terms (see for example the section on the discontinuous Galerkin method). However, since we are solving the second-order wave equation and our timestep is often limited by other considerations, we will not use these techniques in our work.

To summarize, the most important features of these integration schemes are

- The order of approximation, i.e., the exponent k of the error $e = O(\Delta t^k)$;
- The numerical stability, i.e., the maximum allowable timestep.

We will discus in the next section how the second criterion enters the choice of spatial discretization for the problem.

2.2 Spatial discretization

The choice of how to discretize the solution is of paramount importance for the efficacy and efficiency of a numerical scheme.

First of all, the discretization must be able to correctly reproduce all the possible solutions in all the scenarios of interest. For instance, in the case of the acoustic wave equation, if the frequency of the source is peaked around ω_0, then the (local) spatial distance between neighboring degrees of freedom must not be larger than about c/ω_0, where c is the local value of the velocity. In reality, the size must be much smaller to avoid numerical noise.

Moreover, all the differential operators of a PDE need to be translated to the space of discrete degrees of freedom with which the solution is described. The result of this translation, and the numerical and stability properties of the PDE, are therefore crucially dependent on this choice.

One of the first space discretization methods to be adopted in the solution of hyperbolic PDEs was the *finite difference* (FD) method ([93], see, e.g., [94, 95] for some seismic applications). With this choice, degrees of freedom correspond to the values of the solution $p(x_i, t)$, $x_i \in L$ on a lattice of points $L \subset \mathbb{R}^d$. Differential operators D are then replaced by discrete ones, based on stencils relating points close to each other. The stencil is applied to compute the values $(Dp)(x_i, t_j)$, which are used to compute $p(x_i, t_{j+1})$ through one of the time integration schemes seen above. Advantages of this method include its simplicity, both in implementation and in error analysis, and the possibility to employ powerful and very efficient computational techniques such as the Fast Fourier Transform (FFT) for its solution. However, the regular structure of the points requires a very fine mesh for the description of complex geometries, partially negating the advantages of the method in these cases. Nevertheless, finite difference methods remain very popular in industrial applications due to their simplicity.

Another very popular discretization technique is based on the *finite volume* (FV) method [96]. In this paradigm, after recasting the problem as a first-order equation both in time and space,

$$\frac{1}{\lambda}\frac{\partial p}{\partial t} + \nabla \cdot v = s(x,t), \ \frac{\partial v}{\partial t} + \frac{1}{\rho}\nabla p = 0, \tag{2.4}$$

the variables are discretized via their integrated (or average) values,

$$\langle p \rangle_i \coloneqq \int_{C_i} p(x,t)\,\mathrm{d}x, \ \langle v \rangle_i \coloneqq \int_{C_i} v(x,t)\,\mathrm{d}x,$$

over a set of discrete cells C_i, $i = 1, \ldots, n$. Using the divergence theorem, one is able to relate the values in neighboring cells through surface integrals on the shared interface. This results in an algebraic linear system, where integrated values $\langle \partial_t p \rangle$ in neighboring cells i, j are related by *fluxes* that are computed on the surfaces. We will discuss the form of the fluxes later when we introduce the *discontinuous Galerkin* method. Finite volume methods are usually based on an adaptable unstructured mesh, and are thus very well suited for describing complex geometries. However, these methods have a slow order of convergence, and can achieve higher precision only by significantly decreasing the mesh size.

Finally, some techniques use degrees of freedom that are not based on an underlying mesh (*meshless* techniques), but on an unstructured set of points. For example, *smoothed particle*

hydrodynamics (SPH) describes the solution by introducing a point cloud of points $(a_i)_{i=1,\dots,n}$ in $\mathbb{R}^d$ and using *radial basis functions* $\varphi_i(x) \coloneqq \varphi(x - a_i)$ centered on the points, where φ is the general form of a kernel, usually a Gaussian or a harmonic function (*thin plate splines*). The solution p is then expressed as a sum over these functions,

$$p(x,t) \coloneqq \sum_{i=1}^{n} \gamma_i(t)\varphi_i(x).$$

The problem is then reformulated in the weak sense, similarly to the Galerkin method that we discuss in the rest of this section. We will not present these methods here. We will however note that, aside from the great advantage of avoiding the need for a geometric mesh altogether, these methods suffer from a few drawbacks, notably the difficulty in imposing boundary conditions, as the boundary position is not easily described.

A visual representation of the three methods is given in Figure 2.1.

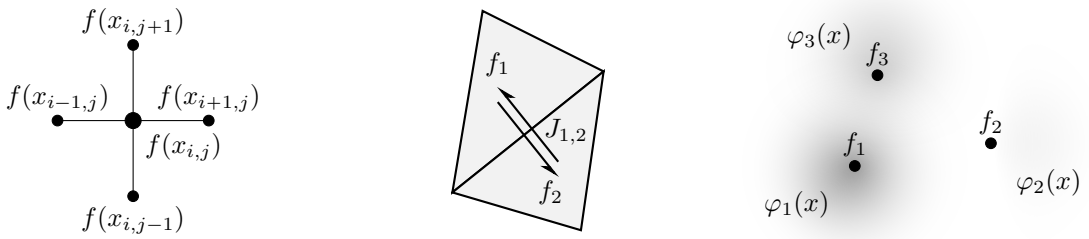


Figure 2.1: (Left) typical 5-point stencil used for the computation of the Laplacian in two-dimensional finite differences. (Middle) typical setup for finite volumes on a triangulation. The flux $J_{1,2}$ between two cells is computed as a function of the values of f on the neighboring cells. (Right) three points and their associated radial basis function in SPH.

2.2.1 The Galerkin method

Many modern schemes for the solution of PDEs, like the finite element (FE) method [97, 98] and isogeometric analysis (IGA) [2], are based on a weak formulation that is formalized by Galerkin theory. This is a very general approach that has the advantage of retaining the generic geometrical description of FV methods, while allowing for a sub-geometric degree of precision thanks to the introduction of appropriately shaped basis functions. Moreover, the weak formulation on which the Galerkin method relies is also used for many error estimates in PDE theory, and these tools are therefore available for the characterization of numerical Galerkin schemes. Galerkin methods have been very successful in treating elliptic and parabolic equations, especially for self-adjoint operators. After the introduction of *discontinuous Galerkin* methods, this approach has been successfully applied to hyperbolic equations [99–101], as we shall see in the next section.

Weak form of the equation

For this subsection and the next one, we refer the reader to a more complete treatment found in the literature, and only give a very partial account. For second-order hyperbolic problems, one can refer to [102, 103] for estimates of errors in Galerkin methods. A good introduction to the subject is given in [104]. For discontinuous Galerkin methods, we refer to, e.g., [105–107] for a more general introduction and to [108, 109] for error analysis in second-order hyperbolic problems.

The starting point for the Galerkin method implementation is reformulating (1.5) in its *weak form*, or *variational form*. Instead of requiring that (1.5) holds exactly, one chooses a finite (but representative) set of *test functions* $\varphi \in \mathcal{T}$, and requires that the projection of the differential equation over each of these functions is satisfied, i.e.,

$$\int_\Omega \frac{1}{\lambda}\varphi\frac{\partial^2 p}{\partial t^2}\,\mathrm{d}\Omega - \int_\Omega \varphi\nabla\cdot\left(\frac{1}{\rho}\nabla p\right)\,\mathrm{d}\Omega = \int_\Omega \varphi s\,\mathrm{d}\Omega,$$

and after integrating by parts and applying the absorbing boundary conditions (1.9) on p, one obtains

$$\int_\Omega \frac{1}{\lambda}\varphi\frac{\partial^2 p}{\partial t^2}\,\mathrm{d}\Omega + \int_{\partial\Omega_A} \frac{1}{\rho c}\varphi\frac{\partial p}{\partial t}\,\mathrm{d}S + \int_\Omega \frac{1}{\rho}\nabla\varphi\cdot\nabla p\,\mathrm{d}\Omega = \int_\Omega \varphi s\,\mathrm{d}\Omega. \tag{2.5}$$

This is the weak form of (1.5), which we can rewrite as

$$\left\langle \varphi, \frac{\partial^2 p}{\partial t^2}\right\rangle_\lambda + \left\langle \varphi, \frac{\partial p}{\partial t}\right\rangle_{\rho c, A} + a_\rho(\varphi, p) = \langle \varphi, s\rangle\,, \tag{2.6}$$

where the subscripts λ, ρc indicate that the integration is done with respect to these measures, and the subscript A indicates that the integration is performed only on the absorbing boundary of Ω. The last term on the left hand side of (2.5) is the only one containing the spatial derivatives. It corresponds to the bilinear form in φ, p

$$a_\rho(\varphi, p) \coloneqq \int_\Omega \frac{1}{\rho}\nabla\varphi\cdot\nabla p\,\mathrm{d}\Omega.$$

Let us make the hypothesis that the physical parameters λ and ρ can be described via positive bounded functions. Then, since $a_\rho(p,p) = \|\nabla p\|^2_{L^2(\Omega,\rho)}$,

$$a_\rho(p,p) \geq C_{1,\rho}\,\|\nabla p\|_{L^2(\Omega)} \text{ for all } p \in H^1_0(\Omega), \tag{2.7}$$

with the nonnegative constant $C_{1,\rho}$ independent of p. Since ρ is positive, Cauchy-Schwartz's inequality also implies

$$a_\rho(\varphi, p)^2 \leq C_{2,\rho}\, a_\rho(\varphi,\varphi) a_\rho(p,p) \text{ for all } p \in H^1(\Omega),$$

again with the nonnegative constant $C_{2,\rho}$ independent of p. Similar inequalities are also valid if λ and ρ are chosen from some more general measure spaces, see, e.g., [110], although this

generality is not required in the rest of this work. Notice that, after choosing

$$\varphi \coloneqq \frac{\partial p}{\partial t},$$

one can rewrite the left hand side of (2.6) as

$$\frac{\mathrm{d}}{\mathrm{d}t} E_{\rho,\lambda}(p) + \int_{\partial\Omega_A} \frac{1}{\rho c} \left| \frac{\partial p}{\partial t} \right|^2 \mathrm{d}S$$

where the *energy* $E_{\rho,\lambda}(p)$ is defined as

$$E_{\rho,\lambda}(p) \coloneqq \frac{1}{2} \int_\Omega \frac{1}{\lambda} \left| \frac{\partial p}{\partial t} \right|^2 \mathrm{d}\Omega + \frac{1}{2} a_\rho(p,p) =: e^2_{\rho\lambda}(p), \tag{2.8}$$

and the integral on the boundary is always positive and represents therefore an absorption (i.e., damping) term.

We now wish to discretize (2.5) by choosing a finite-dimensional subspace $\Phi_h \subset \mathcal{T}$ for the test functions, and another finite-dimensional subspace $\mathcal{Q}_h \subset H^2(\Omega) =: \mathcal{Q}$ for the solution. The subscript h refers to the fact that these spaces usually depend on some geometric construction (e.g., a grid or a mesh) with a typical size h.

In this work, we make the choice $\Phi_h = \mathcal{Q}_h$, which is called the *Bubnov-Galerkin* approach. Choosing a different discretized space for the space of test functions and the space of solutions would result in a *Petrov-Galerkin* scheme. Such schemes have many advantages, including the potential to make some matrices very simple or even diagonal, but for splines, the construction of dual spaces with good numerical properties is an active research topic (see, e.g., [111]), and not mature enough to be used in this work.

A crucial criterion in the choice of the subspace $\mathcal{Q}_h$ is whether the discretized solution p_h approaches p as the discretization becomes finer and finer, i.e., as $h \to 0$, and how quickly. The study of this behavior takes the name of *a priori* error analysis, meaning that these estimates cannot be used to compute actual numerical bounds on the error, since they depend on the knowledge of the non-discretized weak solution p. However, these estimates are fundamental in establishing the correctness (and degree of convergence) of the discretization scheme. The goal of the next few subsections is to illustrate the crucial role played by polynomials, and polynomial-reproducing discretization spaces, in the determination of such estimates. We show in Figure 2.2 the typical shape of a solution to the (continuous) Gaerkin problem.

A priori error analysis

Suppose that p is a solution of the weak problem (2.5) over all $\varphi \in \mathcal{T}$, and suppose that we find a discretized solution $p_h \in \mathcal{Q}_h$ to the semi-discrete (i.e., discretized only in space) problem (2.5) formulated over the subset of test functions $\varphi_h \in \mathcal{Q}_h \subset \mathcal{T}$. The question that we wish to answer in this section is the following: can we place some bounds on the asymptotic behavior of the difference between the semi-discretized solution and weak the solution of (2.5) as the discretization becomes finer, i.e., as $h \to 0$?

A first, unavoidable requirement on the discretized space $\mathcal{Q}_h$ is that it must be capable of

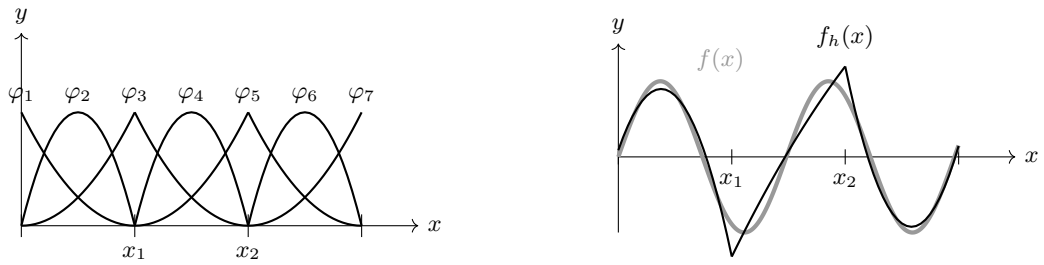


Figure 2.2: (Left) a typical piecewise-polynomial basis of degree 2 on three elements used for the Galerkin method in one-dimension and (right) the L^2 projection of a function on this basis. Note that the projection is only continuous.

representing the weak solutions to (2.5) as the discretization becomes finer, i.e.,

$$\lim_{h\to 0} \inf_{\varphi_h \in \mathcal{Q}_h} \|p - p_h\|_{L^\infty([0,T_f];L^2(\Omega))} = 0.$$

Once this requirement is met, we wish to estimate the error of the Galerkin solution, i.e.,

$$\|p - p_h\|^2_{L^\infty([0,t_f];L^2(\Omega))} \coloneqq \sup_{0\le t\le t_f} \int_\Omega |p(x,t) - p_h(x,t)|^2 \, \mathrm{d}\Omega, \tag{2.9}$$

as a function of the discretization size h. Notice that such estimates are necessary to ensure the numerical stability of the discretization scheme, but need to be supplemented with conditions on the integration scheme in time. This kind of analysis is better suited to a full *a posteriori* error analysis, which we do not perform here as it is not in the scope of the present work.

Let us split the goal of estimating (2.9) in two steps: first, we show that the Galerkin solution is optimal, in the sense that it achieves the same degree of convergence as the best approximating function in the space $\mathcal{Q}_h$. The estimate is then reduced to a pure functional approximation problem. Once this is done, we can derive bounds on the error that depend explicitly on h (and thus determine the order of convergence) after assuming that $\mathcal{Q}_h$ contains all the polynomials up to a given degree in its linear span.

Optimality of the Galerkin solution

A good introduction to the optimality estimates for the Galerkin method in second-order hyperbolic equations can be found in [104]. We also refer the reader to [103], which obtains slightly tighter estimates, to [112] for a different treatment of the absorbing boundary term in the case of mixed finite elements, and to [108] for a similar estimate in the case of discontinuous Galerkin methods.

Here, we apply the results of the classic approach [102], which proves the convergence rate of the Galerkin approximation to the acoustic wave equation with absorbing boundary conditions. The goal of this subsection is to show the philosophy underlying such *a priori* error estimates, and to illustrate the role played by polynomials and polynomial-reproducing discretizations.

Following [102] and many other classical approaches, we use the bilinear form a_ρ to define the *elliptic projection* $\Pi_h p \in \mathcal{Q}_h$ of the solution p onto $\mathcal{Q}_h$. The bilinear form a_ρ is not coercive,

and therefore, it cannot be used alone to define the elliptic projection as it would not yield a unique solution. We can however make the definition unique by requiring that $\Pi_h p$ satisfies

$$a_\rho(p - \Pi_h p, \varphi_h) + \langle p - \Pi_h p, \varphi_h \rangle_\lambda = 0, \tag{2.10}$$

at all times t and for every $\varphi_h \in \mathcal{Q}_h$. The condition (2.10) uniquely defines $\Pi_h p$, since the second term acts as a regularization.

Notice that the projection defined by (2.10) is possible because $\mathcal{Q}_h \subset \mathcal{Q}$, i.e., because the discretized space is *conforming*, which implies the fact that $\Pi_h p$ has a well-defined trace on $\partial\Omega$. Inevitably, some modifications are required in the case of the discontinuous Galerkin methods, whose discretization space is non-conforming, as shown later in this chapter.

Let us use the shorthand $w_h \coloneqq \Pi_h p$ for the elliptic projection, and let us derive an estimate for $\|p - p_h\|_{L^\infty([0,t_f];L^2(\Omega))}$. Since both p and p_h satisfy (2.6) for test functions $\varphi_h \in \mathcal{Q}_h$, by plugging φ_h as into (2.6), one obtains

$$\left\langle \partial_t^2(p - p_h), \varphi_h \right\rangle_\lambda + \langle \partial_t(p - p_h), \varphi_h \rangle_{\rho c, A} + a_\rho(p - p_h, \varphi_h) = 0, \tag{2.11}$$

and, separating $p - p_h = (p - w_h) - (p_h - w_h)$ and using (2.10),

$$\begin{aligned} \left\langle \partial_t^2(p - w_h), \varphi_h \right\rangle_\lambda - \langle p - w_h, \varphi_h \rangle_\lambda + \langle \partial_t(p - w_h), \varphi_h \rangle_{\rho c, A} \\ = \left\langle \partial_t^2(p_h - w_h), \varphi_h \right\rangle_\lambda + a_\rho(p_h - w_h, \varphi_h) + \langle \partial_t(p_h - w_h), \varphi_h \rangle_{\rho c, A}. \end{aligned}$$

From this expression, [102, 104] derive a suitable error estimate for $(p_h - w_h)$ and its time derivatives. We can directly apply these estimates to our case. In particular, our approach corresponds to the choice $\alpha_1 = 0$, $\alpha_2 = 1$, $g = 0$ and $a = c$ in [102]. Furthermore, since we are assuming that ρ, λ and c are piecewise smooth positive functions, the norms $\|p\|_\lambda^2$, $a_\rho(p,p) + \|p\|_{L^2(\Omega)}^2$ and $\|p\|_{\rho c, A}^2$ are equivalent to the norms $\|p\|_{L^2(\Omega)}^2$, $\|p\|_{H^1(\Omega)}^2$ and $\|p\|_{L^2(\partial\Omega_A)}^2$, respectively. Finally, thanks to the initial conditions (1.7), $p(0) = \partial_t p(0) = 0$, we can exactly represent both p and $\partial_t p$ at time zero in the discretized space $\mathcal{Q}_h$ by simply choosing the zero vector. Overall, we recover the following lemma.

Lemma 2.2.1 ([102, Lemma 3])**.** *There exists a constant C such that*

$$\begin{aligned} \|\partial_t(p_h - w_h)\|_{L^\infty([0,t_f];L^2(\Omega))}^2 + \|\partial_t(p_h - w_h)\|_{L^\infty([0,t_f];H^1(\Omega))}^2 + \|\partial_t(p_h - w_h)\|_{L^\infty([0,t_f];L^2(\partial\Omega))}^2 \\ \leq C \Big(\left\|\partial_t^2(p - w_h)\right\|_{L^2([0,t_f];L^2(\Omega))}^2 + \|p - w_h\|_{L^2([0,t_f];L^2(\Omega))}^2 \\ + \|p - w_h\|_{L^\infty([0,t_f];H^{-1/2}(\partial\Omega_A))}^2 + \|\partial_t(p - w_h)\|_{L^\infty([0,t_f];H^{-1/2}(\partial\Omega_A))}^2 \\ + \left\|\partial_t^2(p - w_h)\right\|_{L^2([0,t_f];H^{-1/2}(\partial\Omega_A))}^2 \Big), \end{aligned} \tag{2.12}$$

where $\|\cdot\|_{H^{-1/2}(\partial\Omega_A)}$ is the norm on $H^{-1/2}(\partial\Omega_A)$, the dual of $H^{1/2}(\partial\Omega_A)$.

This result is important because it bounds the difference between the Galerkin solution p_h and the elliptic projection w_h in a fine norm, thus showing that the Galerkin solution is optimal. Furthermore, (2.12) allows to derive directly an estimate for the Galerkin error via the triangular

inequality, i.e.,

$$\|p - p_h\|_{L^\infty([0,t_f];L^2(\Omega))} \leq \|p - w_h\|_{L^\infty([0,t_f];L^2(\Omega))} + \|p_h - w_h\|_{L^\infty([0,t_f];L^2(\Omega))},$$

where the last term can be bounded by (2.12). Thus, the problem of estimating the accuracy of the Galerkin solution has been transformed into a pure functional approximation problem.

Error estimate in the case of polynomial-reproducing spaces

Suppose now that $\mathcal{Q}_h$ contains the space of all the polynomials up to a degree k, $\mathcal{Q}^k$, in its linear span. We say in this case that $\mathcal{Q}_h$ is *polynomial-reproducing*. We can then apply to (2.12) a number of estimates that rely on the central role that polynomials play in approximation theory. More specifically, the Stone-Weierstrass theorem [113] states that polynomials are dense in the space of continuous functions, and thus every continuous function can be uniformly approximated by polynomials. Furthermore, after denoting with h the (maximum) size of a mesh element in the discretization, polynomial approximation can be used to determine the order of convergence in terms of h via Jackson-type inequalities [114] such as the Bramble-Hilbert Lemma [97, 115], which states

$$\inf_{\varphi_h \in \mathcal{Q}_h} \|p - \varphi_h\|_{H^r(\Omega)} \leq C' h^{\min(k+1-r,m)} \|p\|_{H^m(\Omega)},$$

and thus in particular

$$\inf_{\varphi_h \in \mathcal{Q}_h} \|p - \varphi_h\|_{L^2(\Omega)} \leq C h^{\min(k+1,m)} \|p\|_{H^m(\Omega)},$$

where the constants C and C' do not depend on h, but can depend on the polynomial order k. This establishes the degree of approximation of polynomial spaces.

Using this kind of inequalities, and with some regularity assumptions on p and its time derivatives, one can determine the order of convergence of the norms of $p - w_h$ appearing in (2.12), see, e.g., [116]. One then obtains the result

Lemma 2.2.2 ([102, Lemma 5])**.** *Suppose that p, $\partial_t p$ are in $L^\infty([0,t_f];L^2(\Omega))$, and that $\partial_t^2 p$ is in $L^2([0,t_f];L^2(\Omega))$. Then the inequality*

$$\|\partial_t^r (p - w_h)\|_{L^s([0,t_f];L^2(\Omega))} + \|\partial_t^r (p - w_h)\|_{L^s([0,t_f];H^{-1/2}(\partial\Omega_A))} \leq C h^k, \tag{2.13}$$

where C is a constant independent of h, holds for $(r,s) = (0,\infty)$, $(1,\infty)$ and $(2,2)$.

Together with (2.12), (2.13) yields the following final estimate.

Theorem 2.2.3 ([102, Theorem 2])**.** *Let p be a solution of* (1.5) *with boundary conditions* (1.8) *and* (1.9)*, and with initial conditions* (1.7)*, and let p_h be its Galerkin approximation. Suppose that p, $\partial_t p$ are in $L^\infty([0,t_f];L^2(\Omega))$, and that $\partial_t^2 p$ is in $L^2([0,t_f];L^2(\Omega))$. Then there is a constant C independent of h such that*

$$\|\partial_t (p - p_h)\|_{L^\infty([0,T_f];L^2(\Omega))} + \|p - p_h\|_{L^\infty([0,T_f];L^2(\Omega))} \leq C h^k.$$

This result shows the order of convergence of the semi-discrete Galerkin solution with respect to the solution of the (non discretized) weak problem, finalizing the *a priori* error estimate.

A final consideration. The estimates provided in this section are only *a priori* error estimates, since evaluating the magnitude of the error would require the knowledge of the exact solution p. In particular, since no discretization in time was done in this section, the solution p_h is not the discrete solution that one finds in practice, but the solution of a *semi-discrete* scheme. The choice of an appropriate discretization in time is crucial in order to preserve the maximal order of approximation.

One can also produce some *a posteriori* error estimates, based on the fully discretized scheme. Computing such estimates involves the knowledge of details about the type of spatial discretization (grid, mesh, etc.) and also the choice of time integration scheme. For the second order wave equation with explicit timestepping, one might find some estimates, e.g., in [117] for the Runge-Kutta integration scheme and in [118] for the leapfrog scheme.

Discrete form for explicit timestepping

We give now the full form of the discretized acoustic wave equation in the weak form (2.6), using the second-order leapfrog time integration scheme. The space of discretized functions $\mathcal{Q}_h$ is finite-dimensional, and thus admits a basis. Let $(\varphi_i)_{i=1}^n$ be such a basis. By far the most common basis choice is that of piecewise polynomials, for examples defined on a set of volumes that constitute a subdivision (a *mesh*) of Ω. Notice that some form of regularity must be imposed across the whole domain for the error estimates of the previous subsection to be valid. Different bases for polynomials can be chosen, as we will discuss in a later chapter.

Once the choice of basis is made, one can expand the solution on this basis at all times,

$$p(x,t) = \sum_{i=1}^{n} \gamma_i(t)\varphi_i(x),$$

and use any of the basis functions φ_j as a test function. One can then incorporate the time integration scheme on the coefficient vector $\gamma(t)$, which is replaced with a vector of real numbers $\gamma^{(t)}$, where the integer t now simply signifies the iteration number. Using the formula (2.1) for $\partial^2\gamma/\partial t^2$, and using a central approximation for the first derivative in time $\partial\gamma/\partial t$, i.e.,

$$\frac{\partial\gamma}{\partial t}(t_i) = \frac{p(x,t_{i+1}) - p(x,t_{i-1})}{2\Delta t} + O(\Delta t^2),$$

one then obtains the linear system

$$M\frac{\gamma^{(t+1)} - 2\gamma^{(t)} + \gamma^{(t-1)}}{\Delta t^2} = -K\gamma^{(t)} - B\frac{\gamma^{(t+1)} - \gamma^{(t-1)}}{2\Delta t} + S(t), \tag{2.14}$$

where

$$M_{ij} \coloneqq \int_\Omega \frac{1}{\lambda}\varphi_i\varphi_j \,\mathrm{d}\Omega \tag{2.15}$$

is the *mass matrix*,

$$K_{ij} \coloneqq \int_\Omega \frac{1}{\rho}\nabla\varphi_i \cdot \nabla\varphi_j \,\mathrm{d}\Omega \tag{2.16}$$

is the *stiffness matrix*,

$$B_{ij} \coloneqq \int_{\partial\Omega_A} \frac{1}{\rho c}\varphi_i\varphi_j \,\mathrm{d}S \tag{2.17}$$

is the *boundary condition matrix* or *damping matrix*, and

$$S_i(t) \coloneqq \int_\Omega \varphi_i(x)s(x,t)\,\mathrm{d}\Omega$$

is the *source vector*. The explicit timestepping scheme can then be obtained by making $\gamma^{(t+1)}$ explicit in (2.14),

$$\gamma^{(t+1)} = \left(M + \frac{\Delta t}{2}B\right)^{-1}\left(2M\gamma^{(t)} - M\gamma^{(t-1)} - \Delta t^2 K\gamma^{(t)} + \frac{\Delta t}{2}B\gamma^{(t-1)}\right) + \Delta t^2 S(t). \tag{2.18}$$

The mass matrix (2.15) is diagonal if and only if the basis functions φ_i are orthogonal with respect to the scalar product induced by ρ. In fact, M corresponds exactly to the Gramian matrix of these functions, cf. [119, Chapter 7]. Numerical methods that make this matrix diagonal or otherwise easy to invert have therefore a tremendous numerical advantage in computing the time evolution of acoustic waves via (2.18). Notice that, even if only the inverse of the sum $(M + \Delta t/2\, B)$ appears in (2.18), the term $\Delta t/2\, B$ can be regarded as a perturbation of the matrix M, and the inverse of the sum can be computed knowing the inverse of M, for example through the Sherman-Morrison formula [120].

Since the contribution of the matrix B is only important near the boundaries, the numerical stability of the wave propagation algorithm is dominated by the spectral properties of $M^{-1}K$, and the CFL timestep can be tied to the maximum eigenvalue of this matrix (cf. (2.3)). Thus, the behavior of the spectrum of this matrix as the size of the space discretization decreases and the degree of polynomial approximation increases is of paramount importance for the performance results of a discretization scheme.

2.2.2 The interior penalty discontinuous Galerkin method

In real-world applications of seismic wave propagation and inversion, one often does not have the luxury of working in a smooth setting. First, the physical subsurface itself is often full of discontinuities in velocity, density, or both. These zones of high gradient of impedance cannot be simply disregarded or smoothed out without nefarious consequences on the convergence of the inverse problem, since the reflectors are responsible for some substantial features of the recorded data. Second, even if the pressure stays continuous, it will generally no longer be in $H^1(\Omega)$. Moreover, the first-order formulation (2.4) makes use of the wave velocity variable v, whose tangential component is discontinuous across a reflector. For this reason, the choice of discontinuous discretization functions that underlies the *discontinuous Galerkin* (DG) method has found a fertile ground in seismic wave simulation, see, e.g., [1, 89, 108, 121–123], as well as in many other hyperbolic problems.

In this work, we will make use of the standard DG approach. Compared to the continuous Galerkin approach presented in the previous section, the DG method allows to perform computations in a local fashion, leading to the aforementioned advantages in the modelization of

complex geometries, as well as to efficient parallelization schemes. Even though we do not use the standard DG geometric discretization and associated bases, we make use of many high-level characteristics of the DG approach, which allow us to construct a multi-domain, hybrid DG-IGA scheme that we will detail in Chapter 6. Consequently, we keep this introduction abstract enough to be applicable to our work. For a general and complete introduction to discontinuous Galerkin methods see, e.g., [107] or [105].

In a discontinuous Galerkin scheme, one subdivides the domain Ω into a finite set of disjoint subdomains Ω_i, $i = 1, \dots, n_d$, satisfying

$$\Omega_i \cap \Omega_j = \varnothing \text{ for } i \neq j, \bigcup_{i=1}^{n_d} \overline{\Omega}_i = \overline{\Omega},$$

and introduces a set of discretized function spaces $\mathcal{S}_{h,i}$ independently for each Ω_i. Each $\mathcal{S}_{h,i}$ satisfies $\mathcal{S}_{h,i} \subset H^1(\Omega_i)$, but generally

$$\mathcal{S}_h \coloneqq \bigoplus_{i=1}^{n_d} \mathcal{S}_{h,i} \not\subseteq H^1(\Omega),$$

and the violation is large, since a general function obtained by picking a different function $\varphi_{h,i} \in \mathcal{S}_{h,i}$ per domain is not continuous, let alone smooth, at the interfaces between domains. We however assume that $\mathcal{S}_h \subset L^2(\Omega)$, which means that the discretization space is *non-conforming*, and thus (2.10) cannot be derived directly as we did in the previous subsection. An alternative strategy has to be devised.

Let us introduce the usual setup for discontinuous Galerkin methods. Let $\mathcal{F}_I$ be the set of internal *facets* of the domain decomposition of Ω, i.e., the set of all intersections $F_{i,j} \coloneqq \overline{\Omega}_i \cap \overline{\Omega}_j$ between any two neighboring cells Ω_i and Ω_j. Suppose that we have chosen a normal direction N_F for every face $F_{i,j}$. Given a scalar function p and a vector-valued function v on Ω, we can then define their *jump* at the interface F as

$$[\![p]\!]_F \coloneqq (p^+ - p^-)N_F, \; [\![v]\!]_F \coloneqq (v^+ - v^-) \cdot N_F, \tag{2.19}$$

where p^+, v^+ (respectively p^-, v^-) are the traces of p and v taken in the domain which sees N_F as an outward normal (resp., inward normal). Notice that the jump of a scalar function is vector-valued, while the jump of a vector-valued function is a scalar. Similarly, we define their *average*

$$\{\!\{p\}\!\}_F \coloneqq \frac{1}{2}(p^+ + p^-), \; \{\!\{v\}\!\}_F \coloneqq \frac{1}{2}(v^+ + v^-). \tag{2.20}$$

We can extend this definition to the whole set $\mathcal{F}$ of facets of the decomposition, also known as its *skeleton*, by defining these expressions on the set of boundary facets $\mathcal{F}_B$ of Ω. For a boundary facet $B \in \mathcal{F}_B$, we define

$$[\![p]\!]_B \coloneqq pN_B, \; [\![v]\!]_B \coloneqq v \cdot N_B, \; \{\!\{p\}\!\}_B \coloneqq p, \; \{\!\{v\}\!\}_B \coloneqq v, \tag{2.21}$$

with the normal N_B always chosen in the inward direction with respect to Ω. We denote by $[\![p]\!]$, $[\![v]\!]$, $\{\!\{p\}\!\}$ and $\{\!\{v\}\!\}$ the functions defined on the whole skeleton via (2.19), (2.20) and (2.21).

Notice that

$$\begin{aligned}\llbracket p \rrbracket_F \cdot \{\!\{ v \}\!\}_F + \llbracket v \rrbracket_F \{\!\{ u \}\!\}_F &= \frac{1}{2}(p^+v^+ + \cancel{p^+v^-} - \cancel{p^-v^+} - p^-v^- \\ &\quad + p^+v^+ + \cancel{p^-v^+} - \cancel{p^+v^-} - p^-v^-) \cdot N_F, \\ &= \llbracket pv \rrbracket_F,\end{aligned} \tag{2.22}$$

which is sometimes called the *jump identity* or *trace identity*.

Let us now return to the acoustic wave equation (1.5), and repeat the derivation of the weak form of the equation (2.5), but with the test function φ selected in the discontinuous space $\mathcal{S}_h$. Integration by parts cannot be performed on the whole domain Ω this time, but we must split the integrals over all subdomains. We obtain this time

$$\sum_{i=1}^{n_d} \int_{\Omega_i} \frac{1}{\lambda} \varphi \frac{\partial^2 p}{\partial t^2} \,\mathrm{d}\Omega + \sum_{i=1}^{n_d} \int_{\Omega_i} \frac{1}{\rho} \nabla\varphi \cdot \nabla p \,\mathrm{d}\Omega - \sum_{i=1}^{n_d} \int_{\partial\Omega_i} \frac{1}{\rho} \varphi \nabla p \cdot \,\mathrm{d}S = \sum_{i=1}^{n_d} \int_{\Omega_i} \varphi s \,\mathrm{d}\Omega. \tag{2.23}$$

In the third sum (2.23), each internal interface F appears twice, once with the normal N_F, and once with the normal $-N_F$. After applying the usual boundary conditions (1.8) and (1.9) to p, the third term can thus be rewritten as

$$-\sum_{F\in\mathcal{F}_I} \int_F \llbracket \frac{1}{\rho} \varphi \nabla p \rrbracket \,\mathrm{d}F + \int_{\partial\Omega_A} \frac{1}{\rho c} \varphi \frac{\partial p}{\partial t} \,\mathrm{d}S,$$

and using (2.22), one can write

$$\sum_{F\in\mathcal{F}_I} \int_F \llbracket \frac{1}{\rho} \varphi \nabla p \rrbracket \,\mathrm{d}F = \sum_{F\in\mathcal{F}_I} \int_F \llbracket \varphi \rrbracket \cdot \{\!\{ \frac{1}{\rho} \nabla p \}\!\} \,\mathrm{d}F + \sum_{F\in\mathcal{F}_I} \int_F \{\!\{ \varphi \}\!\} \llbracket \frac{1}{\rho} \nabla p \rrbracket \,\mathrm{d}F. \tag{2.24}$$

Assuming that ρ can be represented by piecewise smooth functions, it can be shown (see, e.g., [124, Theorems 8.1 and 8.2]) that $\nabla \cdot (1/\rho \nabla p)$ must belong to $L^2(\Omega)$, and therefore $1/\rho \nabla p \in H(\mathrm{div})$, implying that $1/\rho \nabla p \cdot N_F$ is continuous across the interface F. Consequently, the last term in (2.24) is zero. Putting it all together, we obtain the following tentative weak form:

$$\begin{aligned}\sum_{i=1}^{n_d} \int_{\Omega_i} \frac{1}{\lambda} \varphi \frac{\partial^2 p}{\partial t^2} \,\mathrm{d}\Omega + \sum_{i=1}^{n_d} \int_{\Omega_i} \frac{1}{\rho} \nabla\varphi \cdot \nabla p \,\mathrm{d}\Omega - \sum_{F\in\mathcal{F}_I} \int_F \llbracket \varphi \rrbracket \cdot \{\!\{ \frac{1}{\rho} \nabla p \}\!\} \,\mathrm{d}F \\ + \int_{\partial\Omega_A} \frac{1}{\rho c} \varphi \frac{\partial p}{\partial t} \,\mathrm{d}S = \sum_{i=1}^{n_d} \int_{\Omega_i} \varphi s \,\mathrm{d}\Omega.\end{aligned} \tag{2.25}$$

The term containing the jump of φ is known as a *flux* term.

First and foremost, the bilinear form associated to (2.25) is not symmetric. Numerically, symmetric operators are often easier to treat. For example, when solving a linear system governed by a symmetric matrix, one can apply the Cholesky decomposition (or the LDL decomposition if the matrix is not positive definite) instead of the full LU decomposition, which is much less memory intensive. If instead one opts for an iterative solution via a Krylov subspace method,

then one can apply in the symmetric case the very efficient conjugate gradient algorithm, instead of less efficient and less stable methods such as the biconjugate gradient algorithm or the generalized minimal residual (GMRES) method. Preconditioning also usually becomes more complicated for nonsymmetric operators.

For all these reasons, we wish to restore the symmetry of (2.25). We can do so by adding an additional term, namely the symmetric of the flux term,

$$-\sum_{F\in\mathcal{F}_I}\int_F \{\!\{\frac{1}{\rho}\nabla\varphi\}\!\}\cdot[\![p]\!]\,\mathrm{d}F.$$

Notice that this term is zero on the non-discretized solution, so that the numerical scheme remains coherent.

Another fundamental issue of the bilinear form associated to (2.25) is that it does not satisfy the positivity condition (2.7). The culprits are the terms containing the jumps $[\![\varphi]\!]$ and $[\![p]\!]$, which invariably introduce a negative summand in the bilinear form, and thus prevent any form of positive (semi-)definiteness. This issue is resolved by adding a *penalty term*

$$\alpha\sum_{F\in\mathcal{F}_I}\int_F [\![\varphi]\!]\cdot[\![p]\!]\,\mathrm{d}F$$

for some penalty coefficient $\alpha > 0$. For the non-discretized solution, this term is always zero due to the cited continuity property of p. However, in the discrete problem, this addition bears many consequences. First of all, it is positive semi-definite, so for a high enough choice of the coefficient α, the nonnegativity of the form is restored. Second, note that α is a dimensionful quantity: it has the dimension of $1/\rho\cdot 1/\ell$, where ℓ is the unit of length. Therefore, some scaling with respect to the local parameters is required. Usually, the most penalizing choice is used (cf. [108]):

$$\alpha = \alpha_0\frac{\max(\frac{1}{\rho^+},\frac{1}{\rho^-})}{\min(h^+,h^-)},$$

where h is a typical size of the degrees of freedom across the interface. Which typical size must be taken is not an obvious choice, and has been the object of some research. For example, for Lagrange bases on simplicial meshes in two and three dimensions, [125, 126] have proposed lower bounds that depend on the polynomial degree and on the geometrical properties (the inradius or the angles) of the two neighboring cells. In Chapter 6, we show how the choice of [125] can also be applied to our unstructured spline spaces.

Finally, when the error norm associated to the bilinear form is computed, the presence of a term proportional to the square of the jump of the solution acts like a stabilizer for the method. A drawback is that, when some energy is stored in the corresponding term of the energy norm, the (physical) energy of the system is not exactly conserved, and the penalty term acts as a damping term. This effect can potentially introduce some inaccuracies. Thus, the coefficient α of the penalty term is the object of a trade-off between stability and accuracy.

We can now write the full weak *interior-penalty discontinuous Galerkin* (IPDG) formulation

for the acoustic wave equation,

$$\sum_{i=1}^{n_d}\int_{\Omega_i}\frac{1}{\lambda}\varphi\frac{\partial^2 p}{\partial t^2}\,\mathrm{d}\Omega+\sum_{i=1}^{n_d}\int_{\Omega_i}\frac{1}{\rho}\nabla\varphi\cdot\nabla p\,\mathrm{d}\Omega-\sum_{F\in\mathcal{F}_I}\int_F[\![\varphi]\!]\cdot\{\!\{\frac{1}{\rho}\nabla p\}\!\}\,\mathrm{d}F \tag{2.26}$$
$$-\sum_{F\in\mathcal{F}_I}\int_F\{\!\{\frac{1}{\rho}\nabla\varphi\}\!\}\cdot[\![p]\!]\,\mathrm{d}F+\alpha\sum_{F\in\mathcal{F}_I}\int_F[\![\varphi]\!]\cdot[\![p]\!]\,\mathrm{d}F+\int_{\partial\Omega_A}\frac{1}{\rho c}\varphi\frac{\partial p}{\partial t}\,\mathrm{d}S=\sum_{i=1}^{n_d}\int_{\Omega_i}\varphi s\,\mathrm{d}\Omega.$$

The bilinear form a_ρ of the continuous Galerkin approach thus gets replaced by

$$a_{\mathrm{DG}}(\varphi,p)\coloneqq\sum_{i=1}^{n_d}\int_{\Omega_i}\frac{1}{\rho}\nabla\varphi\cdot\nabla p\,\mathrm{d}\Omega-\sum_{F\in\mathcal{F}_I}\int_F[\![\varphi]\!]\cdot\{\!\{\frac{1}{\rho}\nabla p\}\!\}\,\mathrm{d}F \tag{2.27}$$
$$-\sum_{F\in\mathcal{F}_I}\int_F\{\!\{\frac{1}{\rho}\nabla\varphi\}\!\}\cdot[\![p]\!]\,\mathrm{d}F+\alpha\sum_{F\in\mathcal{F}_I}\int_F[\![\varphi]\!]\cdot[\![p]\!]\,\mathrm{d}F.$$

This version of the interior penalty DG method is called *symmetric*, since the flux terms have the same coefficient. Non-symmetric versions are possible (see, e.g., [127]), and they sometimes offer the added benefit of making the bilinear form a_{DG} nonnegative for all choices of the penalty parameter. However, error analysis in this case is much harder, and remains an open problem in many cases. We show in Figure 2.3 the typical shape of a solution to the discontinuous Galerkin problem.

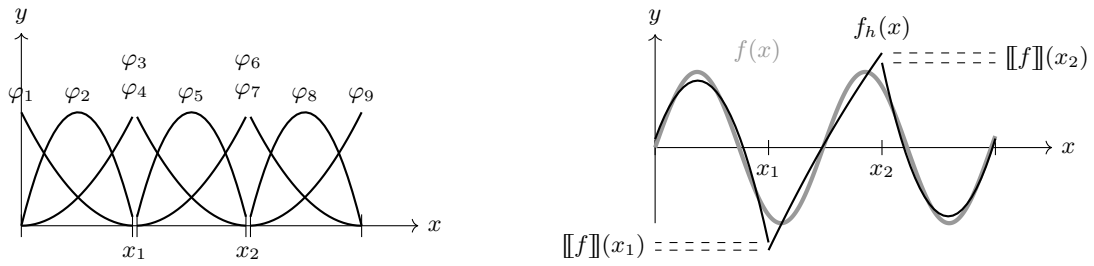


Figure 2.3: (Left) a typical piecewise-polynomial basis of degree 2 on three elements used for the discontinuous Galerkin method in one-dimension and (right) the L^2 projection of a function on this basis. Note that the projection is not continuous. The jump of the function is highlighted in the picture.

A priori error analysis of the IPDG method

After comparing (2.26) with (2.5), one can immediately see that the equivalent of the energy norm (2.8) in the case of the IPDG method is given by

$$|||p|||^2\coloneqq\sum_{i=1}^{n_d}\int_{\Omega_i}\frac{1}{\lambda}\left|\frac{\partial p}{\partial t}\right|^2\,\mathrm{d}\Omega+\sum_{i=1}^{n_d}\int_{\Omega_i}\frac{1}{\rho}\left|\nabla p\right|^2\,\mathrm{d}\Omega+\alpha\sum_{F\in\mathcal{F}_{\mathrm{DG}}}\int_F[\![p]\!]^2\,\mathrm{d}F. \tag{2.28}$$

The term involving the spatial derivatives is naturally formulated on the space of continuously differentiable functions on Ω, $H^1(\Omega)$, over which the last term is zero. Since the space is non-

conforming, in order to capture both the discretized and continuous (smooth) spaces, the error estimate must take place over the space

$$V \coloneqq H^1(\Omega) + \mathcal{S}_h.$$

We run immediately into a problem: the trace of ∇p is not guaranteed to exist for a general $p \in H^1(\Omega)$ (and thus in V), since the Sobolev embedding theorem would require a regularity of degree at least 2 to produce such a trace (see, e.g., [32, Chapter 5]). Consequently, the terms $\{\!\{\frac{1}{\rho}\nabla p\}\!\}$ and $\{\!\{\frac{1}{\rho}\nabla \varphi\}\!\}$ appearing in (2.26) are not generally well defined. The classical solution to this problem takes the form of a projection operator (see, e.g., [108])

$$\Pi_{\mathrm{DG}} : L^2(\Omega) \to \mathcal{S}_h,$$

where $\mathcal{S}_h$ denotes as before the non-conforming discretized space, and Π_{DG} projects orthogonally with respect to the usual L^2 norm. Notice that this has nothing to do with the trivial projection from $V = H^1(\Omega) + \mathcal{S}_h$ onto its second component, but it is a non-trivial projection consisting of finding, for a given function $g \in L^2(\Omega)$, the best-approximating function $\Pi_{\mathrm{DG}} g$ in $\mathcal{S}_h$ in the L^2 sense. This nontrivial projection operator then allows to *lift* the notion of average from $\mathcal{S}_h$ (where the traces are well-defined) to $L^2(\Omega)$, by defining, for $f \in H^1(\Omega)$,

$$\{\!\{\frac{1}{\rho}\nabla f\}\!\} \coloneqq \{\!\{\Pi_{\mathrm{DG}}\frac{1}{\rho}\nabla f\}\!\}.$$

With this lift, the energy norm $|||\cdot|||$ (2.28) and the lifted bilinear form $\hat{a}_{\mathrm{DG}}$, obtained from (2.27) by replacing $\{\!\{\frac{1}{\rho}\nabla f\}\!\}$ with $\{\!\{\Pi_{\mathrm{DG}}\frac{1}{\rho}\nabla f\}\!\}$, become compatible. One can then prove the nonnegativity and continuity of $\hat{a}_{\mathrm{DG}}$ in this norm (see, e.g., [108, 128]), provided that $\alpha > \alpha_{\min}$, with the lower bound depending on the details of the domain and its subdivision. We do not repeat the proofs here, as they are discussed in more detail in Chapter 6 for the DG-IGA multipatch scheme. However, we notice that the classical proofs of this fact are not immediately applicable in our slightly more general setting, since they are formulated on simplicial meshes, i.e., on subdivisions where each subdomain $\Omega_i =: \sigma_i$ is a simplex. With that assumption, one can rely on the *inverse inequality*

$$\|v\|^2_{L^2(\partial\sigma_i)} \leq C\frac{\mathrm{vol}^{d-1}(\partial\sigma_i)}{\mathrm{vol}^d(\sigma_i)}\|v\|^2_{L^2(\sigma_i)} \tag{2.29}$$

valid whenever v is a polynomial of degree k inside σ_i for some constant C depending only on the degree k and the dimension d. This inequality allows to transfer the integrals over internal facets in (2.26) to volume integrals, thus removing the issue of trace definition, and allowing to remove the projection Π_{DG} from the estimate. Even if in our work the shape of the subdomains Ω_i is not restricted to simplices, and in fact the subdomains are not even assumed to be convex, it is still possible to recover an inverse inequality for the multi-patch spline spaces on which we build our numerical scheme in Chapter 6. The inequality is indeed very similar to the case of simplicial subdomains. Consequently, we make here the assumption that this inequality is available even in our more general setting.

Once the nonnegativity and continuity of the bilinear form a_{DG} (2.26) have been established, one is immediately faced with another problem, as the lifted form $\hat{a}_{\mathrm{DG}}$ is no longer compatible with the non-discretized weak problem (2.26), since the discretization space is non-conforming. Specifically, suppose that p is a solution of the non-discretized weak problem (2.26), and suppose furthermore that $p \in H^2(\Omega)$, so that the trace $\nabla p|_F$ is well-defined on every facet F. Let $\varphi_h \in \mathcal{S}_h$ be a test function. Then the lifted form $\hat{a}_{\mathrm{DG}}(\varphi_h, p)$ contains the term

$$-\sum_{F\in\mathcal{F}_I}\int_F [\![\varphi_h]\!]\cdot \{\!\{\Pi_{\mathrm{DG}}\frac{1}{\rho}\nabla p\}\!\}\,\mathrm{d}F,$$

while $a_{\mathrm{DG}}(\varphi_h, p)$ contains

$$-\sum_{F\in\mathcal{F}_I}\int_F [\![\varphi_h]\!]\cdot \{\!\{\frac{1}{\rho}\nabla p\}\!\}\,\mathrm{d}F,$$

Notice that these two terms are indeed different, since $\nabla p \in H^1(\Omega)$, and its best-fitting function in $\mathcal{S}_h$ is not even continuous in general (cf. Figure 2.3). In order to apply the same error estimates as in the case of the continuous Galerkin method, one needs to define the elliptic projection using the lifted form $\hat{a}_{\mathrm{DG}}$. However, due to this incompatibility, when computing (2.11), one obtains an extra term in the form of the *residual*

$$r(p,\varphi_h) \coloneqq \hat{a}_{\mathrm{DG}}(\varphi_h,p) - a_{\mathrm{DG}}(\varphi_h,p) = -\sum_{F\in\mathcal{F}_I}\int_F [\![\varphi_h]\!]\cdot\{\!\{\Pi_{\mathrm{DG}}\frac{1}{\rho}\nabla p - \frac{1}{\rho}\nabla p\}\!\}\,\mathrm{d}F.$$

The error estimate can then proceed exactly as in the previous section. When deriving the final estimate for $\|p-p_h\|_{L^\infty([0,t_f],L^2(\Omega))}$, the additional term $r(p,\varphi_h)$ appearing in (2.9) gets divided by $\|p_h - w_h\|_{L^2(\Omega)}$. One can bound this additional term via

$$\frac{|r(p,p_h-w_h)|}{\|p_h-w_h\|_{L^2(\Omega)}} \leq \sup_{\eta_h\in\mathcal{S}_h}\frac{|r(p,\eta_h)|}{\|\eta_h\|_{L^2(\Omega)}}.$$

Fortunately, this does not hinder the order of convergence of the method, since it turns out to be of the same order of approximation as the error. To see this, notice that

$$|r(p,\eta_h)|^2 \leq \left(\sum_{F\in\mathcal{F}_I}\int_F \alpha[\![\eta_h]\!]^2\,\mathrm{d}F\right)\left(\sum_{F\in\mathcal{F}_I}\int_F \alpha^{-1}\left|\{\!\{\Pi_{\mathrm{DG}}\frac{1}{\rho}\nabla p - \frac{1}{\rho}\nabla p\}\!\}\right|^2\,\mathrm{d}F\right).$$

Using the inverse inequalities seen above, both integrals can be bounded by the L^2 norm of their arguments on $L^2(\Omega)$, and after noticing that the projection Π_{DG} is stable in the L^2 norm, the projection Π_{DG} can be removed from the integral (see e.g. [108, Lemma 4.6] and [108, Lemma 4.7] or also [109, Lemma 1.3.9]).

We do not repeat here the details of this derivation, which is quite standard. We do however remark that two crucial assumptions rely on the fact that the domains Ω_i are simplices: the determination of the minimum penalization coefficient α that restores positivity, and the inverse inequality (2.29). These two properties need to be recovered in the case of our multi-patch DG-IGA scheme, as we do in Chapter 6.

Finally, we do not present here the *a posteriori* error analysis of the IPDG method, and we instead refer the reader to, e.g., [108]. However, we wish to emphasize that the IPDG method, as presented here, retains the same order of approximation of the FE method of the same degree, even with a non-conforming discretization space. We also point out once again the fundamental role that the polynomial reproduction property plays in the computation of error estimates of virtually all Galerkin methods, and therefore in proving their accuracy.

Discrete form for explicit timestepping via the IPDG method

We derive now the explicit form of (2.26) after selecting, as before, a basis $(\varphi_i)_{i=1}^n$ of $\mathcal{S}_h$, and after having chosen a second-order leapfrog timestepping scheme. Comparing (2.26) with (2.6), it is clear that the form of the algebraic equation that one obtains after discretization is very similar. In fact, one obtains

$$\gamma^{(t+1)} = \left(M + \frac{\Delta t}{2}B\right)^{-1}\left(2M\gamma^{(t)} - M\gamma^{(t-1)} - \Delta t^2 K'\gamma^{(t)} + \frac{\Delta t}{2}B\gamma^{(t-1)}\right) + \Delta t^2 S(t),$$

where the modified stiffness matrix

$$K' \coloneqq K + F + P$$

now incorporates the flux and penalty terms of (2.26). The corresponding matrices are given by

$$F_{i,j} \coloneqq -\sum_{F\in\mathcal{F}_I}\int_F \left(\llbracket\varphi_i\rrbracket \cdot \{\!\{\frac{1}{\rho}\nabla\varphi_j\}\!\} + \{\!\{\frac{1}{\rho}\nabla\varphi_j\}\!\} \cdot \llbracket\varphi_j\rrbracket\right) \mathrm{d}F \tag{2.30}$$

and

$$P_{i,j} \coloneqq \alpha_{i,j}\sum_{F\in\mathcal{F}_I}\int_F \llbracket\varphi_i\rrbracket \cdot \llbracket\varphi_j\rrbracket \,\mathrm{d}F. \tag{2.31}$$

For piecewise constant values of ρ, the penalty parameter is often chosen to be

$$\alpha_{i,j} \coloneqq \alpha_0 \frac{\max(\frac{1}{\rho_i}, \frac{1}{\rho_j})}{\min(h_i, h_j)},$$

where $1/\rho_i$, $1/\rho_j$ are the physical parameters in domains Ω_i and Ω_j respectively, and h_i, h_j are two typical dimensions of the two neighboring functions φ_i, φ_j. As discussed above, in the case of simplicial meshes, h_i and h_j are tied to the geometrical properties of the two neighboring simplices supporting the two basis functions, usually their inradius or some other quantity involving their the angles, see, e.g., [125, 126]. We derive an equivalent quantity for unstructured splines in Chapter 6. As for α_0, a commonly chosen value is [89]

$$\alpha_0 = \binom{k+d}{d},$$

where k is the polynomial order of the (polynomial-reproducing) basis, and d is the number of space dimensions. Of course, the goal of the coefficient α_0 is to make the bilinear form (2.26)

nonnegative, and the simulation stable. Therefore, more penalizing values are acceptable, even though they risk making the simulation less accurate due to their damping effect.

Notice that the matrices F and P only involve functions that are supported on the interface between two domains. In the classic mesh-based DG method, this is the case for all basis functions φ_i. However, in view of the DG-IGA scheme that we wish to formulate, the domains Ω_i are usually larger, and they include many functions that are not supported on the internal interfaces, and for which the matrix elements (2.30) and (2.31) are zero. Thus, from a computational standpoint, these matrices are better computed through a loop over the facets $\mathcal{F}_I$. One can also perform a loop on all the internal and external boundary facets of the domains, and compute the damping matrix (2.17) at the same time.

Finally, but far more importantly, we need to emphasize an extremely interesting feature that has made DG methods very successful for transport problems with explicit timestepping. In fact, only functions supported on the same domain Ω_i give a nonzero contribution to the computation of the mass matrix (2.15) (as well as the unmodified stiffness matrix (2.16) and the damping matrix (2.17)). Consequently, these matrices are *block-diagonal*, with one block per subdomain. The computation of the inverse M^{-1} (and $(M + \Delta t/2\, B)^{-1}$) is thus made significantly less expensive. This represents a tremendous numerical advantage, both for iterative and factorization (direct) approaches to matrix inversion. Moreover, the matrix K' only relates functions belonging to different domains by the flux and penalty terms. This has a big relevance in code parallelization, since different tasks dispatched to different nodes only need to communicate a very limited amount of information during a time iteration. Moreover, in usual DG formulations on simplicial meshes, every domain Ω_i has the same shape, and all elements are related by an affinity transformation. Consequently, all the diagonal blocks of these matrices are related by a simple Jacobian, and do not need to be computed independently. We will see that all these advantages, excluding the last one, are retained in our DG-IGA formulation.

2.3 Discretization for the inverse problem

We spend here a few words about two topics that are relevant when performing seismic inversion, namely the choice of discretization for the physical parameters ρ and λ, and the choice of algorithm for the minimization of the cost function.

2.3.1 Discretization of the space of physical parameters

In the previous section, we have seen that the convex dual of a constrained optimization problem can be formulated in terms of the *residuals*. In the case of full waveform inversion, for a given detector datum $d_r(t)$, the residual of a given solution $\overline{p}(x,t)$ reads

$$\varepsilon(t) \coloneqq \overline{p}(x_r,t) - d_r(t).$$

It is a well-known property of L^2 minimization that the residuals must be orthogonal to the image of the time-evolution operator (1.11), i.e., they must be orthogonal to all the solutions to the acoustic wave problem for all possible choices of physical parameters. Working in Fourier space, with a pointlike source and after evaluating the solution at $x = x_r$, the time-evolution

convolution (1.11) becomes a multiplication, and the orthogonality condition $\langle \varepsilon, p \rangle = 0$ then becomes

$$\int_{-\infty}^{+\infty} \overline{\widetilde{\varepsilon}(\omega)} \widetilde{s}(\omega) G_{\rho,\lambda}(\omega) \, \mathrm{d}\omega = 0 \text{ for all Green's functions } G_{\rho,\lambda}(\omega), \tag{2.32}$$

where $\widetilde{\varepsilon}(\omega)$, $\widetilde{s}(\omega)$ are the Fourier transforms of $\varepsilon(t)$ and the source $s(t)$, respectively, and $G_{\rho,\lambda}$ is the Green's function taking source data to receiver data. Thus, the product (2.32) can only be zero if $\widetilde{\varepsilon}(\omega)$ and $\widetilde{s}(\omega) G_{\rho,\lambda}(\omega)$ have disjoint supports for all allowable ρ, λ.

We can restate this conclusion by saying that the residual can be zero (and thus the inversion can be "perfect") only if two conditions are met: the spectrum of the source term contains all the frequencies present in the measurements $d_r(t)$, and these frequencies are not suppressed by the choice of allowable physical parameters. Notice that, if s is a compactly-supported wavelet, its Fourier transform is an entire function, and therefore it only vanishes over a measure-zero set. Therefore, the choice of discretization of physical parameters plays a large role in the successful minimization of the residuals in the inverse problem, since the residual is allowed to contain all the frequencies that are suppressed by $G_{\rho,\lambda}(\omega)$. For example, the choice of piecewise-constant values on a very coarse mesh suppresses the higher frequencies by aliasing, which can therefore be found in the residual.

The inverse problem therefore poses some constraints on the coarseness of the spatial (and temporal) discretization that go beyond the pure numerical stability criteria discussed earlier. These considerations, and many related ones, have been studied in more detail by many authors. We refer the reader to [40], and references therein, for further reading.

Concretely, the most common choice for the space of physical parameters, used in the majority of FWI instances, is a set of piecewise-smooth (and most often piecewise-constant) functions defined over the same mesh as the solution. While this is undoubtedly a convenient solution, there are some cases where this choice might be sub-optimal. In fact, there is no *a priori* reason why the typical variations in size of the physical model, which are tied to the typical size of sedimentary layers, should be correlated with the typical scale of variation of the source, which is tied to the source frequency. Since the number of degrees of freedom on which the inversion is performed is very adversely correlated with the convergence of the inversion process itself [9], one should try to avoid introducing too many unnecessary degrees of freedom.

However, it is impossible to know beforehand which degrees of freedom are useful for describing the shape of the model, and which are instead redundant. For this reason, some authors (see, e.g., [57, 58]) rely on *adaptive meshing* to adjust the shape and size of the mesh during the inversion process. This is a very promising technique, which presents however some difficulties as it is nontrivial to generate a model evolution that retains an acceptable shape of the mesh elements. As we have seen before, the geometric properties of mesh cells (such as their inradius) enter many error estimates, and thus the introduction of degenerated elements may adversely impact the numerical properties of the method.

An approach based on unstructured splines might help circumvent some of these difficulties. These functions are in fact formulated naturally on (unstructured) point clouds, and thus are inherently robust to shape modification. Furthermore, degenerate configurations of the points (e.g., collinear or coplanar) induce a loss of continuity of the basis, which may be a desirable feature when reconstructing the position of sharp reflectors in the subsurface. We shall briefly

discuss these topics in Chapter 7.

2.3.2 Minimization of the cost function

We have shown in Chapter 1 how the adjoint state method can be used to compute a subgradient of the cost function with respect to the model parameters. This is the first step in many local optimization algorithms, since a subgradient gives a direction along which the function is (at least locally) nonincreasing.

Suppose that we wish to minimize a cost function $f : \mathbb{R}^d \mapsto \mathbb{R}$, locally, around a point $\overline{x}$. One might use the Newton method, which is among the most widely used techniques for local minimization. Suppose that f is twice differentiable at $\overline{x}$. Then the Taylor expansion of f gives, at order 2,

$$f(x) = f(\overline{x}) + \nabla f(\overline{x}) \cdot (x - \overline{x}) + (x - \overline{x})^T H_f(\overline{x})(x - \overline{x}) + O(|x - \overline{x}|^3),$$

where $H_f(\overline{x})$ is the *Hessian matrix*

$$(H_f)_{ij}(\overline{x}) \coloneqq \frac{\partial^2 f}{\partial x_i \partial x_j}(\overline{x}).$$

If f were a quadratic function, the minimum could be found with a single *Newton iteration*

$$\overline{x}_{k+1} = x_k - H_f(x_k)^{-1} \nabla f(x_k). \tag{2.33}$$

An algorithm that follows Newton iterations converges quadratically for a strongly convex function f with Lipschitz-continuous Hessian [50]. However, one might risk overstepping the local minimum with a single iteration. For this reason, a step size $\gamma \leq 1$ is chosen for (2.33), i.e.,

$$x_{k+1} = x_k + \gamma \delta_k,$$

where $\delta_k = -H_f(x_k)^{-1} \nabla f(x_k)$ is the descent direction. The step γ must be chosen to satisfy the *Armijo-Wolfe conditions*, [129, 130]

$$f(x_k + \gamma \delta_k) \leq C_1 \gamma \delta_k \cdot \nabla f(x_k), \tag{2.34a}$$
$$-\delta_k \cdot \nabla f(x_k + \gamma \delta_k) \leq -C_2 \delta_k \cdot \nabla f(x_k), \tag{2.34b}$$

for some constants $0 < C_1 < C_2 < 1$, which ensure that f and ∇f are both sufficiently decreased by the iteration.

The full Hessian matrix, let alone its inverse, is too computationally expensive to be computed in most practical cases. Thus, one usually relies on algorithms that reconstruct the second-order information from the available history of first-order information, i.e., the gradients at previous iterations. The most common of these *quasi-Newton* algorithms is the BFGS method, from the name of its inventors, and its limited-memory incarnation L-BFGS (see, e.g., [131]). Using this method, the conjugate direction $\delta_k = -H_f(x_k)^{-1} \nabla f(x_k)$ is approximated iteratively as $\delta_1 = -\nabla f(x_1)$, $\delta_i = -\nabla f(x_i) + \beta_i \nabla f(x_{i-1})$ for $i > 1$. Many choices for β_i are

possible, see, e.g., [49, 50].

In the case of full waveform inversion, the space of physical parameters is so large that in most cases even these optimized algorithms can become computationally very expensive. The scientific debate on whether the inclusion of second-order information makes up for the increase in the computational cost by sufficiently improving the result or the convergence rate of the reconstruction seems to be far from settled [40, 47, 48]. We will show that in the case of unstructured spline functions, all the second-order derivatives needed for the computation of the Hessian are available. Nevertheless, we will only use first-order information, i.e., the pure gradient direction $\delta_k = -\nabla f(x_k)$, in most of our simulations.

Finally, notice that for a general non-differentiable function f, the gradient $\delta_k = -\nabla f(x_k)$ can be replaced by any subgradient $\delta_k \in \partial f(x_k)$. Second-order information, for example in the form of sub-hessians, can be included in a somewhat more subtle way, see, e.g., [132] for the formulation of quasi-Newton optimization over nonsmooth functions.

Once a descent direction δ_k has been selected, one simply has to select a step γ along δ_k such that the Armijo-Wolfe conditions (2.34) are satisfied. Notice that (2.34b) requires the computation of $\nabla f(x_k + \gamma\delta_k)$, which is highly impractical for FWI since it would require a very large number of direct and backward propagations per iteration. Fortunately, the Armijo-Wolfe conditions (2.34) can be replaced by the equivalent conditions (see, e.g., [45])

$$f(x_k) + (1-C)\gamma\delta_k \cdot \nabla f(x_k) \leq f(x_k + \gamma\delta_k) \leq f(x_k) + C\gamma\delta_k \cdot \nabla f(x_k) \tag{2.35}$$

for some constant $0 < C < 1$, which are known as the *Armijo-Goldstein conditions*. The issue with (2.35) is that they are satisfied by any sufficiently small step γ. In order to avoid this pitfall, the selection of an appropriate step is usually done by *backtracking*, i.e., by starting with a large step γ_0 and decreasing it until the conditions (2.35) are met.

2.4 Discussion and further reading

We have presented in this section some of the most common discretization techniques for hyperbolic equations, both in time and space. For the time discretization, only the basic explicit timestepping techniques that are followed in this work have been presented in any detail. No detail of implicit or even hybrid implicit-explicit methods has been given. This is due to the fact that time discretization is not one of the main focal points of this work. The literature on time integration is of course very large, as this is a topic of real importance in all time-dependent PDEs and also for statistical ensembles (for example, the leapfrog scheme is known as the Verlet integrator in molecular dynamics). Nevertheless, the reader is referred to the classic texts of numerical analysis and ODE integration such as [133, 134] for a more general overview.

One aspect that we have not covered here is local timestepping. In fact, thanks to the freedom offered by the non-conforming bases of discontinuous Galerkin methods, one can avoid a global bottleneck on the timestep by using a different timestep for every subdomain. Choosing timesteps that are multiples of a common smallest value helps with synchronization. This technique can be applied both with an explicit timestepping scheme [83, 135–137] and with an implicit one [138, 139].

Concerning finite volumes and finite differences, since these schemes are not the focus of our work, we will only include here the reference [96], which is a standard text for the finite volumes method applied to hyperbolic problems, and the relatively recent text [93] which presents a modern and hands-on approach to finite difference methods.

Regarding the Galerkin method, we have presented here only the basic formulation and *a priori* error analysis. Our objective is simply to pave the way for the formulation of the IPDG method, that shares of course many traits with the standard Galerkin approach, and emphasize the important role played by polynomials and polynomial-reproducing bases in the asymptotic behavior of these methods via the classic estimates. The Galerkin approach underpins much of modern numerical analysis, and for this reason we refer the reader to one of the many comprehensive books on the subject, such as, e.g., [140], for a more detailed treatment of the subject. Furthermore, isogeometric analysis (IGA) is a relatively recent incarnation of FE methods based on B-spline function bases, that we will treat in more detail in Chapter 3. For this technique, we refer to the classic text [2].

Not many resources on the direct application of FE methods to hyperbolic equations can be found, since these methods are usually better suited for self-adjoint operators; we cite here [141]. *A posteriori* error analysis for FE methods, which we have skipped here, can be found in many texts, including, e.g., the detailed and well-known monograph [142].

For the discontinuous Galerkin method, aside from the classical texts [105, 107, 143], we can cite here the seminal papers [99, 100] that introduced the use of DG methods for neutron transport problems, and [101], [144–147] that helped establish it as a prominent technique for hyperbolic problems more in general. We would also like to refer the novice reader to [106], which provides a very accessible introduction to DG methods for elliptic and hyperbolic problems. Moving closer to the focus of this work, [108] contains the formulation of the symmetric interior penalty Galerkin method (including *a priori* and some *a posteriori* error bounds) for the second order wave equation on a mesh, which is very close to the formulation that we adopt here, the only difference being the absorbing boundary conditions, the choice of more general domains and the adoption of splines instead of Lagrange or other standard polynomial bases.

As mentioned in the text, adding a second flux term in (2.26), with the same sign as the first one, leads to the Symmetric Interior Penalty (SIPG, or here simply IPDG) method. However, one can also add a flux term with the opposite sign, obtaining a Nonsymmetric Interior Penalty (NIPG) method [127, 148], or not add any new flux term at all, yielding the Incomplete Interior Penalty (IIPG) method [149, 150]. These two variants have different flavors than the symmetric version, and borrow the concept of upwind fluxes from classic finite volume schemes.

The wave equation can equivalently be formulated in the frequency domain, leading to the standard Helmholtz equation. Even though the numerical advantage tied to the inversion of the mass matrix is less important in this case, discontinuous Galerkin approaches can nonetheless be exploited very efficiently in the frequency domain, for example through the Hybridizable Discontinuous Galerkin framework. We refer the reader to [84, 151] for the forward problem and to [40] for the inversion problem.

Concerning the inverse problem, some relevant references for the adaptive remeshing technique and the various Newton, quasi-Newton and line search methods have been given in the main text of the last two subsections.

3 | Piecewise polynomial approximation and splines

> [...] erano notizie discontinue, disarmoniche, inessenziali, dalle quali non risultava il nesso tra le mie azioni, e una nuova azione non riusciva a spiegare o a correggere l'altra, cosicché esse restavano addizionate l'una all'altra, con segno positivo o negativo, come in un lunghissimo polinomio che non è possibile ridurre a un'espressione più semplice.
>
> *Italo Calvino, Le Cosmicomiche, Gli anni-luce (1965)*

As we have explored in the previous chapter, error estimates for weak formulations based on the (continuous or discontinuous) Galerkin method rely crucially on inequalities, such as Jackson's inequality, involving polynomials. Specifically, polynomial-reproducing bases are often the starting point for the refinement of general *a priori* estimates, as shown in the previous chapter, and are even more important in *a posteriori* estimates, where the choice of degrees of freedom, and in particular the shape of the underlying discretization geometry, plays an important role. In many regards, the very notion of *order of approximation*, expressed through Jackson-type inequalities, relies on polynomial approximation.

Aside from the order of convergence, another important notion is that of uniform convergence. In this regard, polynomials are *universal*, i.e., they are dense in the space of continuous functions by the Stone-Weierstrass theorem [113]. Thus, polynomial functions of increasing degree provide uniform convergence (i.e., convergence in the L^∞ norm) for continuous solutions, such as the pressure in the acoustic wave equation with piecewise-smooth coefficients. However, if the choice is restricted to a subset of polynomials (e.g., the subset of interpolating polynomials at a given, fixed set of points), the rate of uniform convergence, and even whether this convergence is achieved at all, is all but guaranteed: successive approximations of a function, even regular, may exhibit oscillations that increase with the order of the polynomial. This is the well known *Runge's phenomenon* [152] (see Figure 3.1).

Lastly, but not less importantly, polynomials are computationally efficient to evaluate numerically, as they involve only a limited number of basic algebraic operations, and can be evaluated to machine precision if required. Many specific bases of polynomials (such as Bernstein-Bézier polynomials discussed below) possess efficient evaluation schemes capable of evaluating all the polynomials in a basis while reusing many intermediate results. Another numerical advantage

appears when computing superposition matrices such as the mass matrix (2.15) or the stiffness matrix (2.16) of the previous section, for which there exist efficient *quadrature rules* that allow the computation of these integrals via a weighted sum of the values of the polynomial at some specific points.

For these reasons, the theory of discretized weak forms of PDEs, as formulated through the Galerkin method, inevitably intersects with the theory of polynomial approximation. This chapter is dedicated to a brief introduction to the main piecewise-polynomial bases used in numerical analysis, with the goal of providing a quick overview of the subject, and also help place into context the spaces of *smooth* piecewise polynomial functions, a.k.a. B-splines, which are the main actors in isogeometric analysis [2] and in the remainder of this work. In the last part of the chapter, we use the properties of B-spline bases to compute analytically the CFL timestep condition for a one-dimensional infinite and homogeneous medium over cardinal B-splines, showing analytically in this simplified setting a well-known advantage of these functions for explicit timestepping.

3.1 Piecewise-polynomial approximations

The presentation of this chapter somehow follows the logic of the very complete text [153], to which we refer for a very complete analysis of polynomial and spline interpolation, spline properties and spline approximation theory. We strive to present the subject in a self-contained manner, and we will detach ourselves from the reference text wherever needed to emphasize the properties and present some additional considerations that are important in finite element and isogeometric analysis, especially for the wave equation.

3.1.1 Nodal polynomials

One of the basic challenges that a Galerkin basis must overcome is the ability to interpolate the values of the solution in the integration domain. Consider two points p_1 and p_2 in a vector space V. The values of $p \in V$ might be points in Euclidean space, representing the positions of some locations of our simulation domain, or they can be values of variables such as pressure or density. A simple linear interpolation of the two points can be obtained as

$$p(\lambda) \coloneqq \lambda p_1 + (1-\lambda)p_2 \text{ with } 0 \leq \lambda \leq 1. \tag{3.1}$$

As λ goes from 0 to 1, the point $p(\lambda)$ traces out a segment between p_1 and p_2. Notice that, since the two coefficients λ and $1-\lambda$ are nonnegative and sum to one, the combination is *convex*. Convexity of the interpolation is a very important property since it implies that the value of the interpolant never surpasses the minimum or maximum value of the interpolated variables, which generally leads to more stable numerical schemes, preventing the aforementioned Runge's phenomenon. The most general *convex combination* of n points p_i can be defined as:

$$p(\lambda_1, \ldots, \lambda_n) \coloneqq \sum_{i=1}^{n} \lambda_i p_i \text{ with } 0 \leq \lambda_i \leq 1 \text{ and } \sum_{i=1}^{n} \lambda_i = 1,$$

in which case the interpolating variable is confined inside the *convex hull* of the interpolated variables.

We can generalize the linear interpolation (3.1) by changing the parameterization of the interpolating segment. Let $x \in [a_1, a_2]$ be the parametric space in $\mathbb{R}$ that we wish to map onto the segment. The corresponding parametric segment is:

$$\gamma(x) = \frac{a_2 - x}{a_2 - a_1} p_1 + \frac{x - a_1}{a_2 - a_1} p_2. \tag{3.2}$$

We can interpret this segment as a simple immersion of a 1-dimensional simplex (a segment) $\gamma : [a_1, a_2] \to \mathcal{V}$, with a constant *Jacobian* given by:

$$|\,\mathrm{d}\gamma| = \left|\frac{\mathrm{d}\gamma}{\mathrm{d}x}\right| = \frac{|p_2 - p_1|}{a_2 - a_1}.$$

This point of view will become more useful in the next chapter. Suppose now that we wish to interpolate three points p_1, p_2 and $p_3 \in V$ with a curve. Perhaps the simplest way to achieve this is to take two linear interpolations $\gamma_1(x)$, $\gamma_2(x)$ between successive couples of points p_1, p_2 and p_2, p_3 (3.2), with corresponding parameters a_1, a_2 and a_2, a_3, and then perform another linear interpolation between the two segments. The resulting curve reads:

$$\gamma(x) = \frac{a_3 - x}{a_3 - a_1} \gamma_1(x) + \frac{x - a_1}{a_3 - a_1} \gamma_2(x). \tag{3.3}$$

It is easy to check that $\gamma(a_1) = \gamma_1(a_1) = p_1$ and $\gamma(a_3) = \gamma_2(a_3) = p_3$. Moreover:

$$\gamma(a_2) = \frac{a_3 - a_2}{a_3 - a_1} \gamma_1(a_2) + \frac{a_2 - a_1}{a_3 - a_1} \gamma_2(a_2) = \frac{a_3 - a_2 + a_2 - a_1}{a_3 - a_1} p_2 = p_2.$$

Therefore, the curve that we have obtained is *interpolating*, i.e., it passes through the three given points. It is instructive to compute explicitly the form of the three interpolation coefficients:

$$\gamma(x) = \frac{(x - a_2)(x - a_3)}{(a_1 - a_2)(a_1 - a_3)} p_1 + \frac{(x - a_1)(x - a_3)}{(a_2 - a_1)(a_2 - a_3)} p_2 + \frac{(x - a_1)(x - a_2)}{(a_3 - a_1)(a_3 - a_2)} p_3. \tag{3.4}$$

From this expression, it is easy to check that the curve is now a quadratic curve, whose shape depends on the choice of the three parameters a_1, a_2 and a_3. Notice however that the interpolation is *not convex*, since for example the coefficient of p_1 is negative for $x \in [a_2, a_3]$. This is the cause of Runge's phenomenon when higher order interpolations are computed (see Figure 3.1, right).

We can generalize the previous interpolation to $k+1$ points $p_1, \ldots, p_{k+1} \in V$, given the $k+1$ real parameters $a_1 < \ldots < a_{k+1}$. The interpolation we obtain is of the form

$$\gamma(x) = \sum_{i=1}^{k+1} \ell_{i,k}(x) p_i,$$

with the $k+1$ functions $(\ell_{i,k}(x))_{i=1}^{k+1}$ given by the *Lagrange polynomials*,

$$\ell_{i,k}(x) = \prod_{\substack{j=1 \\ j\neq i}}^{k+1} \frac{x - a_j}{a_i - a_j}.$$

It can be easily checked that the resulting polynomial interpolation $\gamma(x)$ has degree k and it is in fact the unique polynomial of degree k such that, given a set of distinct values $(a_1, \ldots, a_{k+1}) =: A$, i.e., a *grid*, satisfies $\gamma(a_i) = p_i$ for $k = 1, \ldots, k+1$. Notice also that the logarithmic derivative of $\ell_{i,k}(x)$ can be easily expressed,

$$\frac{\ell'_{i,k}(x)}{\ell_{i,k}(x)} = \sum_{\substack{j=1 \\ j\neq i}}^{k+1} \frac{1}{x - a_j}.$$

One of the most important properties of Lagrange polynomials is that they are *interpolating*, i.e., $\ell_{i,k}(a_j) = \delta_{ij}$. This means that, if we are given a function $f(x)$ to interpolate over the grid A, it is very easy to determine the coefficients of the Lagrange interpolation for any degree k,

$$f(x) = \sum_{i=1}^{k+1} f(a_i)\ell_{i,k}(x) + R(x), \tag{3.5}$$

where the remainder can be bounded uniformly by

$$|R(x)| \leq \frac{(a_{k+1} - a_1)^{(k+1)}}{(k+1)!} \max_{a_1 \leq x \leq a_{k+1}} \left| f^{(k+1)}(x) \right|. \tag{3.6}$$

Plugging the constant function $f(x) \coloneqq 1$ into (3.5), for which the remainder (3.6) is zero, shows that

$$\sum_{i=1}^{k+1} \ell_{i,k}(x) = 1$$

for all x, i.e., the sum of the Lagrange polynomials is the constant polynomial equal to one. Some low-degree Lagrange polynomials forming the Lagrange basis, as well as an example of interpolation of a Ricker wavelet (1.6) by Lagrange polynomials, are shown in Figure 3.1.

