Mathematical and numerical methods for three-dimensional reflective tomography and for approximation on the sphere

Jean-Baptiste Bellet

Université de Lorraine, Ecole Doctorale IAEM-Lorraine

Mémoire d'Habilitation à Diriger des Recherches

Spécialité Mathématiques

par

Jean-Baptiste Bellet

Université de Lorraine, CNRS, IECL, F-57000 Metz, France

Méthodes mathématiques et numériques pour la tomographie réflective tri-dimensionnelle et pour l'approximation sur la sphère

Mathematical and numerical methods for three-dimensional reflective tomography and for approximation on the sphere

Soutenu le 13 octobre 2023, devant le jury composé de :

Gérard Berginc	Thales LAS	*Examinateur*
Antonin Chambolle	CNRS, Université Paris-Dauphine	*Rapporteur*
Jean-Pierre Croisille	Université de Lorraine	*Parrain scientifique*
Agnès Desolneux	CNRS, ENS Paris-Saclay	*Rapportrice*
Carole Le Guyader	INSA de Rouen	*Présidente du jury*
Victor Nistor	Université de Lorraine	*Examinateur*
Nir Sochen	University of Tel Aviv, Israël	*Rapporteur*

Remerciements

Je remercie les membres du jury pour l'intérêt qu'ils ont porté à ce travail et pour le temps qu'ils ont consacré à son évaluation. Je remercie les rapporteurs, Antonin Chambolle, Agnès Desolneux et Nir Sochen, pour leur examen attentif. Je remercie Carole Le Guyader d'avoir consciencieusement présidé ce jury. Je remercie mon collègue Victor Nistor d'y avoir pris part.

Je remercie mon parrain scientifique Jean-Pierre Croisille. Pour son soutien, depuis ma candidature au poste de Maître de Conférences à l'Université de Metz, jusqu'à la soutenance de ce mémoire. Aussi, pour les nombreux échanges scientifiques, à propos de la Cubed Sphere, mais aussi tant d'autres.

Je remercie Gérard Berginc, collaborateur de longue date. Pour sa confiance, depuis notre premier entretien[1] chez Thales Optronique. Et bien sûr, pour m'avoir initié à la tomographie réflective tri-dimensionnelle.

Je remercie mes collaborateurs pour nos interactions scientifiques et humaines.

Je remercie les collègues de l'Université de Lorraine avec qui j'ai l'occasion de passer d'agréables moments, notamment autour d'un café, ou d'une assiette de frites.

Je remercie mes amis pour leur fidélité, ma famille pour son soutien.

Merci à Céline, Ariane et Joseph.

[1] Entretien d'embauche, combiné à une sortie moto Rouen-Guyancourt, devenue légendaire.

Résumé. Cette thèse porte sur des aspects mathématiques et numériques de la tomographie réflective et de l'approximation sur la sphère. Le premier sujet concerne le calcul de reconstructions tri-dimensionnelles en imagerie optique à l'aide de transformées de type Radon. Les travaux présentés, en partie issus de collaborations avec des entreprises, incluent des aspects appliqués comme le développement et l'implémentation d'algorithmes, des tests numériques, ainsi qu'une méthode brevetée. D'un point de vue plus théorique, nous étayons mathématiquement le sujet en examinant les singularités ; notamment, l'analyse microlocale de la transformation de Radon établit une correspondance entre les singularités de la reconstruction et celles des données. Enfin, nous relions diffusion Lambertienne et transformation de Radon au sens des distributions. La deuxième partie porte sur l'approximation sur la sphère dans des bases d'harmoniques sphériques, dans le cas où la grille de discrétisation est la Cubed Sphere. On construit un interpolant de Lagrange qui est minimal pour un certain ordre lexicographique, avec pour application une formule de quadrature précise. On étudie également des problèmes de moindres carrés non régularisés, avec pour application une transformation de Funk-Radon discrète stable. En parallèle, différents résultats d'invariance par le groupe de symétrie du cube sont montrés et exploités, tandis que la structure en grands cercles de la grille est mise à profit dans l'étude de matrices de Vandermonde.

Abstract. This thesis deals with mathematical and numerical aspects of reflective tomography and approximation on the sphere. The first topic concerns the calculation of three-dimensional reconstructions in optical imaging using Radon-type transforms. The work presented, which is partly the result of collaborations with companies, includes applied aspects such as the development and implementation of algorithms, numerical tests and a patented method. From a more theoretical point of view, we provide mathematical support for the subject by examining the singularities; in particular, the microlocal analysis of the Radon transform establishes a correspondence between the singularities of the reconstruction and those of the data. Finally, we link Lambertian diffusion and the Radon transform extended to distributions. The second part deals with approximation on the sphere in spherical harmonics bases, in the case where the discretization grid is the Cubed Sphere. We construct a Lagrange interpolant which is minimal for a certain lexicographic order, and we use it to design an accurate quadrature rule. We also study non-regularized least squares problems, in particular to define a stable discrete Funk-Radon transform. In parallel, various results on invariance by the symmetry group of the cube are shown and exploited, while the great circles associated to the grid are used to study Vandermonde matrices.

Contents

Introduction

This habilitation thesis is a synthesis of works realized since I was hired as an associate professor (Maître de Conférences) in 2011. Two independent parts are presented. The manuscript is organized as follows.

Part I: Mathematical and numerical aspects of three-dimensional optical imaging based on the Radon transform

The first part deals with mathematical and numerical aspects of three-dimensional (3D) optical imaging based on Radon-kind transforms. This subject has been motivated by collaborations with companies which design innovative imaging systems in visible and infrared optics. My main collaborator on this topic is G. Berginc from the company Thales Optronique S.A.

In Chapter 1, we introduce 3D reflective tomography, which includes an acquisition in visible to near-infrared optics (VIS-NIR), the inversion of a Radon-kind transform from X-ray tomography, and 3D visualization. We present the specific context of this work, and we exhibit various numerical results, including 3D reconstructions from real images. This chapter refers to the project report [21], the proceedings [16, 17], the working document [20], the article [8], and the patent [18].

Chapter 2 deals with a mathematical analysis of reflective tomography. The chosen presentation follows the chronological order of the results. Firstly, some intuitive understanding is presented; a parallel with edge detection by the Hough transform is drawn. Secondly, asymptotic models are sketched to describe the reconstructed geometry and the artifacts. Thirdly, some results extracted from microlocal analysis justifies the method in term of singularities. Lastly, we exhibit a framework where the Radon transform extended to distributions models pure diffuse reflection. This chapter refers to the letter [12], the note [15], the working document [19], and the articles [3, 8].

In Chapter 3, we propose two unconventional algorithms for problems inspired by reflective tomography: a multiresolution greedy algorithm which increases computational efficiency, and an algebraic solver for multi-view reconstruction in a general setting. This chapter is based on the articles [1, 2], and contains new reconstructions obtained by a home-made scanner.

In Chapter 4, we conclude this first part with some perspectives.

Appendix A summarizes standard mathematical results about the Radon transform; they include an inversion formula, reconstruction algorithms, and results from microlocal analysis. These results provide the main mathematical background for tomography.

Appendix B deals with some physical background about electromagnetic imagery, image formation in VIS-NIR optics, and X-ray tomography.

Part II: Mathematical and numerical aspects of spectral computing on the Cubed Sphere

The second part of the thesis deals with the Cubed Sphere, which is a spherical grid widely used for numerical computation on the sphere, in climatology and meteorology. More specifically, my subjects of interest concern geometrical and metric properties of the equiangular Cubed Sphere, and

computing with spherical harmonics on this grid, including the study of Vandermonde matrices. I have been working on the subject since 2020, with J.-P. Croisille (UL) and M. Brachet from the Université de Poitiers.

Chapter 5 deals with mathematical properties of the Cubed Sphere. The shortest geodesic arcs, whose length is the separation distance of the grid, are given: they especially match with the vertices of a cuboctahedron. As a consequence of this metric property, the symmetry group of the Cubed Sphere coincides with the octahedral group. This chapter summarizes the main results of the article [4].

Chapter 6 tackles Lagrange interpolation on the Cubed Sphere by a spherical harmonic. The approach especially factorizes a suitable Vandermonde matrix under an echelon form, in order to eliminate undersampled spherical harmonics. This chapter is a reworking of the article [10] and contains new results; we have improved the theoretical lower bound on the degree which guarantees the existence of an interpolating function, and we now describe the interpolating function as the solution of a lexicographical optimization problem.

In Chapter 7, we design and we study a new octahedral quadrature rule on the Cubed Sphere, taking benefit from Lagrange interpolation. Contrary to Gaussian quadrature, where the set of nodes and weights is solution of a nonlinear system, only the weights are unknown here. Despite this conceptual simplicity, the new quadrature displays an accuracy comparable to optimal quadratures, such as the Lebedev rules. This chapter is extracted from the article [9].

In Chapter 8, we study least squares fitting by a spherical harmonic on the Cubed Sphere. The most important observation is that selecting a degree compatible with the Shannon-Nyquist's frequency along the equatorial great circle provides an approximation problem that is well-conditioned, whereas violating this condition implies that the condition number explodes when the number of nodes tends to infinity. Another point concerns the block diagonal structure of the normal matrix, based on octahedral symmetry consideration; this result permits to improve the computational efficiency. The chapter is extracted from the article [11].

In Chapter 9, we define some discrete Funk-Radon transform in a spectral framework based on least squares fitting on the Cubed Sphere, without regularizing. We exhibit the pseudoinverse and, as above, we argue that the transform is expected to be stable as soon as the Shannon-Nyquist condition is fulfilled along the equator. Various numerical experiments attest to the accuracy and the convergence of the approach, in particular for toy models from diffuse Magnetic Resonance Imaging. This chapter is extracted from the article [5].

Lastly, Chapter 10 concludes this work by a series of perspectives.

Publications

Peer-reviewed international journals

[1] J.-B. Bellet, *Multiresolution greedy algorithm dedicated to reflective tomography*, SMAI Journal of Computational Mathematics, 4 (2018), pp. 259–296.

[2] ——, *Flexible algebraic technique for multiview reconstruction: incremental learning in reflective tomography*, Optical Engineering, 58 (2019), pp. 1–22.

[3] J.-B. Bellet, *An exact Radon formula for Lambertian tomography*, Journal of Mathematical Imaging and Vision, 64 (2022), pp. 939–947.

[4] J.-B. Bellet, *Symmetry group of the equiangular cubed sphere*, Quarterly of Applied Mathematics, 80 (2022), pp. 69–86.

[5] J.-B. Bellet, *A discrete Funk transform on the Cubed Sphere*, Journal of Computational and Applied Mathematics, 429 (2023).

[6] J.-B. Bellet and G. Berginc, *Modèle effectif de couche mince rugueuse périodique sur une structure semi-infinie*, ESAIM: M2AN, 47 (2013), pp. 1367–1386.

[7] ——, *Imaging from monostatic scattered intensities*, Mathematical Methods in the Applied Sciences, 37 (2014), pp. 1772–1783.

[8] ——, *Heuristic imaging from generic projections: backprojection outside the range of the Radon transform*, SMAI MathematicS In Action, 9 (2020), pp. 1–16.

[9] J.-B. Bellet, M. Brachet, and J.-P. Croisille, *Quadrature and symmetry on the Cubed Sphere*, Journal of Computational and Applied Mathematics, 409 (2022).

[10] ——, *Interpolation on the Cubed Sphere with Spherical Harmonics*, Numerische Mathematik, 153 (2023), pp. 249—278.

[11] J.-B. Bellet and J.-P. Croisille, *Least Squares Spherical Harmonics Approximation on the Cubed Sphere*, Journal of Computational and Applied Mathematics, 429 (2023).

[12] G. Rigaud, J.-B. Bellet, G. Berginc, I. Berechet, and S. Berechet, *Reflective Imaging Solved by the Radon Transform*, IEEE Geoscience and Remote Sensing Letters, 13 (2016), pp. 936–938.

Comptes rendus

[13] J.-B. Bellet and G. Berginc, *Imagerie laser*, Comptes Rendus Mathematique, 349 (2011), pp. 315–317.

[14] ——, *Modèle électromagnétique d'objet dissimulé*, Comptes Rendus Mathematique, 349 (2011), pp. 153–155.

[15] ——, *Reflective filtered backprojection*, Comptes Rendus Mathematique, 354 (2016), pp. 960–964.

Conference proceedings

[16] J.-B. Bellet, I. Berechet, S. Berechet, G. Berginc, and G. Rigaud, *Laser Interactive 3D Computer Graphics*, in Proceedings of the 2nd International Conference on Tomography of Materials and Structures, Québec, B. Long, ed., 2015, pp. 109–113.

[17] G. Berginc, J.-B. Bellet, I. Berechet, and S. Berechet, *Optical 3D imaging and visualization of concealed objects*, in Proc. SPIE 9961, Reflection, Scattering, and Diffraction from Surfaces V, L. M. Hanssen, ed., 2016.

International patent

[18] S. Berechet, I. Berechet, J.-B. Bellet, and G. Berginc, *Method for discrimination and identification of objects of a scene by 3-D imaging.* Priority: FR1402929A, 2014–12–19; EP2015080258W, 2015–12–17. Publications: FR3030847B, 2017. EP3234914B1, 2018. US10339698B2, 2019. HK1245975B, 2019. JP6753853B2, 2020. IL252791B, 2021. KR102265248B1, 2021.

Others

[19] J.-B. Bellet, *Analyse asymptotique et géométrique de la tomographie réflective.* Document de travail `https://hal.archives-ouvertes.fr/hal-01571707`, 2017.

[20] J.-B. Bellet and G. Berginc, *Quality control of 3D reflective tomography.* Working document `https://hal.archives-ouvertes.fr/hal-01368354v2`, 2017.

[21] J.-B. Bellet, G. Berginc, and J.-P. Croisille, *Algorithmes de reconstruction et imagerie laser tomographique.* Rapport de fin de projet AMIES, collection HAL Maths-Entreprises, `https://hal.archives-ouvertes.fr/hal-00830979`, 2013.

Part I

Mathematical and numerical aspects of three-dimensional optical imaging based on the Radon transform

Chapter 1

Three-dimensional reflective tomography

1.1 Introduction

This chapter introduces three-dimensional (3D) reflective tomography. We present various aspects of the subject, including the context of our study, the principle of the method, implementation aspects, and a wide variety of numerical experiments on synthetic and real data.

1.2 Context

The subject has been motivated by collaborations with companies which work on novel imaging modalities, in the context described hereafter.

1.2.1 Technological context: three-dimensional active laser imaging

Thales Optronique S.A. (TOSA) is a company specialized in the design and development of innovative optronic systems, in particular in visible and infrared optics. TOSA has patented in 2009 a technology concerning the 3D reconstruction of a scene, based on active laser imagery, [32–35]. Basically, the principle is the following. For the acquisition, a laser source illuminates the scene to be imaged, for several angular position with respect to the scene. At the same time, some camera co-located with the source records backscattered intensities. This provides one bi-dimensional (2D) image of the scene per angular position. In order to clarify expectations, the orders of magnitude for the characteristic lengths are:

- VIS-NIR wavelengths: 0.4-3 µm;
- diameter of the imaged scene: 10 m;
- distance between the device and the scene: 5 km.

Next, the collected images are injected into processing algorithms, in order to compute a 3D reconstruction of the scene. The reconstruction step is based on the Feldkamp-Davis-Kress algorithm [48], from X-ray tomography. Lastly, the reconstructed 3D volume is explored in order to model objects of interest from the scene, for instance under the form of surfaces. In particular, TOSA has worked on surface representations [28,29,31], with the Small and Medium Enterprise SISPIA, specialized in algorithms.

1.2.2 Contractual context

The following contractual context is the initial motivation for my studies on 3D optical imaging.

A collaboration between TOSA and the Université de Lorraine (UL) started in 2012. This collaboration was formalized by means of a one-year project, entitled *Algorithmes de reconstruction*

et imagerie laser tomographique, supported by the Agence pour les Mathématiques en Interaction avec les Entreprises et la Société (AMIES)[1]. I leaded the project; the other participants were G. Berginc from TOSA and J.-P. Croisille from UL. The project served as a starting point for my works about 3D active laser imaging based on algorithms from tomography.

For UL, a second collaboration concerning 3D laser imaging started in 2013; it was formalized by a consortium agreement between SISPIA, TOSA, and UL. The associated two-year project, entitled *DatDriv3D+*, was supported by the Direction Générale de la Compétitivité de l'Industrie et des Services (Ministère du Redressement productif) and the Direction Générale de l'Armement (Ministère de la Défense). The leader of the project was I. Berechet from SISPIA, G. Berginc was the leader for TOSA's part, I was the leader for the UL's part; the other participants were S. Berechet from SISPIA, and G. Rigaud as a postdoctoral researcher UL.

1.3 Principle of reflective tomography

Three-dimensional (3D) reflective tomography deals with the reconstruction of a 3D scene, by a combination of visible to near-infrared (VIS-NIR) optics and algorithms from X-ray tomography.

1.3.1 Acquisition: cone beam scan in VIS-NIR optics

The acquisition consists in measuring a set of bi-dimensional (2D) optical images in the VIS-NIR band. It especially collects images of "reflecting" surfaces under several angles of view; we refer to Appendix B.2 for a modeling of VIS-NIR images, including the integral equation (B.7) for reflecting surfaces. Typically, a motionless scene is observed by a camera with a continuous motion, such as the onboard camera in Figure 1.1. The wavelength range is considered as a parameter of the acquisition device, which can be passive (without light source), or active (with its own light source).

Figure 1.1: Onboard camera.

From a geometrical point of view, the acquisition is a cone beam scan, similar to a tomographic X-ray scan described in Appendix B.3. Indeed, an ideal camera realizes a perspective projection (Figure B.4); hence, each 2D optical image contains projections along a cone beam of rays, analogous to the cone beam of a radiography (Figure B.10).

We assume that this acquisition geometry is known. In other words, any recorded image is assumed to be calibrated, which means that the intrinsic and extrinsic matrices of the camera are known in (B.1). In practice, this may require additional measurements or pre-processing steps. For instance, in [32], some correcting algorithms calibrate the images in order to approach some known geometry (circular cone beam scan). Also, we refer to textbooks in computer vision for calibration procedures [57, 73]; this is outside the scope of the presented thesis.

1.3.2 Solver: cone beam computed tomography

The so-called *reflective tomography* is an ingenious principle based on the geometrical similarity of a tomographic scan and the acquisition described above. It consists in injecting 2D reflective images into a solver of cone beam computed tomography, in order to compute some 3D reconstruction of the initial scene. In this way, reflective tomography is a qualitative inversion procedure which captures the 3D geometry of reflecting surfaces. From a physical point of view, a reconstruction from radiant incidance on pixels [$\mathtt{W.m^{-2}}$] represents a power per unit volume [$\mathtt{W.m^{-3}}$]. Furthermore, the method is not based on a model of the brightness (incidance) of the recorded pixels; so, pre-processing such as rescaling the brightness is tolerated.

[1] `https://www.agence-maths-entreprises.fr`

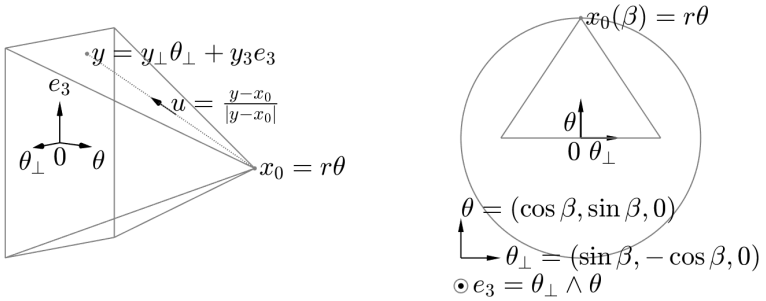


Figure 1.2: Circular cone beam scan. Left: perspective ray (x_0, u) through a fixed point $x_0 = x_0(\beta)$, with direction $u = u(\beta, y_\perp, y_3)$; the parametrization of the ray considers that the image is formed on a point $y = y_\perp \theta_\perp + y_3 e_3$ (in a virtual screen). Right: by rotation around the axis e_3, the point $x_0(\beta)$ scans a horizontal circle. The total set of rays is $\mathcal{L}_{\mathrm{CB}}$ defined in (1.1). In cone beam tomography, $x_0(\beta)$ is the position of an X-ray source; in reflective tomography, $x_0(\beta)$ is the position of the optical center of a camera.

The assumed acquisition geometry impacts the choice of the solver. In this chapter, we focus on the Feldkamp-Davis-Kress (FDK) algorithm [48], dedicated to a circular cone beam scan (Figure 1.2). In 3D reflective tomography [32–35,64], the FDK algorithm computes efficiently a 3D reconstruction, sampled on a 3D grid of voxels, from 2D VIS-NIR images.

1.3.3 Reconstruction algorithm for a circular scan: FDK algorithm

The FDK algorithm [48] is one of the most widely used methods in 3D computed tomography. This method has been designed to "invert" efficiently the X-ray transform $\mathcal{X}$ defined by (B.11), in the case of a circular cone beam scan. The FDK algorithm is a heuristic extension of the 2D Radon inversion (A.9) and the FBP formula (A.14). It provides some *filtered backprojection* operator $\mathcal{B}\Phi$ such that

$$\mathcal{B}\Phi\mathcal{X}[f]|_{\mathcal{L}_{\mathrm{CB}}(r,a,b)} \approx f, \quad f : \mathbb{R}^3 \to \mathbb{R}.$$

Here, $\mathcal{L}_{\mathrm{CB}}(r, a, b)$ denotes the set of rays of a circular cone beam scan, parametrized by

$$\mathcal{L}_{\mathrm{CB}}(r, a, b) := \{(x_0(\beta), u(\beta, y_\perp, y_3)),\ \beta \in [0, 2\pi],\ (y_\perp, y_3) \in [-a, a] \times [-b, b]\}; \tag{1.1}$$

the position $x_0(\beta)$ and the direction $u(\beta, y_\perp, y_3)$ of a ray are specified in Figure 1.2. The weighted filtering Φ and the backprojection operator $\mathcal{B}$ are defined in the descriptive of the algorithm on the following page. We refer to [48, 78] for a comprehensive derivation of these operators. The practical implementation of the FDK algorithm is analogous to the FBP algorithm on page 67.

Among the properties of the algorithm, for $r \to \infty$, any horizontal cross-section $(\mathcal{B}\Phi\mathcal{X}[f])(\cdot, \cdot, z)$ looks like a 2D filtered backprojection (A.14) from the Radon transform $\mathcal{R}[f(\cdot, \cdot, z)]$, in the plane $x_3 = z$. In particular, the FDK algorithm is relevant in X-ray tomography, at least if r is large enough.

1.4 Positioning of the method

1.4.1 Laser reconstruction

Reflective tomography uses algorithms from transmission tomography for reflective data, despite the VIS-NIR wavelengths are much larger than the X-ray ones. To the author's knowledge, this principle emerged at the end of the 1980s for laser radars [67, 68, 82], and more particularly for *laser range profiling*. The original reflective tomography reconstructs a 2D image from one-dimensional (1D) range profiles of a rotating target; the solver is a 2D filtered backprojection. Laser range profiling has been further studied since the 1990s. Various solvers have been tested [74]. The applications

FDK algorithm.

Input. X-ray transform $F : \mathcal{L}_{\mathrm{CB}}(r, a, b) \to \mathbb{R}$, measured by the cone beam scan of Figure 1.2.

Step 1.a: weighting. Compute the weighted data set F_w:

$$F_w(\beta, y_\perp, y_3) = w(y_\perp, y_3) F(x_0(\beta), u(\beta, y_\perp, y_3)), \quad w(y_\perp, y_3) = \tfrac{r}{(r^2 + {y_\perp}^2 + {y_3}^2)^{0.5}}.$$

Step 1.b: filtering. Compute the horizontal filtering ΦF:

$$\Phi F(\beta, y_\perp, y_3) := \mathcal{F}_1^{-1}\{|\sigma| \hat{h}(\sigma) \mathcal{F}_1[F_w(\beta, \cdot, y_3)](\sigma)\}, \quad y_3 \in [-b, b], \beta \in [0, 2\pi],$$

where $\mathcal{F}_1(g)(\sigma) = \int g(y_\perp) \mathrm{e}^{-\mathrm{i}\sigma y_\perp} \mathrm{d}y_\perp$ is the Fourier transform, and $\hat{h}$ is an even windowing function with compact support.

Step 2: backprojection. Compute the backprojection on a grid of voxels: for each voxel location x, compute $\mathcal{B}\Phi F(x)$ where $\mathcal{B}$ is a weighted summation over lines through x,

$$\mathcal{B}G(x) := \int_0^{2\pi} \frac{r^2}{(r - x \cdot \theta)^2} G(\beta, y_\perp, y_3) \mathrm{d}\beta,$$

with $y_\perp = \frac{rx \cdot \theta_\perp}{r - x \cdot \theta}$, $y_3 = \frac{rx_3}{r - x \cdot \theta}$, $\theta = (\cos\beta, \sin\beta, 0)$.

Output. FDK reconstruction $\mathcal{B}\Phi F$, evaluated on a grid of voxels.

include range-resolved imaging of satellites [71, 75, 76]. See also [30, 40, 59, 102] for works realized in the past decade.

In 3D active laser imaging [32–35], the reconstruction principle shares similarities with range profiling, but the input data are 2D optical images instead of 1D range profiles, and the output is a 3D reconstruction, computed with a 3D solver from tomography. This is very similar with the method of [52], where 3D models of an object are computed from photographs in the visible band; the method relies on one 2D filtered backprojection per horizontal cross-section, assuming orthographic rays for a large focal length.

1.4.2 Geometric tomography

Essentially, an optical image contains a perspective projection of a scene. Reconstructing the geometry of the scene from such data enters into the framework of *geometric tomography*:

> "*Geometric tomography* is the area of mathematics dealing with the retrieval of information about a geometric object from data about its sections, or projections, or both", [51].

In 3D reflective tomography, the data are related to the geometry of the scene, but also to physical parameters such as the BRDF (described in (B.6) and Figure B.7). Using algorithms from X-ray tomography such as the FDK algorithm appears to be an efficient way of combining these data for recovering the geometry.

1.4.3 Multi-view stereo

3D reflective tomography enters also in the framework of *multi-view stereo*:

> "The goal of multi-view stereo is to reconstruct a complete 3D object model from a collection of images taken from known camera viewpoints", [95].

There exists a huge variety of methods for multi-view stereo; see for instance [57, 73, 98] and the references therein. We can clarify the position of the method discussed so-far, following the six-point taxonomy of [95]:

1. Scene representation: the reconstruction is computed on a grid of 3D voxels.
2. Photoconsistency measure: the method does not need the comparison of pixel values in different images.
3. Visibility model: the method does not need to predict visibility and occlusions.
4. Shape priors: the method does not need shape priors.
5. Reconstruction algorithm: the FDK algorithm from cone beam computed tomography is the solver.
6. Initialization requirements: the FDK algorithm is direct and does not need an initialization.

This list emphasizes that 3D reflective tomography is robust.

Furthermore, reflective tomography is related to the *shape from silhouette*, where a *visual hull* of the scene [26, 72] is obtained by some backprojection of binarized images containing the silhouettes. In comparison, the FDK algorithm is a backprojection, but it is combined with a filtering, and the images are not required to be binary. The shape from silhouette is a common initialization in multi-view stereo; one may imagine 3D reflective tomography as an alternative method.

1.5 Visualization

We have discussed so far 3D reflective tomography as the computation of a 3D grid of voxels, based on algorithms from transmission tomography. In this section, we further analyze such a 3D volume in order to represent, visualize, or extract some objects of interest. In Figure 1.3, we display a reconstruction using three usual techniques [101]:

(a) Slicing: a 2D cross-section is extracted and directly displayed. This method is exact and simple, but it is difficult to appreciate 3D structures.

(b) Surface rendering: a surface is extracted by the means of thresholding, and a 2D synthesis image, eventually based on radiometric concepts, is displayed. The computation is efficient and 3D solids are nicely represented, but thresholding may be tricky and lacunarities may appear.

(c) Volume rendering: the 2D displayed image is a "projection" of the whole 3D volume. The projection is more or less sophisticated and eventually based on models of light propagation such as radiative transfer. A single view somehow captures simultaneously any 3D structure.

(a) Slice. (b) Surface rendered. (c) Volume rendered (MIP).

Figure 1.3: Three visualizations of a 3D reflective tomography reconstruction.

In reflective tomography, the surfaces of the original scene are observed to be located near the peaks of the reconstruction. In order to extract the objects of interest, we need especially to extract the brightest voxels. Using the coordinates of such voxels in a reference frame, we obtain a point cloud of the scene. Using the extracted voxels, or the point cloud, surfaces are deduced, by processes such as interpolation (or data approximation).

The rest of this section is devoted to volume rendering and the extraction of point clouds. We relate these two fields as in the international patent [18].

1.5.1 Volume rendering

Concerning volume rendering, it appears that we want especially to visualize surfacic points, represented by bright voxels. This motivates the use of the *Maximum Intensity Projection* (MIP), as it is performed in [52], [16, 17], [18]. Indeed, the MIP projects a volumetric "intensity" onto a screen of pixels along straight rays; each pixel records the maximum intensity along the corresponding ray, as in Figure 1.4. Hence, in 3D reflective tomography, the MIP efficiently computes a contrasted image of the reconstructed surfaces (up to artifacts), as in Figure 1.3(c). The simplest form of MIP is free of parameters, and improvements are available, such as removal of unexpected voxels by thresholding, or perception of distance improved by the means of an attenuation coefficient [101]. In practice, the rendered image is often adjusted by the means of thresholding, rescaling, or color mapping.

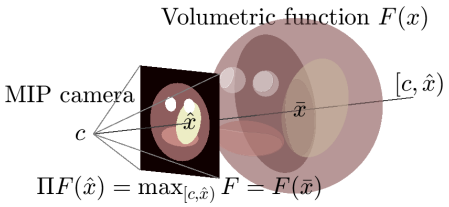


Figure 1.4: Image formation through a MIP camera.

More formally, after some eventual processing such as restriction to a sub-volume of interest, thresholding, rescaling, sign reversal, or whatever, the 3D reconstruction defines a compactly supported function $F : \mathbb{R}^3 \to \mathbb{R}$. As in Figure 1.4, we display such a volumetric reconstruction F using a *perspective MIP camera*, defined for a geometry of projection as in Figure B.4. More precisely, the MIP of F, at a pixel located at $\hat{x}$, is defined by

$$\Pi F(\hat{x}) := \max_{x \in [c,\hat{x})} F(x) = \max_{\lambda \geq 0} F(c + \lambda u_{c,\hat{x}}). \tag{1.2}$$

Here, c represents an optical center, $[c, \hat{x})$ is the ray through the pixel $\hat{x}$ (half-line), $u_{c,\hat{x}} = \frac{\hat{x}-c}{|\hat{x}-c|}$ is the direction of the ray, $\lambda \geq 0$ is the depth of a point $x = c + \lambda u_{c,\hat{x}}$ of the ray. Such a MIP camera has an intrinsic matrix, and an extrinsic matrix, similarly as (B.1); changing the focal length in the intrinsic matrix enables zooming in/out, while changing the position and the orientation in the extrinsic matrix provides several points of view. In practice, automatized scenario of visualization can be used, such as displaying a (MIP) cone beam scan of the whole reconstruction, with a predefined threshold. It is also possible to proceed interactively.

Remark 1.1. Other volume rendering methods can be obtained analogously: fix $\Pi F(\hat{x}) := \|F|_{[c,\hat{x})}\|_p$ with $p \in [1, \infty]$ (for $p = 1$, this is an X-ray transform, for $p = \infty$, this is a MIP).

1.5.2 Point clouds

To go further with the extraction of objects from the reconstruction, representations based on point clouds are also of particular interest. The most intuitive way of extraction is based on thresholding; a point cloud is obtained by the coordinates of voxels with intensity between thresholds. In order to densify the point cloud, this procedure is iterated, on several sub-volumes and with several thresholds. More inventively, an efficient selection of voxels can be operated by the MIP [18]. As explained previously, the MIP displays an intensity map corresponding essentially to surfacic voxels. We select some pixels of the displayed image by thresholding; then, the associated voxels/points are extracted. See Figure 1.5 for an example. Formally, for a MIP image (1.2), this process defines some points under the form

$$\bar{x} = \arg \max_{x \in [c,\hat{x})} F(x) \in \mathbb{R}^3; \tag{1.3}$$

the corresponding pixel of the MIP image (1.2) has the position $\hat{x}$ and the intensity $\Pi F(\hat{x}) = F(\bar{x})$. Here again, the procedure is iterated, for various sub-volumes, and various thresholds. This method, which uses the MIP as a compression method to extract a point cloud, is performed interactively, or some automatic scanning scenario is pre-defined.

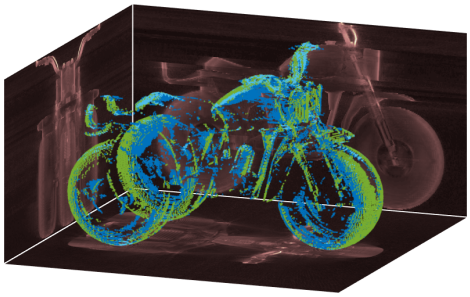


Figure 1.5: Superposition of three point clouds extracted from three (orthographic) MIPs.

1.6 Implementation

As far as X-ray inversion and 3D volume rendering are concerned, I have developed various codes in various languages. Briefly, some codes are essentially dedicated to tests in `Matlab`. Another code developed in `CUDA C` demonstrates the efficiency of the combination FDK-MIP in reflective tomography.

1.6.1 Matlab codes for testing

Numerical tests and numerical illustrations are often performed in `Matlab`, eventually interfaced with compiled codes to reduce the computational time.

In particular, I have developed `Matlab` codes for FBP algorithms in various acquisition geometries [21], following [78, Chap. 5]. These codes include the inversion of the Radon transform by the FBP algorithm for a parallel scan in 2D, and the FDK algorithm for a circular cone beam scan in 3D. Various strategies have been used to reduce the computational time. The first one consists in avoiding loops using vectorization. The second one deals with distributing the computation on several cores (with the function `parcellfun` in `Octave`). The third one tackles the bottleneck of the FDK algorithm: the most time-consuming task is the backprojection step (typically in $\mathcal{O}(N^4)$ operations), so I have developed a `Fortran` code dedicated to this task. It significantly reduces the computational time.

Concerning 3D visualization of a 3D grid of voxels, I have developed a ray tracer in `Matlab` for 3D volume rendering such as the MIP, the X-ray transform, or the attenuated X-ray transform. Also, the `Matlab` code of the MIP has been translated into a `C` code and a `CUDA C` code; here again, calling one of these compiled codes can accelerate the computation (which typically requires $\mathcal{O}(N^4)$ operations for video rendering).

1.6.2 Interactive software in CUDA C

In 2014, I developed an interactive software in `CUDA C`, using [92] as a reference for CUDA programming. The code especially combines the FDK algorithm and a MIP rendering on a Graphics Processing Unit (GPU).

The FDK reconstruction is a grid of voxels which is computed and stored on the GPU. The weighted filtering is based on the Fast Fourier Transform of the `cuFFT` library. For the backprojection, the computations are massively parallelized; the voxels are computed independently. Note that programming efficiently the FDK algorithm on a GPU is a subject of concern by itself; see for example [36, 81, 89, 91, 93].

Concerning the display of the FDK reconstruction, any MIP view is computed directly on the GPU and is displayed on the computer screen using `OpenGL`. Furthermore, the `GLUT` library is used to manage interactions with mouse/keyboard. This provides a rendering software that enables to move as a virtual observer inside the reconstructed volume. The displacements of the camera are managed by the mouse, whereas some other parameters such as thresholding or the limits of the

region of interest are managed by the keyboard. Last but not least, the display is updated in "real time"; the rendering software is interactive.

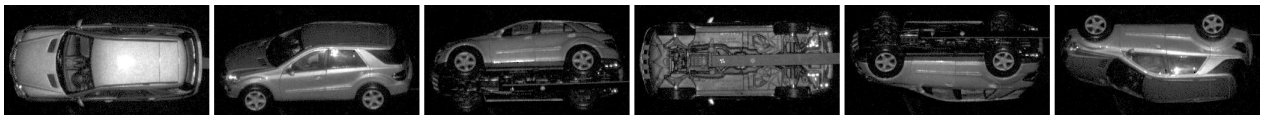
The resulting home-made software has demonstrated the efficiency of FDK-MIP on various test cases from TOSA, during the *DatDriv3D+* project. In 2015, some video clips, produced by means of this software, have been shown during a research exhibition at Thales (Thales Research Days).

1.7 Reconstruction from real data

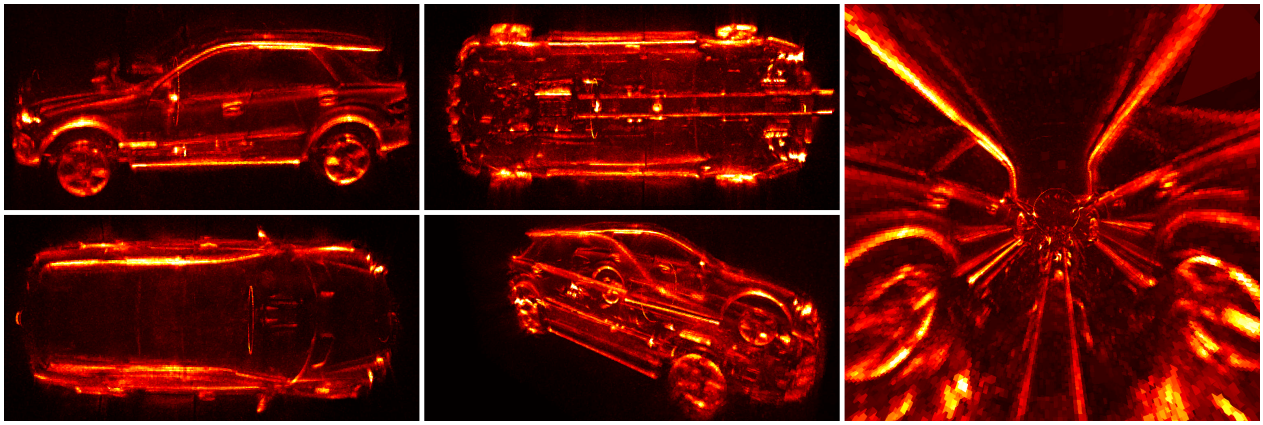
We demonstrate the strength of reflective tomography on two test cases with real data. The 2D images have been measured by TOSA, using active laser imagery in the VIS-NIR band. The acquisition is assimilated to a circular cone beam scan such as Figure 1.2. The tomography solver is the FDK algorithm, and the display is a perspective MIP as in Figure 1.4. The computation and the display are performed with the home-made software described in Subsection 1.6.2.

1.7.1 Circular cone beam scan

For the first test case, displayed in Figure 1.6 and published in [8], we consider a sequence of 360 images of size 181×342; see (a) for some samples. The acquisition geometry is similar with Figure 1.2; the angle β scans a uniform grid, with a one degree step. A FDK tomographic reconstruction $181 \times 181 \times 342$ is computed in 2.6 seconds on a GPU Nvidia Tesla C2075. As can be observed in the snapshots (b) of the interactive display, the reconstruction contains surfaces of the original scene with many features and details, useful for identification purposes.



(a) Input dataset (courtesy of TOSA): 360 VIS-NIR active images of size 181×342. Here, six samples of the sequence are displayed.



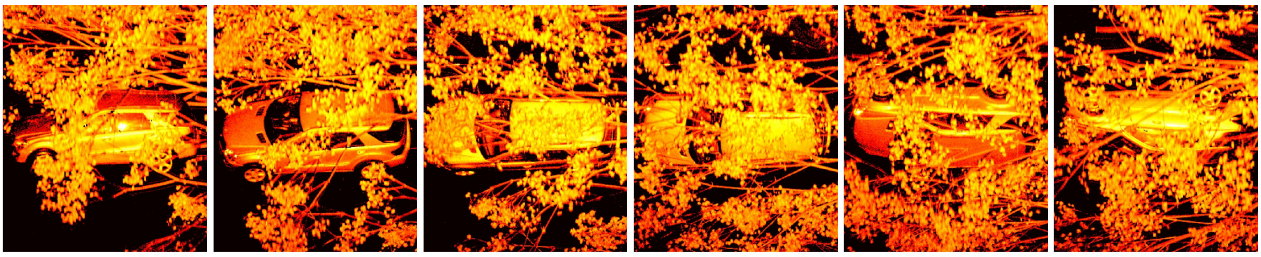
(b) Volume rendering of a 3D tomographic reconstruction (home-made software). The reconstruction is a grid of $181 \times 181 \times 342$ voxels; it is computed by the FDK algorithm in 2.6 seconds on a GPU Nvidia Tesla C2075. The display is a MIP computed interactively.

Figure 1.6: First test case: 3D reflective tomography from a circular cone beam scan in VIS-NIR optics.

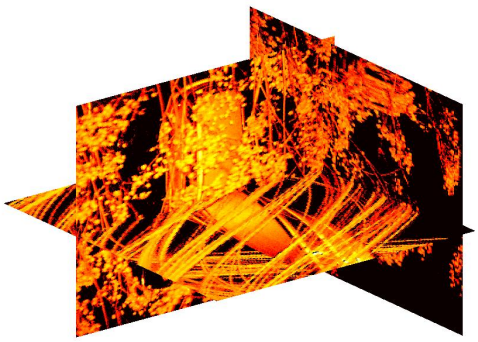
1.7.2 Limited view with occlusions

The second test case, displayed in Figures 1.7-1.8 and published in [16, 17], is more challenging. The scene is a car partially occluded by branches and foliage, and the cone beam scan is performed

only over a half-circle. Some images of the sequence and slices through the dataset are displayed in Figure 1.7. The dataset contains 181 images of size 421×342; in degrees, the angle β ranges from 0 to 180, with a one degree step. Two kinds of incompleteness are noticed: incompleteness due to occlusions, and angular incompleteness due to a restriction of the angular range.



(a) Samples of the input sequence (color mapped).



(b) Slices through the input dataset $181 \times 421 \times 342$.

Figure 1.7: Second test case: 3D reflective tomography from limited view with occlusions in VIS-NIR optics. The input dataset (courtesy of TOSA) contains 181 VIS-NIR active images of size 421×342, associated to a half-circular cone beam scan.

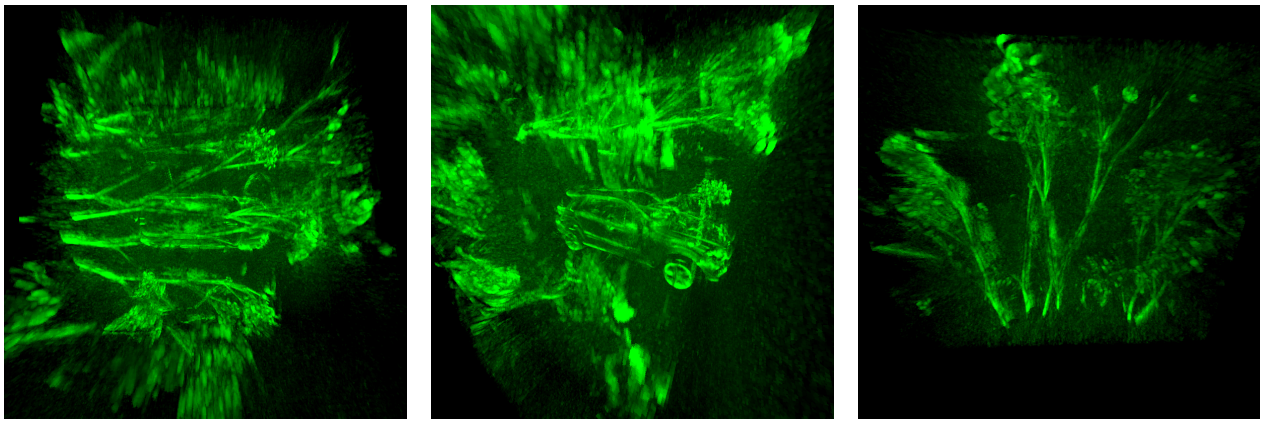
Despite incompleteness, a tomographic reconstruction can still be computed by the FDK algorithm; here, the backprojection integrates only over the known angular range. The home-made software, executed on a Nvidia Tesla C2075, computes the weighted filtering by `cuFFT` in 0.4 second, and the backprojection on a grid of $421 \times 421 \times 342$ voxels, in 3.6 seconds. The reconstruction is displayed in Figure 1.8. Two different MIP views of the whole scene are represented in (a). We have also interactively separated some regions of interest: branches are isolated at the right of (a), whereas the car is displayed alone in (b), after removal of branches and foliage. This test case shows that 3D reflective tomography can automatically overcome issues due to occultation.

1.8 Numerical experiments

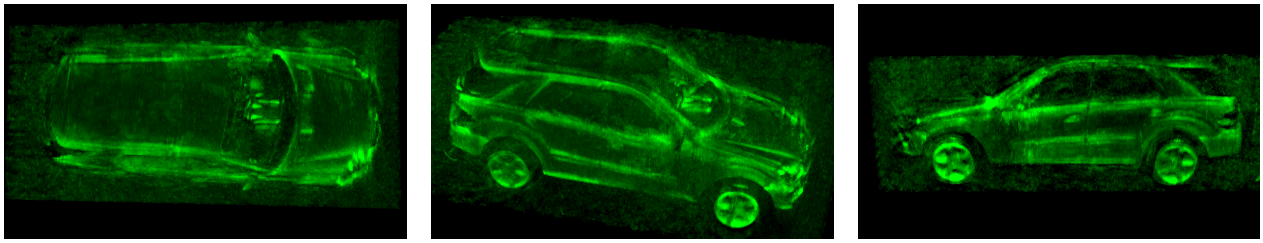
We test the principle of 3D reflective tomography on several classes of images, from a numerical point of view; we consider a Gouraud model, cartoon images with discontinuities, images of a randomized pattern, and noisy images. Overall, these tests show that the method computes the initial geometry, displays it under the form of contrasted images, in a robust and stable way. These results are extracted from [20] and we refer to [17] for similar results.

1.8.1 Reconstruction from a Gouraud model of the Stanford Bunny

We test 3D reflective tomography on the Stanford Bunny [100], enlightened by a Gouraud model [53].



(a) Whole 3D reconstruction, and interactively extracted branches.



(b) Interactively extracted car.

Figure 1.8: Second test case: 3D reflective tomography from limited view with occlusions in VIS-NIR optics. See Figure 1.7 for the input dataset (courtesy of TOSA). The reconstruction is a grid of $421 \times 421 \times 342$ voxels, computed by the FDK algorithm in 4.0 seconds on a GPU Nvidia Tesla C2075 (home-made software). Here, the interactive MIP is used to display and to extract objects of interest from the reconstruction (home-made software).