We have seen that, for any given choice of function f and grid A, the Lagrange polynomial is the unique degree-k polynomial interpolating f on a given set of points, $q(a_i) = f(p_i)$. However, the choice of the grid A is not determined *a priori*, and any other choice of distinct grid A' with $a'_1 < \ldots < a'_{k+1}$ determines an interpolating polynomial. It is therefore legitimate to ask which grid provides the best interpolation for the function f over the whole interval.

For simplicity, let the interpolation interval be $\Omega \coloneqq [-1, 1]$, and let $A \subset \Omega$ be a grid contained in the interval. Let $\hat{f}_k^A$ be the Lagrange polynomial of degree k that interpolates f on A. It can be proven that, if $f \in C^0(\Omega)$,

$$\left\| f - \hat{f}_k^A \right\|_\infty \leq (1 + \Lambda_k(A)) \left\| f - \overline{f} \right\|_\infty,$$

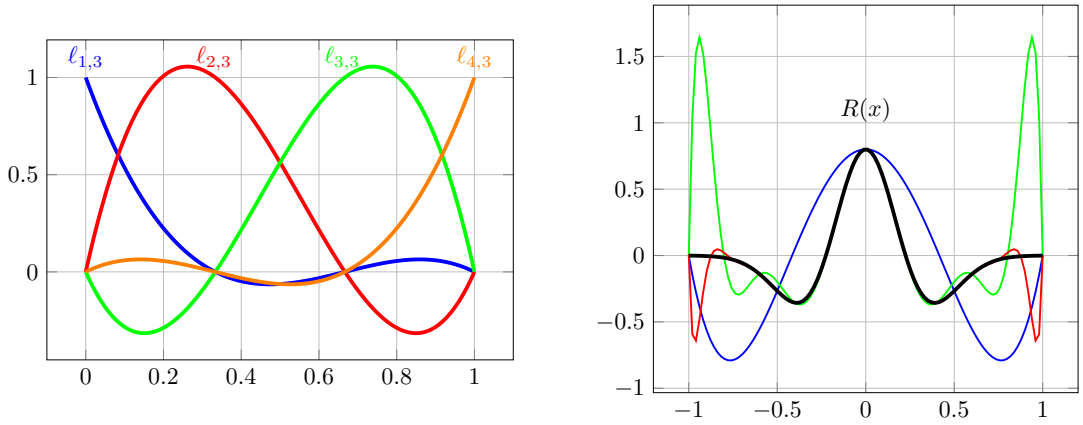


Figure 3.1: (Left) Cubic Lagrange basis functions on $[0, 1]$. (Right) Interpolation of a Ricker wavelet with Lagrange polynomials of order $k = 4$ (blue), 10 (green), 14 (red) over a uniform grid. The Runge phenomenon is well visible near the endpoints of the interval.

where $\overline{f}$ is the best interpolating polynomial in $\mathcal{Q}_k$ in the sense of the maximum norm $\|\cdot\|_\infty$ on Ω, and the *Lebesgue constant*

$$\Lambda_k(A) = \max_{x\in\Omega} \sum_{i=1}^{k+1} \left|\ell_{i,k}^A(x)\right|$$

depends on the choice of the grid A. Intuitively, the more the polynomials $\ell_{i,k}$ oscillate, the less uniform the approximation is. For a uniformly spaced grid, it can be proven that

$$\Lambda_k^{\text{unif}} = O\left(\frac{2^k}{k \log k}\right) \text{ for } k \to \infty,$$

see, e.g., [154] and also [155] for an improved asymptotic expansion. Therefore, there are some functions that cannot be interpolated uniformly with an equally spaced grid of parameters, since the interpolation becomes exponentially worse as the interpolation degree increases.

In general, better Lebesgue constants can be achieved by choosing appropriate grids. One common choice that gives generally good results is to place the nodes a_i in correspondence with the zeros of some family of orthogonal polynomials on Ω. Choosing the zeros of the Legendre polynomial $L_{k+1}(x)$ on the interpolation interval yields the so-called *Gauss-Legendre* nodes, for which the Lebesgue constant only grows asymptotically as

$$\Lambda_k^{\text{GL}} = O\left(\sqrt{k}\right) \text{ for } k \to \infty,$$

see, e.g., [156], whereas choosing the zeros of the *Chebyshev* polynomials $T_{k+1}(x)$ leads to the *Chebyshev-Gauss* nodes, which achieve the best possible asymptotic uniform interpolation, namely

$$\Lambda_k^{\text{CG}} = O\left(\log(k)\right) \text{ for } k \to \infty,$$

see, e.g., [157].

There is however an important drawback with using unequally spaced grids A, which becomes

relevant when solving the time-dependent wave equation. As we have seen in Chapter 2, given the causality of the wave equation, when one computes the value of the solution for a given node a_i at time $t+\Delta t$, starting from the knowledge of the solution at a previous time t, one must take into account the previous values at all the nodes that are able to influence the value of a_i during the timestep Δt, namely, those at a distance of $c\Delta t$ or less from a_i. If this condition is not satisfied, the construction of the new state at time $t+\Delta t$ is missing some important information and it is most likely not accurate. After selecting a time integration scheme, this leads directly to a condition on the largest allowable time step Δt. For a uniform grid, as the degree k of the polynomial increases, the spacing between nodes in a given interval of length h scales like h/k, and therefore the maximum allowable time step scales like $\Delta t_{\min} \sim h/k$. However, zeros of orthogonal polynomials, such as Legendre and Chebyshev polynomials tend to cluster together close to the endpoints of the interval. In fact, if one defines $a_i = \cos\theta_i$, then it can be proven that for any weight function on the interval $[-1,1]$ bounded from zero, the spacing between nearby zeros of the corresponding orthogonal polynomial of degree k satisfies [158, Theorem 6.11.1]

$$\theta_{i+1} - \theta_i < C\frac{\log k}{k}$$

for some constant C. If the weight function is bounded on the closed interval, the finer result [158, Equation 6.11.15]

$$\frac{C_1}{k} < \theta_{i+1} - \theta_i \leq \frac{C_2}{k}$$

holds for some constants C_1, C_2. In other words, the zeros tend to cluster together near the ends of the endpoints of the interval with a spacing that scales like h/k^2. Consequently, the much more stringent condition $\Delta t_{\min} \sim h/k^2$ must be imposed on the simulation time step, as dictated by the CFL stability condition, see, e.g., [159]. This is a well-known scaling property of FE simulations for hyperbolic systems, that is inherited also by discontinuous Galerkin simulations. The zeros of the uniform, Gauss-Legendre and Chebyshev-Gauss grids are shown in Figure 3.2.

All these considerations can be transferred easily to the d-variate setting, $d > 1$. In fact, Lagrange interpolation over a set of points $(a_i)_{i=1}^n \subset \mathbb{R}^d$, $n = \binom{k+d}{d}$, can simply be obtained by building the $n \times n$ *Vandermonde matrix*

$$V \coloneqq \begin{pmatrix} 1 & (a_1)^{I_1} & \dots & (a_1)^{I_k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (a_n)^{I_1} & \dots & (a_n)^{I_k} \end{pmatrix},$$

where $(x)^{I_r}$ is a vector containing the $\binom{r+d}{d}$ d-variate monomials of degree r, $x_1^{i_1}\dots x_d^{i_d}$ with $i_1 + \dots + i_d = r$. The coefficients $C \coloneqq (c_1, \dots, c_n)$ of the Lagrange polynomials are then obtained simply by solving

$$VC = P,$$

where $P = (p_1, \dots, p_n)$ are the $n = \binom{k+d}{d}$ points to be interpolated.

Interpolation nodes for multivariate Lagrange polynomials can sometimes be obtained from their one-dimensional counterparts. For example, Lagrange polynomials for parallelotopal cells (e.g., rectangular or hexahedral) and their deformations can be obtained by simple tensor prod-

uct of one-dimensional polynomials. For simplicial cells (e.g., triangles and tetrahedra), one can obtain an interpolation grid by using a uniform subdivision for each of the $d+1$ barycentric coordinates λ_i, which parameterize the simplex and satisfy $0 \leq \lambda_i \leq 1$ and $\sum_{i=1}^{d+1} \lambda_i = 1$. Alternative interpolation points with better Lebesgue constants can be found, although this is far from a trivial task, see, e.g., [160–164] and Figure 3.2. Mixed strategies can be employed for intermediate elements such as rectangular pyramids or triangular prisms (also known as *wedges*).

In any case, the same trade-off between uniformity of the approximation (expressed by the Lebesgue constant) and clustering of the nodes (which impacts the k-dependence of the timestep) is present in all dimensions, see, e.g., [143, 165].

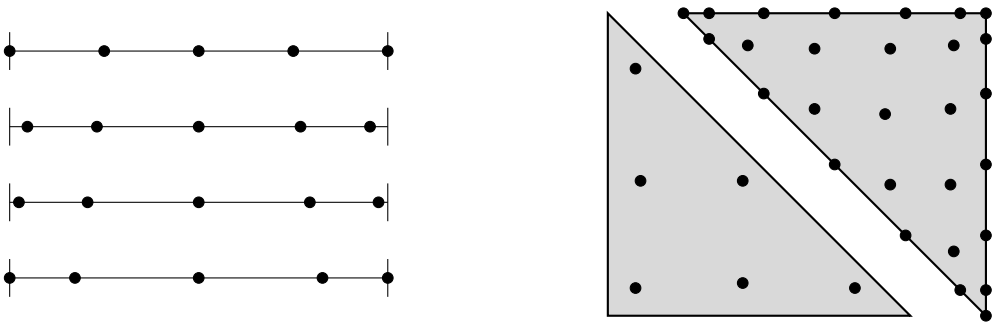


Figure 3.2: (Left) from top to bottom, uniform, Gauss-Legendre, Chebyshev-Gauss and Gauss-Lobatto nodes on the interval for $k = 4$. Notice that nodes with better uniform approximation properties tend to cluster together near the ends of the interval. (Right) Gauss nodes on a triangle for $k = 4$ from [160] (left) and a set of Fekete points for $k = 6$ from [162] (right).

3.1.2 Gauss and Gauss-Lobatto quadratures, spectral elements

The quality and uniformity of the approximation is not the only consideration to be taken into account when choosing the interpolation grid points. As seen in the previous section, discretization schemes based on the Galerkin method lead to the computation of many superposition integrals such as (2.15) and (2.16). The choice of basis has an impact on the cost of the computation of these matrices (i.e., the *matrix assembly* cost), and even more importantly on the form of the resulting matrix. The shape of the Gramian matrix of the basis, i.e., the mass matrix (2.15), is of particular relevance for explicit timestepping, since this matrix needs to be inverted (or a pre-computed factorization reused) at each timestep.

Integrating polynomials over intervals in $\mathbb{R}$ can be achieved through the use of a very efficient computational technique that goes under the name of *Gauss quadrature*. In fact, pick an open finite interval $\Omega \subset \mathbb{R}$, and a measure μ on Ω. Let $q_{k+1}(x)$ be a polynomial of degree $k+1$ (i.e., with nonzero coefficient of x^{k+1}) that is orthogonal to $\mathcal{Q}_r$, $r = 0, \ldots, k$, i.e.,

$$\int_\Omega q_{k+1}(x) g(x) \, \mathrm{d}\mu = 0$$

for all polynomials $g(x)$ of degree strictly less than $k+1$. Let $A_\mathrm{G} \coloneqq (a_i)_{i=1}^{k+1}$ be the roots of q_{k+1}, which must be all real, simple and lie inside Ω [158, Theorem 3.3.1]. One can divide any

polynomial $h(x)$ of degree $s \leq 2k+1$ by $q_{k+1}(x)$, obtaining a quotient $t(x)$ of degree $s-k-1 \leq k$ and a remainder $r(x)$ of degree less than $k+1$. Thus, $q_{k+1}(x)$ is orthogonal to both $t(x)$ and $r(x)$, and

$$\int_\Omega h(x)\,\mathrm{d}\mu = \int_\Omega q_{k+1}(x)t(x) + r(x)\,\mathrm{d}\mu = \int_\Omega r(x)\,\mathrm{d}\mu.$$

Since $r(x)$ has degree less than $k+1$, there exists a set of real weights $(w_i^{\mathrm{G}})_{i=1}^{k+1}$ such that

$$\int_\Omega r(x)\,\mathrm{d}\mu = \sum_{i=1}^{k+1} w_i^{\mathrm{G}} r(a_i),$$

with a_i being the i-th root of $q_{k+1}(x)$. Since $q_{k+1}(a_i) = 0$,

$$r(a_i) = r(a_i) + q_{k+1}(a_i)t(a_i) = h(a_i).$$

We conclude finally that there exists a set of weights $(w_i^{\mathrm{G}})_{i=1}^{k+1}$ such that, for all polynomials $h(x)$ of degree at most $2k+1$,

$$\int_\Omega h(x)\,\mathrm{d}\mu = \sum_{i=1}^{k+1} w_i^{\mathrm{G}} h(a_i),$$

which is known as the *Gauss quadrature rule* for intervals. The computation of the Gaussian weights w_i^{G} is straightforward once one selects a measure μ and determines a set of orthogonal polynomials on (Ω, μ). For example, if $\Omega = [-1, 1]$ and $\mu = 1$, the weights are given by [166]

$$w_i^{\mathrm{G}} = \frac{2}{(1-a_i^2)q'_{k+1}(a_i)^2}.$$

The orthonormal polynomials for the uniform measure $\mu = 1$ are the well-known Legendre polynomials L_k, and for a more general measure $\mu(x) = (1-x)^\alpha(1+x)^\beta$ one obtains various kinds of Jacobi polynomials, such as Chebyshev polynomials for $\alpha = \beta = 1/2$.

If one chooses the Lagrange polynomials of degree k with the nodes placed at the zeros of $q_{k+1}(x)$ as a discretization basis, then one immediately finds using the Gaussian quadrature rule that the mass matrix, if the bulk modulus λ is constant over the interval, reduces to

$$B_{ij} = \frac{1}{\lambda}\int_\Omega \ell_{i,k}^{\mathrm{G}}(x)\ell_{a_j,k}^{\mathrm{G}}(x)\,\mathrm{d}\mu = \frac{1}{\lambda}\sum_{l=1}^{k} w_l \ell_{a_i,k}^{\mathrm{G}}(a_l)\ell_{j,k}^{\mathrm{G}}(a_l) = \frac{1}{\lambda} w_i^{\mathrm{G}}\delta_{ij}, \tag{3.7}$$

where $\ell_{i,k}^{\mathrm{G}}$ is the Lagrange interpolant for the i-th node a_i of A_{G}. Since the calculation of the superposition integral via quadratures is exact in this case, this means that the mass matrix is *diagonal*. Notice that the quadrature rule can be applied since $\ell_{a_i,k}^{\mathrm{G}}(x)\ell_{a_j,k}^{\mathrm{G}}(x)$ has degree $2k$. This fact is an immense computational advantage in view of the previous discussion. However, one cannot realistically use the zeros of such orthogonal polynomials to formulate a proper Galerkin scheme, since all the nodes are located inside Ω, and there is no node placed at the boundaries of the simulation domain to use for the imposition of boundary conditions. The problem is much more prominent for discontinuous Galerkin applications, where one needs nodes

at the boundary of each element in order to compute fluxes and penalty terms.

A solution to this problem consists in modifying the set of quadrature points to explicitly include the endpoints of the interval. Let us denote by $a_2, \dots, a_k$ the $k-1$ roots of the polynomial $q_k'(x)$, where $q_k(x)$ is orthogonal to all polynomials of degree $k-1$ or less on (Ω, μ), and let a_1 and a_{k+1} be the endpoints of Ω. The union of these nodes, $A_{\mathrm{GL}} \coloneqq (a_i)_{i=1}^{k+1}$, is known as the set of *Gauss-Lobatto* quadrature points (see Figure 3.2). Notice that, given a polynomial $g(x)$,

$$\int_\Omega g(x) q_k'(x)\,\mathrm{d}\mu = g(x)q_k(x)|_{x=-1}^{x=1} - \int_\Omega g'(x) q_k(x)\,\mathrm{d}\mu,$$

and the second integral is zero whenever $g(x)$ has degree k or less. The same argument as in the case of the Gaussian quadrature rules can then be applied to conclude that the Gauss-Lobatto quadrature is exact if the integrand has a degree of $2k-1$ or less, instead of $2k+1$ as in the case of pure Gaussian quadrature rules. The weights can be determined as before once Ω and μ are specified. Again in the case $\Omega = [-1, 1]$ and $\mu = 1$, one finds [166]

$$w_1^{\mathrm{GL}} = w_{k+1}^{\mathrm{GL}} = \frac{2}{k(k+1)},\ w_i^{\mathrm{GL}} = \frac{2}{k(k+1)q_k(a_i)^2}.$$

Notice that the Lagrange polynomials $(\ell_{i,k})_{i=1}^{k+1}$ of degree k over these points are all zero on the boundary, except for the first and last, which are nonzero only on the initial and final endpoints of the interval, respectively. This is the ideal configuration for the computation of fluxes, penalty terms and boundary conditions. However, since only the integral of polynomials up to degree $2k-1$ is exact on the grid A^{GL}, the integration of $\ell_{i,k}^{\mathrm{GL}}(x)\ell_{j,k}^{\mathrm{GL}}(x)$ via (3.7) is not exact in general, as the product has degree $2k$. Nonetheless, one can easily evaluate the error in this approximation. We follow here the arguments of [167].

First, notice that the polynomials $(q_j(x))_{j=0}^{k}$ forming the orthonormal basis (Ω, μ) are still orthogonal if the integrals are replaced by Gauss-Lobatto quadratures, the only difference being that $q_k(x)$ is no longer normalized to 1. In fact, in the integrals

$$\int_\Omega q_i(x) q_j(x)\,\mathrm{d}\mu,$$

the integrands are all polynomials of degree strictly less than $2k$, unless $i = j = k$, and therefore they can be computed exactly using the quadrature rule. Let us measure the lack of normalization of $q_k(x)$ using

$$\frac{1}{\beta} \coloneqq \sum_{i=1}^{k+1} w_i^{\mathrm{GL}} q_k(a_i)^2, \tag{3.8}$$

which is always well defined since q_k cannot be zero at any of the locations a_i. One can express $\ell_{i,k}^{\mathrm{GL}}(x)$ as

$$\ell_{i,k}^{\mathrm{GL}}(x) = \sum_{j=0}^{k} \alpha_{ij} q_j(x)$$

with the coefficients

$$\alpha_{ij} = \sum_{r=1}^{k+1} w_r^{\mathrm{GL}} \ell_{i,k}^{\mathrm{GL}}(a_r) q_j(a_r) = w_i^{\mathrm{GL}} q_j(a_i) \text{ for } j < k,$$
$$\alpha_{ik} = \beta w_i^{\mathrm{GL}} q_k(a_i).$$

Using this expression, one can rewrite (3.7) as

$$\begin{aligned}
\lambda M_{ij} &= \int_\Omega \ell_{i,k}^{\mathrm{GL}}(x) \ell_{j,k}^{\mathrm{GL}}(x) \,\mathrm{d}\mu = \sum_{r=0}^{k} \sum_{s=0}^{k} \alpha_{ir} \alpha_{js} \int_\Omega q_r(x) q_s(x) \,\mathrm{d}\mu = \sum_{r=0}^{k} \sum_{s=0}^{k} \alpha_{ir} \alpha_{js} \delta_{rs}, \\
&= \sum_{r=0}^{k} \alpha_{ir} \alpha_{jr} = \sum_{r=0}^{k} \alpha_{ir} w_j^{\mathrm{GL}} q_r(a_j) + (\beta - 1) \alpha_{ik} w_j^{\mathrm{GL}} q_k(a_j), \\
&= w_j^{\mathrm{GL}} \ell_{i,k}^{\mathrm{GL}}(a_j) + \beta(\beta - 1) w_i^{\mathrm{GL}} w_j^{\mathrm{GL}} q_k(a_i) q_k(a_j), \\
&= w_i^{\mathrm{GL}} \delta_{ij} + \beta(\beta - 1) w_i^{\mathrm{GL}} w_j^{\mathrm{GL}} q_k(a_i) q_k(a_j), \\
&=: \lambda M_{ij}^{\mathrm{GL}} + \lambda R_{ij},
\end{aligned}$$

where

$$M_{ij}^{\mathrm{GL}} := \frac{1}{\lambda} w_i^{\mathrm{GL}} \delta_{ij}$$

is the (diagonal) Gauss-Lobatto approximation for the mass matrix and

$$R_{ij} := \frac{1}{\lambda} \beta(\beta - 1) w_i^{\mathrm{GL}} w_j^{\mathrm{GL}} q_k(a_i) q_k(a_j)$$

is the difference between the mass matrix and its approximation. Notice that R_{ij} is a *dyadic matrix*, since it can be expressed as the external product uv^T, using the vectors $u_i := \beta(\beta - 1)/\lambda \, w_i^{\mathrm{GL}} q_k(a_i)$ and $v_i := w_i^{\mathrm{GL}} q_k(a_i)$. Thus, the mass matrix and its Gauss-Lobatto approximation are related by a simple rank-one update [167]. For a finite-element method based on a mesh with N intervals, the preceding argument is still valid inside each mesh element, and the two matrices differ by N rank-one updates, one for every element of the mesh.

One can also evaluate the error in approximating the inverse of the mass matrix M^{-1} by $\left(M^{\mathrm{GL}}\right)^{-1}$, using the Sherman-Morrison formula [120], as

$$M^{-1} = \left(M^{\mathrm{GL}}\right)^{-1} - \frac{\left(M^{\mathrm{GL}}\right)^{-1} uv^T \left(M^{\mathrm{GL}}\right)^{-1}}{1 + v^T \left(M^{\mathrm{GL}}\right)^{-1} u}.$$

One has

$$\left(\left(M^{\mathrm{GL}}\right)^{-1} uv^T \left(M^{\mathrm{GL}}\right)^{-1}\right)_{ij} = \lambda\beta(\beta - 1) q_k(a_i) q_k(a_j)$$

and

$$v^T \left(M^{\mathrm{GL}}\right)^{-1} u = \beta(\beta - 1) \sum_{i=1}^{k+1} w_i^{\mathrm{GL}} q_k(a_i)^2 = \beta - 1,$$

cf. (3.8). Thus,

$$\left(M^{-1}\right)_{ij} = \frac{\lambda}{w_i^{\mathrm{GL}}}\delta_{ij} - \lambda(\beta-1)q_k(a_i)q_k(a_j).$$

Notice that the correction term $-\lambda(\beta-1)q_k(a_i)q_k(a_j)$ is also a dyadic matrix, which can be expressed as wz^T with $w_i \coloneqq -\lambda(\beta-1)q_k(a_i)$ and $z_i \coloneqq q_k(a_i)$. The fact that the polynomial q_k is orthogonal to all the polynomials of degree strictly less than k is an extremely important fact for the numerical properties of the approximation. In fact, suppose that $h(x)$ is a polynomial of degree less than k. Computing the weak form of $h(x)$ by projecting on the i-th test function yields a vector H with components

$$(H)_i \coloneqq \int_\Omega h(x)\ell_{i,k}^{\mathrm{GL}}(x)\,\mathrm{d}\mu = \sum_{j=1}^{k+1} w_j^{\mathrm{GL}} h(a_j)\ell_{i,k}^{\mathrm{GL}}(a_j) = w_i^{\mathrm{GL}} h(a_i).$$

But then, when M^{-1} is applied to this vector,

$$(M^{-1}H)_i = ((M^{\mathrm{GL}})^{-1}H)_i - \lambda(\beta-1)\sum_{j=1}^{k+1} q_k(a_i)q_k(a_j)w_j^{\mathrm{GL}}h(a_j) = ((M^{\mathrm{GL}})^{-1}H)_i,$$

due to the above-mentioned orthogonality property of q_k. Therefore, the correction term to the inverse mass matrix is orthogonal to the subspace generated by the weak form of all polynomials of degree k or less. In other words, the degree of error introduced by this approximation is of the same order as the overall degree of approximation of the discrete problem. Thus, the order of convergence of the method is not affected.

The resulting numerical scheme is known as the *spectral element method* (SEM), and has been employed for the seismic wave propagation problem since the seminal paper [168].

In dimension $d > 1$, a suitable basis that makes the mass matrix diagonal is obtained by simple tensor product of d copies of one-dimensional SEM basis functions. A consequence of this choice is that only tensor-product cells (i.e., parallelotopes such as rectangles or hexahedra) are allowed in a SEM mesh, which can be a severe limitation when dealing with complex geometries. For this reason, some recent approaches such as [169] have tried to couple the SEM method with a method based on an unstructured mesh, such as the DG method. The resulting scheme achieves a good performance in subregions of simple geometry, which are meshed via hexahedra and where SEM is applied, while retaining the geometric flexibility of an unstructured mesh in zones of complex geometry.

3.1.3 Bernstein polynomials

Let us return for a moment to the three-point interpolation scheme (3.3). The logic behind this approximation scheme consisted in an iterated linear interpolation, first between couples of adjacent points, then between couples of adjacent interpolating curves, and so on until a single interpolating curve is obtained. This scheme is always interpolating, and therefore bound to produce the Lagrange polynomials, due to their uniqueness. As pointed out, however, the interpolation is not convex, leading to a lack of uniformity in the approximation, quantified by the Lebesgue constant.

The question then naturally arises: would it be possible to obtain an approximation of the $k+1$ points $p_1, \ldots, p_{k+1}$ that is always convex? Notice that this approximation must necessarily be non-interpolating, and we are merely requesting that the curve passes close to the input points, exchanging nodal interpolation for uniform and convex approximation.

Looking at equation (3.3), the reason why the resulting interpolating polynomial is not convex appears rather clearly. In fact, the global interpolation between the two segments $\gamma_1(x)$ and $\gamma_2(x)$ is convex for all parameters $x \in [a_1, a_3]$, but $\gamma_1(x)$ itself is only convex for $x \in [a_1, a_2]$, and $\gamma_2(x)$ is only convex for $x \in [a_2, a_3]$. Thus, there is no value x at which all three interpolations are convex. Armed with this observation, and following the logic of [153], we can obtain a convex combination over the whole interval $[a_1, a_3]$ by first re-parameterizing the two segments $\gamma_1(x)$ and $\gamma_2(x)$ over the same interval $[a_1, a_3]$, and only then taking the linear combination. The resulting quadratic interpolation is

$$\gamma_B(x) = \frac{(x-a_3)^2}{(a_3-a_1)^2}p_1 + 2\frac{(a_3-x)(x-a_1)}{(a_3-a_1)^2}p_2 + \frac{(x-a_1)^2}{(a_3-a_1)^2}p_3, \tag{3.9}$$

which is the (quadratic) *Bézier curve* between the three points p_1, p_2 and p_3. Comparing (3.9) with (3.4), it is clear that, in the case of the Bézier curve, the three coefficients are positive for every $x \in [a_1, a_3]$, and they sum to one (i.e., they form a partition of unity). Thus, the approximation is convex, and the curve is completely contained in the convex hull $\operatorname{conv}\left((a_i, p_i)_{i=1}^3\right)$ of the three original points. It is easy to check that $\gamma_B(a_1) = p_1$ and $\gamma_B(a_3) = p_2$, so that the curve interpolates the endpoints. However, in general $\gamma_B(a_2) \neq p_2$, so the intermediate point is not interpolated. For this reason, the points p_1, p_2 and p_3 are not called interpolation points but *control variables*, and the corresponding parameter space points are not called nodes but *knots*.

It is easy to extend the rule used in (3.9) to higher order. We obtain a polynomial construction, the *Bézier curve*

$$\gamma_B(x) = \sum_{i=1}^{k+1} b_{i,k}(x) p_i,$$

where the Bézier coefficients $(b_{i,k})_{i=1}^{k+1}$ are given by

$$b_{i,k}(x) = \frac{1}{(a_{k+1}-a_1)^k}\binom{k}{i-1}(x-a_1)^{i-1}(a_{k+1}-x)^{k+1-i}. \tag{3.10}$$

These coefficients, when reparameterized over the interval $[0, 1]$ are also known as *Bernstein polynomials*,

$$\bar{b}_{i,k}(t) = \binom{k}{i-1} t^{i-1}(1-t)^{k+1-i}. \tag{3.11}$$

Bernstein polynomials are another popular choice for basis polynomials in Galerkin finite element analysis. Notice that, since both Bernstein polynomials and Lagrange polynomials span the same space at every degree k, the set of solutions that can be approximated by any of the two polynomial bases at a given degree is exactly the same. For this reason, there exists a simple invertible linear operator that allows to convert between the expansion coefficients in one basis and the other. However, in the case of Lagrange polynomials, the expansion coefficients of

a function also represent the interpolated values of the function at the given parameter values (*nodal* approximation), while the coefficients of Bernstein polynomials do not have such a simple meaning (*modal* approximation).

Bernstein polynomial bases possess many nice computational properties, including some forms of optimal stability, that we will not discuss here. We refer instead the reader to one of the specialized texts (see for example [170]). One thing worth mentioning here, however, is that Bernstein polynomials can be computed at any degree efficiently and robustly using *De Casteljau's algorithm* ([171]),

$$\begin{aligned}\gamma_B^{(i,0)}(x) &\coloneqq p_i,\\ \gamma_B^{(i,j)}(x) &\coloneqq \frac{a_{k+1}-x}{a_{k+1}-a_1}\gamma_B^{(i,j-1)}(x) + \frac{x-a_1}{a_{k+1}-a_1}\gamma_B^{(i+1,j-1)}(x),\end{aligned} \tag{3.12}$$

for $i = 1, \ldots, k+1$, with $\gamma_B(x) \coloneqq \gamma_B^{(1,k)}(x)$.

One of the most interesting aspects of Bernstein polynomials is that the interpolation is easy to do as long as we have a parameter $t \in [0,1]$ that describes our position in the interpolation interval. It is then very easy to write bivariate, or generally multivariate, Bernstein polynomials. To see how this is possible, consider again the Bernstein coefficients given in (3.11). We can create some *barycentric coordinates* λ_1, λ_2 in the interval $[0,1]$ by simply defining $\lambda_1 \coloneqq t, \lambda_2 \coloneqq 1-t$. The coefficients of (3.11) are then simply the binomial coefficients of

$$(\lambda_1 + \lambda_2)^k.$$

By exploiting this construction, we can create multivariate Bernstein polynomials defined on any shape on which barycentric coordinates are available. This is especially useful for triangles and tetrahedra, but can be done in more general settings such as general polygons (see for example [172]). If we have a set of barycentric coordinates $(\lambda_i)_{i=1}^n$ satisfying $0 \le \lambda_i \le 1$ and $\sum_{i=1}^n \lambda_i = 1$, the corresponding Bernstein polynomials are simply given by the multinomial expansion of

$$\left(\sum_{i=1}^{n} \lambda_i\right)^k. \tag{3.13}$$

For this reason, Bernstein polynomials are natural candidates for polynomial basis functions defined on unstructured meshes.

Another important property of Bernstein polynomials can be easily derived from (3.10), namely

$$b'_{i,k}(x) = k\left(b_{i-1,k-1}(x) - b_{i,k-1}(x)\right),$$

with the understanding that $b_{-1,k}(x) = 0$. This formula, which allows to write the derivative of a Bernstein polynomial of order k as a sum of two Bernstein polynomials of order $k-1$, is particularly useful to compute the stiffness matrix (2.16) and other matrices arising in Galerkin methods efficiently. We will see that this important feature is shared by B-spline functions, as well as unstructured splines. We will also see that Bernstein-Bézier polynomials over simplices (and more generally the corresponding DG and FEM spaces on a mesh) can be obtained as

special cases of unstructured spline functions and their spaces.

Finally, notice that although Bernstein-Bézier polynomials and Lagrange polynomials span the same polynomial space, the numerical stability of the resulting superposition matrices (e.g., the mass matrix (2.15) and stiffness matrix (2.16)) can be very different.

We show in Figure 3.3 an example of Bernstein basis on the interval and on a triangle.

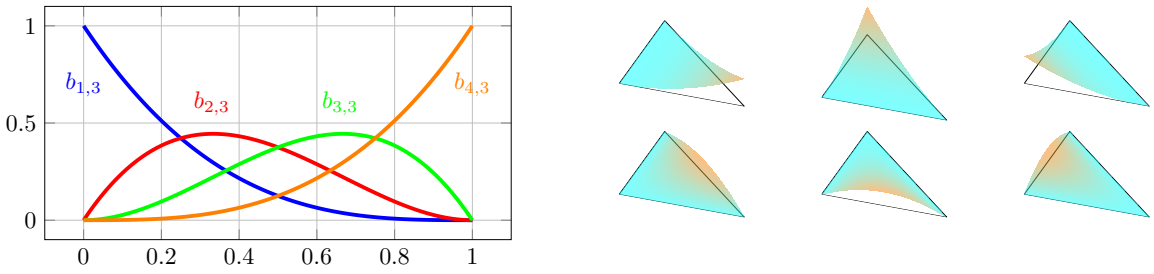


Figure 3.3: Bernstein polynomials of degree $k = 3$ over the segment $[0, 1]$ (left) and for $k = 2$ over the standard unit triangle (right).

3.1.4 B-splines

So far, we have assumed that the domain Ω is subdivided into a series of cells (intervals for $d = 1$), each endowed with an independent polynomials space, built with Lagrange or Bernstein-Bézier polynomials. The resulting space is generally non-conforming across the cell boundaries, as no smoothness is imposed there. FEM spaces generally restore C^0 regularity by performing suitable linear combinations of basis functions. However, low-regularity spaces might be non-conforming for many physical problems involving higher space derivatives. Furthermore, the shape of basis functions is in this case very heterogeneous, including completely smooth functions (with zero trace on the boundary of elements) and non-smooth functions peaked on the boundaries. This can lead to an increase in numerical noise and generally worse numerical properties of the system matrices [2].

Obtaining smooth approximations over $n+1$ of points using the previous techniques requires the construction of a polynomial of degree n, clearly infeasible for real applications. This can be avoided by slightly modifying our interpolation rules in order to glue together smoothly multiple Bézier curves, obtaining a *Bézier spline*. This can be achieved by slightly modifying De Casteljau's recursive construction rule (3.12) as follows. Let us start with a *grid* of $n+k+1$ points $A = (a_1, \ldots, a_{n+k+1})$, $a_1 \leq \ldots \leq a_{n+k+1}$. This vector is known as the *knot vector* of the spline. As in the case of Bézier curves, we start with $n + k$ piecewise constant functions:

$$N_{i,0}(x) = \begin{cases} 1 & \text{if } a_i \leq x < a_{i+1}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.14}$$

This defines the degree-0 basis functions. Note that the interval $a_i \leq x < a_{i+1}$ is open on the right. Starting from (3.14), the higher order basis functions can be obtained by the recursion formula

$$N_{i,k}(x) = \frac{x - a_i}{a_{i+k} - a_i} N_{i,k-1}(x) + \frac{a_{i+k+1} - x}{a_{i+k+1} - a_{i+1}} N_{i+1,k-1}(x). \tag{3.15}$$

The resulting basis functions are called *B-splines*, and the recursion formula is known as the *Cox-De Boor* algorithm (see [173, 174]). Comparing (3.15) to (3.12), the most important difference is that the interpolation at order k is performed only on $k+1$ consecutive knots. Consequently, a basis function $N_{i,k}(x)$ of order k is nonzero only on $k+1$ consecutive knot spans (*locality*). Since the next basis function $N_{i+1,k}(x)$ is shifted to the right by one knot span, this means that, if we want to have n basis functions (corresponding to n control points), we need a knot vector of size $n+k+1$. In this case, we can form our B-spline curve simply by combining the basis functions with the control points,

$$\gamma_N(x) = \sum_{i=1}^{n} N_{i,k}(x) p_i.$$

The locality property of the basis ensures that the modification of a single control point has only a local effect on the curve, changing its value in only $k+1$ knot spans. Finite element analysis based on B-spline functions is called *isogeometric analysis* (IGA) [2].

Another property worth noting is that the recursion formula (3.15) can be evaluated even when a series of m consecutive knots are coincident, i.e., $a_i = a_{i+1} = \ldots = a_{i+m-1}$, provided that the corresponding indeterminate fractions $(x-a_i)/(a_{i+k}-a_i)$ are taken to be identically zero when the denominator is zero. As we will see, this is an important feature that allows the modification of the local regularity of the basis functions.

The recurrence relation (3.15) also allows us to compute the derivatives of any B-spline function in terms of lower-order B-splines. We have:

$$N'_{i,k}(x) = \frac{k}{a_{i+k}-a_i} N_{i,k-1}(x) - \frac{k}{a_{i+k+1}-a_{i+1}} N_{i+1,k-1}(x). \tag{3.16}$$

Repeating the derivation, we obtain a formula for the higher derivatives in terms of lower degree functions:

$$N^{(r)}_{i,k}(x) = \frac{k!}{(k-r)!} \sum_{j=0}^{r} \alpha_{r,j} N_{i+j,k-r}(x),$$

with the coefficients given by:

$$\begin{aligned}
\alpha_{0,0} &= 1, \\
\alpha_{r,0} &= \frac{\alpha_{r-1,0}}{a_{i+k-r+1}-a_i}, \\
\alpha_{r,j} &= \frac{\alpha_{r-1,j}-\alpha_{r-1,j-1}}{a_{i+k+j-r+1}-a_{i+j}} \text{ for } j = 1, \ldots, r-1, \\
\alpha_{r,r} &= \frac{-\alpha_{r-1,r-1}}{a_{i+k+1}-a_{i+r}}.
\end{aligned}$$

The proofs of these statements are by induction on k using (3.15), and can be found in many specialized texts such as [175].

We will now prove a series of properties of B-splines that are important for finite element and isogeometric analysis. The proofs can be found in many texts (see, e.g., [153, 175], but are given here since they are a good introduction to the recursive properties of B-spline functions

that will be useful in Chapter 4.

Theorem 3.1.1. *Let $N_{i,k}(x)$, $i = 1, \ldots, n$ be a set of basis functions defined by the recursion formulas* (3.14) *and* (3.15) *starting from a knot vector $x = (a_1, \ldots, a_{n+k+1})$, $a_1 \leq \ldots \leq a_{i+k+1}$. The functions $N_{i,k}(x)$ have the following properties:*

1. *Local support: the basis function $N_{i,k}(x)$ is only nonzero for $x \in [a_i, a_{i+k+1})$, that is, only in $k+1$ consecutive knot spans. This also means that, for each $x \in [a_1, a_{n+k+1})$, only $k+1$ basis functions are supported at x;*

2. *Bandwidth: the support of a basis function $N_{i,k}(x)$ superposes with that of at most $2k+1$ other basis functions;*

3. *Positivity: all basis functions are nonnegative, $N_{i,k}(x) \geq 0$ for all x;*

4. *Partition of unity: for every degree k, $\sum_{i=1}^{n} N_{i,k}(x) = 1$ for all $x \in [a_k, a_{n+1})$. Thus, if the first and last knot values are repeated k times, the partition of unit is true everywhere;*

5. *Regularity: every basis function $N_{i,k}(x)$ has regularity C^∞ between any two knots, and regularity C^{k-r} across every knot which is repeated r times;*

6. *Interpolation: if a knot a_j is repeated k times, i.e., $a_j = \ldots = a_{j+k-1}$, then the basis functions satisfy $N_{i,k}(a_j) = \delta_{i,j}$, and the corresponding control point is exactly interpolated.*

Proof. Property 1 follows immediately from the recursion relation (3.15), since $\mathrm{supp}(N_{i,0}) = [a_i, a_{i+1})$ and at each degree k, the support grows on the right by 1 knot, i.e., $(\mathrm{supp}(N_{i,k}) = \mathrm{supp}(N_{i,k-1}) \cup \mathrm{supp}(N_{i+1,k-1}))$.

Property 2 follows from property 1 and the fact that $\mathrm{supp}(N_{i+m,k})$ is shifted by m knot spans to the right with respect to $\mathrm{supp}(N_{i,k})$. Thus, the support of $N_{i,k}$ superposes at most with that of itself, k functions to the left and k functions to the right.

Property 3 follows from (3.15) and the fact that the two factors $(x - a_i)/(a_{i+k} - a_i)$, (resp. $(a_{i+k+1} - x)/(a_{i+k+1} - a_{i+1})$) are only negative for $x \notin [a_i, a_{i+k})$ (resp. $x \notin [a_{i+1}, a_{i+k+1})$) where the corresponding basis functions $N_{i,k-1}$ (resp. $N_{i+1,k-1}$) are zero (property 1).

Partition of unity (property 4) is clearly true for degree 0 basis functions, which are just the characteristic functions of their respective (disjoint) supports. Assume now that it is true for all degrees $r \leq k-1$. Consider a non-zero knot span $[a_j, a_{j+1})$ that has at least k knot spans to the right and to the left. According to property 1, only the basis functions $N_{i,k}(x)$ with $i = j-k, \ldots, j$ are nonzero on this span. We have:

$$\begin{aligned}
\sum_{i=j-k}^{j} N_{i,k}(x) &= \sum_{i=j-k}^{j} \left(\frac{x - a_i}{a_{i+k} - a_i} N_{i,k-1}(x) + \frac{a_{i+k+1} - x}{a_{i+k+1} - a_{i+1}} N_{i+1,k-1}(x) \right), \\
&= \frac{x - a_{j-k}}{a_j - a_{j-k}} N_{j-k,k-1}(x) \\
&\qquad + \sum_{i=j-k+1}^{j} \frac{x - a_i + a_{i+k} - x}{a_{i+k} - a_i} N_{i,k-1}(x) + \frac{a_{j+k+1} - x}{a_{j+k+1} - a_{j+1}} N_{j+1,k-1}(x).
\end{aligned}$$

Now, the functions $N_{j-k,k-1}(x)$ and $N_{j+1,k-1}(x)$ are both zero on the knot span $[a_j, a_{j+1})$, and by the induction hypothesis, $\sum_{i=j-k+1}^{j} N_{i,k-1}(x) = 1$, proving the property.

Let us prove now the interpolation property 6. For $k = 0$, the property is clearly true, since $N_{i,0}(a_j) = \delta_{ij}$ by (3.14). Note the importance of the open interval in the definition. Assume now that the property is true for all $r \leq k-1$. Consider k consecutive repeated knots $a_j = \ldots = a_{j+k-1}$. The value $a_j = a_{j+k-1}$ belongs to the knot span $[a_{j+k-1}, a_{j+k})$, which means that the only functions that can be nonzero at a_j, according to property 1, are the $k+1$ functions $N_{i,k}$ with $i = j-1, \ldots, j+k-1$. The function $N_{i,k}(a_j)$ is given by a sum of $N_{i,k-1}$ and $N_{i+1,k-1}$ by (3.15). Now, by induction hypothesis, we have that $N_{i,k-1}(a_j) = \delta_{i,j}$, since the node a_j is repeated $k > k-1$ times. Consequently, the only nonzero order k function is

$$N_{j,k}(a_j) = \frac{a_j - a_{j-1}}{a_{j-1+k} - a_{j-1}} = 1,$$

proving the property.

Finally, we prove property 5. In the proof of property 6, we saw that, if a knot is repeated k times, then the only function that is nonzero at the knot is $N_{j-1,k}(a_j)$. If the knot is repeated once more, the support of this function reduces to zero, and the basis functions are no longer continuous at the knot. Imagine now that a knot is repeated only r times. By using (3.1.4), we see that the m-th derivative of $N_{i,k}(a_j)$ involves the basis functions $N_{j,k-m}$, which are continuous for $m = r$ but discontinuous for $m = r+1$ by the previous argument. □

We will see in Chapter 4 that Properties 1, 3, 4, 6 are valid also for unstructured splines, with minor modifications, while Property 2 is not valid (although the bandwidth is still limited), and 5 is valid in a more general sense, i.e., when points become affinely dependent.

When the first and last points are repeated $k+1$ times each, the knot vector is called *open* or *clamped*. For example, for a cubic spline ($k = 3$), the following is an open knot vector:

$$\{0, 0, 0, 0, 1/3, 2/3, 1, 1, 1, 1\}.$$

By Properties 5 and 6, the functions then interpolate the boundary points, analogously to the Gauss-Lobatto nodes seen above, allowing the imposition of boundary conditions. Working with open knots also means that the partition of unit property (Property 4) is valid everywhere. Thus, applications of B-splines in numerical analysis use almost exclusively open knot vectors. An example of a B-spline basis over an open knot vector is shown in Figure 3.4.

An interesting special case is that of a knot vector containing only the values 0 and 1, repeated $k+1$ times each. By plugging this special knot vector into the recursion rule (3.15), it is easy to see that this basis is nothing else than the Bernstein polynomial basis that we saw in the last section. Consequently, the Bernstein polynomials can be obtained as a special case from B-spline bases, proving that the latter are indeed more general.

Some other properties of B-splines become particularly relevant for their use in numerical analysis. For example, the pointwise positivity of these functions means that the mass matrix (2.15) has all positive entries, making mass lumping and preconditioning strategies easier. On the other hand, positivity also means that the mass matrix cannot be made diagonal by a suitable choice of knots, since the superposition integral between any two non-disjoint functions

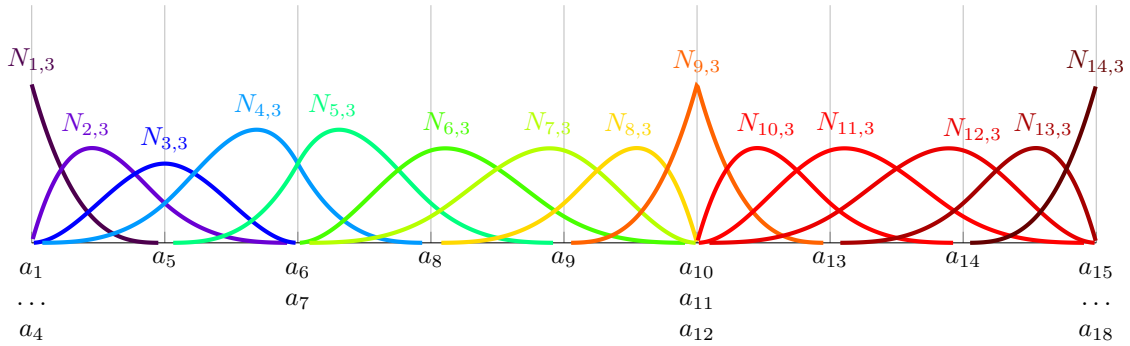


Figure 3.4: Example of a B-spline basis ($k = 3$) over a clamped knot vector with some repetitions.

is necessarily positive and thus nonzero. A linear combination of more than one B-spline function is able to to produce both signs, but usually has a larger support than any of its summands. Consequently, the generalization of the spectral element method to B-spline functions is not straightforward. It is however possible to recover a diagonal mass matrix by introducing a dual basis different to the primal one, see, e.g., [111, 176], a direction that we will not explore in this work.

Spline functions are convex combinations of the control points. This means that the spline curve is less oscillating than the polygonal line connecting the points (also known as the *control net*). B-splines are thus *variation diminishing*, and do not exhibit the overshoots and oscillations typical of Runge's phenomenon that are encountered with interpolating polynomials. These properties conspire to make B-splines a well-conditioned basis, less prone to numerical noise [2]. We refer the reader in particular to an illustrative example contained in [2, Chapter 5], where the high energy vibration modes of a one-dimensional rod are studied, showing that the splitting between acoustic-branch and optical-branch vibration modes is an artifact induced by the choice of basis functions with heterogeneous shape. This artifact disappears when B-spline functions are used.

The possibility to locally change the regularity of the basis by repeating some knot vectors is also particularly interesting for many applications, including seismic wave propagation, as it allows to capture the reduced regularity of the solution in correspondence with sharp reflectors. Moreover, this opens the possibility of recovering discontinuities in the model starting from smooth physical parameters if the knot positions are allowed to evolve during the inversion process (see Chapter 7).

Finally, B-splines possess some efficient computational properties. First of all, the partition of unity property (Property 4) can be extended to a more general polynomial-reproducing property, since the monomial x^k can be expressed as

$$x^k = \sum_{i=1}^{n+k+1} \sigma(a_{i+1}, \ldots, a_{i+k}) N_{i,k}(x), \tag{3.17}$$

where $\sigma(x_1, \ldots, x_k)$ is the totally symmetric polynomial in k variables, normalized such that $\sigma(x, \ldots, x) = x^k$. Notice that only the *internal knots* $(a_{i+1}, \ldots, a_{i+k})$ of the spline function $N_{i,k}$

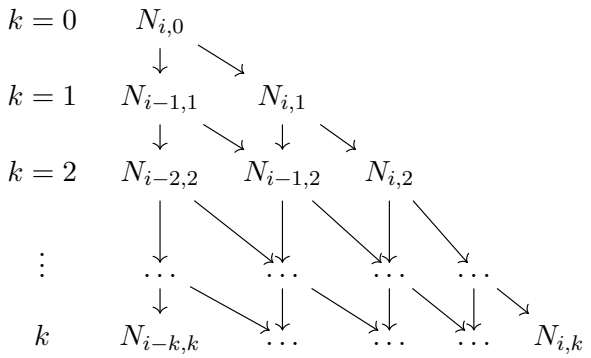


Figure 3.5: Recurrence algorithm used to evaluate all the B-spline functions in a basis supported at a given point x [178, Chapter X, Algorithm 8]. The algorithm starts from the only spline function $N_{i,0}$ of degree 0 supported at x, and computes the value of all the splines supported at x up to a given degree k. Each arrow corresponds to an application of (3.15).

appear as arguments of σ. The expression (3.17) is also known as *Marsden's identity* [177].

Suppose now that we want to compute the system matrices such as (2.15) or (2.16) in practice. Integration can be performed either by numerical quadratures, or as a linear combination of (known) integrals of monomials. However, constructing these matrices element by element has $O(N^2)$ complexity, where N is the number of basis functions. Consequently, some algorithms that allow to evaluate all the B-spline functions supported on a given point x have been developed, most notably [178, Chapter X, Algorithm 8], which underpins virtually all practical applications of B-splines. The scheme is based on the recursion formula (3.15), and is presented in Figure 3.5. Using these algorithms, one only needs to loop over all knot spans once, significantly reducing the matrix assembly complexity.

Extending these properties and algorithms to multivariate unstructured splines is a crucial prerequisite for their use in numerical analysis. We give a contribution towards this goal in Chapter 5.

3.2 CFL condition with cardinal B-splines

In this section, we introduce a very important machinery that will allow us to compute various properties of B-spline functions, such as their Fourier transforms and superposition integrals, that are essential in their use as basis functions for analysis. These formulas will also help us to extend the definition of B-splines to unstructured domains in higher dimensions.

We will often specialize our equations to the case of *cardinal* B-splines. A *cardinal B-spline* basis is defined on $\mathbb{R}$ using a knot vector without repetition (and thus maximal regularity), with uniformly spaced knots:

$$A = h\mathbb{Z},$$

where h is the (constant) knot span, equivalent to the element size in standard finite element analysis. Cardinal B-spline functions of order k are denoted with $B_{i,k}$, $i \in \mathbb{Z}$. From the symmetry of the knot vector, it is clear that all cardinal B-splines of a given order are just shifted copies of each other, i.e., $B_{i+j,k}(x) = B_{i,k}(x - jh)$. This property makes a lot of expressions simpler

and opens up the possibility of using discrete Fourier transforms to explicitly compute many of the properties of this basis. The cardinal B-spline basis is shown in Figure 3.6.

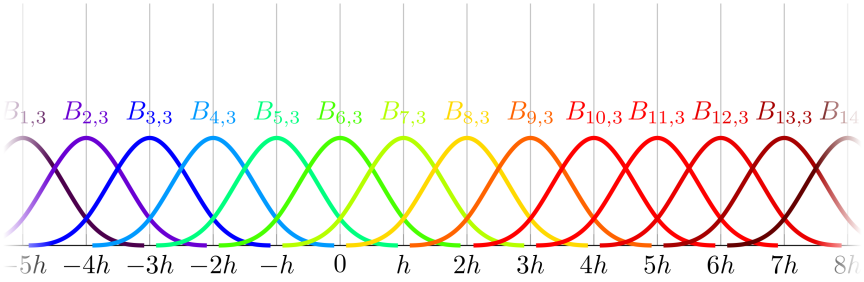


Figure 3.6: The cardinal B-spline basis, i.e., equally spaced and infinite, for $k = 3$.

3.2.1 B-splines and divided differences

We wish to introduce in this section two very important concepts that are extremely useful in a deeper understanding of univariate B-splines. The first is the *truncated power* function x_+^k, defined as

$$x_+^k = \begin{cases} x^k & \text{if } x > 0, \\ 0 & \text{if } x \leq 0, \end{cases}$$

and the second is the *divided difference* of a real function $f : \mathbb{R} \mapsto \mathbb{R}$ over a set of increasing real knots $(a_i, \ldots, a_{i+k})$, defined recursively as follows:

Definition 3.2.1. *Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a real function, and let $\{a_i, \ldots, a_{i+k}\}$ be a sequence of increasing real numbers. The divided difference operator $[a_i, \ldots, a_{i+p}]f$ is defined recursively as follows:*

$$\begin{aligned} [a_i]f &\coloneqq f(a_i), \\ [a_i, \ldots, a_{i+k}]f &\coloneqq \frac{[a_{i+1}, \ldots, a_{i+k}]f - [a_i, \ldots, a_{i+k-1}]f}{a_{i+k} - a_i}. \end{aligned} \tag{3.18}$$

The number of intervals k in the sequence of points is called the *order* of the divided difference. It is clear from the definition that the divided difference operator is a linear functional. For example:

$$\begin{aligned} [a_1]f &= f(a_1), \\ [a_1, a_2]f &= \frac{f(a_2) - f(a_1)}{a_2 - a_1}, \\ [a_1, a_2, a_3]f &= \frac{f(a_1)}{(a_1 - a_2)(a_1 - a_3)} + \frac{f(a_2)}{(a_2 - a_1)(a_2 - a_3)} + \frac{f(a_3)}{(a_3 - a_1)(a_3 - a_2)}. \end{aligned}$$

Divided differences appear, among other things, in the calculation of the Newton polynomial interpolation method and in the calculation of finite difference stencils. In fact, the divided difference of order k of a function f represent an estimate of the k-th derivative of f, as it will be made precise in Theorem 3.2.4.

We try to give here a self-contained presentation of divided differences, and we do not venture into a more general theory, for which the interested reader can find many sources, starting from the more classical articles from De Boor [179] and Whittaker and Robinson [180] to modern texts in numerical analysis such as [181].

We can get an explicit, expanded form for divided differences as follows:

Lemma 3.2.2. *Let $[a_i, \dots a_{i+k}]f$ be the divided difference of f over a set of values. Then:*

$$[a_i, \dots a_{i+k}]f = \sum_{j=0}^{k} \frac{f(a_{i+j})}{\prod_{\substack{r=0 \\ r\neq j}}^{k}(a_{i+j} - a_{i+r})}. \tag{3.19}$$

Proof. By induction over k. The formula is trivially true for $k = 0$, since the divided difference is simply $[a_i]f = f(a_i)$, and in (3.19), only the term $j = 0$ appears in the sum, and no term appears in the product in the denominator. Assume now that (3.19) is true for $k - 1$. We have from the definition (3.18):

$$\begin{aligned}
[a_i, \dots, a_{i+k}]f &= \frac{[a_{i+1}, \dots, a_{i+k}]f - [a_i, \dots, a_{i+k-1}]f}{a_{i+k} - a_i}, \\
&= \frac{1}{a_{i+k} - a_i} \left(\sum_{j=1}^{k} \frac{f(a_{i+j})}{\prod_{\substack{r=1 \\ r\neq j}}^{k}(a_{i+j} - a_{i+r})} - \sum_{j=0}^{k-1} \frac{f(a_{i+j})}{\prod_{\substack{r=0 \\ r\neq j}}^{k-1}(a_{i+j} - a_{i+r})} \right), \\
&= \frac{1}{a_{i+k} - a_i} \left(\sum_{j=1}^{k-1} \frac{f(a_{i+j})(a_{i+k} - a_i)}{\prod_{\substack{r=0 \\ r\neq j}}^{k}(a_{i+j} - a_{i+r})} \right) \\
&\qquad + \frac{1}{a_{i+k} - a_i} \left(\frac{f(a_{i+k})}{\prod_{r=1}^{k-1}(a_{i+k} - a_{i+r})} - \frac{f(a_i)}{\prod_{r=1}^{k-1}(a_i - a_{i+r})} \right), \\
&= \sum_{j=0}^{k} \frac{f(a_{i+j})}{\prod_{\substack{r=0 \\ r\neq j}}^{k}(a_{i+j} - a_{i+r})},
\end{aligned}$$

where in the last equation we have combined the three terms into a single sum for j between 0 and k. Since we obtain (3.19) at order k, we have proven the Lemma. □

This formula takes a particularly simple form if the points on which the divided difference is computed are equally spaced.

Corollary 3.2.3. *Let $(a_i, \dots, a_{i+k})$ be equally spaced points, i.e., $a_n - a_m = (n - m)h$. Then:*

$$[ih, \dots, (i + k)h]f = \frac{(-1)^k}{k!h^k} \sum_{j=0}^{k} (-1)^j \binom{k}{j} f(a_{i+j}). \tag{3.20}$$

Proof. The denominator in (3.19) contains a product on all differences $(a_{i+j} - a_{i+r}) = (j - r)h$ for $r = 0, \dots, k$, $r \neq j$. The differences $j - r$ thus take the values $j, j-1, \dots, 1$ for $r = 0, \dots, j-1$ and $-1, \dots, j - k$ for $r = j + 1, \dots, k$, where the last $k - j$ terms have a negative sign. Thus,

the denominator in (3.19) reduces to

$$h^k(-1)^{k-j}j!(k-j)! = h^k(-1)^{k-j}k!/\binom{k}{j},$$

proving the formula. □

We can now prove the following theorem, which characterizes divided differences as estimators of the derivatives of a function.

Theorem 3.2.4 (Mean value)**.** *Let $A \coloneqq \{a_i, \dots, a_{i+k}\}$ be $(k+1)$ be distinct real numbers, let I be the open interval $I \coloneqq (\min(A), \max(A))$ and let f be a function of class C^k in I. Then there exists a point $x^* \in I$ such that*

$$[a_i, \dots, a_{i+k}]f = \frac{f^{(k)}(x^*)}{k!}. \tag{3.21}$$

Proof. Let $\ell(x)$ be the interpolating Lagrange polynomial of $f(x)$ computed on the interpolation points x,

$$\ell(x) = \sum_{j=0}^{k} \left(\prod_{\substack{r=0 \\ r\neq j}}^{k} \frac{x - a_{i+r}}{a_{i+j} - a_{i+r}} \right) f(a_{i+j}).$$

The (constant) k-th derivative of $\ell(x)$ reads:

$$\begin{aligned} \ell^{(k)} &= k! \sum_{j=0}^{k} \frac{f(a_{i+j})}{\prod_{\substack{r=0 \\ r\neq j}}^{k}(a_{i+j} - a_{i+r})}, \\ &= k![a_i, \dots, a_{i+k}]f. \end{aligned} \tag{3.22}$$

We have used the explicit form given by (3.19). On the other hand, if we consider the remainder of the interpolation $R(x) \coloneqq f(x) - \ell(x)$, we notice that $R(a_{i+j}) = 0$ for every one of the $k+1$ interpolation points a_{i+j}, $j = 0, \dots, k$. By Rolle's theorem, then, $R'(x)$ has a zero in between any two consecutive interpolation points, i.e., it has k zeros. Iterating this reasoning, we realize that $R^{(k)}(x)$ must have (at least) one zero $x^* \in I$, for which $f^{(k)}(x^*) = \ell^{(k)}$. Plugging this value in (3.22) yields the claim. □

The definition of divided difference (3.18) bears a striking resemblance to the Cox-De Boor recursion formula (3.15) that defines the B-spline basis. It is thus not surprising that B-splines can be defined easily in terms of divided differences, see for example [174]. In fact, De Boor used this definition to compute many important properties of B-splines, including the integrals of products of B-spline basis functions in which we are interested (see [182]). We will specialize here those results to the case of cardinal (i.e., equally spaced knots) basis functions.

The B-spline basis functions of degree zero defined in (3.14) are just the characteristic functions of the knot spans $[a_i, a_{i+1})$. Consequently, they can be rewritten as the difference of two

truncated power functions of degree 0, which are just Heaviside functions, as follows:

$$\begin{aligned} N_{i,0}(x) &= (a_{i+1}-x)_+^0 - (a_i - x)_+^0 = [a_{i+1}:y](y-x)_+^0 - [a_i:y](y-x)_+^0, \\ &= (a_{i+1}-a_i)[a_i, a_{i+1}:y](y-x)_+^0, \end{aligned}$$

where the notation $[a_i, \dots, a_{i+k} : y]$ means that the divided difference must be applied to the variable y in the following expression. This pattern extends to higher orders.

Theorem 3.2.5. *The B-splines basis functions can be expressed as divided differences of the truncated power function as follows:*

$$N_{i,k}(x) = (a_{i+k+1} - a_i)[a_i, \dots, a_{i+k+1} : y](y-x)_+^k. \tag{3.23}$$

Proof. We have already proven the theorem for $k = 0$. We proceed by induction on k. From (3.15), we obtain

$$\begin{aligned} N_{i,k}(x) &= \frac{x-a_i}{a_{i+k}-a_i} N_{i,k-1}(x) + \frac{a_{i+k+1}-x}{a_{i+k+1}-a_{i+1}} N_{i+1,k-1}(x), \\ &= \frac{(a_{i+k+1}-x)(a_{i+k}-a_i)N_{i+1,k-1}(x) - (a_i - x)(a_{i+k+1}-a_{i+1})N_{i,k-1}(x)}{(a_{i+k}-a_i)(a_{i+k+1}-a_{i+1})}, \\ &= \left[(a_{i+k+1}-x)[a_{i+1}, \dots, a_{i+k+1}:y](y-x)_+^{k-1} - (a_i - x)[a_i, \dots, a_{i+k}:y](y-x)_+^{k-1} \right], \\ &= (a_{i+k+1}-a_i)[a_i, \dots, a_{i+k+1}:y](y-x)_+^k. \end{aligned}$$

We have used the recursive definition of (3.18) on the function $(y-x)_+^k$, the Leibniz rule and the simple property $(y-x)(y-x)_+^{k-1} = (y-x)_+^k$ □

We can use this expression to derive an explicit formula for the integral of the product of two B-spline basis functions, as done in [182]. This formula is used in [183] to provide an efficient numerical scheme for the calculation of such integrals. However, our goal is be instead to explicitly compute the integral for cardinal basis functions. But first, we have to prove the following Lemma.

Lemma 3.2.6. *We can compute the divided differences of a C^{k+1} function f as*

$$[a_i, \dots, a_{i+k+1}]f = \frac{1}{(a_{i+k+1}-a_i)k!} \int_{-\infty}^{\infty} N_{i,k}(y) f^{(k+1)}(y)\, \mathrm{d}y. \tag{3.24}$$

Proof. The Taylor expansion of f in the interval $[a_i, a_{i+k+1})$ reads:

$$f(x) = \sum_{j=0}^{k} \frac{f^{(j)}(a_i)}{j!}(x-a_i)^j + \int_{a_i}^{x} \frac{f^{(k+1)}(y)}{k!}(x-y)^k \,\mathrm{d}y, \tag{3.25}$$

where the second term in (3.25) is the integral remainder $R(x)$ of the Taylor expansion. We can

replace the power in the remainder by a truncated power and extend the domain of integration,

$$R(x) = \int_{a_i}^{a_{i+k+1}} \frac{f^{(k+1)}(y)}{k!}(x-y)_+^p \,\mathrm{d}y.$$

If we apply the divided difference operator $[a_i, \ldots, a_{i+k+1} : x]$ to this expression, the first term in (3.25) vanishes since the divided difference operator of degree $k+1$ annihilates polynomials of order $j \leq k$ by the mean value theorem (3.21). The remainder reads

$$\begin{aligned}[a_i, \ldots, a_{i+k+1}]f &= \int_{a_i}^{a_{i+k+1}} [a_i, \ldots, a_{i+k+1} : x](x-y)_+^k \frac{f^{(k+1)}(y)}{k!} \,\mathrm{d}y, \\ &= \frac{1}{(a_{i+k+1} - a_i)k!} \int_{a_i}^{a_{i+k+1}} N_{i,k}(y) f^{(k+1)}(y) \,\mathrm{d}y.\end{aligned}$$

Finally, we can extend the integration limits to $(-\infty, \infty)$ because the function $N_{i,k}(y)$ is zero for $y \notin [a_i, a_{i+k+1}]$. □

This is known as the *Peano form* of the divided difference (see, e.g., [179, 184, 185]). Incidentally, by comparing (3.24) and (3.21), we see that this Lemma gives us a way to compute the value of the internal point x^* of Theorem 3.2.4: it is just the weighted average of $f^{(k+1)}(x)$, with a weight proportional to $N_{i,k}(x)$. This Lemma also allows us to compute various superposition integrals of B-spline basis functions, including their Fourier transform. We have the two following important corollaries:

Corollary 3.2.7. *Let $N_{i,k}^A(x)$, $N_{j,r}^B(x)$ be two B-spline basis functions defined over the knot vectors $A = (a_i, \ldots, a_{i+k+1})$ and $B = (b_j, \ldots, b_{j+r+1})$ respectively. The superposition integral*

$$T_{i,j}^{k,r}(A, B) \coloneqq \int_{-\infty}^{\infty} N_{i,k}^A(x) N_{j,r}^B(x) \,\mathrm{d}x$$

is given by

$$T_{i,j}^{k,r}(A, B) = C_{i,j}^{k,r}(A, B)[a_i, \ldots, a_{i+k+1} : x][b_j, \ldots, b_{j+r+1} : y](y-x)_+^{k+r+1}, \tag{3.26}$$

with the factor $C_{i,j}^{k,r}(A, B)$ defined as

$$C_{i,j}^{k,r}(A, B) \coloneqq (-1)^{k+1} \frac{(a_{i+k+1} - a_i)(b_{j+r+1} - b_j)k!r!}{(k+r+1)!}.$$

Proof. We take $f(x) = (b_{j+r+1} - b_j)[b_j, \ldots, b_{j+r+1} : y](y-x)_+^{k+r+1}$, which is a function of class C^∞ in the variable x if $x \neq b_i$ for all i, and class C^{k+r+1} otherwise. The $(k+1)$-th derivative of f is:

$$\begin{aligned}f^{(k+1)}(x) &= (-1)^{k+1} \frac{(k+r+1)!}{r!}(b_{j+r+1} - b_j)[b_j, \ldots, b_{j+r+1} : y](y-x)_+^r, \\ &= (-1)^{k+1} \frac{(k+r+1)!}{r!} N_{j,r}(x).\end{aligned}$$

Plugging the above defined f into (3.24) proves the corollary. □

Corollary 3.2.8. *Let $N_{i,k}(x)$ be a B-spline basis function. Then, its Fourier transform reads:*

$$\widetilde{N}_{i,k}(\nu) = \frac{(a_{i+k+1} - a_i)k!}{(2\pi\mathrm{i}\nu)^{k+1}} \sum_{j=0}^{k+1} \frac{e^{2\pi\mathrm{i}\nu a_{i+j}}}{\prod_{\substack{r=0 \\ r\neq j}}^{k+1}(a_{i+j} - a_{i+r})}.$$

Proof. We have

$$\widetilde{N}_{i,k}(\nu) = \int_{-\infty}^{\infty} N_{i,k}(x) e^{2\pi\mathrm{i}\nu x}\,\mathrm{d}x.$$

We can use (3.24) by setting $f(x) = e^{2\pi\mathrm{i}\nu x}$, obtaining:

$$\widetilde{N}_{i,k}(\nu) = (a_{i+k+1} - a_i)k![a_i, \ldots, a_{i+k+1} : x]\frac{e^{2\pi\mathrm{i}\nu x}}{(2\pi\mathrm{i}\nu)^{k+1}}.$$

Finally, by using the expanded form (3.19) for the divided difference, we obtain the claim. □

This result was already obtained in [186]. For a cardinal B-spline basis function, we can use (3.19) for the knot vector $h\mathbb{Z}$, obtaining:

$$\begin{aligned}
\widetilde{B}_{i,k}(\nu) &= \frac{1}{(2\pi\mathrm{i}\nu h)^{k+1}} e^{2\pi\mathrm{i}\nu(ih)} \sum_{j=0}^{k+1} \binom{k+1}{j} e^{2\pi\mathrm{i}\nu(jh)}(-1)^{k+1-j}, \qquad (3.27)\\
&= e^{2\pi\mathrm{i}\nu(ih)} \left(\frac{e^{2\pi\mathrm{i}\nu h} - 1}{2\pi\mathrm{i}\nu h}\right)^{k+1},\\
&= e^{\pi\mathrm{i}\nu(2i+k+1)h} \left(\frac{\sin \pi\nu h}{\pi\nu h}\right)^{k+1}.
\end{aligned}$$

Corollary 3.26 gives us a very useful explicit form for the superposition integral of two B-splines basis functions, and can already be found for example in [183]. The formula for the Fourier transform (3.27) can be already be found in [186, 187]. It is not surprising that the Fourier transform of a cardinal B-spline basis function is a power of the function $\mathrm{sinc}(x) \coloneqq \sin(x)/x$. In fact, this result could have been obtained by noting that the cardinal B-spline functions can be expressed as repeated convolutions of the characteristic function, $B_{i,k} = \chi([ih, (i+1)h])^{*(n+1)}$. Note that the Fourier transform is of course real if $i = -(k+1)/2$, i.e., if the basis function is centred around the origin and thus even. Since the Fourier transform is a unitary operator, we can compute the scalar product of two B-spline basis functions in Fourier space (Parseval's identity),

$$\begin{aligned}
\int_{-\infty}^{\infty} B_{i,k}(x)B_{j,r}(x)\,\mathrm{d}x &= \int_{-\infty}^{\infty} e^{\pi\mathrm{i}\nu(2(i-j)+k-r)h} \left(\frac{\sin \pi\nu h}{\pi\nu h}\right)^{k+r+2} \mathrm{d}\nu,\\
&= hB_{i,k+r+1}((j+r+1)h) = hB_{j,k+r+1}((i+k+1)h).
\end{aligned}$$

From the definition, it is clear that the superposition integral is zero if $i > j + r + 1$, or if $j > i + k + 1$, which is obvious since $B_{i,k}$ is supported on $[i, i+k+1]h$ and $B_{j,r}$ is supported on

$[j, j+r+1]h$. We will be interested in the special case $k = r$,

$$\int_{-\infty}^{\infty} B_{i,k}(x)B_{j,k}(x)\,\mathrm{d}x = hB_{i,2k+1}((j+k+1)h) = hB_{j,2k+1}((i+k+1)h). \tag{3.28}$$

Finally, in order to derive the stiffness matrix, we can use (3.16) to express the derivative of a B-spline basis function as a combination of B-spline basis functions of a lower order. For the cardinal basis,

$$B'_{i,k}(x) = \frac{1}{h}\left(B_{i,k-1}(x) - B_{i+1,k-1}(x)\right). \tag{3.29}$$

Consequently, we also have:

$$\begin{aligned}
\int_{-\infty}^{\infty} B'_{i,k}(x)B'_{j,k}(x)\,\mathrm{d}x &= \frac{1}{h^2}\int_{-\infty}^{\infty}\left(B_{i,k-1}(x) - B_{i+1,k-1}\right)\left(B_{j,k-1}(x) - B_{j+1,k-1}\right)\,\mathrm{d}x, \qquad (3.30)\\
&= \frac{1}{h}\left(2B_{i,2k-1}((j+k)h) - B_{i+1,2k-1}((j+k)h)\right.\\
&\qquad\qquad \left. -B_{i-1,2k-1}((j+k+1)h)\right),\\
&= \left(B'_{i,2k}((j+k)h) - B'_{i-1,2k}((j+k)h)\right)\\
&= hB''_{i,2k+1}((j+k+1)h).
\end{aligned}$$

3.2.2 Spectral properties of the mass and stiffness matrices

Let us now consider a one-dimensional homogeneous domain, and let us compute the stability conditions

$$\begin{aligned}
\lambda_{\min}(M^{-1}K) > 0,&\\
\Delta t^2 < \frac{4}{\lambda_{\max}(M^{-1}K)},&
\end{aligned} \tag{3.31}$$

for the LF2 (i.e., order-2 leapfrog) time integration of the acoustic wave equation over the basis of cardinal B-splines. Note that all eigenvalues of $M^{-1}K$ must be real since both matrices are symmetric.

The superposition integrals entering the definition of M (2.15) and K (2.16) have been obtained in the previous section, namely,

$$\begin{aligned}
M_{ij} &= hB_{i,2k+1}((j+k+1)h) = M_{ji},\\
K_{ij} &= -hB''_{i,2k+1}((j+k+1)h) = K_{ji}.
\end{aligned}$$

Notice that we can compute explicitly the matrix entries of M and K. For M, one can combine (3.28) with (3.23) and (3.20), obtaining

$$\begin{aligned}
M_{ij} &= \frac{h}{(2k+1)!}\sum_{r=0}^{k-|i-j|}\binom{2k+2}{r}(-1)^r(k+1-|i-j|-r)^{2k+1},\\
&= h\frac{A(2k+1, k-|i-j|)}{(2k+1)!},
\end{aligned}$$

where $A(n,m)$ is the (n,m)-th *Eulerian number*, which corresponds to the number of permutations of n elements where m couples of elements exchange their relative order. In this formula and the next, the Eulerian numbers must be set to zero when m is out of bounds, i.e., $A(n,m) \coloneqq 0$ for $m < 0$ and $m \geq n$. This relationship between B-splines and Eulerian numbers has been noticed before, see for example [188].

The entries of the stiffness matrix K can similarly be computed explicitly, either applying (3.29) twice to the previous result, or by combining (3.30) with the second derivative with respect to x of (3.23), and again using (3.20). One finds

$$\begin{aligned} K_{ij} &= \frac{1}{h(2k-1)!} \sum_{r=0}^{k-|i-j|} \binom{2k+2}{r} (-1)^r (k+1-|i-j|-r)^{2k-1}, \\ &= \frac{A(2k-1, k-|i-j|) - 2A(2k-1, k-1-|i-j|) + A(2k-1, k-2-|i-j|)}{h(2k-1)!}. \end{aligned}$$

Let us now consider a simple uniform, periodic domain of length L, discretized using a knot vector that decomposes the domain into N elements of size $h \coloneqq L/N$. All the B-spline basis functions of a given order are symmetric about their midpoint and are shifted copies of each other. Consequently, the matrix entries i, j only depend on the difference $|i-j|$, which means that both the mass matrix and the stiffness matrix are symmetric *circulant* matrices, with $2k+1$ non-zero terms in every row and column. A circulant matrix is a matrix where the row $j+1$ is a shifted copy of the first row by j positions to the right. If we denote the first row by $(c_0, c_1, \ldots, c_{N-1})$, the eigenvalues of such a matrix are expressed as

$$\lambda_j = \sum_{r=0}^{N-1} c_r e^{2\pi \mathrm{i} jr/N},$$

with corresponding eigenvectors

$$e_j = \frac{1}{\sqrt{N}} (1, e^{2\pi \mathrm{i} j/N}, \ldots, e^{2\pi \mathrm{i} j(N-1)/N}).$$

For symmetric circulant matrices, one has additionally $c_{N-i} = c_i$.

We can thus compute all the eigenvalues and eigenvectors of the mass and stiffness matrices by using the Poisson summation formula,

$$\sum_{r=-\infty}^{\infty} f(rh) e^{2\pi \mathrm{i} rt} = \sum_{r=-\infty}^{\infty} \widetilde{f}((t+r)/h).$$

After defining $\alpha_j \coloneqq j/N$, we can compute:

$$\begin{aligned} \lambda_j^M &= h \sum_{r=-k}^{k} B_{-k,2k+1}(rh) e^{2\pi \mathrm{i} r\alpha_j}, \\ &= h \sum_{r=-\infty}^{\infty} B_{-k,2k+1}(rh) e^{2\pi \mathrm{i} r\alpha_j}, \end{aligned}$$

$$= h \sum_{r=-\infty}^{\infty} \widetilde{B}_{-k,2k+1}((\alpha_j + r)/h).$$

Using the expression (3.27) for the Fourier transform of $B_{i,k}$, we obtain

$$\begin{aligned}
\lambda_j^M &= h \sum_{r=-\infty}^{\infty} \left(\frac{\sin \pi(\alpha_j + r)}{\pi(\alpha_j + r)} \right)^{2k+2}, \\
&= h \left(\frac{\sin \pi\alpha_j}{\pi} \right)^{2k+2} \sum_{r=-\infty}^{\infty} \left(\frac{1}{(\alpha_j + r)} \right)^{2k+2}, \\
&= h \left(\frac{\sin \pi\alpha_j}{\pi} \right)^{2k+2} \left(\frac{1}{(\alpha_j)^{2k+2}} + \sum_{r=1}^{\infty} \frac{1}{(\alpha_j + r)^{2k+2}} + \sum_{r=1}^{\infty} \frac{1}{(-\alpha_j + r)^{2k+2}} \right), \\
&= h \left(\frac{\sin \pi\alpha_j}{\pi} \right)^{2k+2} \left[\sum_{r=0}^{\infty} \left(\frac{1}{(\alpha_j + r)^{2k+2}} + \frac{1}{(1 - \alpha_j + r)^{2k+2}} \right) \right], \\
&= h \left(\frac{\sin \pi\alpha_j}{\pi} \right)^{2k+2} \left(\zeta_H(\alpha_j, 2k+2) + \zeta_H(1 - \alpha_j, 2k+2) \right), \\
&= h \left(\frac{\sin \pi\alpha_j}{\pi} \right)^{2k+2} \left(\frac{1}{(\alpha_j)^{2k+2}} + \frac{1}{(1 - \alpha_j)^{2k+2}} + \zeta_H(1 + \alpha_j, 2k+2) + \zeta_H(2 - \alpha_j, 2k+2) \right),
\end{aligned}$$

where $\zeta_H(z, s) \coloneqq \sum_{r=0}^{\infty} 1/(r + z)^s$ denotes the *Hurwitz zeta function*. The symmetry of λ_j with respect to the exchange $\alpha_j \leftrightarrow (1 - \alpha_j)$ is evident. In the last expression, we have made explicit the finite limits of the expression for $\alpha_j \to 0$ and $\alpha_j \to 1$ by extracting the respective poles from the zeta functions.