At a first step, we read the full resolution Stanford Bunny (69451 faces) with the `read_ply` function from [83]. We "color" the faces of the object with the smooth pattern $x \mapsto 1+0.5\sin(20\pi|x|)$ (in some system of coordinates), computed at the vertices and extended to the faces by interpolation. We generate 1605 images of size 397×312 with the Gouraud model of `Matlab`; a black background is considered. A circular scan is performed: vertical images are obtained by a rotation of the Bunny over 360 degrees, with a constant angular step. We display 6 images of this sequence in Figure 1.9.

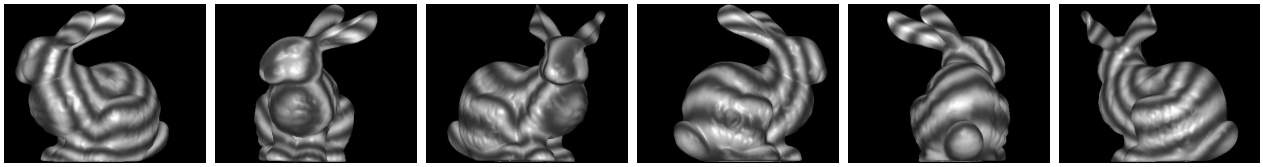


Figure 1.9: Gouraud images of a patterned Stanford Bunny. Here, 6 samples (step of 60 degrees) of a circular cone beam scan comprising 1605 images of size 397×312.

Next, we compute a 3D tomographic reconstruction F. In Figure 1.10, we display a MIP rendering of the reconstruction from several angles of view. The first line of this figure is some *re-projection* associated to Figure 1.9 since the rays of projection are similar. The second line contains novel views of the scene. Contrasted representations of the scene are obtained.

We evaluate now the reconstruction-visualization procedure. In Figure 1.11, we examine the quality of an aerial MIP view ΠF, displayed in (a). We focus here on an horizontal aerial view, since forming such a view from vertical images is the core of the prediction problem. We discriminate the pixels corresponding to surfacic points of the initial object, as follows. For any pixel $\hat{x}$ of the

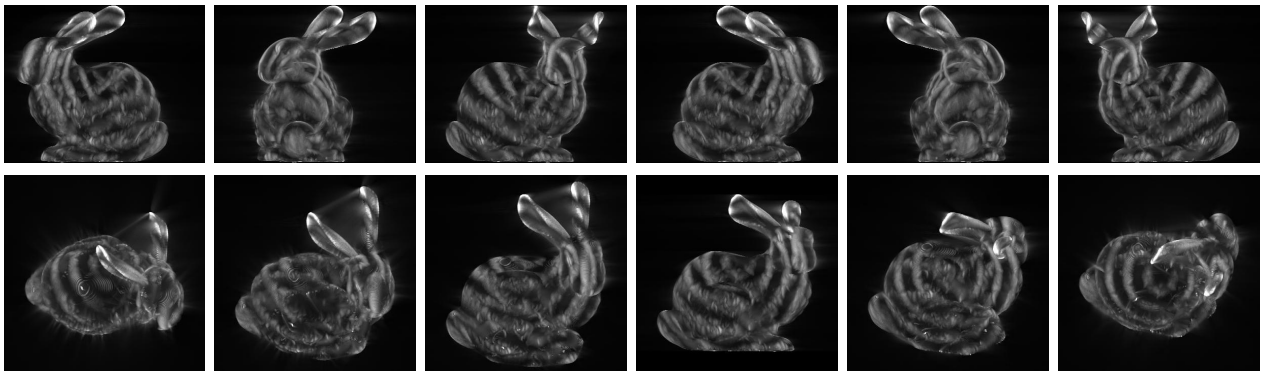
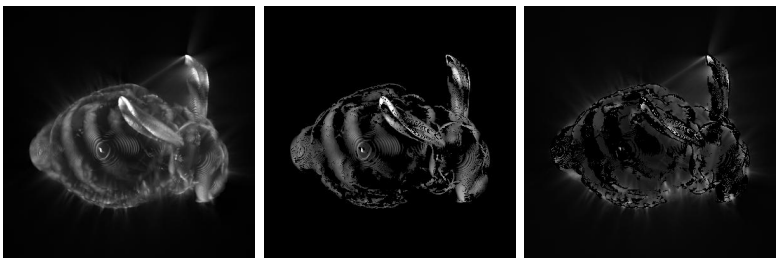


Figure 1.10: 3D tomographic reconstruction from a Gouraud model of a patterned Stanford Bunny. Two sequences of MIP views are displayed. Top: a rotation around the vertical axis (60 degrees step) generates *re-projected* views, associated to the original images of Figure 1.9. Bottom: a rotation around a horizontal axis (30 degrees step) *predicts* novel views of the scene.

MIP image, we compute an estimation of the distance $d(\hat{x})$ between the visualized voxel $\bar{x}$ selected by (1.3) and (the initial faces of) the Stanford Bunny. Then, $\bar{x}$ is considered as a surface point of the initial object if and only if $d(\hat{x}) < \delta$, where δ denotes the edge length of a voxel. Therefore, we form two images based on this criterion: in (b), the map $\hat{x} \mapsto \mathbb{1}_{d(\hat{x})<\delta}\Pi F(\hat{x})$ is an extraction of relevant pixels representing surface points, and in (c), the complementary map $\hat{x} \mapsto \mathbb{1}_{d(\hat{x})\geq\delta}\Pi F(\hat{x})$ contains pixels which do not represent the initial object.



(a) (b) (c)

Figure 1.11: Discrimination of pixels representing surface points. (a) Aerial MIP image of a 3D reconstruction from vertical images of the Stanford Bunny. (b) Pixels of (a) associated to surface points of the Bunny. (c) Pixels of (a) which do not correspond to the Bunny. Here, (a) is an output of the visualization procedure, (b) contains a "true" information about the initial scene, (c)=(a)−(b) represents an error.

We compute some statistics in order to quantify the quality of the aerial MIP image: 15% of the pixels in ΠF correspond to voxels associated to surface points, they explain 43% of the total intensity of the image. The ratio of these values defines a *concentration* of the intensity among the "true" pixels; the value is 2.91. In comparison, among the other pixels, the concentration of the intensity is 0.66. Therefore, in average, the intensity of a true pixel is 4.38 times the intensity of another pixel. This supports the following claim: for a MIP rendering in reflective tomography, the bright points correspond to the surfaces of the initial scene.

1.8.2 Reconstruction from cartoon images of a non-convex object

We test the principle of 3D reflective tomography in the case of cartoon images of a non-convex object. In particular, we illustrate the impact of discontinuities.

We consider a sphere with a dent, defined in spherical coordinates by

$$\rho = 1 + 0.75(r-1)\mathbb{1}_{r<1}, \quad r := \tfrac{1}{0.08}[(\tfrac{\psi}{\pi} + \tfrac{1}{4})^2 + (\tfrac{2\phi}{\pi} + \tfrac{1}{6})^2], \quad \psi \in [-\pi, \pi], \quad \phi \in [-\tfrac{\pi}{2}, \tfrac{\pi}{2}],$$

where ψ is the azimuth, ϕ is the elevation, and $\rho > 0$ is the radius; we compute this object in `Matlab` from a discrete version of the sphere, discretized with 640^2 patches. For a fixed parameter $m \geq 0$, we define on this surface a piecewise constant pattern in spherical coordinates:

$$(\psi, \phi) \mapsto p_m(\psi) p_m(\phi), \quad \text{with} \quad p_m(s) = 0.5 + 0.25 \mathbb{1}_{(ms - \lfloor ms \rfloor) < 0.5}. \tag{1.4}$$

We simulate a circular scan of this patterned object; we generate 801 images of size 201×201 (constant angular step), using plot of surfaces in `Matlab`. *Cartoon* images are considered: the brightness of a pixel is directly the value of the pattern at the visible point. Next, we compute a MIP of a tomographic reconstruction (restricted to a half-space) from this scan. In Figure 1.12, we display one image of the scan and the corresponding MIP re-projection, for severals values of the parameter m.

For $m = 0$, any image of the scan is binary, and contains the *silhouette* of the object. In this case, the concavity cannot be recovered, since the *visual hull* [72] has the same silhouettes than the object itself. On the contrary, for $m > 0$, the discontinuous pattern permits to reconstruct the object with a pattern whose structure is similar. In particular, the concavity appears clearly. Also, it is worth noting that the boundary of the concavity, which corresponds to some geometrical discontinuity, is emphasized in any case.

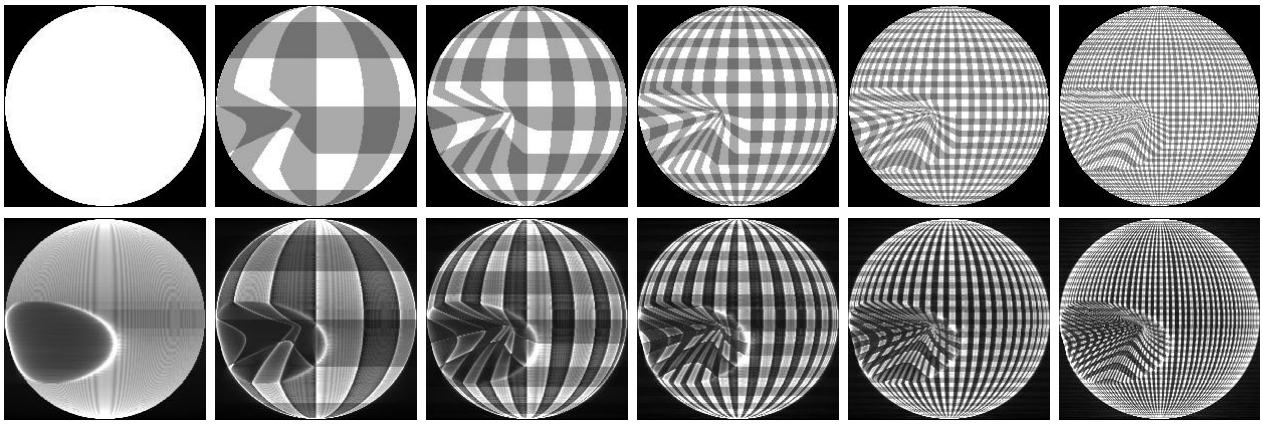


Figure 1.12: 3D tomographic reconstruction from 801 cartoon images 201×201, for a non-convex object with a discontinuous pattern. Top: one image of the input sequence. Bottom: MIP re-projection of the reconstruction. From left to right: the pattern (1.4) on the object has more and more discontinuities, $m = 0$, 1, 2, 4, 8, 16.

1.8.3 Reconstruction from a randomized pattern

We test the principle of reflective tomography on cartoon images of a randomized pattern, which draws some parallel with an active surface whose reflectance varies.

We consider some circular scan $F_\sigma(\beta, y_\perp, y_3)$ of the Stanford Bunny, where $\sigma \geq 0$ is a fixed parameter. This dataset comprises 801 images of size 200×157. For any angle β, we consider a pattern on the Bunny, defined by

$$x \mapsto 1 + (0.2 + \sigma \eta_1(\beta)) \sin(\pi \sigma \eta_2(\beta) + 20\pi |x|), \tag{1.5}$$

where the $\eta_i(\beta)$ are independent realizations of the Gaussian $\mathcal{N}(0,1)$. The image $F_\sigma(\beta, \cdot, \cdot)$ is assumed to be a cartoon image obtained by projection of this pattern. Note that the projected pattern depends on β; this dependence is severe if σ is large. On the first line of Figure 1.13, we represent a slice in the dataset, $(\beta, y_\perp) \mapsto F_\sigma(\beta, y_\perp, 0)$, for several values of σ. In this figure, for $\sigma = 0$, a point which is visible through some angular range appears along a level set. For $\sigma > 0$, this point appears along the same curve, but this is no longer a level set since it is projected with values depending on β.

We compute a tomographic reconstruction from each of these datasets, and we display a MIP image of the reconstruction on the second line of Figure 1.13. As can be observed, the scene is successfully recovered, despite randomized projections. In particular, some "coherent" information such as the silhouette is automatically extracted. In fact, the principle of reflective tomography is robust; it captures the structure of a dataset without measuring photoconsistency.

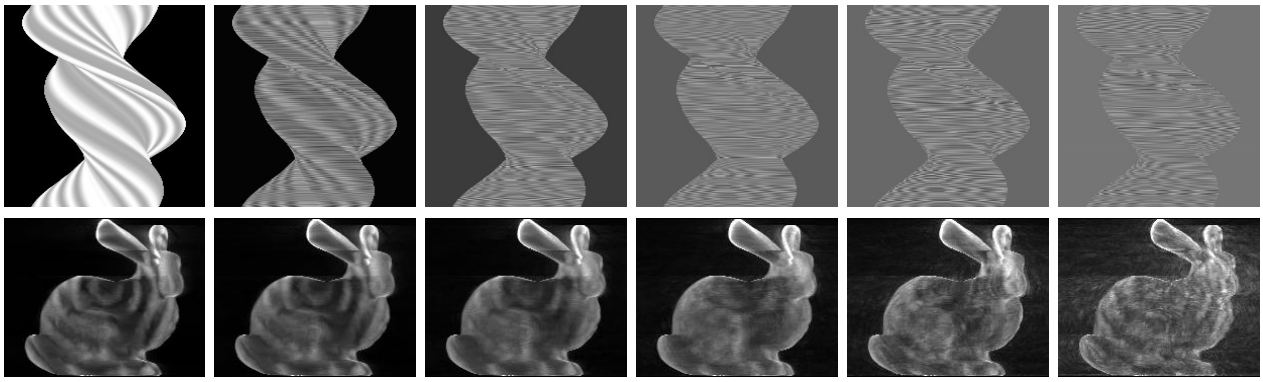


Figure 1.13: 3D tomographic reconstruction from the randomized pattern (1.5). From left to right, the angular dependency of the pattern becomes more and more severe: $\sigma = 0,\ 2^j,\ -2 \leq j \leq 2$. Top: horizontal slice $(\beta, y_\perp) \mapsto F_\sigma(\beta, y_\perp, 0)$ in the dataset. Bottom: MIP view of the reconstruction.

1.8.4 Reconstruction from noisy images

We realize a stability test considering reconstruction from noisy datasets, for speckle noise.

We consider a Gouraud model of the Stanford Bunny as in Subsection 1.8.1; we denote by F the dataset, comprising here 801 images of size 200×157. For normalization purposes, we apply a linear scaling such that the range of F becomes $[1, 2]$ ($F := 1 + \frac{F - \min F}{\max F - \min F}$). For a fixed parameter $\sigma \geq 0$, we introduce a dataset F_σ with a speckle noise of magnitude σ,

$$F_\sigma = F(1 + \sigma\eta), \tag{1.6}$$

where η contains $801 \times 200 \times 157$ independent realizations of the Gaussian $\mathcal{N}(0, 1)$ (and the operations are defined component wise). Then we compute a tomographic reconstruction from the dataset F_σ. In Figure 1.14, we display one image of the input sequence and the corresponding MIP re-projection, for several values of σ. We observe that the visual perception of the reconstructed scene is stable.

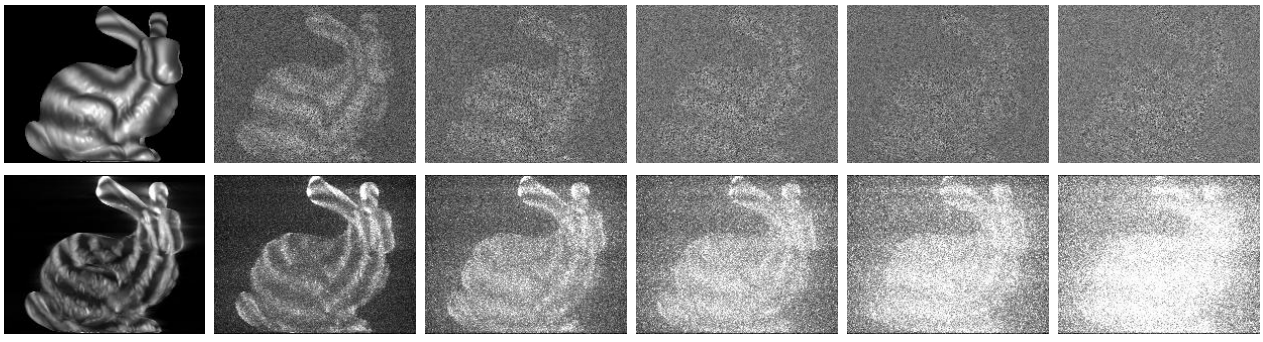


Figure 1.14: 3D tomographic reconstruction from a Gouraud model disturbed by a speckle noise (1.6). From left to right, the level of speckle noise is $\sigma = 0, 0.3, 0.6, 0.9, 1.2, 1.5$. Top: one noisy image of the input sequence. Bottom: MIP re-projection of the reconstruction.

Chapter 2

Mathematical analysis of reflective tomography

2.1 Introduction

Three-dimensional reflective tomography relies on algorithms from X-ray transmission tomography, applied on optical images in the VIS-NIR band; 2D VIS-NIR images are injected into a 3D tomography solver, such as the FDK algorithm, in order to compute a 3D reconstruction. From a mathematical point of view, image formation in VIS-NIR optics may be described by a reflective model, such as the rendering equation (B.7) for the radiance; the records are radiant incidances, modeled by (B.5), whereas the solver aims at inverting the X-ray transform (B.11). This principle of reconstruction introduces some points to be mathematically clarified.

Indeed, there is no guarantee that the recorded data belong to the range of the X-ray transform. A visible point of the scene may appear with different incidances on different images; this is different from the X-ray transform, where a point contributes in the same manner on different rays (by the means of an attenuation coefficient independent from the ray). Furthermore, many materials are opaque for VIS-NIR wavelengths, which produces occlusions. This is a source of non-linearity and it implies that optical images are not expected to be directly in the range of the linear X-ray transform.

Therefore, some mathematical gap to be filled can be expressed as follows: what is the meaning of an X-ray inversion of data which do not belong to the range of the X-ray transform? Moreover, any occlusion introduces some incompleteness in the data [12]. Hence, another question arises concerning the artifacts; the artifacts resulting from the occlusions must be clarified. Such questions are the initial motivation of the works presented in this chapter.

We emphasize some mechanisms of reflective tomography in order to understand the meaning of the reconstruction and to describe the artifacts. We analyze some model problems, where the solver is an X-ray inversion, but the data are not assumed to be in the range of the X-ray transform. To simplify the analysis, we restrict our attention to 2D models, where the X-ray transform is the Radon transform, the acquisition geometry is a parallel scanning, and the inversion procedure is a filtered backprojection (FBP). This is some limit case of the FDK algorithm ($r \to \infty$): for a camera with a large focal length in far field, and a horizontal circular cone beam scan, the FDK algorithm reduces to such a 2D inversion for each horizontal cross-section, analogously to [52], [12]. We refer to Sections 1.3.3 and A.6 for details concerning the FDK algorithm and the FBP.

At a first step, we propose some intuitive interpretation of the FBP as an accumulator array; each pixel value represents an accumulation of "coherent" contrasts along a sinusoid [15]. In this way, reflective tomography looks like a "dual" approach to edge detection by the Hough transform in image processing [44]. In a second step, we formulate a geometrical model of reflective tomography, in the case of piecewise smooth reflective projections [19]. It includes some description of expected artifacts. The model relies on a high frequency asymptotics of the FBP, when the cutoff pulsation tends to infinity. In a third step, we discuss reflective tomography using microlocal analysis of the

Radon transform, as described in Section A.8. This analyzes the problem in term of singularities. The reconstructed geometry, and the artifacts, are encoded by a wavefront set, in correspondence with the wavefront set of the dataset [8]. Lastly, we consider the specific case of a Lambertian convex reflector [3]. In this case, we have more than a correspondence of singularities: a suitable processing of the projections satisfies an exact Radon formula, based on the extension to distributions described in Section A.4; therefore, reflective tomography can be understood here as an extension of transmission tomography, obtained mathematically by extending the Radon transform to distributions.

2.2 Accumulator array of coherent contrasts

2.2.1 Projection of opaque objects

In this section, we consider a model of projection as Figure 2.1. In a plane, we represent the boundary of a collection of opaque objects by a bounded set $\Gamma \subset \mathbb{R}^2$; we assume that Γ is a disjoint union of piecewise smooth Jordan curves. The scene Γ is projected towards a screen of sensors aligned with a direction $\theta = (\theta_1, \theta_2) \in \mathbb{S}^1$, along the orthogonal direction $\theta^\perp = (\theta_2, -\theta_1) \in \mathbb{S}^1$; for instance, the two dotted lines of Figure 2.1 represent two rays of projection. On a line $x \cdot \theta = s$, where $s \in \mathbb{R}$ is a parameter, the visible point $y(\theta, s) \in \Gamma$ is such that the ray from $y(\theta, s)$ does not meet Γ. In this case, the sensor records an information $F(\theta, s)$, which depends on the visible point and on the angle,

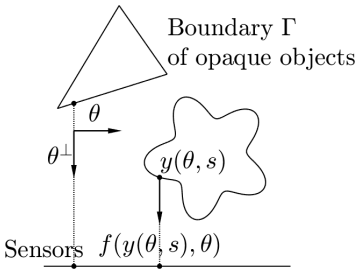


Figure 2.1: Projection of opaque objects.

$$F(\theta, s) = f(y(\theta, s), \theta). \tag{2.1}$$

The function f is analogous to the radiant incidance, or the radiance of the visible point, in (B.5); $f(y, \theta)$ represents the brightness of the visible point y on a screen defined by θ. If the line $x \cdot \theta = s$ does not meet Γ, we fix $F(\theta, s) = 0$ ("background"). To finish with, the same process is repeated for several orientations of the screen; the angle θ scans a finite set $\Theta \subset \mathbb{S}^1$. In this way, the dataset is $F(\theta, s)$, $\theta \in \Theta$, $s \in \mathbb{R}$. By definition, for any angle $\theta \in \Theta$, the projection $s \in \mathbb{R} \mapsto F(\theta, s)$ is a compactly supported function, defined by (2.1) on the support.

For illustration purposes, an example is considered in Figure 2.2. To simplify, the surface function $f(y)$ of this example does not depend on the angle (as for cartoon images). On the left, we display the projection $F(\theta_0, s)$ of $f(y)$, for a fixed angle θ_0. On the right, we display the dataset $F(\theta, s)$; the angle $\theta \in \Theta$ scans a uniform discretization of $\mathbb{S}^1$,

$$\Theta = \{(\cos i\delta\theta, \sin i\delta\theta),\ 0 \leq i \leq 359\},\ \delta\theta = \tfrac{\pi}{180}.$$

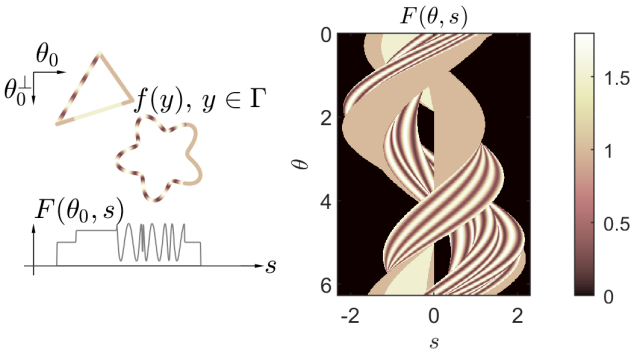


Figure 2.2: Example of dataset.

Any point $y \in \Gamma$ appears in F along pieces of the sinusoid $y \cdot \theta = s$, with brightness $f(y)$; in general, $y \in \Gamma$ does not appear along the whole sinusoid, due to occlusions.