The above expression is clearly decreasing for $0 \leq \alpha_j \leq 1/2$, since the zeta functions decrease more quickly than the two inverse powers. Consequently, the minimum and maximum eigenvalues are attained, respectively, for $j = N/2$, i.e., $\alpha_j = 1/2$ and $j = 0$, i.e., $\alpha_j = 0$:

$$\begin{aligned}
\lambda_{\min}^M &= \frac{2h}{\pi^{2k+2}} \zeta_H(\frac{1}{2}, 2k+2), \\
&= \frac{2h}{\pi^{2k+2}} (2^{2k+2} - 1)\zeta(2k+2), \\
\lambda_{\max}^M &= h.
\end{aligned}$$

We have used here the shift property $\zeta_H(a+1, s) + 1/a^s = \zeta_H(a, s)$ and the identity $\zeta_H(1/2, s) = (2^s - 1)\zeta(s)$, where $\zeta(s)$ is the *Riemann zeta function*. By using the bounds $1 < \zeta(2k+2) \leq \pi^2/6$ and $2^{2k+1} < 2^{2k+2} - 1 < 2^{2k+2}$, valid for all integers $k \geq 0$, we can constrain

$$C_1 h \left(\frac{4}{\pi^2} \right)^k \leq \lambda_{\min}^M \leq C_2 h \left(\frac{4}{\pi^2} \right)^k,$$

with the constants $C_1 = 4/\pi^2$ and $C_2 = 4/3$. Note that, for the mass matrix, this implies that

the condition number $\kappa(M) \coloneqq \lambda^M_{\max}/\lambda^M_{\min}$ is bounded by

$$C_2^{-1}\left(\frac{\pi^2}{4}\right)^k \leq \kappa(M) \leq C_1^{-1}\left(\frac{\pi^2}{4}\right)^k.$$

This is a tighter constraint than the one given for more general B-splines in [189], which in one dimension would correspond to $\kappa(M) \sim 4^k$. This is undoubtedly due to the particularly simple and well-conditioned form of the cardinal B-splines.

A similar calculation can be performed for the stiffness matrix,

$$\begin{aligned}
\lambda^K_j &= -\frac{1}{h}\sum_{r=-k}^{k} B''_{-k,2k+1}(kh)e^{2\pi i r\alpha_j}, \\
&= \frac{1}{h}\sum_{r=-\infty}^{\infty} (2\pi)^2(\alpha_j+r)^2\widetilde{B}_{-k,2k+1}((\alpha_j+r)/h), \\
&= \frac{4}{h}\frac{(\sin\pi\alpha_j)^{2k+2}}{(\pi)^{2k}}\left(\zeta_H(\alpha_j,2k)+\zeta_H(1-\alpha_j,2k)\right), \\
&= \frac{4}{h}\frac{(\sin\pi\alpha_j)^{2k+2}}{(\pi)^{2k}}\left(\frac{1}{(\alpha_j)^{2k}}+\frac{1}{(1-\alpha_j)^{2k}}+\zeta_H(1+\alpha_j,2k)+\zeta_H(2-\alpha_j,2k)\right).
\end{aligned}$$

The minimum eigenvalue is obtained for $\alpha_j = 0$, with $\lambda^K_{\min} = 0$. This is not surprising, since the periodic boundary conditions that we have chosen allow non-zero constant functions, which are annihilated by derivatives. The maximum eigenvalue can be located by setting the derivative of the above expression to zero. This gives two locations $\alpha^{\pm}_j$, whose value can be found numerically by solving the equation:

$$\tan\pi\alpha^-_j = \pi\frac{(k+1)}{k}\frac{\zeta_H(\alpha^-_j,2k)+\zeta_H(1-\alpha^-_j,2k)}{\zeta_H(\alpha^-_j,2k+1)-\zeta_H(1-\alpha^-_j,2k+1)},$$

with other value given by $\alpha^+_j \coloneqq 1-\alpha^-_j$. In the special case $k=0$, only one maximum is found for $\alpha^+_j = \alpha^-_j = 1/2$. In all cases, we can bound the eigenvalues of K from above by noticing that $\lambda^K_j = 4/h^2(\sin\pi\alpha_j)^2\lambda^M_j(k-1)$, and remembering that $\lambda^M_{\max} = h$ is independent of k. Thus:

$$\begin{aligned}
\lambda^K_{\min} &= 0, \\
\lambda^K_{\max} &\leq \frac{4}{h}.
\end{aligned}$$

Finally, the matrices M and K are diagonalized by the same discrete Fourier transform defined above. Consequently, the eigenvalues of $M^{-1}K \coloneqq T$ can be obtained simply as the pointwise ratio of the eigenvalues of K and M. We have

$$\lambda^T_j = \frac{(2\pi)^2}{h^2}\frac{\zeta_H(\alpha_j,2k)+\zeta_H(1-\alpha_j,2k)}{\zeta_H(\alpha_j,2k+2)+\zeta_H(1-\alpha_j,2k+2)}.$$

Again, the minimum eigenvalue is obtained for $\alpha_j = 0$, with $\lambda^T_{\min} = 0$. The numerator grows more rapidly than the denominator for $0 \leq \alpha_j \leq 1/2$, so that the maximum is again located at

$\alpha_j = 1/2$. We obtain finally

$$\lambda_{\min}^T = 0,$$
$$\lambda_{\max}^T = \frac{(2\pi)^2}{h^2}\frac{2^{2k}-1}{2^{2k+2}-1}\frac{\zeta(2k)}{\zeta(2k+2)},$$

from which one can easily extract the following bounds and asymptotic behavior,

$$\frac{\pi^2}{h^2} \leq \lambda_{\max}^T \leq \frac{12}{h^2} \text{ for } k \geq 1,$$
$$\lim_{k\to\infty} \lambda_{\max}^T = \frac{\pi^2}{h^2}.$$

This means that the CFL condition can be bounded by a constant, dependent only on h and not on k. For the leapfrog timestepping scheme, (3.31) becomes

$$\Delta t < \frac{h}{\sqrt{3}}. \tag{3.32}$$

It should be strongly emphasized that, when increasing the polynomial order of a Lagrange basis, new nodes (degrees of freedom) are added to the interior of the elements. Consequently, the minimum wavelength that a Lagrange basis of order k is capable of resolving scales like $\hat{h} \sim h/k$, with the equality if we choose equally spaced nodes. This is in contrast with B-spline basis functions, that can only resolve a minimum wavelength h at all orders k, since no new knots are added to the knot vector. Consequently, it is misguided to compare the CFL condition (3.32) directly to the CFL condition of a FEM scheme of the same order, since the meaning of the spacing h is not the same. In fact, the spatial resolution of the B-spline functions for a fixed knot vector does not change with the degree k, whereas the spatial resolution of a FEM basis scales like $1/k$. Therefore, the correct way to compare the two values is to decrease, in the case of B-splines, the size of the knot vector by a factor k as the order k increases. With this modification, the CFL condition (3.32) becomes

$$\Delta t < \frac{h}{\sqrt{3}k}.$$

This still represents a gain of a factor k in the denominator with respect to FEM schemes, making B-splines advantageous for their use in an explicit time integration scheme. In Figure 3.7 we show the computed CFL timestep stability limit as a function of k for the periodic 1D domain with $c = 1$, $h = 1$ and a second-order central difference leapfrog scheme. We compare it with the maximum timestep over the same configuration for the SEM, FEM (Lagrange or Bernstein-Bézier basis) methods and the Bernstein-Bézier DG method with three different penalty terms, $\alpha, 2\alpha$ and 4α, where α is the minimum value that ensures coercivity (see [89]). Notice the $1/k$ behavior of B-spline bases, contrasted with the $1/k^2$ trend of other methods. In the same figure, we also show a plot of the relative error of the solution to the acoustic wave equation with a sinusoidal source in the same infinite homogeneous model. The solutions are compared at a fixed distance from the source, for multiple mesh sizes h and multiple numbers of degrees of

freedom, i.e., size of the mass and stiffness matrices. The convergence rate $1/h^k$ is evident. At the bottom of the figure, we show the same comparison, but between the IGA (B-spline basis) and DG methods. For the DG method, we have chosen a penalization 2α. Notice that, despite a generally higher error for the IGA method, the order of convergence is the same. Moreover, B-spline functions have a higher efficiency per degree of freedom, achieving the same precision as the DG method with many less degrees of freedom per unit length.

Overall, the cardinal B-spline bases provide an advantage especially for higher orders over practically all other methods except finite differences. The similarity of the B-spline CFL condition with that of finite differences is not completely surprising, since the large stencil, implicit in the divided difference formula (3.26), bears some similarities with the stencils used in finite difference methods.

3.3 Discussion and further reading

The trade-off between the uniformity of the approximation and the spacing of the grid points (and thus conditioning of the problem matrices) constituted the *leitmotif* of the previous section. We refer the reader to [159] for some interesting discussions on this topic.

Concerning the spectral element method, the direct application of the SEM scheme to simplicial cells requires finding the simplicial equivalent of the Gauss-Lobatto quadrature points. This problem can be equivalently formulated as the problem of finding a distribution of point charges that minimizes a certain electrostatic potential, see, e.g., [190]. Generally, such problems are called *Fekete problems*, from the name of the mathematician who first proposed one of them [191]. Unfortunately, these are considered as intrinsically hard problems, as evidenced by the fact that Smale decided to include one of them, namely the Fekete problem on the sphere, as the 7th item of his list of challenging mathematical problems for the 21st century [192]. Thus, as of today, the spectral element approach is generally limited to tensor-product bases. Nonetheless, efficient quadrature rules for simplices have been proposed, see, e.g., [160, 162–164] and Figure 3.2.

The work [169] is dedicated to the coupling of SEM with a method based on an unstructured mesh, such as the DG method. The resulting scheme achieves a good performance in subregions of simple geometry, which are meshed using hexahedra and where SEM is applied, while retaining the geometric flexibility of an unstructured mesh in zones of complex geometry. Spectral elements are not the only approximation method used to simplify the inversion of the mass matrix. We cite here the *mass lumping technique*, see, e.g., [161], and the use of *weight-adjusted* mass matrices, see, e.g., [193].

The use of B-splines as basis functions for the Galerkin method has spawned the field of isogeometric analysis, for which we refer to [2]. Due to the discussed smoothness properties of B-splines, this method is particularly well-suited for physical systems including higher space derivatives, such as for example the Cahn-Hillard equation [194] in multiphase materials, the polyharmonic behavior of thin plates found in Reissner-Mindling plate theory [195, 196], or the Willmore flow of differential geometry [197]. Some of these applications are discussed in [2] and the many references therein.

Dedicated quadratures for B-spline functions and NURBS (Non-Uniform Rational B-Splines)

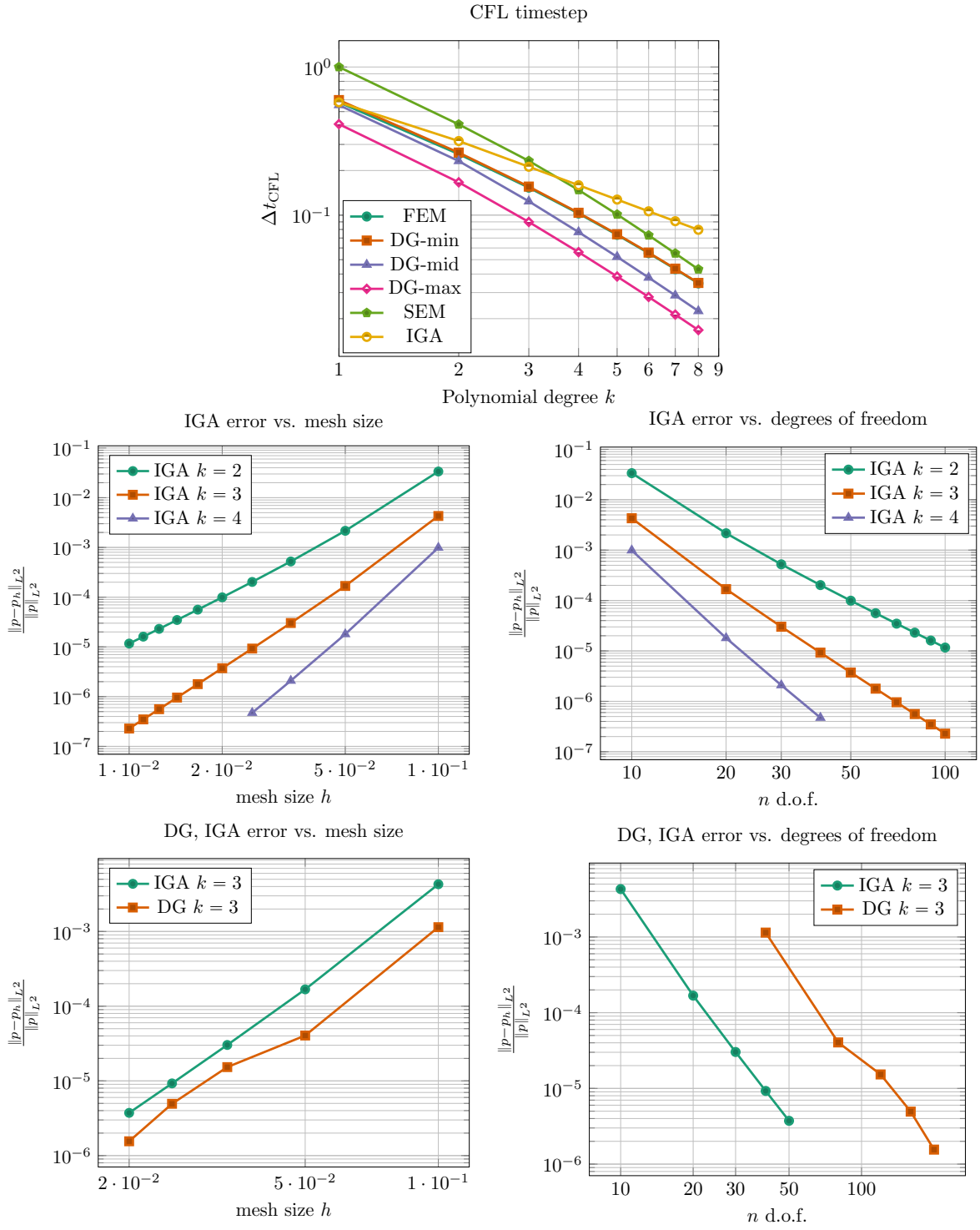


Figure 3.7: (Top) the timestep Δt_{CFL} as a function of polynomial degree k for an infinite 1D medium with $\rho = c = 1$ and $h = 1$, for SEM, FEM, IGA, and the DG method with penalty terms, α (min), 2α (mid) and 4α (max), where α is the minimum value ensuring coercivity. (Middle) relative error of the solution of the acoustic wave equation with a sinusoidal source, as a function of the mesh size h and the number of degrees of freedom per unit length. (Bottom) same comparison, but between IGA and DG. For the DG method, the penalization is 2α.

have also been explored, based on special collocation points and the theory of orthogonal polynomials, see, e.g., [198–201].

Finally, for the CFL condition in DG and FEM method, one can refer for example to [3, 159]. We also wish to point out an interesting comparison between the spectral properties of the SEM and IGA (B-spline) methods, contained in [202].

4 | Simplex spline functions

> Such is our way of thinking–we find beauty not in the thing itself but in the patterns of shadows, the light and the darkness, that one thing against another creates. [...] Were it not for shadows, there would be no beauty.
>
> *Jun'ichirō Tanizaki, In Praise of Shadows (1933)*

Bernstein polynomials and B-splines have originally been developed as versatile tools for modelling and design, capable of reproducing many of the most common shapes, both natural and artificial, within a unified, natural and advantageous framework. When introducing these objects in Chapter 3, we have chosen to follow this historical point of view, which usually begins by assuming some recursive definition, often more or less justified by empirical interpolation mechanisms, from which one then proceeds to derive all the relevant properties of these bases. This is also the approach followed by most introductory texts, both modern and classical (see, e.g., [2, 175]). However, this style of presentation tends to hide some important structure that underlies the complex and beautiful mathematics of B-splines, and makes some of their most important properties, such as regularity and polynomial reproduction, appear somewhat mysteriously from an apparently *ad-hoc*, unrelated definition. Moreover, the axiomatic form of the fundamental recursion relation appears quite rigid and it is unclear how it could be modified while preserving some of the properties of B-spines, discouraging the development of generalizations.

In this chapter, we give a more natural-looking definition of B-splines (and, as a consequence, also Bernstein polynomials) that lays bare some of the simplicity and naturalness of these objects, and helps shed some light on their connections to other fields of mathematics. We discuss some of these connections in the last section. This presentation style allows us to unify some results and points of view that can be found in the literature, but whose connection to one another and to B-splines is sometimes a little weak and indirect. Crucially, this approach allows us to construct a framework for unstructured multivariate B-splines and their use as a geometric tool for numerical analysis. We derive some more or less known numerical properties of a family of unstructured splines called *simplex splines*, which will be pivotal in the construction of unstructured spline spaces in Chapter 5. Furthermore, some important equations relating B-splines that arise naturally in this formulation can be used to formulate a satisfactory set of degrees of freedom for seismic inversion. Some generalizations, which become natural in this language, are also be briefly discussed. Finally, in the last section, we give some useful references

which can be used as a starting point to explore the tantalizing connections of unstructured splines with other branches of mathematics and physics.

4.1 B-splines as shadows of simplices

Soon after the introduction of B-spline functions via the Cox-De Boor recurrence relations (3.14) and (3.15), a curious connection with high-dimensional simplices was found by Curry and Schoenberg [185].

Theorem 4.1.1 (Curry and Schoenberg)**.** *Let Σ^{k+1} be a $(k+1)$-dimensional* simplex*, i.e., the convex hull of $k+2$ affinely independent points $(p_1, \dots, p_{k+2})$ in $\mathbb{R}^{k+1}$, and let $\mu : \mathbb{R}^{k+1} \to \mathbb{R}$ be the projection on the first coordinate. For a given $x \in \mathbb{R}$, define the* shadow *of Σ^{k+1} at x as the intersection of the fiber of the projection at x and the simplex, i.e., $\sigma_x \coloneqq \mu^{-1}(x) \cap \Sigma^{k+1}$. Let*

$$M(x \mid p_1, \dots, p_{k+2}) \coloneqq \operatorname{vol}(\sigma_x) / \operatorname{vol}(\Sigma^{k+1}) \tag{4.1}$$

be the function that associates to every point $x \in \mathbb{R}$ the normalized shadow of the simplex at x. Then, M is proportional to a B-spline basis function of degree k with knots $a_i = \mu(p_i)$, $i = 1, \dots, k+2$, and satisfies the normalization condition:

$$\int_{\mathbb{R}} M(x \mid p_1, \dots, p_{k+2}) \, \mathrm{d}x = 1. \tag{4.2}$$

Proof. The result is well-known, and proofs can be found in the cited papers and in [203, Chapter 18], to which we refer for a more in-depth reading. However, we provide here a constructive proof, as it is instructive to see this geometric connection directly.

For $k = 0$, the simplex Σ^1 reduces to a segment and μ is the identity. The shadow of the simplex is simply the characteristic function of the segment, which corresponds to the B-spline basis function of degree zero defined by (3.14). Normalization is obtained by dividing by the segment length $\operatorname{vol}(\Sigma^1)$.

Consider now a simplex Σ^{k+1} embedded in $\mathbb{R}^{k+1}$, and order its vertices so that the corresponding projections $a_i \coloneqq \mu(p_i)$ are ordered. The edge $\overline{p_1 p_{k+2}}$ connecting the two vertices with the smallest and largest projections must belong to the simplex since every couple of vertices in a simplex is connected. Consider a real value $x \in \mathbb{R}$. If $x \notin [a_1, a_{k+2}]$, then σ_x is empty and $M(x \mid p_1, \dots, p_{k+2}) = 0$, which is in accordance with the recursion formula (3.15) evaluated outside the convex hull of the spline knots. If $x \in [a_1, a_{k+2}]$, then σ_x contains at least a point on the edge $\overline{p_1 p_{k+2}}$, which we will denote by $\hat{x}$. We can split the simplex Σ^{k+1} into two disjoint simplices Σ_1^{k+1} and Σ_2^{k+1} defined respectively as the convex hulls of the sets of points $(\hat{x}, p_2, \dots, p_{k+2})$ and $(p_1, \dots, p_{k+1}, \hat{x})$. To see why this is true, notice that we can take any point $\hat{y}$ in the interior of Σ^{k+1} and connect it to every vertex, thus obtaining a triangulation of the original simplex into $k+2$ disjoint sub-simplices. In the limit where $\hat{y}$ lies on the edge $\overline{p_1 p_{p+2}}$, the volumes of the k simplices containing all three points $\hat{y}$, p_1 and p_{k+1} degenerate to zero. If we now denote by σ_x^1 and σ_x^2 the intersections $\mu^{-1}(x) \cap \Sigma_1^{k+1}$ and $\mu^{-1}(x) \cap \Sigma_2^{k+1}$ respectively,

the following relations hold,

$$\mathrm{vol}(\Sigma^{k+1}) = \mathrm{vol}(\Sigma_1^{k+1}) + \mathrm{vol}(\Sigma_2^{k+1}), \;\; \mathrm{vol}(\sigma_x) = \mathrm{vol}(\sigma_x^1) + \mathrm{vol}(\sigma_x^2),$$

from which we obtain

$$\frac{\mathrm{vol}(\sigma_x)}{\mathrm{vol}(\Sigma^{k+1})} = \frac{\mathrm{vol}(\sigma_x^1)}{\mathrm{vol}(\Sigma_1^{k+1})}\frac{\mathrm{vol}(\Sigma_1^{k+1})}{\mathrm{vol}(\Sigma^{k+1})} + \frac{\mathrm{vol}(\sigma_x^2)}{\mathrm{vol}(\Sigma_2^{k+1})}\frac{\mathrm{vol}(\Sigma_2^{k+1})}{\mathrm{vol}(\Sigma^{k+1})}. \tag{4.3}$$

The ratio $\mathrm{vol}(\Sigma_1^{k+1})/\,\mathrm{vol}(\Sigma^{k+1})$ must depend linearly on x, since the volume of Σ_1^{k+1} can be written as a product $\mathrm{vol}(B_1^k)h_x/(k+1)$ of the volume of the base simplex B_1^k, defined as the convex hull of $(p_2, \ldots, p_{k+2})$, times the height h_x of the simplex, which is just the difference $\hat{x} - p_{k+2}$ projected onto the normal to B_1^k. Since this ratio is 0 for $x = a_{k+2}$ and 1 for $x = a_1$, we simply have $\mathrm{vol}(\Sigma_1^{k+1})/\,\mathrm{vol}(\Sigma^{k+1}) = (a_{k+2} - x)/(a_{k+2} - a_1)$. Similarly, $\mathrm{vol}(\Sigma_2^{k+1})/\,\mathrm{vol}(\Sigma^{k+1}) = (x - a_1)/(a_{k+2} - a_1)$.

Let us now consider the ratio $\mathrm{vol}(\sigma_x^1)/\,\mathrm{vol}(\Sigma_1^{k+1})$. Notice that σ_x^1 is itself a k-dimensional simplex, since it is obtained by slicing a $(k+1)$-dimensional simplex with a plane passing through one of its vertices, namely $\hat{x}$. Its volume can therefore be expressed as

$$\mathrm{vol}(\sigma_x^1) = \frac{1}{k}\,\mathrm{vol}(\beta_x^1)\widetilde{h}_x, \tag{4.4}$$

where $\beta_x^1 = \mu^{-1}(x) \cap B_1^k$ is the base of σ_x^1 and the height $\widetilde{h}_x$ is the distance between $\hat{x}$ and the $(k-1)$-dimensional hyperplane containing β_x^1. Let ρ be the affine shear transformation that preserves the fibers of μ, leaves σ_x^1 unchanged and transforms the base B_1^k into $\rho(B_1^k)$, orthogonal to the height $\widetilde{h}_x$. We have that $\rho(\beta_x^1) = \beta_x^1$, $\rho(\hat{x}) = \hat{x}$ and

$$\mathrm{vol}(\rho(\Sigma_1^{k+1})) = \frac{1}{k+1}\widetilde{h}_x\,\mathrm{vol}(\rho(B_1^k)) = \mathrm{vol}(\Sigma_1^{k+1}), \tag{4.5}$$

since shear transformations are volume-preserving. Dividing (4.4) by (4.5) we finally obtain

$$\frac{\mathrm{vol}(\sigma_x^1)}{\mathrm{vol}(\Sigma_1^{k+1})} = \frac{k+1}{k}\frac{\mathrm{vol}(\beta_x^1)}{\mathrm{vol}(\rho(B_1^k))}.$$

Since $\mu^{-1}(x) \cap B_1^k \subset \sigma_x^1$, then $\mu^{-1}(x) \cap \rho(B_1^k) = \mu^{-1}(x) \cap B_1^k = \beta_x^1$, and the second member of 4.1 is proportional to a lower-dimensional M function:

$$\frac{\mathrm{vol}(\sigma_x^1)}{\mathrm{vol}(\Sigma_1^{k+1})} = \frac{k+1}{k}M(x \mid \widetilde{p}_2, \ldots, \widetilde{p}_{k+2}), \tag{4.6}$$

where $\widetilde{p}_i$ is just $\rho(p_i)$, seen as a k-dimensional point after identifying the hyperplane containing B_1^k with $\mathbb{R}^k$. Notice that $\mu(\widetilde{p}_i) = \mu(p_i) = a_i$, so the knot values are preserved.

The same process can be repeated for the other ratio $\mathrm{vol}(\sigma_x^2)/\,\mathrm{vol}(\Sigma_2^{k+1})$. Substituting into (4.6), one obtains

$$\frac{k}{k+1}M(x \mid p_1, \ldots, p_{k+2}) = \frac{x - a_1}{a_{k+2} - a_1}M(x \mid \widetilde{p}_1, \ldots, \widetilde{p}_{k+1}) + \frac{a_{k+2} - x}{a_{k+2} - a_1}M(x \mid \widetilde{p}_2, \ldots, \widetilde{p}_{k+2}).$$

We can recognize the Cox-De Boor recursion formula (3.15) after relating the unnormalized and normalized B-spline basis functions as follows:

$$N(x \mid a_1, \ldots, a_n) = \frac{a_n - a_1}{n-1} M(x \mid p_1, \ldots, p_n),$$

thus proving that the M functions are proportional to B-spline basis functions.

Finally, the normalization condition (4.2) is a simple consequence of the definition given in (4.1) and Fubini's theorem, since the integrand is always nonnegative. □

In other words, a B-spline basis function of degree k can be naturally seen as the one-dimensional *shadow* of a $(k+1)$-dimensional simplex. We show in Figure 4.1 the geometric intuition behind Theorem 4.1.1.

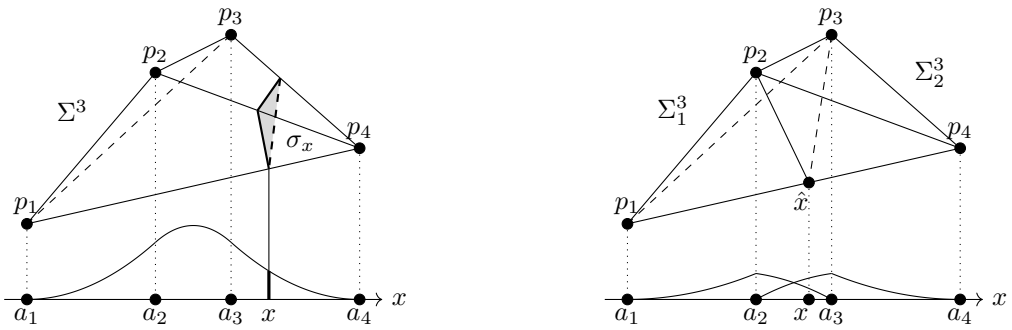


Figure 4.1: (Left) interpretation of a B-spline function of degree k as projection of a $(k+1)$-dimensional simplex. (Right) the geometric splitting process at the origin of the recurrence formula. The two plotted functions correspond to the two summands on the right hand side of (4.3).

For obvious reasons, splines obtained under this definition have been called *simplex splines* (see, e.g., [203–207]). Theorem 4.1.1 indirectly implies that, even though the definition (4.1) depends on the choice of $k+2$ points $p_i \in \mathbb{R}^{k+1}$, the resulting spline only depends on the knots, i.e., the projections $a_i = \mu(p_i)$. This can also be directly seen by noticing that any affine transformation that leaves the projections $\mu(p_i)$ unaffected will produce the same simplex spline, since the ratios of volumes are preserved. Consequently, we will denote the simplex splines by only giving the positions $A \coloneqq (a_i)_{i=1}^{k+2} \subset \mathbb{R}$ of the projections of the simplex vertices,

$$M(x \mid A) \coloneqq M(x \mid a_1, \ldots, a_{k+2}).$$

As a consequence of 4.1.1, given any continuous function $f : \mathbb{R} \to \mathbb{R}$, its superposition integral with a simplex spline can be computed as follows,

$$\int_{\mathbb{R}} f(x) M(x \mid A)\, \mathrm{d}x = \int_{\Sigma} f \circ \mu_A \, \mathrm{d}\Sigma,$$

with $\mu_A : \Sigma \to \mathbb{R}$, also called the *moment map*, denoting the affine map

$$\mu_A(\lambda_1, \ldots, \lambda_{k+2}) \coloneqq \sum_{i=1}^{k+2} \lambda_i a_i =: \Lambda \cdot A,$$

where $\lambda_1, \ldots, \lambda_{k+2}$ are real-valued *barycentric coordinates*, Σ is the *standard simplex* defined by imposing the constraints $\lambda_i \geq 0$ for all i and $\sum_{i=1}^{k+2} \lambda_i = 1$ on the barycentric coordinates, and $\mathrm{d}\Sigma = \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_{k+2} / \operatorname{vol}(\Sigma)$ is the standard Lebesgue uniform measure on the simplex divided by the simplex volume. In other words, the measure induced by a B-spline function can be interpreted as the regular flat Euclidean measure on a unit simplex after composition with an affine projection mapping the $k+2$ simplex vertices to the knots defining the spline. It should be noted that this construction is valid, without modification, even with repeated knot values: in this case, more than one simplex vertex will be mapped to the same knot. In fact, (4.7) can be taken as a *functional definition* of normalized B-splines, skipping entirely the recursive definition presented in Chapter 3. This approach is very fruitful in view of the application of splines in Galerkin methods, which simplifies the derivation of many properties and lends itself to many generalizations, and it is also the approach taken for example by Micchelli in his seminal paper [205]. For this reason, this is the point of view that we will adopt from this point onward: we elevate (4.7) to a definition of a normalized (simplex) spline as a linear functional on $L^2(\mathbb{R})$, that associates to f its integral over a simplex under the moment map defined by its knots. We therefore let

$$\langle f, M(A) \rangle \coloneqq \int_{\mathbb{R}} f(x) M(x \mid A) \, \mathrm{d}x = \int_{\Sigma} f \circ \mu_A \, \mathrm{d}\Sigma \tag{4.7}$$

be the functional definition of the function $M(A) \coloneqq M(\cdot \mid A)$.

4.2 Generalization via Dirichlet measures

We have seen in the previous section a very natural geometric interpretation of B-splines as shadows of higher-dimensional simplices. However, the derivation of the recurrence formula for B-splines from this purely geometrical formulation has proven to be quite awkward and convoluted. One of the main reasons is that, when B-splines of lower order are introduced, they no longer correspond to simplices of the same dimension, but to facets or other lower-dimensional objects. One can avoid such problems by incorporating the idea of *multiplicity* or *repetitions* of a knot, which is a natural number and can become zero if the knot is removed. In order to achieve this, the standard Lebesgue measure $\mathrm{d}\Sigma$ must be replaced by a *Dirichlet measure*, a generalization borrowed from statistics.

4.2.1 Dirichlet measures and Dirichlet averages

Let $A \coloneqq (a_i)_{i=1}^n \subset \mathbb{R}$ be a vector of non necessarily distinct knots, and let $\Sigma \coloneqq \Sigma^{n-1}$ represent, as in (4.7), the standard simplex in $\mathbb{R}^{n-1}$ defined by the barycentric coordinates $\lambda_1, \ldots, \lambda_n$ satisfying $0 \leq \lambda_i \leq 1$ for each i and $\sum_{i=1}^n \lambda_i = 1$. Define a vector of *multiplicities* $R = (r_1, \ldots, r_n)$ of length n, such that $r_i \in \mathbb{C}_+$, where $\mathbb{C}_+$ represents the set of complex numbers with positive

real part. The *Dirichlet measure* on Σ is then defined as

$$\mathrm{d}\Sigma_R \coloneqq \frac{1}{B(R)} \lambda_1^{r_1-1} \cdots \lambda_n^{r_n-1} \, \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_n, \tag{4.8}$$

where $B(R)$ is the *multivariate Beta function* defined by

$$B(R) \coloneqq \frac{\Gamma(r_1) \cdots \Gamma(r_n)}{\Gamma(r_1 + \cdots + r_n)}, \tag{4.9}$$

and Γ is the Euler Gamma function. Note that the conditions $\mathrm{Re}(r_i) > 0$ are required to avoid the appearance of poles in the Gamma functions, although they will later need to be relaxed. We show a few examples of Dirichlet measures in Figure 4.2.

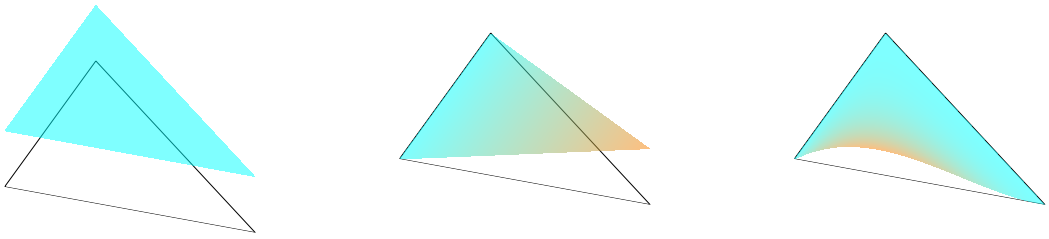


Figure 4.2: Three examples of Dirichlet distributions on a two-dimensional simplex Σ^2, corresponding to different choices of repetitions r_1, r_2 and r_3 in (4.8). (Left) the case $r_1 = r_2 = r_3 = 1$ (i.e., no repetitions). (Middle) The case $r_1 = r_2 = 1$, $r_3 = 2$. (Right) the case $r_1 = 1$, $r_2 = 3$ and $r_3 = 2$.

Given the vector R of multiplicities and the Dirichlet measure defined by (4.8), one can generalize the functional definition (4.7) as

$$\langle f, M(A, R) \rangle \coloneqq \int_\Sigma f \circ \mu_A \, \mathrm{d}\Sigma_R = \frac{1}{B(R)} \int_\Sigma f(\Lambda \cdot A) \lambda_1^{r_1-1} \cdots \lambda_n^{r_n-1} \, \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_n. \tag{4.10}$$

Notice that if $R = (1, \ldots, 1)$, then $B(R) = 1/(n-1)! = 1/\operatorname{vol}(\Sigma)$, and the measure reduces to the standard normalized Lebesgue measure $\mathrm{d}\Sigma$ on the $(n-1)$-dimensional simplex. One then recovers (4.7), which indeed corresponds to the spline $M(x \mid A)$ with all knot multiplicities equal to one. Notice also that the Dirichlet measure (4.8) is normalized to one for all valid choices of multiplicities, i.e., $\int_\Sigma \mathrm{d}\Sigma_R = 1$. For general parameters $r_i \in \mathbb{C}_+$, the term on the right hand side of (4.10) is called the *Dirichlet average* of f [208].

Using this expression, one can change the number of repetitions of any knot in A by changing the corresponding multiplicity in R. When applying (4.10), the corresponding integral is still formulated over the same $(n-1)$-dimensional simplex Σ, even if the degree of the corresponding spline function is different. This feature of the Dirichlet average formulation allows to derive much more easily many of the properties of unstructured (simplex) splines.

4.2.2 Integer parameters and knot multiplicities

Let us now state a few more properties of Dirichlet averages that cement the interpretation of R as a vector of knot multiplicities. Many of these properties can be recovered from equivalent statements on Dirichlet averages found in [208], but since the connection between Dirichlet averages and spline functions was only recognized many years after the publication of this book [209, 210], these statements were never directly translated to the language of spline functions. Therefore, we give here a self-contained presentation.

First, notice that the simplex spline definition (4.10) does not depend on the order of the knots.

Theorem 4.2.1 (Carlson)**.** *Let $\sigma \in S_n$ be a permutation of the knot vector A. Then*

$$\langle f, M(\sigma(A), \sigma(R))\rangle = \langle f, M(A, R)\rangle.$$

Remark 4.2.2. *This theorem corresponds to [208, Theorem 5.2-3].*

Proof. It is clear from the definition(4.10). In fact, since $\Lambda \cdot A = \sigma(\Lambda) \cdot \sigma(A)$, then permuting the knots in the scalar product can be undone by permuting the corresponding barycentric coordinates. If we also apply the same permutation to the multiplicity vector R, then both the Dirichlet measure (4.8) and the simplex Σ are left unchanged. □

Second, notice that nothing prevents us from increasing the multiplicity of a knot $a_i \in A$ by inserting a_i (anywhere) as a new component of A, obtaining a new vector of length $n+1$, which we denote by $A \sqcup \{a_i\}$. The same effect can be achieved by instead increasing the multiplicity of a_i in R.

Theorem 4.2.3 (Carlson)**.** *Suppose that the function f is continuous, and that the two nodes a_i and a_j in A are equal. Let $R_{i\overleftarrow{+}j}$ be equal to R with the value r_i replaced by $r_i + r_j$ and the value r_j omitted. Then*

$$\langle f, M(A, R)\rangle = \big\langle f, M(A \setminus \{a_j\}, R_{i\overleftarrow{+}j})\big\rangle.$$

Remark 4.2.4. *This theorem corresponds to [208, Theorem 5.2-4].*

Proof. Because of Theorem 4.2.1, we can without loss of generality assume $i = n-1$ and $j = n$. We have

$$\langle f, M(A, R)\rangle = \frac{1}{B(R)} \int_\Sigma f(\Lambda \cdot A)\lambda_1^{r_1-1} \cdots \lambda_n^{r_n-1}\, \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_n.$$

Let us introduce the change of variables $\nu_i \coloneqq \lambda_i$ for $i = 1, \ldots, n-2$ and $w_1\nu_{n-1} \coloneqq \lambda_{n-1}$, $w_2\nu_{n-1} \coloneqq \lambda_n$, where $\nu_{n-1} = 1 - \nu_1 - \cdots - \nu_{n-2}$, and the variables are subject to the constraints $0 \le \nu_i \le 1$ for $i = 1, \ldots, n-1$, $w_1, w_2 \ge 0$ and $w_1 + w_2 = 1$. The Jacobian associated to this change of variables can be easily computed to be $J = \nu_{n-1}$, and the variables w_1, w_2 simply

span the interval $[0,1]$. We have

$$\begin{aligned}
&\langle f, M(A,R)\rangle = \\
&= \frac{1}{B(R)} \int_{\Sigma^{n-1}} f(\Lambda \cdot A) \lambda_1^{r_1-1} \cdots \lambda_n^{r_n-1} \,\mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_n, \\
&= \frac{1}{B(R)} \int\limits_{\Sigma^{n-2}\times\Sigma^1} f(\sum_{i=1}^{n-1} \nu_i a_i) \nu_1^{r_1-1} \cdots \nu_{n-1}^{r_{n-1}+r_n-2+1} w_1^{r_{n-1}-1} w_2^{r_n-1} \,\mathrm{d}\nu_1 \cdots \mathrm{d}\nu_n \,\mathrm{d}w_1 \,\mathrm{d}w_2.
\end{aligned}$$

The integration domain is factorized into a $(n-2)$-dimensional simplex Σ^{n-2} times the one-dimensional simplex (interval) Σ^1 spanned by the auxiliary variables w_1 and w_2, subject to $w_1 + w_2 = 1$. Since the integrand is separable, the real parts of the r_i are positive, and f is continuous, we can apply Fubini's theorem to integrate out the two auxiliary variables,

$$\int_{\Sigma^1} w_1^{r_{n-1}-1} w_2^{r_{n-1}-1} \,\mathrm{d}w_1 \,\mathrm{d}w_2 = \frac{\Gamma(r_{n-1}+r_n)}{\Gamma(r_{n-1})\Gamma(r_n)}.$$

When this factor is multiplied by $1/B(R)$, the terms $\Gamma(r_{n-1})\Gamma(r_n)$ at the numerator get replaced by $\Gamma(r_{n-1}+r_n)$ (see (4.9)). Since the sum of all the elements of R is equal to the sum of all the elements of $R_{(n-1)\overleftarrow{+}(n)}$, the resulting factor is simply $1/B(R_{(n-1)\overleftarrow{+}(n)})$. Thus,

$$\begin{aligned}
\langle f, M(A,R)\rangle &= \frac{1}{B(R_{ij})} \int_{\Sigma^k} f(\nu_1 a_1 + \cdots + \nu_{n-1} a_{n-1}) \nu_1^{r_1-1} \cdots \nu_{n-1}^{r_{n-1}+r_n-1} \,\mathrm{d}\nu_1 \cdots \mathrm{d}\nu_n, \\
&= \Big\langle f, M(A \setminus \{a_n\}, R_{(n-1)\overleftarrow{+}(n)}) \Big\rangle,
\end{aligned}$$

proving the theorem. □

Notice that this property is valid also for noninteger multiplicities, and in fact it is valid for general vectors R with components in $\mathbb{C}_+$, therefore representing a genuine generalization of usual knot repetitions.

Finally, we have not yet discussed how a knot can be removed from A using the multiplicity vector R. In fact, the condition $\mathrm{Re}(r_i) > 0$, due to the pole of the function $\Gamma(r_i)$ in $B(R)$, seems to preclude such a possibility. In general, it should not be possible to completely eliminate a knot in (4.10) in every case. For example, assume that there are exactly two knots a_1 and a_2 in A, with respective multiplicities $R = (1, r)$, and consider $f(x) = 1$ if $a_1 < x < a_2$ and zero otherwise. Then,

$$\langle f, M(A,R)\rangle = \frac{\Gamma(1+r)}{\Gamma(r)} \int_0^1 f(a_1 + \lambda(a_2 - a_1)) \lambda^{r-1} \,\mathrm{d}\lambda = r \int_0^1 \lambda^{r-1} \,\mathrm{d}\lambda = 1,$$

for all positive values of r, in accordance with the normalization of the Dirichlet measure. Thus, the limit of this expression as $r \to 0$ is 1. In contrast, removing a_2 from A altogether yields $\langle f, M(A,R)\rangle = f(a_1) = 0$. The reason for this discrepancy is clearly the discontinuity of f.

We now prove that, for continuous functions f, whenever $r_i \to 0$ the corresponding knot can be removed in A. In order to do this, we need two lemmas that will be very useful in the rest of the chapter. These two lemmas can also be derived from some of the statements in Chapters

4, 5 and 6 of [208]. The first lemma is a simple property of the multivariate Beta distribution.

Lemma 4.2.5. *Let R be a vector with $r_i \in \mathbb{C}_+$, and let*

$$\begin{aligned} c &\coloneqq r_1 + \cdots + r_n, \\ w_i &\coloneqq \frac{r_i}{c}. \end{aligned}$$

We denote by E_i the unit vector of size n with components $(E_i)_j = \delta_{ij}$, so that $R + E_i$ can be seen as the result of increasing the multiplicity of knot i by 1. We have

$$B(R + E_i) = w_i B(R). \tag{4.11}$$

Proof. From the definition (4.9),

$$B(R+E_i) = \frac{\Gamma(r_1)\cdots\Gamma(r_i+1)\cdots\Gamma(r_n)}{\Gamma(r_1+\cdots+r_n+1)} = \frac{r_i}{r_1+\cdots+r_n}\frac{\Gamma(r_1)\cdots\Gamma(r_i)\cdots\Gamma(r_n)}{\Gamma(r_1+\cdots+r_n)} = w_i B(R).$$

□

The second lemma expresses the integral of the product of a function with a spline in terms of splines of higher degree.

Lemma 4.2.6. *The following identities hold:*

$$\langle f, M(A,R)\rangle = \sum_{i=1}^{n} w_i \langle f, M(A, R+E_i)\rangle, \tag{4.12}$$

$$\langle xf, M(A,R)\rangle = \sum_{i=1}^{n} w_i a_i \langle f, M(A, R+E_i)\rangle, \tag{4.13}$$

$$\langle (x-a_j)f, M(A,R)\rangle = \sum_{i=1}^{n} w_i (a_i - a_j) \langle f, M(A, R+E_i)\rangle. \tag{4.14}$$

Proof. We have

$$\begin{aligned} \langle f, M(A,R)\rangle &= \frac{1}{B(R)} \int_\Sigma \left(\sum_{i=1}^{n} \lambda_i\right) f(\Lambda\cdot A)\lambda_1^{r_1-1}\cdots\lambda_n^{r_n-1}\,\mathrm{d}\lambda_1\cdots\mathrm{d}\lambda_n, \\ &= \sum_{i=1}^{n} w_i \langle f, M(A, R+E_i)\rangle, \end{aligned}$$

and

$$\begin{aligned} \langle xf, M(A,R)\rangle &= \int_\Sigma \left(\sum_{i=1}^{n} \lambda_i a_i\right) f(\Lambda\cdot A)\lambda_1^{r_1-1}\cdots\lambda_n^{r_n-1}\,\mathrm{d}\lambda_1\cdots\mathrm{d}\lambda_n, \\ &= \sum_{i=1}^{n} w_i a_i \langle f, M(A, R+E_i)\rangle, \end{aligned}$$

where we have used (4.11). Combining these two equations yields the third one. Notice that, over finite domains, if $f(x)$ is integrable, then $xf(x)$ is integrable as well. □

We can now prove a variation of one of Carlson's theorems about analytic continuation of Dirichlet averages.

Theorem 4.2.7 (Carlson)**.** *Let f be a continuous function inside the convex hull of the spline knots. Then,*

$$\lim_{r_i \to 0} \langle f, M(A, R) \rangle = \langle f, M(A \setminus \{a_i\}, R \setminus \{r_i\}) \rangle.$$

Remark 4.2.8. *This theorem is proven for polynomials in [208, Theorem 6.2-4] and for holomorphic functions in [208, Theorem 6.3-3] , although it is never explicitly stated in the book that it also applies to all continuous functions.*

Proof. From (4.13), we know that, for $k > 0$,

$$\Big\langle x^k, M(A, R) \Big\rangle = \sum_{i=1}^{n} w_i a_i \Big\langle x^{k-1}, M(A, R + E_i) \Big\rangle.$$

We can apply (4.13) k times, after which we obtain only terms of the form $\Big\langle 1, M(A, \hat{R}) \Big\rangle = 1$, for some multiplicity vector $\hat{R}$. At this point the recurrence stops, since $\Big\langle 1, M(A, \hat{R}) \Big\rangle = 1$ for all $\hat{R}$. If $r_i \to 0$, but $c = r_1 + \cdots + r_n \not\to 0$, then all the terms containing the variable w_i vanish. Since only these terms contain higher multiplicities of the knot a_i, the chain of recursions can therefore be read backwards discarding the knot a_i at every step, thus proving the theorem for monomials and hence for polynomials.

The proof for a general continuous f then follows from the Stone-Weierstrass theorem, which states that polynomials are uniformly dense in the set of continuous functions. □

Theorems 4.2.1, 4.2.3 and 4.2.7 prove that, for continuous test functions f, there is no distinction between repeating multiple (or zero) times a knot value in A or setting a higher (respectively, zero) multiplicity in R. Thus, we will not make such a distinction, and denote instead by

$$M(x \mid A) \coloneqq M(x \mid A, R)$$

any such choice, in order to avoid redundant and cumbersome notation. We will similarly indicate by

$$M(x \mid A \sqcup \{a_i\}) \coloneqq M(x \mid A, R + E_i)$$

an increase in the multiplicity of the knot a_i by one, and by

$$M(x \mid A \setminus \{a_i\}) \coloneqq M(x \mid A, R - E_i)$$

a decrease in multiplicity by one, including complete removal, of the knot a_i. Finally, for ease of reading, we will write pointwise equalities between spline functions, which should instead be interpreted weakly, i.e.,

$$M_1(x) + M_2(x) = M_3(x) \text{ should be interpreted as } \langle f, M_1 + M_2 \rangle = \langle f, M_3 \rangle.$$

This notation is somewhat justified, since all spline functions of degree $k > 0$ are continuous (as proven by the recurrence formulas that we will derive later), and thus the equality is not valid in the strong sense only for zero-degree splines, which are discontinuous.

After these simplifications, we are ready to present a number of important properties of univariate and (later) multivariate simplex splines.

4.3 Univariate simplex splines and B-splines

In this section, we prove that simplex splines reproduce the usual (normalized) B-splines whenever the multiplicities are all positive integers, and we recover a few important properties of univariate B-spline functions in this new context. Aside from providing a deeper, geometric understanding of some familiar properties, this section and the next one pave the way for the derivation of the corresponding properties of multivariate simplex splines, which are the topic of the next section. Multivariate simplex splines built on knot clouds are the main actors of the formulation of our numerical scheme.

4.3.1 Recurrence formulas

We start by proving a recurrence relation which can be used to compute simplex splines, and reduces to the usual Cox-De Boor rule (3.15) in the case of positive integer multiplicities. We use the following lemma.

Lemma 4.3.1. *The Dirichlet measure defined in* (4.8) *satisfies the following identity for every* $1 \leq i, j \leq n$*:*

$$\left(\frac{\partial}{\partial \lambda_i} - \frac{\partial}{\partial \lambda_j}\right)(\mathrm{d}\Sigma_R) = (c-1)\left(\mathrm{d}\Sigma_{R-E_i} - \mathrm{d}\Sigma_{R-E_j}\right). \tag{4.15}$$

Proof. From definition (4.8), we can see that

$$\begin{aligned}
\frac{\partial}{\partial \lambda_i}(\mathrm{d}\Sigma_{R+E_i+E_n}) &= \frac{1}{B(R+E_i+E_n)}\frac{\partial}{\partial \lambda_i}\left(\lambda_1^{r_1-1}\cdots\lambda_i^{r_i}\cdots\lambda_n^{r_n}\right)\mathrm{d}\lambda_1\cdots\mathrm{d}\lambda_{n-1}, \\
&= r_i\frac{B(R+E_n)}{B(R+E_i+E_n)}\,\mathrm{d}\Sigma_{R+E_n} - r_n\frac{B(R+E_i)}{B(R+E_i+E_n)}\,\mathrm{d}\Sigma_{R+E_i}, \\
&= (c+1)\left(\mathrm{d}\Sigma_{R+E_n} - \mathrm{d}\Sigma_{R+E_i}\right),
\end{aligned}$$

since λ_i appears in $\mathrm{d}\Sigma_{R+E_i+E_n}$ in the term $\lambda_i^{r_i}$, but also in the term $\lambda_n^{r_n}$ through the relation $\lambda_n = 1 - \lambda_1 - \cdots - \lambda_{n-1}$. In the last step, we have made use of (4.11). Substituting the multiplicity vector R in place of $R + E_i + E_n$, which implies substituting $c - 2$ in place of c, yields

$$\frac{\partial}{\partial \lambda_i}(\mathrm{d}\Sigma_R) = (c-1)\left(\mathrm{d}\Sigma_{R-E_i} - \mathrm{d}\Sigma_{R-E_n}\right). \tag{4.16}$$

By computing expression (4.16) with free indices i and j and taking the difference we obtain (4.15). □

Armed with this result, we derive the simplex spline generalization of Cox-De Boor's formula. This relation is equivalent to the expression [208, Relation 5.6-3], using the insight from [209].

Lemma 4.3.2 (Carlson)**.** *The identity*

$$(a_i - a_j)M(x \mid A \setminus \{a_k\}) + (a_j - a_k)M(x \mid A \setminus \{a_i\}) + (a_k - a_i)M(x \mid A \setminus \{a_j\}) = 0 \quad (4.17)$$

holds for all $i, j, k = 1, \ldots, n$.

Proof. Consider the following family of one-parameter transformations R^{ε}_{ijk} of the λ variables,

$$\begin{aligned} \lambda_i' &= \lambda_i + (a_j - a_k)\varepsilon, \\ \lambda_j' &= \lambda_j + (a_k - a_i)\varepsilon, \\ \lambda_k' &= \lambda_k + (a_i - a_j)\varepsilon, \\ \lambda_l' &= \lambda_l,\ l \neq i, j, k, \end{aligned}$$

generated by the differential operator

$$\begin{aligned} \nabla_{ijk} &\coloneqq (a_i - a_j)\frac{\partial}{\partial \lambda_k} + (a_j - a_k)\frac{\partial}{\partial \lambda_i} + (a_k - a_i)\frac{\partial}{\partial \lambda_j}, \\ &= a_i\left(\frac{\partial}{\partial \lambda_k} - \frac{\partial}{\partial \lambda_j}\right) + a_j\left(\frac{\partial}{\partial \lambda_i} - \frac{\partial}{\partial \lambda_k}\right) + a_k\left(\frac{\partial}{\partial \lambda_j} - \frac{\partial}{\partial \lambda_i}\right). \end{aligned}$$

Notice that these transformations are affinely well defined, i.e.,

$$\sum_{m=0}^{n} \lambda_m' = 1 + \varepsilon\left[(a_j - a_k) + (a_k - a_i) + (a_i - a_j)\right] = 1,$$

and they preserve the moment map,

$$\Lambda' \cdot A = \Lambda \cdot A + \left[(a_j - a_k)a_i + (a_k - a_i)a_j + (a_i - a_j)a_k\right]\varepsilon = \Lambda \cdot A.$$

Therefore, even if f is not itself differentiable, we can compute the derivative $\nabla_{ijk} f(\Lambda \cdot A)$ in the direction of R^{ε}_{ijk} as

$$\nabla_{ijk} f(\Lambda \cdot A) \coloneqq \lim_{\varepsilon \to 0} \frac{f(R^{\varepsilon}_{ijk}(\Lambda) \cdot A) - f(\Lambda \cdot A)}{\varepsilon} = 0.$$

We can therefore compute the integral

$$\int_{\Sigma} f(\Lambda \cdot A)\nabla_{ijk}\,(\,\mathrm{d}\Sigma_R)$$

using integration by parts on λ_i, λ_j and λ_k. The limits of integration for the variable λ_i are $\lambda_i = 0$ and $\lambda_n = 0$. Since $\mathrm{Re}(r_i)$ and $\mathrm{Re}(r_n)$ are positive, the boundary term vanishes when integrating by parts. The same argument applies to λ_j and λ_k, leading to

$$\int_{\Sigma} f(\Lambda \cdot A)\nabla_{ijk}\,(\,\mathrm{d}\Sigma_R) = -\int_{\Sigma} \nabla_{ijk}\,(f(\Lambda \cdot A))\,\mathrm{d}\Sigma_R = 0. \quad (4.18)$$

Expanding the first term of (4.18) and inserting (4.15) yields

$$\int_\Sigma f(\Lambda \cdot A)\left((a_i - a_j)\,\mathrm{d}\Sigma_{R-E_k} + (a_j - a_k)\,\mathrm{d}\Sigma_{R-E_i} + (a_k - a_i)\,\mathrm{d}\Sigma_{R-E_j}\right) = 0,$$

proving the lemma. □

From this symmetric relation, one can obtain the following expression, more similar to the standard recursion formulas used for B-splines.

Theorem 4.3.3 (Carlson)**.** *Let $c \coloneqq r_1 + \cdots + r_n$, and let E_i again denote the unit vector of size n with components $(E_i)_j = \delta_{ij}$. Then the following recursion relation holds:*

$$(c-2)(a_n - a_1)M(x \mid A) = (c-1)\left((x - a_1)M(x \mid (A) \setminus \{a_n\}) + (a_n - x)M(x \mid A \setminus \{a_1\})\right). \tag{4.19}$$

Remark 4.3.4. *This identity can also be derived from the second equation in [208, Exercise 5.9-6], which ultimately traces back to [211], using the correspondence to B-splines made in [209].*

Proof. We compute the quantity

$$(\langle f, (x - a_1)M(A \setminus \{a_n\}\rangle + \langle f, (a_n - x)M(A \setminus \{a_1\})\rangle) \tag{4.20}$$

by substituting (4.14) in (4.20), obtaining

$$\langle f, M_1 + M_2 + M_3\rangle$$

with

$$\begin{aligned}
\langle f, M_1\rangle &= \sum_{i=1}^{n} \frac{r_i}{c-1}(a_i - a_1)\langle f, M(A \setminus \{a_n\} \sqcup \{a_i\})\rangle,\\
\langle f, M_2\rangle &= \sum_{i=1}^{n} \frac{r_i}{c-1}(a_n - a_i)\langle f, M(A \setminus \{a_1\} \sqcup \{a_i\})\rangle,\\
\langle f, M_3\rangle &= -\frac{2}{c-1}(a_n - a_1)\langle f, M(A)\rangle.
\end{aligned}$$

Thanks to (4.2.1), we can rewrite

$$(a_i - a_1)\langle f, M(A \setminus \{a_n\} \sqcup \{a_i\})\rangle = (a_n - a_1)\langle f, M(A)\rangle + (a_i - a_n)\langle f, M(A \setminus \{a_1\} \sqcup \{a_i\})\rangle.$$

Therefore,

$$\langle f, M_1 + M_2\rangle = (a_n - a_1)\sum_{i=1}^{n} \frac{r_i}{c-1}\langle f, M(A)\rangle = (a_n - a_1)\frac{c}{c-1}\langle f, M(A)\rangle,$$

and, finally,

$$(c-2)(a_n - a_1)\langle f, M(A)\rangle = (c-1)\left(\langle f, (x - a_1)M(A \setminus \{a_n\})\rangle + \langle f, (a_n - x)M(A \setminus \{a_1\})\rangle\right),$$

proving the claim. □

Corollary 4.3.5. *In the same conditions as in Theorem 4.3.3, the following more general identity holds for $i, j = 1, \dots, n$:*

$$(c-2)(a_j - a_i)M(x \mid A) = (c-1)\left[(x-a_i)M(x \mid A \setminus \{a_j\}) + (a_j - x)M(x \mid A \setminus \{a_i\})\right],$$

Proof. Simply note that, thanks to the permutation invariance of simplex splines with respect to their knots (Theorem 4.2.1), there is nothing special about the subscripts 1 and n in (4.19), which can be replaced by arbitrary subscripts $i \neq j$. The case $i = j$ is trivial. □

Equation (4.19) directly translates into the Cox-De Boor recursion formula (3.15) when the multiplicities r_i are integers. In fact, after ordering the knots in ascending order, the functions $M(x \mid A \setminus \{a_1\})$ and $M(x \mid A \setminus \{a_n\})$ can be interpreted as B-spline functions of degree k where the first and last knot have been removed, respectively, using the correspondences

$$\begin{aligned} N(x \mid A) &\leftarrow \frac{a_n - a_1}{c-1} M(x \mid A), \\ k &\leftarrow c - 2 = |A| - 2. \end{aligned} \tag{4.21}$$

Notice that, thanks to (4.19), the result obtained via simplex splines is actually more general than the Cox-De Boor recursion formula, as it allows to remove any two (distinct) knots in order to compute a spline of order k from two splines of order $k-1$, and not just the first and last knots. This is because simplex splines are by definition independent of the ordering of the knots (as proven in Theorem 4.2.1). This feature is important when generalising splines to unstructured point clouds in higher dimension, where no natural ordering of the knot locations is possible. However, ordering of one-dimensional knots also accomplishes the task of providing suitable grouping of $k+2$ consecutive knots, which is essential for the definition of a complete spline basis. Thus, a new geometric structure will have to assume this role in higher dimensions, as will be discussed in Chapter 5.

Notice that repeated application of (4.19) and Theorem 4.2.7 allows to completely eliminate the knots in A one by one. Once we have eliminated all the knots except two, we can compute the remaining integral directly via

$$\langle f, M(a_1, a_2) \rangle \coloneqq \int_\Sigma f(\lambda_1 a_1 + \lambda_2 a_2)\, \mathrm{d}\lambda_1 = \int_{[a_1, a_2]} f(x)\, \mathrm{d}x.$$

The recursion relation (4.19) can therefore be used as a practical means to compute simplex splines, just as (3.15) can be used for the usual B-splines.

4.3.2 Knot insertion formulas

We derive now two important relations that will generalize easily to the multivariate case, and will be very important in the generalization of simplex splines to higher dimensions. First, we derive a *knot insertion formula*, that allows us to insert a new knot into a simplex spline and

express it as a linear combination of other simplex splines in which one knot has been replaced by the new knot.

Corollary 4.3.6 (Carlson, Micchelli)**.** *Let A have at least 2 distinct knots, and let f be continuous. Let β_i, $i = 1, \dots, n$ be the barycentric coordinates of a point $\overline{a} \in \mathbb{R}$, i.e., real numbers such that*

$$\sum_{i=1}^{n} \beta_i a_i = \overline{a}, \ \sum_{i=1}^{n} \beta_i = 1.$$

Then, the following identity holds:

$$M(x \mid A) = \sum_{i=1}^{n} \beta_i M(x \mid A \sqcup \{\overline{a}\} \setminus \{a_i\}). \tag{4.22}$$

Remark 4.3.7. *This formula corresponds to Micchelli's knot insertion formula, see [205]. A simple geometric interpretation can be found in Chapter 18 of [203].*

Proof. Write (4.17) with knot vector $\overline{A} \coloneqq A \sqcup \{\overline{a}\}$, and setting $k = n+1$. Notice that $a_{n+1} = \overline{a}$. Then, multiplying by β_i, and summing on i from 1 to n yields directly

$$(\overline{a} - a_j) M(x \mid A) + (a_j - \overline{a}) \sum_{i=1}^{n} \beta_i M(x \mid A \sqcup \{\overline{a}\} \setminus \{a_i\}) = 0.$$

We can remove the factor $(\overline{a} - a_j)$ directly if $\overline{a} \neq a_j$, and lift this restriction by rewriting the relation with another free index different from j. □

Notice that if $\beta_i > 0$ for all i, the combination is convex. In a very similar way, we can obtain another recursion formula for simplex splines, more computationally complex than the formula in Theorem 4.3.3, but with the advantage of being easier to extend to higher dimensions.

Corollary 4.3.8 (Carlson, Micchelli)**.** *Let β_i, $i = 1, \dots, n$, be the barycentric coordinates for x, i.e., real numbers such that*

$$\sum_{i=1}^{n} \beta_i a_i = x, \ \sum_{i=1}^{n} \alpha_i = 1.$$

Then, the following identity holds:

$$(c-2) M(x \mid A) = (c-1) \sum_{i=1}^{n} \beta_i M(x \mid A \setminus \{a_i\}). \tag{4.23}$$

Remark 4.3.9. *This formula corresponds to the univariate version of Corollary 3 in [209], and can also be found in [205].*

Proof. Starting from the result of Corollary 4.3.5, simply multiply each side by β_i and sum. After noticing that

$$\sum_{i=1}^{n} \beta_i (x - a_i) = 0,$$

we obtain

$$(c-2)(a_j-x)M(x \mid A) = (c-1)(a_j-x)\sum_{i=1}^{n}\beta_i M(x \mid A \setminus \{a_i\}).$$

This proves the identity for $x \neq a_j$. If A contains only one knot $\overline{x}$, then the identity is not valid for $x = \overline{x}$ (indeed, as we show in the last section of this chapter, the spline has a pole there). Else, we can repeat the derivation switching $i \leftrightarrow j$, and since $a_i \neq a_j$, this proves the identity for $x = a_j$ as well. □

4.3.3 Spatial derivatives

Finally, we can easily prove two useful formulas for the derivative of a univariate simplex spline, which easily generalizes to higher dimensions as shown in the next section, allowing us to compute general directional derivatives of multivariate simplex splines. The first relation expresses the (weak) derivative of a simplex spline of degree k as a sum of two splines of degree $k-1$.

Corollary 4.3.10. *The following identity holds weakly over functions in* $C^1(\mathrm{conv}(A))$*:*

$$(a_i - a_j)M'(x \mid A) = (c-1)\left(M(x \mid A \setminus \{a_i\}) - M(x \mid A \setminus \{a_j\})\right).$$

Remark 4.3.11. *This is a more general statement of the corresponding derivative formula for the usual B-splines. If* $i = 1$ *and* $j = k+2$*, we recover* (3.16)*.*

Proof. Starting from (4.18), multiply by $f(\Lambda \cdot A)$ and integrate. Under the hypothesis that f is continuously differentiable, we can integrate by parts on the variables λ_i and λ_j. The boundary term vanishes since $\mathrm{Re}(r_i)$ and $\mathrm{Re}(r_j)$ are positive. We can thus evaluate

$$\langle f, M(A \setminus \{a_i\})\rangle - \langle f, M(A \setminus \{a_j\})\rangle$$

as

$$\begin{aligned}\int_\Sigma f(\Lambda \cdot A)\left(\frac{\partial}{\partial\lambda_i} - \frac{\partial}{\partial\lambda_j}\right)(\,\mathrm{d}\Sigma_R) &= -\int_\Sigma \left(\frac{\partial}{\partial\lambda_i} - \frac{\partial}{\partial\lambda_j}\right)(f(\Lambda \cdot A))\,\mathrm{d}\Sigma_R,\\ &= -(a_i - a_j)\int_\Sigma f'(\Lambda \cdot A)\,\mathrm{d}\Sigma_R,\\ &= (a_i - a_j)\big\langle f, M'(A)\big\rangle,\end{aligned}$$

where we have applied the usual definition of the distributional derivative, whereby M' is implicitly defined by

$$\big\langle f, M'(A)\big\rangle \coloneqq -\big\langle f', M(A)\big\rangle.$$

□

The second relation that we derive can be used to compute the directional derivative of a spline function. First, we need the following property of the Dirichlet measure.

Corollary 4.3.12 (Carlson, Micchelli)**.** *Let ν_i, $i = 1, \dots, n$ be real numbers such that*

$$\sum_{i=1}^{n} \nu_i = 0.$$

Then, the following identity holds:

$$\sum_{i=1}^{n} \nu_i \frac{\partial}{\partial \lambda_i} \, \mathrm{d}\Sigma_R = (c-1) \sum_{i=1}^{n} \nu_i \, \mathrm{d}\Sigma_{R-E_i}. \tag{4.24}$$

Proof. According to (4.15), the term

$$\frac{\partial}{\partial \lambda_i} \, \mathrm{d}\Sigma_R - (c-1) \, \mathrm{d}\Sigma_{R-E_i}$$

is actually independent of i. Therefore, multiplying this expression by ν_i and summing yields

$$\sum_{i=1}^{n} \nu_i \frac{\partial}{\partial \lambda_i} \, \mathrm{d}\Sigma_R - (c-1) \sum_{i=1}^{n} \nu_i \, \mathrm{d}\Sigma_{R-E_i} = \left(\sum_{i=1}^{n} \nu_i \right) \left(\frac{\partial}{\partial \lambda_j} \, \mathrm{d}\Sigma_R - (c-1) \, \mathrm{d}\Sigma_{R-E_j} \right) = 0,$$

which proves the claim. □

The directional derivative of a simplex spline can then be evaluated as follows.

Corollary 4.3.13. *Let ν_i be defined as in Corollary 4.3.12, and let*

$$v = \sum_{i=1}^{n} \nu_i a_i,$$

Then, the following identity holds weakly over functions in $C^1(\mathrm{conv}(A))$:

$$v M'(x \mid A) = (c-1) \sum_{i=1}^{n} \nu_i M(x \mid A \setminus \{a_i\}). \tag{4.25}$$

Proof. Starting from (4.24), we multiply by $f(x)$ and integrate. We can then apply the functional definition of a simplex spline (4.10):

$$\begin{aligned} \int_\Sigma f(\Lambda \cdot A) \frac{\partial}{\partial \lambda_i} \left(\mathrm{d}\Sigma_R \right) &= - \int_\Sigma \frac{\partial}{\partial \lambda_i} \left(f(\Lambda \cdot A) \right) \mathrm{d}\Sigma_R, \\ &= - \int_\Sigma (a_i - a_n) f'(\Lambda \cdot A) \, \mathrm{d}\Sigma_R, \\ &= (a_i - a_n) \big\langle f, M'(A) \big\rangle. \end{aligned}$$

Multiplying by ν_i and summing over i, the left hand side becomes:

$$\int_\Sigma f(\Lambda \cdot A) \sum_{i=1}^{n} \nu_i \frac{\partial}{\partial \lambda_i} \left(\mathrm{d}\Sigma_R \right) = (c-1) \sum_{i=1}^{n} \nu_i \int_\Sigma f(\Lambda \cdot A) \, \mathrm{d}\Sigma_{R-E_i},$$

thanks to (4.24). Applying the same operation to the right hand side yields:

$$\sum_{i=1}^{n} \nu_i (a_i - a_n) \big\langle f, M'(A) \big\rangle = v \big\langle f, M'(A) \big\rangle.$$

Equating the last two expressions yields the desired result. □

As we discuss below, in higher dimension, the variable $v = \sum_{i=1}^{n} \nu_i a_i$ becomes a vector, so that the left hand side of (4.24) and (4.25) can be interpreted as a directional derivative. This formula then becomes the multivariate analog of the B-spline derivative formula (3.16).

4.3.4 Knot dependence of simplex splines

In virtually all shape optimization applications, the modification of the geometry of a spline (or NURBS) model is performed over the control points, with the knot positions remaining fixed. A one-dimensional B-spline curve, for example, can be represented over a B-spline basis as:

$$\boldsymbol{\gamma}(t) = \sum_{i=1}^{n} \boldsymbol{\gamma}_i N_{i,k}(t).$$

Classical shape optimization would keep the basis (and thus the knots) fixed, and focus on determining the optimal control variables $\boldsymbol{\gamma}_i$ minimizing a given cost function. These problems are often focused on the determination of a complex shape, on which the relevant physical parameters are relatively homogeneous, and the only adjustment parameter is the shape itself. In contrast, in the context of full waveform inversion, the geometric shape of the simulation domain is usually very simple, often just a parallelepiped, while the distribution of physical properties inside it can be very complex, and with varying degrees of continuity. In many approaches to waveform inversion, the domain is first meshed, and some (constant or variable) values of physical parameters assigned to each cell. During inversion, the values in each cell are adjusted to match the measured data, and optionally the shape of the cells themselves is modified.

A similar configuration can be obtained with a spline-based model. For example, consider a simple one-dimensional velocity model on a segment, with a variable density represented on a B-spline basis over a given knot vector A with n knots,

$$\rho(x) = \sum_{i=1}^{n} \rho_i N_{i,k}(x \mid A).$$

During the inversion step, the equivalent separation into physical and geometrical degrees of freedom can be obtained by letting the control variables ρ_i govern the values of the relevant physical quantities (velocity, etc...), and devolve the description of the shape to the positions of the spline knots $(a_i)_{i=1}^{n}$. In the one-dimensional example, moving the knots would change the shape of the zone of influence of each physical value ρ_i, just like the classical approach. However, many advantages arise from this perspective, including a larger freedom in the localization and shape of the reconstructed irregularities, without requiring an explicit mesh. While these

advantages will be better discussed in Chapter 7, these considerations nonetheless stimulate a proper analysis of the dependence of B-splines on the position of their knots, which is the purpose of this section.

As seen above, the Dirichlet average of a function can be defined once a vector of knots A, possibly with repetitions, is given. The Dirichlet average can thus be viewed as a map from real functions to functions of the n real parameters $a_1, \dots, a_n$. Using a notation similar to Carlson's [208], we can write

$$f(x) \mapsto F(A),$$

with the Dirichlet average of f being denoted as F. Once the real variable x has been integrated out, Dirichlet averages can be seen as pure multivariate functions of their knots positions, and the question of the precise relation of this dependence naturally arises.

The most important tool that will be used repeatedly in this section is the recognition that the knots A only appear in the definition of a simplex spline (4.10) through the moment map μ_A, and more specifically as an argument of the function f. Therefore, if $f(x)$ has a degree of differentiability C^k with respect to x, we expect the whole expression (4.10) to have the same degree of differentiability with respect to any of the knots a_i, which can then be transferred to the spline function M. Notice that

$$\frac{\partial}{\partial a_i} f(\Lambda \cdot A) = \lambda_i f'(\Lambda \cdot A) \tag{4.26}$$

and, more generally,

$$\frac{\partial^m}{\partial a_{i_1}^{m_1} \cdots \partial a_{i_k}^{m_k}} f(\Lambda \cdot A) = \lambda_{i_1}^{m_1} \cdots \lambda_{i_k}^{m_k} f^{(m)}(\Lambda \cdot A), \text{ with } \sum_{j=1}^{k} m_j = m.$$

Derivatives with respect to single vs. repeated knots

In the rest of this section, we use the definition of a spline function via Dirichlet measures to compute the derivatives of a spline function with respect to the position of its knots. Recall that knots can be repeated in A, and that this is equivalent to modifying the corresponding multiplicities in the multiplicity vector R. Any combination is possible, and thus one can in principle compute the derivative with respect to a single copy of a repeated knot a_i, leaving its other copies untouched,

$$\frac{\partial}{\partial a_i} M(x \mid A),$$

or one can compute the derivative with respect to multiple (or all) copies of a_i. Due to the permutation symmetry of simplex splines expressed via Theorem (4.2.1), the result of this operation is simply

$$\sum_{j \in I} \frac{\partial}{\partial a_j} M(x \mid A) = r_i \frac{\partial}{\partial a_i} M(x \mid A),$$

where I is the set containing all the indices of the collocated points that are considered in the derivative, and $r_i \coloneqq |I|$ is the corresponding multiplicity. Given the extreme simplicity in switching from one point of view to the other, we opt here for allowing freedom of choice with

regard to how many copies of a knot are included in the computation of the derivative. This can be conveniently expressed by including in A only knots that are considered as distinct, and adjusting their corresponding multiplicities in R accordingly.

After this necessary (but perhaps slightly too scrupulous) explanation, we are ready to proceed.

Knot dependence and the Euler-Darboux equations

If the function f is at least of class C^m, it can be proven that the derivatives can indeed be taken inside the integral (4.10) defining the Dirichlet average, which we express as follows.

Theorem 4.3.14 (Carlson [208, Theorem 5.3-2])**.** *Let $f \in C^m(\operatorname{conv}(A))$, and let the Dirichlet average of a function be defined by* (4.10)*. Then,*

$$\frac{\partial^m}{\partial a_{i_1}^{m_1} \cdots \partial a_{i_k}^{m_k}} \int_\Sigma f(\Lambda \cdot A)\, \mathrm{d}\Sigma_R = \int_\Sigma f^{(m)}(\Lambda \cdot A) \lambda_{i_1}^{m_1} \cdots \lambda_{i_k}^{m_k}\, \mathrm{d}\Sigma_R, \tag{4.27}$$

where $m \coloneqq \sum_{j=1}^k m_j$.

The proof can be found in the cited reference. This result directly gives a simple formula for the derivative of a simplex spline with respect to a knot.

Corollary 4.3.15. *The following identities hold weakly over functions in $C^1(\operatorname{conv}(A))$:*

$$\frac{\partial}{\partial a_i} M(x \mid A) = -w_i M'(x \mid A \sqcup \{a_i\}), \tag{4.28}$$

$$\sum_{i=1}^n \frac{\partial}{\partial a_i} M(x \mid A) = -M'(x \mid A). \tag{4.29}$$

Remark 4.3.16. *These identities correspond to equations 5.6-5 and 5.3-3 (with $n = 1$) of [208], respectively.*

Proof. Using (4.27) with $k = 1$ and $m_1 = 1$, we can take the derivative inside the integral. Thus:

$$\begin{aligned}
\frac{\partial}{\partial a_i} \langle f, M(A) \rangle &= \int_\Sigma \frac{\partial}{\partial a_i} f(\Lambda \cdot A)\, \mathrm{d}\Sigma_R, \\
&= \int_\Sigma f'(\Lambda \cdot A) \lambda_i\, \mathrm{d}\Sigma_R, \\
&= w_i \int_\Sigma f'(\Lambda \cdot A)\, \mathrm{d}\Sigma_{R+E_i}, \\
&= w_i \big\langle f', M(x \mid A \sqcup \{a_i\}) \big\rangle, \\
&= -w_i \big\langle f, M' A \sqcup \{a_i\} \big\rangle,
\end{aligned}$$

where in the third step we have applied the identity

$$\lambda_i\, \mathrm{d}\Sigma_R = w_i\, \mathrm{d}\Sigma_{R+E_i},$$

a simple consequence of (4.8) and (4.11), and in the last step we have used the definition of distributional derivative. The second identity can be obtained from the first one by summing over i and then applying (4.12). □

Notice that the derivative with respect to a_i has been essentially transformed into a derivative with respect to x. By repeated application of (4.28) and (4.29), we easily obtain similar expressions for higher order derivatives:

Corollary 4.3.17. *The following identities hold weakly over functions in* $C^m(\mathrm{conv}(A))$,

$$\frac{\partial^m}{\partial a_{i_1}^{m_1}\cdots\partial a_{i_k}^{m_k}} M(x \mid A) = (-1)^m \frac{(r_1)_{m_1}\cdots(r_k)_{m_k}}{(c)_m} M^{(m)}(x \mid A \sqcup \{a_{i_1}\}^{m_1} \sqcup \cdots \sqcup \{a_{i_k}\}^{m_k}),$$

$$\left(\sum_{i=1}^{n} \frac{\partial}{\partial a_i}\right)^m M(x \mid A) = (-1)^m M^{(m)}(x \mid A),$$

where $m \coloneqq m_1 + \cdots + m_k$ *and*

$$(a)_b \coloneqq a \cdot (a+1) \cdots (a+b-1) = \frac{\Gamma(a+b)}{\Gamma(a)}$$

denotes the Pochhammer symbol.

Remark 4.3.18. *These identities correspond to Eqs. 5.6-6 and 5.3-3 (with general n) of [208], respectively.*

One of the most interesting aspects of interpreting simplex splines as multivariate functions of the knot locations is the existence of a hidden symmetry, which can be expressed via a set of Euler-Darboux equations[1], as proven by Carlson in his work [209]. We can reformulate this fact using our notation as follows.

Theorem 4.3.19 (Carlson)**.** *The following relation holds weakly over functions in* $C^2(\mathrm{conv}(A))$, *for all* $1 \le i, j \le n$*:*

$$\left[(a_i - a_j)\frac{\partial^2}{\partial a_i \partial a_j} + r_i \frac{\partial}{\partial a_j} - r_j \frac{\partial}{\partial a_i}\right] M(x \mid A) = 0. \tag{4.30}$$

Proof. Recalling that $\Lambda \cdot A = \lambda_1 a_1 + \cdots + \lambda_n a_n$ and that $\lambda_n = 1 - \lambda_1 - \cdots - \lambda_{n-1}$, we find

$$\lambda_i \frac{\partial}{\partial \lambda_i} f(\Lambda \cdot A) = (a_i - a_n)\frac{\partial}{\partial a_i} f(\Lambda \cdot A). \tag{4.31}$$

Then, differentiating (4.31) with respect to a_n and using (4.26) yields

$$\begin{aligned}(a_i - a_n)\frac{\partial^2}{\partial a_i \partial a_n} f(\Lambda \cdot A) &= \frac{\partial}{\partial a_i} f(\Lambda \cdot A) + \lambda_i \frac{\partial}{\partial \lambda_i}(\lambda_n f'(\Lambda \cdot A)) \\ &= \lambda_i \lambda_n \frac{\partial}{\partial \lambda_i} f'(\Lambda \cdot A).\end{aligned}$$

[1] Carlson calls these Euler-Poisson equations. However, over time, the name for the most general form of this equation seems to have settled on Euler-Darboux, with the name Euler-Poisson-Darboux reserved for the special case $r_i = r_j$.

We integrate over Σ with respect to the measure $\mathrm{d}\Sigma_R$. On the left hand side, we can take the derivatives with respect to the knots outside the integral, thanks to (4.27). On the right hand side, we integrate by parts over λ_i, and the boundary term disappears since $\mathrm{Re}(r_i), \mathrm{Re}(r_n) > 0$ and therefore $\lambda_i \lambda_n \, \mathrm{d}\Sigma_R$ is zero at the endpoints $\lambda_i = 0$, $\lambda_n = 0$. Consequently,

$$\begin{aligned}
(a_i - a_n) \frac{\partial^2}{\partial a_i \partial a_n} \int_\Sigma f(\Lambda \cdot A) \, \mathrm{d}\Sigma_R &= - \int_\Sigma f'(\Lambda \cdot A) \frac{\partial}{\partial \lambda_i} \left(\lambda_i \lambda_n \, \mathrm{d}\Sigma_R \right), \\
&= - \int_\Sigma f'(\Lambda \cdot A) (r_i \lambda_n - r_n \lambda_i) \, \mathrm{d}\Sigma_R, \\
&= - \int_\Sigma \left(r_i \frac{\partial}{\partial a_n} - r_n \frac{\partial}{\partial a_i} \right) f(\Lambda \cdot A) \, \mathrm{d}\Sigma_R,
\end{aligned}$$

where in the last step we have again used (4.26). Using the permutation symmetry proven in Theorem 4.2.1, we can replace the index n with an arbitrary index j. Plugging this result into the definition (4.10) proves the theorem. □

We show in Chapter 7 how this property can be used to speedup the computation of the Hessian matrix, with potential applications in optimization and seismic inversion.

Knot derivatives of the usual B-spline functions

The expressions for first and second derivatives of simplex splines can be directly translated into equivalent expressions valid for the usual Cox-De Boor B-splines, with some tedious but straightforward calculations. The formula for the first derivative was already derived by direct computation in [212], although in a slightly less general form.

Theorem 4.3.20. *Let $N_k(x \mid A)$ be a B-spline function of order k defined over the knot vector $A = (a_1, \ldots, a_n)$ with respective multiplicities $R = (r_1, \ldots, r_n)$, $\sum_{i=1}^n r_i = k + 2$. Then, the first and second derivatives of the spline function with respect to the knot positions can be expressed as follows:*

$$\begin{aligned}
\frac{\partial}{\partial a_i} N_k(x \mid A) =& \frac{\delta_{in} - \delta_{i1}}{a_n - a_1} N_k(x \mid A) - \frac{r_i}{k+1} N'_{k+1}(x \mid A \sqcup \{a_i\}), \\
\frac{\partial^2}{\partial a_i^2} N_k(x \mid A) =& \frac{r_i}{k+1} \left(\frac{r_i + 1}{k+2} N''_{k+2}(x \mid A \sqcup \{a_i\}^2) - 2 \frac{\delta_{in} - \delta_{i1}}{a_n - a_1} N'_{k+1}(x \mid A \sqcup \{a_i\}) \right), \\
\frac{\partial^2}{\partial a_i \partial a_j} N_k(x \mid A) =& - \frac{r_i}{k+1} N'_{k+1}(x \mid A \sqcup \{a_i\}) \left(\frac{r_j}{a_i - a_j} + \frac{\delta_{jn} - \delta_{j1}}{a_n - a_1} \right) \\
& - \frac{r_j}{k+1} N'_{k+1}(x \mid A \sqcup \{a_j\}) \left(\frac{r_i}{a_j - a_i} + \frac{\delta_{in} - \delta_{i1}}{a_n - a_1} \right).
\end{aligned} \tag{4.32}$$

Proof. Recall that, according to (4.21), the usual B-spline basis functions can be obtained from simplex splines with positive integer multiplicities simply by changing the normalization. If the knots are in increasing order,

$$N_k(x \mid A) = \frac{a_n - a_1}{k+1} M(x \mid A), \; c = k + 2.$$

We can thus directly apply (4.28) and (4.30), taking (4.21) into account. We begin by computing the first derivative,

$$\begin{aligned}\frac{\partial}{\partial a_i}N_k(x \mid A) &= \frac{\partial}{\partial a_i}\left(\frac{a_n - a_1}{k+1}\right)M(x \mid A),\\ &= \frac{\delta_{in} - \delta_{i1}}{k+1}M(x \mid A) - \frac{r_i}{k+2}\frac{a_n - a_1}{k+1}M'(x \mid A \sqcup \{a_i\}),\\ &= \frac{\delta_{in} - \delta_{i1}}{a_n - a_1}N_k(x) - \frac{r_i}{k+1}N'_{k+1}(x \mid A \sqcup \{a_i\}).\end{aligned}$$

In the last step, we have made use of the following relation:

$$\begin{aligned}M'(x \mid A \sqcup \{a_i\}) &= \frac{\partial}{\partial x}\left(\frac{k+2}{a_n - a_1}N_{k+1}(x \mid A \sqcup \{a_i\})\right),\\ &= \frac{k+2}{a_n - a_1}N'_{k+1}(x \mid A \sqcup \{a_i\}).\end{aligned}$$

Differentiating again with respect to a_i yields

$$\begin{aligned}\frac{\partial^2}{\partial a_i^2}N_k(x \mid A) &= \frac{\partial^2}{\partial a_i^2}\left(\frac{a_n - a_1}{k+1}M(x \mid A)\right),\\ &= 2\left(\frac{\partial}{\partial a_i}\frac{a_n - a_1}{k+1}\right)\left(\frac{\partial}{\partial a_i}M(x \mid A)\right) + \frac{a_n - a_1}{k+1}\frac{\partial^2}{\partial a_i^2}M(x \mid A),\\ &= -2\frac{\delta_{in} - \delta_{i1}}{k+1}\frac{r_i}{k+2}M'(x \mid \sqcup\{a_i\}) + \frac{a_n - a_1}{k+1}\frac{r_i(r_i+1)}{(k+2)(k+3)}M''(x \mid A \sqcup \{a_i\}^2),\\ &= \frac{r_i}{k+1}\left(\frac{r_i+1}{k+2}N''_{k+2}(x \mid A \sqcup \{a_i\}^2) - 2\frac{\delta_{in} - \delta_{i1}}{a_n - a_1}N'_{k+1}(x \mid A \sqcup \{a_i\})\right),\end{aligned}$$

whereas differentiating with respect to a_j yields the mixed second derivative

$$\begin{aligned}\frac{\partial^2}{\partial a_i \partial a_j}N_k(x \mid A) &= \frac{\partial^2}{\partial a_i \partial a_j}\left(\frac{a_n - a_1}{k+1}M(x \mid A)\right),\\ &= P_{ij} + P_{ji} + Q_{ij},\end{aligned}$$

with P_{ij} given by

$$\begin{aligned}P_{ij} &= \frac{\partial}{\partial a_i}\left(\frac{a_n - a_1}{k+1}\right)\frac{\partial}{\partial a_j}\left(M(x \mid A)\right),\\ &= -\frac{\delta_{in} - \delta_{i1}}{k+1}\frac{r_j}{k+2}M'(x \mid A \sqcup \{a_j\}),\\ &= -\frac{\delta_{in} - \delta_{i1}}{a_n - a_1}\frac{r_j}{k+1}N'_{k+1}(x \mid A \sqcup \{a_j\}),\end{aligned}$$

and where Q_{ij} can be computed using (4.30),

$$Q_{ij} = \frac{a_n - a_1}{k+1}\frac{\partial^2}{\partial a_i \partial a_j}M(x \mid A),$$

$$
\begin{aligned}
&= \frac{a_n - a_1}{(k+1)(a_i - a_j)} \left(r_j \frac{\partial}{\partial a_i} M(x \mid A) - r_i \frac{\partial}{\partial a_j} M(x \mid A) \right), \\
&= \frac{a_n - a_1}{(k+1)(a_i - a_j)} \frac{r_i r_j}{k+2} \left(M'(x \mid A \sqcup \{a_j\}) - M'(x \mid A \sqcup \{a_i\}) \right), \\
&= \frac{r_i r_j}{(k+1)(a_i - a_j)} \left(N'_{k+1}(x \mid A \sqcup \{a_j\}) - N'_{k+1}(x \mid A \sqcup \{a_i\}) \right).
\end{aligned}
$$

Adding together the terms P_{ij}, P_{ji} and Q_{ij} yields the second member of (4.32). □

Note that, according to (4.32), if i and j correspond to the first and last knots and both have multiplicity 1, then the mixed derivative with respect to a_i and a_j is identically zero. This is due to the fact that according to De Boor's recursion formula, the value of the spline can be expressed as a sum of two pieces, one dependent on the first knot but independent of the last one, and one dependent on the last knot but not on the first one. Notice also that the computational advantage in the calculation of the Hessian matrix mentioned above, which we detail in Chapter 7, also applies to the usual B-spline basis functions, since once one can compute the full Hessian (i.e., all the mixed derivatives $\partial^2 N_k(x \mid A)/\partial a_i \partial a_j$) only by knowing the values of $N_k(x \mid A)$, $N'_{k+1}(x \mid A \sqcup \{a_i\})$ and $N''_{k+2}(x \mid A \sqcup \{a_i\}^2)$ for each knot a_i.

4.4 Multivariate simplex splines

The previous section has been devoted to exploring some of the most basic properties of univariate simplex splines. However, the conclusion of Theorem 4.3.3 is that, for (positive) integer values of the multiplicity vector R, simplex splines merely reproduce the usual B-spline basis, albeit with a different normalization. There is, however, a very important reason why the Dirichlet-measure approach to splines is very interesting for the present work, and it has to do with the simplicity of obtaining a multivariate, unstructured generalization of B-splines.

It is very easy in fact to generalize the definition (4.10) of a simplex spline to dimension d: simply modify the moment map μ_A so that it maps $\mathbb{R}^n$ into $\mathbb{R}^d$. We continue to denote by $\Lambda = (\lambda_1, \ldots, \lambda_n)$ the vector containing the barycentric coordinates for the standard simplex $\Sigma \coloneqq \Sigma^{n-1}$ in $\mathbb{R}^{n-1}$, but we replace the $n \times 1$ knot vector A by a $n \times d$ knot matrix $A \coloneqq (a_1, \ldots, a_n)$, where now each knot a_i lies in $\mathbb{R}^d$. We thus refer to the j-th component of the i-th knot as $a_{i,j}$. Then, the moment map can still be expressed as

$$\mu_A(\Lambda) \coloneqq \Lambda \cdot A = \lambda_1 a_1 + \cdots + \lambda_n a_n$$

and, given a function $f : \mathbb{R}^d \to \mathbb{R}$, integrable over $\operatorname{conv}(A)$, the corresponding multivariate simplex spline is defined by

$$\int_{\mathbb{R}^d} f(x) M(x \mid A) \, \mathrm{d}x = \langle f, M(A) \rangle \coloneqq \int_{\Sigma} f(\Lambda \cdot A) \, \mathrm{d}\Sigma_R, \tag{4.33}$$

with no modification required on the Dirichlet measure (4.8). We show a few multivariate simplex splines in Figure 4.3.

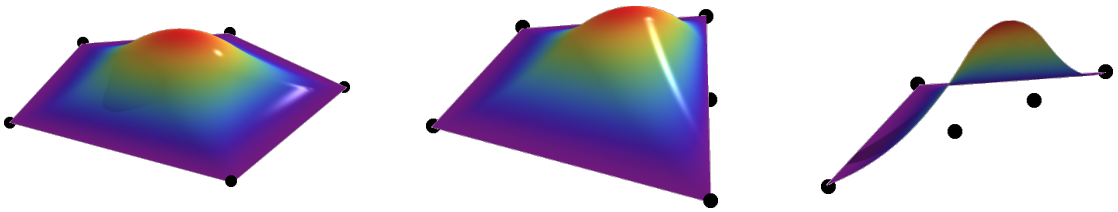


Figure 4.3: Three examples of bi-variate simplex splines of degree $k = 2$, built on $d + k + 1 = 5$ points. As the points become collinear, the regularity of the spline function decreases.

4.4.1 Recurrence and knot insertion formulas

As we will see shortly, a very large portion of theorems and identities can be seamlessly transported to multivariate splines. In fact, Theorem 4.2.1, Theorem 4.2.3, Lemma 4.2.6, Theorem 4.2.7 and their respective proofs carry over to multivariate simplex splines without any modification, and there is no need to restate them here. However, Lemma 4.3.2 and its proof require some modification. Since the change is very instructive for what is to follow, we do it here explicitly.

Lemma 4.4.1. *Let $M(x \mid A)$ be a multivariate simplex spline defined weakly on functions over $\mathbb{R}^d$, as in (4.33), with at least $d+2$ distinct knots in A. Then, for any choice of $d+2$ indices $1 \le i_1, \dots, i_{d+2} \le n = |A|$, the following identity holds:*

$$\begin{vmatrix} a_{i_1,1} & \cdots & a_{i_1,d} & 1 & M(x \mid A \setminus \{a_{i_1}\}) \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ a_{i_{d+2},1} & \cdots & a_{i_{d+2},d} & 1 & M(x \mid A \setminus \{a_{i_{d+2}}\}) \end{vmatrix} = 0. \tag{4.34}$$

Proof. If two indices are repeated, then the identity is trivial. We will therefore consider distinct indices. Moreover, to simplify the notation, we will recur to Theorem 4.2.1 and consider, without loss of generality, the sequence of indices $1, \dots, d+2$.

Recall that the proof of Lemma 4.3.2 was based on the introduction of the differential operator $\nabla_{ijk} \coloneqq (a_i - a_j)\partial/\partial\lambda_k + (a_j - a_k)\partial/\partial\lambda_i + (a_k - a_i)\partial/\partial\lambda_j$, which can be also written as

$$\nabla_{ijk} = \begin{vmatrix} a_i & 1 & \frac{\partial}{\partial\lambda_i} \\ a_j & 1 & \frac{\partial}{\partial\lambda_j} \\ a_k & 1 & \frac{\partial}{\partial\lambda_k} \end{vmatrix}.$$

We extend this definition to the multivariate case by introducing the operator

$$\nabla_{1,\cdots,d+2} \coloneqq \begin{vmatrix} a_{1,1} & \cdots & a_{1,d} & 1 & \frac{\partial}{\partial\lambda_1} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ a_{d+2,1} & \cdots & a_{d+2,d} & 1 & \frac{\partial}{\partial\lambda_{d+2}} \end{vmatrix} =: \det(D), \tag{4.35}$$

where D is the matrix defined implicitly in (4.35). Each derivative $\partial/\partial\lambda_i$ in (4.35) is the generator of translations in the corresponding variable λ_i. Therefore, the operator $\nabla_{1,\dots,d+2}$

generates the transformation

$$\lambda_i' = \lambda_i + (-1)^{i+d}\varepsilon D_{i,d+2},\ i = 1, \ldots, d+2,$$
$$\lambda_j' = \lambda_j,\ j = d+2, \ldots, n,$$

where $D_{i,j}$ is the minor of the matrix D formed by removing the i-th row and j-th column. This transformation, just like the transformation R_{ijk}^{ε} used in the proof of Lemma 4.3.2, is affinely well-defined, i.e.,

$$\sum_{i=1}^{n} \lambda_i' = \sum_{i=1}^{n} \lambda_i + \varepsilon \sum_{j=1}^{d+2} (-1)^{j+d} D_{i,d+2} = 1.$$

In fact, the second term is simply the determinant of a matrix obtained from D by replacing all the values in the last column by ε, and it is zero since the last two columns are multiples of one another. Furthermore, this transformation leaves the moment map unchanged, since

$$(\Lambda' \cdot A)_k = \sum_{i=1}^{n} \lambda_i' a_{i,k} = \sum_{i=1}^{n} \lambda_i a_{i,k} + \varepsilon \sum_{j=1}^{d+2} (-1)^{j+d} D_{j,d+2} a_{j,k} = (\Lambda \cdot A)_k,$$

after recognizing in the second sum the determinant of a matrix obtained from D by replacing its last column with its k-th column, which is again zero. Consequently, by the same arguments of Lemma 4.3.2,

$$\int_{\Sigma} f(\Lambda \cdot A) \nabla_{1,\ldots,d+2} \left(\mathrm{d}\Sigma_R \right) = - \int_{\Sigma} \nabla_{1,\ldots,d+2} \left(f(\Lambda \cdot A) \right) \mathrm{d}\Sigma_R = 0.$$

The operator $\nabla_{1,\ldots,d+2}$ is a linear function of the partial derivatives $\partial/\partial\lambda_i$, which can be easily seen by expanding (4.35) with respect to its last column. Moreover, if we substitute every differential operator $\partial/\partial\lambda_i$ with a constant, the whole determinant becomes zero. A linear map L that contains the constant vector in its kernel can be written as a linear map on pairwise differences, i.e.,

$$\sum_{j=1}^{n} L_{ij} v_j = \sum_{j=1}^{n} L_{ij}(v_j - v_1) + \left(\sum_{j=1}^{n} L_{ij} \right) v_1 = \sum_{j=1}^{n} L_{ij}(v_j - v_1).$$

Consequently, only the differences of derivatives $\partial/\partial\lambda_i - \partial/\partial\lambda_j$ appear in the determinant of D. We can thus use (4.18) to replace each differential operator $\partial/\partial\lambda_i$ in the determinant with $(c-1)M(x \mid A \setminus \{a_i\})$, proving the lemma. □

Lemma 4.4.1 allows us to compute a generalization of the Cox-De Boor recursion formula for multivariate simplex splines. Furthermore, the determinant (4.34) is equal to the determinant that can be used to check if the point a_{d+2} lies on the circumsphere determined by the point $a_1, \ldots, a_{d+1}$, after replacing the spline function $M(x \mid A \setminus \{a_i\})$ by $|a_i|^2$, cf. [213, Lemma 8.1]. This form is often used for the computation of Voronoi diagrams and Delaunay triangulations. This is a foreshadowing of the important role that these structures will play in Chapter 5.

From Lemma 4.4.1, we can readily obtain a generalization of Corollary 4.3.5, and thus of the

Cox-De Boor recursion formula (3.15), to multivariate simplex splines:

Theorem 4.4.2. *Suppose that there are at least $d+2$ distinct knots in A. Then, the following identity holds weakly for any choice of $d+1$ indices $1 \leq i_1, \ldots, i_{d+1} \leq n$:*

$$\begin{vmatrix} x_i & \cdots & x_d & 1 & (c-d-1)M(x \mid A) \\ a_{i_1,1} & \cdots & a_{i_1,d} & 1 & (c-1)M(x \mid A \setminus \{a_{i_1}\}) \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ a_{i_{d+1},1} & \cdots & a_{i_{d+1},d} & 1 & (c-1)M(x \mid A \setminus \{a_{i_{d+1}}\}) \end{vmatrix} = 0. \tag{4.36}$$

Remark 4.4.3. *This theorem provides a generalization of [4, Proposition 3.1] for knots with arbitrary multiplicity.*

Proof. If any of the $d+1$ indices $i_1, \ldots, i_{d+1}$ are repeated, the equation is trivially satisfied. Thanks to Theorem 4.2.1, we can without loss of generality choose the indices to be equal to $1, \ldots, d+1$.

In order to simplify the notation, we introduce, for every knot a_i, an extra coordinate $a_{i,d+1} = 1$. Notice that in this case the multivariate versions of (4.12) and (4.13) yield

$$\sum_{k=1}^{n} \frac{r_k - \delta_{kj}}{c-1} a_{k,i} \langle f, M(x \mid A \sqcup \{a_k\} \setminus \{a_j\}) \rangle = \langle f, M(x \mid A \setminus \{a_j\}) \rangle$$

for $i = 1, \ldots, d+1$. We can rewrite this expression as

$$\begin{aligned} \sum_{k=1}^{n} \frac{r_k}{c-1} a_{k,i} \langle f, M(x \mid A \sqcup \{a_k\} \setminus \{a_j\}) \rangle &= \langle f, M(x \mid A \setminus \{a_j\}) \rangle \\ &\quad + \frac{a_{j,i}}{c-1} \langle f, M(x \mid A) \rangle, \end{aligned} \tag{4.37}$$

again for $i = 1, \ldots, d+1$.

Let us rewrite (4.34) with variable x, indices $k, 1, \ldots, d+1$ and knot vector $A \sqcup \{a_k\}$,

$$\begin{vmatrix} a_{k,1} & \cdots & a_{k,d+1} & M(x \mid A) \\ a_{1,1} & \cdots & a_{1,d+1} & M(x \mid A \sqcup \{a_k\} \setminus \{a_1\}) \\ \vdots & \ddots & \vdots & \vdots \\ a_{d+1,1} & \cdots & a_{d+1,d+1} & M(x \mid A \sqcup \{a_k\} \setminus \{a_{d+1}\}) \end{vmatrix} =: \det(P) = 0,$$

and expand the determinant with respect to the first row,

$$\sum_{i=1}^{d+1} (-1)^{i+1} a_{k,i} P_{1,i} + (-1)^{d+1} M(x \mid A) P_{1,d+2} = 0, \tag{4.38}$$

where $P_{i,j}$ is the (i,j) minor of the above-defined matrix P. We introduce the matrix

$$Q = \begin{pmatrix} a_{1,1} & \cdots & a_{1,d+1} \\ \vdots & \ddots & \vdots \\ a_{d+1,1} & \cdots & a_{d+1,d+1} \end{pmatrix},$$

so that the determinant of the minor $P_{1,d+2}$ is equal to $\det(Q)$. Let us now consider in detail one of the terms in the sum (4.38),

$$\begin{aligned} (-1)^{i+1} a_{k,i} A_{1,i} = & \qquad (4.39) \\ &= (-1)^{i+1} a_{k,i} \begin{vmatrix} a_{1,1} & \cdots & \hat{a}_{1,i} & \cdots & a_{1,d+1} & M(x \mid A \sqcup \{a_k\} \setminus \{a_2\}) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ a_{d+1,1} & \cdots & \hat{a}_{d+1,i} & \cdots & a_{d+1,d+1} & M(x \mid A \sqcup \{a_k\} \setminus \{a_{d+1}\}) \end{vmatrix}, \\ &= (-1)^{i+1} a_{k,i} \sum_{j=2}^{d+2} (-1)^{j+1} M(x \mid A \sqcup \{a_k\} \setminus \{a_j\}) Q_{j,i}, \end{aligned}$$

where the hat denotes that a particular term is omitted, and $Q_{j,i}$ is the (j,i) minor of the matrix Q.

If we now multiply (4.39) by $r_k/(c-1)f(x)$, sum over k, and integrate over x, we can apply (4.37) and replace each spline function $M(x \mid A \sqcup \{a_k\} \setminus \{a_j\})$ with $x_i M(x \mid A \setminus \{a_j\}) + a_{j,i}/(c-1)M(x \mid A)$. After summing over j, the first term produces the following determinant:

$$(-1)^{i+1} x_i \begin{vmatrix} a_{1,1} & \cdots & \hat{a}_{1,i} & \cdots & a_{1,d+1} & M(x \mid A \setminus \{a_2\}) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ a_{d+1,1} & \cdots & \hat{a}_{d+1,i} & \cdots & a_{d+2,d+1} & M(x \mid A \setminus \{a_{d+1}\}) \end{vmatrix} =: (-1)^{i+1} x_i B_{1,i},$$

where the matrix B is derived from P by replacing each spline function $M(x \mid A \sqcup \{a_k\} \setminus \{a_i\})$ with $M(x \mid A \setminus \{a_i\})$. The second term produces the determinant:

$$\begin{aligned} (-1)^{i+1} \frac{a_{j,i}}{c-1} M(x \mid A) \begin{vmatrix} a_{1,1} & \cdots & \hat{a}_{2,i} & \cdots & a_{1,d+1} & a_{1,i} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ a_{d+1,1} & \cdots & \hat{a}_{d+1,i} & \cdots & a_{d+1,d+1} & a_{d+1,i} \end{vmatrix} = \\ = (-1)^d \frac{a_{j,i}}{c-1} \det(Q) M(x \mid A), \\ = (-1)^d \frac{a_{j,i}}{c-1} B_{i,d+2} M(x \mid A). \end{aligned}$$

Plugging these two terms back into (4.38), after summing over k and integrating, we obtain

$$\left\langle f, \left[\sum_{i=1}^{d+1} (-1)^{i+1} x_i B_{1,i} + (-1)^d \frac{d+1}{c-1} B_{1,d+2} M(x \mid A) + \right. \right. \\ \left. \left. + (-1)^{d+1} \frac{c}{c-1} B_{1,d+2} M(x \mid A) \right] \right\rangle = 0,$$

or, after rearranging,

$$\left\langle f, \left[\sum_{i=1}^{d+1} (-1)^{i+1}(c-1)x_i B_{1,i} + (-1)^{d+1}(c-d-1)B_{1,d+2}M(x \mid A) \right] \right\rangle = 0,$$

which is just the expansion of the weak form of (4.36) with respect to the first row. □

Notice that, if we expand (4.36) with respect to the last column, the coefficient of $M(x \mid A)$ corresponds to the volume of the d dimensional simplex defined by the $d+1$ chosen knots, whereas the coefficient of $M(x \mid A \setminus \{a_i\})$ is equal (up to sign) to the volume of the simplex obtained by replacing the i-th knot with x.

Equation (4.36) can be used to recursively compute the value of a simplex spline, and is the direct analog of the Cox-De Boor recurrence formula (4.19). We will see in Chapter 5 that, in the spline spaces that we consider, every defining spline function possesses a natural subset of $d+1$ affinely independent knots. Choosing this subset in (4.36) allows to compute $M(x \mid A)$ recursively, making it a multivariate generalization of (4.19), and a slight generalization of known equivalent formulas such as [4, Proposition 3.1] in the case of higher knot multiplicities.

Moving on to the other identities that have already been proven for univariate simplex splines, the knot insertion formula proven in Corollary 4.3.6 can be proven also for the multivariate simplex splines by a very similar proof, by multiplying the last column of the determinant in (4.34) by β_{i_1} and summing over i_1. We refer to [209] for a complete proof of this fact. The statement of the corollary is unchanged in the multivariate case. The same argument applies to Corollary 4.3.8, whose proof for multivariate simplex splines can also be found in [209]. The multivariate analog of (4.23) reads

$$(c-d-1)M(x \mid A) = (c-1)\sum_{i=1}^{n} \beta_i M(x \mid A \setminus \{a_i\}), \quad \text{whenever } \sum_{i=1}^{n} \beta_i a_i = x, \sum_{i=1}^{n} \beta_i = 1. \tag{4.40}$$

4.4.2 Spatial derivatives

The derivative formulas in Corollary 4.3.12 and Corollary 4.3.13 can be proven analogously to the univariate version, after noticing that in the multivariate case

$$\left(\frac{\partial}{\partial \lambda_i} - \frac{\partial}{\partial \lambda_j} \right) f(\Lambda \cdot A) = (a_i - a_j) \cdot \nabla_x f(\Lambda \cdot A),$$

where ∇_x denotes the derivative with respect to the spatial coordinate x, from which we can derive the multivariate version of (4.24),

$$(a_i - a_j) \cdot \nabla_x M(x \mid A) = (c-1)\left(M(x \mid A \setminus \{a_i\}) - M(x \mid A \setminus \{a_j\}) \right), \tag{4.41}$$

as well as that of (4.25),

$$v \cdot \nabla_x M(x \mid A) = (c-1) \sum_{i=1}^{n} \nu_i M(x \mid A \setminus \{a_i\}), \tag{4.42}$$
$$\text{whenever } \sum_{i=1}^{n} \nu_i a_i = v, \sum_{i=1}^{n} \nu_i = 0.$$

Notice that (4.41) and (4.42) are just scalar equations, and not vectorial ones. Equations (4.40), (4.41) and (4.42) can all be derived from [209, Theorem 7], and two of them are explicitly obtained in the cited work.

4.4.3 Knot dependence

Just like the functional relations of the previous section, the differential relations relating univariate simplex splines to their knots that were presented in Section 4.3.4 can be generalized to the multivariate case. The starting point for this derivation is, as in Section 4.3.4,

$$\frac{\partial}{\partial a_{i,j}} f(\Lambda \cdot A) = \lambda_i \frac{\partial}{\partial x_j} f(\Lambda \cdot A), \tag{4.43}$$

and therefore

$$\nabla_{a_i} f(\Lambda \cdot A) = \lambda_i \nabla_x f(\Lambda \cdot A),$$

which is valid for all functions $f \in C^1(\mathrm{conv}(A))$. Once again, the derivative can be taken inside integrals of the form (4.27), as proven in [208, Theorem 5.3-2]. The same arguments as in Corollary 4.3.12 and Corollary 4.3.13 then allow us to prove the following theorem.

Theorem 4.4.4. *Let $m_1, \ldots, m_k$ be positive integers, and let $m \coloneqq \sum_{k=1}^{k} m_k$. Then the following identities hold weakly over functions in $C^1(\mathrm{conv}(A))$ (for the first two identities) or in $C^m(\mathrm{conv}(A))$ (for the last one):*

$$\frac{\partial}{\partial a_{i,j}} M(x \mid A) = -w_i \frac{\partial}{\partial x_j} M(x \mid A \sqcup \{a_i\}), \tag{4.44}$$
$$\sum_{i=1}^{n} \frac{\partial}{\partial a_{i,j}} M(x \mid A) = -\frac{\partial}{\partial x_j} M(x \mid A),$$
$$\frac{\partial^m}{\partial a_{i_1,j_1}^{m_1} \cdots \partial a_{i_k,j_k}^{m_k}} M(x \mid A) =$$
$$= (-1)^m \frac{(r_1)_{m_1} \cdots (r_k)_{m_k}}{(c)_m} \frac{\partial^m}{\partial x_{j_1}^{m1} \cdots \partial x_{j_k}^{m_k}} M(x \mid A \sqcup \{a_{i_1}\}^{m_1} \sqcup \cdots \sqcup \{a_{i_k}\}^{m_k}).$$

Remark 4.4.5. *The first identity corresponds to (6.6) in [209], and the others are simple consequences of the first one.*

From the theorem above, we can easily derive the following identities, which express various derivatives of a spline function with respect to its knots in terms of its spatial derivatives,

$$\nabla_{a_i} M(x \mid A) = -w_i \nabla_x M(x \mid A \sqcup \{a_i\}),$$

$$\begin{aligned}
\sum_{i=1}^{n} \nabla_{a_i} M(x \mid A) &= -\nabla_x M(x \mid A), \\
\Delta_{a_i} M(x \mid 1) &= \frac{r_i(r_i+1)}{c(c+1)} \Delta_x M(x \mid A \sqcup \{a_i\}^2), \\
\sum_{k=1}^{d} \left(\sum_{i=1}^{n} \frac{\partial}{\partial a_{i,k}} \right)^2 M(x \mid A) &= \sum_{k=1}^{d} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2}{\partial a_{i,k} \partial a_{j,k}} M(x \mid A), \\
&= \Delta_x M(x \mid A).
\end{aligned}$$

Finally, it is also possible to derive a multivariate version of (4.30), first proposed in [209, Theorem 5].

Theorem 4.4.6 (Carlson)**.** *The following relation holds weakly over functions in* $C^2(\mathrm{conv}(A))$ *for all* $1 \leq i, j \leq n$ *and* $1 \leq k \leq d$*:*

$$\left[\sum_{m=1}^{d} (a_{i,m} - a_{j,m}) \frac{\partial^2}{\partial a_{i,m} \partial a_{j,k}} + r_i \frac{\partial}{\partial a_{j,k}} - r_j \frac{\partial}{\partial a_{i,k}} \right] M(x \mid A) = 0. \tag{4.45}$$

Proof. Very similar to the proof of Theorem 4.3.19. The multivariate equivalent of (4.31) reads

$$\lambda_i \frac{\partial}{\partial \lambda_i} f(\Lambda \cdot A) = (a_i - a_n) \cdot \nabla_{a_i} f(\Lambda \cdot A). \tag{4.46}$$

Differentiating (4.46) with respect to $a_{n,k}$ and using (4.43) yields

$$\begin{aligned}
(a_i - a_n) \cdot \nabla_{a_i} \frac{\partial}{\partial a_{n,k}} f(\Lambda \cdot A) &= \frac{\partial}{\partial a_{i,k}} f(\Lambda \cdot A) + \lambda_i \frac{\partial}{\partial \lambda_i} \left(\lambda_n \frac{\partial}{\partial x_k} f(\Lambda \cdot A) \right), \\
&= \lambda_i \lambda_n \frac{\partial}{\partial \lambda_i} \frac{\partial}{\partial x_k} f(\Lambda \cdot A).
\end{aligned}$$

If we integrate using the measure $\mathrm{d}\Sigma_R$, taking the knot derivatives outside the integral, we find

$$\begin{aligned}
(a_i - a_n) \cdot \nabla_{a_i} \frac{\partial}{\partial a_{n,k}} \int_\Sigma f(\Lambda \cdot A) \, \mathrm{d}\Sigma_R &= -\int_\Sigma \frac{\partial}{\partial x_k} f(\Lambda \cdot A) \frac{\partial}{\lambda_i} \left(\lambda_i \lambda_n \, \mathrm{d}\Sigma_R \right), \\
&= -\int_\Sigma \frac{\partial}{\partial x_k} f(\Lambda \cdot A)(r_i \lambda_n - r_n \lambda_i) \, \mathrm{d}\Sigma_R, \\
&= -\int_\Sigma \left(r_i \frac{\partial}{\partial a_{n,k}} - r_n \frac{\partial}{\partial a_{i,k}} \right) f(\Lambda \cdot A) \, \mathrm{d}\Sigma_R,
\end{aligned}$$

after integrating by parts in the first term, with the boundary term being zero for the same reasons as in Theorem 4.3.19, and applying (4.43) again in the last step. We can now invoke Theorem 4.2.1 to exchange the index n with an arbitrary index j, completing the proof, since

$$(a_i - a_j) \cdot \nabla_{a_i} \frac{\partial}{\partial a_{j,k}} = \sum_{m=1}^{d} (a_{i,m} - a_{j,m}) \frac{\partial^2}{\partial a_{i,m} \partial a_{j,k}}.$$

□

Notice that, by exchanging i with j in (4.45), we can obtain the following alternative form,

$$\left[\sum_{m=1}^{d}(a_{i,m}-a_{j,m})\frac{\partial^2}{\partial a_{i,k}\partial a_{j,m}}+r_i\frac{\partial}{\partial a_{j,k}}-r_j\frac{\partial}{\partial a_{i,k}}\right]M(x\mid A)=0, \tag{4.47}$$

which differs from (4.45) because the sum in the first term is over the partial derivative with respect to $a_{j,m}$ instead of $a_{i,m}$.

Again, we will explore some possible implications of (4.47) for seismic inversion and optimization in Chapter 7.

4.5 Discussion and further reading

Historically, the opportunity of using (4.7) as a functional definition of multivariate splines was already recognized by De Boor [204], and some later works by Dahmen and Micchelli [205, 214, 215] showed how to derive many of the useful relations presented in this chapter. A quick introduction can be found in [203, Chapter 18]. This definition did not change significantly in the following works, with the notable exception of the extension through Dirichlet measures [209, 210] used in this chapter. The attention later turned to the construction of polynomial-reproducing spline spaces, which is the subject of Chapter 5.

The definition of multivariate splines as shadows of simplices can be generalized to more general polyhedra. Most of the familiar properties of B-splines, such as their piecewise-polynomial character, regularity and partition of unity are also satisfied by these more general projections, see, e.g., [216–222].

We give hereafter a few results and references related to the theory of generalized simplex spline functions. The goal of this section is simply to give a few pointers in the relevant literature to the interested reader on a few topics that pique our interest, but the content of this section has no bearing at all on the rest of this work, and can thus be entirely skipped without prejudice.

Simplex splines in other areas of mathematics and physics

Simplex splines are defined using shadows (i.e., fibers of projections) of higher dimensional simplices. Other than their immediate geometric interpretation, the concept of polyhedral shadow employed in Theorem 4.1.1 is analogous to the integration process that underlies the *Radon transform*, see, e.g., [223]. This connection is exploited in some algorithms that use B-spline convolution kernels for the computation of the Radon transform and its inverse (see, e.g., [205, 224]).

Dirichlet measures, also called *Dirichlet distributions*, appear in many areas of statistics, and are related to the Gamma and Beta distributions. Their most typical use is when the barycentric coordinates, which are nonnegative and sum to one, are regarded as probabilities assigned to a given set of events. The Dirichlet distribution is then a distribution over all possible ways to assign probabilities to these events, i.e., a distribution over different (probabilistic) models. For this reason, the Dirichlet distribution (and therefore spline functions) often appear in Bayesian statistics to infer probability models from data (see, e.g., [225]), as is the case for example of the well-known *Pólya's urn* model [226].

Dirichlet averages were introduced by Carlson in his comprehensive book [208] in order to tie together many special functions used in applied mathematics, including hypergeometric and confluent hypergeometric functions, Bessel functions and Jacobi polynomials. Even though the term "spline" does not appear at all in the book, the very strong ties between Dirichlet averages and B-splines were later recognized [209, 210], and this connection has since been used to derive some very important properties of univariate and multivariate B-splines, an approach that we have used extensively in this chapter.

Simplex splines and their generalizations make their appearance in many areas of mathematics and physics, often in surprising ways, and sometimes without being recognized as such. For example, the term *moment map* employed in this text originates from the use of analogous maps in Hamiltonian actions, and especially their use in localization formulas of equivariant cohomology [227, 228], where they appear as the pushforward measure defined through the Duistermaat-Heckman formula [229]. Analogously to Dirichlet measures, which pushes forward a measure from a polytope (specifically, a simplex) to euclidean space, this formula is used to push forward measures defined on symplectic manifolds. The gap between these two formulations can be closed using Delzant's theorem [230], which directly relates toric actions on symplectic manifolds with polytopes. For example, if one takes the complex projective space $\mathbb{C}P^{n-1}$ and the toric action parameterized by the real parameter $t \in \mathbb{R}$ and defined by $(z_1, \dots, z_n) \to (e^{\mathrm{i}ta_1} z_1, \dots, e^{\mathrm{i}ta_n} z_n)$, the corresponding Delzant polytope (i.e., the image of the moment map) is a $(n-1)$-dimensional simplex, and the Duistermaat-Heckman measure on $\mathbb{R}$ is exactly a simplex spline. In this context, splines are used for example to efficiently compute the number of integer points in rational polytopes, see, e.g., [219, 231].

Splines also make their (sometimes unexpected) appearance in quantum physics. In [232], a connection between Bézier curves (and thus Bernstein polynomials) and expectation values of spin systems is derived. A similar connection exists more largely for B-splines. Suppose that we are given a self-adjoint operator H (an observable) with real eigenvalues $(a_1, \dots, a_n)$, and suppose that we randomly select a pure state $|\psi\rangle$ from an n-level quantum system and compute the expectation value $E_\psi \coloneqq \langle\psi|H|\psi\rangle$. What is the probability distribution of the expectation values E_ψ as ψ is selected uniformly with respect to the unitarily invariant Fubini-Study metric? It turns out that this is none other than the normalized simplex spline with knots $(a_1, \dots, a_n)$, see, e.g., [233–237]. Since a general finite-dimensional quantum system of dimension n can be represented by the complex projective space $\mathbb{C}P^{n-1}$, and the action of H can be given in a suitable eigenbasis as the toric action described above, this result is not completely unexpected. Splines have also made their appearance in high-energy physics, as (Laplace transforms of) integrands of one-loop calculations in anti-de Sitter (AdS) space, see, e.g., [238]. Some higher-loop calculations provided in this paper also uncover some new, previously unknown covolution formulas for splines, which we do not restate here as they are out of scope for our work.

All these fascinating connections seem to gravitate around the natural link between spline functions and polytopes, to which we give a small contribution in Chapter 5. Nevertheless, if one considers that splines were originally introduced for modeling shapes in engineering applications, their appearance in so many different physical and mathematical subfields is undoubtedly very surprising.

Knot dependence of spline functions in statistics

The derivation of formulas for the derivatives of spline functions with respect to knot positions was done following the thorough analysis of [208], and especially its sections 5.3–5.6, where a handful of extremely useful identities and differential equations are derived for Dirichlet averages. These relations were later explicitly extended to splines (both univariate and multivariate) in [209].

The derivation of these formulas requires switching the point of view from splines with given knots evaluated at a variable location, to splines with variable knots evaluated at a fixed location. This approach is most often encountered in statistical applications, where splines often appear as multivariate probability distributions (see for example [239, 240]). In this context, the roles of variable and parameter are exchanged, and x is seen merely as a real parameter distinguishing between various multivariate distributions on the knots a_i. For example, in [210], this point of view is exploited to obtain general simplex splines as fractional integrals of B-splines, defined using Fourier analysis. Additionally, B-splines and simplex splines seen as functions of their knots have proven invaluable in the study of functions which are radial in the ℓ^1 norm, see, e.g., [241, 242], and especially [243], where this connection is used to compute the Fourier transform of spline functions with respect to their knots.

Analytical continuation and negative multiplicities

Usually, knot multiplicity values for a spline are taken to be positive integers. However, most of the formulas presented in this chapter are valid for complex values of multiplicities with positive real part. Additionally, Carlson [208] proved that an analytical continuation can be performed to relax this positivity condition. The logic underlying this extension is as follows:

1. Find the broadest possible analytical extension for the Dirichlet average (4.10) when f is a polynomial;

2. Whenever f is an analytical function, use its Taylor series representation to prove that the analytical extension is valid if the radius of convergence contains the convex hull of the knots;

3. By the principle of permanence of functional equations, all the functional relations valid in the halfplane $\mathbb{C}_+$ are automatically valid in the extended zone of the complex plane for all multiplicities.

This extension can be used to derive some interesting results on spline functions, that we give here in an informal and very incomplete fashion. We refer the reader to Chapter 6 of [208] for a more complete and detailed analysis.

Consider the Dirichlet average

$$\int_{\mathbb{R}} x^k M(x \mid A, R)\,\mathrm{d}x = \int_{\Sigma} (\Lambda \cdot A)^k \,\mathrm{d}\Sigma_R =: \left\langle x^k \right\rangle_{A,R}, \tag{4.48}$$

which represents the k-th *moment* of the simplex spline $M(x \mid A, R)$ seen as a distribution. We have made explicit the dependency on the knot multiplicities as we will perform an analytic

extension on them. The integral in (4.48) can be computed explicitly, thus giving a closed-form expression for the moments of simplex splines [208, Representation 6.2-1],

$$\left\langle x^k \right\rangle_{A,R} = \frac{k!}{(c)_k} \sum_m \frac{(r_1)_{m_1} \cdots (r_n)_{m_n}}{m_1! \cdots m_n!} a_1^{m_1} \cdots a_n^{m_n}, \tag{4.49}$$

where the sum is extended to all size-n vectors m of nonnegative integers such that $\sum_{i=1}^n m_i = k$. This form can be easily obtained by plugging the multinomial expansion of $f = (\Lambda \cdot A)^k = (\sum_{i=1}^n \lambda_i a_i)^k$ into the definition (4.10) and then applying (4.11) m_i times for every knot a_i. An equivalent recurrence relation is given in [244],

$$\left\langle x^k \right\rangle_{A,R} = \frac{(k-1)!}{(c)_k} \sum_{i=1}^k \left(\sum_{j=1}^n r_j a_j^i \right) \frac{(c)_{k-i}}{c+k-i} \left\langle x^{k-i} \right\rangle_{A,R}, \tag{4.50}$$

with $\langle 1 \rangle_{A,R} = 1$ due to the normalization condition of simplex splines. Notice that, by using the relation (4.21), we can also compute the moments of the usual B-spline functions, simply by multiplying (4.49) and (4.50) by $(a_n - a_1)/(c-1)$.

According to (4.49), the k-th moment of a simplex spline is simply given by a homogeneous polynomial of degree k in the knots $a_1, \ldots, a_n$. Crucially, it is also a rational function of the multiplicities r_i, with the denominator given by $(c)_k$. Therefore, an analytical extension is possible to all values of multiplicities except if $c = \sum_{i=1}^n r_i \in \mathbb{Z}_{\leq 0}$, in which case the denominator is zero for all $k > -c$ and (4.48) develops a pole. This represents the maximum possible extended analytical domain in terms of multiplicities[2].

If f is analytical in a disk $D \subset \mathbb{C}$ with center $\overline{x}$ that contains all the knots in its interior, then for $x \in D$ its Taylor series converges and we can thus write

$$\begin{aligned} \langle f, M(A,R) \rangle &= \sum_{i=0}^\infty \frac{f^{(n)}(\overline{x})}{n!} \int_{\mathbb{R}} (x - \overline{x})^i M(x \mid A, R) \, \mathrm{d}x, \\ &= \sum_{i=0}^\infty \frac{f^{(n)}(\overline{x})}{n!} \left\langle (x - \overline{x})^i \right\rangle_{A,R}. \end{aligned}$$

The series is analytical term by term in the multiplicities $r_1, \ldots, r_n$ and in the knot vector A. As proven in [208, Section 6.3], the convergence of this series is uniform, implying that the series itself is analytical. Therefore, if we restrict the set of test functions f to analytic functions, all the functional equations derived so far (including all the relations from Section 4.2 onward) can be extended to all sets of multiplicities $r_1, \ldots, r_n$ satisfying

$$\sum_{i=1}^n r_i =: c \notin \mathbb{Z}_{\leq 0}.$$

We focus now on the analytical extension of the simplex splines to complex values of its knots. As noticed in [209], we can recover (and generalize) the simplex splines and their action

[2]Unless regularization is achieved by pre-multiplying each side of the equation by $\Gamma(c)$, thus avoiding the poles. We will not consider this technique here.

as distributions via the Dirichlet average of x^{-1}. We start by applying Cauchy's integral formula to the action of a simplex spline (4.10) on an analytical function f:

$$\begin{aligned}\langle f, M(A,R)\rangle &= \frac{1}{2\pi\mathrm{i}}\int_{\mathbb{R}}\oint_{\gamma}\frac{f(z)}{z-x}M(x\mid A,R)\,\mathrm{d}z\,\mathrm{d}x,\\ &= \frac{1}{2\pi\mathrm{i}}\oint_{\gamma} f(z)\left\langle (z-x)^{-1}\right\rangle_{A,R}\,\mathrm{d}z,\\ &= \frac{1}{2\pi\mathrm{i}}\oint_{\gamma} f(z)\left\langle x^{-1}\right\rangle_{z-A,R}\,\mathrm{d}z,\end{aligned} \tag{4.51}$$

where $(z-A)_i = z-a_i$ and the contour γ lies inside D and contains all the knots of A in its interior. Since the knots all lie on the real axis, we can deform γ into two parallel lines, one running from $\min(A)$ to $\max(A)$ below the real line, and the other running in the opposite direction above it, plus two small half circles. It can be proven [208, 209] that the integral along the two infinitesimally small circles vanishes. Thus, the shape of the spline at each point x is determined by the behavior of the average $\left\langle x^{-1}\right\rangle_{z-A,R}$, seen as a function of z, on the real line, and specifically in the interval $(\min(A),\max(A))$. In particular:

- A *branch cut* over an interval $(x_1,x_2)\subset(\min(A),\max(A))$, with the jump at x being equal to $J(x)$, gives a contribution to the integral (4.51):

  $$\int_{x_1}^{x_2} f(x)J(x)\,\mathrm{d}x,$$

  and hence the spline at x is equal to the *jump* across the branch cut at x;

- A pole of the form $r/(z-x_0)$ gives, by Cauchy's integral formula, a contribution:

  $$rf(x_0) = \int_{\mathbb{R}} f(x)\cdot r\delta(x-x_0)\,\mathrm{d}x,$$

  and thus the spline function at x_0 behaves as $r\delta(x-x_0)$. Similarly, a pole of the form $r/(z-x_0)^{k+1}$ contributes a term $rk!\delta^{(k)}(x-x_0)$ to the shape of the spline function;

- Finally, the sub-intervals without branch cuts or poles do not contribute to the integral, since the contour γ can be shrunk to a point. Hence, they do not contribute to the shape of the spline function.

Summing the contributions of all branch cut jumps and poles, we obtain the final shape of the spline.

A lot of functional relations involving the Dirichlet average $\left\langle x^{-1}\right\rangle_{z-A,R}$ are known, and plenty can be found in Chapters 5, 6 and 8 of [208]. These formulas give rise to recurrence relations that can be used to reduce the Dirichlet average to a set of simple averages, whose branch cuts and poles are known. For example, Section 8.5 of [208] contains a recurrence relation capable of simplifying any Dirichlet average with integer (positive or negative) multiplicities. By using the above-defined dictionary to convert branch cut jumps and poles into distributions, we obtain corresponding recurrence relations that enable the calculation of simplex splines with generalized

(negative or even fractional) multiplicities. As an example, we present in Algorithm 2 a recursive procedure, unpublished as far as we know, that can be deduced to compute simplex splines with both positive and negative integer multiplicities $R = (r_1, \ldots, r_i)$ satisfying the constraint $c = \sum_{i=1}^{n} r_i > 0$. The algorithm is based on a series of identities that can be found in [208], together with the necessary proofs.

Algorithm 2 Evaluation of a simplex spline $M(x \mid A, R)$ with integer (but possibly negative) multiplicities, with $c \coloneqq \sum_{i=1}^{n} r_i > 0$. The references in the comments to the right refer to the equations in [208] used to derive the relation.

Require: $R = (r_1, \ldots, r_n)$ is a vector of integer multiplicities, not necessarily positive, with $\sum_{i=1}^{n} r_i = c > 0$; A is the knot vector

return DSpline$(x, 1, A, R)$, defined below.

function DSpline(x, u, A, R)
  if $u = 1$ and all multiplicities are 0 except for two that are 1 **then**
    $i \leftarrow$ index of first knot with multiplicity 1
    $j \leftarrow$ index of second knot with multiplicity 1
    return $\chi_{[a_i, a_j]}(x)$ ▷ Eq. 8.5-2
  else if all multiplicities are 0 except for one that is positive **then**
    $i \leftarrow$ index of only knot with nonzero multiplicity
    return $(u-1)!\delta^{(u-1)}(x - a_i)$ ▷ Eq. 5.9-22 with $\beta' = 0$
  else if there are negative multiplicities **then**
    $i \leftarrow$ index of first knot with negative multiplicity
    return $\frac{c-u}{c}$DSpline$(x, u, A, R + E_i) + \frac{u}{c} a_i$DSpline$(x, u+1, A, R + E_i)$ ▷ Eq. 5.9-7
  else if $u > 1$ and there are at least two knots with positive multiplicity **then**
    $i \leftarrow$ index of first knot with positive multiplicity
    $j \leftarrow$ index of first knot with positive multiplicity
    return $\frac{c-1}{u-1} \frac{1}{a_i - a_j}$(DSpline$(x, u-1, A, R - E_i)$
      $-$DSpline$(x, u-1, A, R - E_j)$) ▷ Eq. 8.5-1
  else at least two multiplicities are positive and $c > u + 1$
    $i \leftarrow$ index of first knot with positive multiplicity
    $j \leftarrow$ index of last knot with positive multiplicity
    return $\frac{c-1}{c-u-1} \frac{1}{a_i - a_j}$($a_i$DSpline$(x, u, A, R - E_j)$
      $- a_j$DSpline$(x, u, A, R - E_i)$) ▷ exercise 5.9-6

The unifying perspective of allowing negative multiplicities may be able to bring a useful point of view to many different topics. One such example is the theory of *quasi interpolants*, which are used in signal processing to iteratively compute the best-fitting spline to a set of data points that are slowly evolving, because they are being acquired and/or they are being deleted from a limited-space buffer. A best fit based on the minimisation of the ℓ^2 norm would require an inversion of a global matrix each time a data point is modified. In contrast, quasi-interpolants are used to compute the best fitting spline Qf to a given function f,

$$(Qf)(x) = \sum_{i=1}^{n} q_i(f) N_{i,k}(x),$$

where every coefficient $q_i(f)$ is obtained thorough the application of a linear operator q_i depending only on the behavior of the function f in or around the support of the basis function $N_{i,k}$. Typically, the linear operators q_i are chosen to be either *integral*, i.e., based on the superposition integral of f with a given function, *discrete*, i.e., based on punctual values of f, or *differential*, that is, based on the values of the derivatives of f at a few points. Simplex splines with both positive and negative multiplicities are able to resume naturally all three types of linear functionals in a single linear operator, belonging to a wide family of functions interconnected by many recurrence relations. This approach might therefore make it easier to develop new quasi-interpolants satisfying the all-important properties of projection, polynomial reproduction and locality. For an introduction to quasi-interpolants, see [245] and references therein.

Furthermore, splines with negative multiplicities might be useful for the derivation of dual bases to B-splines, i.e., polynomial-reproducing bases of functions orthogonal to the usual B-splines, with potential applications in Petrov-Galerkin approaches to isogeometric analysis.

Finally, notice that the extension of B-splines to complex order has already been pursued by other authors, both using simplex splines (see [246, 247]) and independently of it (see, e.g., [248]).

Simplex splines as Green's functions of a hyperbolic operator

Let us go back to the Euler-Darboux equation (4.30) satisfied by simplex splines. After the change of variables

$$\begin{aligned}\rho &= a_i - a_j,\\ \tau &= a_i + a_j,\end{aligned}$$

the differential equation can be rewritten as

$$\left[\frac{\partial^2}{\partial\tau^2} + \frac{r_i - r_j}{\rho}\frac{\partial}{\partial\tau} - \frac{\partial^2}{\partial\rho^2} - \frac{r_i + r_j}{\rho}\frac{\partial}{\partial\rho}\right] M(x \mid A) = 0. \tag{4.52}$$

Consider now a function $g : \mathbb{R}^n \to \mathbb{R}$ that depends only on the radial coordinate ρ. Then, the Laplacian operator of $\mathbb{R}^n$ acting on g only contains the radial part,

$$\Delta g = \frac{\partial^2}{\partial\rho^2} g + \frac{n-1}{\rho}\frac{\partial}{\partial\rho} g.$$

We can therefore rewrite (4.52) in $\mathbb{R}^n$ as:

$$\left[\frac{\partial^2}{\partial\tau^2} + \frac{r_i - r_j}{\rho}\frac{\partial}{\partial\tau} - \Delta\right] M(x \mid A) = 0. \tag{4.53}$$

This shows that a simplex spline can be interpreted as the Green's function of a hyperbolic operator. If we restrict ourselves to the case $r_i = r_j = r \in \mathbb{N}$, (4.53) reduces to the wave equation in dimension $2r + 1$. A general form of the solution of (4.30) can be found in [249].

Other connections with differential equations of mathematical physics can be found in [208, Section 5.4].

Projective extension of simplex spline functions

Consider once more the Euler-Darboux equation (4.30) obeyed by simplex splines with respect to any couple of knots,

$$\left[(a_i - a_j)\frac{\partial^2}{\partial a_i \partial a_j} + r_i \frac{\partial}{\partial a_j} - r_j \frac{\partial}{\partial a_i}\right] M(x \mid A) =: PM(x \mid A) = 0. \tag{4.54}$$

The symmetries of this equation have been thoroughly studied in [250], where it is shown that the space of solutions of (4.54) is invariant with respect to the action of a Lie group, whose Lie algebra is generated by the following operators:

$$\begin{aligned} X &= a_i^2 \frac{\partial}{\partial a_i} + a_j^2 \frac{\partial}{\partial a_j} + r_i a_i + r_j a_j, \\ H &= 2a_i \frac{\partial}{\partial a_i} + 2a_j \frac{\partial}{\partial a_j} + r_i + r_j, \\ Y &= -\frac{\partial}{\partial a_i} - \frac{\partial}{\partial a_j}. \end{aligned}$$

It is easy to check that these linear first-order differential operators satisfy

$$\begin{aligned} [X, P] &= -(a_i + a_j)P, \\ [H, P] &= -2P, \\ [Y, P] &= 0, \end{aligned}$$

and thus $[X, P]M(x \mid A) = [H, P]M(x \mid A) = [Y, P]M(x \mid A) = 0$. This means that, whenever $M(x \mid A)$ is a solution of $PM(x \mid A) = 0$, then $PXM(x \mid A) = (X + a_i + a_j)PM(x \mid A) = 0$, implying that $XM(x \mid A)$ is also a solution. The same argument works for H and Y, and in fact, since the equation is homogeneous, for the whole Lie algebra generated by these operators, as argued in [250]. The Lie group obtained from this algebra is therefore a *symmetry* of the differential equation (4.54), at least in a neighbourhood of the identity. In order to understand the shape of this symmetry, we can compute the mutual commutators of the operators $\{X, H, Y\}$. A simple calculation shows that the commutators are as follows:

$$[X, Y] = H,\ [H, Y] = -2Y,\ [H, X] = 2X.$$

We recognize here the generators of the Lie algebra $\mathfrak{sl}(2, \mathbb{R})$. Together with the identity, which is trivially a symmetry of (4.54), the Lie algebra extends to the full $\mathfrak{gl}(2, \mathbb{R})$, implying that symmetry group of (4.30) is none other than $GL(2, \mathbb{R})$. Straightforward calculations lead to an explicit form of the representation of $GL(2, \mathbb{R})$ in the space of solutions of (4.54). Let $f(a_i, a_j)$ be a solution of (4.54), and consider as an ansatz the following *projective linear* (*Möbius*) transformation on the argument of a function f, expanded for a small value of its parameter:

$$f\left(\frac{(1 + at)x + ct}{btx + (1 + dt)}\right) = f(x) + f'(x)\left(c + (a - d)x - bx^2\right)t + O(t^2).$$

Comparing this with the definition of the operators X, H and Y above, we see that we can obtain a generator for this transformation on both knots by combining the differential operators of the generators in the following Lie algebra element:

$$L = \frac{a-d}{2}H - bX + cY.$$

Applying the transformation $\exp(tL)$ to the function f, solution to (4.54), we obtain

$$\exp(tL)f(a_i, a_j) = R_i(t)R_j(t)f(a_i(t), a_j(t)),$$

with

$$a_k(t) \coloneqq \frac{(1+at)a_k + ct}{bta_k + (1+dt)}.$$

The normalization factor $R_k(x)$, $k = 1, 2$ is produced by the terms in L lacking a derivative operator. The term $(r_i + r_j)(a-d)/2$ does not contribute, since it represents a constant factor that can be discarded since (4.54) is homogeneous. From the term $-b(r_i a_i + r_j a_j)$ we obtain that $R_k(x)$ must obey

$$\frac{dR_k(t)}{dt} = -r_k R_k(t) b \frac{1}{bta_k + (1+dt)},$$

which together with the condition $R_k(0) = 1$ has the solution $R_k(t) = (bta_k + (1+dt))^{-r_k}$. A more detailed and general derivation of these transformations can be found in [250] and in [251, Chapter 1].

Let us now rewrite the Möbius transformation by introducing a 2×2 matrix $g \in GL(2, \mathbb{C})$,

$$g = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}, \; g_{11}g_{22} - g_{21}g_{12} \neq 0,$$

and making the correspondence $g_{11} = 1+a$, $g_{12} = b$, $g_{21} = c$ and $g_{22} = 1+d$. We can summarize what we have obtained so far as follows.

Theorem 4.5.1 (Miller)**.** *Let $f(a_i, a_j)$ be a solution of the Euler-Darboux equation* (4.54)*, and let $g \in GL(2, \mathbb{R})$ be an invertible 2×2 matrix. Then the function*

$$(g_{12}a_i + g_{22})^{-r_i}(g_{12}a_j + g_{22})^{-r_j} f(\rho(g)a_i, \rho(g)a_j),$$

where

$$\rho(g)x = \frac{g_{11}x + g_{21}}{g_{12}x + g_{22}},$$

is also a solution.

A simplex spline $M(x \mid A)$ satisfies the Euler-Darboux equation (4.30) for every couple of knots. Therefore, if we denote by $\mathcal{S}_R$ the space of simultaneous solutions to all equations (4.54) for all $1 \leq i, j \leq n$, we can obtain a representation $\rho : GL(2, \mathbb{R}) \to \mathrm{Aut}(\mathcal{S}_R)$ by letting an element g act on all knot variables simultaneously. We obtain thus the following

Corollary 4.5.2. *Let $M(x \mid A)$ be a simplex spline, and let $g \in GL(2, \mathbb{R})$ be an invertible 2×2*

matrix. Then the function

$$\prod_{k=1}^{n}(g_{12}a_k + g_{22})^{-r_k} M(x \mid \rho(g)A), \tag{4.55}$$

where

$$\rho(g)a_k = \frac{g_{11}a_k + g_{21}}{g_{12}a_k + g_{22}},$$

is a solution to the Euler-Darboux equations (4.54) *for all* $1 \leq i, j \leq n$.

Up to a rescaling of the spline function, ρ acts projectively on the knot variables. Therefore, it is more natural to rewrite its action in projective coordinates, where $\rho(g)$ acts simply a matrix multiplication by g on the right, $\rho(g)(x) = xg$,

$$\begin{pmatrix} a_k, & 1 \end{pmatrix} \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} \simeq \begin{pmatrix} \rho(g)a_k, & 1 \end{pmatrix}.$$

If we take $g, h \in GL(2, \mathbb{R})$, the projective action of gh according to (4.55) reads:

$$\rho(gh)M(x \mid A) = \prod_{k=1}^{n} \left((g_{11}h_{12} + g_{12}h_{22})a_k + (g_{21}h_{12} + g_{22}h22)\right)^{-r_k} M(x \mid \rho(gh)A)),$$

which has a different multiplying coefficient than the action of $\rho(g)\rho(h)$. Therefore, the representation cannot be lifted to a linear representation, and it is truly projective in character.

We now investigate the symmetries of multivariate splines in the case $d = 2$. From the Euler-Darboux equation (4.45), we can build the two differential operators

$$P_k = (a_{i,1} - a_{j,1})\frac{\partial^2}{\partial a_{i,1}\partial a_{j,k}} + (a_{i,2} - a_{j,2})\frac{\partial^2}{\partial a_{i,2}\partial a_{j,k}} + r_i \frac{\partial}{\partial a_{j,k}} - r_j \frac{\partial}{\partial a_{i,k}},\ k = 1, 2,$$

for $k = 1, 2$. The two following equations are obeyed simultaneously by M:

$$P_1 M(x \mid A) = P_2 M(x \mid A) = 0. \tag{4.56}$$

Consider now the following collection of eight operators,

$$\begin{aligned}
X_3 &= a_{i,1}\frac{\partial}{\partial a_{i,2}} + a_{j,1}\frac{\partial}{\partial a_{j,2}},\ \ Y_3 = a_{i,2}\frac{\partial}{\partial a_{i,1}} + a_{j,2}\frac{\partial}{\partial a_{j,1}}, \\
H_1 &= 2a_{i,1}\frac{\partial}{\partial a_{i,1}} + 2a_{j,1}\frac{\partial}{\partial a_{j,1}} + a_{i,2}\frac{\partial}{\partial a_{i,2}} + a_{j,2}\frac{\partial}{\partial a_{j,2}} + r_i + r_j, \\
H_2 &= 2a_{i,2}\frac{\partial}{\partial a_{i,2}} + 2a_{j,2}\frac{\partial}{\partial a_{j,2}} + a_{i,1}\frac{\partial}{\partial a_{i,1}} + a_{j,1}\frac{\partial}{\partial a_{j,1}} + r_i + r_j, \\
Y_1 &= -\frac{\partial}{\partial a_{i,1}} - \frac{\partial}{\partial a_{j,1}},\ \ Y_2 = -\frac{\partial}{\partial a_{i,2}} - \frac{\partial}{\partial a_{j,2}}, \\
X_1 &= a_{i,1}^2\frac{\partial}{\partial a_{i,1}} + a_{j,1}^2\frac{\partial}{\partial a_{j,1}} + a_{i,1}a_{i,2}\frac{\partial}{\partial a_{i,2}} + a_{j,1}a_{j,2}\frac{\partial}{\partial a_{j,2}} + r_i a_{i,1} + r_j a_{j,1}, \\
X_2 &= a_{i,2}^2\frac{\partial}{\partial a_{i,2}} + a_{j,2}^2\frac{\partial}{\partial a_{j,2}} + a_{i,1}a_{i,2}\frac{\partial}{\partial a_{i,1}} + a_{j,1}a_{j,2}\frac{\partial}{\partial a_{j,1}} + r_i a_{i,2} + r_j a_{j,2}.
\end{aligned}$$

It can be easily checked that

$$
\begin{aligned}
&[X_1, P_1] = -(a_{i,1} + a_{j,1})P_1 - a_{j,2}P_2, \;\; [X_1, P_2] = -a_{i,1}P_2, \\
&[X_2, P_1] = -a_{i,2}P_1, \;\; [X_2, P_2] = -(a_{i,2} + a_{j,2})P_2 - a_{j,1}P_1, \\
&[H_1, P_1] = [H_2, P_1] = -P_1, \;\; [H_1, P_2] = [H_2, P_2] = -P_2, \\
&[Y_1\ , P_1] = [Y_2, P_1] = [Y_1, P_2] = [Y_2, P_2] = 0, \\
&[X_3, P_1] = -P_2, \;\; [X_3, P_2] = 0, \;\; [Y_3, P_2] = -P_1, \;\; [Y_3, P_1] = 0.
\end{aligned}
$$

Once again, these operators preserve the space of functions which are a solution to *both* equations (4.56), and so does the Lie algebra spanned by them. The commutation relations in Table 4.1 are also straightforward (albeit tedious) to check.

$[\,\cdot\,,\,\cdot\,]$	X_1	Y_1	H_1	X_2	Y_2	H_2	X_3	Y_3
X_1	0	H_1	$-2X_1$	0	X_3	$-X_1$	0	$-X_2$
Y_1	$-H_1$	0	$2Y_1$	$-Y_3$	0	Y_1	Y_2	0
H_1	$2X_1$	$-2Y_1$	0	X_2	$-Y_2$	0	X_3	$-Y_3$
X_2	0	Y_3	$-X_2$	0	H_2	$-2X_2$	$-X_1$	0
Y_2	$-X_3$	0	Y_2	$-H_2$	0	$2Y_2$	0	Y_1
H_2	X_1	$-Y_1$	0	$2X_2$	$-2Y_2$	0	$-X_3$	Y_3
X_3	0	$-Y_2$	$-X_3$	X_1	0	X_3	0	$H_1 - H_2$
Y_3	X_2	0	Y_3	0	$-Y_1$	$-Y_3$	$H_2 - H_1$	0

Table 4.1: Commutation relations of the symmetry operators for 2-dimensional simplex splines. They are equivalent to the commutation relations of $\mathfrak{sl}(3, \mathbb{R})$.

Once again, after adding the identity, this algebra is isomorphic to the algebra of $\mathfrak{gl}(3, \mathbb{R})$, suggesting a hidden projective structure for the nodes in two dimensions, just like in one dimension. The action of the projective representation of $GL(3, \mathbb{R})$ on a spline function can be computed as suggested in [250]. This construction might be generalizable to all dimensions, allowing to define general multivariate splines on projective spaces. We are not aware of any current research in this direction.

5 | Simplex spline spaces for numerical analysis

> No hay en la vasta Biblioteca, dos libros idénticos. De esas premisas incontrovertibles dedujo que la Biblioteca es total y que sus anaqueles registran todas las posibles combinaciones de los veintitantos símbolos ortográficos (número, aunque vastísimo, no infinito) o sea todo lo que es dable expresar: en todos los idiomas.
>
> *Jorge Luis Borges, El Jardín de senderos que se bifurcan, La Biblioteca de Babel (1941)*

As discussed in Chapter 3, spline functions have recently entered the realm of partial differential equations via the new field of *isogeometric analysis* [2, 252]. Under this new analytical paradigm, B-spline (and NURBS) basis functions replace the more traditional polynomial bases used in finite element (FE) simulations.

We have seen in Chapter 2 that, when evaluating the order of spatial approximation of a function space via Jackson-type inequalities, one usually requires the space to be *polynomial-reproducing*, i.e., containing in its linear span all the polynomials up to degree k. This property is certainly satisfied by univariate B-splines (see (3.17)), and maximally so, since degree-k univariate B-spline spaces reproduce degree-k polynomials.

Another property of B-spline functions, crucial for their application to real-world problems, is the availability of robust and efficient algorithms to construct the polynomial-reproducing function spaces, as well as efficiently determine and evaluate all the functions supported at a given point. In fact, the classic Cox-De Boor recurrence formula (3.15) immediately leads to an efficient evaluation algorithm capable of reusing intermediate results, such as the classic pyramidal evaluation scheme presented in [178, Chapter X, Algorithm 8], which underpins virtually all practical implementations of B-splines (see Figure 3.5).

In more than one dimension, one needs multivariate equivalents of all these characteristics. Perhaps the simplest solution, used in almost all CAD (Computer Aided Design) software and models, is through the use of tensor products. This corresponds to defining, starting from a collection of d univariate B-spline bases indexed by $i_1, \dots, i_d$ and with orders $k_1, \dots, k_d$, the multivariate B-spline basis functions

$$N_{i_1,\dots,i_d,k_1,\dots,k_d}(x_1, \dots, x_d) \coloneqq N_{k_1,i_1}(x_1) \cdots N_{i_d,k_d}(x_d). \tag{5.1}$$

The d-dimensional knots defining these splines can be thought of as forming a rectilinear d-dimensional grid. This definition guarantees that the one-dimensional properties discussed above directly and trivially extend to any number of dimensions, and also simplifies many calculations due to the separability of superposition integrals.

Even with all its advantages, the tensor product structure of standard multivariate B-spline basis functions can be too rigid for some specific applications, as for example the simulation of PDEs in the natural sciences, where the physical parameters of a domain can have complex geometrical interfaces of reduced regularity or discontinuity, with arbitrary topology, and no CAD model is available.

Consider for example the seismic waveform inversion problem, in which an unknown model, containing discontinuities, needs to be reconstructed starting from a smooth initial guess. Geological features such as salt domes, diapirs, hiatuses and faults can greatly contribute to the complex shape and topology of these discontinuities. At some point, the inversion process would require a localized reduction in the regularity of the model, in order to reproduce the reflectors seen in the measured data. The only way to obtain this with the multivariate splines defined as in (5.1) would be to duplicate some of the knots in one or more dimensions, according to property 5 of Theorem 3.1.1. However, due to the rigid structure of the tensor product, it would then be impossible to localize the discontinuity in a small subregion of the domain. One might try to avoid this issue by using a curved geometry, and rely on control points to reproduce the shape of the irregularity. Even then, it would still be impossible to change the simple topology of a B-spline patch, and some geometries are too complex to be easily defined with only a few isogeometric patches.

This issue is greatly amplified when dealing with inverse problems such as seismic inversion, since the location, number and topology of the discontinuities is *a priori* unknown. Thus, one would need a mechanism to dynamically insert and remove many isogeometric patches during the inversion process, with arbitrary topology, and with a sufficiently good geometrical fit between them. For this reason, we believe that the use of unstructured splines such as the simplex splines introduced in Chapter 4 might simplify some of these issues, and therefore be a good choice for the description of the heterogeneous, non-smooth and complex structure of the subsurface.

Many relevant properties for the construction and evaluation of a single simplex spline function have been given in the previous chapter. The issue that we tackle in this chapter is precisely the construction of polynomial-reproducing spline spaces based on simplex splines in more than one dimension.

5.1 State of the art

After the introduction of simplex splines, it was immediately apparent that their definition led quite naturally to the construction of an unstructured spline space on d-dimensional simplicial complexes [253]. Starting from the triangulation of a $(d+k)$-dimensional polytope, often simply the cartesian product between a d-dimensional domain Ω with a k-dimensional slab, one can employ the usual definition of a simplex spline (4.1), but with a projection $\pi : \mathbb{R}^{d+k} \to \mathbb{R}^d$ on the first d coordinates instead of just the first one. This produces a set of spline functions of degree k, one for every $(d+k)$-dimensional simplex of the triangulation of the polytope, with each

spline supported over a number of adjacent d-dimensional simplices. The fact that this basis constitutes a partition of unity is clear from the definition, but, perhaps more surprisingly, the basis can be proven to reproduce *all* polynomials of degree k [254]. However, this construction was immediately seen to be highly impractical, requiring the triangulation of high-dimensional polytopes of arbitrary shape. An alternative formulation based on the triangulation of simploids, i.e., cartesian products of simplices, was later proposed [255], only requiring the triangulation of the d-dimensional domain of interest instead of a $(d+k)$-dimensional polytope. However, even this formulation, aside from being dependent on the chosen triangulation and other algorithmic details, does not satisfy all the important properties required of a spline basis, and does not naturally reduce to the usual univariate B-spline basis for $d = 1$. A good review of these and other attempts can be found in [256].

A more recent approach to the problem of unstructured multivariate spline bases focuses not on the definition of a spline basis on a a given triangulation, but simply on the selection of an appropriate family of n_s subsets $(S_i)_{i=1}^{n_s}$, $S_i \subset A$, $i = 1, \dots, n_s$, of the knot vector A, such that the set of splines $\{M(x \mid S_i)\}_{i=1}^{n_s}$ is polynomial-reproducing. A first attempt [257], still originating from an underlying triangulation but requiring the selection of a special class of auxiliary knots, was put forward by Dahmen and Micchelli. A more symmetrical result was obtained by Neamtu [4, 256, 258], who proved that, if all the *Delaunay configurations* of order k are chosen as subsets of knots, the resulting spline basis has order k and indeed reproduces all the polynomials up to degree k. Delaunay configurations are a generalization of the usual (order-0) Delaunay triangulations, and correspond to all the sets of indices (B, I), with $|B| = d+1$ and $|I| = k$, such that the circumsphere of the points $(a_i)_{i\in B}$ contains exactly the points $(a_i)_{i\in I}$ in its interior, and no other point of A. The corresponding spline is then based on the union $(a_i)_{i\in I\sqcup B}$ of the $d+1$ *boundary* knots $(a_i)_{i\in B}$ with the k *internal* knots $(a_i)_{i\in I}$.

More recently, in the case of bivariate splines (i.e., $d = 2$), Liu and Snoeyink [5, 259] have provided a constructive algorithm that allows to build order-k Delaunay configurations iteratively, starting from $k = 0$. Their algorithm is derived as the dual of an algorithm by Lee [260] that allows the construction of k-th order Voronoi subdivisions, and involves so-called *centroid triangulations*. We will discuss Liu's algorithm in the next section. In their work, Liu and Snoeyink do not use directly the above definition, but instead propose an algorithm capable of computing regular configurations of order k starting from those of order $k-1$. They are able to prove that their algorithm produces a compatible family of (generalized) Delaunay configurations of order k for $k \leq 3$. More recently, Schmitt [6] introduced another generalization of Delaunay configurations by replacing circles with families of convex Jordan curves that intersect pairwise at most twice. He proved that the *regular convex* Delaunay configurations defined in this way correspond exactly to those produced by the algorithm of Liu and Snoeyink, and that they satisfy the properties used by Neamtu in his proof of polynomial reproduction, thus establishing the validity of the algorithm for all degrees k in dimension 2. Since then, many interesting spline spaces, although not as general as Neamtu's approach, have been built on suitable triangulations and subdivisions (see, e.g., [261, 262]).

Overall, we find that the current state-of-the-art algorithms and approaches for simplex splines are still lacking some important features compared with their tensor-product counterparts, especially since the usual geometrical approaches seem to complicate significantly in $d > 2$.

Moreover, the current formulations fall short of treating the case of repeated knots, which is required in order to impose boundary conditions and to locally control the regularity of the solution. Finally, no simple and general evaluation scheme is known for simplex spline spaces, and the structure of these spaces has not been investigated in sufficient detail to formulate efficient numerical quadratures.

The rest of the chapter is devoted to providing a few contributions to these issues.

5.2 Background

This chapter uses notions coming from combinatorial geometry, for which we adopt some standard concepts and notation. We give here a summary of this notation, a brief introduction of the combinatorial objects used in this chapter, as well as a quick reminder of the properties of simplex splines that are relevant here.

5.2.1 Notation

Given $n \in \mathbb{Z}^+$, we define the range $[n] \coloneqq \{1, \dots, n\}$. The union between two disjoint sets R and S is denoted by $R \sqcup S$, and its complement $[n] \setminus (R \sqcup S)$ is denoted by $\overline{R \sqcup S}$. Note that $|R \sqcup S| = |R| + |S|$, where $|\cdot|$ denotes the cardinality of a set. We also borrow some convenient notation from [4]. In particular, given a configuration of $n \geq d+1$ points $A \coloneqq (a_1, \dots, a_n)$ in $\mathbb{R}^d$ and a set of indices $I \subseteq [n]$ such that the points $(a_i)_{i \in I}$ are affinely independent, we denote by $\det(I)$ the $(d+1) \times (d+1)$ determinant $\det((a_i, 1)_{i \in I})$, with the rows ordered so that $\det(I) > 0$. Similarly, we denote by $\det({}^k_j I)$ the result of replacing the row corresponding to $(a_j, 1)$ in $\det(I)$ with $(a_k, 1)$ in the same position. Notice that $\det({}^k_j I)$ is not necessarily positive. Similarly, for $x \in \mathbb{R}^d$, $\det({}^x_j I)$ is obtained by replacing the row $(a_j, 1)$ in $\det(I)$ with $(x, 1)$.

Let now $\mathcal{R}(A) \subset \mathbb{R}^d$ be any compact polytopal region with vertices in A. A *subdivision* $\mathcal{T}$ of $\mathcal{R}(A)$ is a collection of d-dimensional polytopes Δ with vertices in A such that $\bigcup_{\Delta \in \mathcal{T}} \Delta = \mathcal{R}(A)$ and any two distinct polytopes in $\mathcal{T}$ have disjoint interiors and share a common face, possibly empty. If all the polytopes are simplices, then $\mathcal{T}$ is a *triangulation* of $\mathcal{R}(A)$. Notice that only points in A are allowed to be vertices of a subdivision. Consequently, for $d \geq 3$, there exist polytopal regions that cannot be triangulated in our sense, such as for example Schönhardt's polyhedron [263], see Figure 5.1 and also [264, 265] for some generalizations and further combinatorial aspects of triangulations of polyhedra.