2.2.2 Assumptions: piecewise smoothness

In this section, we assume for technical reasons that for any $\theta \in \Theta$, the projection $F(\theta, \cdot) : s \mapsto F(\theta, s)$ is a piecewise smooth function, with a finite number of singularities. In particular, this assumption tolerates the following discontinuities.

- At the boundary of the support of the projection $F(\theta,\cdot)$, $F(\theta,\cdot)$ jumps from the value 0 of the background to the value $f(y(\theta,s),\theta)$ of the visible point $y(\theta,s)\in\Gamma$. For instance, on the left of Figure 2.2, such a discontinuity occurs for the value of s associated to the bottom-left corner of the triangle. This kind of jump is directly related to the geometry of Γ.

- In the support of the projection $F(\theta,\cdot)$, the visible point $s\mapsto y(\theta,s)$ may jump. For instance, in Figure 2.1, there exists a critical ray $x\cdot\theta=s$ such that $y(\theta,s)$ jumps from the triangle to the star. Such a jump for $y(\theta,s)$ may imply a discontinuity for $F(\theta,s)=f(y(\theta,s),\theta)$, even if f is smooth. This kind of jump is directly related to the geometry of Γ.

- In the support of the projection $F(\theta,\cdot)$, even if $s\mapsto y(\theta,s)$ is continuous, the surface function $y\in\Gamma\mapsto f(y,\theta)$ may be discontinuous, which may introduce discontinuities in the projection $F(\theta,s)=f(y(\theta,s),\theta)$. For example, in Figure 2.2, a jump of f on the triangle introduces a jump for $F(\theta_0,s)$. This kind of jump can depend on the geometry of Γ; it can also be related to physical parameters (discontinuities at the interface between separate materials).

Notice that our considerations does not use any equation on $f(y,\theta)$; in particular, f is not constrained to satisfy a rendering equation such as (B.7).

2.2.3 A reconstruction formula

Following the principle of reflective tomography, we introduce an X-ray transform for the considered acquisition geometry. Here, the scene is projected along lines $x\cdot\theta=s$ with $(\theta,s)\in\mathbb{S}^1\times\mathbb{R}$. Therefore, the well-suited X-ray transform is the Radon transform $\mathcal{R}$ defined by (A.1),

$$\mathcal{R}a(\theta,s)=\int_{x\cdot\theta=s}a(x)\,\mathrm{d}\ell=\int_{\mathbb{R}}a(s\theta+t\theta^{\perp})\,\mathrm{d}t,\quad(\theta,s)\in\mathbb{S}^1\times\mathbb{R},\quad a:\mathbb{R}^2\to\mathbb{R}.\tag{2.2}$$

Then, we reconstruct the scene from the dataset F defined in (2.1), by a FBP inspired from the Radon formula (A.9) and the FBP (A.14). Indeed, we define a reconstruction formula by

$$\mathcal{I}[F](x):=\mathcal{B}[F\star\psi_\Omega](x),\quad x\in\mathbb{R}^2,\tag{2.3}$$

where, $\star$ denotes the convolution with respect to the variable s, the convolution kernel $\psi_\Omega(s)$ is defined by (A.13), and the operator $\mathcal{B}$ is the discrete backprojection

$$\mathcal{B}g(x)=\sum_{\theta\in\Theta}g(\theta,x\cdot\theta),\quad x\in\mathbb{R}^2.\tag{2.4}$$

In comparison with the original FBP (A.14), the tomographic projection $\mathcal{R}f$ has been replaced by the reflective projection F, and the backprojection $\mathcal{R}^*$ defined in (A.5) has been replaced by an analogous discrete operator. Note that the reconstruction (2.3) is a smooth function

$$\mathcal{I}[F]\in\mathcal{E}(\mathbb{R}^2).$$

Indeed, for any $\theta\in\Theta$, the projection $F(\theta,\cdot)$ is a function with compact support, and is assumed to be piecewise smooth. Then, for similar reasons than (A.16),

$$F(\theta,\cdot)\star\psi_\Omega=\mathcal{F}_1^{-1}\{\mathcal{F}_1[F(\theta,\cdot)]\mathcal{F}_1[\psi_\Omega]\}\in\mathcal{E}(\mathbb{R})\cap\mathcal{S}'(\mathbb{R}),$$

where $\mathcal{F}_1$ denotes the Fourier transform (A.2). Therefore, $x\mapsto F(\theta,\cdot)\star\psi_\Omega(x\cdot\theta)\in\mathcal{E}(\mathbb{R}^2)$, and we conclude by a finite summation over θ.

Equivalently, the reconstruction formula (2.3) is given by

$$\mathcal{I}[F](x)=\mathcal{B}[\tfrac{1}{4\pi}\partial_sF\star\phi_\Omega](x),\quad x\in\mathbb{R}^2,\tag{2.5}$$

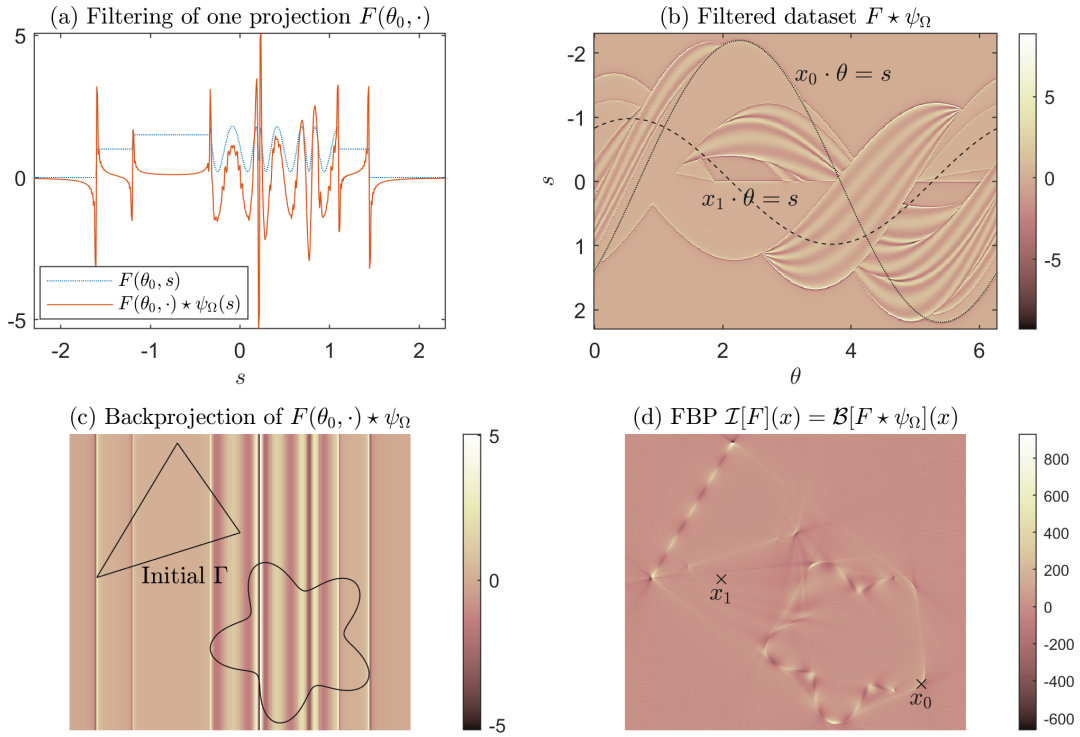


Figure 2.3: Reconstruction by FBP from the projection F of Figure 2.2. The FBP from a single projection $F(\theta_0,\cdot)$ is a backprojection (c) from the filtered projection $F(\theta_0,\cdot)\star\psi_\Omega$ (a). The total FBP $\mathcal{I}[F](x)$ (d) is a summation through the sinusoid $x\cdot\theta=s$ in the filtered dataset $F\star\psi_\Omega$ (b).

where ϕ_Ω is a regularization of the Hilbert transform, defined by

$$\phi_\Omega=\mathcal{F}_1^{-1}[-\mathrm{i}\,\mathrm{sign}(\sigma)\hat{h}_\Omega(\sigma)]\in\mathcal{E}(\mathbb{R})\cap\mathcal{S}'(\mathbb{R}); \tag{2.6}$$

here, $\hat{h}_\Omega:\mathbb{R}\to[0,1]$ denotes a windowing function which is even with compact support $[-\Omega,\Omega]$. In (2.5), $\frac{1}{4\pi}\partial_s F\star\phi_\Omega=F\star\psi_\Omega$, and for any $\theta\in\Theta$, $\partial_s F(\theta,\cdot)\in\mathcal{E}'(\mathbb{R})$ is a distributional derivative with compact support.

For illustration purposes, Figure 2.3 deals with the reconstruction from the projection of Figure 2.2. Filtering is illustrated for a single projection in (a), and for the whole dataset in (b). The backprojection from a single filtered projection is displayed in (c), superimposed with the initial Γ. The final reconstruction is the backprojection from the full filtered dataset, displayed in (d).

2.2.4 Accumulation of coherent contrasts

We discuss the proposed formula (2.5), using Figure 2.3 as an illustration. The first comments deal with filtering. Due to piecewise smoothness, a filtered projection is equal to

$$\tfrac{1}{4\pi}\partial_s F(\theta,\cdot)\star\phi_\Omega(t)=\tfrac{1}{4\pi}\int_{\mathbb{R}}\{\partial_s F(\theta,s)\}\phi_\Omega(t-s)\mathrm{d}s+\tfrac{1}{4\pi}\sum_{s\in j(\theta)}[F(\theta,s)]\phi_\Omega(t-s), \tag{2.7}$$

where $\{\partial_s F(\theta,s)\}$ denotes the usual derivative of $F(\theta,s)$ (defined almost everwhere), the finite set $j(\theta)$ contains the points s where $F(\theta,\cdot)$ jumps, and $[F(\theta,s)]=F(\theta,s^+)-F(\theta,s^-)$ is the amplitude of a jump. Therefore, filtering especially enhances variations and discontinuities, as in (a-b). Also, due to the shape of the kernel ϕ_Ω, the right term in (2.7) behaves like a zero-crossing detection of contours in the projection $F(\theta,\cdot)$, as in (a); recall that some associated jumps are directly related to the initial geometry Γ, as described in Subsection 2.2.2.

Secondly, the backprojection of a single filtered projection defines a plane wave

$$x\mapsto\tfrac{1}{4\pi}\partial_s F(\theta,\cdot)\star\phi_\Omega(x\cdot\theta),$$

as in (c). Due to the structure of $\partial_s F(\theta,\cdot) \star \phi_\Omega$, the most significant values of this plane wave are related to contrasts in $F(\theta,\cdot)$. Next, the reconstruction $\mathcal{I}[F]$ is a summation of these plane waves (d). It contains essentially "small" values. Nevertheless, "large" values appear when the summation contains significant values which are constructively added. These values correspond to "coherent" significant contrasts, and this explains intuitively why some part of Γ is bright in the reconstruction.

This phenomenon appears clearly in (b) and (d). Indeed, for any reconstruction point x, $\mathcal{I}[F](x)$ represents a summation along the sinusoid $x \cdot \theta = s$ in the filtered dataset $\frac{1}{4\pi}\partial_s F \star \phi_\Omega$ (b). For a "generic" point x_1, the filtered values, along the sinusoid $x_1 \cdot \theta = s$, are "incoherent", and their summation $\mathcal{I}[F](x_1)$ is "small". On the contrary, for some specific points x_0, the sinusoid $x_0 \cdot \theta = s$ contains a portion of significant values which are "coherent"; in this case, $\mathcal{I}[F](x_0)$ is "large". These specific points especially include points which are close to Γ, because if $x_0 \in \Gamma$, then x_0 is visible along some portions of the sinusoid $x_0 \cdot \theta = s$, so "coherence" is expected along these portions.

Finally, the reconstruction looks like some "accumulation of coherent contrasts, backprojected at their initial location in space". Peaks are expected near the initial surfaces, and especially near points which appear along coherent contrasts in the projections. For the considered example, one reconstruct some contrasted image of Γ, even for non-convex portions. Also, the right of the star appears uniformly in the dataset (f is constant); such a lack of contrast implies that the associated geometry is not reconstructed.

2.2.5 Parallel with edge detection by the Hough transform

We draw some parallel between reflective tomography and edge detection by the Hough transform.

In image processing, the Hough transform is a standard way of finding edges in an image. Basically, a binary image of contours is computed. Then an accumulator array $A(\theta, s)$ is computed by the Hough transform: along any line $x \cdot \theta = s$, $A(\theta, s)$ counts the number of pixels labeled as contour. Any peak in the array $A(\theta, s)$ is associated to a line $x \cdot \theta = s$ that may contain an edge of the original image.

Reflective tomography looks like a similar approach for sinusoid detection. Filtering enhances contrasts and contours in the known image $F(\theta, s)$. Then the backprojection $\mathcal{B}$ sums along sinusoids $x \cdot \theta = s$. The resulting reconstruction $\mathcal{I}[F](x)$ plays the role of an accumulator array; the peaks are associated to sinusoids which may correspond to visible points of the initial scene.

Both methods start by enhancing the desired structures. And both methods compute some accumulator array by a summation along sets $x \cdot \theta = s$. For edge detection, the summation is performed by the Hough transform, along lines with parameter (θ, s). For tomography, the summation is performed by the backprojection $\mathcal{B}$, along sinusoids with parameter x. Therefore, a FBP looks like a "dual" approach to edge detection by the Hough transform.

2.3 Asymptotic and geometrical modeling

2.3.1 Introduction

In Section 2.2, we have essentially interpreted some model of reflective tomography as an accumulation/cancellation of coherent/incoherent waves. We further investigate this point, from a quantitative point of view. Summation of waves is often studied in a framework of high frequency asymptotics. Arguments such as stationary phase approximation lead to geometrical modeling; this is for instance the basis of geometrical optics. We follow such a strategy in order to study a FBP on some models of reflective projections.

For that purpose, we consider projections $F(\theta,\cdot)$ of opaque objects, similarly as Figure 2.1. and Subsection 2.2.1. We assume now that the angle θ scans the continuous circle $\mathbb{S}^1$, and we still assume that F has a compact support. Following the principle of reflective tomography, we reconstruct the scene by a FBP such as (A.14). Therefore, we consider

$$\mathcal{I}_\Omega(x) := \mathcal{R}^*[F \star \psi_\Omega](x), \tag{2.8}$$

where $\mathcal{R}^*$ is the backprojection (A.5), and ψ_Ω is the Ram-Lak filter defined by (A.13), with $\hat{h}_\Omega(\sigma) = \mathbb{1}_{[-\Omega,\Omega]}(\sigma)$. In practice, the cut-off frequency Ω is bounded above by the Shannon-Nyquist frequency associated to the radial discretization (variable s). It plays the role of a resolution parameter which is ideally very large. That is the reason why we look for an asymptotic expansion of the reconstruction (2.8), when Ω tends to infinity.

Overall, the asymptotic behavior of (2.8) is especially related to the geometry of the dataset F, and it provides some geometrical model of the reconstruction. It is obtained by asymptotic expansion of suitable integrals. In this section, we summarize the shape of the leading order terms, and we insist on the results for two toy models.

2.3.2 Highly oscillatory integrals

Deriving an asymptotics for $\mathcal{I}_\Omega(x)$, $\Omega \to \infty$, means studying the oscillatory integral

$$\mathcal{I}_\Omega(x) = \int_{\mathbb{S}^1}\int_{\mathbb{R}} F(\theta,s)\psi_\Omega(x\cdot\theta - s)\mathrm{d}s\mathrm{d}\theta = \tfrac{\Omega^2}{4\pi^2}\int_0^1 \nu \int_{\mathbb{S}^1}\int_{\mathbb{R}} F(\theta,s)\cos(\Omega\nu(x\cdot\theta-s))\mathrm{d}s\mathrm{d}\theta\mathrm{d}\nu, \quad (2.9)$$

because the Ram-Lak filter (A.13) satisfies $\psi_\Omega(s) = \frac{\Omega^2}{4\pi^2}\int_0^1 \nu\cos(\Omega\nu s)\mathrm{d}\nu$.

Such integrals are studied in the working document [19], from asymptotic techniques described in the textbook [37]. The results are obtained in three steps. The first step establishes some asssumptions about the geometrical structure of F, especially concerning the location of singularities. The second step derives an asymptotic expansion of

$$g(\lambda) = \int_{\mathbb{S}^1}\int_{\mathbb{R}} F(\theta,s)\cos(\lambda(x\cdot\theta-s))\mathrm{d}s\mathrm{d}\theta, \quad \lambda\to\infty;$$

the proof is based on an iterated divergence formula and a stationary phase method. The third step derives an asymptotic expansion of

$$\tfrac{4\pi^2}{\Omega^2}\mathcal{I}_\Omega(x) = \int_0^1 \nu g(\Omega\nu)\mathrm{d}\nu, \quad \Omega\to\infty;$$

this is a study of a g-transform, based on the Mellin transform and a calculation of residues.

The asymptotic expansions are used in order to predict the expected orders of magnitude of the reconstruction $\mathcal{I}_\Omega(x)$, depending on x. Roughly speaking, the expected orders are the following:

- $\mathcal{O}(\sqrt{\Omega})$ on some convex portions of the original scene Γ;
- $\mathcal{O}(\log\Omega)$ for some isolated points of the original scene Γ, such as corners;
- $\mathcal{O}(\log\Omega)$ on some straight lines, which generally represent artifacts;
- $\mathcal{O}(1)$ almost everywhere, which represents a noise.

In particular, some parts of the initial scene Γ are expected to be bright in the reconstruction $\mathcal{I}_\Omega(x)$, with order $\mathcal{O}(\sqrt{\Omega})$. Moreover, some description of the artifacts resulting from the occlusions is obtained; straight lines (associated to corners in F) appear with order $\mathcal{O}(\log\Omega)$.

Remark 2.1. Artifacts due to a limited angle (when θ scans only a portion of $\mathbb{S}^1$) or a spatial truncation (when s scans only an interval $[-R,R]$) are also considered in [19]; the associated order of magnitude is again $\mathcal{O}(\log\Omega)$.

2.3.3 Toy models

We present expected asymptotics of the reconstruction $\mathcal{I}_\Omega = \mathcal{R}^*[F \star \psi_\Omega]$, for two toy models extracted from [19]. Also, we check some results from a numerical point of view.

The first case deals with the reconstruction from *silhouettes* of a smooth convex object, as in Figure 2.4. The projected function is $f(y,\theta) = 1$, $y \in \Gamma$, $\theta \in \mathbb{S}^1$, and the set Γ is an ellipse. Denoting the curvature of Γ by κ, the reconstruction (essentially) satisfies

$$\mathcal{I}_\Omega(x) = \begin{cases} \left(\frac{1}{\pi^{3/2}}\sqrt{\kappa(x)} + o(1)\right)\sqrt{\Omega}, & \text{if } x \in \Gamma, \\ \mathcal{O}(1), & \text{if } x \notin \Gamma. \end{cases} \tag{2.10}$$

Due to this model, the ellipse Γ is expected to appear bright in the reconstruction; this is observed in Figure 2.5(a). Up to a factor, the brightness of the ellipse is expected to be the square root of the curvature; this is confirmed in Figure 2.5(b). Lastly, the residuals $o(1)$ and $\mathcal{O}(1)$ are observed in Figure 2.5(c), for some $x = x_0 \in \Gamma$, and $x = x_1 \notin \Gamma$.

The second toy model, displayed in Figures 2.6-2.7, deals with occlusions. The set Γ is a union of two disjoint circles. *Cartoon* projections are considered: the first circle, resp. second circle, appears in the projections with $f(y,\theta) = f_1$, resp. $f(y,\theta) = f_2$, where f_1 and f_2 are two fixed values. In comparison with Figure 2.4(b), the dataset 2.6(b) looks like two interlaced weighted silhouettes. For the reconstruction $\mathcal{I}_\Omega$, several asymptotic regimes are expected. In brief, the circles should appear with order $\mathcal{O}(\sqrt{\Omega})$, the four straight lines which are tangent to the two circles should appear with order $\mathcal{O}(\log\Omega)$, whereas the order is $\mathcal{O}(1)$ almost everywhere. Therefore, the circles are expected to be peaks in the reconstruction $\mathcal{I}_\Omega$. So are the four mentioned tangent lines; these lines are expected artifacts. This is in agreement with the numerical reconstruction of Figure 2.7(a). Furthermore, in Figure 2.6(c) and Figure 2.7(a), we have selected four points, x_i, $0 \leq i \leq 3$, in order to illustrate the shape of the asymptotics. The point x_0 belongs to the first circle; in the singularities of Figure 2.6(b), x_0 appears independently of the second circle. Then, analogously to 2.10,

$$\mathcal{I}_\Omega(x_0) = \left(\frac{f_1\sqrt{\kappa_1}}{\pi^{3/2}} + o(1)\right)\sqrt{\Omega}, \tag{2.11}$$

where κ_1 denotes the curvature of the first circle. The point x_1 belongs also to the first circle, but appears only once in the singularities of Figure 2.6(b), with a jump $f_1 - f_2$; the associated asymptotics is given by

$$\mathcal{I}_\Omega(x_1) = \left(\frac{(f_1 - f_2)\sqrt{\kappa_1}}{2\pi^{3/2}} + o(1)\right)\sqrt{\Omega}. \tag{2.12}$$

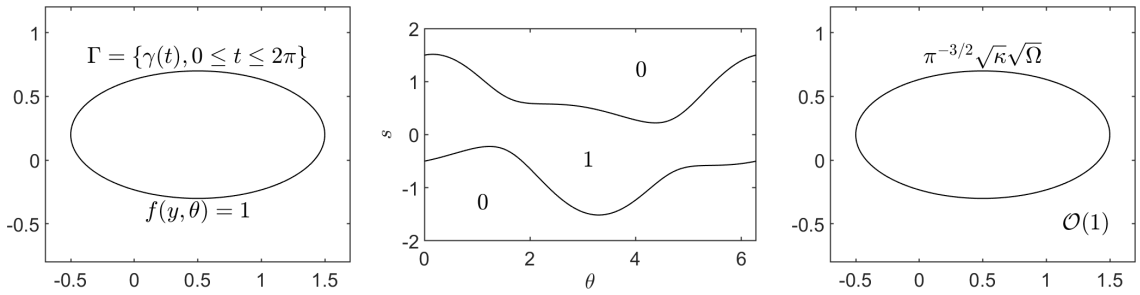
The point $x_2 = (1-\lambda)a + \lambda b$ belongs to a line which is tangent to the first and the second circle, on a point a and a point b; this line corresponds to two corners in Figure 2.6(b). The asymptotics is

$$\mathcal{I}_\Omega(x_2) = \left(\frac{-(f_1 + f_2)}{4\pi^2|a-b|\lambda(1-\lambda)} + o(1)\right)\log\Omega. \tag{2.13}$$

Lastly, the point x_3 is neither on the circles, neither on the four tangent lines. The asymptotics is

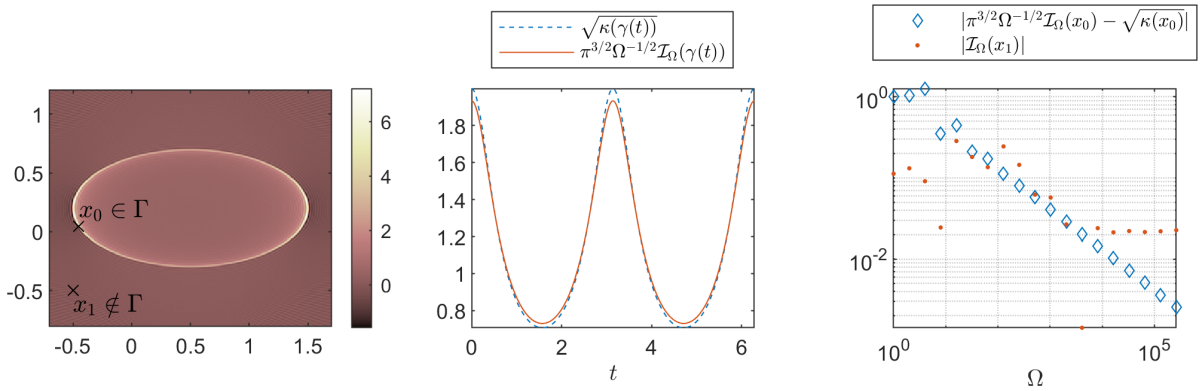
$$\mathcal{I}_\Omega(x_3) = \mathcal{O}(1). \tag{2.14}$$

These four asymptotic regimes are numerically checked in Figure 2.7(b).



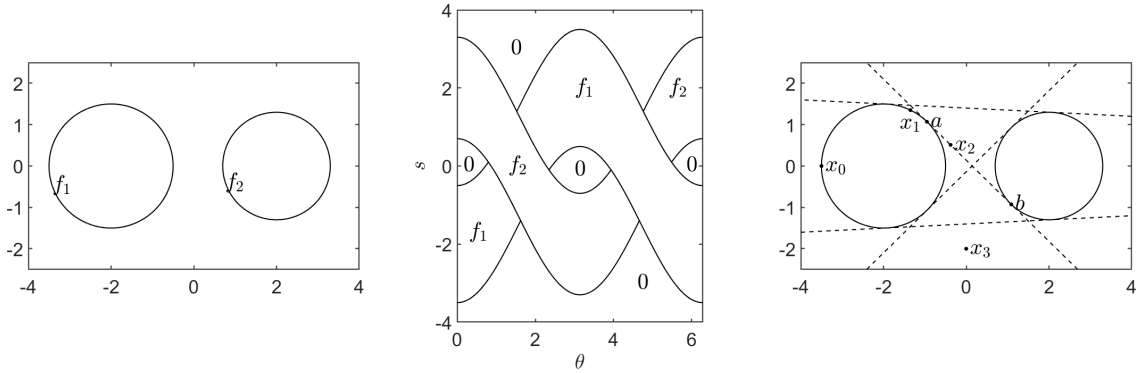
(a) Initial scene Γ (b) Projection $F(\theta, s)$ (silhouette) (c) Asymptotic model of $\mathcal{I}_\Omega$

Figure 2.4: Asymptotic model (2.10) of the reconstruction $\mathcal{I}_\Omega$ from the silhouette F of an ellipse Γ with curvature κ. The ellipse Γ is expected to be bright in the reconstruction.



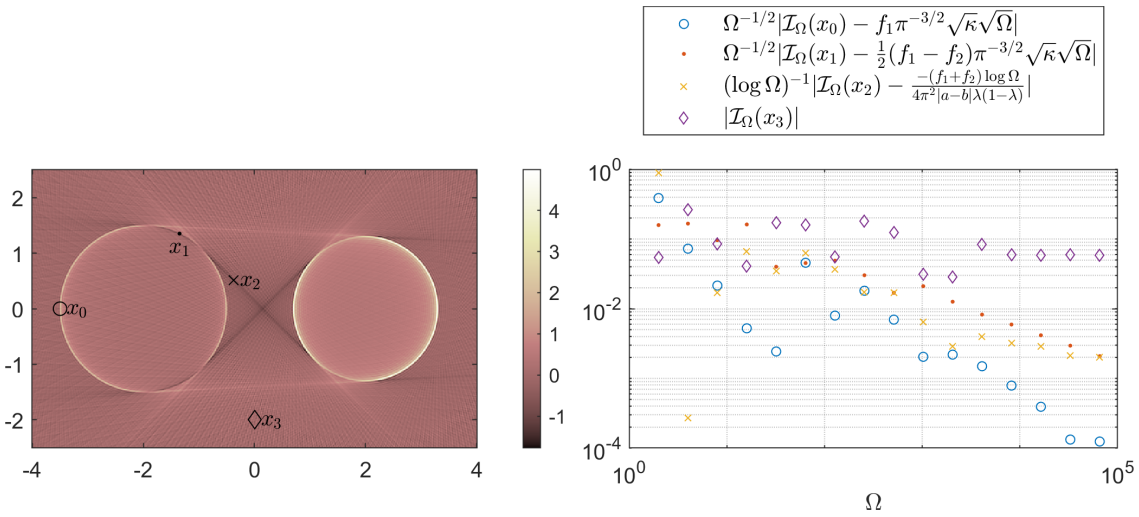
(a) (b) (c)

Figure 2.5: Numerical check of the asymptotic model (2.10) of Figure 2.4, for an ellipse Γ with curvature κ.
(a) Reconstruction $\mathcal{I}_\Omega$, computed with $\Omega = 300$. The ellipse Γ is bright.
(b) Check of the asymptotic constant $\sqrt{\kappa}$ along $\Gamma = \{\gamma(t), 0 \le t \le 2\pi\}$, for $\Omega = 2^{11}$.
(c) For $x_0 \in \Gamma$, resp. $x_1 \notin \Gamma$, plotted in (a), the computed residual is $o(1)$, resp. $\mathcal{O}(1)$, as expected.



(a) Initial scene Γ (b) Projection $F(\theta, s)$ (cartoon) (c) Geometrical model of the reconstruction $\mathcal{I}_\Omega$

Figure 2.6: Modeling of the reconstruction $\mathcal{I}_\Omega$ from cartoon projections F of two circles. The circles are expected to be bright in the reconstruction; the lines tangent to both circles are expected artifacts.



(a) (b)

Figure 2.7: Numerical check of the asymptotic model, for the two circles of Figure 2.6.
(a) Reconstruction $\mathcal{I}_\Omega$, computed with $\Omega = 167.55$. Peaks appear for the circles and the four common tangent lines, as expected.
(b) Numerical residuals of the asymptotic models (2.11)-(2.14), at x_i, $0 \leq i \leq 3$.

2.4 Imaging of singularities

2.4.1 Introduction

Chapter 1 and Sections 2.2-2.3 reveal that singularities play a central role in reflective tomography. This suggests to take a closer look at the singularities of a Radon transform. Hence, we propose now a study based on the microlocal analysis of the Radon transform. We refer to Appendix A.8 for a summary of the mathematical background, including some bibliographic references.

In this section, we define a general principle of imaging which extends the principle of reflective tomography to generic projections, as in [8]. This principle includes incomplete data tomography [38] and is analyzed analogously; the deep mathematical support is Theorem A.8, about the singularities of a tomographic reconstruction. We illustrate the principle on a toy model of reflective tomography, comprising two Lambertian disks.

2.4.2 A general principle

We formulate a general principle to recover a scene from projections along lines.

Principle (Tomographic reconstruction from generic projections)**.** *Let $F \in L^1(\mathbb{S}^1 \times \mathbb{R}) \cap \mathcal{E}'(\mathbb{S}^1 \times \mathbb{R})$ be a function with compact support. Let $\mathcal{R}^*$ denote the backprojection defined in* (A.8)*, and let $\Lambda : \mathcal{E}'(\mathbb{S}^1 \times \mathbb{R}) \to \mathcal{D}'(\mathbb{S}^1 \times \mathbb{R})$ be a pseudodifferential operator. Assuming that $F(\theta, s)$ represents some projection along the ray $x \cdot \theta = s$, $(\theta, s) \in \mathbb{S}^1 \times \mathbb{R}$, the FBP $\mathcal{R}^*\Lambda F$ is considered as a reconstruction of the projected scene. If $f \in \mathcal{E}'(\mathbb{R}^2)$ is a suitable representation of the initial scene, $\mathcal{R}^*\Lambda F$ is expected to share similarities with f.*

In this generic principle, $F(\theta, s)$ can represent any kind of projection with parameter (θ, s). In general, the projection $F(\theta, s)$ depends on the geometry of the projected scene, and eventually on physical parameters. It can be given by a measurement or a computation. Eventually, F is known on a compact subset of $\mathbb{S}^1 \times \mathbb{R}$ and is extended by 0. In particular, the principle includes:

- reflective tomography in optics, where $F(\theta, s)$ represents a radiant incidance such as (B.5),

- tomography from silhouettes or from cartoon images, such as the toy models of Subsection 2.3.3,
- X-ray tomography, where $F = \mathcal{R}f$ is a Radon transform,
- incomplete data tomography, where $F = \mathbb{1}_A \mathcal{R}f$ is a truncated Radon transform ($A \subsetneq \mathbb{S}^1 \times \mathbb{R}$).

Concerning the operator Λ, here are some classical choices:

- in X-ray tomography, with $F = \mathcal{R}f$ and $\Lambda = \frac{1}{4\pi}\mathcal{H}_s\partial_s$, one inverts the Radon transform due to (A.12), $\mathcal{R}^*\Lambda F = f$;
- in local tomography [88], with $F = \mathcal{R}f$ and $\Lambda = -\frac{1}{4\pi}\partial_s^2$, one reconstructs $\mathcal{R}^*\Lambda F = \sqrt{-\Delta} f$, which has the same wavefront set than f, but sharper singularities;
- in reflective tomography, ones usually considers the FBP from X-ray tomography, $\Lambda = \frac{1}{4\pi}\mathcal{H}_s\partial_s$.

In general, a function $F(\theta, s)$, $(\theta, s) \in \mathbb{S}^1 \times \mathbb{R}$, belongs to the range of the Radon transform $\mathcal{R}$ only under exceptional conditions, and we do not have an explicit formula to describe the content of a tomographic reconstruction $\mathcal{R}^*\Lambda F$. Nevertheless, we claim that a FBP $\mathcal{R}^*\Lambda F$, as proposed by the principle, is expected to be relevant.

2.4.3 Mathematical background

We motivate the principle by an analysis of singularities, based on Theorem A.8. For that purpose, we assume that $F \in L^1(\mathbb{S}^1 \times \mathbb{R}) \cap \mathcal{E}'(\mathbb{S}^1 \times \mathbb{R})$ contains some projection of a scene, we assume that $f \in \mathcal{E}'(\mathbb{R}^2)$ is a suitable representation of this scene, and we fix a pseudodifferential operator $\Lambda : \mathcal{E}'(\mathbb{S}^1 \times \mathbb{R}) \to \mathcal{D}'(\mathbb{S}^1 \times \mathbb{R})$.

Firstly, F and the Radon transform $\mathcal{R}f \in \mathcal{E}'(\mathbb{S}^1 \times \mathbb{R})$ define two projections of the same scene, along the same rays. Despite F and $\mathcal{R}f$ may not represent the same quantities, there is a hope that they have geometrical similarities. More precisely, the wavefront sets $\mathrm{WF}\, F$ and $\mathrm{WF}\, \mathcal{R}f$ should have a significant intersection. Using (A.19), the intersection $\mathrm{WF}\, F \cap \mathrm{WF}\, \mathcal{R}f$ should capture some $(\theta, s; \hat\theta, \hat s)$ such that $(s\theta + \frac{\hat\theta}{\hat s}\theta^\perp; \hat s\theta) \in \mathrm{WF}(f)$. This is a precise way of defining similarities between F and $\mathcal{R}f$, and it explains why F is "not so far" from the range of $\mathcal{R}$. On the contrary, if $(\theta, s; \hat\theta, \hat s) \in \mathrm{WF}\, F \setminus \mathrm{WF}\, \mathcal{R}f$ with $\hat s \neq 0$, then $(s\theta + \frac{\hat\theta}{\hat s}\theta^\perp; \hat s\theta) \notin \mathrm{WF}(f)$.

Secondly, we deduce from (A.21) that the reconstruction $\mathcal{R}^*\Lambda F$ is such that

$$\mathrm{WF}(\mathcal{R}^*\Lambda F) \subset \big\{(s\theta + \tfrac{\hat\theta}{\hat s}\theta^\perp; \hat s\theta), \text{ with } (\theta, s; \hat\theta, \hat s) \in \mathrm{WF}(F) \text{ and } \hat s \neq 0\big\}.$$

We split the right member into the two following subsets.

- The first subset contains singularities of f captured by F,

$$S_{F,f} := \big\{(s\theta + \tfrac{\hat\theta}{\hat s}\theta^\perp; \hat s\theta), \text{ with } (\theta, s; \hat\theta, \hat s) \in \mathrm{WF}(F) \cap \mathrm{WF}\, \mathcal{R}f,\ \hat s \neq 0\big\} \subset \mathrm{WF}\, f.$$

- The second subset does not correspond to any singularity of f,

$$A_{F,f} := \big\{(s\theta + \tfrac{\hat\theta}{\hat s}\theta^\perp; \hat s\theta), \text{ with } (\theta, s; \hat\theta, \hat s) \in \mathrm{WF}(F) \setminus \mathrm{WF}\, \mathcal{R}f,\ \hat s \neq 0\big\} \subset \mathbb{R}^4 \setminus \mathrm{WF}\, f.$$

As a result, the wavefront set of the reconstruction $\mathcal{R}^*\Lambda F$ contains two complementary parts.

- The first one is $\mathrm{WF}(\mathcal{R}^*\Lambda F) \cap S_{F,f} \subset \mathrm{WF}\, f$; it contains the singularities of the scene f which are successfully recovered by the imaging process ("S" for success).
- The second one is $\mathrm{WF}(\mathcal{R}^*\Lambda F) \cap A_{F,f} \subset \mathbb{R}^4 \setminus \mathrm{WF}\, f$; the singularities of this set correspond to artifacts of the imaging process, because they are not related to the representation f of the scene ("A" for artifacts).

Finally, this framework gives a meaning to a FBP $\mathcal{R}^*\Lambda F$, even if F does not belong to the range of the Radon transform. A reconstruction $\mathcal{R}^*\Lambda F$ has a wavefront set $\mathrm{WF}\,\mathcal{R}^*\Lambda F$ which contains partially the wavefront set of the initial scene, augmented by a set which represents artifacts. Remarkably, the process captures efficiently the initial geometry, without modeling the content of F (F could represent a Radon transform, an incomplete Radon transform, a time of flight, a radiant incidance, a geometrical quantity, and so on).

2.4.4 Reflective tomography of two Lambertian disks

We present the principle of tomographic reconstruction for a toy model of reflective tomography, with occlusions [8].

We consider a projection F which models the radiant incidance from two Lambertian disks. The scene $K = K_1 \cup K_2$ contains two disjoint disks K_1 and K_2. They define Lambertian reflectors with constant albedo $\rho_1, \rho_2 > 0$. The projection is analogous to the projection of Figure 2.1, with $f(y(\theta, s), \theta)$ given by a Lambert's cosine law (B.8),

$$F(\theta, s) = \begin{cases} \rho(y(\theta, s))\, \theta^\perp \cdot \nu(\theta, s), & \text{if } \{x \cdot \theta = s\} \cap K \neq \emptyset, \\ 0, & \text{if } \{x \cdot \theta = s\} \cap K = \emptyset. \end{cases} \tag{2.15}$$

Here, $y(\theta, s) \in \partial K$ denotes the visible point (*i.e.* $y(\theta, s)$ maximizes $x \cdot \theta^\perp$ on the set $\{x \cdot \theta = s\} \cap K$), $\nu(\theta, s) \in \mathbb{S}^1$ denotes the exterior normal vector to ∂K at $y(\theta, s)$, the cosine $\theta^\perp \cdot \nu(\theta, s)$ represents the cosine of an angle of incidence, and $\rho(y) \in \{\rho_1, \rho_2\}$ denotes the albedo of the point $y \in \partial K$,

$$\rho(y) = \rho_1 \mathbb{1}_{K_1}(y) + \rho_2 \mathbb{1}_{K_2}(y), \quad y \in K = K_1 \cup K_2. \tag{2.16}$$

Note that the Lambert's model (2.15) coincides with a cartoon projection of the albedo coefficient, $\rho(y(\theta, s))$, but weighted by the cosine $\theta^\perp \cdot \nu(\theta, s)$; the cartoon projection $\rho(y(\theta, s))$ is a piecewise constant function analogous to the cartoon projection of Figure 2.6(b), whereas the weight is a function depending on the geometry of the scene K.

Following the strategy of Subsection 2.4.3, we analyze the singularities of F, which results in a theorem which estimates the singularities of the reconstruction $\mathcal{R}^*F$. The approach is especially based on a comparison with the singularities of the Radon transform $\mathcal{R}\mathbb{1}_K$; it identifies artifacts due to the occlusions. We refer to Appendix A.A for preliminary results about $\mathcal{R}\mathbb{1}_K$.

Theorem 2.2. *Let K be the union of two Lambertian disjoint disks, with albedo (2.16). Let T_K denote the union of the four straight lines which are tangent to the two disks. Then, the reconstruction $\mathcal{R}^*F$ from the Lambert's cosine law F defined in (2.15) is such that*

$$\mathrm{WF}\,\mathbb{1}_K \subset \mathrm{WF}\,\mathcal{R}^*F \subset \mathrm{WF}\,\mathbb{1}_K \cup A_K, \tag{2.17}$$

where $\mathrm{WF}\,\mathbb{1}_K$ contains the circles in ∂K with their normal vectors,

$$\mathrm{WF}\,\mathbb{1}_K = \{(x; \hat{x}) \in \partial K \times \mathbb{R}^2 \setminus \{0\} : \hat{x} \text{ is a normal vector to } \partial K \text{ at } x \in \partial K\},$$

and A_K is defined by the lines in T_K and their normal vectors,

$$A_K := \{(x; \hat{x}) \in T_K \times \mathbb{R}^2 \setminus \{0\} : \{y : y \cdot \hat{x} = x \cdot \hat{x}\} \subset T_K\}.$$

In particular, any singularity of the initial geometry K is reconstructed, whereas A_K represents a set of possible artifacts corresponding to additional singularities located in T_K.

Proof. Fix two disjoint disks $K_i = \{x \in \mathbb{R}^2 : |x - z_i| \leq r_i\}$, with radius $r_i > 0$ and center $z_i \in \mathbb{R}^2$, $i = 1, 2$, such that $K = K_1 \cup K_2$. If the visible point is on K_i, *i.e.* $y(\theta, s) \in \partial K_i$, then the normal vector and the cosine of the angle of incidence are given by

$$\nu(\theta, s) = \tfrac{s - z_i \cdot \theta}{r_i}\theta + [1 - (\tfrac{s - z_i \cdot \theta}{r_i})^2]^{1/2}\theta^\perp, \quad \theta^\perp \cdot \nu(\theta, s) = [1 - (\tfrac{s - z_i \cdot \theta}{r_i})^2]^{1/2}.$$

Here, we recognize that the cosine coincides with a Radon transform of a disk (A.22); therefore, the Lambert's model (2.15) satisfies

$$F(\theta,s) = \left[1 - \mathbb{1}_{|s-z_2\cdot\theta|\leq r_2}\mathbb{1}_{(z_2-z_1)\cdot\theta^\perp\geq 0}\right]\frac{\rho_1}{2r_1}\mathcal{R}[\mathbb{1}_{K_1}](\theta,s) + \left[1 - \mathbb{1}_{|s-z_1\cdot\theta|\leq r_1}\mathbb{1}_{(z_1-z_2)\cdot\theta^\perp\geq 0}\right]\frac{\rho_2}{2r_2}\mathcal{R}[\mathbb{1}_{K_2}](\theta,s), \quad (2.18)$$

where the occlusion of K_i by K_j $(i \neq j)$ is encoded by

$$1 - \mathbb{1}_{|s-z_j\cdot\theta|\leq r_j}\mathbb{1}_{(z_j-z_i)\cdot\theta^\perp\geq 0} = \begin{cases} 0, & \text{if the disk } K_j \text{ is visible for the ray } (\theta,s), \\ 1, & \text{otherwise.} \end{cases}$$

Remark 2.3. The expression (2.18) reveals that F looks like the Radon transform $\mathcal{R}[\frac{\rho_1}{2r_1}\mathbb{1}_{K_1}+\frac{\rho_2}{2r_2}\mathbb{1}_{K_2}]$, but with modifications which takes into account occlusions. Therefore, this toy model is very closed to a problem of transmission tomography with incomplete data.

The function F is bounded with compact support, so $F \in L^1_{\mathrm{loc}}(\mathbb{S}^1\times\mathbb{R})$ and the backprojection $\mathcal{R}^*F$ is defined by (A.5),

$$\mathcal{R}^*F \in L^1_{\mathrm{loc}}(\mathbb{R}^2), \quad \mathcal{R}^*F(x) = \int_{\mathbb{S}^1} F(\theta, x\cdot\theta)\,\mathrm{d}\theta.$$

Without loss of generality, we rather analyze the backprojection of the even part

$$F'(\theta,s) = \tfrac{1}{2}[F(\theta,s) + F(-\theta,-s)] \in L^1_{\mathrm{loc}}(\mathbb{S}^1\times\mathbb{R}), \quad (2.19)$$

since it satisfies $\mathcal{R}^*F' = \mathcal{R}^*F$. By Theorem A.8 and Remark A.9, we can already claim that

$$\mathrm{WF}\,\mathcal{R}^*F = \mathrm{WF}\,\mathcal{R}^*F' = \{(s\theta + \tfrac{\hat\theta}{\hat s}\theta^\perp; \hat s\theta) : \hat s \neq 0 \text{ and } (\theta,s;\hat\theta,\hat s) \in \mathrm{WF}\,F'\}. \quad (2.20)$$

Therefore, we analyze $\mathrm{WF}\,F'$.