5.2.2 Simplex splines

We have introduced multivariate simplex spline functions in Chapter 4, along with some of their properties. In this chapter, we rely mainly on the recurrence formula (4.36), which we express as follows. Let $A = (a_1, \dots, a_n)$ be a vector of points in $\mathbb{R}^d$ and let $X \subseteq [n]$ be a subset of size $|X| = k + d + 1$. Then the normalized multivariate spline function $M(x \mid (a_i)_{i \in X})$ of degree k

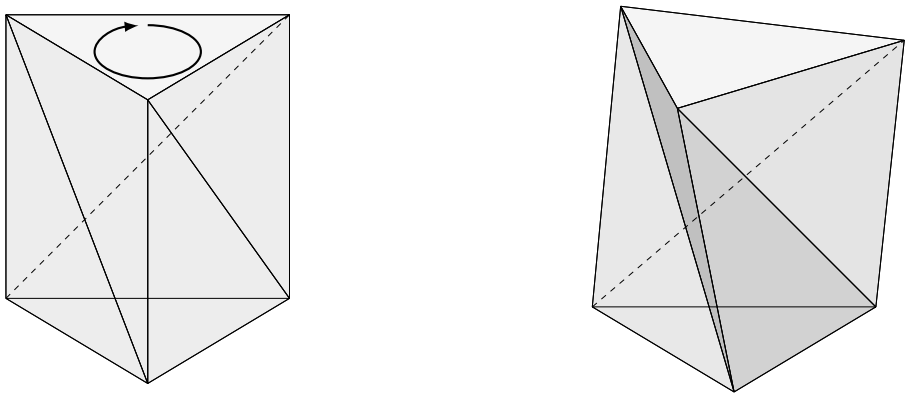


Figure 5.1: (Right) Schönhardt's non-triangulable polyhedron, obtained by twisting one of the faces of a triangular prism (left).

can be evaluated for all $x \in \mathbb{R}^d$ as

$$M(x \mid (a_i)_{i\in X}) \coloneqq \begin{cases} \dfrac{d!}{\det(X)} \mathbf{1}_X(x) & \text{if } k = 0, \quad (5.2a) \\ \dfrac{k+d}{k} \displaystyle\sum_{b\in B} \frac{\det({}^{x}_{b}B)}{\det(B)} M(x \mid (a_i)_{i\in B\setminus\{b\}}) & \text{otherwise,} \quad (5.2b) \end{cases}$$

where $\mathbf{1}_X(x) \coloneqq \mathbf{1}_{\mathrm{conv}(\{a_i\}_{i\in X})}(x)$ is the indicator function of the convex hull of the points indexed by X, the determinants $\det(\cdot)$ are defined in Section 5.2.1 above, and B is any subset $B \subseteq X$ with $|B| = d+1$ such that the points $(a_b)_{b\in B}$ are affinely independent. If no such B exists, then the affine rank of the points indexed by X is less than $d+1$ and the spline, supported on a zero-measure set, is considered to be zero everywhere by continuity. It is a remarkable consequence of (5.2) that the result is independent of the choice of B at each step (see, e.g., [205]). Also notice that, as discussed in Chapter 4, simplex splines integrate to one, i.e., $\int_{\mathbb{R}^d} M(x \mid (a_i)_{i\in X})\mathrm{d}x = 1$. Recall that the functions $M(x \mid \cdot)$ are multivariate piecewise-polynomial functions of $x \in \mathbb{R}^d$ of maximum degree k, with regularity C^{k-1} if all the points are affinely independent, and with reduced regularity otherwise.

Another expression we rely on is the multivariate version of the *knot insertion* formula (4.22). Specifically, if $|X| \geq d+2$ (i.e., if $k \geq 1$), we can select another index $c \in X \setminus B$. We then have

$$\det(B) M(x \mid (a_i)_{i\in X\setminus\{c\}}) = \sum_{b\in B} \det({}^{c}_{b}B) M(x \mid (a_i)_{i\in X\setminus\{b\}}). \quad (5.3)$$

Just like (5.2b) relates splines of order k and $k-1$, allowing for a recurrent evaluation scheme, (5.3) relates splines with the same order $k-1$.

5.2.3 Vector configurations and zonotopal tilings

We give here a very quick introduction to zonotopal tilings, and we refer the reader to [266] or [267, Chapter 6] for a thorough introduction to these combinatorial objects.

Let $A = (a_1, \ldots, a_n)$ be a configuration of points $a_i \in \mathbb{R}^d$, not necessarily affinely independent

or even distinct, but which affinely span $\mathbb{R}^d$. For each point a_i, define its *projective lift* as $v_i \coloneqq (a_i, 1) \in \mathbb{R}^{d+1}$, and let $V \coloneqq (v_1, \dots, v_n)$ be the associated *vector configuration*.

Given two subsets $P, Q \subset \mathbb{R}^d$, their *Minkowski sum* is defined as the set $P + Q \coloneqq \{x + y \in \mathbb{R}^d : x \in P, y \in Q\}$. The Minkowski sum of a set of segments is a special convex polytope known as a *zonotope*. There is a natural zonotope $Z(V) \subset \mathbb{R}^{d+1}$ associated to each point configuration V, defined as follows. For every index $i \in [n]$, define the segment $[0, v_i] \coloneqq \{\alpha_i v_i \in \mathbb{R}^{d+1} : 0 \leq \alpha_i \leq 1\}$. Then, $Z(V)$ is given by the Minkowski sum

$$Z(V) \coloneqq \sum_{i=1}^{n} [0, v_i]. \tag{5.4}$$

Given two subsets of indices $I, B \subseteq [n]$ with $I \cap B = \varnothing$, $|B| = d + 1$ and $\det(B) > 0$, the parallelepiped $\Pi_{I,B} \subset \mathbb{R}^{d+1}$ is defined as

$$\Pi_{I,B} \coloneqq \sum_{i \in I} v_i + \sum_{b \in B} [0, v_b]. \tag{5.5}$$

A collection $\mathcal{P}$ of parallelepipeds $\Pi_{I,B}$ forming a polyhedral subdivision of $Z(V)$ is known as a *fine zonotopal tiling* of $Z(V)$ (see [268] or [267, Chapter 7]). Notice that the $(d+1)$-dimensional volume of the tile $\mathrm{vol}^{d+1}(\Pi_{I,B})$ is equal to $\det(B)$, and that only B determines the shape of $\Pi_{I,B}$, while I simply shifts its position. An example is shown in Figure 5.2. Notice that the set I of each tile $\Pi_{I,B}$ of $\mathcal{P}$ can be read off as the set of vectors in any shortest path connecting the origin to the base of the tile. In the present work, we call $|I|$ the *order* of the tile $\Pi_{I,B}$, and we denote by $\mathcal{P}^{(k)}$ for any integer $k \geq 0$ the subset $\{\Pi_{I,B} \in \mathcal{P} : |I| = k\}$.

The faces of a tile $\Pi_{I,B}$ are themselves parallelepipeds that are obtained by setting α_i equal to 0 or 1 in some of the segments $[0, v_b]$ of (5.5). Clearly, if $\Pi_{J,C}$ is a face of $\Pi_{I,B}$ then $C \subseteq B$ and $I \subseteq J \subseteq I \sqcup B$. If $|C| = d$ then $\Pi_{J,C}$ is called a *facet* of $\Pi_{I,B}$. Since $\mathcal{P}$ is a subdivision, a facet is either shared between exactly two tiles of $\mathcal{P}$, or it is an external facet of $Z(V)$. It is easily checked that two tiles $\Pi_{I,B}$ and $\Pi_{I',B'}$ share a facet if and only if there are two indices $b \in B$, $b' \in B'$ such that $B \setminus \{b\} = B' \setminus \{b'\} = B \cap B'$ and either $I = I'$, $I = I' \sqcup \{b'\}$, $I' = I \sqcup \{b\}$ or $I \sqcup \{b\} = I' \sqcup \{b'\}$. The shared facet $\Pi_{J,C}$ then satisfies $C = B \cap B'$ and $J = I \cup I'$.

Fine zonotopal tilings possess a number of remarkable properties. First, all such tilings of $Z(V)$ are simply different arrangements of the same set of tile shapes.

Theorem 5.2.1 (Shephard [269, Theorem 56])**.** *Every zonotope $Z(V)$ admits a fine zonotopal tiling, and all fine zonotopal tilings of $Z(V)$ have the same number of tiles, namely one full-dimensional tile for each maximal linearly independent subset of V.*

Moreover, one can remove a point a_i corresponding to an index $i \in [n]$ from A and consider the corresponding zonotope $Z(V \setminus \{v_i\})$. Then, any tiling $\mathcal{P}$ of $Z(V)$ induces a tiling $\mathcal{P}_{[n]\setminus\{i\}}$ on $Z(V \setminus \{v_i\})$, as well as on any zonotope built on a subset of V, as follows.

Lemma 5.2.2. *Let $\mathcal{P}$ be a fine zonotopal tiling of $Z(V)$. Then,*

$$\mathcal{P}_{[n]\setminus\{i\}} \coloneqq \{\Pi_{I,B} \in \mathcal{P} : i \notin I \sqcup B\} \sqcup \left\{\Pi_{I\setminus\{i\},B} : \Pi_{I,B} \in \mathcal{P},\, i \in I\right\} \tag{5.6}$$

is a fine zonotopal tiling of $Z(V \setminus \{v_i\})$*. Similarly, for any* $Q \subseteq [n]$*,*

$$\mathcal{P}_{[n]\setminus Q} \coloneqq \left\{\Pi_{I\setminus Q,B} : \Pi_{I,B} \in \mathcal{P},\, B \cap Q = \varnothing\right\} \tag{5.7}$$

is a fine zonotopal tiling of $Z(V \setminus \{v_q\}_{q\in Q})$*.*

Proof. The induced tiling (5.6), also found in [270, Proposition 4.3], can be recovered from the first half of [266, Lemma 4.2] after noticing that $Z(V)$ and the centrally-symmetric zonotope $\mathcal{Z}$ of [266] are related by the linear equation $2 \cdot Z(V) = \mathcal{Z} + \sum_{i=1}^{n} v_i$. Applying (5.6) repeatedly then yields (5.7). □

Since the tiles in $\mathcal{P}$ form a polyhedral subdivision of $Z(V)$, we can form its dual graph $\mathcal{G}$ by associating to each tile $\Pi_{I,B}$ a vertex in $\mathcal{G}$ and by connecting two tiles $\Pi_{I,B}$ and $\Pi_{I',B'}$ with an edge if and only if the tiles share a facet.

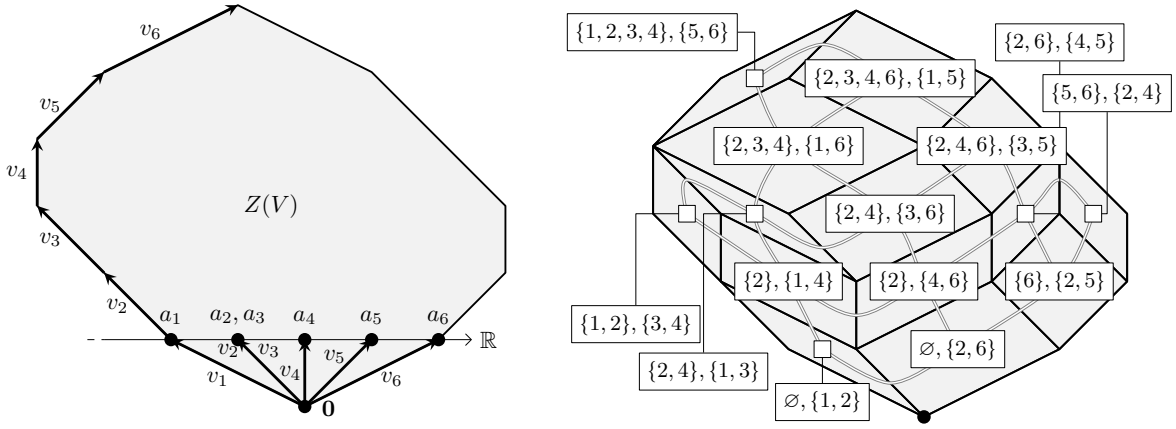


Figure 5.2: (Left) a point configuration $a_1, \ldots, a_6$ in $\mathbb{R}$, with $a_2 = a_3$, their projective lifts $v_1, \ldots, v_6$ and the zonotope $Z(V)$. (Right) a fine zonotopal tiling of $Z(V)$, the subsets I, B associated to each tile $\Pi_{I,B}$, and the dual graph $\mathcal{G}$.

5.3 Polynomial-reproducing spline spaces

As briefly discussed in Section 5.1, Neamtu showed [4] that Delaunay configurations of order k can be used to construct a space of simplex splines which is indeed polynomial-reproducing up to degree k. We introduce here briefly his results, before proposing a generalization. First, let us recall the definition of the *polar form* of a polynomial (see, e.g., [271]):

Definition 5.3.1. *Let* $k \geq 0$ *and let* $q(x)$*,* $x \in \mathbb{R}^d$*, be a* d*-variate polynomial of degree at most* k*. Then there exists a unique function* $Q(x_1, \ldots, x_k)$ *of the* d*-dimensional variables* $(x_1, \ldots, x_k)$ *that is symmetric under permutation of its arguments, affine in each of them, and that agrees with* q *on the diagonal, i.e.,* $Q(x, \ldots, x) = q(x)$*. The function* Q *is called the* polar form *of* q*.*

Let $A = \{a_i\}_{i\in\mathbb{N}}$ be a countably infinite set of points in $\mathbb{R}^d$ in *general position*, i.e., where no subset of $d+1$ points is affinely dependent and no subset of $d+2$ points is co-spherical,

and with no accumulation point. A *Delaunay configuration* $X_{I,B}$ of order $k \geq 0$ is any disjoint couple of sets B, $I \subset \mathbb{N}$ with $|B| = d+1$, $|I| = k$ such that the sphere circumscribed to the simplex $\Delta_B \coloneqq \mathrm{conv}(\{a_b\}_{b\in B})$ contains in its interior the points $\{a_i\}_{i\in I}$ and no other point of A. Notice that this definition depends crucially on the points being in general position. To each such configuration, we can associate through (5.2) the d-variate spline function of order k

$$M(x \mid X_{I,B}) \coloneqq M(x \mid \{a_i\}_{i\in I\sqcup B}).$$

Neamtu's result can be stated as follows:

Theorem 5.3.2 (Neamtu [4])**.** *Let $q(x)$ be a polynomial of degree at most k. Then, for all $x \in \mathbb{R}^d$,*

$$q(x) = \binom{k+d}{d}^{-1} \sum_{X_{I,B}\in D_k} Q((a_i)_{i\in I}) \,\mathrm{vol}^d(\Delta_B) M(x \mid X_{I,B}), \tag{5.8}$$

where Q is the polar form associated to q and the sum is extended to the set D_k of Delaunay configurations of A of order k.

Neamtu's result is based upon some strong assumptions on A, notably the infiniteness and the general position of points in A, which we are able to relax by using the combinatorial nature of zonotopal tilings to our advantage.

Let now $A = (a_1, \ldots, a_n)$ be any finite point configuration in $\mathbb{R}^d$. Assume that the affine span of the points in A is the whole $\mathbb{R}^d$. Let V be the associated vector configuration and $Z(V)$ its associated zonotope, as in Section 5.2.3. Then the following, more general, statement holds:

Theorem 5.3.3. *Let $\mathcal{P}$ be a fine zonotopal tiling of $Z(V)$, let $0 \leq k \leq n-d-1$ and let $\mathcal{P}^{(k)} \coloneqq \{\Pi_{I,B} \in \mathcal{P} : |I| = k\}$. Each tile $\Pi_{I,B} \in \mathcal{P}^{(k)}$ can be associated via (5.2) to the d-variate spline of degree $k = |I|$*

$$M(\,\cdot \mid \Pi_{I,B}) \coloneqq M(\,\cdot \mid (a_i)_{i\in I\sqcup B}). \tag{5.9}$$

Then, for any polynomial $q(x)$ of degree at most k,

$$q(x) = \frac{k!}{(k+d)!} \sum_{\Pi_{I,B}\in\mathcal{P}^{(k)}} Q((a_i)_{i\in I}) \,\mathrm{vol}^{d+1}(\Pi_{I,B}) M(x \mid \Pi_{I,B}) \text{ for } x \in \mathrm{conv}_k(A), \tag{5.10}$$

where Q is the polar form of $q(x)$ and

$$\mathrm{conv}_k(A) = \bigcap_{\substack{S\subseteq[n]\\ |S|=n-k}} \mathrm{conv}(\{a_i\}_{i\in S})$$

is the intersection of the convex hulls of all subsets of A of size $n-k$.

The generalization with respect to Neamtu's result is twofold. First, for a given point configuration A, many different fine zonotopal tilings of $Z(V)$ can be constructed. Each tiling then yields a family of polynomial-reproducing spline spaces for all degrees up to $n-d-1$. In fact, Delaunay configurations can be seen as a special case of this construction, as discussed in the next section.

A second generalization is that the point configuration A is allowed to contain affinely dependent subsets and repeated points. In this case, some of the spline functions have reduced regularity [205], and thus the spline spaces that can be constructed in this way are more generic. Observe that, if all the vertices of $\mathrm{conv}(A)$ are repeated at least $k+1$ times in A, then $\mathrm{conv}_k(A) = \mathrm{conv}(A)$. We obtain therefore a multivariate generalization of the behavior of *clamped* (also called *open*) knot vectors in one dimension.

Corollary 5.3.4. *Assume that each vertex of* $\mathrm{conv}(A)$ *is repeated at least* $k+1$ *times in* A*. Then, in the same conditions as Theorem 5.3.3, the splines* $M(\cdot \mid \Pi_{I,B})$ *for* $\Pi_{I,B} \in \mathcal{P}^{(k)}$ *reproduce polynomials up to order* k *on the whole* $\mathrm{conv}(A)$*.*

The resulting spline space contains functions that are non-vanishing on the boundary of $\mathrm{conv}(A)$. This is a highly desirable property in view of practical applications, especially in numerical analysis, where it can be used to impose non-homogeneous Dirichlet boundary conditions (see, e.g., [2, Section 3.4]).

Finally, notice the similarity between (5.8), (5.10) and the univariate Marsden's identity (3.17). In particular, notice that only the internal knots of the spline (where the notion of internal knot is specific to each of the three cases) appear in the polar form of the polynomial. This aspect will become relevant in Chapter 6.

5.3.1 Proof of Theorem 5.3.3

Neamtu's original proof of the fact that splines associated to Delaunay configurations are polynomial-reproducing (Theorem 4.1 of [4]) rests on a crucial structural property regarding neighbouring pairs of configurations, namely the *edge matching* property proven in [4, Proposition 2.1]. This property underpins also other formulations such as the algorithmic generalization proposed by Liu and and Snoeyink [5, 259] and the geometric description of Schmitt in terms of families of convex Jordan curves [6]. We prove hereafter that a similar property also holds for zonotopal tilings.

Proposition 5.3.5. *Let* $\Pi_{J,C}$ *be a facet of a tile* $\Pi_{I,B} \in \mathcal{P}$*, with* $|J| = k$*. Then* $|I| = k$ *or* $|I| = k-1$*, and exactly one of the following is true:*

(i) $\Pi_{J,C}$ *is shared between* $\Pi_{I,B}$ *and exactly one other tile* $\Pi_{I',B'} \in \mathcal{P}$*, with either* $|I'| = k$ *or* $|I'| = k-1$*. Moreover, if* $\{b\} = B \setminus B'$ *and* $\{b'\} = B' \setminus B$*, the two points* a_b *and* $a_{b'}$ *are separated by the hyperplane* $H \coloneqq \mathrm{aff}(\{a_c\}_{c\in C})$ *if and only if* $|I| = |I'|$*;*

(ii) there exists an index $b \in B$ *such that, for a suitable orientation of the hyperplane* $H \coloneqq \mathrm{aff}(\{a_c\}_{c\in C})$*, the points* $\{a_i\}_{i\in I}$ *are in the positive closed halfspace of* H*, the points* $\{a_i\}_{i\in\overline{I\sqcup B}}$ *are in the negative closed halfspace of* H*, and* a_b *is in the positive open halfspace of* H *if* $b \in J$ *and in the negative open halfspace of* H *if* $b \notin J$*.*

Proof. A facet $\Pi_{J,C}$ of a tile $\Pi_{I,B}$ is obtained by choosing an index $b \in B$ and setting the corresponding coefficient α_b of segment $[0, v_b]$ in (5.5) to either 0, in which case $J = I$, or 1, in which case $J = I \sqcup \{b\}$. Thus, $k \coloneqq |J| = |I|$ or $k \coloneqq |J| = |I| + 1$. Since the tiles in $\mathcal{P}$ form a subdivision of $Z(V)$, $\Pi_{J,C}$ is either a shared facet between $\Pi_{I,B}$ and exactly one other tile $\Pi_{I',B'}$, or it is a boundary facet of $Z(V)$.

In the first case, $C = B \cap B'$, and the previous argument also implies that either $J = I'$ or $J = I' \sqcup \{b'\}$, with $\{b'\} = B' \setminus B$ and $\{b\} = B \setminus B'$, and thus $|I'| = k$ or $|I'| = k - 1$. Since both parallelepipeds are convex polytopes, their interiors are separated by the hyperplane spanned by their common facet, and we can choose a nonzero vector $N \in \mathbb{R}^{d+1}$, normal to the facet, satisfying $\langle v_c, N\rangle = 0$ for all $c \in C = B \cap B'$, and

$$\big\langle z - z', N\big\rangle \geq 0 \tag{5.11}$$

for all $z \in \Pi_{I,B}$ and $z' \in \Pi_{I',B'}$. Notice that necessarily $\langle v_b, N\rangle \neq 0$ and $\langle v_{b'}, N\rangle \neq 0$, since the vectors in B and B' must be linearly independent. The case $|I| = |I'|$ corresponds to either $I = I'$ or $I \sqcup \{b\} = I' \sqcup \{b'\}$. If $I = I'$, then setting $(z, z') = (v_b + \sum_{i\in I} v_i, \sum_{i\in I'} v_i)$ in (5.11) yields $\langle v_b, N\rangle > 0$, while choosing $(z, z') = (\sum_{i\in I} v_i, v_{b'} + \sum_{i\in I'} v_i)$ yields $\langle v_{b'}, N\rangle < 0$. Thus,

$$\operatorname{sign}(\langle v_b, N\rangle) = -\operatorname{sign}(\langle v_{b'}, N\rangle).$$

If $I \sqcup \{b\} = I' \sqcup \{b'\}$, the same choices of (z, z') lead to the same conclusion. The case $|I| \neq |I'|$ is very similar, since it implies either $I = I' \sqcup \{b'\}$ or $I \sqcup \{b\} = I'$. In both cases, plugging the couples $(z, z') = (\sum_{i\in I} v_i, \sum_{i\in I'} v_i)$ and $(z, z') = (v_b + \sum_{i\in I} v_i, v_{b'} + \sum_{i\in I'} v_i)$ in (5.11) leads to

$$\operatorname{sign}(\langle v_b, N\rangle) = \operatorname{sign}(\langle v_{b'}, N\rangle).$$

Thus, the hyperplane $H = \{x \in \mathbb{R}^d : \langle N, (x, 1)\rangle = 0\}$ satisfies the first part of the proposition.

Suppose now that $\Pi_{J,C}$ is a boundary facet of $Z(V)$, with $\{b\} = B \setminus C$. Since $Z(V)$ is a convex polytope, all points $z \in Z(V)$ lie in the same closed halfspace of $\Pi_{J,C}$, and we can choose a nonzero vector $N \in \mathbb{R}^{d+1}$, normal to $\Pi_{J,C}$, so that $\langle v_c, N\rangle = 0$ for all $c \in C$ and

$$\langle z - \sum_{j\in J} v_j, N\rangle \leq 0 \tag{5.12}$$

for all $z \in Z(V)$. Plugging into (5.12), respectively, $z = v_e + \sum_{j\in J} v_j$ with $e \notin J$ and $z = \sum_{j\in J, j\neq f} v_j$ with $f \in J$ shows that

$$\langle v_c, N\rangle = 0,\ \langle v_e, N\rangle \leq 0, \langle v_f, N\rangle \geq 0$$

for all $c \in C$, $e \notin J$ and $f \in J$. Moreover, as before, $\langle v_b, N\rangle \neq 0$ otherwise the vectors indexed by B would be linearly dependent. Therefore, $\langle v_b, N\rangle > 0$ if $b \in J$, and $\langle v_b, N\rangle < 0$ if $b \notin J$. Since $I \subseteq J \subseteq I \sqcup B$, the hyperplane $H = \{x \in \mathbb{R}^d : \langle N, (x, 1)\rangle = 0\}$ satisfies the second part of the proposition. □

Alternative (i) of Proposition 5.3.5 corresponds exactly to (a generalization of) essential and non-essential faces between Delaunay configurations that are described in [4, Proposition 2.1]. However, in Proposition 5.3.5 above, the underlying point set A is finite, leading to the additional case (ii). Notice that the points are not required to be in general position, and can even be repeated multiple times in A.

Armed with this result, we are ready to establish the polynomial reproduction property for spline functions associated to $\mathcal{P}$. The proof is similar to that of [4, Theorem 4.1]; nonetheless,

we give here the full derivation in order to point out the contribution of boundary facets. We start by proving the case $k = 0$.

Proposition 5.3.6. *Let* $\mathcal{P}^{(0)} \coloneqq \{\Pi_{\varnothing,B} \in \mathcal{P}\}$. *Then the set of simplices* $\mathcal{T}^{(0)} = \{\Delta_B \coloneqq \operatorname{conv}(\{a_b\}_{b\in B}) : \Pi_{\varnothing,B} \in \mathcal{P}^{(0)}\}$ *triangulates* $\operatorname{conv}(A)$.

Proof. Define the hyperplane $H_1 \coloneqq \{x \in \mathbb{R}^{d+1} : x_{d+1} = 1\}$ and let π be its canonical identification with $\mathbb{R}^d$ using the first d coordinates. It follows from (5.4) that $\pi(Z(V) \cap H_1) = \operatorname{conv}(A)$, since this set corresponds exactly to all the convex combinations of points in A. Similarly, for any tile $\Pi_{I,B} \in \mathcal{P}$, (5.5) implies that $\pi(\Pi_{I,B} \cap H_1)$ is empty if $|I| > 1$, equal to the single point a_i if $I = \{i\}$, or equal to the simplex $\operatorname{conv}(\{a_b\}_{b\in B})$ if $I = \varnothing$. The proposition then follows from the fact that the tiles $\Pi_{I,B}$ of $\mathcal{P}$ form a subdivision of $Z(V)$. □

The indicator functions of simplices in $\mathcal{T}^{(0)}$ correspond exactly to degree-zero splines via (5.2a). Proposition 5.3.6 then provides the root of the recurrence in the following proof.

Proof of Theorem 5.3.3. Similarly to the proof of [4, Theorem 4.1], we simply have to prove that, for $x \in \operatorname{conv}_k(A)$, the expression

$$\sum_{\Pi_{I,B}\in\mathcal{P}^{(k)}} Q((a_i)_{i\in I}) \operatorname{vol}^{d+1}(\Pi_{I,B}) M(x \mid \Pi_{I,B}) \tag{5.13}$$

can be rewritten in terms of the tiles in $\mathcal{P}^{(k-1)}$ as

$$\frac{k+d}{k} \sum_{\Pi_{I',B'}\in\mathcal{P}^{(k-1)}} Q((a_i)_{i\in I'}, x) \operatorname{vol}^{d+1}(\Pi_{I',B'}) M(x \mid \Pi_{I',B'}). \tag{5.14}$$

In fact, iterating until $k = 0$ directly leads to the expression

$$\binom{k+d}{k} \sum_{\Pi_{\varnothing,B}\in\mathcal{P}^{(0)}} Q(x, \dots, x) \operatorname{vol}^{d+1}(\Pi_{\varnothing,B'}) M(x \mid \Pi_{\varnothing,B'}),$$

which is simply equal to $(k+d)!/k!\, q(x)$ thanks to (5.2a), the definition of polar form (Definition 5.3.1), and the fact that the simplices defined by splines in $\mathcal{P}^{(0)}$ triangulate $\operatorname{conv}(A)$ (Proposition 5.3.6).

In order to prove that (5.13) is equal to (5.14), similarly to [4], we first apply the spline recurrence formula (5.2b) to (5.13), obtaining

$$\frac{k+d}{k} \sum_{\Pi_{I,B}\in\mathcal{P}^{(k)}} Q((a_i)_{i\in I}) \sum_{b\in B} \det({}^{x}_{b}B) M(x \mid \Pi_{I,B\setminus\{b\}}), \tag{5.15}$$

since $\operatorname{vol}^{d+1}(\Pi_{I,B}) = \det(B)$. We can associate every term in (5.15) with a facet $\Pi_{I,B\setminus\{b\}}$ of $\mathcal{P}$. Following Proposition 5.3.5, there are three possibilities:

(i) The facet is shared with exactly one other tile $\Pi_{I',B'} \in \mathcal{P}^{(k)}$, with $I' = I$, $B' \setminus \{b'\} = B \setminus \{b\} = B \cap B'$ for some $b' \in B'$, and with a_b and $a_{b'}$ lying on opposite sides of

$H \coloneqq \mathrm{aff}(\{v_i\}_{i\in B\cap B'})$. Therefore $\det(\,{}^{x}_{b'}B) = -\det(\,{}^{x}_{b'}B')$, and the two corresponding terms in the sum cancel each other;

(ii) The facet is shared with exactly one other tile $\Pi_{I',B'} \in \mathcal{P}^{(k-1)}$, with $I'\sqcup\{b'\} = I$, $B'\setminus\{b'\} = B \setminus \{b\} = B \cap B'$ for some $b' \in B'$, and with a_b and $a_{b'}$ lying on the same side of $H \coloneqq \mathrm{aff}(\{a_i\}_{i\in B\cap B'})$. After noticing that $I \sqcup B \setminus \{b\} = I' \sqcup B'$, the corresponding term in (5.15) can be rewritten as

$$\frac{k+d}{k} Q((a_i)_{i\in I'\sqcup\{b'\}}) \det(\,{}^{x}_{b'}B') M(x \mid (a_i)_{i\in I'\sqcup B'}). \tag{5.16}$$

(iii) The facet lies on the boundary of $Z(V)$. In this case the hyperplane $H \coloneqq \mathrm{aff}(\{a_i\}_{i\in B\setminus\{b\}})$ contains all the points $\{a_i\}_{i\in I\sqcup B\setminus\{b\}}$ in its positive closed halfspace, out of which at most $|I| = k$ points are in its positive open halfspace. All other points of A lie in its negative closed halfspace. Consequently, if x is in the interior of $\mathrm{conv}_k(A)$, then necessarily $x \notin \mathrm{conv}(\{a_i\}_{i\in I\sqcup B\setminus\{b\}})$ and therefore

$$M(x \mid \Pi_{I,B\setminus\{b\}}) = M(x \mid (a_i)_{i\in I\sqcup B\setminus\{b\}}) = 0.$$

Focusing now on (5.14), and again similarly to [4], we rewrite x in barycentric coordinates with respect to the simplex $\mathrm{conv}(\{a_{b'}\}_{b'\in B'})$ as

$$x = \sum_{b'\in B'} \frac{\det(\,{}^{x}_{b'}B')}{\det(B')} a_{b'}, \tag{5.17}$$

and since Q is multiaffine and $\mathrm{vol}^{d+1}(\Pi_{I',B'}) = \det(B')$, using (5.17), we can rewrite (5.14) as

$$\frac{k+d}{k} \sum_{\Pi_{I',B'}\in\mathcal{P}^{(k-1)}} M(x \mid \Pi_{I',B'}) \sum_{b'\in B'} Q((a_i)_{i\in I'\sqcup\{b'\}}) \det(\,{}^{x}_{b'}B'). \tag{5.18}$$

Similarly as before, by Proposition 5.3.5, we can associate each term in (5.18) with a facet $\Pi_{I'\sqcup\{b'\},B'\setminus\{b'\}}$ of $\mathcal{P}$. If such a facet is shared with exactly one other tile $\Pi_{I,B} \in \mathcal{P}^{(k-1)}$, then it appears twice in the sum, and the two contributions cancel each other since $I' \sqcup \{b'\} = I \sqcup \{b\}$, $I \sqcup B = I' \sqcup B'$ and a_b, $a_{b'}$ are separated by $H \coloneqq \mathrm{aff}(\{a_i\}_{i\in B\cap B'})$. Terms corresponding to facets on the boundary of $Z(V)$ again do not contribute to the sum, since the corresponding hyperplane $H \coloneqq \mathrm{aff}(\{a_i\}_{i\in B'\setminus\{b'\}})$ separates at most the k points in $I' \sqcup \{b'\}$ from the other $n-k$ points of A, and since $b' \notin I'$, the points $\{a_i\}_{i\in I'\sqcup B'}$ either lie on H or on the positive side of H. Thus, if $x \in \mathrm{conv}_k(A)$, we have once more

$$M(x \mid \Pi_{I',B'}) = M(x \mid (a_i)_{i\in I'\sqcup B'}) = 0.$$

The remaining terms correspond to facets shared with exactly one other tile $\Pi_{I,B} \in \mathcal{P}^{(k)}$, and they are equal to the terms (5.16), completing the proof. □

Two examples of families of spline spaces associated to fine zonotopal tilings are shown in Figure 5.3.

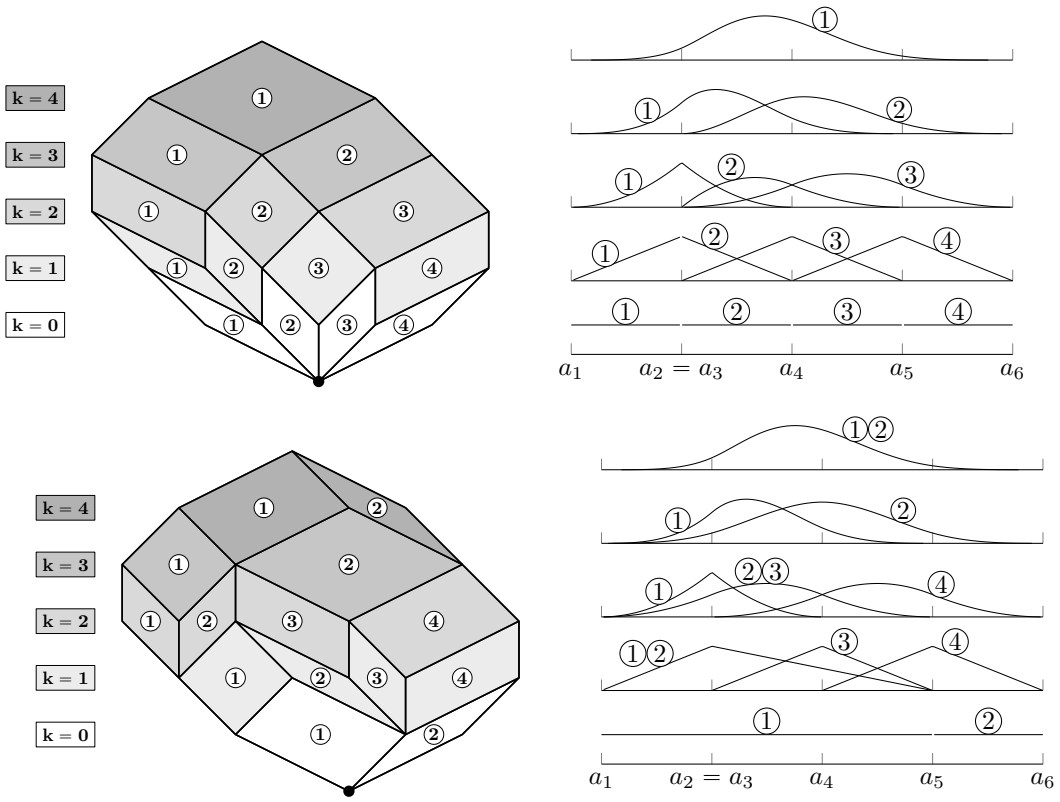


Figure 5.3: Two possible fine zonotopal tilings of $Z(V)$ for the point configuration of Figure 5.2 and their associated spline spaces of degrees $k = 0, \ldots, 4$ for the standard one-dimensional B-spline basis (top) and an alternative tiling (bottom).

5.3.2 Spline space construction

As discussed in Section 5.1, Liu and Snoeyink's algorithm [5, 259] for the iterative construction of generalized Delaunay configurations of A is based on the concept of the order-k *centroid triangulation* [259, 272–274], which is a triangulation of the point set $A^{(k)}$ whose elements are the averages of k-element subsets of A. The order-1 centroid triangulation is simply an (arbitrary) triangulation of A, and an order-k centroid triangulation is obtained from an order-$(k-1)$ centroid triangulation by a subdivision of the polygonal neighborhood of every vertex (its *link region*), with complete freedom in the choice of triangulation for each polygon. Every triangle obtained in this way is then associated to a spline function of degree k.

One major hurdle for the extension of this algorithm to dimensions $d > 2$ lies in the existence of non-convex regions that do not admit any triangulation without introducing new vertices. If such a region is encountered, the algorithm cannot continue, and there is no known condition under which the link regions are all guaranteed to be triangulable. Moreover, the case of affinely dependent and/or repeated points is excluded from the proofs and treated with symbolic perturbation, which creates ambiguous cases and does not allow to extend the proofs of convergence easily. This problem becomes even harder to address as the number of space dimensions grows.

Given a fine zonotopal tiling $\mathcal{P}$ of $Z(V)$, we prove in this section that there exists a construction algorithm similar to Liu and Snoeyink's, with a suitable choice of triangulations, that is able to iteratively construct $\mathcal{P}$. This result rests on a natural definition of the *link region* $\mathcal{R}(I)$ associated to each subset $I \subset [n]$ (Definition 5.3.7), which generalizes naturally Liu and Snoeyink's notion of vertex link.

5.3.3 Relationship with centroid triangulations

Letting r be a natural number and denoting by H_r the hyperplane $H_r \coloneqq \{x \in \mathbb{R}^{d+1} : x_{d+1} = r\}$, the intersection

$$Q^{(r)} \coloneqq Z(V) \cap H_r$$

corresponds to the set $Q^{(r)} \coloneqq \{\sum_{i=1}^n \alpha_i v_i : 0 \leq \alpha_i \leq 1, \sum_{i=1}^n \alpha_i = r\}$, which is just the convex hull of the points $V^{(r)} \coloneqq \{\sum_{b\in B}(a_b, 1),\, B \subseteq [n],\, |B| = r\}$. The region $Q^{(r)}$ is also known as (a multiple of) the *r-set polytope* of A [275, 276]. Just as vectors in V are in bijection with points of A, the set of vectors $V^{(r)}$ can be recast as the set $A^{(r)}$ of all possible averages of r points in A. The intersection $\mathcal{P} \cap H_r$ of a zonotopal tiling of $Z(V)$ with H_r then produces a subdivision of $V^{(r)}$ [270, 277] and therefore of $A^{(r)}$, which corresponds to a *centroid subdivision* in the sense of [259, 272–274].

Recall that the standard hypersimplex $\Delta_{m,n}$ is defined as the convex hull of the points $(x_1, \ldots, x_m) \in \mathbb{R}^m$ such that $0 \leq x_i \leq 1$ and $x_1 + \ldots + x_m = n$. Then, according to (5.5), the intersection of a tile $\Pi_{I,B}$, $|I| = k$ with the hyperplane H_r is an affine transformation of the hypersimplex $\Delta_{d+1,r-k}$, which has a positive dimension if and only if $k < r < k+d+1$. Translated in the language of spline spaces, this means that the cells in the r-th centroid subdivision induced by $\mathcal{P}$ are slices of tiles associated via (5.9) to the basis splines

$$\mathcal{SP}^{(r)} \coloneqq \{M(\,\cdot \mid \Pi_{I,B}),\, r - d - 1 < k \coloneqq |I| < r\}.$$

For $d = 2$, only two types of cells appear in each r-th centroid triangulation for $r > 1$, corresponding to splines of degree $k = r - 1$ and $k = r - 2$. The corresponding hypersimplices $\Delta_{3,1}$ and $\Delta_{3,2}$ are just triangles, and therefore the subdivision is a so-called bi-colored triangulation. This fact is widely known in the context of centroid triangulations [259, 260, 272–274], where the corresponding triangles are called type-I and type-II triangles, respectively. In dimension $d > 2$, the induced subdivision is no longer a triangulation, and the splines of all orders $r - d + 1 \leq k \leq r - 2$ appear in the r-th centroid subdivision as hypersimplices, e.g., octahedra for $d = 3$, $k = r - 2$.

5.3.4 Link regions

We define the *link region* of a subset $Q \subseteq [n]$ as follows:

Definition 5.3.7. *Given a fine zonotopal tiling $\mathcal{P}$ of $Z(V)$ and a subset $Q \subseteq [n]$, $|Q| = k$, the regions $E^{(r)}(Q)$, $r \geq 0$, are defined as the union of simplices*

$$E^{(r)}(Q) \coloneqq \bigcup_{\Pi_{I,B} \in \mathcal{E}^{(r)}(Q)} \operatorname{conv}(\{a_b\}_{b \in B}), \tag{5.19}$$

with

$$\mathcal{E}^{(r)}(Q) \coloneqq \left\{ \Pi_{I,B} \in \mathcal{P}^{(r)} : B \cap Q = \varnothing,\, I \subseteq Q \right\}. \tag{5.20}$$

The link region $\mathcal{R}(Q)$ *of Q is defined as $\mathcal{R}(Q) \coloneqq E^{(k)}(Q)$.*

An example of link region, and its relation to the regions (5.19), is shown in Figure 5.4. Notice that $\mathcal{E}^{(k)}(Q) = \{\Pi_{I,B} \in \mathcal{P} : I = Q\}$ and that $\mathcal{E}^{(r)}(Q) = \varnothing$ for $r > k$. It can be easily checked, though we will not do it explicitly here, that in two dimensions the above defined link region coincides with the interior of a vertex link as used in [5, 6, 259]. However, Definition 5.3.7 is more straightforward, more general, and can be applied to all point configurations in any dimension, allowing to easily prove some important properties, as we do presently.

Proposition 5.3.8. *For any subset $Q \subseteq [n]$, $|Q| = k$, define*

$$\operatorname{conv}_Q(A) \coloneqq \operatorname{conv}(\{a_i\}_{i \notin Q})$$

and let $r \geq 0$. Then, the following holds:

(i) The set of simplices $\mathcal{T}^{(r)}(Q) \coloneqq \{\operatorname{conv}(\{a_b\}_{b \in B}) : \Pi_{I,B} \in \mathcal{E}^{(r)}(Q)\}$ forms a triangulation of $E^{(r)}(Q)$;

(ii) The regions $E^{(r)}(Q)$ form a subdivision of $\operatorname{conv}_Q(A)$;

(iii) The union of all simplices $\bigcup_{r \geq 0} \mathcal{T}^{(r)}(Q)$ triangulates $\operatorname{conv}_Q(A)$;

(iv) The simplices $\mathcal{T}^{(k)}(Q)$ triangulate the link region $\mathcal{R}(Q)$.

Proof. Obviously, (i) implies (iv) via Definition 5.3.7. Notice also that (iii) implies both (ii) and (i), since it is clear from (5.20) that $\mathcal{E}^{(r)}(Q) \cap \mathcal{E}^{(s)}(Q) = \varnothing$ if $r \neq s$. Therefore, the triangulation of $\operatorname{conv}_Q(A)$ decomposes into disjoint triangulations of the subregions $E^{(r)}(Q)$, $r = 1, \ldots, k$.

Let now $\mathcal{P}(Q)$ be the induced tiling of $Z(V \setminus \{v_q\}_{q\in Q})$ via (5.7). Comparing (5.20) with (5.7) shows that the tiles $\{\Pi_{I,B} \in \bigsqcup_{r\geq 0} \mathcal{E}^{(r)}(Q)\}$ are in bijection with the tiles $\{\Pi_{\varnothing,B} \in \mathcal{P}(Q)\} =: \mathcal{P}^{(0)}(Q)$. Therefore, by Proposition 5.3.6, the simplices $\{\mathrm{conv}(\{a_b\}_{b\in B}) : \Pi_{\varnothing,B} \in \mathcal{P}^{(0)}(Q)\}$ form a triangulation of $\mathrm{conv}_Q(A)$, proving (iii). □

Based on these facts, we can replace Definition 5.3.7 of the link region of Q, $|Q| = k$, with

$$\mathcal{R}(Q) \coloneqq \mathrm{conv}_Q(A) \setminus \left(\bigcup_{r=0}^{k-1} E^{(r)}(Q) \right), \tag{5.21}$$

which is preferred from an algorithmic standpoint because it expresses $\mathcal{R}(Q)$ only in terms of the tiles $\Pi_{I,B} \in \mathcal{P}_r$ with $r < k$. Given that the simplex $\mathrm{conv}(\{a_b\}_{b\in B})$ is non-degenerate for any tile $\Pi_{I,B}$, Proposition 5.3.8 implies that the region $\mathcal{R}(Q) \coloneqq E^{(k)}(Q)$ is empty if and only if its triangulation does not contain any simplex, i.e., if and only if $\mathcal{E}^{(k)}(Q)$ is empty. We have therefore the following corollary.

Corollary 5.3.9. *$\mathcal{R}(Q)$ is nonempty if and only if there is a tile $\Pi_{I,B} \in \mathcal{P}$ with $I = Q$.*

Proposition 5.3.8 and Corollary 5.3.9 together imply that any fine zonotopal tiling $\mathcal{P}$ of $Z(V)$, and therefore the associated family of spline spaces, can be obtained iteratively by triangulating the link region associated to each set I for every tile $\Pi_{I,B}$ through some choice of triangulation, similarly to Liu and Snoeyink's algorithm in two dimensions. This statement is made precise in the following theorem. In its proof, we make use of Stiemke's Theorem [278], a variation of Farkas' Lemma stating that, given any set $\{x_1, \ldots, x_m\}$ of m vectors in $\mathbb{R}^n$, exactly one of the following alternatives is true: either there exist $\alpha_1, \ldots, \alpha_m > 0$ such that $\sum_{i=1}^m \alpha_i x_i = 0$, or there exists $y \in \mathbb{R}^n$ such that $\langle y, x_i\rangle \leq 0$, for $i = 1, \ldots, m$, and $\langle y, x_i\rangle \neq 0$ for at least one index.

Theorem 5.3.10. *There exists a choice of triangulations $\mathcal{T}_I$ such that any zonotopal tiling $\mathcal{P}$ of $Z(V)$ (and its associated spline spaces at all orders $0 \leq k \leq n-d-1$) can be iteratively constructed using the following procedure:*

(i) Let $\mathcal{I}^{(0)} = \{\varnothing\}$;

(ii) For every $0 \leq k \leq n-d-1$ and for every $I \in \mathcal{I}^{(k)}$, let $\mathcal{R}(I)$ be the link region computed via (5.21), and let $\mathcal{T}_I$ be its triangulation. Denoting the simplex $\Delta_B \coloneqq \mathrm{conv}(\{a_b\}_{b\in B})$, the subset of tiles $\mathcal{P}^{(k)} \coloneqq \{\Pi_{I,B} \in \mathcal{P} : |I| = k\}$ is given by

$$\mathcal{P}^{(k)} = \{\Pi_{I,B} : I \in \mathcal{I}^{(k)},\, \Delta_B \in \mathcal{T}_I\};$$

(iii) Let

$$\mathcal{I}^{(k+1)} = \{I \sqcup \{b\} : \Pi_{I,B} \in \mathcal{P}^{(k)},\, b \in B,\, \mathcal{R}(I \sqcup \{b\}) \neq \varnothing\}; \tag{5.22}$$

(iv) Repeat (ii) and (iii) until $k = n-d-1$, $\mathcal{I}^{(k+1)} = \varnothing$. Then $\mathcal{P} = \bigsqcup_{k=0}^{n-d-1} \mathcal{P}^{(k)}$.

Proof. Let $\mathcal{P}$ be a fine zonotopal tiling of $Z(V)$. Item (iv) of Proposition 5.3.8 directly states that the tiles $\Pi_{I,B} \in \mathcal{P}^{(k)}$ (i.e., splines of degree k) are in bijection with the simplices $\mathrm{conv}(\{a_b\}_{b\in B})$ of a triangulation of the link region $\mathcal{R}(I)$. Furthermore, due to Corollary 5.3.9, all the tiles $\Pi_{I,B} \in \mathcal{P}^{(k)}$ are associated with a nonempty link region, which is always triangulable since Proposition 5.3.8 exhibits one such triangulation. The only thing left to determine is the set $\{I : \Pi_{I,B} \in \mathcal{P}\}$.

Notice that $I \in \mathcal{I}^{(0)}$ implies $I = \varnothing$, and by (5.21), $\mathcal{R}(\varnothing) = \mathrm{conv}(A)$. Therefore, the tiles $\Pi_{\varnothing,B}$ (i.e., splines of degree 0) are in bijection with the simplices of a triangulation of $\mathrm{conv}(A)$, in accordance with Proposition 5.3.6.

Assume now that we have obtained all the tiles $\Pi_{I,B} \in \mathcal{P}^{(r)}$ for $r = 0, \dots, k$, and we want to determine the set $\mathcal{I}^{(k+1)} \coloneqq \{I : \Pi_{I,B} \in \mathcal{P}^{(k+1)}\}$.

Let $Q \subset [n]$, $|Q| = k+1$, be a set of indices such that $\mathcal{R}(Q) \neq \varnothing$, let $\{\Delta_f, f = 1, \dots, F\}$ be the F boundary facets of $\mathcal{R}(Q)$, and for every $f = 1, \dots, F$, let Π_{Q,B_f} and $b_f \in B_f$ be a tile in $\mathcal{P}^{(k+1)}$ such that $\Delta_f = \mathrm{conv}(\{a_i\}_{i\in B_f\setminus\{b_f\}})$. By Proposition 5.3.6, this tile is unique. Suppose that all the facets $\{\Pi_{Q,B_f\setminus\{b_f\}}, f = 1, \dots, F\}$ lie on the boundary of $Z(V)$, let $|\Delta_f|$ be the volume of Δ_f and let $N_f \in \mathbb{R}^d$ be its normalized normal vector. Without loss of generality, we can choose either all inward or all outward normal vectors so that $\sum_{f=1}^{F} |\Delta_f| \left\langle N_f, a_{b_f} \right\rangle \leq 0$. Since $\mathcal{R}(Q)$ is a nonempty, bounded polytopal region, we know that $\sum_{f=1}^{F} |\Delta_f| N_f = 0$, and we can therefore write the following linear dependency with positive coefficients,

$$\sum_{f=1}^{F} |\Delta_f| \left(N_f, -\left\langle N_f, a_{b_f} \right\rangle\right) + (0, \sum_{f=1}^{F} |\Delta_f| \left\langle N_f, a_{b_f} \right\rangle) = 0. \tag{5.23}$$

Fix a point a_q with $q \in Q$. If, for all $f = 1, \dots, F$, a_q were separated from a_{b_f} by the hyperplane $\mathrm{conv}(\{a_i\}_{i\in B_f\setminus\{b_f\}})$, then we would have

$$\begin{aligned} (a_q, 1) \cdot \left(N_f, -\left\langle N_f, a_{b_f} \right\rangle\right) &= \left\langle N_f, a_q - a_{b_f} \right\rangle < 0, \\ (a_q, 1) \cdot (0, \sum_{f=1}^{F} |\Delta_f| \left\langle N_f, a_{b_f} \right\rangle) &= \sum_{f=1}^{F} |\Delta_f| \left\langle N_f, a_{b_f} \right\rangle \leq 0. \end{aligned} \tag{5.24}$$

By Stiemke's Lemma, (5.23) and (5.24) cannot both be true. Therefore, there must be an index f such that the facet $\Pi_{Q,B_f\setminus\{b_f\}}$ does not lie on the boundary of $Z(V)$. Observe also that $\Pi_{Q,B_f\setminus\{b_f\}}$ cannot be shared with another tile $\Pi_{I',B'} \in \mathcal{P}^{(k+1)}$, since otherwise $I' = Q$ and Δ_f would not be a boundary facet of $\mathcal{R}(Q)$. Therefore, by Proposition 5.3.5, there must be a tile $\Pi_{I,B} \in \mathcal{P}^{(k)}$ with $B_f \setminus \{b_f\} = B \setminus \{b\} = B_f \cap B$ and $Q = I \sqcup \{b\}$ for some $b \in B$. We conclude that

$$\mathcal{I}^{(k+1)} \subseteq \{I \sqcup \{b\} : \Pi_{I,B} \in \mathcal{P}^{(k)},\, b \in B\}.$$

After filtering out the sets $\{I \sqcup \{b\} : \mathcal{R}(I \sqcup \{b\}) = \varnothing\}$, we are left exactly with (5.22).

Finally, when $|Q| = n - d$, the set $\mathrm{conv}_Q(A)$ only contains d points, and therefore the link region $\mathcal{R}(Q)$ has an empty interior. Therefore, $\mathcal{I}^{(n-d)} = \varnothing$, and the process stops. □

This theorem states essentially that any fine zonotopal tiling of $Z(V)$ can be built using a version of Liu and Snoeyink's algorithm, provided that we know in advance which triangulation

needs to be applied to each subset $\{I : \Pi_{I,B} \in \mathcal{P}\}$. In other words, it proves that their algorithm is a universal way of constructing fine zonotopal tilings of $Z(V)$ and their associated spline spaces. However, this result stops short of providing a fully-formed construction algorithm, as it does not guarantee that any given choice of triangulations leads to a valid construction, only that such a choice exists. In the next section, we show that *regular* fine zonotopal tilings can be obtained by choosing a weighted Delaunay triangulation at each step, providing a sufficient condition on the triangulations that guarantees the convergence of the construction process.

Finally, we give a couple of interesting results regarding the combinatorial structure of spline spaces built by Theorem 5.3.10. First, as a direct consequence of Theorem 5.2.1, we obtain the following simple characterization of the total number of spline functions:

Corollary 5.3.11. *The total number of spline functions built by the process described in Theorem 5.3.10 on a point set A with $|A| = n$, summed over all orders $k = 0, \dots, n-d-1$, is always equal to the number of maximal affinely independent subsets of A.*

Next, we provide a characterization of the set of simplices

$$\mathcal{T}^{(k)} \coloneqq \{\mathrm{conv}(\{a_b\}_{b\in B}) : \Pi_{I,B} \in \mathcal{P}^{(k)}\}.$$

The intersection of these simplices defines the zones where all the spline functions are pure polynomials, and their boundaries define the zones of reduced regularity of spline functions, i.e., knots in $d = 1$, knot lines in $d = 2$ and more generally knot hypersurfaces in $d > 2$.

Proposition 5.3.12. *For all $0 \le k \le n-d-1$, the simplices in $\mathcal{T}^{(k)}$ cover $\binom{k+d}{d}$ times the set $\mathrm{conv}_k(A)$.*

Proof. By induction over k. The simplices in $\mathcal{T}^{(0)}$ form a triangulation of $\mathrm{conv}(A)$ by Proposition 5.3.6, and therefore cover it exactly once. Assume now that the proposition is true for every $r < k$. By Property (iii) of Proposition 5.3.8, for any subset $Q \subset [n]$ with $|Q| = k$, the simplices $\{\mathrm{conv}(\{a_b\}_{b\in B}) : \Pi_{I,B} \in \mathcal{E}^{(r)}(Q),\, r \le k\}$ triangulate $\mathrm{conv}_Q(A)$, i.e.,

$$\sum_{r=0}^{k} \sum_{\Pi_{I,B}\in\mathcal{E}^{(r)}(Q)} \mathbf{1}_B = \mathbf{1}_{\mathrm{conv}_Q(A)},$$

where $\mathbf{1}_{\mathrm{conv}_Q(A)} : \mathbb{R}^d \mapsto \mathbb{R}$ is the indicator function of the set $\mathrm{conv}_Q(A) \subset \mathbb{R}^d$ and $\mathbf{1}_B$ is the indicator function of $\mathrm{conv}(\{a_b\}_{b\in B})$. We sum this expression over all subsets $Q \subset [n]$, $|Q| = k$. Each tile $\Pi_{I,B} \in \mathcal{P}^{(r)}$ appears in the sum whenever $I \sqcup J = Q$ for some subset $J \subset [n]$, $|J| = k-r$ with $J \cap B = \varnothing$. Therefore, the occurrences of a tile of $\mathcal{P}^{(r)}$ in the sum correspond to the possible choices of $|Q \setminus I| = k - r$ indices among the $\left|\overline{I \sqcup B}\right| = n - r - d - 1$ which are available. We obtain

$$\sum_{\Pi_{I,B}\in\mathcal{P}^{(k)}} \mathbf{1}_B + \sum_{r=0}^{k-1} \binom{n-r-d-1}{k-r} \sum_{\Pi_{I,B}\in\mathcal{P}^{(r)}} \mathbf{1}_B = \sum_{Q\subset[n],|Q|=k} \mathbf{1}_{\mathrm{conv}_Q(A)}. \tag{5.25}$$

By induction, the simplices derived from the tiles in $\mathcal{P}^{(r)}$ cover the region $\mathrm{conv}_r(A) \supseteq \mathrm{conv}_k(A)$ exactly $\binom{r+d}{d}$ times, and the sum on the right covers $\mathrm{conv}_k(A)$ exactly $\binom{n}{k}$ times. Using multiset

notation and the Vandermonde identity, we can derive

$$\sum_{r=0}^{k} \binom{n-r-d-1}{k-r}\binom{r+d}{r} = \sum_{r=0}^{k} \left(\!\!\binom{n-k-d}{k-r}\!\!\right)\left(\!\!\binom{d+1}{r}\!\!\right), \tag{5.26}$$
$$= \left(\!\!\binom{n-k+1}{k}\!\!\right) = \binom{n}{k}.$$

Separating the term with $r = k$ in the first sum in (5.26), we conclude that the first term in (5.25), i.e., the set of all simplices in $\mathcal{T}^{(k)}$, must cover the region $\mathrm{conv}_k(A)$ exactly $\binom{k+d}{d}$ times. □

Notice that in general it is not possible to extract a collection of $\binom{k+d}{d}$ independent triangulations from the set $\mathcal{T}^{(k)}$, as these simplices form in general a *branched cover* of $\mathrm{conv}(A)$. In practice, $\mathcal{T}^{(k)}$ forms a complex web of overlapping simplices that contains many complex intersections, see, e.g., Figure 5.4.

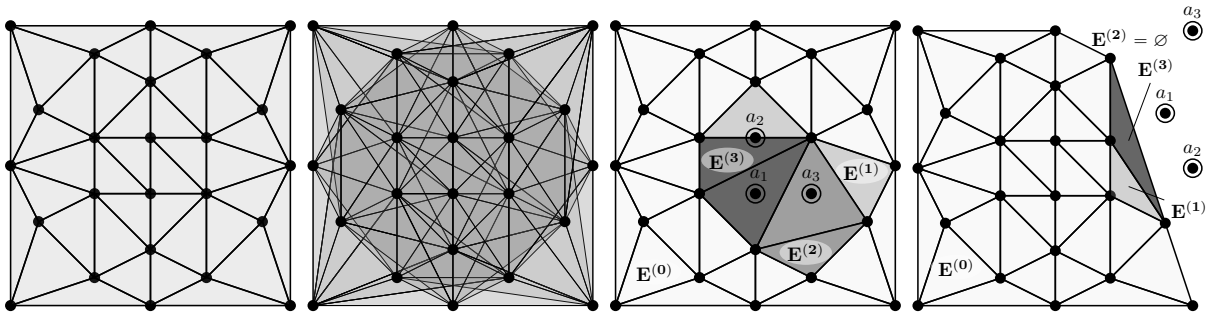


Figure 5.4: For a point configuration $A \subset \mathbb{R}^2$ with collinear points, the sets $\mathcal{T}^{(k)}$ for $k = 0$ (left) and $k = 2$ (center left), with the shading indicating the number of simplices covering each point, and the regions $E^{(r)}(Q)$ of (5.19) for two possible choices of $Q \coloneqq (a_1, a_2, a_3)$ (right, center right).

5.4 Spline spaces from regular fine zonotopal tilings

We specialize the results of the previous section to spline spaces derived from *regular* fine zonotopal tilings. Given a polytope $P \subset \mathbb{R}^{d+2}$, we define its *upper convex hull* as the set of faces of P whose outward normal vector has a positive $(d+1)$-th component.

Definition 5.4.1. *A zonotopal tiling $\mathcal{P}$ of $Z(V) \subset \mathbb{R}^{d+1}$ is* regular *if its tiles are precisely the projections along the $(d+1)$-th coordinate of the faces in the upper convex hull of another zonotope $\hat{Z} \subset \mathbb{R}^{d+2}$.*

We show that this special case corresponds exactly to simplex splines associated to weighted Delaunay configurations. The special properties of these tilings then allow us to derive a set of practical algorithms for the construction of the spline spaces and the determination and evaluation of all spline functions that are supported at a given point $x \in \mathbb{R}^d$.

5.4.1 Delaunay triangulations and regular zonotopal tilings

Let $h : A \mapsto \mathbb{R}$ be a *height function* over A. Let $\mathcal{T}$ be a set of simplices that triangulate $\text{conv}(A)$ with vertices in A. For every subset $B \subseteq [n]$, $|B| = d+1$ such that there is a simplex $\Delta \coloneqq \text{conv}(\{a_b\}_{b\in B}) \in \mathcal{T}$, let us order B such that $\det((a_b, 1)_{b\in B}) > 0$. If, for every $i \in A \setminus B$,

$$\det((a_b, h(a_b), 1)_{b\in B}, (a_i, h(a_i), 1)) < 0, \tag{5.27}$$

then the triangulation $\mathcal{T}$ is called a *weighted Delaunay triangulation* with height function h. If the points of A are in general position, plugging $h(a) = |a|^2$ in (5.27) yields the usual Delaunay triangulation, see, e.g., [213].

In order for the Delaunay triangulation to exist and to be unique, a bit of care is required when choosing the height function h.

Definition 5.4.2. *A height function h is* generic *if, given the lifted point cloud*

$$\hat{A} \coloneqq \{(a, h(a)),\, a \in A\} \subset \mathbb{R}^{d+1},$$

the only affinely dependent subsets of $d+2$ points in $\hat{A}$ lie on a vertical plane, i.e., a plane whose normal $N \in \mathbb{R}^{d+1}$ satisfies $N_{d+1} = 0$.

Notice that affinely dependent subsets are indeed allowed on vertical planes, and thus the points in A can be repeated or affinely dependent. If h is generic, then the determinant in (5.27) is always nonzero, and the weighted Delaunay triangulation is unique. Hereafter, we will only consider generic height functions. We can now use (5.27) to specialize Theorem 5.3.10 to weighted Delaunay triangulations.

Theorem 5.4.3. *Let h be a generic height function on A, and, for every set $Q \subseteq [n]$, let $\mathcal{T}_Q(h)$ be the weighted Delaunay triangulation of $\mathcal{R}(Q)$ with height function h. Then the procedure outlined in Theorem 5.3.10 with the choice $\mathcal{T}_I = \mathcal{T}_I(h)$ always produces a regular fine zonotopal tiling $\mathcal{P}(h)$.*

Proof. It is easy to prove this theorem using the lifting property (5.27). See also [279, 280] and especially [281] for similar constructions and an interesting generalization.

Let $\hat{A} = \{\hat{a}_i \coloneqq (a_i, h(a_i)),\, i = 1, \dots, n\} \subset \mathbb{R}^{d+1}$ be the point cloud lifted by h, $\hat{V} \coloneqq \{(a_i, h(a_i), 1) : i = 1, \dots, n\}$ be the associated vector configuration, and $Z(\hat{V})$ be the zonotope built on $\hat{V}$. Denoting by $\pi : \mathbb{R}^{d+2} \mapsto \mathbb{R}^{d+1}$ the projection that removes the $(d+1)$-th coordinate, it is easy to check that $\pi(Z(\hat{V})) = Z(V)$. We define $\mathcal{P}(h)$ as the regular zonotopal tiling

$$\mathcal{P}(h) \coloneqq \{\pi(\hat{\Pi}_{I,B}) : \hat{\Pi}_{I,B} \text{ is in the upper convex hull of } Z(\hat{V})\}. \tag{5.28}$$

The fact that (5.28) is indeed a regular zonotopal tiling of $Z(V)$ was proven, e.g., in [282, Lemma 2.2]. Since $\hat{\Pi}_{I,B}$ is a boundary facet of $Z(\hat{V})$, we can follow the same reasoning as in the proof of item (ii) of Proposition 5.3.5. After selecting the face normal N_B of $\hat{\Pi}_{I,B}$ with $(N_B)_{d+1} > 0$, given that h is generic and the face is not vertical, we conclude that the determinant

$$\det((a_b, h(a_b), 1)_{b\in B}, (a_i, h(a_i), 1)) \tag{5.29}$$

is positive for all $i \in I$ and negative for all $i \in \overline{I \sqcup B}$, while the condition $(N_B)_{d+1} > 0$ translates to $\det((a_b, 1)_{b\in B}) > 0$. Since only the points $\{a_i\}_{i\in\overline{I\sqcup B}}$ appear in the link region $\mathcal{R}(I)$, the weighted Delaunay condition (5.27) is satisfied for all the points in $\mathcal{R}(I)$. □

Theorems 5.3.10 and 5.4.3 together give a practical construction algorithm for all regular fine zonotopal tilings of $Z(V)$, and therefore for their associated spline spaces. Restricting the construction to the special case $d = 2$ and to points in generic position, this process reduces to a version of Liu and Snoeyink's construction algorithm [5, 6, 259].

5.4.2 Splines supported at a point

In this subsection we show that, in the case of spline spaces associated to regular fine zonotopal tilings, there exists an efficient process to determine all the spline functions up to a given degree $k \geq 0$ that are supported at a given point $x \in \mathbb{R}^d$. This is equivalent, by (5.9), to finding all the tiles $\Pi_{I,B} \in \mathcal{P}(h)$ such that $x \in \mathrm{conv}(\{a_i\}_{i\in I\sqcup B})$. In this case, by extension, we say that the tile $\Pi_{I,B}$ is supported at x.

For spline functions of degree 0, the task is particularly simple. In fact, since the simplices $\mathcal{T}^{(0)}$ triangulate $\mathrm{conv}(A)$ (Proposition 5.3.6), whenever $x \in \mathrm{conv}(A)$ there is one and only one tile $\Pi_{\varnothing,Z}$ supported at x. Computationally, $\Pi_{\varnothing,Z}$ can be found efficiently via a point location query on a triangulation, for which many efficient algorithms exist, see, e.g., [283, 284]. We prove in the remainder of this section that all the other tiles $\Pi_{I,B}$ (and hence spline functions) supported at x can be found from $\Pi_{\varnothing,Z}$ using a suitable orientation, induced by x, of the dual graph $\mathcal{G}$ of $\mathcal{P}(h)$, i.e., the simple, connected graph having the tiles of $\mathcal{P}(h)$ as vertices and their connecting internal facets as edges.

We assume hereafter that the test point $x \in \mathbb{R}^d$ is *generic*, i.e., it satisfies the following condition:

$$x \notin \mathrm{aff}(\{a_c\}_{c\in C}) \text{ for all internal facets } \Pi_{J,C} \text{ of } \mathcal{P}.$$

This excludes from the possible values of x a zero-measure subset of $\mathbb{R}^d$, and as a consequence, all the following results must be understood to hold almost everywhere. This restriction can be easily lifted using some well-known techniques such as symbolic perturbation. We can define an orientation o_x, depending on x, on the dual graph $\mathcal{G}$ of $\mathcal{P}$ as follows. Let $\Pi_{J,C}$ be a facet shared by two tiles $\Pi_{I,B}$ and $\Pi_{I',B'}$, with normal vector $N_C \in \mathbb{R}^{d+1}$. Then we define the orientation of the corresponding edge in $\mathcal{G}$ as $\Pi_{I,B} \to \Pi_{I',B'}$ if and only if

$$\mathrm{sign}\left(\langle N_C, (x, 1)\rangle\right) = \mathrm{sign}\left(\left\langle N_C, z' - z\right\rangle\right) \tag{5.30}$$

for any $z' \in \Pi_{I',B'}$, $z \in \Pi_{I,B}$. In other words, we pick the direction of N_C that leads to a positive scalar product with $(x, 1)$, and we use it to orient the corresponding edge.

The orientation o_x defined by (5.30) yields a directed graph $(\mathcal{G}, o_x)$. In the case of regular tilings, this graph is acyclic.

Lemma 5.4.4. *Let $\mathcal{P}(h)$ be a regular fine zonotopal tiling of $Z(V)$ with generic height function h. Then the directed graph $(\mathcal{G}, o_x)$ is acyclic for every generic $x \in \mathbb{R}^d$. The same is true for any fine zonotopal tiling $\mathcal{P}$ of $Z(V)$, regular or not, when $d = 1$.*

Proof. Let $\Pi_i \coloneqq \Pi_{I_i,B_i}$, $i = 1, \dots, r$ be a family of r tiles of $\mathcal{P}(h)$ and let $F_i \coloneqq \Pi_{J_i,C_i}$, $i = 1, \dots, r$ be a family of facets such that F_i is shared between the tiles Π_i and Π_{i+1}. Let us assume that the tiles form a cycle in $\mathcal{G}$, i.e., $\Pi_{r+1} = \Pi_1$. For each $1 \leq i \leq r$, let $N_i \coloneqq N_{C_i}$ be a vector normal to the i-th facet and pointing from the tile Π_i to the tile Π_{i+1}.

Since $\mathcal{P}(h)$ is regular, by Theorem 5.4.3, for each tile Π_i there is a vector $y_i \in \mathbb{R}^{d+2}$ with $(y_i)_{d+1} > 0$ such that $\langle y_i, (a_s, h(a_s), 1)\rangle$ is positive if $s \in I_i$, zero if $s \in B_i$, and negative if $s \in \overline{I_i \sqcup B_i}$. Define the point $g_i \in \mathbb{R}^{d+1}$ component-wise as

$$(g_i)_j \coloneqq \frac{(y_i)_j}{(y_i)_{d+1}},\ j = 1, \dots, d,\ (g_i)_{d+1} \coloneqq \frac{(y_i)_{d+2}}{(y_i)_{d+1}}, \tag{5.31}$$

which is possible since $(y_i)_{d+1} > 0$. For all $b \in B_i$, $\langle y_i, (a_b, h(a_b), 1)\rangle = 0$ implies

$$\langle g_i, v_b\rangle = -h(a_b), \tag{5.32}$$

and as a consequence, for all $c \in B_i \cap B_{i+1} = C_i$,

$$\langle g_{i+1} - g_i, v_c\rangle = 0, \tag{5.33}$$

i.e., the vector $(g_{i+1} - g_i)$ is parallel to N_i. Let now $z_i \in \Pi_i$ be the point

$$z_i \coloneqq \sum_{j \in I_i} v_j + \frac{1}{2} \sum_{b \in B_i} v_b, \tag{5.34}$$

and let $b \in B_i$, $b' \in B_{i+1}$ be the two indices such that $B_i \setminus \{b\} = B_{i+1} \setminus \{b'\}$. Let $\sigma_1 = +1$ or -1 if $b \in I'$ or $b \notin I'$, respectively, and similarly $\sigma_2 = +1$ or -1 if $b' \in I$ or $b' \notin I$ respectively. Using (5.29), (5.31) and (5.34), it is easy to check that $\operatorname{sign}(\langle g_i, v_{b'}\rangle + h(a_{b'})) = \sigma_2$, $\operatorname{sign}(\langle g_{i+1}, v_b\rangle + h(a_b)) = \sigma_1$ and $z_{i+1} - z_i = \sigma_1 v_b - \sigma_2 v_{b'}$. Therefore, according to (5.32) and (5.33),

$$\begin{aligned}
\operatorname{sign}(\langle g_{i+1} - g_i, z_{i+1} - z_i\rangle) &= \operatorname{sign}(\langle g_{i+1} - g_i, \sigma_1 v_b - \sigma_2 v_{b'}\rangle), \\
&= \operatorname{sign}(\sigma_1 \langle g_{i+1}, v_b\rangle + \sigma_2 h(a_{b'}) + \sigma_1 h(a_b) + \sigma_2 \langle g_i, v_{b'}\rangle), \\
&= \sigma_1^2 + \sigma_2^2 > 0.
\end{aligned}$$

In other words, $(g_{i+1} - g_i)$ always points in the same direction as N_i, and thus $g_{i+1} - g_i = \mu_i N_i$ for some $\mu_i > 0$. We can therefore write:

$$0 = \sum_{i=1}^{r} (g_{i+1} - g_i) = \sum_{i=1}^{r} \mu_i N_i \text{ with } \mu_1, \dots, \mu_r > 0. \tag{5.35}$$

Taking the scalar product of (5.35) with $(x, 1)$, $x \in \mathbb{R}^d$ shows that, for at least one facet F_i, we must have $\langle N_i, (x, 1)\rangle < 0$ and therefore

$$\operatorname{sign}\left(\langle N_i, (x, 1)\rangle\right) \neq \operatorname{sign}\left(\langle N_i, z_{i+1} - z_i\rangle\right),$$

i.e., (5.30) fails. In other words, this orientation cannot be induced by any generic point $x \in \mathbb{R}^d$.

All orientations $(\mathcal{G}, o_x)$ are therefore acyclic.

In the one-dimensional case, we can obtain the positive linear combination of normals (5.35) without assuming the existence of the vectors y_i. We only give a sketch of the proof. First, there is at least one tile Π_i such that $F_i \neq F_{i+1}$, else the tiles cannot form a loop. Furthermore, since each tile is convex, each angle $N_i \angle N_{i+1}$ can only be strictly less than π, but the total angle along the cycle must be equal to $2k\pi$, $k \in \mathbb{Z} \setminus \{0\}$. These conditions imply that there is a closed path in $\mathbb{R}^2$ whose j-th displacement vector is directed along N_j. Defining g_i as the i-th vertex of the path then yields (5.35).

□

Remark 5.4.5. *The construction used in the proof of Lemma 5.4.4 is similar to the affinization of central hyperplane arrangements, see, e.g., [285, Chapter 7].*

As a directed acyclic graph, $(\mathcal{G}, o_x)$ can be topologically sorted, and the (only) tile $\Pi_{\varnothing,Z}$ supported at x can be used as the root of an oriented path that follows the topological sorting. We prove now that the other tiles $\Pi_{I',B'}$ supported at x are all reachable from $\Pi_{\varnothing,Z}$ using such a path. First, we need a lemma in convex theory, very similar (although not equivalent) to Carathéodory's theorem.

Lemma 5.4.6. *Let $A = (a_1, \ldots, a_n)$ be a configuration of $n > d+1$ points in $\mathbb{R}^d$, and let $B \subset [n]$ be a set of $|B| = d+1$ indices such that the points $(a_i)_{i\in B}$ are affinely independent. Then, for every $x \in \operatorname{conv}(A)$ there exists an index $b \in B$ such that a_b and x are on the same closed halfspace of $\operatorname{aff}(\{a_i\}_{i\in B\setminus\{b\}})$ and $x \in \operatorname{conv}(\{a_i\}_{i\in[n]\setminus\{b\}})$.*

Proof. First, assume that $x \in \operatorname{conv}(\{a_i\}_{i\in B})$. In this case, for all $b \in B$, x is on the same closed halfspace of $\operatorname{aff}(\{a_i\}_{i\in B\setminus\{b\}})$ as a_b. We can then pick any index $c \in [n] \setminus B$, and the (possibly degenerate) simplices $\operatorname{conv}(\{a_i\}_{i\in B\setminus\{b\}\sqcup\{c\}})$ for all $b \in B$ cover the set $\operatorname{conv}(\{a_i\}_{i\in B})$. Thus, for at least one index $b \in B$, $x \in \operatorname{conv}(\{a_i\}_{i\in B\setminus\{b\}\sqcup\{c\}})$, and the lemma is satisfied.

Assume now that $x \notin \operatorname{conv}(\{a_i\}_{i\in B})$. Then, $x \in \operatorname{conv}(A)$ if and only if

$$x = \sum_{i=1}^{n} \mu_i a_i$$

for some real numbers μ_i satisfying $\mu_i \geq 0$ and $\sum_{i=1}^n \mu_i = 1$. Since the points indexed by B are affinely independent, we can also express $x = \sum_{b\in B} \lambda_b a_b$, with $\sum_{b\in B} \lambda_b = 1$. We extend this to a linear combination $x = \sum_{i=1}^n \lambda_i a_i$ by defining $\lambda_i \coloneqq 0$ for $i \notin B$. We have

$$\sum_{i=1}^{n} \mu_i = 1 = \sum_{i=1}^{n} \lambda_i,$$

and therefore $\sum_{i=1}^n (\mu_i - \lambda_i) = 0$. The expression $\mu_i - \lambda_i$ cannot be identically zero for all $i \in [n]$, since otherwise $x \in \operatorname{conv}(\{a_j\}_{j\in B})$, which has been excluded. Thus, there must be at least one $b \in B$ with $\lambda_b > \mu_b \geq 0$. If we pick an index $c \in B$ such that

$$c \in \operatorname*{arg\,min}_{b\in B} \left\{ \alpha_b \coloneqq \frac{\mu_b}{\lambda_b - \mu_b} : \lambda_b > \mu_b \right\},$$

we can write the nonnegative linear combination

$$\sum_{i=1}^{n} [\mu_i - (\lambda_i - \mu_i)\alpha_c]\, a_i = x, \tag{5.36}$$

where clearly $\mu_i - (\lambda_i - \mu_i)\alpha_c \geq 0$ and $\mu_c - (\lambda_c - \mu_c)\alpha_c = 0$. Thus, the point a_c satisfies the lemma, since $\lambda_c > \mu_c \geq 0$ implies that a_c and x are on the same open halfspace of $\operatorname{aff}(\{a_i\}_{i \in B \setminus \{c\}})$, and x can be expressed as the convex combination (5.36) with the point a_c having a zero coefficient. □

We can now prove that there is always a directed path in $(\mathcal{G}, o_x)$ from $\Pi_{\varnothing,Z}$ to any tile $\Pi_{I',B'}$ supported at x.

Proposition 5.4.7. *Let $\mathcal{P}(h)$ be a regular fine zonotopal tiling of $Z(V)$ with generic height function h, let $x \in \operatorname{conv}(A)$ be a generic point, and let $\Pi_{\varnothing,Z}$ be the only tile in $\mathcal{P}^{(0)}(h)$ supported at x. Then for every tile $\Pi_{I',B'} \in \mathcal{P}(h)$ supported at x, there is a directed path in $(\mathcal{G}, o_x)$ from $\Pi_{\varnothing,Z}$ to $\Pi_{I',B'}$ with every tile $\Pi_{I,B}$ in the path satisfying $|I| \leq |I'|$.*

Proof. If $I' = \varnothing$, then necessarily $\Pi_{I',B'} = \Pi_{\varnothing,Z}$, and we are done. Else, we complete the proof by finding another tile $\Pi_{I,B}$ and an oriented edge $\Pi_{I,B} \to \Pi_{I',B'}$ in $(\mathcal{G}, o_x)$ such that $\Pi_{I,B}$ is supported at x and $I \subseteq I'$. The same reasoning can then be applied to $\Pi_{I,B}$ and again repeatedly, yielding an oriented path of tiles supported at x and with non-increasing $|I|$. Since the graph is acyclic (Lemma 5.4.4) and the number of tiles is finite, the process must eventually end with $\Pi_{I,B} = \Pi_{\varnothing,Z}$ as the root of the path.