We deduce from (2.18) that

$$F'(\theta,s) = \left[1 - \tfrac{1}{2}\mathbb{1}_{|s-z_2\cdot\theta|\leq r_2}\right]\frac{\rho_1}{2r_1}\mathcal{R}[\mathbb{1}_{K_1}](\theta,s) + \left[1 - \tfrac{1}{2}\mathbb{1}_{|s-z_1\cdot\theta|\leq r_1}\right]\frac{\rho_2}{2r_2}\mathcal{R}[\mathbb{1}_{K_2}](\theta,s). \quad (2.21)$$

Here, the products by $1 - \frac{1}{2}\mathbb{1}_{|s-z_i\cdot\theta|\leq r_i}$, $i = 1,2$, are a consequence of the initial occultations. Such a factor has the same singularities than $\mathcal{R}[\mathbb{1}_{K_i}]$ given in (A.22); the corresponding wavefront set is given by (A.23). By Proposition A.12.(iii), the sum of products in (2.21) has a wavefront set such that

$$\mathrm{WF}\,F' \subset \mathrm{WF}\,\mathcal{R}\mathbb{1}_{K_1} \cup \mathrm{WF}\,\mathcal{R}\mathbb{1}_{K_2} \cup \left[(\mathrm{sing\,supp}\,\mathcal{R}\mathbb{1}_{K_1} \cap \mathrm{sing\,supp}\,\mathcal{R}\mathbb{1}_{K_2}) \times (\mathbb{R}^2\setminus\{0\})\right],$$

and we deduce from Proposition A.12.(i-ii) that

$$\mathrm{WF}\,F' \subset \mathrm{WF}\,\mathcal{R}\mathbb{1}_K \cup \{(\theta,s;\hat\theta,\hat s) \in (\mathbb{S}^1\times\mathbb{R})\times(\mathbb{R}^2\setminus\{0\}) : \{x\cdot\theta = s\} \subset T_K\}. \quad (2.22)$$

We prove now that

$$\mathrm{WF}\,\mathcal{R}\mathbb{1}_K \subset \mathrm{WF}\,F'. \quad (2.23)$$

Firstly,

$$\mathrm{WF}\,\mathcal{R}\mathbb{1}_K \setminus \left[(\mathrm{sing\,supp}\,\mathcal{R}\mathbb{1}_{K_1} \cap \mathrm{sing\,supp}\,\mathcal{R}\mathbb{1}_{K_2}) \times (\mathbb{R}^2\setminus\{0\})\right] \subset \mathrm{WF}\,F'. \quad (2.24)$$

In other words, the functions F' and $\mathcal{R}\mathbb{1}_{K_i}$ are singular in the same directions, except for a few points. To prove this result, fix $(\theta_0,s_0) \in \mathrm{sing\,supp}\,\mathcal{R}\mathbb{1}_{K_i} \setminus \mathrm{sing\,supp}\,\mathcal{R}\mathbb{1}_{K_j}$, *i.e.* $|s_0 - z_i\cdot\theta_0| = r_i$ and $|s_0 - z_j\cdot\theta_0| \neq r_j$, with $1 \leq i \neq j \leq 2$. Then, it can be seen that

$$\{(\hat\theta,\hat s) : (\theta_0,s_0;\hat\theta,\hat s) \in \mathrm{WF}\,F'\} = \{(\hat\theta,\hat s) : (\theta_0,s_0;\hat\theta,\hat s) \in \mathrm{WF}\,\mathcal{R}\mathbb{1}_{K_i}\}.$$

Indeed, if $|s_0 - z_j \cdot \theta_0| > r_j$, which is equivalent to $(\theta_0, s_0) \notin \operatorname{supp} \mathcal{R}\mathbb{1}_{K_j}$, then F' coincides with $\frac{\rho_i}{2r_i}\mathcal{R}\mathbb{1}_{K_i}$ in a neighborhood of (θ_0, s_0). On the contrary, if $|s_0 - z_j \cdot \theta_0| < r_j$, then (θ_0, s_0) is in the interior of $\operatorname{supp} \mathcal{R}\mathbb{1}_{K_j}$, but is not in $\operatorname{sing\,supp} \mathcal{R}\mathbb{1}_{K_j}$. In this case, $F'(\theta, s)$ coincides with $\frac{1}{2}\frac{\rho_i}{2r_i}\mathcal{R}[\mathbb{1}_{K_i}](\theta, s) + [1 - \frac{1}{2}\mathbb{1}_{|s - z_i\cdot\theta| \leq r_i}]\frac{\rho_j}{2r_j}\mathcal{R}[\mathbb{1}_{K_j}](\theta, s)$ in some neighborhood of (θ_0, s_0), which achieves the proof of (2.24). Secondly, a wavefront set is closed, so (2.23) is a consequence of (2.24).

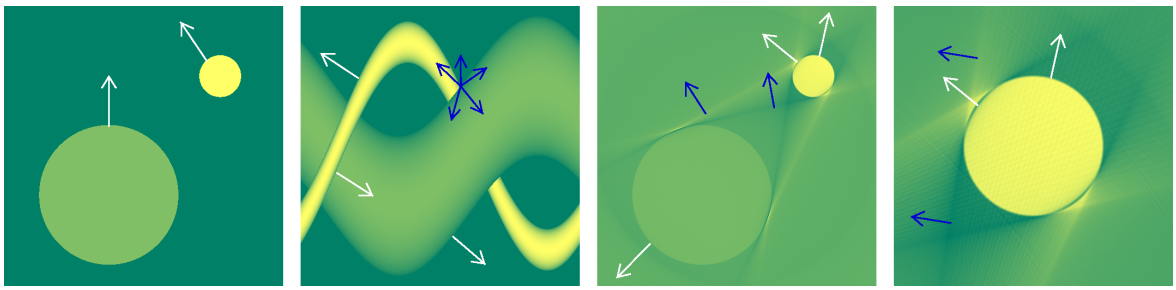
We conclude with (A.19), (2.23), (2.20), and (2.22),

$$\operatorname{WF} \mathbb{1}_K = \{(s\theta + \tfrac{\hat\theta}{\hat s}\theta^\perp; \hat s\theta) : \hat s \neq 0 \text{ and } (\theta, s; \hat\theta, \hat s) \in \operatorname{WF} \mathcal{R}\mathbb{1}_K\} \subset \operatorname{WF} \mathcal{R}^* F' \subset \operatorname{WF} \mathbb{1}_K \cup A_K. \quad \square$$

Note that the usual tomographic reconstruction $\mathcal{R}^*\Lambda F$ satisfies also

$$\operatorname{WF} \mathcal{R}^*\Lambda F \subset \operatorname{WF} \mathbb{1}_K \cup A_K, \quad \Lambda := \tfrac{1}{4\pi}\mathcal{H}_s\partial_s. \tag{2.25}$$

Indeed, $\mathcal{R}^*\Lambda F = \mathcal{R}^*\Lambda F'$, where F' is defined by (2.19); as in (A.21), we have the inclusion $\operatorname{WF} \mathcal{R}^*\Lambda F' \subset \operatorname{WF} \mathcal{R}^* F'$, and (2.25) follows from (2.20) and (2.17).



(a) (b) (c) (d)

Figure 2.8: Reflective tomography for two Lambertian disks K_1, K_2. (a) Albedo $\rho = \rho_1\mathbb{1}_{K_1} + \rho_2\mathbb{1}_{K_2}$. (b) Lambertian projection $F = \rho(y)\,\theta^\perp \cdot \nu$. (c) Tomographic reconstruction $\mathcal{R}^*\Lambda F$, $\Lambda = \frac{1}{4\pi}\mathcal{H}_s\partial_s$. (d) Zoom on $\mathcal{R}^*\Lambda F$. In any case, the white arrows are elements of a wavefront which is compatible with the representation ρ of the scene, whereas the blue arrows are additional singularities which introduce artifacts. Any singularity of the circles ∂K_1 and ∂K_2 is recovered, whereas artifacts appear on the four lines which are tangent to K_1 and K_2. These four lines are associated to eight corners in F. The observed results are in agreement with the theoretical inclusion (2.25) (and Theorem 2.2).

In Figure 2.8, we display an albedo function ρ such as (2.16), the associated Lambert's cosine law F defined in (2.15), and a tomographic reconstruction $\mathcal{R}^*\Lambda F$. As can be observed, and as expected, the reconstruction contains any singularity of the initial circles ∂K_1 and ∂K_2; it contains also elements of A_K, on the four lines which are tangent to the two circles.

2.4.5 Reflective tomography from cartoon images of two disks

It is instructive to consider the cartoon projection of ρ defined in (2.16). It is analogous to (2.15), but without the geometrical weight,

$$F(\theta, s) = \begin{cases} \rho(y(\theta, s)), & \text{if } \{x \cdot \theta = s\} \cap K \neq \emptyset, \\ 0, & \text{if } \{x \cdot \theta = s\} \cap K = \emptyset. \end{cases}$$

If $\rho_1 \neq \rho_2$, a proof similar with the proof of Theorem 2.2 shows that the estimation (2.17) is still valid. This is in agreement with the geometrical model of Figure 2.6. Note that here, Lambertian projections and cartoon projections result in the same kind of singularities.

On the contrary, if $\rho_1 = \rho_2$, the cartoon projection becomes a silhouette. This annihilates some part of the wavefront set. As a consequence, the inclusion $\operatorname{WF} \mathbb{1}_K \subset \mathcal{R}^* F$ is no longer valid and some part of the circles is lost in the reconstruction. This case is enlightening. Indeed, a standard approach

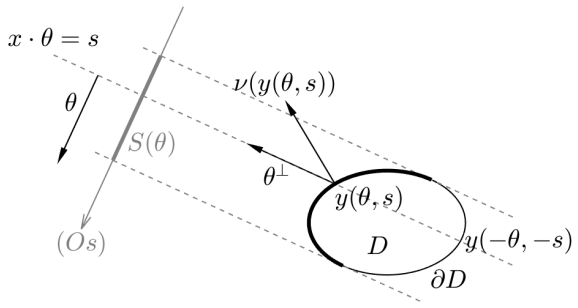


Figure 2.9: Geometrical setup corresponding to the Lambertian projection in Definition 2.4. For any angle $\theta \in \mathbb{S}^1$, the set $S(\theta)$ contains the values $s \in \mathbb{R}$ such that the line $x \cdot \theta = s$ intersects D. For any $s \in S(\theta)$, $\partial D \cap \{x \cdot \theta = s\} = \{y(\theta,s), y(-\theta,-s)\}$ with $\nu(y(\theta,s)) \cdot \theta^\perp > 0$. For $s \in \partial \overline{S(\theta)}$, the line $x \cdot \theta = s$ is tangent to ∂D at a unique $y(\theta,s)$, and $\nu(y(\theta,s)) \cdot \theta^\perp = 0$. For any $s \notin \overline{S(\theta)}$, $\bar{D} \cap \{x \cdot \theta = s\} = \emptyset$. The bold curve represents a set of visible points, $\{y(\theta,s),\ s \in \overline{S(\theta)}\}$.

in computer vision binarizes images in order to get silhouettes at a first step, and computes some backprojection to get a visual hull in a second step. On this test case, we see that a backprojection (or FBP) directly applied on Lambertian projections can reconstruct more relevant singularities than the visual hull.

We refer to [8] for more details about cartoon projections and silhouettes.

2.5 An exact Radon formula for a Lambertian reflector

2.5.1 Introduction

In this section, we summarize the article [3]. We restrict our attention to a Lambertian convex reflector, as depicted in Figure B.9. We consider purely diffuse reflection, modeled by the Lambert's cosine law (B.8), in a 2D setup. We exhibit an explicit Radon formula which relates the geometry $\partial D \subset \mathbb{R}^2$, the albedo coefficient $\rho : \partial D \to \mathbb{R}$, and values of the "radiance" $\rho(y)\cos\alpha$. This formula is based on the extension of the Radon transform that is described in Section A.4. In particular, it models diffuse reflection as the Radon transform of a distribution. This is to be contrasted with transmission tomography from Section B.3, where transmission is modeled by the Radon transform of a classical function (described in Section A.3).

2.5.2 Lambert's cosine law with convexity assumption

We specify the model of diffuse reflection that is considered. It is a parametrization of some Lambert's cosine law, under convexity assumption.

Unless otherwise stipulated, we assume in the reminder of this section:

(H1) $D \subset \mathbb{R}^2$ is a bounded open set with $\mathscr{C}^1$ boundary ∂D, such that the closure $\bar{D}$ is strictly convex, *i.e.* $\forall x, y \in \bar{D}, \forall t \in (0,1), tx + (1-t)y \in D$; $\nu(y) \in \mathbb{S}^1$ denotes the exterior unit normal vector to D at $y \in \partial D$, and μ denotes the length measure on ∂D.

(H2) $\rho \in L^\infty(\partial D)$ is a positive function, bounded and bounded away from zero, *i.e.* $\rho(\partial D) \subset [c, C]$ where c, C are two positive constants.

These assumption permit to project the object as in Figure 2.9, and to define a *Lambertian projection* parametrized as follows.

Definition 2.4. Assume (H1-H2). The *Lambertian projection* of (D, ρ), denoted by $\mathcal{L}[D,\rho]$, is a function defined on $\mathbb{S}^1 \times \mathbb{R}$ as follows. For every $(\theta, s) \in \mathbb{S}^1 \times \mathbb{R}$,

- if the line $x \cdot \theta = s$ does not intersect $\bar{D}$, then $\mathcal{L}[D,\rho](\theta,s) := 0$;

- if the line $x \cdot \theta = s$ intersects $\bar{D}$, the *visible point* $y(\theta, s)$ is defined as the unique point $y \in \partial D \cap \{x \cdot \theta = s\}$ such that $\nu(y) \cdot \theta^\perp \geq 0$, and

$$\mathcal{L}[D, \rho](\theta, s) := \rho(y(\theta, s))\, \nu(y(\theta, s)) \cdot \theta^\perp.$$

2.5.3 Diffuse reflection as the Radon transform of a distribution

The Lambertian projection is related to the Radon transform of a distribution as follows.

Theorem 2.5. *Assume (H1-H2). Let $\mathcal{L}[D, \rho]$ denote the Lambertian projection of Definition 2.4, and let $\frac{1}{\rho}\mathrm{d}\mu \in \mathcal{E}'(\mathbb{R}^2)$ be the Radon measure defined by $\left\langle \frac{1}{\rho}\mathrm{d}\mu, \psi \right\rangle = \int_{\partial D} \frac{\psi(y)}{\rho(y)}\mathrm{d}\mu(y)$, $\psi \in \mathcal{E}(\mathbb{R}^2)$. Then,*

$$\mathcal{R}\left[\frac{1}{\rho}\mathrm{d}\mu\right] = \frac{\mathbb{1}_{\mathcal{L}[D,\rho](\theta,s)>0}}{\mathcal{L}[D, \rho](\theta, s)} + \frac{\mathbb{1}_{\mathcal{L}[D,\rho](\theta,s)>0}}{\mathcal{L}[D, \rho](-\theta, -s)}, \tag{2.26}$$

where $\mathcal{R} : \mathcal{E}'(\mathbb{R}^2) \to \mathcal{E}'(\mathbb{S}^1 \times \mathbb{R})$ denotes the extended Radon transform (A.7), and the right member takes the value 0 if $\mathcal{L}[D, \rho](\theta, s) = 0$ (by convention).

Proof. The function $\frac{1}{\rho}$ is bounded on ∂D, so the Radon measure $\frac{1}{\rho}\mathrm{d}\mu \in \mathcal{E}'(\mathbb{R}^2)$ is defined. Then, the Radon transform of $\frac{1}{\rho}\mathrm{d}\mu$ is a distribution in $\mathcal{E}'(\mathbb{S}^1 \times \mathbb{R})$ defined by (A.7); it is such that

$$\left\langle \mathcal{R}\left[\frac{1}{\rho}\mathrm{d}\mu\right], \phi \right\rangle = \left\langle \frac{1}{\rho}\mathrm{d}\mu, \mathcal{R}^*\phi \right\rangle = \int_{\partial D} \frac{\mathcal{R}^*\phi(y)}{\rho(y)}\mathrm{d}\mu(y), \quad \phi \in \mathcal{E}(\mathbb{S}^1 \times \mathbb{R}).$$

Here, for any test function $\phi \in \mathcal{E}(\mathbb{S}^1 \times \mathbb{R})$, the smooth function $\mathcal{R}^*\phi \in \mathcal{E}(\mathbb{R}^2)$ is the backprojection defined in (A.5). By Fubini's theorem, we obtain that

$$\left\langle \mathcal{R}\left[\frac{1}{\rho}\mathrm{d}\mu\right], \phi \right\rangle = \int_{\mathbb{S}^1} \int_{\partial D} \frac{\phi(\theta, y \cdot \theta)}{\rho(y)}\mathrm{d}\mu(y)\mathrm{d}\theta.$$

In this expression, we can prove with [3, Lemma 3] that for any $\theta \in \mathbb{S}^1$, the inner integral satisfies

$$\int_{\partial D} \frac{\phi(\theta, y \cdot \theta)}{\rho(y)}\mathrm{d}\mu = \int_{S(\theta)} \left(\frac{\phi(\theta, s)}{\rho(y(\theta, s))\nu(y(\theta, s)) \cdot \theta^\perp} + \frac{\phi(\theta, s)}{\rho(y(-\theta, -s))\nu(y(-\theta, -s)) \cdot (-\theta)^\perp} \right) \mathrm{d}s,$$

where $S(\theta) = \{x \cdot \theta,\, x \in D\} \subset \mathbb{R}$ represents a projection of D to a line oriented by θ, as depicted in Figure 2.9. Therefore,

$$\left\langle \mathcal{R}\left[\frac{1}{\rho}\mathrm{d}\mu\right], \phi \right\rangle = \int_{\mathbb{S}^1} \int_{\mathbb{R}} \phi(\theta, s) \left(\frac{\mathbb{1}_{s \in S(\theta)}}{\mathcal{L}[D, \rho](\theta, s)} + \frac{\mathbb{1}_{s \in S(\theta)}}{\mathcal{L}[D, \rho](-\theta, -s)} \right) \mathrm{d}s\mathrm{d}\theta.$$

This computation shows that the compactly supported (non-negative) function

$$\frac{\mathbb{1}_{s \in S(\theta)}}{\mathcal{L}[D, \rho](\theta, s)} + \frac{\mathbb{1}_{s \in S(\theta)}}{\mathcal{L}[D, \rho](-\theta, -s)},$$

is in $L^1(\mathbb{S}^1 \times \mathbb{R})$ and coincides as a distribution with $\mathcal{R}\left[\frac{1}{\rho}\mathrm{d}\mu\right] \in \mathcal{E}'(\mathbb{S}^1 \times \mathbb{R})$. To conclude the proof, we remark that $\mathbb{1}_{s \in S(\theta)} = \mathbb{1}_{\mathcal{L}[D,\rho](\theta,s)>0}$. □

We can deduce an inversion formula which expresses some representation of the scene in term of its Lambertian projection. It is a filtered backprojection extended to distributions, as follows.

Corollary 2.6. *Assume (H1-H2). Then, the Radon measure $\frac{1}{\rho}\mathrm{d}\mu \in \mathcal{E}'(\mathbb{R}^2)$ is uniquely determined by the Lambertian projection $\mathcal{L}[D, \rho]$; moreover, it satisfies the inversion formula*

$$\frac{1}{\rho}\mathrm{d}\mu = 2\mathcal{R}^*\Lambda \frac{\mathbb{1}_{\mathcal{L}[D,\rho](\theta,s)>0}}{\mathcal{L}[D, \rho](\theta, s)}.$$

Here, the dual transform $\mathcal{R}^* : \mathcal{D}'(\mathbb{S}^1 \times \mathbb{R}) \to \mathcal{D}'(\mathbb{R}^2)$ is defined in (A.8), and the operator $\Lambda : \mathcal{E}'(\mathbb{S}^1 \times \mathbb{R}) \to \mathcal{D}'(\mathbb{S}^1 \times \mathbb{R})$ is the one from the extended inversion formula (A.12).

2.5.4 Discussion

Theorem 2.5 provides a way to extract a Radon transform from a Lambert's cosine law $\rho(y)\cos\alpha$, for a Lambertian convex reflector (D,ρ) in two dimensions. The formula (2.26) exhibits an appropriate pre-processing to obtain an element in the range of the transform $\mathcal{R}$ from the cosine law $\mathcal{L}[D,\rho]$: $\mathcal{L}[D,\rho]$ must be inverted, and must be made symmetrical in a second step. In this case, the relevant mathematical object to represent the scene is the Radon measure $\frac{\mathrm{d}\mu}{\rho}$. This object contains simultaneously the geometry and the physics of the problem. The support of this Radon measure is exactly the boundary ∂D of the reflector, while the density is directly the inverse of the albedo ρ.

2.6 Conclusion

We have enlightened various mathematical aspects of a principle dealing with the tomographic reconstruction from projections outside the range of the Radon transform, which includes reflective tomography. Intuitively, for discontinuous projections, such a method looks like a contour detection based on coherent contrasts; this is confirmed by asymptotic models. More rigorously, the microlocal analysis of the Radon transform provides a deep insight: the principle enters into the framework of imaging by the means of a Fourier Integral Operator. The singularities of the reconstruction correspond to singularities of the data. In reflective tomography, they are expected to contain partially the geometry of the initial scene, with artifacts resulting from the occultations; on some toy models, we have seen that artifacts appear along lines corresponding to corners in the dataset. Lastly, we have also found an exact Radon formula to support reflective tomography in a canonical framework dealing with a model of pure diffuse reflection in 2D.

Chapter 3

Unconventional algorithms in reflective tomography

3.1 Introduction

Reflective tomography deals with the reconstruction of a scene from VIS-NIR images, using algorithms from X-ray tomography. So far, we have especially encountered the FBP algorithm, or the FDK algorithm, in the case of a parallel scan, or a circular cone-beam scan. In general, the surfaces of the original scene are extracted from a reconstruction computed on a whole grid of pixels, or voxels. Two natural questions arise.

The first question concerns the computational efficiency. A whole reconstruction contains many useless voxels. There is a desire to compute only the voxels in a neighborhood of the wanted surfaces. A multiresolution algorithm dedicated to such a problem has been proposed in [1]; it takes benefit from the asymptotic models of Section 2.3. This algorithm, which is a lead to increase efficiency, is presented in the first part of this chapter.

The second question concerns the geometry of acquisition. We would like algorithms which tolerate general situations, where the calibration parameters of the camera can be arbitrary. This could, for instance, tolerate the merge of datasets from several distinct cameras. That is the reason why an algebraic technique based on the X-ray transform and Kaczmarz-type iterations has been proposed in [2]. This method is an unconventional way of tackling the problem of multi-view reconstruction; it is summarized and tested in the second part of the chapter.

3.2 Multiresolution greedy algorithm

3.2.1 Introduction

Assume that an imaging functional $\mathcal{J}_\Omega : \mathbb{R}^2 \to \mathbb{R}$ is given, where $\Omega > 0$ represents a resolution parameter, and assume that $\mathcal{J}_\Omega$ has the following asymptotic behavior, when $\Omega \to \infty$,

$$\mathcal{J}_\Omega(x) = \begin{cases} \mathcal{O}(1), & x \in S, \\ o(1), & x \notin S, \end{cases} \tag{3.1}$$

where $S \subset \mathbb{R}^2$ is a wanted set. Considering such an imaging functional is directly motivated by Section 2.3, where some asymptotic models of reflective tomography are discussed. In this case, $\mathcal{J}_\Omega(x) = \Omega^{-1/2}\mathcal{R}^*[F \star \psi_\Omega](x)$ is an adequate normalization of $\mathcal{I}_\Omega$ defined in (2.8), and S corresponds to a portion of surfaces to be reconstructed.

A natural way to estimate the set S is the following. The evaluation of $\mathcal{J}_\Omega$ on a regular cartesian grid, with step associated to the resolution Ω, provides a pixelized image. Due to (3.1), S is expected to appear under the form of bright pixels. Therefore, S is estimated by an extraction of the brightest

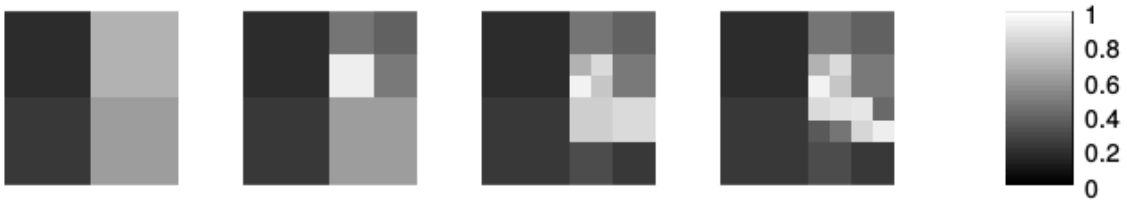


Figure 3.1: Principle of the multiresolution algorithm: compute an image with coarse pixels, then refine sets of bright pixels, iteratively. From left to right: initialization, iteration 1, 2, 3.

pixels. In this procedure, $\mathcal{J}_\Omega$ is computed on a whole grid, even if the wanted set S is very small. Therefore, limiting the computational effort for the pixels far from the unknown set S is a desire to increase efficiency.

This problem is tackled in [1] by a multiresolution algorithm based on the asymptotic behavior (3.1). In this section, we describe the principle of this multiresolution algorithm, and we propose numerical experiments on toy models. We refer to the original text [1] for more details.

3.2.2 Principle

We consider $(\mathcal{J}_\Omega, S)$ satisfying (3.1). The set S is unknown but we assume that we can evaluate $\mathcal{J}_\omega(x)$, for any resolution $0 < \omega \leq \Omega$, where the maximal resolution $\Omega > 0$ is fixed. In this case, we aim at computing an estimation S' of S, with fine pixels associated to the maximal resolution Ω. In order to clarify expectations, we fix an area A for the wanted estimation S' of S. To reach this goal without computing $\mathcal{J}_\Omega$ on a whole fine grid, we propose the algorithm described on this page, and illustrated in Figure 3.1. Basically, the method increases iteratively the resolution of bright pixels.

Multiresolution greedy algorithm, for $(\mathcal{J}_\Omega, S)$ satisfying (3.1).

Input. Area A for the wanted estimation S' of S, maximal resolution Ω, coarse resolution $\Omega_0 = 2^{-k}\Omega$.

Initialization.
(a) Evaluate $\mathcal{J}_{\Omega_0}$ on a coarse grid of pixels, associated to the resolution Ω_0.
(b) Estimate S by a set S' of area A, obtained by selection of the brightest pixels.

Iterations.
While the estimation S' contains pixels with resolution $< \Omega$,
(a) refinement: in S', double the resolution of any pixel with resolution $< \Omega$, *i.e.*, if $\omega < \Omega$ is the resolution of a pixel, replace the pixel by four sub-pixels computed with $\mathcal{J}_{2\omega}$;
(b) update the estimation S' of S: select any pixel with resolution Ω, and the brightest pixels with resolution $< \Omega$, so that the total area of S' is A.

Output. Estimation of S by a set S' of pixels with resolution Ω, and with total area A.

The initialization computes a coarse image $\mathcal{J}_{\Omega_0}(x)$, where x scans a coarse grid; the mesh size is associated to a given initial resolution $\Omega_0 = 2^{-k}\Omega$. Then we iteratively refine some pixels. Due to (3.1), the bright pixels are expected to belong to S, whereas the other pixels are expected to be noise. Therefore, we select a set of bright pixels. We divide them into four (sub-)pixels. For all of the new pixels, the imaging functional $\mathcal{J}_\omega$ is computed with an adequate value of ω: from a pixel to a sub-pixel, ω is doubled. Then we iterate. Obviously, this method is multiresolution, because the resolution parameter ω varies between Ω_0 and Ω.

Concerning the pixels to be refined, each iteration refines the brightest pixels, among the pixels

whose resolution is not maximal, and we constrain the total area: adding these pixels with the pixels at maximal resolution (already computed) must fill approximately an area A. In this way, the method is a greedy algorithm to compute a set S' of fine pixels, with prescribed area A, and whose accumulated intensity is maximal. See Figure 3.1, where A represents the area of the 12 finest pixels.

Concerning the area A to be filled, it is defined in practice as a percentage α of the area of a full reconstruction. Eventually, after convergence, it is possible to increase the value of A (or α) and to continue the iterations.

3.2.3 Comments

According to the asymptotic assumption (3.1), refining the brightest pixels should especially refine the $\mathcal{O}(1)$, so the method is expected to focus on S. Also, the pixels should have the same order of magnitude near S, even if they correspond to several values of ω. Furthermore, if a pixel is not closed to S but is selected for refinement (*false positive*), it is expected that the corresponding brightness decreases, due to the behavior in $o(1)$; somehow, refining a "noisy" pixel reduces the associated noise, which should avoid further refinement at the corresponding position.

Note that the method does not eliminate pixels during the process. The algorithm decides itself which zones must be refined, and it has a desired behavior for noise (false positive mentioned above). For any computed pixel, there may exist a future iteration which refines it. This has two advantages. This avoids eliminating prematurely pixels that should be preserved. Secondly, if the aimed area A corresponds to the whole fine grid, then the algorithm converges exactly to $\mathcal{J}_\Omega$ evaluated on the whole fine grid.

3.2.4 Numerical results for cartoon projections

We perform numerical experiments concerning the FBP from cartoon projections. Analogously to Figure 2.6, we consider the cartoon projection $F(\theta, s)$ of two circles, displayed in Figure 3.2. The projected values are $f_1 = 1$ (on the smallest circle) and $f_2 = 0.77$ (on the other one); the dataset contains 1609 images of size 512. The imaging functional is defined by a FBP

$$\mathcal{J}_\omega(x) = \omega^{-1/2}\mathcal{R}^*[F \star \psi_\omega](x), \quad \omega \leq \Omega, \tag{3.2}$$

where Ω corresponds to the Shannon-Nyquist frequency associated to the discretization in s; due to(2.11)-(2.14), $\mathcal{J}_\Omega$ is expected to satisfy (3.1). The corresponding multiresolution algorithm has been implemented in `Fortran`, and is executed on a workstation HP Z820, processors Intel Xeon E5-2609, 2.40 GHz.

Figure 3.2: Toy model for the multiresolution greedy algorithm. Left: scene containing two circles with values $f_1 = 1$ and $f_2 = 0.77$. Right: cartoon projection 1609×512 of the scene.

First, we execute the algorithm for several sizes of the initial grid. The area A is defined as $\alpha = 1\%$ of the area of a whole reconstruction. The computational times are reported in Table 3.1, and the reconstructions are displayed in Figure 3.3. The first main observation is that the multiresolution greedy algorithm succeeds in extracting the scene with an improvement of the computational time. The second one is that dividing by two or by four the full resolution for the initialization achieves a good compromise between speed and quality.

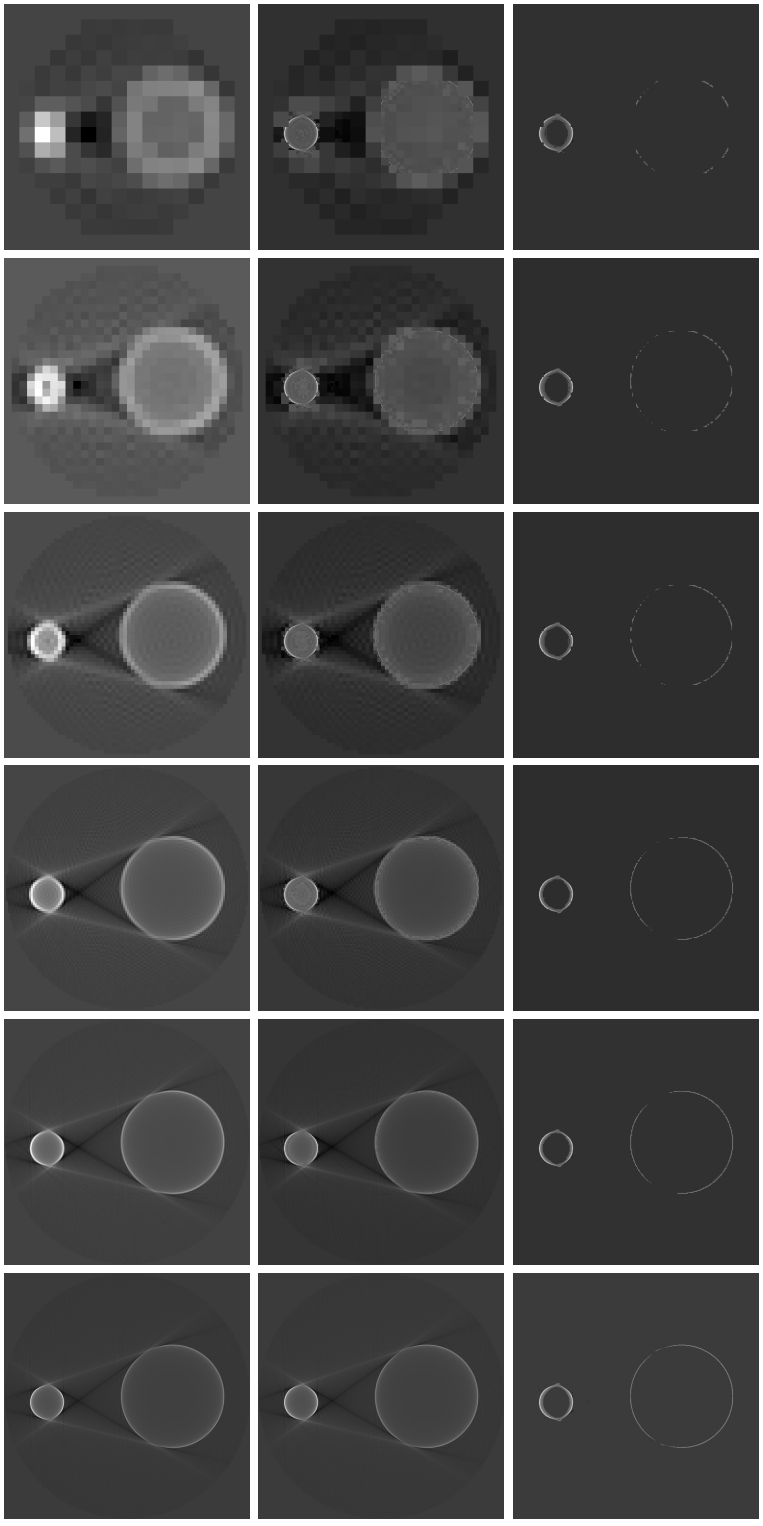


Figure 3.3: Multiresolution reconstructions from the cartoon projection displayed in Figure 3.2. From top to bottom, the size of the initial grid is 16×16, 32×32, 64×64, 128×128, 256×256, and 512×512 (full resolution). From left ro right: initial reconstruction, final multiresolution reconstruction, extracted thin pixels. The number of final pixels has been set to $\alpha \cdot 512^2$, $\alpha = 1\%$.

Initial grid	16×16	32×32	64×64	128×128	256×256	512×512
Time (s)	0.728	0.752	0.784	1.12	4.14	31.8

Table 3.1: Computational time in function of the initial grid, for the reconstructions of Figure 3.3.

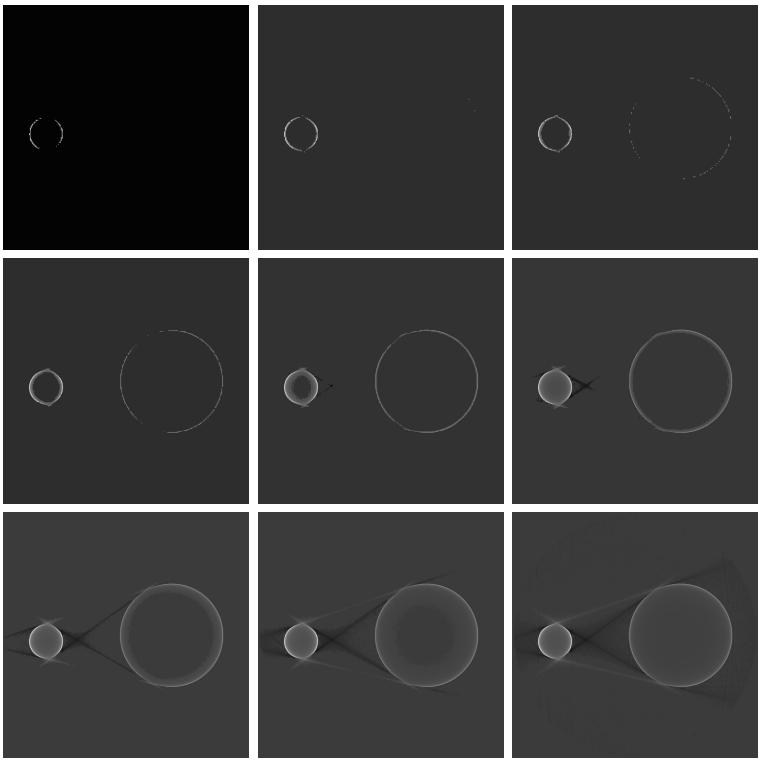


Figure 3.4: Reconstructions with $\alpha \cdot 512^2$ pixels from the cartoon projection of Figure 3.2, computed with the multiresolution greedy algorithm. The initial grid is fixed, with 128×128 pixels. From left to right, and top to bottom, the rate of pixels is $\alpha = 0.32 \cdot 2^\beta$, $-8 \leq \beta \leq 0$.

Second, we consider an initial size 128×128, and we execute the algorithm for various areas A, defined by various percentages $\alpha = 0.32 \cdot 2^\beta$ of a total area. The extracted reconstructions are displayed in Figure 3.4, and the computational times are reported in Table 3.2. As expected with (2.11)-(2.14), the multiresolution process extracts in priority the circles, then the artifacts, and then the noise. Here, the values $\alpha = 1\%, 2\%$, realize a good compromise to get efficiently a complete reconstruction of the circles without artifacts; the reconstruction has missing parts for smaller values of α, and captures artifacts for larger values. Also, the computational time significantly increases for larger values of α, but is not really reduced for smaller values (some costs are incompressible).

Rate of pixels α (%)	0.125	0.25	0.5	1	2	4	8	16	32
Time (s)	0.676	0.736	0.860	1.12	1.61	2.57	4.62	8.41	16.2

Table 3.2: Computational time in function of the rate α, for the reconstructions of Figure 3.4.

3.2.5 Numerical test of a 3D extension

The multiresolution greedy algorithm based on the FBP (3.2) has been extended in 3D for some orthographic[1] scan [1]. Here, we present some numerical reconstructions from noisy Gouraud images.

The dataset is obtained as follows, in `Matlab`. Analogously to Subsection 1.8.1 and Figure 1.9, we start with a series F of 805 noisy images 512×512 of the Stanford Bunny; on these images, the surface contains a pattern projected with the Gouraud model of `Matlab`. Next, exactly as in Subsection 1.8.4, we add a speckle noise (1.6), with $\sigma = 1$ for the magnitude of the noise. A few images of the resulting sequence are displayed in Figure 3.5. Reconstructing the surface with such a level of noise is quite challenging.

We compute two reconstructions from the noisy dataset, using a `Fortran` implementation of the 3D multiresolution greedy algorithm, on a workstation HP Z820, processors Intel Xeon E5-2609, 2.40 GHz. First, we compute a reference reconstruction on a 3D grid of $512 \times 512 \times 512$ voxels, using one FBP per horizontal cross-section. Then we extract $1‰512^3$ voxels (the brigthest ones). Second, we use the multiresolution greedy algorithm, even if, strictly speaking, the asymptotic behavior of $\mathcal{J}_\Omega$ has not been studied for the Gouraud model. The size of the initial grid is $64 \times 64 \times 512$ (the vertical step is not reduced), and the wanted volume corresponds to the volume of $1‰512^3$ voxels.

In both cases, three orthogonal MIPs of the obtained voxels are displayed in Figure 3.6. The reconstruction of the greedy method looks less diffuse. In fact, the multiresolution method starts from a regularized reconstruction, since it contains a low-pass filter related to the initial resolution; therefore, the final reconstruction is computed as a refinement of a "cleaned" reconstruction. Concerning the computational time, 8760 seconds for the reference versus 237 seconds for the greedy method; the ratio of time is about 37.

3.3 Flexible algebraic reconstruction

3.3.1 Introduction

This section deals with tomography for 3D multi-view reconstruction from 2D images in VIS-NIR optics. The images are assumed to be calibrated: the intrinsic and extrinsic matrices of the cameras, in (B.1), are assumed to be known, but they can be arbitrary. This last property is the main difference with the experiments considered so far, where the geometry of acquisition was assumed to be a parallel scan, or a circular cone beam scan (after an eventual pre-processing [32]). As a result, it is not possible anymore to use an analytical formula such as a FBP or FDK formula.

In X-ray tomography, projections for a general geometry of acquisition can be inverted by an Algebraic Reconstruction Technique (ART) with Kaczmarz iterations, as described in Section A.7. Such an approach, based on linear algebra, is flexible and tolerates any setting for the rays of projection; the main requirement is the knowledge of the rays. Therefore, the principle of reflective tomography suggests to use such a solver for calibrated VIS-NIR images, since the rays are known in this case. This has been achieved in [2], with a Kaczmarz-type method based on the X-ray transform. In this section, we present the principle of this method, and we apply this principle on real photographs. We refer to the original text [2] for a more comprehensive study and tests on CCD images extracted from the Middlebury datasets [94, 95].

3.3.2 Algebraic iterative solver

Consider a collection of N calibrated 2D images, b_j, $1 \leq j \leq N$. For any image b_j, the pixels values are arranged in a vector, so that $b_j \in \mathbb{R}^{m_j}$, where m_j denotes the total number of pixels; for any pixel $1 \leq p \leq m_j$, the associated ray of projection is a known straight line, denoted by L_j^p.

[1] The considered orthographic scan, with parallel rays, is a simplification of a circular cone-beam scan for a camera in far field.

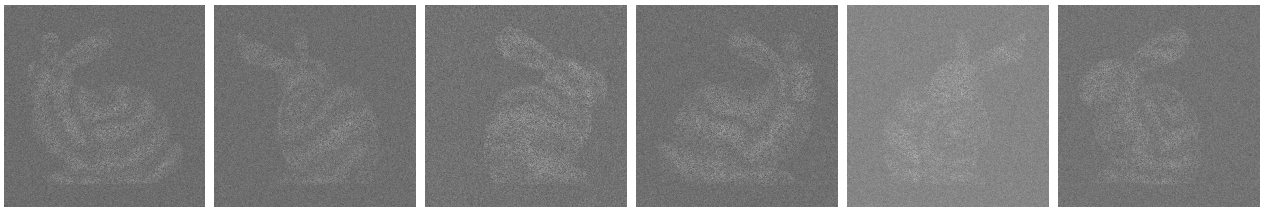


Figure 3.5: Samples of a sequence of 805 Gouraud images 512×512, with a speckle noise.

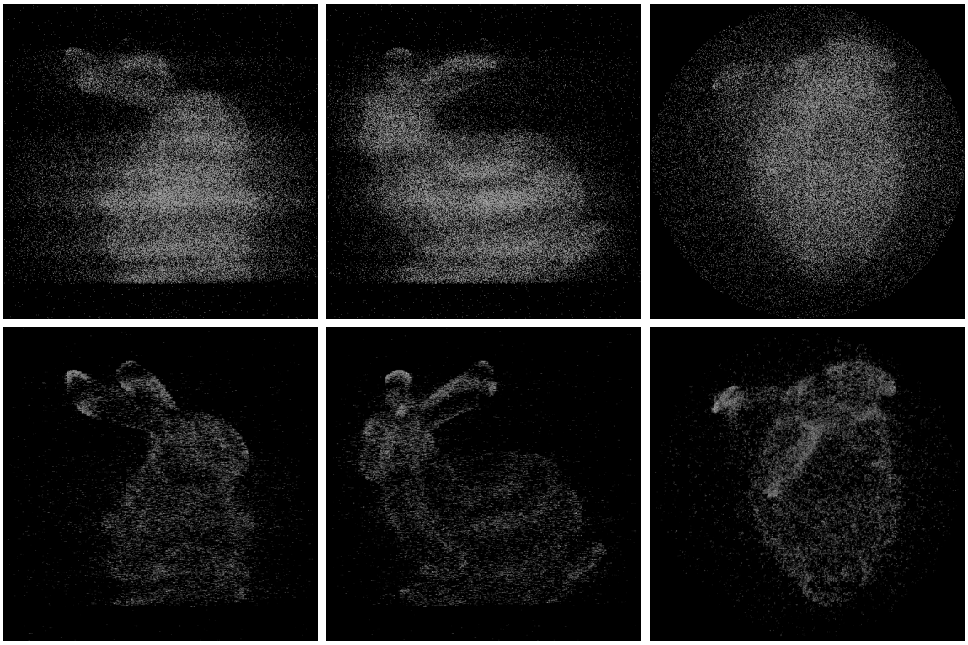


Figure 3.6: Reference method (top) versus multiresolution greedy algorithm (bottom) for 3D reconstruction. The display is a MIP of $1‰512^3$ voxels reconstructed from 805 noisy Gouraud images 512×512 (Figure 3.5). Top: voxels extracted from a FBP $512 \times 512 \times 512$, in 8760 s. Bottom: voxels computed by a 3D multiresolution greedy algorithm with initial grid $64 \times 64 \times 512$, in 273 s.

We aim at computing some 3D reconstruction of the scene which appears on the images b_j. Therefore, we introduce a 3D grid of n voxels, and we look for a reconstruction defined on this grid. Equivalently, we introduce basis functions ϕ_i, $1 \leq i \leq n$, associated to the voxels, so that ϕ_i denotes the characteristic function of the voxel numbered i; we look for a reconstruction $\phi \in \mathrm{span}\{\phi_i, 1 \leq i \leq n\}$. We denote by $x = (x_i)_{1 \leq i \leq n} \in \mathbb{R}^n$ the values of ϕ on the voxels, *i.e.* $\phi = \sum_{i=1}^n x_i \phi_i$.

Following the principle of reflective tomography, we propose to model the images b_j, $1 \leq j \leq N$, by the means of the X-ray transform defined in (B.11). For that purpose, for any $1 \leq j \leq N$, we define $A_j \in \mathbb{R}^{m_j \times n}$ as the matrix of the X-ray transform, in the basis $\phi_i, 1 \leq i \leq n$, and for the rays L_j^p associated to the image b_j,

$$A_j := \left[\int_{L_j^p} \phi_i \, \mathrm{d}\ell \right]_{1 \leq p \leq m_j, 1 \leq i \leq n};$$

the element at position (p, i) of A_j corresponds to the length (in `[m]`) of the intersection of the voxel i with the ray associated to the pixel p of the image b_j. Ideally, we would like to find a reconstruction $\phi = \sum_i x_i \phi_i$ such that

$$A_j x = b_j, \quad 1 \leq j \leq N,$$

which would mean that each image b_j is a cone beam radiography of ϕ. Note that here, there is no guarantee that this linear system is compatible; in particular, we rather look for some generalized solution.

We recognize a problem decomposed into block of rows as in (A.17), which suggests to use Kaczmarz iterations described on page 68. Also, for safety reasons, we slightly regularize the recurrence relation (A.18). Therefore, we define one cycle of iterations by

$$\begin{cases} x^{(0)} \in \mathbb{R}^n, \\ x^{(j)} := x^{(j-1)} + \omega A_j^* (A_j A_j^* + \sigma \mathrm{I})^{-1} (b_j - A_j x^{(j-1)}), \quad 1 \leq j \leq N; \end{cases} \tag{3.3}$$

with $\sigma > 0$ and $\omega > 0$. Then, the iterate $x^{(N)}$ is used to initialize a second cycle of iterations ($x^{(0)} := x^{(N)}$). And so on.

The recurrence relation can be explained as follows. At step j, $x^{(j-1)}$ represents a 3D model of the scene, deduced from $x^{(0)}$ and the calibrated images b_i, $i \leq j-1$. We update this model using a constraint based on the calibrated image b_j. In the case $\omega = 1$, the new model $x^{(j)}$ is defined as the unique solution of

$$\min_{x \in \mathbb{R}^n} \|A_j x - b_j\|^2 + \sigma \|x - x^{(j-1)}\|^2; \tag{3.4}$$

therefore, the X-ray transform of $\phi^{(j)} = \sum_i x_i^{(j)} \phi_i$ must reproduce the data b_j (as much as possible), with $x^{(j)}$ close to the previous model $x^{(j-1)}$. If $\omega \neq 1$, $x^{(j)}$ is defined as a barycenter between the solution of (3.4) and $x^{(j-1)}$.

In practice, the images b_j are arranged in a "randomized" order. We start (3.3) with $x^{(0)} = 0$. For any iteration j, the matrices A_j and A_j^* are not stored; we rather evaluate products matrix-vector, by ray tracing. The vectors $(A_j A_j^* + \sigma \mathrm{I})^{-1}(b_j - A_j x^{(j-1)})$ are computed approximately, with a few iterations of the conjugate gradient method. Usually, a few cycles of iteration are enough to get satisfactory results; numerical convergence can be checked by a control of the Root Mean Square Error,

$$\left(\frac{1}{\sum_{j=1}^N m_j} \sum_{j=1}^N \|A_j x - b_j\|^2 \right)^{1/2} \quad \texttt{[pixel intensity]}, \tag{RMSE}$$

at the end of each cycle of iteration.