According to Lemma 5.4.6, and since x is generic, there exists an index $b' \in B'$ such that

$$x \in \operatorname{conv}(\{a_i\}_{i \in I' \sqcup B' \setminus \{b'\}}) \text{ and } a_{b'},\, x \text{ are on the same side of } H_{b'}, \tag{5.37}$$

where $H_{b'} \coloneqq \operatorname{aff}(\{a_i\}_{i \in B' \setminus \{b'\}})$. Necessarily, this means that there is an index $j \in I'$ such that a_j is on the same side of $H_{b'}$ as $a_{b'}$, otherwise $H_{b'}$ would separate x from the convex hull $\operatorname{conv}(\{a_i\}_{i \in I' \sqcup B' \setminus \{b'\}})$ and (5.37) would be false. Proposition 5.3.5 then guarantees that there is a tile $\Pi_{I,B}$, connected to $\Pi_{I',B'}$ with an edge in $\mathcal{G}$, such that $B \setminus \{b\} = B' \setminus \{b'\}$ for some $b \in B$ and either $I' = I$ or $I' = I \sqcup \{b\}$. The point a_b is on the opposite side of $H_{b'}$ as $a_{b'}$ and x in the first case, and on the same side in the second case. It is easy to check, using (5.30) and taking the representative points $z \in \Pi_{I,B}$ and $z' \in \Pi_{I',B'}$ defined as in (5.34), that in both cases the edge associated to the tile $\Pi_{J,C}$ with $J = I'$, $C = B \cap B'$ is oriented from $\Pi_{I,B}$ to $\Pi_{I',B'}$. Furthermore, in both cases, $B \sqcup I \supseteq I' \sqcup B' \setminus \{b'\}$, implying that $\Pi_{I,B}$ is supported at x, and $I \subseteq I'$. This completes the proof. □

Proposition 5.4.7 is important because it shows that every tile $\Pi_{I,B}$ of order k can be connected to $\Pi_{\varnothing,Z}$ in $(\mathcal{G}, o_x)$ using only tiles of order k or less (see, e.g., Figure 5.5). In practical applications, this implies that all the spline functions of degree k supported at any given point can be found efficiently using only the knowledge of spline functions of degree $r \leq k$. Therefore, when constructing a spline space using the process delineated in Theorems 5.3.10 and 5.4.3, the iterations can be safely stopped at the desired degree, without requiring any higher-degree functions.

Furthermore, Theorem 5.4.7 suggests a simple and efficient algorithm to find all the spline functions supported at a point x. The first step, which requires finding the spline of degree $k = 0$ having x in its support, can be efficiently implemented via any search tree constructed on the simplices in $\mathcal{T}^{(0)}$ [283, 284]. Such trees typically have a $O(n \log(n))$ construction complexity and a $O(\log(n))$ query complexity, n being the number of degree-zero splines. After this first step, the complexity is simply linear in the number of spline functions (of all degrees $r \leq k$) which are nonzero on x, and does not depend on the total number of functions in the spline space.

Notice however that there is still a need to check explicitly if every visited spline function is actually supported at x, albeit only for a limited number of functions.

We show an example of the directed graph $(\mathcal{G}, o_x)$ in Figure 5.5.

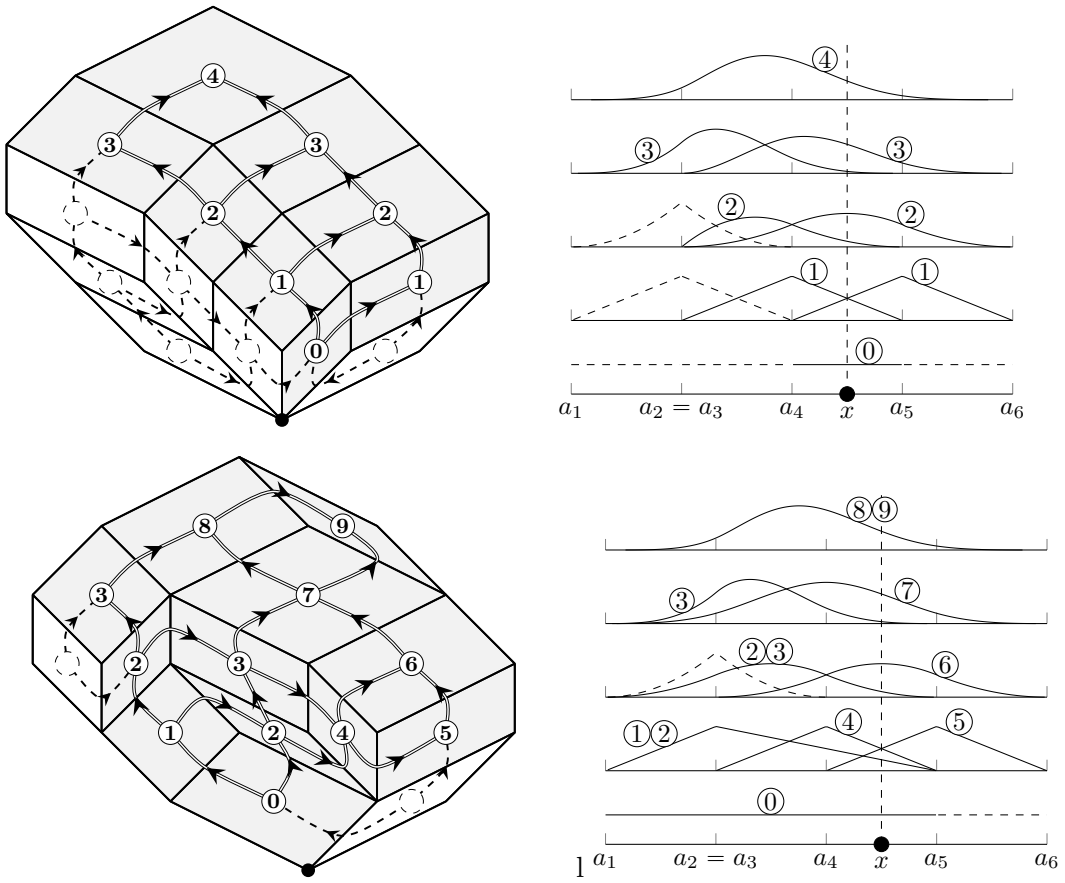


Figure 5.5: (Left) oriented dual graph $(\mathcal{G}, o_x)$ for the tilings of Figure 5.3, with the orientation induced by a point $x \in (a_4, a_5)$. The subgraph determined by the tiles supported at x is drawn with solid lines, and the tiles are numbered according to their position in a topological sorting of $(\mathcal{G}, o_x)$, starting with 0 for the tile $\Pi_{\varnothing,Z}$. (Right) corresponding spline functions supported at x.

5.4.3 Spline evaluation

Once all the spline functions supported at a given point x have been determined, one might be tempted to use the oriented graph $(\mathcal{G}, o_x)$ and its topological sorting to compute the value of all the spline functions on x.

Imagine that we want to compute, for some tile $\Pi_{I,B}$ supported at x, the value of $\overline{M}_b \coloneqq M(x \mid (a_i)_{i\in I\sqcup B\setminus\{b\}})$ for all $b \in B$, which can in turn be used to compute the value of the spline itself $\overline{M} \coloneqq M(x \mid \Pi_{I,B})$ using (5.2b). For every $b \in B$ and every point $x \in \mathbb{R}^d$, if $\overline{M}_b(x) \neq 0$, there is exactly one edge $\Pi_{I',B'} \to \Pi_{I,B}$ with $B \setminus \{b\} = B' \setminus \{b'\}$ and either $I = I'$, $I = I' \sqcup \{b'\}$, $I' = I \sqcup \{b\}$ or $I \sqcup \{b\} = I' \sqcup \{b'\}$. Suppose that the values of $M(x \mid (a_i)_{i\in I'\sqcup B'})$ and $M(x \mid (a_i)_{i\in I'\sqcup B'\setminus\{b'\}})$ for all $b' \in B'$ are known. Are we able to compute the value of $\overline{M}_b$? The answer depends on which case is realized. In particular:

(i) If $I = I'$, then $\overline{M}_b = M(x \mid (a_i)_{i\in I'\sqcup B'\setminus\{b'\}})$, which is known;

(ii) if $I = I' \sqcup \{b'\}$, then $\overline{M}_b = M(x \mid \Pi_{I',B'})$, which is also known;

(iii) if $I \sqcup \{b\} = I' \sqcup \{b'\}$, then $\overline{M}_b$ can be computed from the set of known values $M(x \mid (a_i)_{i\in I'\sqcup B'\setminus\{b'\}})$, $b' \in B'$ via a single application of (5.3).

However, in the case $I' = I \sqcup \{b\}$, there does not seem to be any obvious way to directly obtain $\overline{M}_b$. If this case happens only for a single $b \in B$, then it is still possible to obtain $\overline{M}_b$ via (5.2b), after noticing that $\overline{M} = M(x \mid (a_i)_{i\in I'\sqcup B'\setminus\{b'\}})$. In general, however, this case may happen more than once for a given point x and a given spline $M(x \mid \Pi_{I,B})$ when $d \geq 3$. Thus, it is essentially impossible to build an efficient recurrent evaluation scheme without the use of some auxiliary functions.

We propose here a slightly different construction, based on the following observation. First, notice that the problematic case $I' = I \sqcup \{b\}$ cannot arise if $M(x \mid \Pi_{I,B})$ is a spline of maximal degree for $\mathcal{P}$ (see Figure 5.3). However, if we consider a zonotopal tiling $\mathcal{P}_{I,B}$ of the zonotope $Z(V_{I,B})$ built on the reduced point configuration $A_{I,B} \coloneqq (a_i)_{i\in I\sqcup B}$, then $M(x \mid \Pi_{I,B})$ can indeed be obtained from any maximal-degree tile of $\mathcal{P}_{I,B}$. Thus, if in the evaluation of each spline $M(x \mid \Pi_{I,B})$ we use the reduced tiling $\mathcal{P}_{I,B}$, the problematic case $I' = I \sqcup \{b\}$ cannot occur, and neither can the case $I' = I$. Notice that an induced tiling $\mathcal{P}_{I,B}$ of $Z(V_{I,B})$ can simply be obtained from $\mathcal{P}$ via Lemma 5.2.2.

The reasoning of the previous paragraph suggests a simple procedure to build a set of auxiliary spline functions that are sufficient to compute, via recurrence, the value of any function $M(x \mid \Pi_{I,B})$:

(i) Build the tiling $\mathcal{P}_{I,B}$ induced by $\mathcal{P}$ on the reduced point configuration $A_{I,B} \coloneqq (a_i)_{i\in I\sqcup B}$ via Lemma 5.2.2;

(ii) For each $b \in B$, find the unique tile $\Pi_{I',B'} \in \mathcal{P}_{I,B}$, if any, such that $B \cap B' = B \setminus \{b\}$. If the tile exists, the value of $M(x \mid (a_i)_{i\in I\sqcup B\setminus\{b\}})$ can then be computed from the values of $M(x \mid \Pi_{I',B'})$ and $M(x \mid (a_i)_{i\in I'\sqcup B'\setminus\{b'\}})$, $b' \in B'$, either directly or through (5.3), otherwise the value is zero;

(iii) Store the subsets (I', B') found in step (ii), and repeat the same process from step (i) starting from each corresponding tile $\Pi_{I',B'}$.

The set of stored subsets (I', B') obtained during this process corresponds to a set of auxiliary spline functions that are sufficient to compute the value of the spline $M(x \mid \Pi_{I,B})$ for all x. Applying this process to all tiles $\Pi_{I,B} \in \mathcal{P}^{(k)}$ then yields a complete set of auxiliary functions sufficient for the evaluation of all the basis functions of order k via (5.2b) and (5.3). Notice that the same couple (I', B') can be obtained starting from multiple basis functions, in which case, it should obviously be stored only once.

So far, we have not detailed how the subsets corresponding to the tiles connected to $\Pi_{I,B}$ in the induced tiling $\mathcal{P}_{I,B}$ can be found efficiently in step (ii). Naively, one can start from the knowledge of the whole tiling $\mathcal{P}$ and apply Lemma 5.2.2, but this is obviously computationally infeasible in most applications. Thankfully, in the case of regular tilings, there is a more efficient way to compute them.

Lemma 5.4.8. *Let $\mathcal{P}(h)$ be a regular fine zonotopal tiling of $Z(V)$ with height function h, and let $\Pi_{I,B}$ and $\Pi_{I',B'}$ be two of its tiles, sharing a facet $\Pi_{J,C}$ with normal vector N_C. Define, for convenience,*

$$\begin{aligned} \sigma_{ij} &\coloneqq \operatorname{sign}\left(\det((a_c, h(a_c), 1)_{c\in C}, (a_i, h(a_i), 1), (a_j, h(a_j), 1))\right), \\ \sigma_i &\coloneqq \operatorname{sign}\left(\det((a_c, 1)_{c\in C}, (a_i, 1))\right). \end{aligned} \tag{5.38}$$

Then $b' \in I$ if and only if $\sigma_{bb'} \cdot \sigma_b > 0$, $b \in I'$ if and only if $\sigma_{bb'} \cdot \sigma_{b'} < 0$, and, choosing the orientation of N_C such that $\langle N_C, (x, 1)\rangle = \det((a_c, 1)_{c\in C}, (x, 1))$, $\operatorname{sign}(\langle N_C, z - z'\rangle) = \sigma_{bb'} \cdot \sigma_b \cdot \sigma_{b'}$ for all $z \in \Pi_{I,B}$, $z' \in \Pi_{I',B'}$.

Proof. The first two facts follow immediately from the Delaunay property (5.27), since, if $\sigma_b > 0$, then $b' \in I$ if and only if $\sigma_{bb'} > 0$, and the same is true if both signs are reversed. The same reasoning applies to the condition $b \in I'$ using $\sigma_{b'b} = -\sigma_{bb'}$ and $\sigma_{b'}$. If we now consider the representative points $z \in \Pi_{I,B}$ and $z' \in \Pi_{I',B'}$ defined as in (5.34), we can express their difference as

$$z - z' = \frac{1}{2}\left(\sigma_{bb'}\sigma_{b'} v_b + \sigma_{bb'}\sigma_b v_{b'}\right),$$

and therefore

$$\operatorname{sign}(\langle N_C, z - z'\rangle) = \frac{1}{2}\operatorname{sign}\left(\sigma_{bb'}\sigma_{b'}\langle N_C, (a_b, 1)\rangle + \sigma_{bb'}\sigma_b\langle N_C, (a_{b'}, 1)\rangle\right), \tag{5.39}$$

but since b and b' are on the same side of $\operatorname{aff}(\{a_c\}_{c\in C})$ if and only if $0 < \sigma_b \cdot \sigma_{b'} = (\sigma_{bb'}\sigma_b)\cdot(\sigma_{bb'}\sigma_{b'})$, the two terms in the sum on the right hand side of (5.39) always have the same sign, and we can thus rewrite (5.39) as

$$\frac{1}{2}\left(\sigma_{bb'}\sigma_{b'}\operatorname{sign}(\langle N_C, (a_b, 1)\rangle) + \sigma_{bb'}\sigma_b \operatorname{sign}(\langle N_C, (a_{b'}, 1)\rangle)\right) = \sigma_{bb'} \cdot \sigma_b \cdot \sigma_{b'},$$

since $\operatorname{sign}(\langle N_C, (a_b, 1)\rangle) = \sigma_b$, and similarly for b'. This completes the proof. □

In the case of regular tilings, Lemma 5.4.8 can be used to build any induced tiling $\mathcal{P}_{I,B}$, its dual graph and the induced orientations simply by taking the collection $\mathcal{B} \coloneqq \{B' \subseteq I \sqcup B :$

$|B'| = d+1, \det(B') \neq 0\}$ of all affinely independent subsets of size $d+1$ of $(a_i)_{i \in I \sqcup B}$, and using for each subset B' the signs $\sigma_{bb'}$, σ_b and $\sigma_{b'}$, $b' \in B'$ to construct the associated subset I' and form the tile $\Pi_{I',B'} \in \mathcal{P}_{I,B}$. The evaluation graph for $\Pi_{I,B}$ will then contain all the tiles directly adjacent to $\Pi_{I,B}$ in $\mathcal{P}_{I,B}$. Notice that, when all auxiliary functions are taken into account, the splines of degree zero do *not* constitute in general a triangulation of $\mathrm{conv}(A)$. However, it is still possible to build search trees capable of efficiently finding all the (possibly overlapping) simplices that contain a given point x, for example using structures such as bounding volumes hierarchies (BVH), of which the R-tree and $R^\star$-tree [283, 284] are prominent examples. We illustrate the construction of auxiliary functions and the corresponding evaluation obtained via the process outlined above in Figures 5.6 and 5.7 respectively.

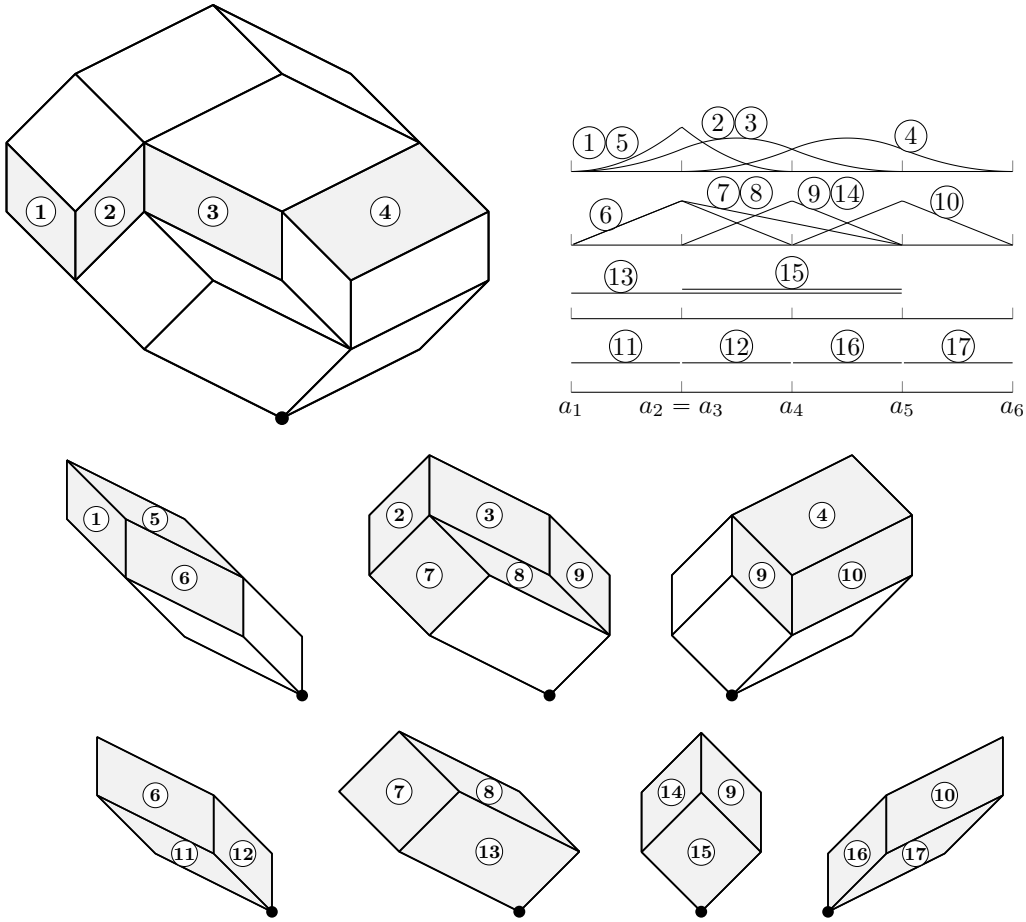


Figure 5.6: (Top left) a regular fine zonotopal tiling and the associated spline space over the point configuration of Figure 5.3, with the tiles corresponding to the splines of degree $k = 2$ (i.e., $\mathcal{P}^{(2)}$) highlighted and numbered from 1 to 4. (Top right) corresponding spline functions and auxiliary functions, numbered 5 through 17, computed by the process of Section 5.4.3. (Bottom) the induced zonotopal tilings $\mathcal{P}_{I,B}$ encountered during the construction of auxiliary spline functions. Highlighted tiles correspond to stored functions.

We end this section with a couple of final considerations. First, notice that it is not necessary to explicitly prove that the evaluation graph is acyclic, as this is evident from its construction. In particular, the evaluation graph for splines of order k clearly generates a k-partite oriented

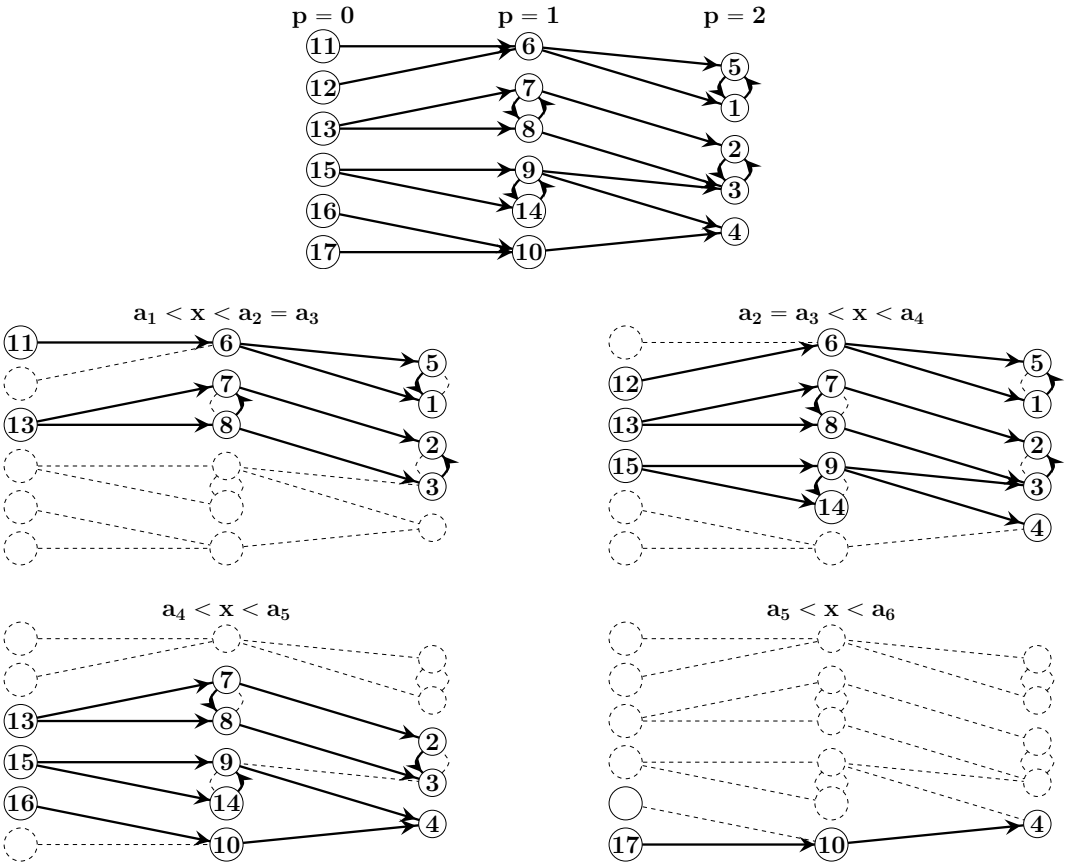


Figure 5.7: (Top) the complete graph containing all the auxiliary functions obtained via the construction presented in Section 5.4.3 in the case of the example of Figure 5.6. (Bottom) the actual evaluation graph obtained when computing the value of the spline functions at different locations x.

graph, to which some connections between splines of the same order are added (Figure 5.7). Since the connections between tiles of the same order are always a subset of those in the full dual graph $\mathcal{G}$ of $\mathcal{P}$, no cycle can be created by the orientation o_x induced by any point x.

Second, notice that in the special case where *every* point in A is repeated at least $k+1$ times, the construction process of Theorems 5.3.10 and 5.4.3 yields the usual Bernstein-Bézier functions [286] over a triangulation of $\mathrm{conv}(A)$, and the evaluation graph reduces to the usual de Casteljau algorithm [287] over each simplex.

Finally, notice that, as can be gleaned from Figure 5.6, the procedure outlined here does not lead in general to a minimal amount of auxiliary spline functions. In particular, each tile $\Pi_{I,B}$ for which there is an index $i \in I$ such that $a_i \notin \mathrm{conv}(\{a_b\}_{b\in B})$ can lead to an increased number of auxiliary functions. How often this happens is determined by the chosen height function h, either globally or locally in each induced tiling $\mathcal{P}_{I,B}$, and is related to the presence of *slivers*, i.e., simplices with skewed aspect ratios, in the associated weighted Delaunay triangulations. Some techniques exist to optimize the Delaunay height function in order to reduce the number of these elements, see, e.g., [288, 289]. We defer to a future work the investigation of how these techniques can help optimize the number of auxiliary functions required in the evaluation of simplex splines.

5.5 Discussion and further reading

The contents of this chapter were published in [290]. We wish to express our gratitude to the anonymous reviewers of the article for providing valuable feedback, and in particular for their proposal of an alternative, much simpler proof of Proposition 5.3.6, that we have integrated in our presentation here.

In this chapter, we have uncovered an interesting combinatorial structure capable of producing spaces of polynomial-reproducing multivariate (simplex) splines built atop any point configuration A, which ties them to the well studied fine zonotopal tilings of the associated zonotope $Z(V)$. This correspondence allows to generalize the set of known multivariate spline spaces and to adapt a known construction algorithm to a more general setting. When the tiling is regular, its dual graph provides a way to efficiently determine all the spline functions supported at any given point x, and to devise a recurrent evaluation scheme that reuses some intermediate results, thus providing a useful first step in the practical application of simplex spline bases in approximation and analysis.

Only fine zonotopal tilings have been explored in the present work. Possible connections between more general zonotopal tilings and other kinds of multivariate splines, such as box splines or more general polyhedral splines [216, 220] might be possible by relaxing this restriction.

From a computational standpoint, it is possible that the correspondence uncovered in the present work can be used to obtain further optimized algorithms for multivariate splines. Two aspects in particular deserve a particular attention in our opinion.

First, the evaluation scheme proposed in this work does not guarantee a minimal number of auxiliary functions. This aspect can perhaps be improved by using optimized weighted Delaunay triangulations coming from computer graphics applications (see, e.g., [288, 289]). In fact, some of the improved height functions used in these applications are able to create well-centered

simplices, which in our applications would lead to a significant improvement in the efficiency of the evaluation algorithm.

Second, the freedom given by the possibility of constructing spline spaces over point sets with repeated knots can be exploited to spaces with variable regularity and localized or arbitrarily-shaped discontinuities, with interesting applications in function approximation and numerical analysis. We use some of these features in Chapter 6.

We conclude with a few interesting references. First, notice that some connections between zonotopal tilings and box splines have been drawn in the past, see, e.g., [221, 222, 291, 292]. These connections are, as far as we know, unrelated to those presented here, as they pertain mainly to the shape of a single box spline, and not to spline spaces. Some deeper connections might however hide beneath the surface.

Second, zonotopal tilings are intimately related with another combinatorial structure, *oriented matroids*. Specifically, the Bohne-Dress theorem [266, 293, 294] states that zonotopal tilings are in bijection with single-element liftings of oriented matroids. Thus, spline spaces obtained in this chapter can be equivalently reformulated in terms of the oriented matroid built atop the vector configuration V. Other connections with many other combinatorial objects are known, although it is not immediately clear if any of them can be used to derive some other efficient algorithm for simplex spline computation.

6 | Fully unstructured multi-patch DG-IGA scheme

[...] prenez un petit bout de sens
puis un grand morceau d'innocence
faites chauffer à petit feu
au petit feu de la technique
versez la sauce énigmatique
saupoudrez de quelques étoiles
poivrez et mettez les voiles [...]

Raymond Queneau, Le Chien à la Mandoline,
Pour un art poétique (1958)

As we have discussed in Chapter 2 and detailed in Chapter 3, one of the most successful techniques for the numerical solution of PDEs is the standard finite element (FE) analysis. This technique relies on piecewise-polynomial functions defined over an underling mesh which have global C^0 regularity over the simulation domain Ω. Starting from this technique, two seemingly opposing tendencies have arisen over the years.

In discontinuous Galerkin (DG) schemes, the basis functions are replaced by independent polynomial bases over each mesh element, making the functions discontinuous on mesh faces. Continuity is then restored via the imposition of suitable numerical fluxes and penalty terms between elements. This fully unstructured approach offers a great deal of flexibility and combines a good modelization of complex geometries with local h (mesh size), k (polynomial degree) and even t (timestep)-adaptivity, which is especially appreciated in the physical sciences (see e.g. [1] for a recent application). In the case of time-domain wave propagation, the block-diagonal nature of the mass matrix also allows to use efficient explicit time discretization schemes. As the degree k of the basis increases, the CFL restriction on the timestep scales like $O(h/k^2)$, where h is the spatial discretization step.

Going in a seemingly opposite direction, isogeometric analysis (IGA) [2, 252] replaces the standard FE basis by B-splines, i.e., piecewise-polynomial functions of degree k with increased regularity, up to order C^{k-1}. Since these functions (and their rational counterparts, NURBS functions) are routinely used in engineering to represent the (exact) geometry of mechanical pieces, IGA obviates the need to mesh the simulation domain before performing analysis, eliminating the associated potential discrepancy as a source of error. Moreover, IGA has been

proven to possess superior numerical properties, including a CFL condition timestep for wave propagation that scales like $O(h/k)$ (cf. the last section of Chapter 3).

Aiming at bridging the chasm between these two worlds, some recent approaches have focused on formulating independent IGA schemes over different B-spline patches, which are then coupled through the introduction of DG-like fluxes and penalties. These attempts, which are often named multi-patch DG-IGA schemes, have proven fruitful, with a recent work [3] highlighting how its application to time-domain wave propagation allows to retain the parallelization potential of DG (and its associated block-diagonal mass matrix) with the improved CFL condition typical of IGA.

However, the parameterization of IGA patches suitable for numerical analysis is far from a trivial task, and even ensuring its injectivity requires a careful placement of control points (see, e.g., [295, 296]). This problem is even more relevant in many applications to the natural sciences, where the geometry of discontinuities can be complex and of arbitrary topology and pre-existing CAD models are lacking or nonexistent. The problem is compounded when dealing with the seismic inverse problem, which requires a highly flexible geometric description whose features and discontinuities are unknown at the beginning of the inversion, making it extremely hard to devise a suitable CAD patch structure. This is unfortunate, as problems like seismic full-waveform inversion simultaneously demand a high efficiency per degree of freedom in the solution of the PDEs, and the accurate reproduction of the geometry of sharp contrasts and boundaries, two tasks at which IGA excels.

In this chapter, we present a simple but highly flexible numerical scheme that reproduces the features of multi-patch DG-IGA approaches but relies on unstructured multivariate splines. The corresponding IGA patches can have arbitrary topology, and are built starting from a simple set of points, without further structure, simplifying considerably their use in inverse problems. Moreover, our basis functions reproduce the usual FE and DG bases as special cases, making it very easy to seamlessly couple all three numerical schemes in a single simulation.

In Section 6.1, we introduce our numerical scheme, based on the symmetric interior penalty DG scheme (IPDG), and we point out its main differences with the usual IPDG and IGA approaches. In Section 6.2, we show how to complete the results of Chapter 5 to create separate spline spaces over each subdomain. We also discuss an insightful limiting case of our scheme, showing that it can reproduce the usual FE and DG bases as special cases. In Section 6.3 we show that, near the domain boundaries, unstructured spline spaces behave very similarly to the usual DG bases, and thus the same well-known inverse inequalities can be used to guarantee the positivity of the bilinear form associated to the weak formulation, and the applicability of the usual *a priori* error analysis estimates. A few two- and three-dimensional numerical experiments that prove the efficacy and flexibility of our method are presented in Section 6.4, and some of the advantages and drawbacks of the method are discussed in Section 6.5.

6.1 Overview of the numerical scheme

The starting point for our fully unstructured multi-patch DG-IGA scheme for wave propagation is similar to the usual IPDG scheme presented in Chapter 2, i.e., the weak problem

$$\sum_{i=1}^{n_d} \int_{\Omega_i} \frac{1}{\lambda} \varphi \frac{\partial^2 p}{\partial t^2} \,\mathrm{d}\Omega + \sum_{i=1}^{n_d} \int_{\Omega_i} \frac{1}{\rho} \nabla\varphi \cdot \nabla p \,\mathrm{d}\Omega - \sum_{F \in \mathcal{F}_{\mathrm{DG}}} \int_F [\![\varphi]\!] \cdot \{\!\{ \frac{1}{\rho} \nabla p \}\!\} \,\mathrm{d}F \tag{6.1}$$

$$- \sum_{F \in \mathcal{F}_{\mathrm{DG}}} \int_F \{\!\{ \frac{1}{\rho} \nabla\varphi \}\!\} \cdot [\![p]\!] \,\mathrm{d}F + \alpha \sum_{F \in \mathcal{F}_{\mathrm{DG}}} \int_F [\![\varphi]\!] \cdot [\![p]\!] \,\mathrm{d}F + \sum_{F \in \mathcal{F}_{\mathrm{A}}} \int_F \frac{1}{\rho c} \varphi \frac{\partial p}{\partial t} \,\mathrm{d}S = \sum_{i=1}^{n_d} \int_{\Omega_i} \varphi s \,\mathrm{d}\Omega,$$

where the domain of interest Ω follows the hypotheses of Theorem 1.2.3 and s is the source term.

Compared to the usual IPDG approach, there are a few differences. First of all, the subdomains $\Omega_i \subseteq \Omega$ appearing in the sums of (6.1) are in general neither simplices, nor even necessarily convex. In fact, we wish to build complex domains whose shape follows the irregularities of the propagation media. We do not even assume that the domains are simply connected, and we allow the presence of internal boundaries. However, we still assume that the domains are *polyhedral* (or polygonal in two dimensions), and that they form a *subdivision* of Ω, i.e.,

$$\Omega_i \cap \Omega_j = \varnothing \text{ for } i \neq j, \quad \bigcup_{i=1}^{n_d} \overline{\Omega}_i = \overline{\Omega}.$$

As in many classic implementations of the IPDG method, we assume that the physical parameters λ, ρ and c are positive and constant over each domain, with $c = \sqrt{\lambda/\rho}$. Notice that in our case the subdomains can be large. In order to account for variable physical parameters in a domain, one might either subdivide the domain into suitable subdomains, or introduce space-dependent physical parameters, possibly by expressing them using the same basis spline functions as the solution, or, more likely, spline functions of a lower degree. We do not follow this approach here, for simplicity, even tough our presentation could easily be modified to accommodate it.

Similarly to the usual IPDG approach, and in contrast with methods such as continuous finite elements, we can be quite liberal with the imposition of boundary conditions. For example, one can introduce internal boundaries into any domain, and treat them either as physical barriers by imposing suitable (internal) boundary conditions (e.g., absorbing, reflecting, or any other), or as numerical barriers, where the physical propagation of the wave is unimpeded, but where continuity is only imposed weakly via the usual DG fluxes and penalty terms.

More precisely, we denote by $\mathcal{F}$ the set of all internal and external boundary facets of the compatible domains Ω_i, i.e.,

$$\mathcal{F} \coloneqq \bigcup_{i=1}^{n_d} \{F : F \in \partial\Omega_i\},$$

and we select four disjoint subsets $\mathcal{F}_{\mathrm{DG}}$, $\mathcal{F}_{\mathrm{A}}$, $\mathcal{F}_{\mathrm{N}}$, and $\mathcal{F}_{\mathrm{D}}$ of $\mathcal{F}$, on which we impose, respectively, a DG flux (DG), a lowest-order absorbing boundary condition (A) (1.9), a free-surface Neumann condition (N) (1.8), or the Dirichlet condition $p|_F = 0$ (D), see Figure 6.7 for an example. Notice that free-surface and Dirichlet conditions do not appear directly in (6.1). In fact, the contribution

of free-surface facets simply vanishes in (6.1), while the Dirichlet condition is imposed implicitly by selecting a suitable discrete spline space that satisfies them.

We impose a few constraints on the allowed internal and external boundary conditions. Clearly, we require that

$$\mathcal{F}_a \cap \mathcal{F}_b = \varnothing \text{ for } a, b \in \{\mathrm{DG}, \mathrm{A}, \mathrm{N}, \mathrm{D}\}, a \neq b \quad \text{and} \quad \bigsqcup_{a \in \{\mathrm{DG},\mathrm{A},\mathrm{N},\mathrm{D}\}} \mathcal{F}_a = \mathcal{F},$$

i.e., that there are no holes in the boundary conditions. Furthermore, we require that

$$\mathcal{F}_{\mathrm{DG}} \cap \partial\Omega = \varnothing,$$

i.e., that DG fluxes are never imposed on the external boundaries of the simulation domain Ω. These fluxes can however be imposed on boundaries between domains or on internal boundaries of a given domain, in which case we denote them as *transparent* boundaries. Notice that there is no restriction on the use of absorbing, free-surface or Dirichlet boundary conditions on internal boundaries within a subdomain, or on partial boundaries between subdomains. This allows the creation of non-simply-connected subdomains. We illustrate this point with some numerical examples in Section 6.4.

Notwithstanding the great variety of allowed combinations of boundary conditions, most applications will be restricted to absorbing or free-surface boundary conditions on external or internal boundaries (i.e., holes) of Ω, and DG conditions between neighboring subdomains.

Notice also that it would be possible to impose uni-directional constraints, as for example a surface that is absorbing as seen from one side, but reflecting or transparent as seen from the other. This would simply require associating boundary conditions to *half-facets* instead of facets, see e.g. [297, 298]. However, we are not aware of any physical application of these possibilities, and thus we do not explore them here, for ease of presentation.

After the subdomains and boundary conditions have been selected and the weak form (6.1) has been written, a suitable spline space $\mathcal{S}_i$ is introduced in each subdomain Ω_i. We describe how this spline space is obtained in the next section.

6.2 Multi-patch spline space construction

In the previous chapter, polynomial-reproducing spline spaces have been tied to some combinatorial geometry objects known as zonotopal tilings. Specifically, the structure of these objects has been exploited to build a set of simplex splines containing all the polynomials over suitable subsets of $\mathrm{conv}(A)$ in their linear span. This combinatorial approach works in any space dimension, and can be used even when points in A are affinely dependent or repeated more than once. In particular, if the points lying directly on the convex hull of A are repeated $k+1$ times, then the spline space of degree k reproduces all the polynomials up to degree k over $\mathrm{conv}(A)$, and thus allows the imposition of boundary conditions.

However, this scheme is not directly applicable to our case, since we are not assuming that the domains Ω_i are convex, and therefore we cannot simply introduce a point cloud in each domain and build a spline space on its convex hull. We therefore introduce in this section the

assumptions on the domain boundaries and on the height function that are required in order to produce a suitable space of multi-patch splines.

6.2.1 Gabriel property

One potential roadblock lies with the fact that the construction algorithm is based on successive weighted Delaunay triangulation steps. The issue of building a Delaunay triangulation that respects a given set of constraints (constrained Delaunay triangulation, or CDT) is notoriously hard [299], and even determining whether a domain is triangulable or not is in general NP-hard (non polynomial hardness, cf. [300]). For this reason, we shift this burden to a suitable pre-processing step, and we require that a suitable point configuration and a suitable height function are selected.

The first step in this construction consists in ensuring that all the boundary facets used for the definition of the domains and the boundary conditions are represented in triangulations of the point configuration A. Let h be a generic height function in the sense of Definition 5.4.2, and $\mathcal{T}_A(h)$ be the associated weighted Delaunay triangulation of $\mathrm{conv}(A)$, as defined in the second half of Chapter 5. We introduce the following weighted version of the well-known Gabriel property [301] of Delaunay facets, see also [302].

Definition 6.2.1 (Gabriel property)**.** *Let $C \subset [n]$ be a subset of indices such that $|C| = d$. Suppose that there exists a point $\gamma \in \mathbb{R}^d$ such that $\det((a_c, 1)_{c\in C}, (\gamma, 1)) = 0$ and*

$$\gamma \cdot (a_i - a_c) - \frac{h(a_i) - h(a_c)}{2} \leq 0, \tag{6.2}$$

for all $i \in [n]$, with equality if and only if $i \in C$. Then the facet $\mathrm{aff}((a_c)_{c\in C})$ is said to have the Gabriel property.

Gabriel facets are interesting because they are automatically included in the weighted Delaunay triangulation. To see this, notice that (6.2) is affine in $(a_i, h(a_i))$, and describes a hyperplane $H \subset \mathbb{R}^{d+1}$ that passes through the points $(a_c, h(a_c))_{c\in C}$ and has all the points $(a_i, h(a_i))_{i\in[n]\setminus C}$ on its negative side. By rotating this plane around $\mathrm{aff}((a_c, h(a_c))_{c\in C}$ until it touches another point $(a_b, h(a_b))$ for some $b \notin C$, one finds that the points $C \sqcup \{b\}$ satisfy the weighted Delaunay condition (5.27), and therefore the corresponding simplex $\mathrm{conv}((a_i)_{i,\in C\sqcup\{b\}})$ (and thus the facet $\mathrm{conv}((a_i)_{i,\in C})$) appears in the corresponding weighted Delaunay triangulation.

Moreover, given a Gabriel facet, one can find the corresponding point γ by choosing an index $c \in C$ and finding a solution to the d equations $\det((a_c, 1)_{c\in C}, (\gamma, 1)) = 0$ and $\gamma \cdot (a_i - a_c) = (h(a_i) - h(a_c))/2$ for $i \in C$, $i \neq c$. The point γ represents the weighted version of the center of the diametrical sphere circumscribed to $\mathrm{conv}((a_c)_{c\in C})$. Therefore, if (6.2) is not satisfied, then one may proceed to *subdivide* the facet by introducing a new point $e = \sum_{c\in C} \lambda_c a_c$, with barycentric coordinates λ_c satisfying $0 \leq \lambda_c \leq 1$ and $\sum_{c\in C} \lambda_c = 1$ and an associated weight $h(e) \leq \sum_{c\in C} \lambda_c h(a_c)$. The sublinearity of the weight then makes it more likely that the Gabriel condition (6.2) is satisfied. If not, one may proceed to further subdivide the facets. For classical Delaunay triangulations, this process is known as making the triangulation *conforming Gabriel*, see, e.g., [303, 304] and Figure 6.1. We will not enter here the details of how one can most

efficiently refine the constraint facets so that they are conforming Gabriel and instead assume that this pre-processing step has already been performed.

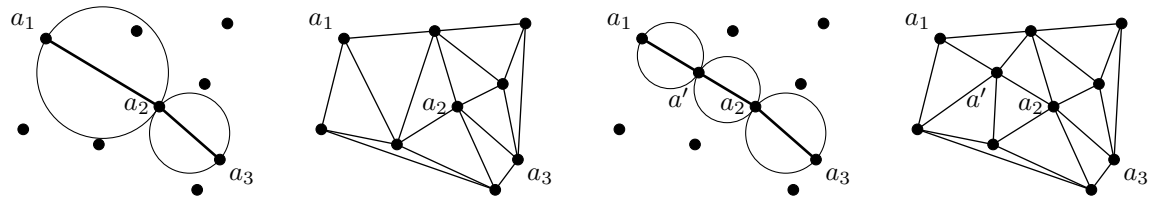


Figure 6.1: Example of Gabriel facets in two dimensions for the standard Delaunay weights. (Left) the facet a_2a_3 is Gabriel, since its diametral sphere is empty, but the facet a_1a_2 is not, and in fact it is not part of the Delaunay triangulation (center left). (Center right) The Gabriel property can however be restored by refining the facet a_1a_2, introducing the additional point a'. The facet a_1a_2 is then guaranteed to be represented in the modified Delaunay triangulation (right).

6.2.2 Fine-grained height function

After making sure that all the constraint facets $F \in \mathcal{F}$ are Gabriel, one still has to make sure that they are included not only in the order-zero Delaunay triangulation of A, but also in all the triangulations at orders $r \leq k$ that are required for the construction process of Theorem 5.4.3. We show that repeating all the points belonging to constraint facets $k+1$ times in A is sufficient, after an appropriate height function h has been chosen.

Let us define an *independent set* of indices B as a subset $B \subseteq [n]$, $|B| = d+1$ such that the points $(a_i)_{i\in B}$ are all affinely independent. These subsets are in bijection with the tiles of any fine zonotopal tiling of $Z(V)$. To any independent set of indices B, we can associate the (non-vertical) hyperplane $H_B \subset \mathbb{R}^{d+1}$ passing through all the points $(a_i, h(a_i))_{i\in B}$. We can now define a *fine-grained* height function as follows.

Definition 6.2.2 (Fine-grained height function)**.** *A height function h is said to be* fine-grained *if, for every independent set B and for every set of indices $I \subseteq [n]$ such that the points $(a_i)_{i\in I}$ are all coincident, either $I \cap B \neq \varnothing$, or all the points indexed by I are on the same side of H_B.*

In other words, with a fine-grained height function, the different copies of a point in A have very close heights, so that the corresponding lifted points in $\hat{A}$ can only be separated by a hyperplane defined using one of the points themselves. In practice, this property is easy to achieve using symbolic perturbation, by defining a custom comparison function that considers the height of coincident points in A as equal when comparing against another distinct point of A. One can then simply order consistently the heights of these points when they are compared against each other. Notice that the notion of a fine-grained height function is compatible with the Gabriel property. In fact, when defining a Gabriel facet using points in A that are repeated more than once, one can use in the definition the copy of the point with the lowest value of the height function, which satisfies (6.2) automatically. The following property immediately follows from Definition 6.2.2.

Lemma 6.2.3. *Suppose that the height function h is fine-grained, and let (I, B), $|I| = k$, be a couple of knot indices defining a tile $\Pi_{I,B} \in \mathcal{P}(h)$. Suppose that a point a_i associated to an index $i \in I$ is repeated at least $k+1$ times in A. Then, there exists exactly one index $b \in B$ such that $a_b = a_i$.*

Proof. Suppose that the points $(a_b)_{b\in B}$ are all distinct from a_i. Then, since h is fine-grained, the hyperplane $H_B \coloneqq \mathrm{aff}((a_b, h(a_b))_{b\in B})$ contains all the (lifted) copies of a_i on the same side. Due to (5.27) and Theorem 5.4.3, this implies that either none of the $k+1$ copies of a_i are in I, which contradicts the hypothesis, or that they all are, which is impossible since $|I| = k$. Therefore, there must be at least an index $b \in B$ such that $a_b = a_i$, and there cannot be more than one since the points indexed by B define a tile and thus $\det(B) > 0$. □

The previous Lemma states that, if a point is given a multiplicity of at least $k+1$ in A, then it can be an internal knot of a spline of degree k only if one of its copies is also one of the boundary knots of the spline. This suggests that repeating points in A is an effective way of carving out sub-domains in A, as proven in the following subsection.

6.2.3 Multi-patch spline space

We prove now that, if the properties discussed in the preceding sections are satisfied, the spline space built via the process detailed in Chapter 5 indeed yields a spline space for each subdomain.

We first prove the following proposition, that characterizes the behavior of these spline spaces near the constraint facets $F \in \mathcal{F}$, and therefore near the boundary of Ω and the interfaces between subdomains. By pointing out the similarity with the usual IPDG bases, we pave the way for a simple adaptation of many proofs of the numerical properties of this method to our approach.

Proposition 6.2.4. *Let h be a generic, fine-grained height function, and let F be a Gabriel facet, arbitrarily oriented, whose vertices are repeated at least $k+1$ times in A. If there are at least $k+1$ points in A on the positive side of F, then there are exactly $\binom{k+d}{k}$ spline functions $M(x \mid I \sqcup B)$ of degree k such that $\mathrm{conv}((a_b)_{b\in B})$ lies on the positive side of F and contains F as a facet.*

Proof. Let $(a_c)_{c\in C}$ be a set of $|C| = d$ points such that $F = \mathrm{conv}((a_c)_{c\in C})$. By hypothesis, each point a_c is repeated $k+1$ times in A, and any of these copies can be used to define the facet F. Suppose that we have made a choice, and for each point a_c, define the subset of indices R_c as

$$R_c \coloneqq \{i \in [n] : a_i = a_c,\ h(a_i) > h(a_c)\}.$$

Let us then choose the indices $c \in C$ such that $\sum_{c\in C} |R_c| = k' \leq k$. Notice that since $|R_c| \leq k$ and $|C| = d$, there are exactly $\binom{k+d}{k}$ distinct ways to perform such a choice.

Since F is Gabriel, there exists a hyperplane $H_C \subset \mathbb{R}^{d+1}$ passing through $(a_c, h(a_c))_{c\in C}$ and containing all the other distinct points of A on its negative side. Let x be a point on the positive side of F, and let us rotate H_C around $\mathrm{aff}((a_c, h(a_c))_{c\in C})$ so that the quantity $(N_{d+1}\hat{x}_{d+1})$ decreases, where N is the oriented normal of H_C, $\hat{x}$ is the projection of x onto H_C, and the subscript $d+1$ denotes the $(d+1)$-th coordinate. Denote by $H_C(\theta)$ the hyperplane H_C rotated

by an angle θ, such that $H_C(0) = H_C$ and $H_C(\overline{\theta})$ is a vertical hyperplane for some $0 < \overline{\theta} \leq \pi/2$. Since there are at least $k+1$ points of A on the positive side of F, and h is generic (Definition 5.4.2), the rotating plane $H_C(\theta)$ must encounter at least $k+1$ points $(a_i, h(a_i))$ before becoming vertical. Let us denote by

$$S \coloneqq \{(\theta, i) : (a_i, h(a_i)) \in H_C(\theta) \text{ for some } 0 < \theta < \overline{\theta}\}$$

the set of encountered points, ordered by increasing angle θ. Notice that the corresponding points in A must be distinct from the points in $(a_c)_{c\in C}$, otherwise the associated hyperplane would be vertical. Let $S_{\leq k-k'}$ denote the set of the first $k-k'$ entries of S, and let (θ_b, b) be the following (i.e., the $(k-k'+1)$-th) entry. Then, the hyperplane $H_C(\theta_b)$ passes through all the points $(a_j, h(a_j))_{j\in B}$, where $B \coloneqq C \sqcup \{b\}$, and it contains on its positive side exactly the points indexed by

$$I \coloneqq \left(\bigsqcup_{c\in C} R_c\right) \sqcup \{i : (\theta, i) \in S_{\leq k-k'}\},$$

i.e., the disjoint union of the k' points in the sets R_c, $c \in C$, with the indices in the first $k-k'$ entries of S. Thus, by (5.27) the simplex spline $M(x \mid I \sqcup B)$ has degree $|I| = k$ and satisfies the property of this proposition. We can build in this way exactly $\binom{k+d}{k}$ distinct simplex splines, completing the proof. □

Notice that the definition of positive side of F is arbitrary. Thus, if the facet F has at least $k+1$ points of A on each side, then Proposition 6.2.4 can be applied to both sides of F. An example of the simplices of Proposition 6.2.4 is shown in Figure 6.3, top right.

Proposition 6.2.4 is a generalization of a property of the usual mesh-based discontinuous Galerkin methods, where for every facet F in the mesh having a simplicial mesh element on its positive side, there are exactly $\binom{k+d}{k}$ polynomials supported on the mesh element adjacent to F. We show in the next subsection that this is not an accident.

Proposition 6.2.4 immediately allows to subdivide the spline functions in the spline space according to the subdomains $(\Omega_i)_{i=1}^{n_d}$. In fact,

Corollary 6.2.5. *Let h satisfy the hypotheses of Proposition 6.2.4, and suppose that each facet $F \in \mathcal{F}$ is Gabriel and its vertices are repeated at least $k+1$ times in A. Then, for every spline $M(x \mid I \sqcup B)$ of degree k, the interior of the simplex $\operatorname{conv}((a_b)_{b\in B})$ cannot intersect the boundary of any subdomain.*

Proof. For every facet $F \in \mathcal{F}$ on the boundary of a subdomain Ω_i, define the positive side as the side on which Ω_i lies. Then, there are at least $k+1$ points on the positive side of F, and Proposition 6.2.4 ensures that there are $\binom{k+d}{k}$ spline functions $M(x \mid I \sqcup B)$ whose simplex $\operatorname{conv}((a_b)_{b\in B})$ is adjacent to F on its positive side. Notice that these simplices do not cross F. Let C be the intersection of their interiors. Let $M(x \mid J \sqcup B')$ be a distinct simplex spline whose associated simplex $\operatorname{conv}((a_b)_{b\in B'})$ intersects F. Then, the interior of the associated simplex must also intersect C, in contradiction with Proposition 5.3.12. The same reasoning can be repeated for all the facets on the boundary of Ω_i, since they are all Gabriel by hypothesis, proving the corollary. □

If one chooses a set of domain boundaries and a height function that satisfy the hypotheses of Corollary 6.2.5, then one can use the process detailed in the previous section to build a spline space over the whole Ω, and assign each spline function to a subdomain. This yields one spline space per subdomain, according to the criterion

$$M(x \mid I \sqcup B) \in \mathcal{S}_i \text{ if and only if } \operatorname{conv}((a_b)_{b\in B}) \subseteq \Omega_i.$$

This is the criterion that we choose to construct our discontinuous spline space. Notice that the internal points of the splines in $\mathcal{S}_i$ are also contained in Ω_i. This is trivially true for splines of degree 0, and since step (iii) in Theorem 5.3.10 states that the set of internal points of splines of degree k is a subset of the set of points $I \sqcup B$ of splines of order $k-1$, it is true by induction at all degrees k. Notice also that the spline spaces produced under the assumptions of Corollary 6.2.5 are indeed discontinuous. In fact, they include splines of degree k with $k+d$ points (out of the total $k+d+1$) lying on the same facet F on the boundary of the domain. Using the recurrence relation (5.2), it is easy to show that these functions have a non-zero trace on F, i.e., on the boundary of the subdomain. This allows to evaluate boundary conditions, inter-domain fluxes and penalty terms. In all cases, the splines in $\mathcal{S}_i$ are taken to be zero outside Ω_i.

Finally, notice that Corollary 6.2.5 has an interesting special case.

Corollary 6.2.6. *Let the height function h and the subdomains $(\Omega_i)_{i=1}^{n_d}$ satisfy the hypotheses of Corollary 6.2.5, and suppose furthermore that each subdomain Ω_i is a simplex, such that $(\Omega_i)_{i=1}^{n_d}$ forms a triangulation of Ω. Then, the spline space built by Corollary 6.2.5 produces the Bernstein-Bézier discontinuous Galerkin basis over each simplex Ω_i.*

Proof. One can simply follow the proof of Proposition 6.2.4, but notice that, within the assumptions of this corollary, the $k'-k$ points first encountered by the rotating plane $H_C(\theta)$, i.e., the points in in $S_{\leq k-k'}$, must be all liftings of coincident points in A, i.e., points $(a_i, h(a_i))$ with the same first d coordinates. Thus, the knot vector of the resulting spline consists of $d+1$ distinct points $\hat{A} \coloneqq (a_i)_{i=1}^{d+1} \subset \mathbb{R}^d$, each repeated a number of times $1 \leq r_i \leq k+1$, with $\sum_{i=1}^{d+1} r_i = d+k+1$. Definition (4.10) then shows that these splines act on a function f as

$$\left\langle f, M(\hat{A}) \right\rangle = \frac{1}{B(r_1, \ldots, r_{d+1})} \int_{\Sigma^d} f\left(\sum_{i=1}^{d+1} \lambda_i a_i\right) \lambda_1^{r_1-1} \cdots \lambda_{d+1}^{r_{d+1}-1} \, \mathrm{d}\Sigma,$$

where the integration is over the d-dimensional simplex Σ^d described by the barycentric coordinates $\lambda_1, \ldots, \lambda_{d+1}$ with $0 \leq \lambda_i \leq 1$ and $\sum_{i=1}^{d+1} \lambda_i = 1$. The product $\lambda_1^{r_1-1} \cdots \lambda_{d+1}^{r_{d+1}-1}$ can be recognized as the barycentric representation of a Bernstein-Bézeir polynomial of degree $\sum_{i=1}^{d+1}(r_i-1) = k$, cf. (3.13). Moreover, all these splines are supported on the same simplex, and all the $\binom{k+d}{k}$ combinations of multiplicities that sum to $k+d+1$ are obtained, as shown in the proof of Proposition 6.2.4. Thus, the whole Bernstein-Bézier basis over the simplex is obtained. Finally, the same simplex is obtained for the splines at any degree $r \leq k$, and in particular for $r = 0$, in which case one obtains the weighted Delaunay triangulation of $\operatorname{conv}(A)$ with height function h (see Proposition 5.3.6). By the Gabriel property, the triangulation created by the simplices Ω_i must be a subset of the weighted Delaunay triangulation of A, completing the proof. □

This result proves that one can obtain the usual interior-penalty discontinuous Galerkin scheme, on a Bernstein-Bézier basis, as a special case of our method. Thus, one expects these two numerical schemes to be naturally compatible, and easy to mix in a single simulation. We show that this is indeed the case with some numerical experiments in Section 6.4.

6.2.4 Some algorithmic aspects

We discuss here a few implementation details of our method.

From simplex splines to multivariate B-splines

We do not use directly the simplex splines described in the previous section as functions for our discretized scheme. In fact, the polynomial-reproduction equation (5.10) of Theorem 5.3.3 shows that the coefficient of a simplex spline in the development of a given polynomial only depends on the internal knots of the spline function. But in our spaces, many different splines can share the same internal knot indices I, namely all those obtained from the triangulation of the link region $\mathcal{R}(I)$ of Definition (5.3.7). Consequently, the decomposition is not unique, since any linear combination of these splines can be used.

We remove this unwanted freedom by fixing a given linear combination of splines sharing the same internal indices. Specifically, as done, e.g., in [259, Chapter 8], we define for every set of internal indices I the *multivariate B-spline function* of degree k

$$N(x \mid I) \coloneqq \frac{k!}{(k+d)!} \sum_{\{B:\Pi_{I,B}\in\mathcal{P}^{(k)}\}} \det(B) M(x \mid I \sqcup B). \tag{6.3}$$

It is important to remark that, in order to preserve the multi-patch property of the spline space, the sum (6.3) is done independently for every subdomain Ω_i, and it only includes spline functions in $\mathcal{S}_i$. Simplex splines with the same internal knots but belonging to different patches $\mathcal{S}_i$ are not summed, to preserve the domain decomposition. We illustrate this sum in Figure 6.2.

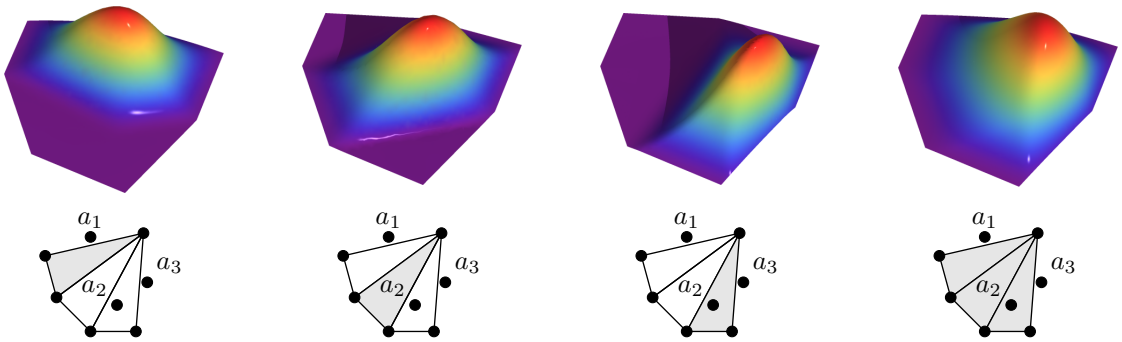


Figure 6.2: (Top) multivariate B-spline function of degree $k = 3$ (last picture), obtained as the linear combination of the three simplex splines on its left. The splines share the same set of internal knot indices $I = \{1, 2, 3\}$. (Bottom) the corresponding link region $\mathcal{R}(I)$ and its triangulation, yielding the three simplex splines in the sum.

Finally, multivariate B-splines can also be used to recover, as a special case, another well-known basis.

Corollary 6.2.7. *Suppose that the height function h satisfies the hypotheses of Corollary 6.2.6, and there is only one subdomain $\Omega_1 = \Omega$. Suppose that all the points have multiplicity $k+1$ in A, and define the multivariate B-spline functions via* (6.3)*. Then the multivariate B-spline functions defined via* (6.3) *correspond to the usual C^0 Bernstein-Bézier finite element basis on the weighted Delaunay triangulation $\mathcal{T}(h)$ of Ω with height function h.*

Proof. Since all the points in A are repeated $k+1$ times, the same reasoning as in the proof of Corollary 6.2.6 can be applied to each facet of $\mathcal{T}(h)$ to show that the simplex splines correspond to a complete basis of Bernstein-Bézier polynomials of degree k on $\mathcal{T}(h)$. Two such splines $M(x \mid I \sqcup B_1)$ and $M(x \mid I \sqcup B_2)$ can share the same set of internal knot indices $I \neq \varnothing$ if and only if the knots $(a_i)_{i\in I}$ are vertices of the simplex $\sigma_I \coloneqq \sigma_{B_1} \cap \sigma_{B_2}$ obtained as the intersection of the two (closed) supporting simplices $\sigma_{B_1} \coloneqq \operatorname{conv}((a_b)_{b\in B_1})$ and $\sigma_{B_2} \coloneqq \operatorname{conv}((a_b)_{b\in B_2})$. Furthermore, these knots must have the same multiplicity in both splines. Let $d' < d$ be the dimension of σ_I, and consider the spline $M(x \mid I \sqcup B_1)$, ordering its knots such that the vertices of σ_I appear first. Then, since all the other knots must have multiplicity 1 in $I \sqcup B_1$, the corresponding exponents in (4.10) are zero, and thus the trace of the simplex spline function $M(x \mid I \sqcup B_1)$ on σ_I can be simply be obtained from (4.10) after setting the barycentric coordinates $\lambda_{d'+2} = \ldots, = \lambda_{d+1} = 0$, and multiplying by the Jacobian $\operatorname{vol}^{d'}(\sigma_I)/\operatorname{vol}^{d}(\sigma_{B_1})$,

$$\langle f, M(I \sqcup B_1)\rangle_{\sigma_I} = \frac{\operatorname{vol}^{d'}(\sigma_I)}{\operatorname{vol}^{d}(\sigma_{B_1})} \frac{1}{B(r_1, \ldots, r_{d'+1}, 1, \ldots, 1)} \int_{\sigma_I} f\left(\sum_{i=1}^{d'+1} \lambda_i a_i\right) \lambda_1^{r_1-1} \cdots \lambda_{d'+1}^{r_{d'+1}-1} \,\mathrm{d}\sigma_I, \tag{6.4}$$

and similarly for $M(x \mid I \sqcup B_2)$. Here, $(a_i)_{i=1}^{d'+1}$ is the set of distinct vertices of σ_I and the integers $(r_i)_{i=1}^{d'}$ represent their multiplicity in both $I \sqcup B_1$ and $I \sqcup B_2$, while $B(\,\cdot\,)$ stands for the multivariate beta function (4.9). Examining (6.4), one can conclude that multiplying each spline by $\det(B_1) = d!\operatorname{vol}^{d}(\sigma_{B_1})$ makes the trace dependent only on the shared face σ_I and not on the originating simplex. Thus, the multivariate B-spline functions obtained through the sum (6.3) are continuous across the interfaces between supporting simplices. □

This corollary shows that the function basis of the finite element method, using Bernstein-Bézier polynomials, can also be found as a special case of our simplex spline space.

Basis construction, evaluation and quadratures

After ensuring that all the constraint facets are Gabriel, we can build the spline space $\mathcal{S}_h$ using the algorithm presented in Theorem 5.4.3, where every Delaunay triangulation is realized with the height function h. The fine-grained property of h is quite simple to achieve in practice. One can in fact simply assign the same height value to all coincident points, and then re-insert, in the triangulation of a link region $\mathcal{R}(Q)$, all the knots a_i, $i \in Q$, such that the multiplicity of the knot in Q is less than the corresponding multiplicity in A. This guarantees that the Delaunay triangulation of $\mathcal{R}(Q)$ includes these knots as vertices, which are then included in the subsets B of the resulting splines $M(x \mid Q \sqcup B)$, satisfying Lemma 6.2.3. This is equivalent to computing

the comparisons of height functions of coincident points according to the *symbolic perturbation* rule

$$h(a_i) > h(a_j) \Leftrightarrow i > j \text{ whenever } a_i = a_j. \tag{6.5}$$

For spline evaluation, we rely on the auxiliary functions constructed as explained in Section 5.4.3. Specifically, we start from the tiles $\Pi_{I,B}$ associated to spline functions of degree k, as constructed by the previous process, and we build the auxiliary functions and evaluation graph described in that section. We then construct an $R^\star$-tree [284] using all the degree-zero splines produced by this process. This search tree is then used to find all the degree-zero splines supported at a point, from which the evaluation graph can be followed up to degree k. Notice that these splines can overlap, and therefore the choice of a bounding volume hierarchy method like the $R^\star$-tree is appropriate.

When building a set of auxiliary functions for spline evaluation, the construction of the evaluation graphis done as suggested by Lemma 5.4.8, and the direction of the graph between two splines $M(x \mid I \sqcup B)$ and $M(x \mid I' \sqcup B')$ is determined using the signs of the determinants $\sigma_{bb'}$, σ_b and $\sigma_{b'}$ defined there, with $C \coloneqq B \cap B'$. Once again, one can easily include the fine-grained property of h in these evaluations. In fact, determining the signs of the determinant σ_{ij} appearing in (5.38) is equivalent to determining the position of the point $(a_j, h(a_j)) \in \mathbb{R}^{d+1}$ with respect to the plane $H \coloneqq \operatorname{aff}((a_c, h(a_c))_{c\in C}, (a_i, h(a_j))) \subset \mathbb{R}^{d+1}$. One can then simply apply the perturbation rule (6.5) to assign a consistent sign to σ_{ij} whenever a_j coincides with some other point in $(a_r)_{r\in C\sqcup\{i\}}$.

Finally, we need to spend a few words on the assembly of the mass matrix (2.15), stiffness matrix (2.16) and damping matrix (2.17). The integral

$$\int_\Omega M(x \mid I \sqcup B) M(x \mid I' \sqcup B') \,\mathrm{d}\Omega$$

is done by splitting the integration domain over a set of *cells* on which both spline functions are pure polynomials. Thus, the assembly of any matrix element can be split as follows,

$$\int_\Omega M(x \mid I \sqcup B) M(x \mid I' \sqcup B') \,\mathrm{d}\Omega = \sum_{i=1}^{n_c} \int_{C_i} M(x \mid I \sqcup B) M(x \mid I' \sqcup B') \,\mathrm{d}\Omega,$$

where $(C_i)_{i=1}^{n_c}$ is a family of n_c cells that form a subdivision of Ω and such that any spline of the space is a pure polynomial on C_i. We call this subdivision the *quadrature subdivision*. Using the recurrence formula (5.2b) and the validity of the evaluation scheme of Section 5.4.3, it is clear that these cells can be obtained as intersections of simplices $\operatorname{conv}((a_b)_{b\in B})$ associated to splines $M(x \mid I \sqcup B)$ (i.e., tiles $\Pi_{I,B} \in \mathcal{P}(h)$) of degree k or less. This can be constructed via Algorithm 3. Once the cells are computed, the integration of the product of two splines becomes simply the integral of a polynomial over each cell, which can be computed using standard algorithms.

One major drawback of this approach is that, since there are exactly $\sum_{r=0}^{k} \binom{r+d}{d} = \binom{k+d+1}{d}$ simplices supported over any point x, the number of cells in the quadrature subdivision can be very large. This issue is discussed in the last section of this chapter.

Algorithm 3 Construction of the quadrature cells for the assembly of system matrices.

$\mathcal{T} \leftarrow R^\star$-tree constructed on the simplices $\{\mathrm{conv}((a_b)_{b\in C}) : \Pi_{I,B} \in \mathcal{P}, |I| \leq k\}$
$x \leftarrow$ a point in the interior of Ω
$Q = \varnothing$ queue of seed points
$\mathcal{C} = \varnothing$ set of cells
PUSH(Q, x) insert x into Q
while Q is not empty **do**
    $x \leftarrow$POP(Q)
    $\{\Sigma_1, \ldots, \Sigma_m\} \leftarrow$QUERY$(\mathcal{T}, x)$ simplices supported at x
    $C \leftarrow \bigcap_{i=1}^m \Sigma_m$
    if C not empty and $\mathcal{C}$ does not contain C **then**
        insert C into $\mathcal{C}$
        for facet F of C **do**
            $f \leftarrow$ centroid of F
            $N_F \leftarrow$ outward normal of F, $|N_F| = 1$
            $x_f \leftarrow f + \varepsilon N_F$ for a small ε
            PUSH(Q, x_f)
return $\mathcal{C}$

6.3 Some numerical properties

We discuss in this section some numerical properties of the proposed multi-patch DG-IGA scheme that are useful for understanding the numerical behavior of the approach, especially in comparison to the closely related IPDG method.

6.3.1 Splines having a nonzero trace on constraint facets

Proposition 6.2.4 allows to characterize the spline functions having a nonzero trace on a facet, as follows.

Proposition 6.3.1. *In the hypotheses of Proposition 6.2.4, the splines on the positive side of F having a nonzero trace on F are exactly the Bernstein-Bézier polynomials having a nonzero trace on F and supported on a simplex adjacent to F.*

Proof. Due to the regularity of splines, in order for a spline function $M(x \mid I \sqcup B)$ of degree k to have a nonzero trace on F, it must have exactly $k + d$ knots on F. The knots $(a_b)_{b\in B}$ describe a non-degenerate simplex, and therefore they cannot all be on F. Thus, there is exactly one $b \in B$ such that $a_b \notin F$, and F is thus a facet of $\mathrm{conv}((a_b)_{b\in B})$. All the knots $(a_i)_{i\in I}$ are on F. Let us now prove that they are actually vertices of F.

Suppose that a knot $a_i \in F$ for some $i \in I$. Then, there exist d coefficients $\lambda_1, \ldots, \lambda_d$ satisfying $0 \leq \lambda_c \leq 1$, $\sum_{c\in C} \lambda_c = 1$ and $\sum_{c\in C} \lambda_c a_c = a_i$. If a_i is not a vertex of F, then one can multiply (6.2) by λ_c and sum over c, yielding $h(a_i) \geq \sum_{c\in C} \lambda_c h(a_c)$. This contradicts the fact that $i \in I$. In fact, if one takes any point a_j affinely independent of $(a_c)_{c\in C}$, and orders the set

$B \coloneqq C \sqcup \{j\}$ so that $\det(B) > 0$, the Delaunay condition (5.27) is satisfied, since

$$\begin{aligned}\det((a_j, h(a_j), 1)_{j\in B}, (a_i, h(a_i), 1)) &= \sum_{c\in C} \lambda_c \det((a_j, h(a_j), 1)_{j\in B}, (a_c, h(a_c), 1)) \\ &\quad + \det((a_j, h(a_j), 1)_{j\in B}, (0, h(a_i) - \sum_{c\in C} \lambda_c h(a_c), 0)), \\ &= -\left(h(a_i) - \sum_{c\in C} \lambda_c h(a_c)\right) \det(B) < 0,\end{aligned}$$

which implies that $i \notin I \sqcup B$. We conclude the proof by noticing that the point a_b (i.e. the only point with index $b \in B$ that does not lie on F) must be the same for all these splines, as it corresponds to the first knot encountered by the rotating plane used in the proofs of Proposition 6.2.4 and Corollary 6.2.6. Thus, all these splines are supported on the same simplex $\mathrm{conv}((a_b)_{b\in B})$. This completes the proof. □

Proposition 6.3.1 is very important, since it shows that the spline functions obtained when building the space $\mathcal{S}_h \coloneqq \bigoplus_{i=1}^{n_d} \mathcal{S}_i$ have exactly the same trace on the constraint facets $F \in \mathcal{F}$ as the usual discontinuous Galerkin Bernstein-Bézier basis functions. Indeed, splines satisfying the proposition are simply Bernstein-Bézier polynomials over the same simplex $\mathrm{conv}((a_b)_{b\in B})$. The corresponding simplices are shown in Figure 6.3, bottom left. This property also implies that the sets of splines having a nonzero trace on any constraint facet are the same as in the case of the Bernstein-Bézier discontinuous Galerkin. Therefore, the usual inverse inequalities from [305], i.e.,

$$\int_F f^2 \,\mathrm{d}F \leq \frac{(k+1)(k+d)}{d} \frac{\mathrm{vol}^{d-1}(F)}{\mathrm{vol}^d(\Sigma)} \int_\Sigma f^2 \,\mathrm{d}\Sigma, \tag{6.6}$$

which are valid for any polynomial function f of degree k over the simplex Σ having F as one of its facets, can be directly used for any constraint facet. This allows to apply in an extremely straightforward way many of the proofs of the numerical properties of the IPDG method. We show two such instances in the next section.

6.3.2 Positivity of the bilinear form

As discussed in Chapter 2, the IPDG method relies on a penalty term that makes its bilinear form positive. This is necessary to ensure the stability of the method. However, a large penalty term can adversely impact the performance of the numerical scheme. More specifically, the condition number of the bilinear form scales linearly in the penalty constant (see, e.g., [306]), which affects the maximum allowed timestep as well as the precision and numerical stability of the method as the timestepping iterations increase. Thus, the determination of a general criterion for choosing a reasonably small but effective penalty coefficient is crucial.

In most applications, a different penalty term is associated with each facet of the mesh. Some early works tie the value of the penalization coefficient to the size of the facet [307] or the size of the adjacent elements [308]. A sufficient value for the penalty term, which uses the ratio between these two quantities and does not contain undetermined constant factors, was given in [125]. For simplicial cells, the resulting value can be equivalently described in terms of the

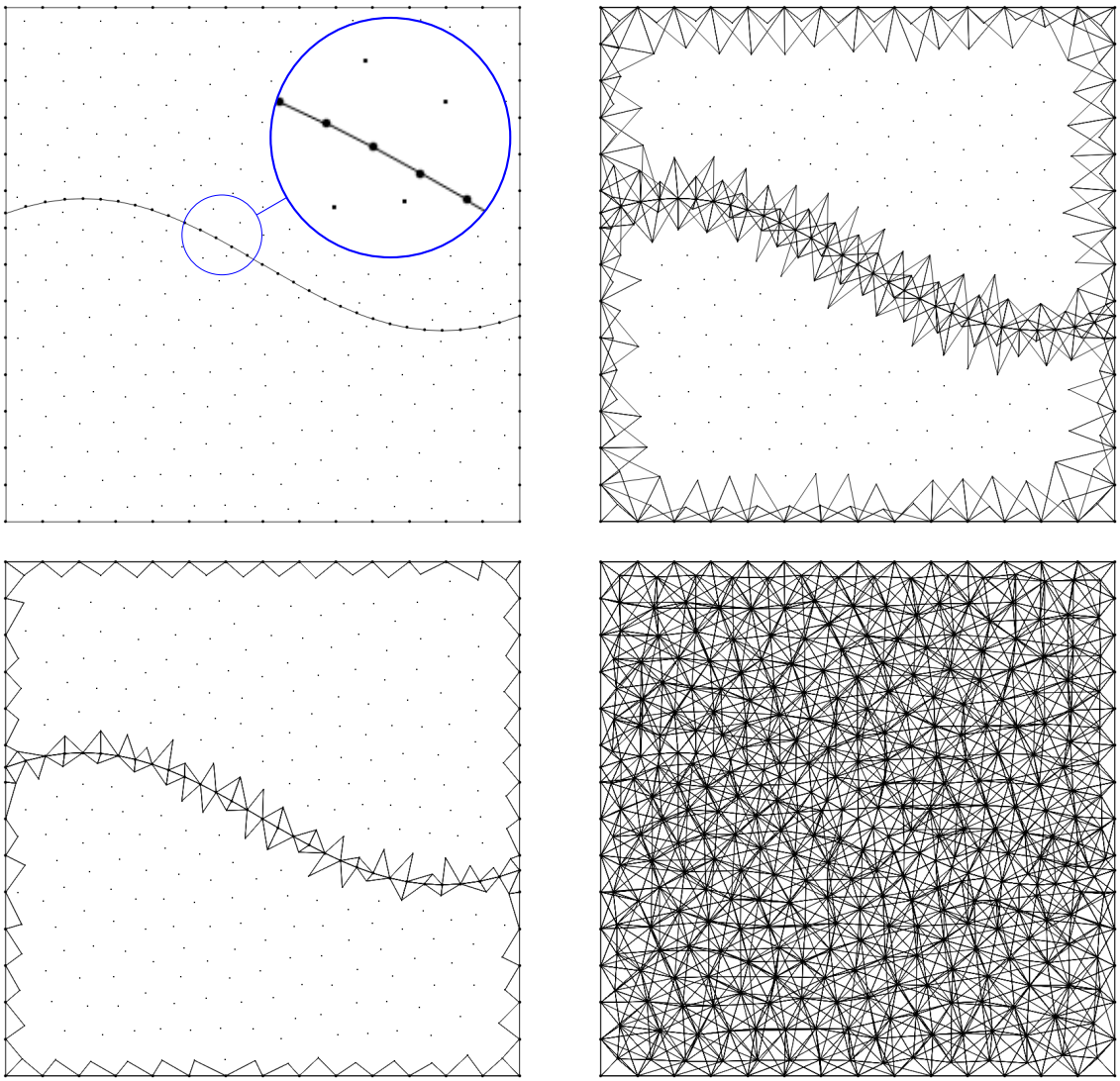


Figure 6.3: (Top left) a point configuration $A \subset \mathbb{R}^2$, with two domains. We use A to build a spline space of degree $k = 2$. The Gabriel facets are shown in the picture. Points on these facets are repeated $2+1$ times. (Top right) the simplices associated to the splines of Proposition 6.2.4, that protect the boundaries and allow the decomposition of the spline space. (Bottom left) the simplices associated to the splines of Proposition 6.3.1, corresponding to splines that have a nonzero trace on constraint facets. (Bottom right) the simplices associated to all splines of degree ≤ 2. Their intersection determines the quadrature decomposition.

radius of the inscribed sphere, see, e.g., [89]. In [126], lower bounds on the penalty term are related to the shape of the neighboring cells.

The weak form of the problem (6.1) is expressed using the bilinear form

$$a(\varphi,p) \coloneqq \sum_{i=1}^{n_d} \int_{\Omega_i} \frac{1}{\rho}\nabla\varphi\cdot\nabla p \,\mathrm{d}\Omega - \sum_{F\in\mathcal{F}_{\mathrm{DG}}} \int_F \left([\![\varphi]\!]\cdot \{\!\{\frac{1}{\rho}\nabla p\}\!\} + [\![p]\!]\cdot\{\!\{\frac{1}{\rho}\nabla\varphi\}\!\}\right) \mathrm{d}F + \sum_{F\in\mathcal{F}_{\mathrm{DG}}} \alpha(F)\int_F [\![\varphi]\!]\cdot[\![p]\!]\,\mathrm{d}F, \tag{6.7}$$

and we need to determine a sufficiently large, but reasonably small, value for α that makes the form (6.7) positive. In other words, after defining

$$\|p\|_h^2 \coloneqq \sum_{i=1}^{n_d} \int_{\Omega_i} \frac{1}{\rho}\,|\nabla p|^2 \,\mathrm{d}\Omega + \sum_{F\in\mathcal{F}_{\mathrm{DG}}} \int_F [\![p]\!]^2\,\mathrm{d}F,$$

one requires

$$a(p,p) = \sum_{i=1}^{n_d} \int_{\Omega_i} \frac{1}{\rho}\,|\nabla p|^2 \,\mathrm{d}\Omega - 2\sum_{F\in\mathcal{F}_{\mathrm{DG}}} \int_F [\![p]\!]\cdot\{\!\{\frac{1}{\rho}\nabla p\}\!\}\,\mathrm{d}F + \alpha(F)\sum_{F\in\mathcal{F}_{\mathrm{DG}}}\int_F [\![p]\!]^2\,\mathrm{d}F \geq C\,\|p\|_h$$

for all p in our multi-patch spline space $\mathcal{S}_h \coloneqq \bigoplus_{i=1}^{n_d}\mathcal{S}_i$, where the constant C does not depend on p.

Thanks to Proposition 6.3.1, the results of [125] can be directly applied to our method. Specifically, in the cited paper, the authors introduce a penalty term $\alpha(F)$ on each facet F of the mesh, and use the inverse inequality (6.6) on the simplices adjacent to F to determine a value of $\alpha(F)$ that guarantees the positivity of a. We can apply exactly their reasoning to our case, the only exception being that for our method, the penalty term $\alpha(F)$ is only computed on the facets $F\in\mathcal{F}_{\mathrm{DG}}$, and that we have to take into account the density values. Since we use a piecewise-constant density value over each domain, we can easily adapt the expression (7) of [125] and choose, for $F\in\mathcal{F}_{\mathrm{DG}}$,

$$\alpha(F) = \frac{(k+1)(k+d)}{2}\frac{\max(\frac{1}{\rho^+},\frac{1}{\rho^-})}{\min(r(\Sigma_{B^+}),r(\Sigma_{B^-}))}, \tag{6.8}$$

where the subscripts $+$ and $-$ identify the two sides of F, $M(x \mid I^\pm \sqcup B^\pm)$ are two splines with a nonzero trace on F, one on each side of F, $\Sigma_{B^\pm} \coloneqq \mathrm{conv}((a_b)_{b\in B^\pm})$ and $r(\Sigma_{B^\pm})$ is the inradius (i.e., the radius of the inscribed sphere) of $\Sigma_{B^\pm}$. In practical cases, one often chooses a larger penalty term to ensure the good numerical conditioning of the system matrices.

6.3.3 *A priori* error analysis

The applicability of the inverse inequality (6.6), guaranteed by Proposition 6.3.1, also allows to directly apply the *a priori* error estimates of the IPDG method, presented in Chapter 2, with very minimal modifications.

As we have seen in Chapter 2, the proof that the *a priori* error of the IPDG method is the same as the usual continuous Galerkin approach relies on an estimate of the residual

$$r(p,\eta_h) \coloneqq \hat{a}(\eta_h,p) - a(\eta_h,p) = -\sum_{F\in\mathcal{F}_{\mathrm{DG}}}\int_F [\![\eta_h]\!]\cdot \{\!\{\Pi_{\mathrm{DG}}\frac{1}{\rho}\nabla p - \frac{1}{\rho}\nabla p\}\!\}\,\mathrm{d}F.$$

We bound this residual in the energy semi-norm

$$|||p|||^2 \coloneqq a(p,p) = \int_\Omega |\nabla p|^2\,\mathrm{d}\Omega + \sum_{F\in\mathcal{F}_D G}\int_F \alpha[\![p]\!]^2\,\mathrm{d}F.$$

We compute

$$\sup_{\eta_h\in\mathcal{S}_h}\frac{|r(p,\eta_h)|}{|||\eta_h|||}$$

for $p\in H^1(\Omega)+\mathcal{S}_h$.

Using the inverse inequality (6.6) and the Cauchy-Schwartz inequality,

$$|r(p,\eta_h)|^2 \le C_1\left(\sum_{F\in\mathcal{F}_I}\int_F \alpha[\![\eta_h]\!]^2\,\mathrm{d}F\right)\left(\sum_{\substack{F\in\mathcal{F}_{\mathrm{DG}},\\ \Sigma:\partial\Sigma\supset F}}\frac{\mathrm{vol}^{d-1}(F)}{\mathrm{vol}^d(\Sigma)}\int_\Sigma \alpha^{-1}\left|\{\!\{\Pi_{\mathrm{DG}}\frac{1}{\rho}\nabla p - \frac{1}{\rho}\nabla p\}\!\}\right|^2\,\mathrm{d}F\right),$$

$$\begin{aligned}
&\le C_2|||\eta_h|||^2 h^{-1}\int_\Omega\left(\alpha^{-1}\left|\Pi_{\mathrm{DG}}\frac{1}{\rho}\nabla p\right|^2 + \left|\frac{1}{\rho}\nabla p\right|^2\right)\mathrm{d}\Omega,\\
&\le C_2|||\eta_h|||^2|||p|||^2.
\end{aligned}$$

We have used here the following facts: the geometric term $\mathrm{vol}^{d-1}(F)/\mathrm{vol}^d(\Sigma)$ scales like h^{-1}, the penalization coefficient $\alpha(F)$ (6.8) is proportional to the radius of the inscribed sphere of the simplices incident to F and thus also scales like h^{-1}, and the projection Π_{DG} is done in the $L^2(\Omega)$ norm and is thus stable in this norm. Thus,

$$\sup_{\eta_h\in\mathcal{S}_h}\frac{|r(p,\eta_h)|}{|||\eta_h|||} \le C|||p|||,$$

which is the same *a priori* estimate as the IPDG method [108].

6.4 Some numerical results

We present in this section a few results showcasing the capabilities of this approach.

6.4.1 Block-diagonal mass matrix

In Figure 6.4 we show three different choices for the constrained facets (and thus the repeated knots) on the same point configuration A. We compare a meshed point configuration where each mesh element is a subdomain and all points are repeated $k+1$ times, a multi-patch DG-IGA

approach with 6 subdomains, and a pure IGA approach, i.e., a single subdomain. Notice that the pure DG case is obtained as in Corollary 6.2.6, and the pure IGA case is obtained by only constraining the facets on $\partial\Omega$. Consequently, all three numerical schemes are obtained through our construction, with pure DG and pure IGA obtained as the limiting cases. This is reflected in the sparsity pattern of the mass matrix (Figure 6.4, right), which is always block-diagonal, but where one can arbitrarily vary the size and number of diagonal blocks from one per mesh element to a single block for the whole domain. Notice also that the blocks are sparse (except near the DG limit) and have a limited bandwidth, comparable with the usual DG case (cf. Proposition (5.3.12)). The pure IGA ad DG-IGA cases have a very similar number of nonzero entries, $\sim 3.5 \cdot 10^5$, despite their rather different appearance.

From an implementation point of view, this flexibility can allow a suitable distribution of computational tasks (load balancing). Ideally, the blocks can be made small enough so that a single computational node is capable of storing the factorization of the mass matrix corresponding to a single subdomain.

6.4.2 Validation

We have first validated the method on a simple two-dimensional homogeneous model Ω of size 9.2km $\times$ 3.0km with $\rho = 1000\mathrm{kg\,m^{-3}}$, $c = 1500\mathrm{m\,s^{-1}}$, absorbing boundary conditions on all sides, a single source point and an array of receivers. We have used for the source, as in all other results of this section, a Ricker wavelet (1.6) in time. We have compared the values at the receivers with the analytical result, which we have computed using the Gar6more software [309]. We have computed the L^2 error of the simulation at the position $x_r \in \mathbb{R}^d$, $d = 2, 3$ of a receiver r as

$$e_r^2 \coloneqq \frac{\int_0^{t_f} (p(x_r, t) - p_{\mathrm{A}}(x_r, t))^2 \,\mathrm{d}t}{\int_0^{t_f} p_{\mathrm{A}(x_r,t)}^2 \,\mathrm{d}t}, \tag{6.9}$$

where p_{A} is the analytical solution computed with Gar6more. The point configuration A includes around $3.4 \cdot 10^4$ points. We have performed an IGA simulation, defining a single subdomain and increasing the multiplicity of points on $\partial\Omega$, a finite element method (FEM) simulation obtained by repeating $k+1$ times all the points in A (Corollary 6.2.7), and a DG simulation obtained from this by additionally defining a subdomain for each simplex in a triangulation of Ω (Corollary 6.2.6). The simulation was repeated with k ranging from 1 to 4. In Figure 6.5, we show a snapshot from the simulation, and we compare the CFL timestep for the LF2 time integration scheme (2.3), the number of degrees of freedom (i.e., the number of multivariate B-splines in the basis), the relative error (6.9) and the error times the number of degrees of freedom, which represents the inverse of the precision per degree of freedom.

As can be seen, multivariate spline spaces share many properties with their more usual tensor-product counterparts. Specifically, the number of degrees of freedom increases only linearly with the order k (due to the fact that no new nodes are inserted, cf. discussion at the end of Chapter 3), and the CFL condition only decreases as h/k instead of h/k^2 as in the case of FEM and DG methods. The precision per number of degrees of freedom is comparable for FEM and IGA, and has the same behavior as a function of k in all three methods.

Notice that, for $k = 1$, FEM and IGA coincide, and that at all orders the DG simulation

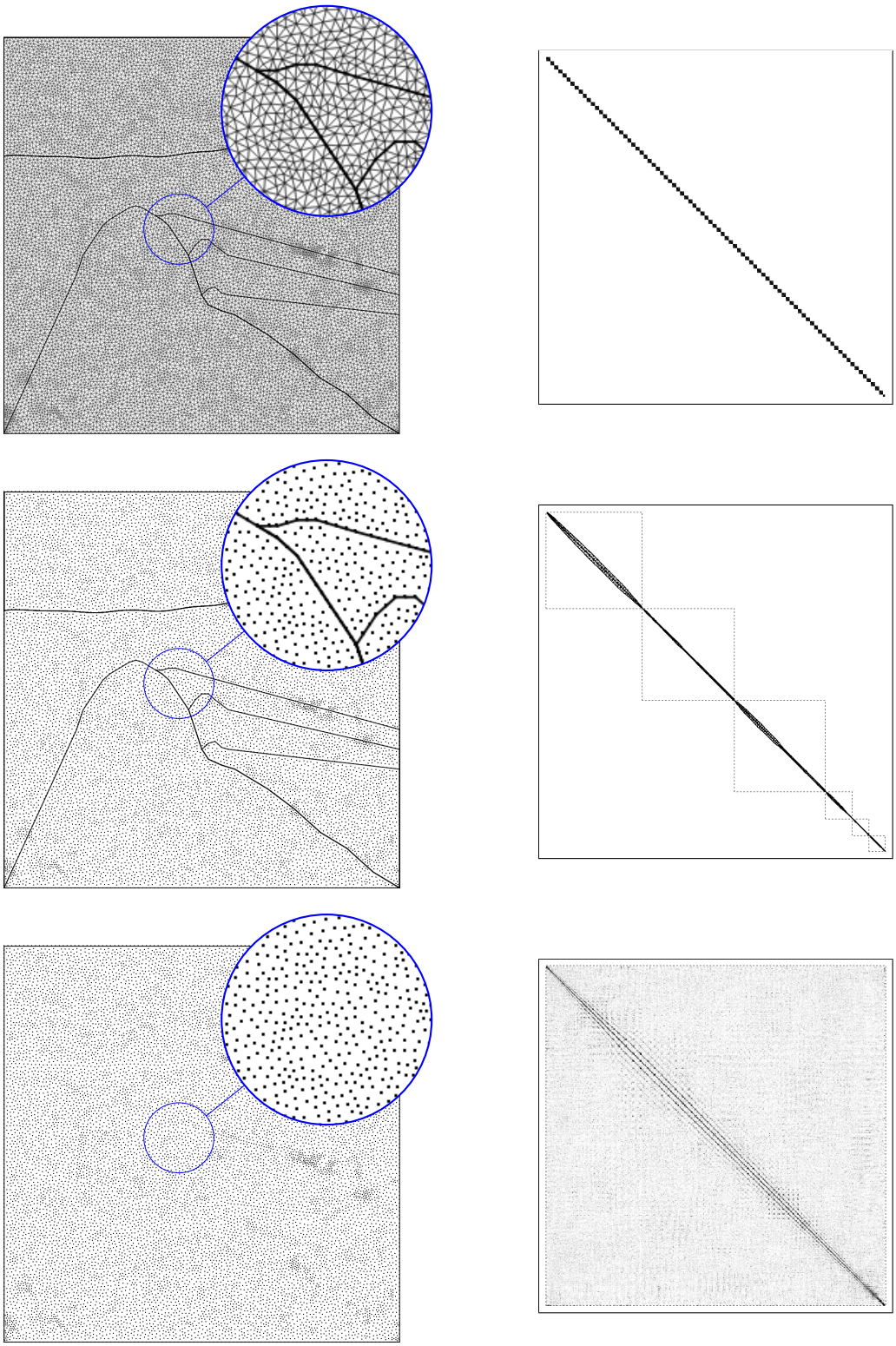


Figure 6.4: Three choices of constraints for a point configuration A (left), and the resulting sparsity pattern of the mass matrix (right). (Top) DG approach, with $n = 27954$ simplices. (Middle) multi-patch DG-IGA with 6 subdomains. (Bottom) pure IGA with a single subdomain.

is penalized by the excessive number of degrees of freedom with respect to both FEM and IGA, which is a well-known drawback of the method. Also notice that the penalization term α used in Figure 6.5 for the DG calculation is twice the minimum necessary for positivity, which further penalizes the maximum allowable timestep of the DG calculation (although it makes the simulation very stable).

In order to investigate whether the introduction of subdomain boundaries impacts the behavior of the CFL condition, we have performed a similar calculation on a simple two-dimensional bi-layered model, with the same dimensions as the two-dimensional homogeneous model above, but split horizontally into halves. The finit element simulation in this case is to be interpreted as a multi-patch FEM calculation, where one FEM basis is introduced in each subdomain, and DG fluxes are used to couple subdomains.

The two media have the same density $\rho = 1000\text{kg}\,\text{m}^{-3}$, but the medium containing the source has a velocity of $c = 1500\text{m}\,\text{s}^{-1}$, while the second medium has a velocity of $c = 2500\text{m}\,\text{s}^{-1}$. The point configuration contains about $2.3 \cdot 10^4$ points, of which about $1.7 \cdot 10^4$ are in the region of lower velocity, since the point density was adapted to the local seismic wavelength. In this simulation, we have computed the same quantities as in the homogeneous case. As one can see, the presence of an interface does in fact penalize the CFL maximum timestep of all the methods. The multi-patch DG-IGA simulation still achieves, however, the best timestep.

In general, as noticed in [3], the superior CFL condition timestep seen in multi-patch methods is due to the increased support of the basis functions. The same behavior is experienced when the usual C^0 finite element bases are modified to have a larger support [310]. In particular, we expect this behavior to be present as long as the number of functions in each patch of degree k is much larger than $\binom{k+d}{k}$.

6.4.3 Multi-patch simulation and blending with DG

We have tested the multi-patch DG-IGA approach on a simple two-dimensional 3km $\times$ 3km synthetic seismic model, the same used in [169]. This model, shown in Figure 6.7, is composed of 6 layers, including water on the top and a salt body in the interior.

We have performed three simulations. The first two simulations correspond to a multi-patch DG-IGA model with 6 domains, shown in Figure 6.7, and a hybrid simulation whereby four domains are treated via the IGA approach, and two domains are meshed and simulated using the DG basis obtained via Corollary 6.2.6. The third simulation is also performed in this hybrid configuration, but with the same physical medium in all the subdomains, in order to show that no numerical artifacts appear at the boundaries between two different numerical schemes. Results are shown in Figure 6.10.

6.4.4 Non-simply-connected domains

The usual tensor-product spline spaces used to define multivariate spline functions are limited to a simple topology, namely, that of a topological sphere. Obtaining a non-simply-connected domain then requires gluing together multiple patches. This is a tedious and sometimes very difficult step that often results in reduced regularity along seam lines. Instead, the approach proposed in this chapter allows to perform full IGA simulations on a non-simply connected

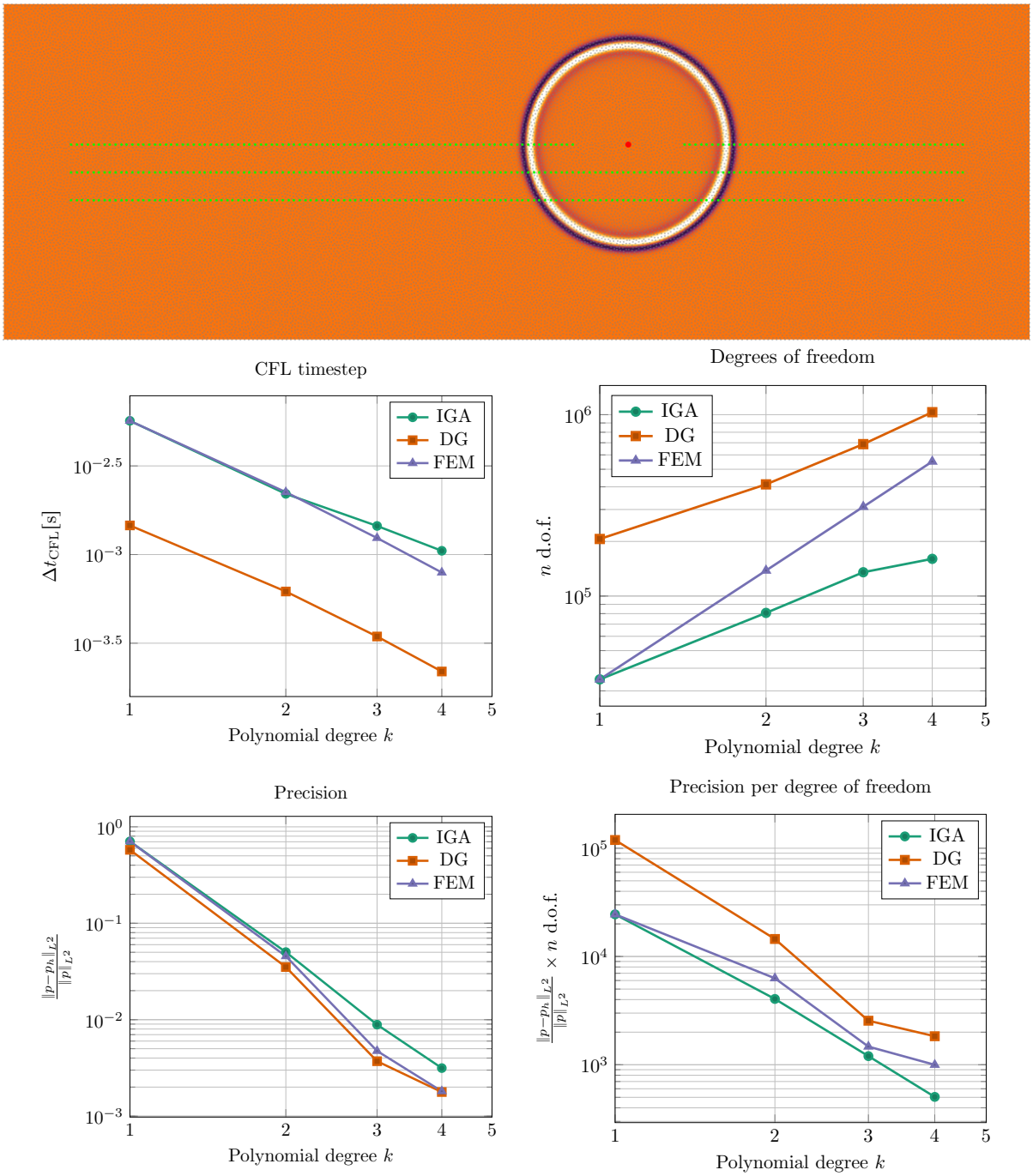


Figure 6.5: A two-dimensional homogeneous simulation (top) comparing, for IGA, FEM and DG simulations and for $k = 1, \ldots, 4$, the CFL timestep condition (middle left), the number of degrees of freedom (middle right), the relative error (bottom left) and the relative error times the number of degrees of freedom (bottom right).

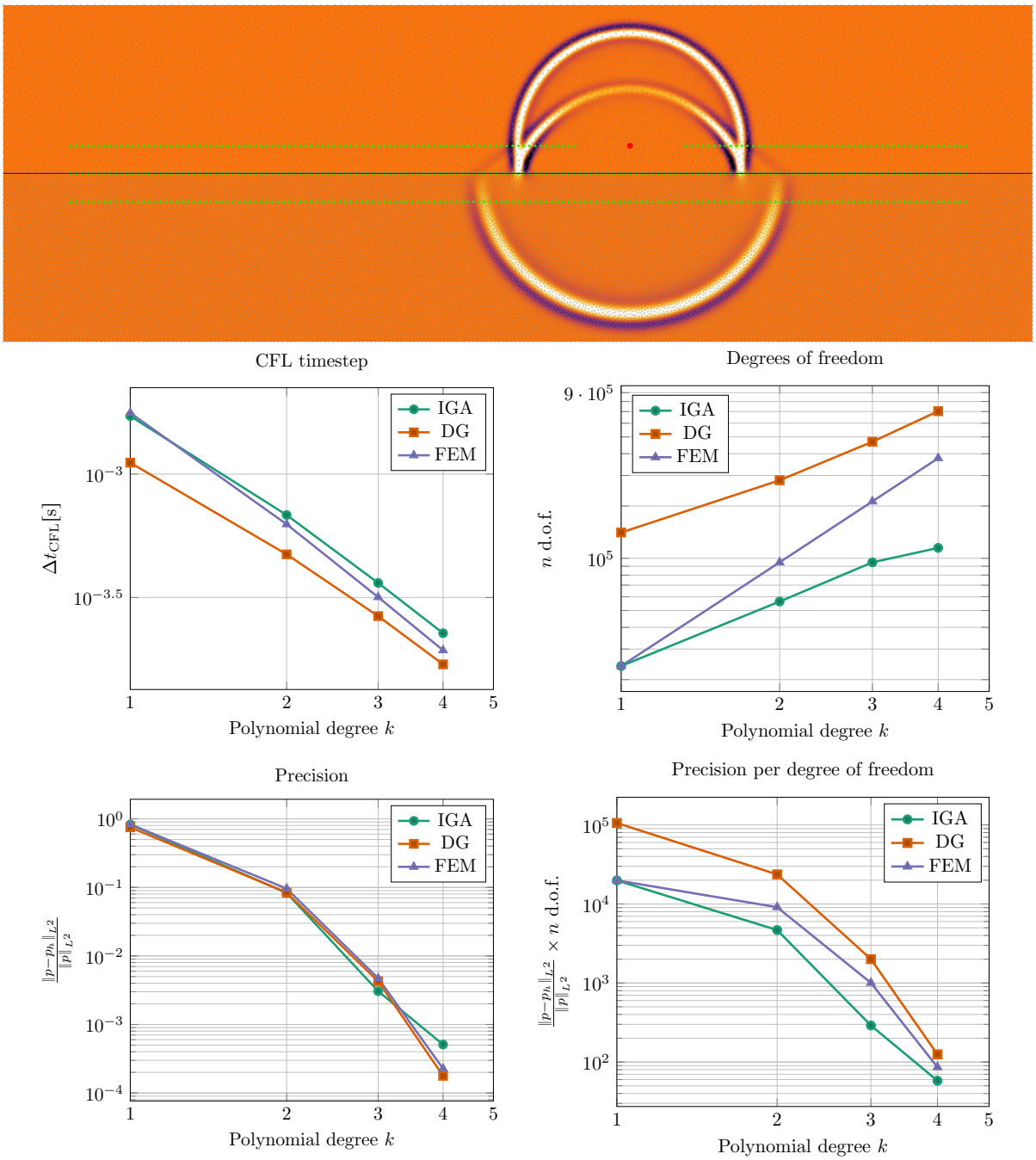


Figure 6.6: A two-dimensional bi-layered simulation (top) comparing, for multi-patch IGA, multi-patch FEM and DG simulations and for $k = 1, \ldots, 4$, the same quantities as Figure 6.5.

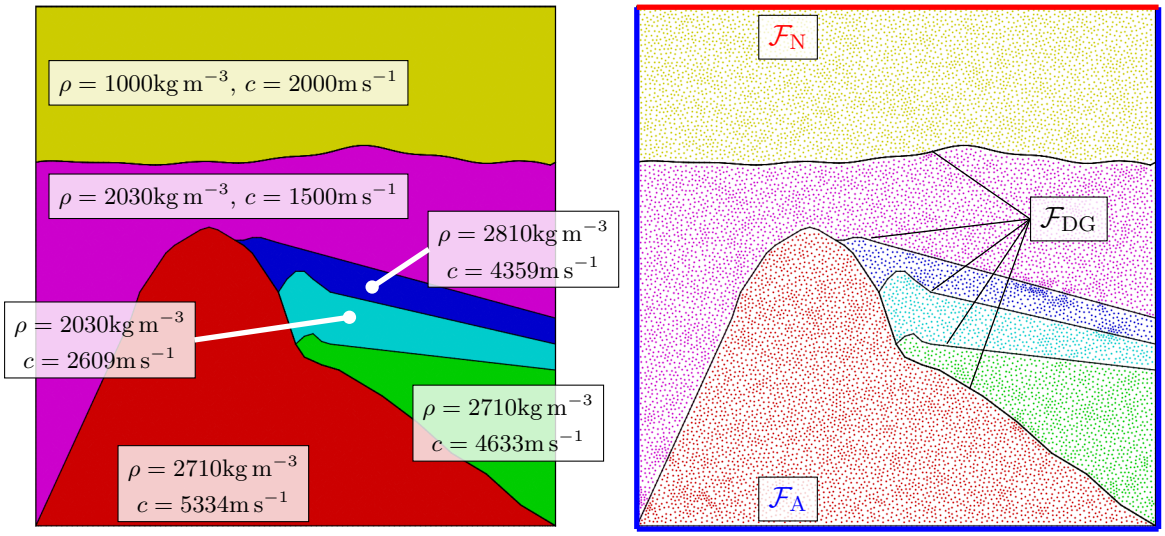


Figure 6.7: The synthetic model used for testing our multi-patch DG-IGA approach, with its associated physical parameters (left) and the point configuration used to construct our splines and DG basis (right). We also show the location of the absorbing ($\mathcal{F}_\text{A}$), free-surface ($\mathcal{F}_\text{N}$) and DG ($\mathcal{F}_\text{DG}$) boundaries.

domain, simply by placing absorbing (or other) boundary conditions on the internal boundaries, and excluding the subdomains representing holes from the simulation.

We show here three examples, all in two dimensions, of this feature. In Figure 6.11, we show a simple model inspired by helioseismology applications, with three domains, one with genus zero and two with genus one. The model was adapted from [311], and the spline space was built on a point configuration containing around 6400 points. Notice that the density of points has been adapted to the local wavelength. The simulation has degree $k = 3$.

In Figure 6.12 we present a simulation of a very simplified acoustic instrument, consisting of a single domain of genus three. We impose free-surface boundary conditions on the sides of the instrument, absorbing boundary conditions at the beginning of the embuch, and transparent (i.e., DG flux) boundary conditions at the exit of the bell. The whole model has approximately 4200 points. The simulation has degree $k = 4$.

In Figure 6.13 we show a simple application to the propagation of acoustic waves in closed spaces, by simulating a two-dimensional model of the interior of a church, namely, the Santa Croce basilica in Florence, Italy. The large amount of columns and other obstacles increase the genus of the simulation domain to 99, which would be extremely difficult to obtain by gluing together tensor-product spline patches. Using unstructured spline spaces, the regularity of the space is kept maximal (i.e., $k - 1$ at degree k) inside the domain. The model comprises around $1.1 \cdot 10^4$ distinct points. The simulation has degree $k = 2$.

Finally, in Figure 6.14 we show how different kinds of boundary conditions can be imposed on a constraint internal to a domain. We have simulated a simple homogeneous domain with a single fault inside, on which we have imposed absorbing, Dirichlet, Neumann and transparent (e.g, DG flux) conditions. The latter simulation illustrates the lack of numerical noise due to

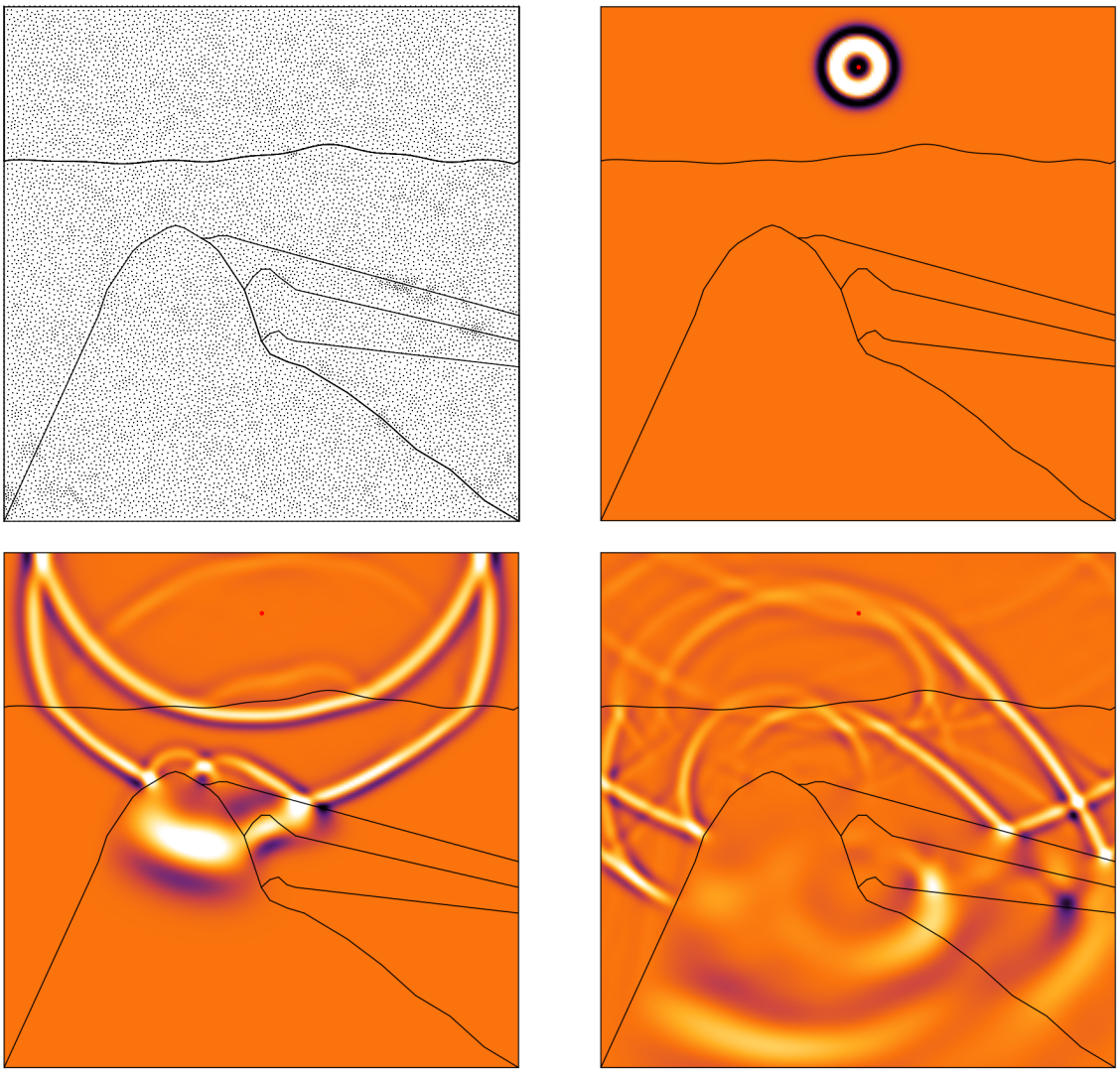


Figure 6.8: Simulation of the model of Figure 6.7 using a multi-patch IGA approach based on 6 domains. The spline space used has degree $k = 3$.

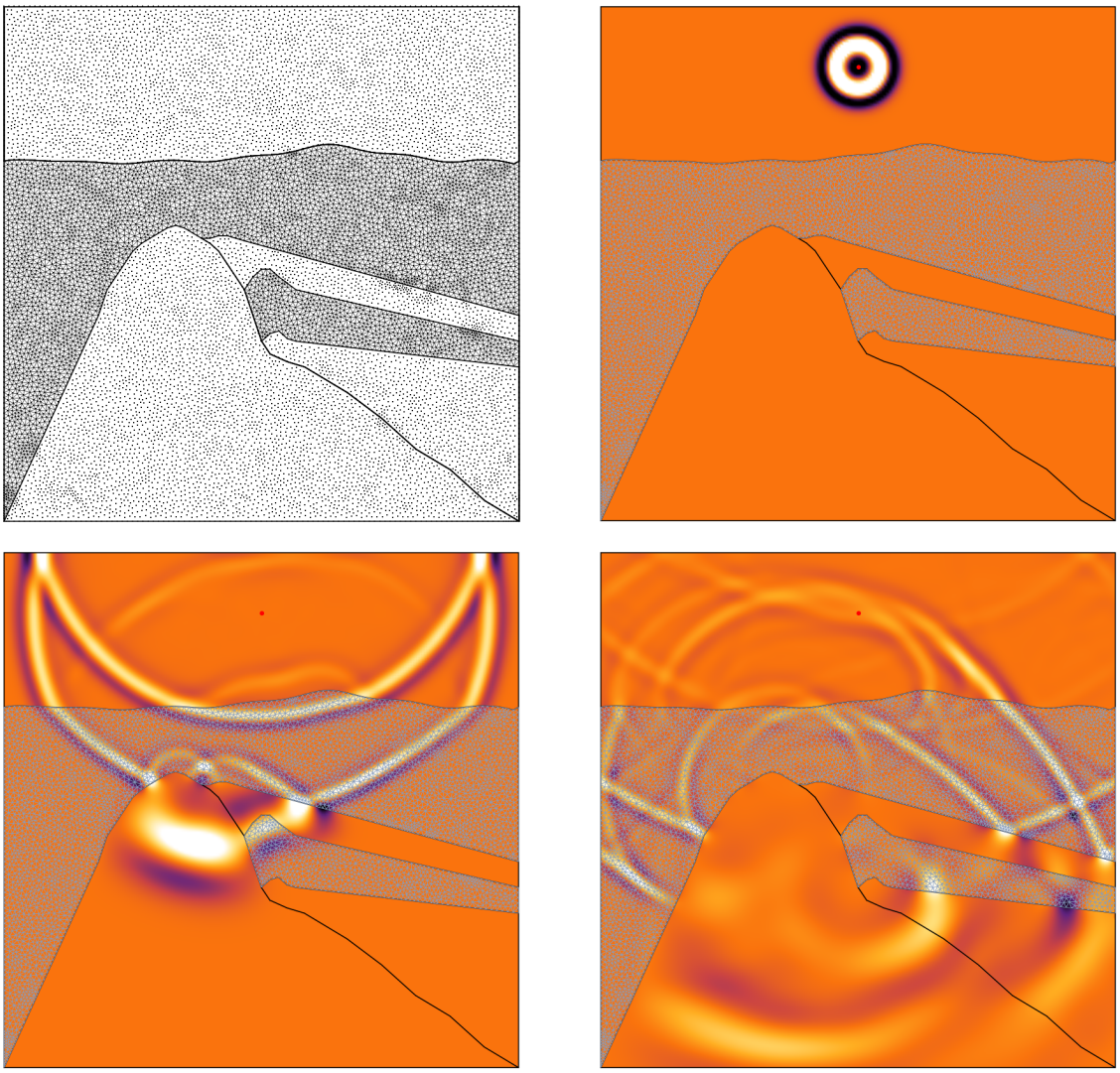


Figure 6.9: Simulation of the model of Figure 6.7 using a hybrid multi-patch DG-IGA approach based on 2 meshed (DG) domains and 4 meshless (IGA) domains. The spline space used has degree $k = 3$.

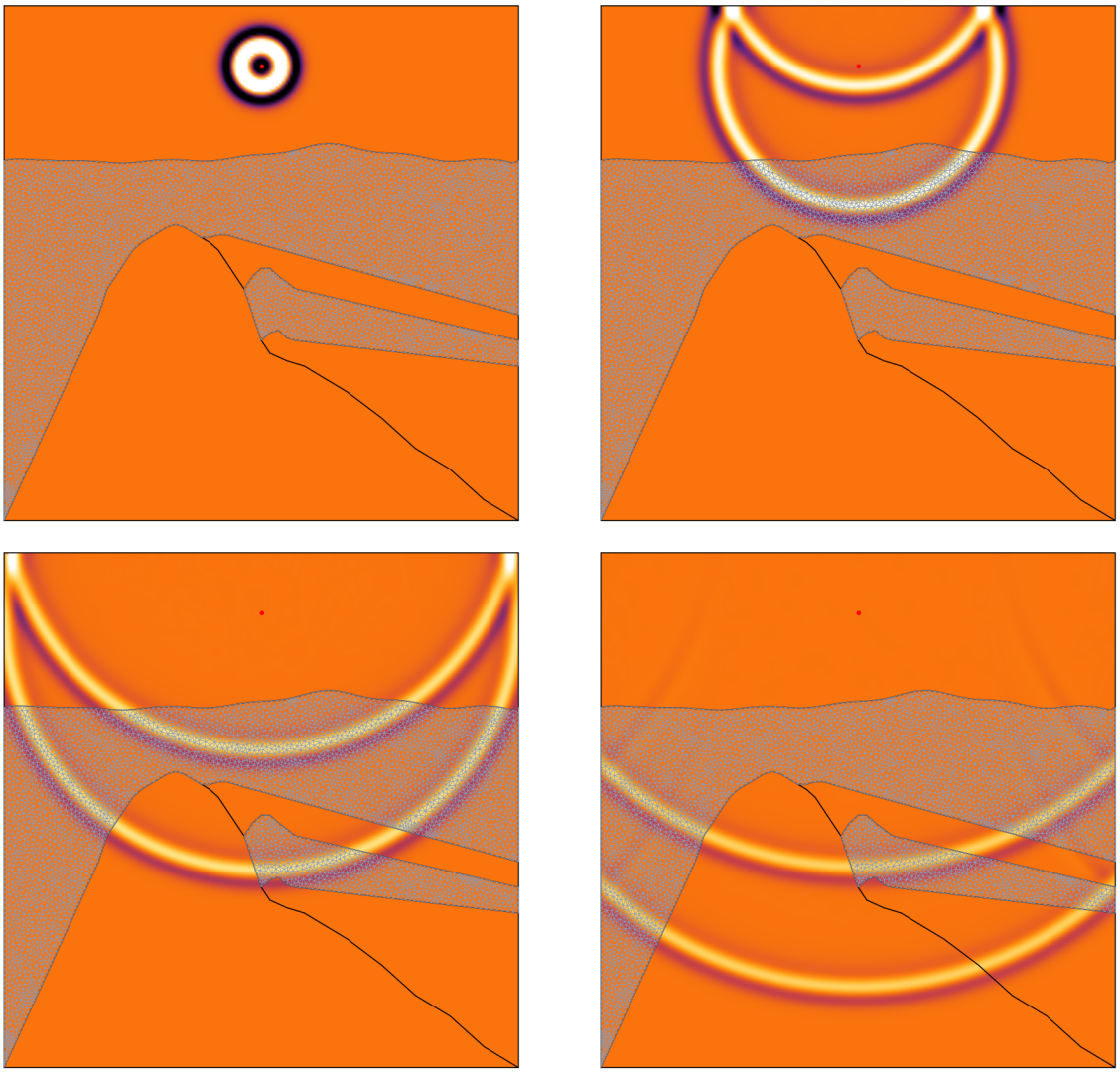


Figure 6.10: Simulation of the model of Figure 6.7 using a hybrid multi-patch DG-IGA approach based on 2 meshed (DG) domains and 4 meshless (IGA) domains. The physical parameters have been set to the same values in all domains, in order to detect numerical artifacts. The spline space used has degree $k = 3$.

the coupling of DG fluxes with the underlying IGA simulation scheme. This simple model is made of around 5000 points. In all cases, the simulation has degree $k = 2$.

We have performed a few numerical simulations using some simple three-dimensional models, with the goal of testing the capabilities of the method in three-dimensions and proving the feasibility of the quadrature algorithm 3. In practical cases, the algorithm is capable of correctly computing the quadrature cells, which has been determined by inspection of the wave equation solution. However, for $k > 3$, in three dimensions it seems that the computational cost of determining the quadrature cells and subsequent matrix assembly tends to become the dominant cost of the whole simulation. For this reason, we believe that a different approach for the computation of quadratures is required. We defer this investigation to a future work, and discuss briefly some possible solutions in the concluding section.

We show in Figure 6.15 the results of the simulation of a simple three-dimensional 1km $\times$ 1km $\times$ 1km bi-layered domain with density $\rho = 1000\text{kg}\,\text{m}^{-3}$ everywhere and velocities $c_1 = 2000\text{m}\,\text{s}^{-1}$ in the upper half and $c_2 = 3000\text{m}\,\text{s}^{-1}$ in the lower half of the model. The simulation was performed at $k = 2$ on the point configuration shown in the picture. The model comprises around $2.6 \cdot 10^4$ points. No significant numerical noise was detected.

6.4.5 A simple application to hyperelasticity

We have also briefly explored the applicability of our approach to hyperelasticity. In particular, we have simulated the linear elastic deformation of a simple two-dimensional gasket that undergoes a large deformation induced by the surrounding constraints. For each position of the constraints, we have computed the displacement field associated to the new equilibrium configuration, and we have applied the corresponding displacement to the original point locations. Our simulation can therefore be considered as quasi-static. No dynamic or other time-dependent quantity was computed, and no friction was added to the simulation, which makes it somewhat unrealistic. We have used a value of 0.3 for the Poisson's ratio of the material. Results are shown in Figure 6.16.

In order to perform hyperelasticity calculations via meshed methods, one usually needs to design a special mesh capable of accommodating the large deformation throughout the simulation without undergoing excessive mesh element degeneration. This is a costly pre-processing step that requires many exploratory simulations and can lead to an over-refined mesh. Thus, for these applications, meshless methods (or almost meshless methods such as the one proposed in this work) might be helpful.

6.5 Discussion and further reading

One of the most interesting features of the proposed method, in our view, is that it provides a natural bridge between DG and IGA methods, which can both be recovered as special cases of our construction. Numerically, this translates into the possibility of retaining the block-diagonal structure of the DG approach (cf. Figure 6.4) while improving the CFL condition thanks to the well-behaved shape of spline functions (cf. Figure 6.5). Furthermore, the choice can be made locally at the level of a sub-domain, providing a natural way to couple the DG and (unstructured)

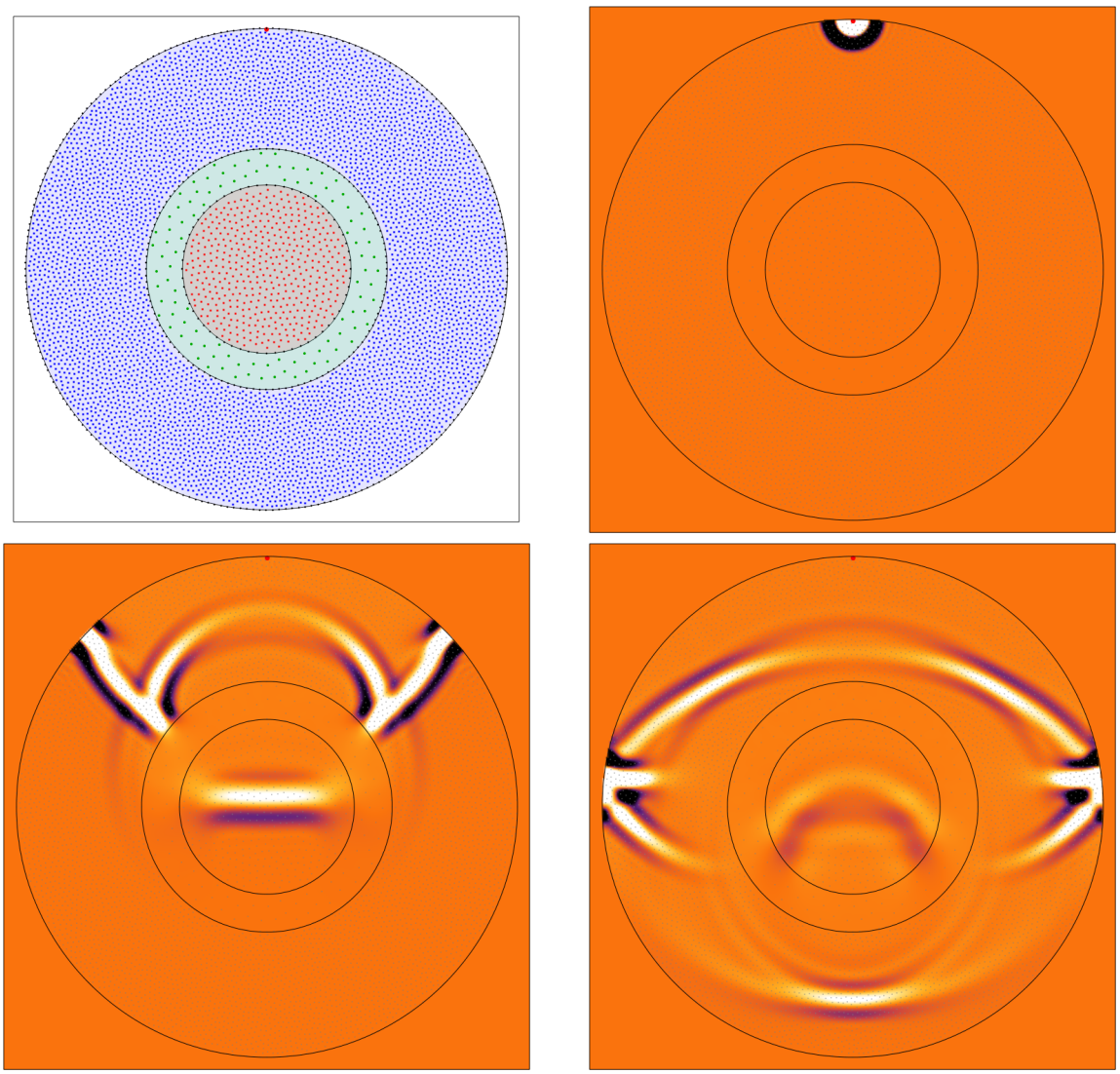


Figure 6.11: Simulation of a helioseismology-inspired model comprising three domains, two of which are not simply-connected.

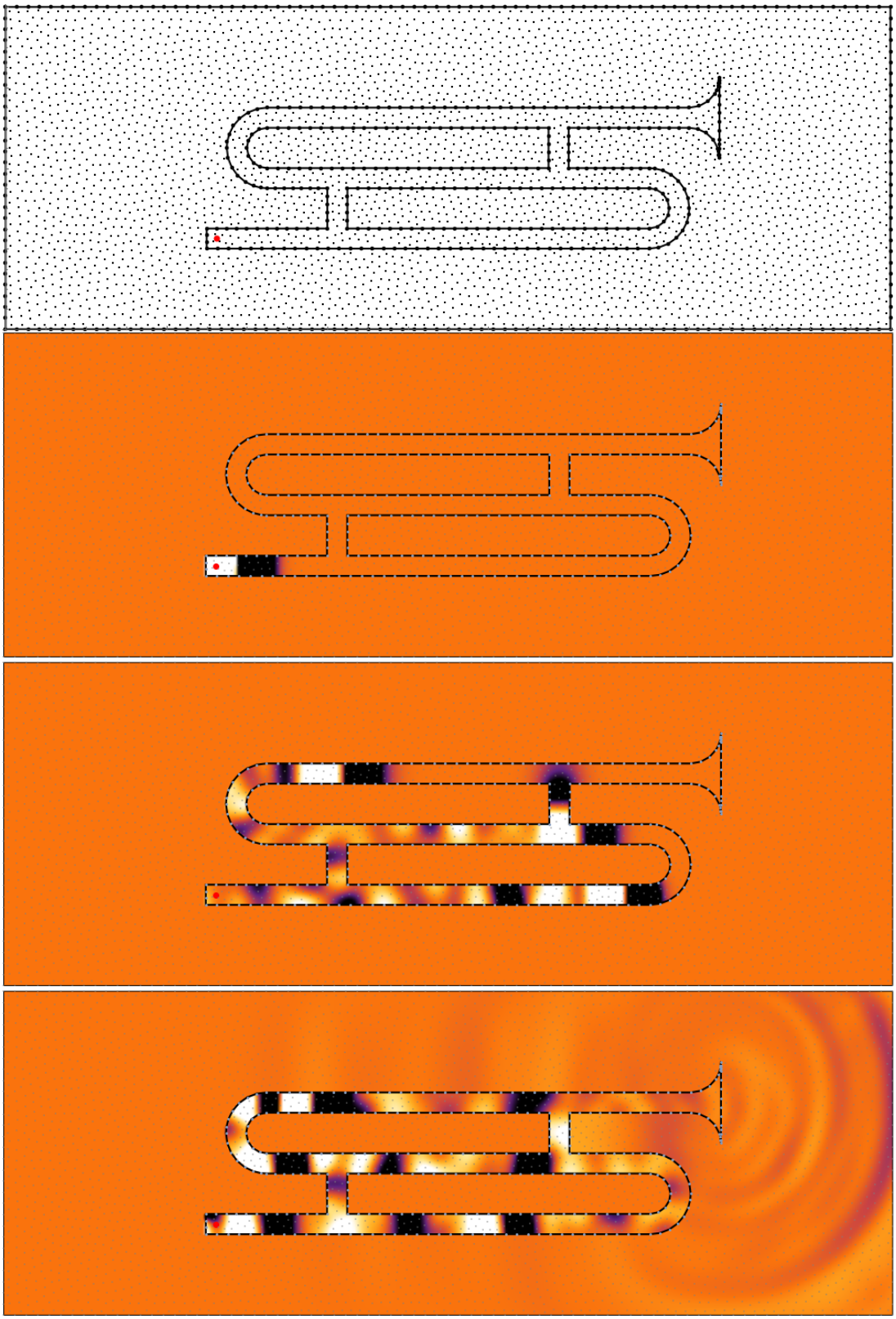


Figure 6.12: Simulation of a simple music instrument (a trumpet), comprising a single domain with genus three.

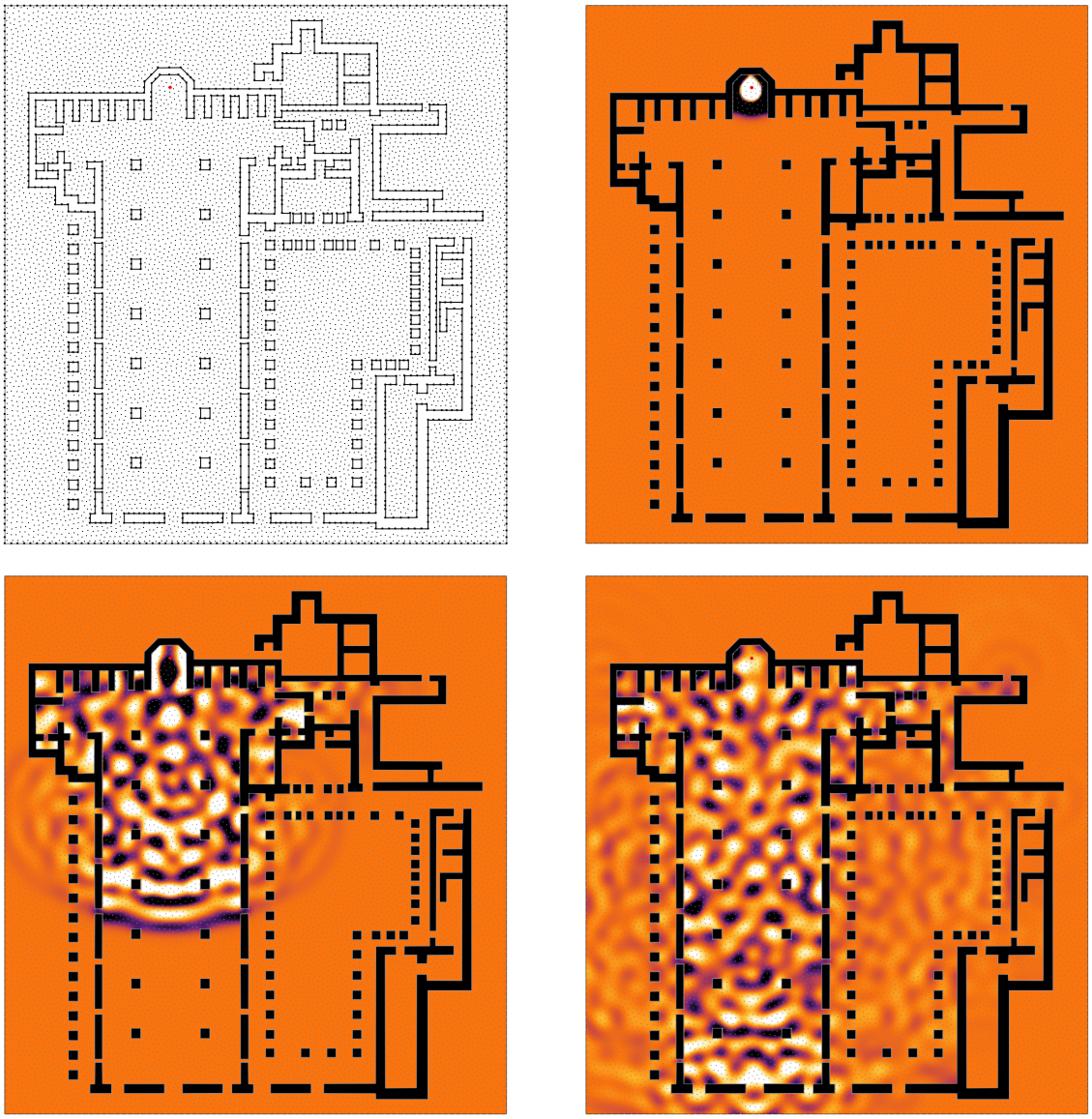


Figure 6.13: Simulation of wave propagation in the 2D model of a church. The simulation comprises a single domain, with a high genus (99) due to the large number of columns and other obstacles. The regularity of the spline space is maximal inside the simulation domain, and only reduced next to the domain boundaries.

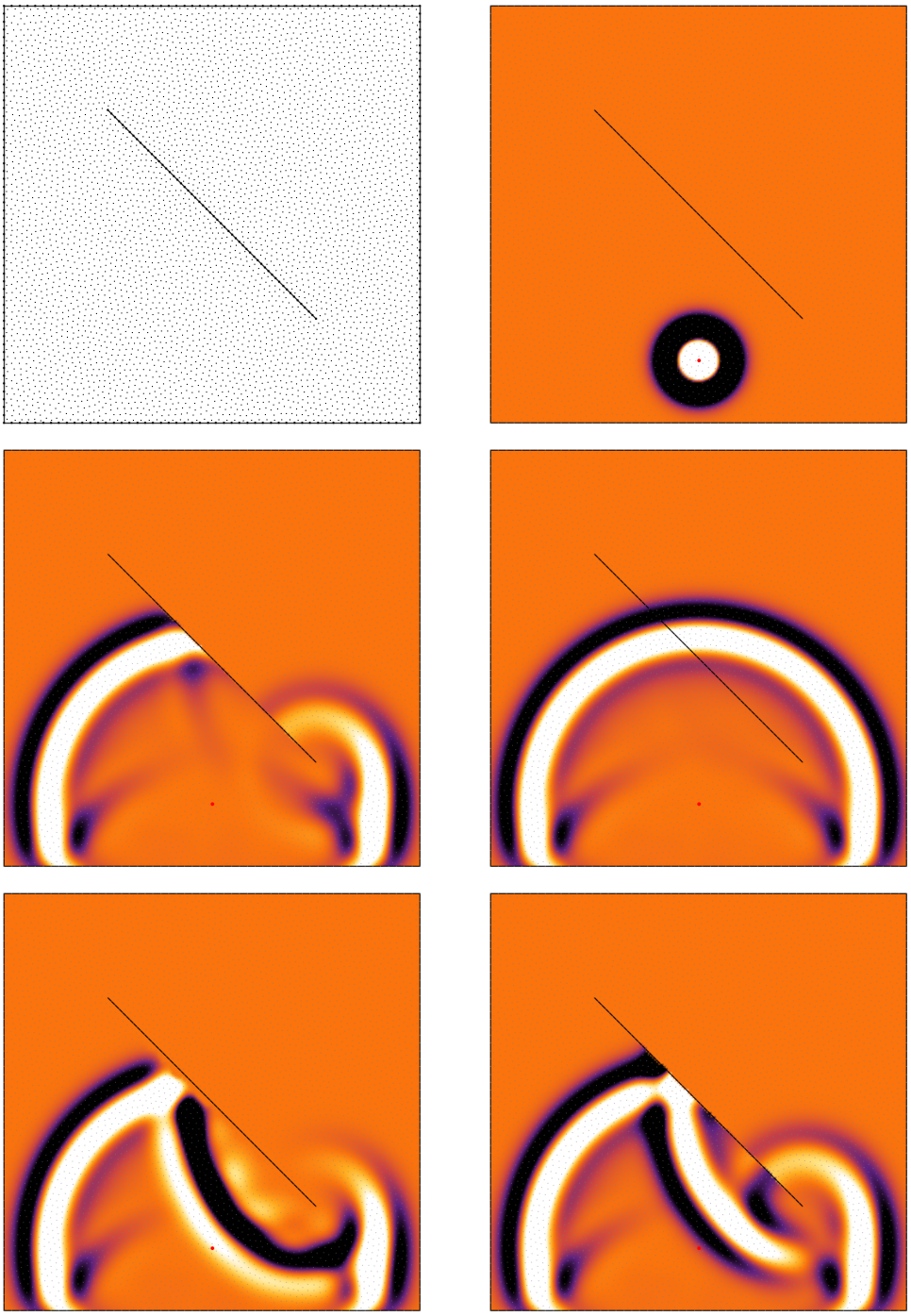


Figure 6.14: Simulation of a simple homogeneous model with a single fault on which we impose a set of different boundary conditions. Top row: model and incident wave. Middle row: absorbing (left) and transparent (right) boundary conditions. Bottom row: Dirichlet (left) and Neumann (right) boundaries.

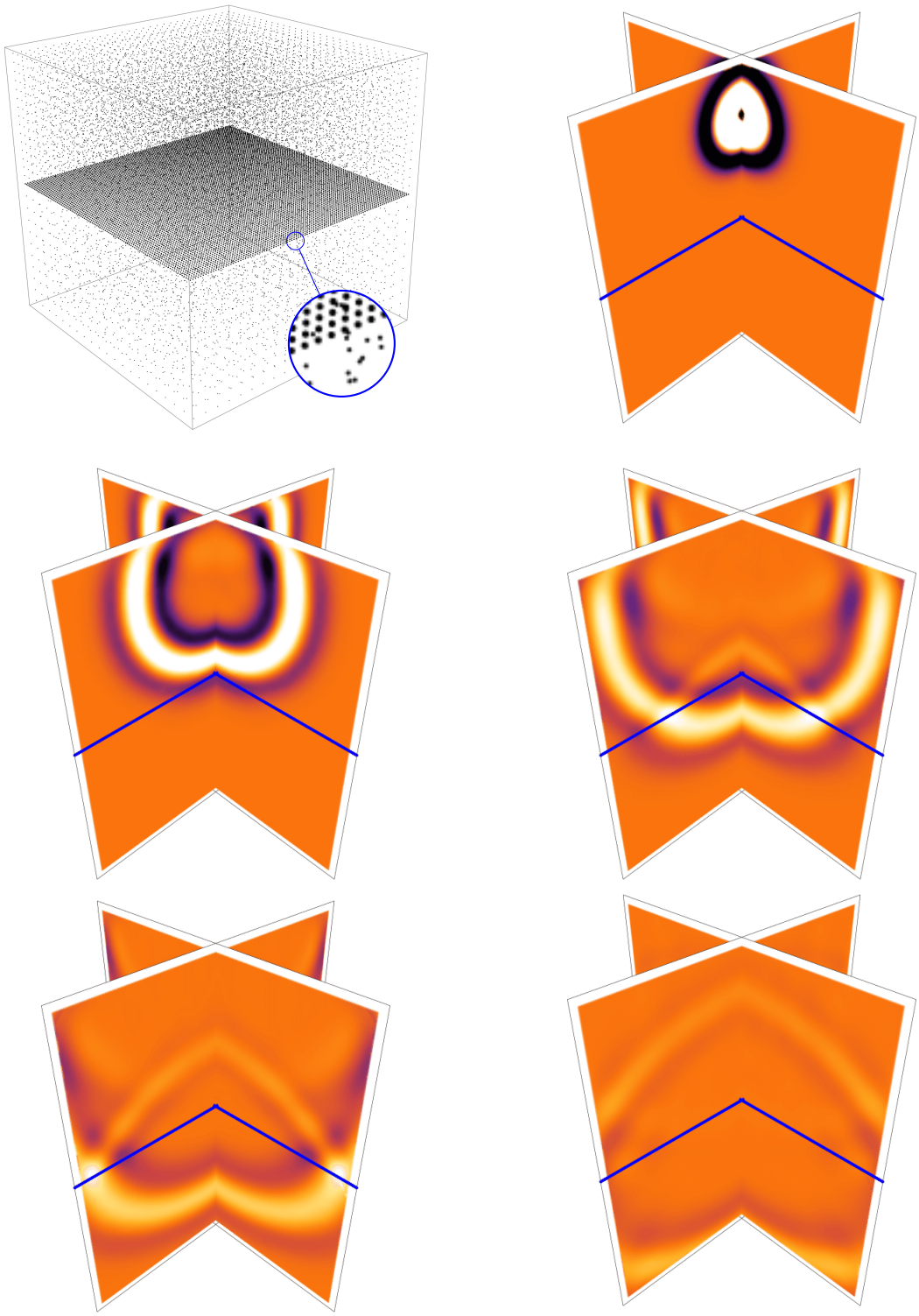


Figure 6.15: Wave propagation in a simple three-dimensional bi-layered model, composed of two IGA domains. We show the point configuration (interface points are thicker in the image), and five simulation snapshots. The interface between the layers is shown in blue.

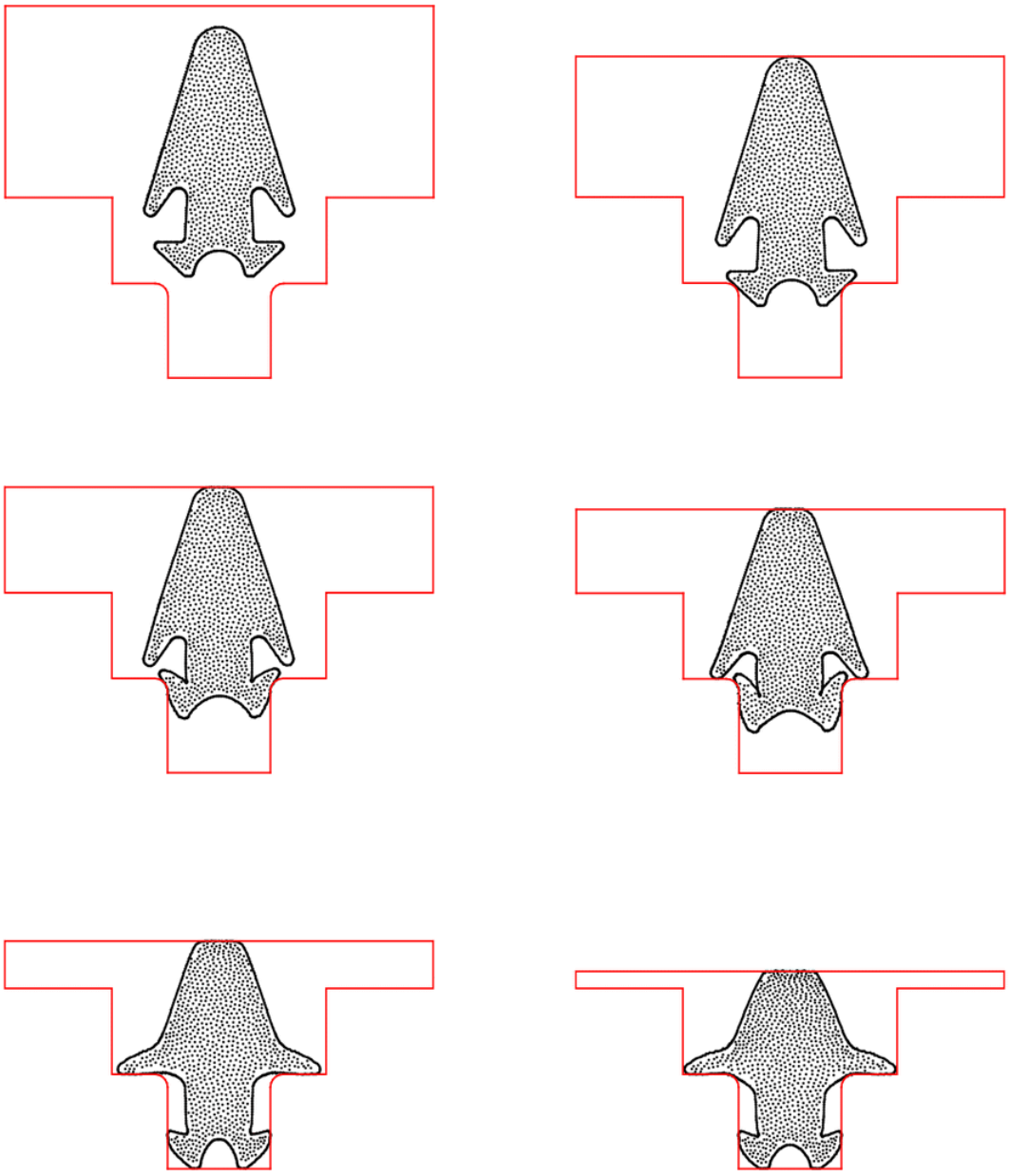


Figure 6.16: Six successive time steps of the simulation of the deformation of a simple two-dimensional gasket undergoing linear elasticity. The gasket's boundary and the point configuration are shown in black, and the constraints are shown in red.

IGA approaches, without any additional numerical noise (cf. Figure 6.10).

Since the spline functions have reduced regularity on the interfaces between subdomains, one can define the shape of the subdomains according to the physics of the problem, by aligning the zones of reduced regularity with the discontinuities of the physical parameters. In addition, one can introduce more subdomains than necessary, in order to reduce the size of the diagonal blocks of the system matrices, if required. One can also imagine fine-tuning the size of the subdomains to facilitate load balancing in parallel machines for HPC applications.

One of the major drawbacks of this method is the cost of computing the quadrature cells and assembling the system matrices themselves. In Algorithm 3, the intersection of simplices can be easily computed via convex hull algorithms, [312, 313]. Many efficient implementations exist, see, e.g., [314]. However, the number of cells in the quadrature decomposition can be very high, and can increase combinatorially with the degree k. One small improvement can be achieved using multivariate B-splines instead of simplex splines, as the linear combination (6.3) recovers continuity, and thus quadrature cells can be computed using Algorithm 3 with degree $k-1$. Furthermore, the computation of quadrature cells, and subsequent matrix assembly, are very easy to parallelize, and can therefore make full use of high-performance computational resources, when available. However, we consider this problem to be a major drawback of the proposed method. Compared to the usual IPDG method, the mass matrix associated to each domain is also in general larger, and thus its inversion represents another additional cost, albeit one that can be mitigated by choosing a finer subdomain decomposition.

One might be able to avoid a sizable amount of computation by using techniques such as that proposed in [315, 316]. In this work, the inversion of the mass matrix is avoided, while retaining a favorable timestep CFL condition, by using a modified timestepping scheme based on defect correction (DeC) techniques (see, e.g., [317]). Within this paradigm, one recovers the accurate solution of the problem by performing a small and convergent set of iterations over approximate, cheaper problems with appropriate residuals on the right hand side. It is conceivable that one might define a similar approximate solution for unstructured splines, one that not only avoids the inversion of the mass matrix but possibly much of the computational complexity of the definition of exact quadrature cells. The approximate problem would be similar to that obtained through *mass lumping* techniques, but the method accuracy would be recovered by the deferred correction method. We plan on investigating this possibility in a future work.

An improvement in the computational cost associated to matrix assembly might also come from the use of alternative integration techniques based on some cone spline decomposition which have been developed over the years (see e.g. [318–321]). Since cone splines can be naturally associated to some features of zonotopes, it is possible that these classic approaches can be made more efficient by exploiting this connection. Other approaches, based on triangulations of simploids [322], may also be successful in improving our method. We defer the exploration of these possibilities to some future work.

7 | Towards Full Waveform Inversion

Wovon man nicht sprechen kann, darüber muß man schweigen.

Ludwig Wittgenstein, Tractatus Logico-Philosophicus (1921)

So far, we have explored the possible use of unstructured spline spaces for the discretization of partial differential equations. We wish to discuss, in this short chapter, the possible consequences of choosing an unstructured spline space to describe the space of physical parameters instead, in view of possible applications to the seismic inversion problem. This section is intended to be only exploratory, and only gives a possible roadmap for future development, even though we have indeed performed some very preliminary numerical experiments to test the practical applicability of the proposed method.

In the first section, we discuss the implications of associating the position of the spline knots to degrees of freedom for inversion. We discuss in particular how the gradient can be computed in this case, and how the reduction in regularity of the basis functions can be used to reconstruct discontinuities in the physical model from a smooth initial guess.

In the following section, we discuss how the Euler-Darboux equations (4.30) and (4.45) can be used to simplify the calculation of the Hessian matrix and reduce the complexity of the Newton method.

Knot locations as degrees of freedom

In Chapter 1, we have briefly presented the seismic inversion process, and the associated optimization problem of the cost function with respect to the change in the physical model. We have discussed the fact that one can associate some inversion degrees of freedom to the problem geometry, for example to the vertices of a mesh (see, e.g., [57, 58]), in which case the cost function is not necessarily differentiable with respect to the model parameters when degenerate configurations are reached.

Using the results of Chapter 4, one can similarly construct a spline space over a point configuration A (see Chapters 5 and 6) and then compute the subgradient of the cost function with respect to the position of the knots in A. One can still expect the cost function to be non-differentiable with respect to these variables, since the spline functions themselves become less regular and even discontinuous as the points become affinely dependent (cf. Figure 3.4). However, since one does not have to guarantee the existence of a valid mesh at all times, one

can possibly avoid the somewhat costly remeshing process that is sometimes necessary in mesh-adaptive approaches (see [57, 58]).

The reduction in regularity, experienced by the spline functions on A as the points become affinely dependent (see Figure 3.4), can also be an advantage if the underlying model is expected to have discontinuities, as is the case in seismic imaging applications. In this case, one might start from a smooth physical model, with points of A in general position (except at the boundaries), and then let the minimization of the cost function (1.22) displace the points until the resulting model is discontinuous. We show below an example of this approach for a one-dimensional seismic inversion process as well as a simple two-dimensional model, and we defer more sophisticated two- and three-dimensional numerical experiments to a future work.

Gradient computation

The adjoint state method, through Theorem 1.3.3, gives a subgradient of the cost function $J(m)$ at a given point $\overline{m}$ with respect to the model m as a simple scalar product involving the solution to the forward problem $\overline{p}(x,t)$ and the solution to the adjoint problem $\overline{u}(x,t)$, namely,

$$\partial J(\overline{m}) \subseteq \partial \left\langle -\overline{u}, L(m)\overline{p} \right\rangle (\overline{m}), \tag{7.1}$$

where $m \coloneqq (\lambda, \rho)$ is the physical model, and L is the acoustic wave operator

$$L(\lambda, \rho) \coloneqq \frac{1}{\lambda}\frac{\partial^2}{\partial t^2} - \nabla \cdot \left(\frac{1}{\rho}\nabla\right). \tag{7.2}$$

The right hand side of (7.1) involves the derivatives of the linear operator L with respect to the model. Suppose that we discretize (7.2) over a set of functions $\varphi_h \in \mathcal{Q}_h$, and suppose that we use an unstructured spline space to discretize the physical parameters appearing in L and its associated boundary conditions, over a finite set of spline functions over A. We can then write

$$\frac{1}{\lambda} = \sum_{i=1}^{n_s} \gamma_i M(x \mid (a_s)_{s\in S_i}), \tag{7.3}$$

and similarly for ρ and c, where $A \coloneqq (a_i)_{i=1}^n$ is a point configuration and $(S_i)_{i=1}^{n_s} \subset [n]$ are appropriate subsets of indices defining the spline functions in the space. Using an explicit timestepping scheme, we can express (7.2) via a set of matrices, including the mass (2.15) and stiffness (2.16) matrices. In order to compute the right hand side of (7.1), one has to compute the derivatives of these matrices with respect to the inversion parameters. All these matrices can be expressed as integrals of the form

$$\mathcal{O} = \sum_{i=1}^{n_s} \gamma_i \int_\Omega M(x \mid (a_s)_{s\in S_i})\psi(x)\,\mathrm{d}\Omega, \tag{7.4}$$

where ψ is the product of functions $\varphi_h \in \mathcal{Q}_h$ or their derivatives. Boundary matrices such as the damping (2.17) matrices, as well as the flux and penalty matrices (2.30), (2.31) in the case of DG or multi-patch schemes, can be expressed as similar integrals over the boundary, and can

be treated in a very similar way.

Computing the derivative of (7.4) with respect to the coefficients γ_i of (7.3) is straightforward,

$$\frac{\partial \mathcal{O}}{\partial \gamma_i} = \int_\Omega M(x \mid (a_s)_{s \in S_i}) \psi(x) \, \mathrm{d}\Omega.$$

Let us now compute the derivative of $\mathcal{O}$ with respect to the position of a knot in $a_j \in A$. Suppose that a_j has multiplicity $r_{j,i}$ in S_i, and let $w_{j,i} \coloneqq r_{j,i} / |S_i|$. Then, one can use (4.44) to transfer the derivative of each spline with respect to a_j to a derivative with respect to x,

$$\nabla_{a_j} \mathcal{O} = -\sum_{i=1}^{n_s} \gamma_i w_{j,i} \int_\Omega (\nabla_x M(x \mid (a_s)_{s \in S_i} \sqcup \{a_j\}) \psi(x) \, \mathrm{d}\Omega, \tag{7.5}$$

where the disjoint union $\sqcup$ expresses the fact that the multiplicity of a_j is increased by one in $(a_s)_{s \in S_i}$. Using (4.42), one can then compute the directional derivative of M with respect to any direction $v \in \mathbb{R}^d$, after expressing v in terms of the knots of each spline function appearing in (7.5). For example, in the case of splines constructed as in Chapter 5, one can express $S_i \coloneqq I_i \sqcup B_i$, and use the subset B_i to write the direction v, using $|B_i| = d + 1$ terms, as

$$v = \sum_{b \in B_i} \nu_{b,i} a_b, \text{ with } \nu_{b,i} \coloneqq \frac{\det((a_c, 1)_{c \in B_i} \setminus \{(a_b, 1)\} \sqcup \{(v, 0)\})}{\det((a_c, 1)_{c \in B_i})},$$

where the row vector $(v, 0)$ replaces the row vector $(a_b, 1)$ in the determinant, similarly to the case of $\det(B_i)$ introduced in Chapter 5. One then obtains the simple expression

$$v \cdot \nabla_{a_j} \mathcal{O} = -\sum_{i=1}^{n_s} \gamma_i r_{j,i} \int_\Omega \sum_{b \in B_i} \nu_{b,i} M(x \mid I_i \sqcup B_i \sqcup \{a_j\} \setminus \{a_b\}) \psi(x) \, \mathrm{d}\Omega, \tag{7.6}$$

and, using d orthogonal vectors $v_1, \ldots, v_d$ in (7.6), one can compute the derivative of $\mathcal{O}$ with respect to each of the components of the knot vector a_j. This gradient, although computationally expensive, can be used to minimize the cost function with respect to the knot positions.

In Figure 7.1, we consider the minimization of the simple cost function

$$J \coloneqq \int_0^1 |f(x) - \mu(x)|^2 \, \mathrm{d}x, \tag{7.7}$$

where $\mu(x) = 1 + \theta(x - 1/2)$ is a step function, $A = \{0, 0, 0, a_1, \ldots, a_n, 1, 1, 1\}$ is a clamped knot vector, and

$$f(x) = \sum_{i=1}^{n+3} \gamma_i N_{i,2},$$

where $N_{i,2}$ represents the usual one-dimensional B-spline basis of degree 2 over the clamped knot vector A. We show the result of the minimization of the cost function over both the position of the internal knots $(a_i)_{i=1}^n$ and the spline coefficients $(\gamma_i)_{i=1}^{n+3}$, in the case of $n = 9$ internal knots.

We also show the plot of the cost function J, in the special case $n = 3$ and

$$A = \{0, 0, 0, \frac{1}{2}, y, z, 1, 1, 1\}, \tag{7.8}$$

with respect to the knot locations y and z. Notice that in this case the cost function attains its minimum, equal to zero, when $y = z = 1/2$, in which case the spline basis becomes discontinuous and thus capable of reproducing the step function. However, as shown in Figure 7.1, the cost function is not differentiable there, as anticipated in Chapter 1. For this reason, near the minimum, the usage of Newton methods incorporating second-order information can become counterproductive.

In Figure 7.2, we show the result of performing a simple one-dimensional full waveform inversion to recover a piecewise constant velocity model starting from a simple gradient. The model is expressed over the usual B-spline basis over the segment, starting with a clamped knot vector with equally spaced knots. The position of the knots, as well as the expansion coefficients, are recovered through inversion. Notice that, even though the number of degrees of freedom (i.e., the number of knots) is very limited, one is able to recover precisely the discontinuities without any *a priori* knowledge about their location.

In Figure 7.3, we show the result of a simple L^2 minimization of a bivariate spline function, expressed as over the unstructured simplex spline space over a very simple domain. The minimization is performed over both the coefficients of the 186 spline functions in the space, and over the two-dimensional locations of the 17 non-boundary knots in the point set. The target model is a simple step function. As can be seen, the discontinuity tends to be better and better resolved by the optimized function, although the convergence rate seems to slow significantly as the basis becomes more and more degenerate.

Further analysis and numerical tests are needed to assess whether this capability of recovering the location of discontinuities with a low number of inversion degrees of freedom carries over to more complex two- and three-dimensional problems.

Hessian matrix computation

The one-dimensional Euler-Darboux equations presented in Theorem 4.3.19 were first derived by Carlson for Dirichlet averages [208, Theorem 5.4-1] and later reinterpreted in the context of splines [209]. We have detailed some of the theoretical aspects of these relations in Chapter 4. These equations however also have some practical implications, as they directly relate the (mixed) second-order derivatives of simplex splines (and thus B-splines) to a simple linear combination of their first derivatives. In fact, combining (4.30) with (4.28) we can obtain the explicit expression

$$\frac{\partial^2}{\partial a_u \partial a_v} M(x \mid A) = \frac{r_u r_v}{c} \frac{M'(x \mid A \sqcup \{a_u\}) - M'(x \mid A \sqcup \{a_v\})}{a_u - a_v},$$

which can be used in the minimization above to reduce the computational complexity of the calculation of the Hessian matrix of the cost function from $O(n_s^2)$ to $O(n_s)$.

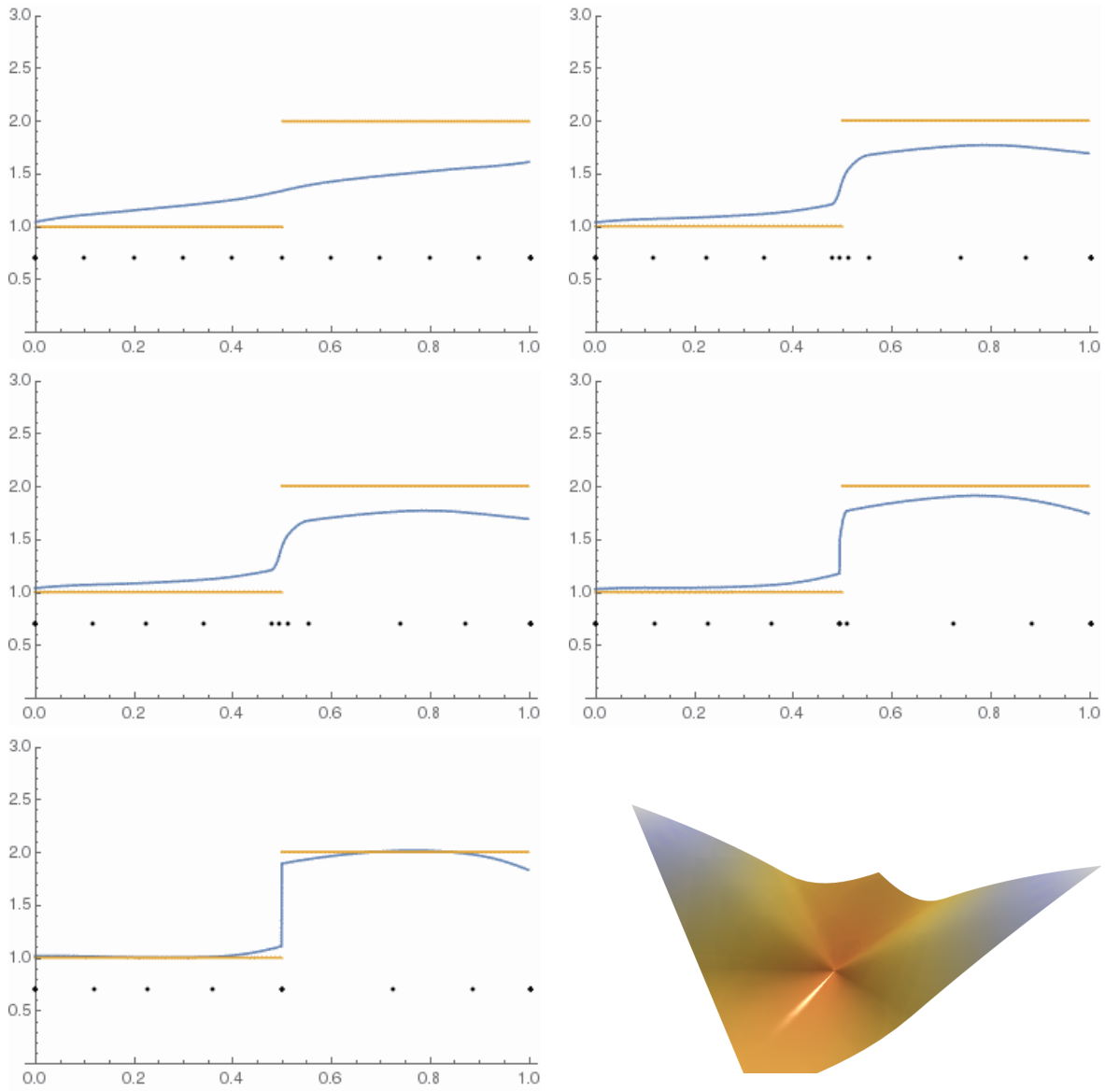


Figure 7.1: One-dimensional minimization of (7.7) with respect to the coefficients and knot positions of the basis. We show the target velocity model in yellow and the evolving model in blue. We also show, under these curves, the knot locations in black. (Bottom right) plot of (7.7) over the knot vector (7.8) with respect to the position of two knots y and z. Notice that the cost function is not differentiable at the minimum.

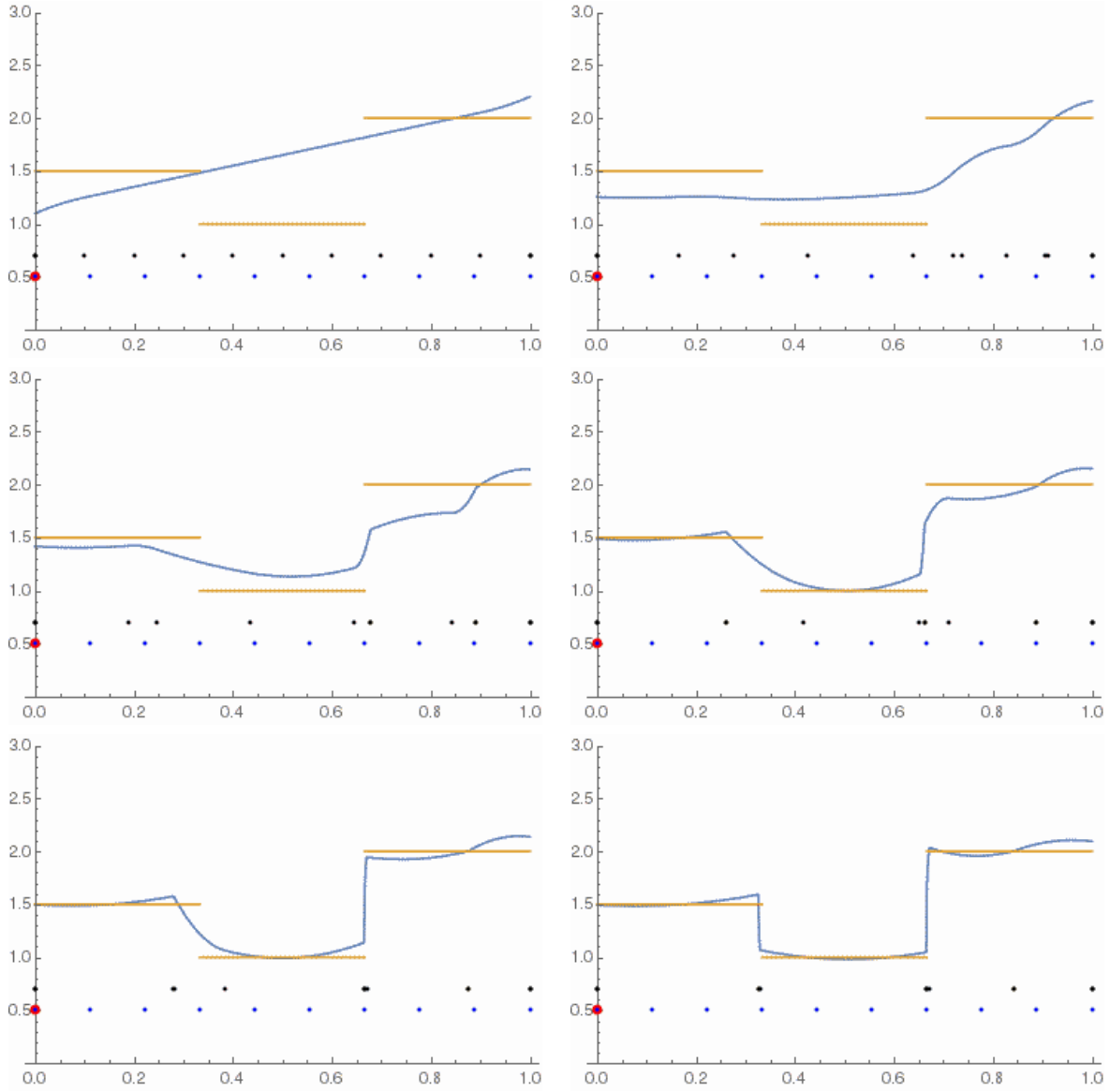


Figure 7.2: One-dimensional FWI with degrees of freedom on the spline coefficients and knot locations. The knots on the boundary are fixed, whereas the 9 internal knots are left free to move. We show the target velocity model in yellow and the evolving model in blue. We also show, under these curves, the knot locations in black, the receiver locations in blue, and the source location in red.

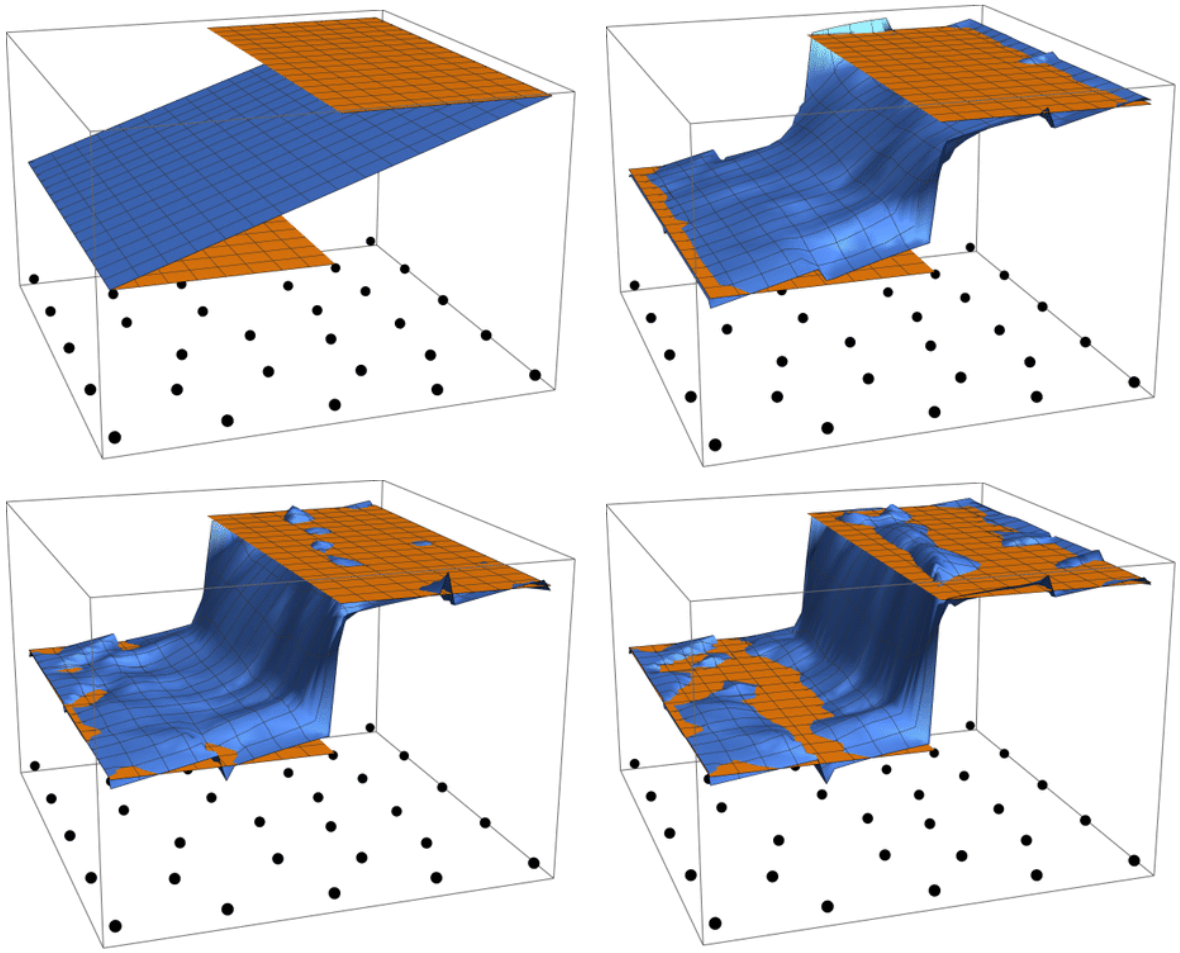


Figure 7.3: Two-dimensional optimization with degrees of freedom on the spline coefficients and knot locations. The knots on the boundary are fixed, whereas the 17 internal knots are left free to move. We show the target velocity model in yellow and the evolving model in blue. We also show, under the function graphs, the knot locations in black.

Let us start with the one-dimensional case. Starting from (7.5),

$$\frac{\partial \mathcal{O}}{\partial a_j} = -\sum_{i=1}^{n_s} \gamma_i w_{j,i} \int_\Omega M'(x \mid (a_s)_{s\in S_i} \sqcup \{a_j\}) \psi(x) \,\mathrm{d}\Omega,$$

one more iteration of (4.28) yields the diagonal part of the Hessian:

$$H_{jj} \coloneqq \frac{\partial^2}{\partial a_j^2}\mathcal{O} = \sum_{i=1}^{n_s} \gamma_i \frac{r_{j,i}(r_{j,i}+1)}{|S_i|\,(|S_i|+1)} \int_\Omega M''(x \mid (a_s)_{s\in S_i} \sqcup \{a_i\}^2) \,\mathrm{d}\Omega.$$

Thanks to (4.30), the off-diagonal part of the Hessian matrix can be evaluated without computing any second-order derivatives:

$$\begin{aligned} H_{uv} &\coloneqq \frac{\partial^2 \mathcal{O}}{\partial a_u \partial a_v}, \\ &= \sum_{i=1}^{n_s} \gamma_i \frac{r_{u,i} r_{v,i}}{|S_i|} \int_\Omega \frac{1}{a_u - a_v} \left(M'(x \mid (a_s)_{s\in S_i} \sqcup \{a_u\}) - M'(x \mid (a_s)_{s\in S_i} \sqcup \{a_v\}) \right) \psi(x) \,\mathrm{d}\Omega. \end{aligned}$$

Essentially, the computation of both the gradient and the Hessian matrix only requires computing, for each spline function of degree k defined by the indices S_i, the $2\,|S_i| = 2(k+2)$ spline functions $M'(x \mid (a_s)_{s\in S_i} \sqcup \{a_u\})$ for all $u \in S_i$, and $M''(x \mid (a_s)_{s\in S_i} \sqcup \{a_u\}^2)$, again for all $u \in S_i$. In contrast, a naive computation would require the evaluation of $(k+2)(k+3)$ different spline functions. This reduction in complexity might represent a speedup for certain applications, including full waveform inversion, where the computation of the full Hessian matrix is often too costly to be performed in practice.

Contrary to the univariate case (4.30), (4.45) cannot be directly used to compute all the off-diagonal terms of the Hessian matrix. For instance, the equation is trivial whenever $i = j$, so the mixed derivatives with respect to two different components of the same knot cannot be computed from (4.45). In fact, in dimension d, the Euler-Darboux equations (4.45) and (4.47),

$$\begin{aligned} \left(\sum_{m=1}^{d} (a_{u,m} - a_{v,m}) \frac{\partial^2}{\partial a_{u,m} \partial a_{v,k}} + r_{u,i} \frac{\partial}{\partial a_{v,k}} - r_{v,i} \frac{\partial}{\partial a_{u,k}} \right) M(x \mid (a_s)_{s\in S_i}) = 0, \\ \left(\sum_{m=1}^{d} (a_{u,m} - a_{v,m}) \frac{\partial^2}{\partial a_{u,k} \partial a_{v,m}} + r_{u,i} \frac{\partial}{\partial a_{v,k}} - r_{v,i} \frac{\partial}{\partial a_{u,k}} \right) M(x \mid (a_s)_{s\in S_i}) = 0, \end{aligned} \tag{7.9}$$

only yield $2d$ linear equations relating the d^2 variables $a_{u,m}$, $a_{v,k}$ for $m, k = 1 \ldots, d$. If $d = 2$, these equations can be solved and the mixed terms of the Hessian involving two different knots can be computed from the terms involving a single knot. Unfortunately, we do not know yet of a way of applying this simplification to the case $d \geq 3$. Even so, (7.9) can be used to reduce the computational cost of the Hessian matrix in all dimensions.

In real-world applications, the use of these techniques always needs to be carefully considered, since, as shown in Figure 7.1, the cost function is not generally differentiable when the knots become affinely dependent. Consequently, the application of the Hessian matrix via Newton's method is more likely to be useful at the beginning of the minimization, before the knots align

to form discontinuities, and can be expected to be less useful and even problematic as the discontinuity interfaces start to form.

All the arguments presented in this section are still partial and under development, and are to be interpreted as hints to a certain usefulness of the Euler-Darboux equations (4.30) and (4.45) for optimization problems, rather than a full-formed numerical approach.

Conclusions

A viagem não acaba nunca. Só os viajantes acabam.

José Saramago, Viagem a Portugal (1983)

We have explored in this work the applicability of simplex spline spaces to the time-explicit numerical analysis of hyperbolic problems, and more specifically, to the simulation of acoustic wave propagation with absorbing boundary conditions and the related inverse problem known as seismic imaging via Full Waveform Inversion (FWI).

The main contributions of this work are as follows. First, we provided a generalization to some already proven results on the construction of polynomial-reproducing simplex spline spaces, based on combinatorial objects known as zonotopal tilings. Using the properties of these objects, we were able to generalize a known result on Delaunay configurations [4], and to extend a formerly established two-dimensional spline space construction algorithm [5, 259] to an arbitrary number of dimensions, as well as to the case of repeated and affinely dependent points. This generalization is needed for applications in numerical analysis, since spline functions cannot interpolate the problem boundary unless their regularity is reduced there. Furthermore, these combinatorial structures possess a natural graph, whose properties can be exploited to devise algorithms for efficient spline evaluation.

Armed with these tools, we used knot multiplicities to subdivide the problem domain into subdomains, each associated to a spline space, and we used standard techniques coming from DG methods to connect the domains with fluxes and symmetric interior penalty terms. We showed that, near subdomain boundaries, our spline spaces reproduce the usual Bernstein-Bézier discontinuous Galerkin (DG) basis over simplices adjacent to the interfaces. Thus, we were able to apply the same inverse inequalities of [305], which are the basis for the derivation of the known positivity constraints and *a priori* error estimates for the usual symmetric interior-penalty DG method (IPDG). Furthermore, if each subdomain is a simplex, we proved that we recover the usual Bernstein-Bézier IPDG scheme, while on the other hand of the spectrum, when a single domain is defined, we derived an unstructured version of the usual B-spline isogeometric analysis (IGA) framework. Thus, numerically, our proposed method can interpolate between these two approaches and their relative advantages, recovering the improved maximum timestep condition of IGA methods, while retaining the block-diagonal structure of the mass matrix typical of DG. Overall, our method resembles a fully unstructured version of the multi-patch DG-IGA approach of [3], with the possibility of simulating domains of complex topology, and to naturally couple DG and (unstructured) IGA schemes in the same simulation. We illustrated these points with

a few numerical experiments.

Finally, we turned our attention to possible applications of unstructured spline functions to FWI. In particular, we explored some of the consequences of minimizing the usual FWI cost function with respect to the position of the knots defining the spline functions. We showed that the gradient of the cost function can be computed using some known formulas derived originally in [208] for Dirichlet averages, and later applied to simplex splines [209]. We noticed that the FWI cost function is not generally differentiable in this case, which is a common feature of optimization processes with geometrical degrees of freedom. For this reason, we introduced a slight generalization of the adjoint state method of FWI using subdifferentials and a known convex duality. We also briefly discussed the possibility of using the Euler-Darboux equations, relating second and first derivatives of spline functions with respect to their knots, to the computation of the Hessian, in one and two dimensions. Some early one- and two-dimensional numerical experiments give hints on the feasibility of this approach, although more experiments in two and three dimensions are needed to further explore the interest of this method for shape optimization problems.

The work presented here can be further extended and improved in many possible directions.

First, for the construction of spline spaces, only fine zonotopal tilings have been explored. It is possible that similar connections could be found between more general zonotopal tilings and other kinds of multivariate splines, such as box splines or more general polyhedral splines [216, 220]. Furthermore, the spline evaluation scheme proposed in this work does not optimize the number of auxiliary functions needed. Optimized weighted Delaunay triangulations coming from computer graphics applications could potentially provide some improvements in this regard.

Second, and more importantly, the assembly of system matrices in our approach is much more computationally expensive compared to other numerical schemes. In fact, if these matrices are assembled using the maximal cells over which all the spline functions in the space are pure polynomials, then one has to deal with a large number of integration cells. Thus, the total number of elementary integrals that need to be evaluated is much larger than in the usual IPDG or FE methods. Even though, computationally, this task can be easily parallelized, we believe that other integration schemes are worth exploring, and notably those connected with combinatorial structures such as cone splines [319] or simploids [322]. Hopefully, the connections with zonotopal tilings and triangulations uncovered in this work might be extended and further exploited to this end. Another possibility is to employ a method similar to that presented in [315, 316]. With this approach, one might be able to avoid a potentially large amount of computation by only evaluating an approximation of the system matrices. One then would recover the full convergence rate by using an appropriate defect-correcting timestepping scheme.

Finally, we have only provided a very early draft for the use of unstructured spline functions for full waveform inversion. Much more work is needed to bring this approach to fruition, notably in two and three dimensions. As for the forward problem, the scalability of the time integration scheme needs to be verified in an actual large-scale HPC setting, and some real-world tests are needed to assess the impact of the inter-node communication associated with the degrees of freedom on the boundaries between the sub-domains. We plan on performing these extensions and analyses in future works.

Bibliography

[1] H. Barucq, R. Djellouli, and E. Estecahandy. "Efficient DG-like formulation equipped with curved boundary edges for solving elasto-acoustic scattering problems". In: *International Journal for Numerical Methods in Engineering* 98.10 (2014), pp. 747–780.

[2] T. J. R. Hughes, J. A. Cottrell, and Y. Bazilevs. "Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement". In: *Computer methods in applied mechanics and engineering* 194.39-41 (2005), pp. 4135–4195.

[3] J. Chan and J. A. Evans. "Multi-patch discontinuous Galerkin isogeometric analysis for wave propagation: Explicit time-stepping and efficient mass matrix inversion". In: *Computer Methods in Applied Mechanics and Engineering* 333 (2018), pp. 22–54.

[4] M. Neamtu. "Delaunay configurations and multivariate splines: a generalization of a result of BN Delaunay". In: *Transactions of the American Mathematical Society* 359.7 (2007), pp. 2993–3004.

[5] Y. Liu and J. Snoeyink. "Quadratic and cubic B-splines by generalizing higher-order Voronoi diagrams". In: *Proceedings of the twenty-third annual symposium on Computational geometry*. ACM. 2007, pp. 150–157.

[6] D. Schmitt. "Bivariate B-Splines from convex pseudo-circle configurations". In: *International Symposium on Fundamentals of Computation Theory*. Springer. 2019, pp. 335–349.

[7] J. Cao, Z. Chen, X. Wei, and Y. J. Zhang. "A finite element framework based on bivariate simplex splines on triangle configurations". In: *Computer Methods in Applied Mechanics and Engineering* 357 (2019), p. 112598.

[8] J. Virieux and S. Operto. "An overview of full-waveform inversion in exploration geophysics". In: *Geophysics* 74.6 (2009), WCC1–WCC26.

[9] E. Beretta, M. V. De Hoop, F. Faucher, and O. Scherzer. "Inverse boundary value problem for the Helmholtz equation: quantitative conditional Lipschitz stability estimates". In: *SIAM Journal on Mathematical Analysis* 48.6 (2016), pp. 3962–3983.

[10] K. Aki and P. G. Richards. *Quantitative seismology*. 2002.

[11] K. E. Bullen and B. A. Bolt. *An introduction to the theory of seismology*. Cambridge university press, 1985.

[12] X. Campman, K. Van Wijk, C. Riyanti, J. Scales, and G. Herman. "Imaging scattered seismic surface waves". In: *Near Surface Geophysics* 2.4 (2004), pp. 223–230.

[13] P. Bergamo, L. Bodet, L. V. Socco, R. Mourgues, and V. Tournat. "Physical modelling of a surface-wave survey over a laterally varying granular medium with property contrasts and velocity gradients". In: *Geophysical Journal International* 197.1 (2014), pp. 233–247.

[14] C. Shen. "Experimental and numerical studies of seismic wave propagation in carbonate rocks at the laboratory scale". PhD thesis. Université de Pau et des Pays de l'Adour, 2020.

[15] J. F. Claerbout. *Imaging the Earth's interior*. Vol. 1. Blackwell scientific publications Oxford, 1985.

[16] B. L. Biondi. *3D seismic imaging*. Society of Exploration Geophysicists, 2006.

[17] R. Feynman, R. Leighton, and M. Sands. *The Feynman lectures on physics, Vol. II: The new millennium edition: mainly electromagnetism and matter*. The Feynman Lectures on Physics. Basic Books, 2011. ISBN: 9780465024940.

[18] D. P. Challa and M. Sini. "On the justification of the Foldy-Lax approximation for the acoustic scattering by small rigid bodies of arbitrary shapes". In: *Multiscale Modeling & Simulation* 12.1 (2014), pp. 55–108.

[19] A. Bendali, P.-H. Cocquet, and S. Tordeux. "Approximation by multipoles of the multiple acoustic scattering by small obstacles in three dimensions and application to the Foldy theory of isotropic scattering". In: *Archive for Rational Mechanics and Analysis* 219.3 (2016), pp. 1017–1059.

[20] H. Barucq, F. Faucher, and H. Pham. "Localization of small obstacles from back-scattered data at limited incident angles with full-waveform inversion". In: *Journal of Computational Physics* 370 (2018), pp. 1–24.

[21] H. Barucq, J. Diaz, V. Mattesi, and S. Tordeux. "Asymptotic behavior of acoustic waves scattered by very small obstacles". In: *ESAIM: Mathematical Modelling and Numerical Analysis* (2020), p. 32.

[22] N. Ricker. "The form and laws of propagation of seismic wavelets". In: *Geophysics* 18.1 (1953), pp. 10–40.

[23] J. Diaz and P. Joly. "A time domain analysis of PML models in acoustics". In: *Computer methods in applied mechanics and engineering* 195.29-32 (2006), pp. 3820–3853.

[24] R. Clayton and B. Engquist. "Absorbing boundary conditions for acoustic and elastic wave equations". In: *Bulletin of the seismological society of America* 67.6 (1977), pp. 1529–1540.

[25] F. Nataf. "Absorbing boundary conditions and perfectly matched layers in wave propagation problems". In: *Direct and inverse problems in wave propagation and applications*. de Gruyter, 2013, pp. 219–232.

[26] H. Barucq. "Étude asymptotique du système de Maxwell avec conditions aux limites absorbantes". PhD thesis. Bordeaux 1, 1993.

[27] K.-J. Engel and R. Nagel. "One-parameter semigroups for linear evolution equations". In: *Semigroup forum*. Vol. 63. 2. Springer. 2001, pp. 278–280.

[28] A. Pazy. *Semigroups of linear operators and applications to partial differential equations*. Vol. 44. Springer Science & Business Media, 2012.

[29] M. H. Stone. "Linear transformations in Hilbert space: III. Operational methods and group theory". In: *Proceedings of the National Academy of Sciences of the United States of America* 16.2 (1930), p. 172.

[30] G. Lumer and R. S. Phillips. "Dissipative operators in a Banach space." In: *Pacific Journal of Mathematics* 11.2 (1961), pp. 679–698.

[31] S. Teufel and R. Tumulka. "Existence of Schrödinger evolution with absorbing boundary condition". In: *arXiv preprint arXiv:1912.12057* (2019).

[32] R. A. Adams and J. J. Fournier. *Sobolev spaces*. Elsevier, 2003.

[33] P.-A. Raviart, J.-M. Thomas, and P. G. Ciarlet. *Introduction à l'analyse numérique des équations aux dérivées partielles*. Mathématiques appliquées pour la maîtrise. Dunod, 2004.

[34] P. Shvartsman. "On extensions of Sobolev functions defined on regular subsets of metric measure spaces". In: *Journal of Approximation Theory* 144.2 (2007), pp. 139–161.

[35] P. Hajłasz, P. Koskela, and H. Tuominen. "Sobolev embeddings, extensions and measure density condition". In: *Journal of Functional Analysis* 254.5 (2008), pp. 1217–1234.

[36] R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology: Volume 1 - physical origins and classical methods*. Springer Science & Business Media, 2012.

[37] A. Tarantola. "Inversion of seismic reflection data in the acoustic approximation". In: *Geophysics* 49.8 (1984), pp. 1259–1266.

[38] C. Shin, S. Pyun, and J. B. Bednar. "Comparison of waveform inversion, part 1: conventional wavefield vs logarithmic wavefield". In: *Geophysical Prospecting* 55.4 (2007), pp. 449–464.

[39] B. Engquist and B. D. Froese. "Application of the Wasserstein metric to seismic signals". In: *Communications in Mathematical Sciences* 12.5 (2014), pp. 979–988.

[40] F. Faucher. "Contributions to seismic full waveform inversion for time harmonic wave equations: Stability estimates, convergence analysis, numerical experiments involving large scale optimization algorithms". PhD thesis. Université de Pau et des Pays l'Adour, 2017.

[41] L. Métivier, R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux. "Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion". In: *Geophysical Supplements to the Monthly Notices of the Royal Astronomical Society* 205.1 (2016), pp. 345–377.

[42] L. Métivier, R. Brossier, Q. Merigot, É. Oudet, and J. Virieux. "An optimal transport approach for seismic tomography: Application to 3D full waveform inversion". In: *Inverse Problems* 32.11 (2016), p. 115008.

[43] Y. Yang, B. Engquist, J. Sun, and B. F. Hamfeldt. "Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion". In: *Geophysics* 83.1 (2018), R43–R62.

[44] J. Chen, Y. Chen, H. Wu, and D. Yang. "The quadratic Wasserstein metric for earthquake location". In: *Journal of Computational Physics* 373 (2018), pp. 188–209.

[45] G. Chavent. *Nonlinear least squares for inverse problems: theoretical foundations and step-by-step guide for applications*. Springer Science & Business Media, 2010.

[46] E. Landa and S. Treitel. "Seismic inversion: what it is, and what it is not". In: *The Leading Edge* 35.3 (2016), pp. 277–279.

[47] R. G. Pratt, C. Shin, and G. Hick. "Gauss–Newton and full Newton methods in frequency–space seismic waveform inversion". In: *Geophysical Journal International* 133.2 (1998), pp. 341–362.

[48] A. Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

[49] J. Nocedal. "Updating quasi-Newton matrices with limited storage". In: *Mathematics of computation* 35.151 (1980), pp. 773–782.

[50] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[51] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock. "An introduction to total variation for image analysis". In: *Theoretical foundations and numerical methods for sparse recovery*. de Gruyter, 2010, pp. 263–340.

[52] S. Becker, J. Bobin, and E. J. Candès. "NESTA: A fast and accurate first-order method for sparse recovery". In: *SIAM Journal on Imaging Sciences* 4.1 (2011), pp. 1–39.

[53] K. Bredies and H. K. Pikkarainen. "Inverse problems in spaces of measures". In: *ESAIM: Control, Optimisation and Calculus of Variations* 19.1 (2013), pp. 190–218.

[54] Y. Soliman, D. Slepčev, and K. Crane. "Optimal cone singularities for conformal flattening". In: *ACM Transactions on Graphics (TOG)* 37.4 (2018), pp. 1–17.

[55] R. Brossier, S. Operto, and J. Virieux. "Which data residual norm for robust elastic frequency-domain full waveform inversion?" In: *Geophysics* 75.3 (2010), R37–R46.

[56] C. C. Lopez. "Speed up and regularization techniques for seismic full waveform inversion". PhD thesis. Université Nice Sophia Antipolis, 2014.

[57] S. Thrastarson, M. van Driel, L. Krischer, C. Boehm, M. Afanasiev, D.-P. Van Herwaarden, and A. Fichtner. "Accelerating numerical wave propagation by wavefield adapted meshes. Part II: full-waveform inversion". In: *Geophysical Journal International* 221.3 (2020), pp. 1591–1604.

[58] P. Jacquet. "Time-Domain Full Waveform Inversion using advanced Discontinuous Galerkin method." PhD thesis. Université de Pau et des Pays de l'Adour, 2021.

[59] O. Nachum and B. Dai. "Reinforcement learning via Fenchel-Rockafellar duality". In: *arXiv preprint arXiv:2001.01866* (2020).

[60] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Vol. 408. Springer, 2011.

[61] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

[62] C. Villani. *Topics in optimal transportation*. 58. American Mathematical Soc., 2003.

[63] B. S. Mordukhovich and Y. Shao. "On nonconvex subdifferential calculus in Banach spaces". In: *J. Convex Anal* 2.1-2 (1995), pp. 211–227.

[64] M. Kern. *Numerical methods for inverse problems*. John Wiley & Sons, 2016.

[65] G. Allaire. *Numerical analysis and optimization: an introduction to mathematical modelling and numerical simulation*. Oxford university press, 2007.

[66] A. Griewank and A. Walther. "Algorithm 799: revolve: an implementation of checkpointing for the reverse or adjoint mode of computational differentiation". In: *ACM Transactions on Mathematical Software (TOMS)* 26.1 (2000), pp. 19–45.

[67] W. W. Symes. "Reverse time migration with optimal checkpointing". In: *Geophysics* 72.5 (2007), SM213–SM221.

[68] R.-E. Plessix. "A review of the adjoint-state method for computing the gradient of a functional with geophysical applications". In: *Geophysical Journal International* 167.2 (2006), pp. 495–503.

[69] R. G. Clapp. "Reverse time migration with random boundaries". In: *Seg technical program expanded abstracts 2009*. Society of Exploration Geophysicists, 2009, pp. 2809–2813.

[70] F. Faucher, H. Barucq, H. Calandra, and G. Chavent. "Quantitative convergence and stability of seismic inverse problems." In: *Reconstruction Methods for Inverse Problems*. 2018.

[71] M. A. Slawinski. *Waves and rays in elastic continua*. World Scientific, 2010.

[72] I. Tsvankin and V. Grechka. *Seismology of azimuthally anisotropic media and seismic fracture characterization*. Society of Exploration Geophysicists, 2011.

[73] R. P. Fletcher, X. Du, and P. J. Fowler. "Reverse time migration in tilted transversely isotropic (TTI) media". In: *Geophysics* 74.6 (2009), WCA179–WCA187.

[74] J. Yan and P. Sava. "Improving the efficiency of elastic wave-mode separation for heterogeneous tilted transverse isotropic media". In: *Geophysics* 76.4 (2011), T65–T78.

[75] L. Boillot. "Contributions à la modélisation mathématique et à l'algorithmique parallèle pour l'optimisation d'un propagateur d'ondes élastiques en milieu anisotrope". PhD thesis. Université de Pau et des Pays l'Adour, 2014.

[76] B. Duquet, K. J. Marfurt, and J. A. Dellinger. "Kirchhoff modeling, inversion for reflectivity, and subsurface illumination". In: *Geophysics* 65.4 (2000), pp. 1195–1209.

[77] E. Baysal, D. D. Kosloff, and J. W. Sherwood. "Reverse time migration". In: *Geophysics* 48.11 (1983), pp. 1514–1524.

[78] G. A. McMechan. "Migration by extrapolation of time-dependent boundary values". In: *Geophysical prospecting* 31.3 (1983), pp. 413–420.

[79] N. D. Whitmore. "Iterative depth migration by backward time propagation". In: *SEG Technical Program Expanded Abstracts 1983*. Society of Exploration Geophysicists, 1983, pp. 382–385.

[80] R. G. Pratt. "Seismic waveform inversion in the frequency domain, Part 1: Theory and verification in a physical scale model". In: *Geophysics* 64.3 (1999), pp. 888–901.

[81] H. Jun, Y. Kim, J. Shin, and C. Shin. "2D elastic time-Laplace-Fourier-domain hybrid full waveform inversion". In: *SEG Technical Program Expanded Abstracts 2013*. Society of Exploration Geophysicists, 2013, pp. 1008–1013.

[82] K. Xu and G. A. McMechan. "2D frequency-domain elastic full-waveform inversion using time-domain modeling and a multistep-length gradient approach". In: *Geophysics* 79.2 (2014), R41–R53.

[83] C. Baldassari. "Modélisation et simulation numérique pour la migration terrestre par équation d'ondes." PhD thesis. Université de Pau et des Pays de l'Adour, 2009.

[84] M. Bonnasse-Gahot, H. Calandra, J. Diaz, and S. Lanteri. "Performances analysis of a Hybridizable discontinuous Galerkin solver for the 3D Helmholtz equations in geophysical context". In: *Third EAGE workshop on High performance computing for Upstream*. Vol. 2017. 1. European Association of Geoscientists & Engineers. 2017, pp. 1–5.

[85] H. Barucq, H. Calandra, J. Diaz, and E. Shishenina. "Space-time Trefftz-discontinuous Galerkin approximation for elasto-acoustics". PhD thesis. Inria Bordeaux Sud-Ouest; UPPA (LMA-Pau); Total E&P, 2017.

[86] E. Shishenina. "Space-Time Discretization of Elasto-Acoustic Wave Equation in Polynomial Trefftz-DG Bases". PhD thesis. Université de Pau et des Pays l'Adour, 2018.

[87] E. Isaacson and H. B. Keller. *Analysis of numerical methods*. Courier Corporation, 2012.

[88] R. Courant, K. Friedrichs, and H. Lewy. "Über die partiellen Differenzengleichungen der mathematischen Physik". In: *Mathematische annalen* 100.1 (1928), pp. 32–74.

[89] C. Agut and J. Diaz. "Stability analysis of the Interior Penalty Discontinuous Galerkin method for the wave equation". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 47.3 (2013), pp. 903–932.

[90] J. C. Gilbert and P. Joly. "Higher order time stepping for second order hyperbolic problems and optimal CFL conditions". In: *Partial differential equations*. Springer, 2008, pp. 67–93.

[91] C. Agut, J. Diaz, and A. Ezziani. "High-order discretizations for the wave equation based on the modified equation technique". In: *10ème Congrès Français d'Acoustique*. 2010.

[92] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Section 17.1 Runge-Kutta Method*. 2007.

[93] H. P. Langtangen and S. Linge. *Finite difference computing with PDEs: a modern software approach*. Springer Nature, 2017.

[94] J. Robertsson and K. Holliger. "Modeling of seismic wave propagation near the earth's surface". In: *Physics of the Earth and Planetary Interiors* 104.1-3 (1997), pp. 193–211.

[95] M. E. Dougherty and R. A. Stephen. "Seismic energy partitioning and scattering in laterally heterogeneous ocean crust". In: *Scattering and Attenuations of Seismic Waves, Part I*. Springer, 1988, pp. 195–229.

[96] R. J. LeVeque. *Finite volume methods for hyperbolic problems*. Vol. 31. Cambridge university press, 2002.

[97] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*. Vol. 3. Springer, 2008.

[98] J. N. Reddy. *Introduction to the finite element method*. McGraw-Hill Education, 2019.

[99] W. H. Reed and T. R. Hill. *Triangular mesh methods for the neutron transport equation*. Tech. rep. Los Alamos Scientific Lab., N. Mex.(USA), 1973.

[100] P. Lesaint and P.-A. Raviart. "On a finite element method for solving the neutron transport equation". In: *Publications mathématiques et informatique de Rennes* S4 (1974), pp. 1–40.

[101] C. Johnson and J. Pitkäranta. "An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation". In: *Mathematics of computation* 46.173 (1986), pp. 1–26.

[102] T. Dupont. "L^2-estimates for Galerkin methods for second order hyperbolic equations". In: *SIAM Journal on Numerical Analysis* 10.5 (1973), pp. 880–889.

[103] G. A. Baker. "Error estimates for finite element methods for second order hyperbolic equations". In: *SIAM Journal on Numerical Analysis* 13.4 (1976), pp. 564–576.

[104] E. Bécache, P. Joly, and C. Tsogka. "An analysis of new mixed finite elements for the approximation of wave propagation problems". In: *SIAM Journal on Numerical Analysis* 37.4 (2000), pp. 1053–1084.

[105] B. Rivière. *Discontinuous Galerkin methods for solving elliptic and parabolic equations: theory and implementation*. SIAM, 2008.

[106] E. H. Georgoulis. "Discontinuous Galerkin methods for linear problems: An introduction". In: *Approximation Algorithms for Complex Systems*. Springer, 2011, pp. 91–126.

[107] B. Cockburn, G. E. Karniadakis, and C.-W. Shu. *Discontinuous Galerkin methods: theory, computation and applications*. Vol. 11. Springer Science & Business Media, 2012.

[108] M. J. Grote, A. Schneebeli, and D. Schötzau. "Discontinuous Galerkin finite element method for the wave equation". In: *SIAM Journal on Numerical Analysis* 44.6 (2006), pp. 2408–2431.

[109] C. Agut. "Schémas numériques d'ordre élevé en espace et en temps pour l'équation des ondes". PhD thesis. Université de Pau et des Pays l'Adour, 2011.

[110] J. Heinonen and P. Koskela. "Quasiconformal maps in metric spaces with controlled geometry". In: *Acta Mathematica* 181.1 (1998), pp. 1–61.

[111] C. Anitescu, C. Nguyen, T. Rabczuk, and X. Zhuang. "Isogeometric analysis for explicit elastodynamics using a dual-basis diagonal mass formulation". In: *Computer Methods in Applied Mechanics and Engineering* 346 (2019), pp. 574–591.

[112] L. C. Cowsar, T. F. Dupont, and M. F. Wheeler. "A priori estimates for mixed finite element approximations of second-order hyperbolic equations with absorbing boundary conditions". In: *SIAM Journal on Numerical Analysis* 33.2 (1996), pp. 492–504.

[113] K. Weierstrass. "Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen". In: *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin* 2 (1885), pp. 633–639.

[114] D. Jackson. *Über die Genauigkeit der Annäherung stetiger Funktionen durch ganze rationale Funktionen gegebenen Grades und trigonometrische Summen gegebener Ordnung.* Dieterich'schen Universität–Buchdruckerei, 1911.

[115] J. H. Bramble and S. Hilbert. "Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation". In: *SIAM Journal on Numerical Analysis* 7.1 (1970), pp. 112–124.

[116] J. Nitsche. "Verfahren von Ritz und Spline Interpolation bei Sturm-Liouville Randwertproblemen". In: *Numerische Mathematik* 13.3 (1969), pp. 260–265.

[117] C. Makridakis and R. H. Nochetto. "A posteriori error analysis for higher order dissipative methods for evolution problems". In: *Numerische Mathematik* 104.4 (2006), pp. 489–514.

[118] E. H. Georgoulis, O. Lakkis, C. G. Makridakis, and J. M. Virtanen. "A posteriori error estimates for leap-frog and cosine methods for second order evolution problems". In: *SIAM Journal on Numerical Analysis* 54.1 (2016), pp. 120–136.

[119] R. A. Horn and C. R. Johnson. *Matrix analysis.* Cambridge university press, 2012.

[120] J. Sherman. "Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix". In: *Annals of mathematical statistics* 20.4 (1949), p. 621.

[121] J. Diaz and M. J. Grote. "Energy conserving explicit local time stepping for second-order wave equations". In: *SIAM Journal on Scientific Computing* 31.3 (2009), pp. 1985–2014.

[122] H. Barucq, J. Diaz, and V. Duprat. "Stability analysis of the interior penalty discontinuous Galerkin method for solving the wave equation coupled with high-order absorbing boundary conditions". In: *Monografías del Seminario Matemático García de Galdeano* 35 (2010), pp. 49–56.

[123] C. Agut, J. Diaz, and A. Ezziani. "High-order schemes combining the modified equation approach and discontinuous Galerkin approximations for the wave equation". In: *Communications in Computational Physics* 11.2 (2012), pp. 691–708.

[124] J. L. Lions and E. Magenes. *Non-homogeneous boundary value problems and applications: Vol. 1.* Vol. 181. Springer Science & Business Media, 2012.

[125] K. Shahbazi. "An explicit expression for the penalty parameter of the interior penalty method". In: *Journal of Computational Physics* 205.2 (2005), pp. 401–407.

[126] Y. Epshteyn and B. Rivière. "Estimation of penalty parameters for symmetric interior penalty Galerkin methods". In: *Journal of Computational and Applied Mathematics* 206.2 (2007), pp. 843–872.

[127] B. Rivière, M. F. Wheeler, and V. Girault. "A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems". In: *SIAM Journal on Numerical Analysis* 39.3 (2001), pp. 902–931.

[128] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. "Unified analysis of discontinuous Galerkin methods for elliptic problems". In: *SIAM Journal on Numerical Analysis* 39.5 (2002), pp. 1749–1779.

[129] P. Wolfe. "Convergence conditions for ascent methods". In: *SIAM review* 11.2 (1969), pp. 226–235.

[130] P. Wolfe. "Convergence conditions for ascent methods. II: Some corrections". In: *SIAM review* 13.2 (1971), pp. 185–188.

[131] R. Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

[132] J. Yu, S. Vishwanathan, S. Günter, and N. N. Schraudolph. "A quasi-Newton approach to nonsmooth convex optimization problems in machine learning". In: *The Journal of Machine Learning Research* 11 (2010), pp. 1145–1200.

[133] K. E. Atkinson. *An introduction to numerical analysis*. John Wiley & sons, 2008.

[134] J. C. Butcher. *Numerical methods for ordinary differential equations*. John Wiley & Sons, 2016.

[135] N. Gödel, S. Schomann, T. Warburton, and M. Clemens. "GPU accelerated Adams-Bashforth multirate discontinuous Galerkin FEM simulation of high-frequency electromagnetic fields". In: *IEEE Transactions on magnetics* 46.8 (2010), pp. 2735–2738.

[136] J. Diaz and M. J. Grote. "Multi-level explicit local time-stepping methods for second-order wave equations". In: *Computer methods in applied mechanics and engineering* 291 (2015), pp. 240–265.

[137] M. Rietmann, D. Peter, O. Schenk, B. Uçar, and M. Grote. "Load-balanced local time stepping for large-scale wave propagation". In: *2015 IEEE International Parallel and Distributed Processing Symposium*. IEEE. 2015, pp. 925–935.

[138] M. N'Diaye. "On the study and development of high-order time integration schemes for ODEs applied to acoustic and electromagnetic wave propagation problems." PhD thesis. Université de Pau et des Pays l'Adour, 2017.

[139] H. Barucq, M. Duruflé, and M. N'Diaye. "High-order locally A-stable implicit schemes for linear ODEs". In: *Journal of Scientific Computing* 85.2 (2020), pp. 1–33.

[140] K. Atkinson and W. Han. *Theoretical numerical analysis*. Vol. 39. Springer, 2005.

[141] P. Monk. *Finite element methods for Maxwell's equations*. Oxford University Press, 2003.

[142] M. Ainsworth and J. T. Oden. "A posteriori error estimation in finite element analysis". In: *Computer methods in applied mechanics and engineering* 142.1-2 (1997), pp. 1–88.

[143] J. S. Hesthaven and T. Warburton. *Nodal discontinuous Galerkin methods: algorithms, analysis, and applications*. Springer Science & Business Media, 2007.

[144] B. Cockburn and C.-W. Shu. "TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework". In: *Mathematics of computation* 52.186 (1989), pp. 411–435.

[145] B. Cockburn, S.-Y. Lin, and C.-W. Shu. "TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One-dimensional systems". In: *Journal of Computational Physics* 84.1 (1989), pp. 90–113.

[146] B. Cockburn, S. Hou, and C.-W. Shu. "The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. IV. The multidimensional case". In: *Mathematics of Computation* 54.190 (1990), pp. 545–581.

[147] B. Cockburn and C.-W. Shu. "The Runge-Kutta discontinuous Galerkin method for conservation laws V: Multidimensional systems". In: *Journal of Computational Physics* 141.2 (1998), pp. 199–224.

[148] B. Riviere. *Discontinuous Galerkin methods for solving the miscible displacement problem in porous media*. The University of Texas at Austin, 2000.

[149] S. Sun. *Discontinuous Galerkin methods for reactive transport in porous media*. The University of Texas at Austin, 2003.

[150] C. Dawson, S. Sun, and M. F. Wheeler. "Compatible algorithms for coupled flow and transport". In: *Computer Methods in Applied Mechanics and Engineering* 193.23-26 (2004), pp. 2565–2580.

[151] M. Bonnasse-Gahot. "High order discontinuous Galerkin methods for time-harmonic elastodynamics". PhD thesis. Université Nice Sophia Antipolis, 2015.

[152] C. Runge. "Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten". In: *Zeitschrift für Mathematik und Physik* 46.224-243 (1901), p. 20.

[153] T. Lyche and K. Mørken. "Spline methods draft". In: *Department of Informatics, Center of Mathematics for Applications, University of Oslo, Oslo* (2008).

[154] A. Turetskii. "The bounding of polynomials prescribed at equally distributed points". In: *Proc. Pedag. Inst. Vitebsk*. Vol. 3. 1940, pp. 117–127.

[155] T. Mills and S. Smith. "The Lebesgue constant for Lagrange interpolation on equidistant nodes". In: *Numerische Mathematik* 61.1 (1992), pp. 111–115.

[156] H. Wang and S. Xiang. "On the convergence rates of Legendre approximation". In: *Mathematics of Computation* 81.278 (2012), pp. 861–877.

[157] S. J. Smith. "Lebesgue constants in polynomial interpolation". In: *Annales Mathematicae et Informaticae*. Vol. 33. 109-123. Eszterházy Károly College, Institute of Mathematics and Computer Science. 2006, pp. 1787–5021.

[158] G. Szegő. *Orthogonal polynomials*. Vol. 23. American Mathematical Soc., 1939.

[159] T. Warburton and T. Hagstrom. "Taming the CFL number for discontinuous Galerkin methods on structured meshes". In: *SIAM Journal on Numerical Analysis* 46.6 (2008), pp. 3151–3180.

[160] D. Dunavant. "High degree efficient symmetrical Gaussian quadrature rules for the triangle". In: *International Journal for Numerical Methods in Engineering* 21.6 (1985), pp. 1129–1148.

[161] G. Cohen, P. Joly, J. E. Roberts, and N. Tordjman. "Higher order triangular finite elements with mass lumping for the wave equation". In: *SIAM Journal on Numerical Analysis* 38.6 (2001), pp. 2047–2078.

[162] M. Blyth and C. Pozrikidis. "A Lobatto interpolation grid over the triangle". In: *IMA Journal of Applied Mathematics* 71.1 (2006), pp. 153–169.

[163] Y. Liu, J. Teng, T. Xu, and J. Badal. "Higher-order triangular spectral element method with optimized cubature points for seismic wavefield modeling". In: *Journal of Computational Physics* 336 (2017), pp. 458–480.

[164] R. Pasquetti and F. Rapetti. "Cubature versus Fekete-Gauss nodes for spectral element methods on simplicial meshes". In: *Journal of Computational Physics* (2017).

[165] P. J. Davis. *Interpolation and approximation*. Courier Corporation, 1975.

[166] M. Abramowitz, I. A. Stegun, and R. H. Romer. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. 1988.

[167] S. A. Teukolsky. "Short note on the mass matrix for Gauss–Lobatto grid points". In: *Journal of Computational Physics* 283 (2015), pp. 408–413.

[168] D. Komatitsch and J.-P. Vilotte. "The spectral element method: an efficient tool to simulate the seismic response of 2D and 3D geological structures". In: *Bulletin of the seismological society of America* 88.2 (1998), pp. 368–392.

[169] A. Citrain. "Hybrid finite element methods for seismic wave simulation: coupling of discontinuous Galerkin and spectral element discretizations". PhD thesis. Normandie Université, 2019.

[170] R. T. Farouki. "The Bernstein polynomial basis: A centennial retrospective". In: *Computer Aided Geometric Design* 29.6 (2012), pp. 379–419.

[171] P. De Casteljau. "Outillages méthodes calcul". In: *André Citroën Automobiles SA, Paris* (1959).

[172] M. S. Floater. "Mean value coordinates". In: *Computer aided geometric design* 20.1 (2003), pp. 19–27.

[173] M. G. Cox. "The numerical evaluation of B-splines". In: *IMA Journal of Applied Mathematics* 10.2 (1972), pp. 134–149.

[174] C. De Boor. "On calculating with B-splines". In: *Journal of Approximation theory* 6.1 (1972), pp. 50–62.

[175] L. Piegl and W. Tiller. *The NURBS book*. Springer Science & Business Media, 2012.

[176] L. Schumaker. *Spline functions: basic theory*. Cambridge University Press, 2007.

[177] M. J. Marsden. "An identity for spline functions with applications to variation diminishing spline approximation". In: *Journal of Approximation Theory* 3.1 (1970), pp. 7–49.

[178] C. De Boor. *A practical guide to splines*. Vol. 27. Springer-Verlag New York, 1978.

[179] C. de Boor. "Divided differences". In: *Surveys in Approximation Theory* 1 (2005), pp. 46–69.

[180] E. T. Whittaker and G. Robinson. "Divided differences & theorems on divided differences". In: *The Calculus of Observations: A Treatise on Numerical Mathematics, 4th ed., New York* (1967), pp. 20–24.

[181] M. B. Allen and E. L. Isaacson. *Numerical analysis for applied science*. Vol. 35. John Wiley & Sons, 2011.

[182] C. de Boor, T. Lyche, and L. L. Schumaker. "On calculating with B-splines II. Integration". In: *Numerische Methoden der Approximationstheorie/Numerical Methods of Approximation Theory*. Springer, 1976, pp. 123–146.

[183] A. H. Vermeulen, R. H. Bartels, and G. R. Heppler. "Integrating products of B-splines". In: *SIAM Journal on Scientific and Statistical Computing* 13.4 (1992), pp. 1025–1038.

[184] T. N. E. Greville. "Introduction to spline functions". In: *Theory and applications of spline functions* (1969), pp. 1–35.

[185] H. B. Curry and I. J. Schoenberg. "On Pólya frequency functions IV: the fundamental spline functions and their limits". In: *Journal d'analyse mathématique* 17.1 (1966), pp. 71–107.

[186] E. Neuman. "Moments and Fourier transforms of B-splines". In: *Journal of Computational and Applied Mathematics* 7.1 (1981), pp. 51–62.

[187] I. J. Schoenberg. *Cardinal spline interpolation*. Vol. 12. Siam, 1973.

[188] T.-X. He. "Eulerian polynomials and B-splines". In: *Journal of Computational and Applied Mathematics* 236.15 (2012), pp. 3763–3773.

[189] K. Gahalaut and S. Tomar. "Condition number estimates for matrices arising in the isogeometric discretizations". In: *RICAM report* 23.2012 (2012), pp. 1–38.

[190] J. S. Hesthaven. "From electrostatics to almost optimal nodal sets for polynomial interpolation in a simplex". In: *SIAM Journal on Numerical Analysis* 35.2 (1998), pp. 655–676.

[191] M. Fekete. "Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten". In: *Mathematische Zeitschrift* 17.1 (1923), pp. 228–249.

[192] S. Smale. "Mathematical problems for the next century". In: *The mathematical intelligencer* 20.2 (1998), pp. 7–15.

[193] J. Chan, R. J. Hewett, and T. Warburton. "Weight-adjusted discontinuous Galerkin methods: wave propagation in heterogeneous media". In: *SIAM Journal on Scientific Computing* 39.6 (2017), A2935–A2961.

[194] J. W. Cahn and J. E. Hilliard. "Free energy of a nonuniform system. I. Interfacial free energy". In: *The Journal of chemical physics* 28.2 (1958), pp. 258–267.

[195] E. Reissner. "The effect of transverse shear deformation on the bending of elastic plates". In: *Journal of Applied Mechanics* 12.2 (1945), A69–A77.

[196] R. D. Mindlin. "Influence of rotatory inertia and shear on flexural motions of isotropic, elastic plates". In: *Journal of Applied Mechanics* 18.1 (1951), pp. 31–38.

[197] T. J. Willmore. "Note on embedded surfaces". In: *An. Sti. Univ."Al. I. Cuza" Iasi Sect. I a Mat.(NS) B* 11.493-496 (1965), p. 18.

[198] F. Auricchio, F. Calabro, T. J. Hughes, A. Reali, and G. Sangalli. "A simple algorithm for obtaining nearly optimal quadrature rules for NURBS-based isogeometric analysis". In: *Computer Methods in Applied Mechanics and Engineering* 249 (2012), pp. 15–27.

[199] M. Bartoň and V. M. Calo. "Optimal quadrature rules for odd-degree spline spaces and their application to tensor-product-based isogeometric analysis". In: *Computer Methods in Applied Mechanics and Engineering* 305 (2016), pp. 217–240.

[200] F. Calabro, G. Sangalli, and M. Tani. "Fast formation of isogeometric Galerkin matrices by weighted quadrature". In: *Computer Methods in Applied Mechanics and Engineering* 316 (2017), pp. 606–622.

[201] M. Bartoň, V. Puzyrev, Q. Deng, and V. Calo. "Efficient mass and stiffness matrix assembly via weighted Gaussian quadrature rules for B-splines". In: *Journal of Computational and Applied Mathematics* 371 (2020), p. 112626.

[202] P. Gervasio, L. Dedè, O. Chanon, and A. Quarteroni. "A computational comparison between isogeometric analysis and spectral element methods: accuracy and spectral properties". In: *Journal of Scientific Computing* 83.1 (2020), pp. 1–45.

[203] H. Prautzsch, W. Boehm, and M. Paluszny. *Bézier and B-spline techniques*. Springer Science & Business Media, 2013.

[204] C. de Boor. "Splines as linear combinations of B-splines". In: *Approximation Theory II* (1976), pp. 1–47.

[205] C. A. Micchelli. "A constructive approach to Kergin interpolation in $\mathbb{R}^k$: multivariate B-splines and Lagrange interpolation". In: *The Rocky Mountain Journal of Mathematics* (1980), pp. 485–497.

[206] T. A. Grandine. "The computational cost of simplex spline functions". In: *SIAM Journal on Numerical Analysis* 24.4 (1987), pp. 887–890.

[207] T. A. Grandine. "The stable evaluation of multivariate simplex splines". In: *Mathematics of Computation* 50.181 (1988), pp. 197–205.

[208] B. C. Carlson. *Special Functions of Applied Mathematics*. Academic Press, 1977. ISBN: 9780121601508.

[209] B. C. Carlson. "B-splines, hypergeometric functions, and Dirichlet averages". In: *Journal of approximation theory* 67.3 (1991), pp. 311–325.

[210] W. Zu Castell. "Dirichlet splines as fractional integrals of B-splines". In: *The Rocky Mountain Journal of Mathematics* (2002), pp. 545–559.

[211] D. G. Zill and B. C. Carlson. "Symmetric elliptic integrals of the third kind". In: *Mathematics of computation* 24.109 (1970), pp. 199–214.

[212] L. A. Piegl and W. Tiller. "Computing the derivative of NURBS with respect to a knot". In: *Computer aided geometric design* 15.9 (1998), pp. 925–934.

[213] L. Guibas and J. Stolfi. "Primitives for the manipulation of general subdivisions and the computation of Voronoi diagrams". In: *ACM transactions on graphics (TOG)* 4.2 (1985), pp. 74–123.

[214] W. Dahmen and C. A. Micchelli. *Recent progress in multivariate splines.* IBM Thomas J. Watson Research Center, 1983.

[215] C. A. Micchelli. *Mathematical aspects of geometric modeling.* SIAM, 1995.

[216] T. N. T. Goodman. "Polyhedral splines". In: *Computation of curves and surfaces.* Springer, 1990, pp. 347–382.

[217] M. D. McCool. "Analytic antialiasing with prism splines". In: *SIGGRAPH.* Vol. 95. Citeseer. 1995, pp. 429–436.

[218] M. D. McCool. *Analytic signal processing for computer graphics using multivariate polyhedral splines.* University of Toronto, 1995.

[219] C. De Concini, C. Procesi, and M. Vergne. "Box splines and the equivariant index theorem". In: *Journal of the Institute of Mathematics of Jussieu* 12.3 (2013), pp. 503–544.

[220] C. De Boor, K. Höllig, and S. Riemenschneider. *Box splines.* Vol. 98. Springer Science & Business Media, 2013.

[221] M. Lenz. "Interpolation, box splines, and lattice points in zonotopes". In: *International Mathematics Research Notices* 2014.20 (2014), pp. 5697–5712.

[222] N. Villamizar, A. Mantzaflaris, and B. Jüttler. "Completeness characterization of Type-I box splines". In: *arXiv preprint arXiv:2011.01919* (2020).

[223] J. Radon. "On the determination of functions from their integral values along certain manifolds". In: *IEEE transactions on medical imaging* 5.4 (1986), pp. 170–176.

[224] S. Horbelt, M. Liebling, and M. Unser. "Discretization of the Radon transform and of its inverse by spline convolutions". In: *IEEE Transactions on medical imaging* 21.4 (2002), pp. 363–376.

[225] A. Narayanan. "Algorithm AS 266: maximum likelihood estimation of the parameters of the Dirichlet distribution". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 40.2 (1991), pp. 365–374.

[226] D. Blackwell and J. B. MacQueen. "Ferguson distributions via Pólya urn schemes". In: *The annals of statistics* 1.2 (1973), pp. 353–355.

[227] N. Berline and M. Vergne. "Classes caractéristiques équivariantes. Formule de localisation en cohomologie équivariante". In: *CR Acad. Sci. Paris* 295.2 (1982), pp. 539–541.

[228] M. F. Atiyah and R. Bott. "The moment map and equivariant cohomology". In: *Topology* 23.1 (1984), pp. 1–28.

[229] J. J. Duistermaat and G. J. Heckman. "On the variation in the cohomology of the symplectic form of the reduced phase space". In: *Inventiones mathematicae* 69.2 (1982), pp. 259–268.

[230] T. Delzant. "Hamiltoniens périodiques et images convexes de l'application moment". In: *Bulletin de la Société mathématique de France* 116.3 (1988), pp. 315–339.

[231] M. Vergne. "Applications of equivariant cohomology". In: *Proceedings of the International Congress of Mathematicians Madrid, August 22–30, 2006*. 2007, pp. 635–664.

[232] M. Vaitkus. "A Physical Perspective on Control Points and Polar Forms: Bézier Curves, Angular Momentum and Harmonic Oscillators". In: *arXiv preprint arXiv:1809.07287* (2018).

[233] C. F. Dunkl, P. Gawron, J. A. Holbrook, Z. Puchała, and K. Życzkowski. "Numerical shadows: measures and densities on the numerical range". In: *Linear Algebra and its Applications* 434.9 (2011), pp. 2042–2080.

[234] C. F. Dunkl, P. Gawron, J. A. Holbrook, J. A. Miszczak, Z. Puchała, and K. Życzkowski. "Numerical shadow and geometry of quantum states". In: *Journal of Physics A: Mathematical and Theoretical* 44.33 (2011), p. 335301.

[235] C. F. Dunkl, P. Gawron, Ł. Pawela, Z. Puchała, and K. Życzkowski. "Real numerical shadow and generalized B-splines". In: *Linear Algebra and its Applications* 479 (2015), pp. 12–51.

[236] T. Gallay and D. Serre. "Numerical measure of a complex matrix". In: *Communications on Pure and Applied Mathematics* 65.3 (2012), pp. 287–336.

[237] L. C. Venuti and P. Zanardi. "Probability density of quantum expectation values". In: *Physics Letters A* 377.31-33 (2013), pp. 1854–1861.

[238] M. F. Paulos. "Loops, polytopes and splines". In: *Journal of High Energy Physics* 2013.6 (2013), p. 7.

[239] W. Dahmen and C. A. Micchelli. "Statistical encounters with B-splines". In: *Contemp. Math* 59 (1986), pp. 17–48.

[240] S. Karlin, C. A. Micchelli, and Y. Rinott. "Multivariate splines: A probabilistic perspective". In: *Journal of Multivariate Analysis* 20.1 (1986), pp. 69–90.

[241] S. Cambanis, R. Keener, and G. Simons. "On α-symmetric multivariate distributions". In: *Journal of Multivariate Analysis* 13.2 (1983), pp. 213–233.

[242] W. Z. Castell. "Fractional derivatives and the inverse Fourier transform of ℓ_1-radial functions". In: *Integral Transforms and Special Functions* 15.3 (2004), pp. 209–223.

[243] H. Berens and Y. Xu. "ℓ_1 summability of multiple Fourier integrals and positivity". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 122. 1. Cambridge University Press. 1997, pp. 149–172.

[244] T. Glüsenkamp. "Probabilistic treatment of the uncertainty from the finite size of weighted Monte Carlo data". In: *The European Physical Journal Plus* 133.6 (2018), p. 218.

[245] P. Sablonnière. "Univariate spline quasi-interpolants and applications to numerical analysis". In: *Rendiconti del Seminario Matematico* 63.3 (2005), pp. 211–222.

[246] B. Forster and P. Massopust. "Statistical encounters with complex B-Splines". In: *Constructive Approximation* 29.3 (2009), pp. 325–344.

[247] P. Massopust and B. Forster. "Multivariate complex B-splines and Dirichlet averages". In: *Journal of Approximation Theory* 162.2 (2010), pp. 252–269.

[248] B. Forster, T. Blu, and M. Unser. "Complex B-splines". In: *Applied and Computational Harmonic Analysis* 20.2 (2006), pp. 261–282.

[249] M. Saigo. "A certain boundary value problem for the Euler-Darboux equation". In: *Mathematica Japonica* 24 (1979), pp. 377–385.

[250] W. Miller Jr. "Symmetries of differential equations. The hypergeometric and Euler-Darboux equations". In: *SIAM Journal on Mathematical Analysis* 4.2 (1973), pp. 314–328.

[251] W. Miller. *Lie theory and special functions*. Academic Press, 1968.

[252] J. A. Cottrell, T. J. Hughes, and Y. Bazilevs. *Isogeometric analysis: toward integration of CAD and FEA*. John Wiley & Sons, 2009.

[253] T. Goodman and S. Lee. "Spline approximation operators of Bernstein-Schoenberg type in one and two variables". In: *Journal of Approximation Theory* 33.3 (1981), pp. 248–263.

[254] W. Dahmen. "Polynomials as linear combinations of multivariate B-splines". In: *Mathematische Zeitschrift* 169.1 (1979), pp. 93–98.

[255] W. Dahmen and C. A. Micchelli. "On the linear independence of multivariate B-splines, I. Triangulations of simploids". In: *SIAM Journal on Numerical Analysis* 19.5 (1982), pp. 993–1012.

[256] M. Neamtu. "What is the natural generalization of univariate splines to higher dimensions?" In: *Mathematical methods for curves and surfaces* (2001), pp. 355–392.

[257] W. Dahmen, C. A. Micchelli, and H.-P. Seidel. "Blossoming begets B-pline bases built better by B-patches". In: *Mathematics of computation* 59.199 (1992), pp. 97–115.

[258] M. Neamtu. "Bivariate simplex B-splines: A new paradigm". In: *Proceedings Spring Conference on Computer Graphics*. IEEE. 2001, pp. 71–78.

[259] Y. Liu. "Computations of Delaunay and higher order triangulations, with applications to splines". PhD thesis. University of North Carolina, Chapel Hill, 2008.

[260] D.-T. Lee. "On k-nearest neighbor Voronoi diagrams in the plane". In: *IEEE transactions on computers* 100.6 (1982), pp. 478–487.

[261] T. Lyche and G. Muntingh. "Stable simplex spline bases for C^3 quintics on the Powell-Sabin 12-split". In: *Constructive approximation* 45.1 (2017), pp. 1–32.

[262] C. Bracco, T. Lyche, C. Manni, F. Roman, and H. Speleers. "Generalized spline spaces over T-meshes: Dimension formula and locally refined generalized B-splines". In: *Applied Mathematics and Computation* 272 (2016), pp. 187–198.

[263] E. Schönhardt. "Über die Zerlegung von Dreieckspolyedern in Tetraeder". In: *Mathematische Annalen* 98.1 (1928), pp. 309–312.

[264] J. Rambau. "On a generalization of Schönhardt's polyhedron". In: *Combinatorial and computational geometry* 52 (2003), pp. 510–516.

[265] J. Pellerin, K. Verhetsel, and J.-F. Remacle. "There are 174 Subdivisions of the Hexahedron into Tetrahedra". In: *ACM Transactions on Graphics (TOG)* 37.6 (2018), pp. 1–9.

[266] J. Richter-Gebert and G. M. Ziegler. "Zonotopal tilings and the Bohne-Dress theorem". In: *Contemporary Mathematics* 178 (1994), pp. 211–211.

[267] G. M. Ziegler. *Lectures on polytopes*. Vol. 152. Springer Science & Business Media, 2012.

[268] A. Björner, M. Las Vergnas, B. Sturmfels, N. White, and G. M. Ziegler. *Oriented matroids*. 46. Cambridge University Press, 1999.

[269] G. C. Shephard. "Combinatorial properties of associated zonotopes". In: *Canadian Journal of Mathematics* 26.2 (1974), pp. 302–321.

[270] P. Galashin, A. Postnikov, and L. Williams. "Higher secondary polytopes and regular plabic graphs". In: *arXiv preprint 1909.05435* (2019).

[271] L. Ramshaw. "Blossoms are polar forms". In: *Computer Aided Geometric Design* 6.4 (1989), pp. 323–358.

[272] D. Schmitt and J.-C. Spehner. "On Delaunay and Voronoi diagrams of order k in the plane". In: *Proc. 3rd Canad. Conf. Comput. Geom.* 1991, pp. 29–32.

[273] D. Schmitt and J.-C. Spehner. "Order-k Voronoi diagrams, k-sections, and k-sets". In: *Japanese Conference on Discrete and Computational Geometry*. Springer. 1998, pp. 290–304.

[274] W. El Oraiby, D. Schmitt, and J.-C. Spehner. "Centroid triangulations from k-sets". In: *International Journal of Computational Geometry & Applications* 21.06 (2011), pp. 635–659.

[275] H. Edelsbrunner, P. Valtr, and E. Welzl. "Cutting dense point sets in half". In: *Discrete & Computational Geometry* 17.3 (1997), pp. 243–255.

[276] D. Schmitt and J.-C. Spehner. "k-set polytopes and order-k Delaunay diagrams". In: *2006 3rd International Symposium on Voronoi Diagrams in Science and Engineering*. IEEE. 2006, pp. 173–185.

[277] J. A. Olarte and F. Santos. "Hypersimplicial subdivisions". In: *Selecta Mathematica* 28.1 (2022), pp. 1–34.

[278] E. Stiemke. "Über positive Lösungen homogener linearer Gleichungen". In: *Mathematische Annalen* 76.2 (1915), pp. 340–342.

[279] H. Edelsbrunner and G. Osang. "The multi-cover persistence of euclidean balls". In: *34th International Symposium on Computational Geometry (SoCG 2018)*. Vol. 99. Leibniz International Proceedings in Informatics (LIPIcs). 2018, 34:1–34:14. ISBN: 978-3-95977-066-8.

[280] H. Edelsbrunner and A. Nikitenko. "Poisson-Delaunay mosaics of order k". In: *Discrete & computational geometry* 62.4 (2019), pp. 865–878.

[281] F. Santos. "On Delaunay oriented matroids for convex distance functions". In: *Discrete & Computational Geometry* 16.2 (1996), pp. 197–210.

[282] L. J. Billera and B. Sturmfels. "Fiber polytopes". In: *Annals of Mathematics* (1992), pp. 527–549.

[283] A. Guttman. "R-trees: a dynamic index structure for spatial searching". In: *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*. 1984, pp. 47–57.

[284] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. "The $R^\star$-tree: an efficient and robust access method for points and rectangles". In: *Proceedings of the 1990 ACM SIGMOD international conference on Management of data*. 1990, pp. 322–331.

[285] M. Beck and R. Sanyal. *Combinatorial reciprocity theorems*. Vol. 195. American Mathematical Soc., 2018.

[286] T. Lyche and K. Scherer. "On the p-norm condition number of the multivariate triangular Bernstein basis". In: *Journal of computational and applied mathematics* 119.1-2 (2000), pp. 259–273.

[287] P. De Casteljau. "Courbes et surfaces à pôles". In: *André Citroën, Automobiles SA, Paris* (1963).

[288] H. Edelsbrunner and D. Guoy. "An experimental study of sliver exudation". In: *Engineering with computers* 18.3 (2002), pp. 229–240.

[289] P. Mullen, P. Memari, F. de Goes, and M. Desbrun. "HOT: Hodge-optimized triangulations". In: *ACM SIGGRAPH 2011 papers*. 2011, pp. 1–12.

[290] H. Barucq, H. Calandra, J. Diaz, and S. Frambati. "Polynomial-reproducing spline spaces from fine zonotopal tilings". In: *Journal of Computational and Applied Mathematics* 402 (2022), p. 113812.

[291] C. De Concini and C. Procesi. *Topics in hyperplane arrangements, polytopes and box-splines*. Springer Science & Business Media, 2010.

[292] M. Lenz. "Zonotopal algebra and forward exchange matroids". In: *Advances in Mathematics* 294 (2016), pp. 819–852.

[293] J. Bohne. "Eine kombinatorische analyse zonotopaler raumaufteilungen". PhD thesis. Sonderforschungsbereich 343, 1992.

[294] A. W. Dress. "Oriented matroids and Penrose-type tilings". In: *Lecture at the "Symposium on Combinatorics and Geometry", organized by A. Björner, KTH Stockholm*. 1989.

[295] G. Xu, B. Mourrain, R. Duvigneau, and A. Galligo. "Parameterization of computational domain in isogeometric analysis: methods and comparison". In: *Computer Methods in Applied Mechanics and Engineering* 200.23-24 (2011), pp. 2021–2031.

[296] G. Xu, B. Mourrain, R. Duvigneau, and A. Galligo. "Analysis-suitable volume parameterization of multi-block computational domain in isogeometric applications". In: *Computer-Aided Design* 45.2 (2013), pp. 395–404.

[297] T. J. Alumbaugh and X. Jiao. "Compact array-based mesh data structures". In: *Proceedings of the 14th International Meshing Roundtable*. Springer. 2005, pp. 485–503.

[298] G. Damiand and P. Lienhardt. *Combinatorial maps: efficient data structures for computer graphics and image processing*. CRC Press, 2014.

[299] J. R. Shewchuk. "General-dimensional constrained Delaunay and constrained regular triangulations, I: Combinatorial properties". In: *Twentieth Anniversary Volume:* Springer, 2009, pp. 1–58.

[300] J. Ruppert and R. Seidel. "On the difficulty of triangulating three-dimensional nonconvex polyhedra". In: *Discrete & Computational Geometry* 7.3 (1992), pp. 227–253.

[301] K. R. Gabriel and R. R. Sokal. "A new statistical approach to geographic variation analysis". In: *Systematic zoology* 18.3 (1969), pp. 259–278.

[302] M. Alexa. "Conforming weighted Delaunay triangulations". In: *ACM Transactions on Graphics (TOG)* 39.6 (2020), pp. 1–16.

[303] J. R. Shewchuk. "Mesh generation for domains with small angles". In: *Proceedings of the sixteenth annual Symposium on Computational Geometry*. 2000, pp. 1–10.

[304] D. Cohen-Steiner, E. C. De Verdière, and M. Yvinec. "Conforming Delaunay triangulations in 3D". In: *Computational Geometry* 28.2-3 (2004), pp. 217–233.

[305] T. Warburton and J. S. Hesthaven. "On the constants in *hp*-finite element trace inverse inequalities". In: *Computer methods in applied mechanics and engineering* 192.25 (2003), pp. 2765–2773.

[306] P. Castillo. "Performance of discontinuous Galerkin methods for elliptic PDEs". In: *SIAM Journal on Scientific Computing* 24.2 (2002), pp. 524–547.

[307] D. N. Arnold. "An interior penalty finite element method with discontinuous elements". In: *SIAM Journal on Numerical Analysis* 19.4 (1982), pp. 742–760.

[308] D. Schötzau, C. Schwab, and A. Toselli. "Mixed *hp*-DGFEM for incompressible flows". In: *SIAM Journal on Numerical Analysis* 40.6 (2002), pp. 2171–2194.

[309] J. Diaz and A. Ezziani. *Gar6more 2d*. 2008. URL: `http://gar6more2d.gforge.inria.fr`.

[310] J. W. Banks and T. Hagstrom. "On Galerkin difference methods". In: *Journal of Computational Physics* 313 (2016), pp. 310–327.

[311] L. Gizon, H. Barucq, M. Duruflé, C. S. Hanson, M. Leguèbe, A. C. Birch, J. Chabassier, D. Fournier, T. Hohage, and E. Papini. "Computational helioseismology in the frequency domain: acoustic waves in axisymmetric solar models with flows". In: *Astronomy & Astrophysics* 600 (2017), A35.

[312] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa. "The quickhull algorithm for convex hulls". In: *ACM Transactions on Mathematical Software (TOMS)* 22.4 (1996), pp. 469–483.

[313] S. Hert and S. Schirra. "3D Convex Hulls". In: *CGAL User and Reference Manual*. 5.3. CGAL Editorial Board, 2021. URL: `https://doc.cgal.org/5.3/Manual/packages.html#PkgConvexHull3`.

[314] The CGAL Project. *CGAL User and Reference Manual*. 5.3. CGAL Editorial Board, 2021. URL: `https://doc.cgal.org/5.3/Manual/packages.html`.

[315] R. Abgrall, P. Bacigaluppi, and S. Tokareva. "How to avoid mass matrix for linear hyperbolic problems". In: *Numerical Mathematics and Advanced Applications ENUMATH 2015*. Springer, 2016, pp. 75–86.

[316] R. Abgrall. "High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices". In: *Journal of Scientific Computing* 73.2 (2017), pp. 461–494.

[317] M. L. Minion. "Semi-implicit spectral deferred correction methods for ordinary differential equations". In: *Communications in Mathematical Sciences* 1.3 (2003), pp. 471–500.

[318] M. Neamtu. "Multivariate divided differences and B-splines". In: *Approximation Theory VI* (1990), pp. 445–448.

[319] M. Neamtu and C. R. Traas. "On computational aspects of simplicial splines". In: *Constructive approximation* 7.1 (1991), pp. 209–220.

[320] E. Cohen, T. Lyche, and R. F. Riesenfeld. "Cones and recurrence relations for simplex splines". In: *Constructive Approximation* 3.1 (1987), pp. 131–141.

[321] C. K. Chui. *Multivariate splines*. Vol. 54. Siam, 1988.

[322] T. A. Grandine. "The evaluation of inner products of multivariate simplex splines". In: *SIAM Journal on Numerical Analysis* 24.4 (1987), pp. 882–886.