3.3.3 Home-made scanner: reconstruction from passive digital photographs

We realize an experiment from scratch to test the algebraic technique (3.3) on our own images.

Acquisition

A face has been photographed without flash from 72 angles of view. Each photograph is a RGB image 384×480; see Figure 3.7 for an image of the sequence. This acquisition enters in the framework of passive imagery. The photographer is between the light sources and the scene, which explains why some shadows appear on the images. This increases the difficulty of the reconstruction task: as observed in the initial samples of Figure 3.8, the brightness of a portion of the face strongly varies from an image to another one.

Calibration

For calibration purpose, the scene is equipped with a calibration pattern (a checkerboard with squares of 8.45 size `[mm]`). This pattern is used to estimate the extrinsic and intrinsic parameters of the camera with the `Camera Calibrator App` in `Matlab`, in a reference frame; the distorsions are estimated simultaneously, then corrected. Finally, the original RGB images are converted into grayscale undistorted calibrated images (values in [0,255]), and the calibration pattern is removed by means of a mask. See Figure 3.7; a pre-preprocessed image is displayed, so are the estimated locations/orientations of the camera.

Implementation

The full algorithm has been sequentially implemented in `Fortran 2003`, in double precision. It includes the frame-driven Kaczmarz iterations (3.3), combined with the conjugate gradient. It also includes ray tracing on a grid of voxels, for the computation of X-ray images, for the backprojection, and for MIP rendering. We measure the total time dedicated to the computation of the reconstruction, which includes the initialization, the iterative updates of the model, iterative loading of the images, and evaluations of RMSEs.

Reconstruction

We define a box which roughly estimates the face (see Figure 3.7), and we compute a 3D reconstruction inside this box, from the grayscale calibrated images. In the reference frame, the box is $[-120, 170] \times [-300, -30] \times [-180, 100]$ `[mm]`, and is decomposed into voxels of size $h = 1.5$ `[mm]`; hence the reconstruction contains about 186^3 voxels.

We perform one cycle of iterations (3.3) with $x^{(0)} = 0$, $\omega = 0.5$, and $\sigma = 5dh$ `[m²]`, where d `[m]` is the diagonal of the box. On a laptop Dell Precision M4400, processor Intel Core 2 Duo T9600, 2.80 GHz, the `Fortran` code mentioned above returns the reconstruction in 1220 seconds. The initial RMSE, with $x = x^{(0)}$ is equal to 67.2; the final RMSE, with $x = x^{(72)}$, is equal to 33.8. In Figure 3.8, six initial calibrated views are visually compared with MIP re-projections of the reconstruction $x^{(72)}$; here, the reconstruction has been restricted to the box $[-110, 160] \times [-300, -30] \times [-170, 100]$ `[mm]`, and a lower threshold fixed to 700 has been applied. In Figure 3.9, we display a circular cone-beam scan of the reconstruction $x^{(72)}$; here again, thresholded MIPs of a box are used. As observed, the method succeeds in catching automatically some features and details.

Cross validation

The algebraic method (3.3) can be understood as online machine learning from calibrated images in optics [2]. This suggests to evaluate the quality of a reconstruction by means of a generalization error [25]. That is the reason why we perform a four-fold cross-validation, as follows.

The dataset with 72 images is randomly divided into four subsets with 18 images. We select one of these subsets. We consider it as a *test* set, while the three other ones define a *training* set. We compute a reconstruction exactly as before, but from the training set only (54 images). Then, we evaluate the quality of the reconstruction with a RMSE computed on the training set,

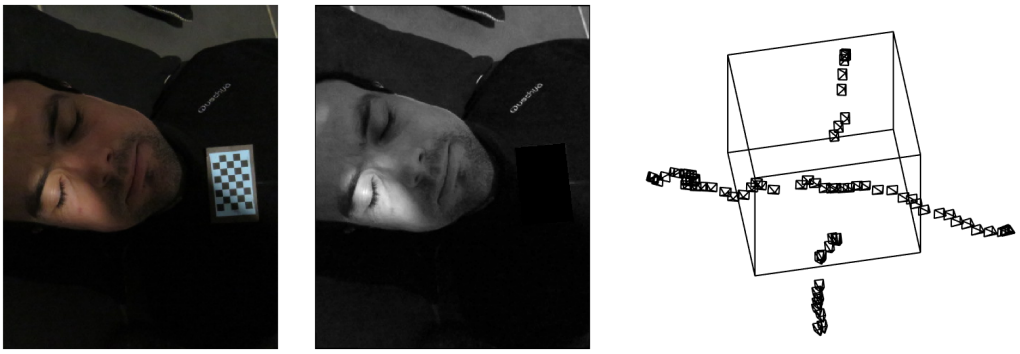


Figure 3.7: Setup of an acquisition to test the algebraic technique (3.3). A face, equipped with a calibration pattern, is photographed from several angles of view without flash; 72 RGB images of size 384×480 are captured. The sequence is calibrated with `Camera Calibrator App` in `Matlab`. Left: one recorded image. Middle: this image is converted into grayscale, after correction of distorsion; the calibration pattern is masked. Right: the position/orientation of any calibrated camera is represented by a cone, the box is a rough estimation of the face.

Figure 3.8: Reconstruction with the algebraic technique (3.3), from 72 calibrated images 384×480 (see Figure 3.7 for the acquisition). Top: six images of the input sequence. Bottom: MIP "re-projection". The computation is performed on voxels of size $h = 1.5$ `[mm]`.

Figure 3.9: Reconstruction with the algebraic technique (3.3), from 72 calibrated images 384×480 (see Figure 3.7 for the acquisition). Here, a circular cone beam scan of an algebraic reconstruction (voxels of size $h = 1.5$ `[mm]`) "predicts" twelve novel views; the rendering is a MIP.

Test number	1	2	3	4
Time [s]	909	912	951	905
RMSE [pixel intensity] on the training set (54 images)	34.0	34.3	33.9	34.4
RMSE [pixel intensity] on the test set (18 images)	35.9	34.5	36.3	35.5

Table 3.3: Four-fold cross validation of one cycle of algebraic iterations (3.3), from 72 calibrated images 384×480 (see Figure 3.7 for the acquisition).

and a RMSE computed on the test dataset. Lastly, we repeat this procedure four times: the test set browses the four subsets selected at the beginning. The RMSEs on the test sets correspond to some generalization error of the method; the variation of the RMSEs indicates the sensitivity of the quality with respect to the training set. Numerical results are summarized in Table 3.3.

Furthermore, we have selected one image per test set. For each trained reconstruction, we compute a MIP view associated to each one of these images. In this way, the MIP view of the i'th reconstruction, associated to the image of the i'th test set, is a prediction; the other views are re-projections. Here, the full box is projected, with a lower threshold fixed to 500. See Figure 3.10.

The main conclusion here is that the RMSE does not strongly depend on the training set.

Figure 3.10: Four-fold cross validation of one cycle of algebraic iterations (3.3), from 72 calibrated images 384×480 (see Figure 3.7 for the acquisition). A line $1 \leq i \leq 4$ contains MIP views of the i'th reconstruction, trained with 54 images; the views are predictions on the diagonal, re-projections otherwise. The last line contains initial images for visual comparison.

Chapter 4

Conclusion and perspectives

4.1 Synthesis of the results

Three-dimensional (3D) reflective tomography reconstructs a motionless 3D scene, from several calibrated bi-dimensional (2D) optical images injected into a Radon-kind inversion. For N calibrated images, denoted by $b_j \in \mathbb{R}^{m_j}$, $1 \leq j \leq N$, the reconstruction on a grid of n voxels, denoted by $x \in \mathbb{R}^n$, is computed such that

$$A_j x \approx b_j, \quad 1 \leq j \leq N, \tag{4.1}$$

where the matrix $A_j \in \mathbb{R}^{m_j \times n}$ represents the X-ray transform along the m_j rays associated to the m_j pixels of the image b_j. The reconstruction is then rendered by suitable 3D visualization methods.

At a first sight, reflective tomography is an empirical method, because the data $[b_j]_{1\leq j\leq N} \in \mathbb{R}^{m_1+\cdots+m_N}$ is not in the range of the transform $[A_j]_{1\leq j\leq N} \in \mathbb{R}^{(m_1+\cdots+m_N)\times n}$. But the principle has been further motivated by mathematical arguments. We have proved how the Radon transform extended to distributions models pure diffuse reflection from a Lambert's cosine law on a 2D convex reflector. And more generally, the microlocal analysis of the Radon transform reveals that the singularities of the reconstruction may be expected to be relevant (up to artifacts).

Concerning the practice, a variety of numerical experiments, on real or synthetic images, attests to the relevance of the approach. It reveals that the method automatically captures the geometry of the initial scene, even if the scene has occlusions, or if the dataset is corrupted by noise. Various solvers have been tested; they include solvers based on analytical inversion formulas such as the FDK algorithm implemented on a Graphics Processing Unit, and an iterative algebraic method based on a block Kaczmarz's method. Various rendering methods have been investigated, including the Maximum Intensity Projection (MIP), and the extraction of point clouds.

The rest of the chapter describes some perspectives.

4.2 Learning dynamical geometry in vision

4.2.1 Reflective tomography and artificial intelligence

In a modern langage, reflective tomography "learns" the geometry of a 3D scene from multiple-view projections in optics. The topic is related to fields of active research such as machine learning and vision, which suggests going further.

Reflective tomography such as (4.1) proceeds as follows. A reconstruction step captures relevant singularities, by means of a Fourier Integral Operator that is tuned for a projection geometry. The reconstruction is rendered by the MIP, so that the global procedure "predicts" novel views of the initial scene. In this case, the MIP appears to compress the 3D reconstruction onto suitable 2D images and point clouds. This approach has some similarities with neural networks, since it injects some "non-physical" model into a non-linear compression method for prediction purposes. This suggests "optimizing" the reconstruction algorithm, or the matrices A_j, in order to improve the

final rendering (if possible). One may also wonder what learning approach could be inspired, or initialized, by reflective tomography.

Such a topic is related to recent works on imaging and machine learning. Indeed, digital wavefront sets and limited-angle tomography based on deep learning have been introduced in [23, 24], whereas the MIP has been introduced in deep learning algorithms for imaging in [103, 104]. Combining or tuning such approaches could be a first step to mimic reflective tomography in deep learning.

4.2.2 Image calibration for multiple-view reconstruction

Camera calibration is an important subject in multiple-view geometry: the location and the orientation of the used cameras, and more generally the rays of projection, must be suitably approximated. For instance, the matrix A_j in (4.1) is based on the rays corresponding to the image b_j. Classical methods based on correspondences between features are available [57, 73], whereas modern approaches based on deep learning are still being developed [27]. In this thesis, camera calibration has been almost skiped. So, a natural question concerns the use, or the design, of a state-of-the-art method, in the specific framework of reflective tomography.

4.2.3 Four-dimensional reconstruction of a dynamical 3D scene in optics

In practical applications, a scene to be imaged is often dynamical. This includes the case of deformable materials, such as a patient who is breathing in the field of medical imaging; this also includes moving rigid solids, such as a moving car in road safety. Furthermore, the subject meets the problem of image calibration: in a coordinate system attached to a moving camera, the location and the orientation of the camera become known and fixed, while the scene appears as a moving one.

Taking into account the motion is a current hot subject; we refer for instance to [66] for a recent book about various topics in time-dependant inverse problems, and to [80] for motion correction in Magnetic Resonance Imaging. This suggests studying reflective tomography in the case of a dynamical scene, in order to recover the geometry and the motion. In this case, the problem (4.1) becomes

$$A_j(t)\, x(t) = b_j(t), \quad 1 \leq j \leq N, \tag{4.2}$$

where $b_j(t)$ is the image of the camera j at time t, $A_j(t)$ is the matrix of the associated X-ray transform, and the unknown $x(t)$ represents the scene a time t. Note that the case $N = 1$ deals with a single moving camera that observes a dynamical scene.

A "naive" approach consists in recording videos of the scene, with several cameras which are fixed, calibrated and synchronized; in this case, a 3D reconstruction can be computed at every time step. To go further, the redundancies and the differences between the several time steps must be taken into account. Finally, the ultimate goal (or dream) would consist in considering moving cameras which are neither calibrated nor synchronized (in this case, t and $A_j(t)$ are somehow unknown).

4.3 New Radon-kind transforms in radiometry

Over the past decade, there has been a considerable interest in developing new imaging modalities based on scattering of light and extensions of the Radon transform, such as Compton scattering tomography [79, 90] and Bragg scattering tomography [105]. Reflective tomography for a Lambertian reflector, as described by Theorem 2.5, enters in this framework, since it considers the Radon transform extended to distributions for modeling diffuse reflection. This result potentially opens new perspectives concerning optical tomographic imaging in a more general context.

A natural question concerns the extension to more general models in radiometry. For instance, studying the following question is of particular interest: given a set of 2D optical images of a 3D scene with occlusions, is there a distribution supported by surfaces, and some Radon-like transform (or

X-ray transform), which can model some transformation of the optical images? This topic is related to image formation in vision; it has its own mathematical interest, and may open new practical applications in vision, in optics, or in any field dealing with radiation. This could potentially lead to novel reconstruction algorithms of physical properties such as the albedo.

4.4 Maximum Intensity Projection in Convolutional Neural Networks

The MIP appears to be an efficient way of compressing some 3D volumes; this has motivated the patent [18], and recent deep learning methods [103, 104]. Some parallel can be drawn with *max-pooling* in Convolutional Neural Networks (in CNN), which downsamples the input by selection of the maximum value over patches. The MIP may be an alternative or complementary way for downsampling/compressing. Further investigations may be needed to develop extensively such a principle.

Appendix A

Radon transform

A.1 Introduction

The Radon transform emerged at the beginning of the XXth century in pioneering works of Lorentz, Funk and Radon [43], and has been extensively studied since the advent of computed tomography in the 1970s. By definition, the Radon transform $\mathcal{R}$ integrates a reasonable function f over straight lines in a plane,

$$\mathcal{R}f(\theta, s) = \int_{x\cdot\theta=s} f(x)\,\mathrm{d}\ell = \int_{\mathbb{R}} f(s\theta + t\theta^{\perp})\,\mathrm{d}t, \quad \theta \in \mathbb{S}^1,\ s \in \mathbb{R}. \tag{A.1}$$

Here, ℓ denotes the length measure on the line of integration $\{x \in \mathbb{R}^2 : x \cdot \theta = s\}$; this line is orthogonal to the vector $\theta \in \mathbb{S}^1$, with signed distance $s \in \mathbb{R}$ from the origin, and oriented by $\theta^{\perp} \in \mathbb{S}^1$.

In this chapter, we summarize some classical results about the Radon transform (A.1), since they lay the foundation for the works presented in this thesis. We present the Radon transform as a continuous linear map on various classes of functions, and we extend it on distributions with compact support. We recall an exact inversion formula, and its practical implementation by the filtered backprojection algorithm. We describe the principle of the algebraic reconstruction technique, which has the advantage of flexibility. Then, we mention microlocal properties, including correspondences of singularities between a distribution and its Radon transform. To finish with, we apply such properties in the specific case of characteristic functions supported by disks.

A.2 Notation

The Fourier transform on $\mathbb{R}^n$, with $n = 1, 2$ is denoted by $\mathcal{F}_n$, its inverse is denoted by $\mathcal{F}_n^{-1}$; they are normalized as follows:

$$\mathcal{F}_1 g(\sigma) = \int_{\mathbb{R}} g(s)\mathrm{e}^{-\mathrm{i}\sigma s}\,\mathrm{d}s, \qquad \mathcal{F}_1^{-1}\hat{g}(s) = \frac{1}{2\pi}\int_{\mathbb{R}} \hat{g}(\sigma)\mathrm{e}^{\mathrm{i}\sigma s}\,\mathrm{d}\sigma, \qquad g,\, \hat{g} \in L^1(\mathbb{R}), \tag{A.2}$$

$$\mathcal{F}_2 f(\xi) = \int_{\mathbb{R}^2} f(x)\mathrm{e}^{-\mathrm{i}x\cdot\xi}\,\mathrm{d}\xi, \qquad \mathcal{F}_2^{-1}\hat{f}(x) = \frac{1}{4\pi^2}\int_{\mathbb{R}^2} \hat{f}(\xi)\mathrm{e}^{\mathrm{i}x\cdot\xi}\,\mathrm{d}\xi, \qquad f,\, \hat{f} \in L^1(\mathbb{R}^2). \tag{A.3}$$

The notation for test function spaces and their dual is usual: $\mathcal{D}$ for $\mathscr{C}^\infty$ functions with compact support, $\mathcal{S}$ for the Schwartz space of rapidly decreasing $\mathscr{C}^\infty$ functions, $\mathcal{E}$ for $\mathscr{C}^\infty$ functions, $\mathcal{E}'$ for distributions with compact support, $\mathcal{S}'$ for tempered distributions, $\mathcal{D}'$ for distributions.

For any $\theta = (\theta_1, \theta_2) \in \mathbb{S}^1$, we fix $\theta^{\perp} := (\theta_2, -\theta_1) \in \mathbb{S}^1$.

A.3 Radon transform for functions

Several fundamental properties can be deduced from Fubini's theorem. In particular, the Radon transform $\mathcal{R}$ in (A.1) defines a continuous linear map $\mathcal{R} : L^1(\mathbb{R}^2) \to L^1(\mathbb{S}^1 \times \mathbb{R})$ [69, Theorem 8]. This

map is closely related to Fourier transforms, as stated by the Fourier slice theorem [69, Theorem 2].

Theorem A.1 (Fourier slice theorem). *Let $f \in L^1(\mathbb{R}^2)$, and $\theta \in \mathbb{S}^1$. Then, the Fourier transform of $\mathcal{R}f(\theta,\cdot)$ coincides with the Fourier transform of f, along the radial line $\sigma\theta$, $\sigma \in \mathbb{R}$,* i.e.

$$\mathcal{F}_1[\mathcal{R}f(\theta,\cdot)](\sigma) = \int_{\mathbb{R}} \mathcal{R}f(\theta,s)\mathrm{e}^{-\mathrm{i}s\sigma}\mathrm{d}s = \int_{\mathbb{R}^2} f(x)\mathrm{e}^{-\mathrm{i}\sigma\theta\cdot x}\mathrm{d}x = \mathcal{F}_2[f](\sigma\theta), \quad \sigma \in \mathbb{R}.$$

Proof. Fix $\sigma \in \mathbb{R}$. By definition, $\mathcal{F}_2[f](\sigma\theta) = \int_{\mathbb{R}^2} f(x)\mathrm{e}^{-\mathrm{i}\sigma\theta\cdot x}\mathrm{d}x$. Therefore, by Fubini's theorem, with $x = s\theta + t\theta^\perp$, $\mathcal{F}_2[f](\sigma\theta) = \int_{\mathbb{R}}\int_{\mathbb{R}} f(s\theta+t\theta^\perp)\mathrm{e}^{-\mathrm{i}\sigma s}\mathrm{d}t\mathrm{d}s = \int_{\mathbb{R}} \mathcal{R}f(\theta,s)\mathrm{e}^{-\mathrm{i}\sigma s}\mathrm{d}s = \mathcal{F}_1[\mathcal{R}f(\theta,\cdot)](\sigma)$, with $s \in \mathbb{R} \mapsto \mathcal{R}f(\theta,s) = \int_{\mathbb{R}} f(s\theta+t\theta^\perp)\mathrm{d}t \in L^1(\mathbb{R})$. □

Another consequence of Fubini's theorem is the duality relation [58, Chap. I, Lemma 5.1]

$$\int_{\mathbb{S}^1\times\mathbb{R}} \int_{x\cdot\theta=s} f(x)\mathrm{d}\ell\, g(\theta,s)\,\mathrm{d}\theta\mathrm{d}s = \int_{\mathbb{R}^2} f(x)\int_{\mathbb{S}^1} g(\theta,x\cdot\theta)\mathrm{d}\theta\,\mathrm{d}x, \tag{A.4}$$

valid for several classes of functions; in particular, it is true if $f \in L^1(\mathbb{R}^2)$ is compactly supported and $g \in \mathcal{E}(\mathbb{S}^1\times\mathbb{R})$, or if $f \in \mathcal{D}(\mathbb{R}^2)$ and $g \in L^1_{\mathrm{loc}}(\mathbb{S}^1\times\mathbb{R})$. The duality relation motivates the introduction of a *backprojection* operator $\mathcal{R}^*$, which integrates a line function over lines through a fixed $x \in \mathbb{R}^2$:

$$\mathcal{R}^* : L^1_{\mathrm{loc}}(\mathbb{S}^1\times\mathbb{R}) \to L^1_{\mathrm{loc}}(\mathbb{R}^2), \quad \mathcal{R}^*g(x) = \int_{\mathbb{S}^1} g(\theta,x\cdot\theta)\,\mathrm{d}\theta. \tag{A.5}$$

The Radon transform an the backprojection act continuously on smooth functions as follows [77, Chap. 2].

Theorem A.2. *By restriction, the Radon transform $\mathcal{R}$ in* (A.1) *defines a continuous linear map $\mathcal{R} : \mathcal{D}(\mathbb{R}^2) \to \mathcal{D}(\mathbb{S}^1\times\mathbb{R})$, the backprojection $\mathcal{R}^*$ in* (A.5) *defines a continuous linear map $\mathcal{R}^* : \mathcal{E}(\mathbb{S}^1\times\mathbb{R}) \to \mathcal{E}(\mathbb{R}^2)$, which is the transpose of $\mathcal{R}$,* i.e.

$$\int_{\mathbb{S}^1\times\mathbb{R}} \mathcal{R}f(\theta,s)g(\theta,s)\,\mathrm{d}\theta\mathrm{d}s = \int_{\mathbb{R}^2} f(x)\mathcal{R}^*g(x)\,\mathrm{d}x, \quad f \in \mathcal{D}(\mathbb{R}^2),\ g \in \mathcal{E}(\mathbb{S}^1\times\mathbb{R}). \tag{A.6}$$

A.4 Radon transform on distributions

The Radon transform and the backprojection are extended to distributions by duality [58, Chap. I], [77, Chap. 2].

Definition-Theorem A.3. Extending the duality relation (A.6) by

$$\langle \mathcal{R}f, g\rangle := \langle f, \mathcal{R}^*g\rangle, \quad f \in \mathcal{E}'(\mathbb{R}^2), \quad g \in \mathcal{E}(\mathbb{S}^1\times\mathbb{R}), \tag{A.7}$$

$$\langle \mathcal{R}^*g, f\rangle := \langle g, \mathcal{R}f\rangle, \quad g \in \mathcal{D}'(\mathbb{S}^1\times\mathbb{R}), \quad f \in \mathcal{D}(\mathbb{R}^2), \tag{A.8}$$

defines the unique continuous linear extension $\mathcal{R} : \mathcal{E}'(\mathbb{R}^2) \to \mathcal{E}'(\mathbb{S}^1\times\mathbb{R})$ of the Radon transform $\mathcal{R} : \mathcal{D}(\mathbb{R}^2) \to \mathcal{D}(\mathbb{S}^1\times\mathbb{R})$, and the unique continuous linear extension $\mathcal{R}^* : \mathcal{D}'(\mathbb{S}^1\times\mathbb{R}) \to \mathcal{D}'(\mathbb{R}^2)$ of the backprojection $\mathcal{R}^* : \mathcal{E}(\mathbb{S}^1\times\mathbb{R}) \to \mathcal{E}(\mathbb{R}^2)$.

Note that for any function $f \in L^1(\mathbb{R}^2)$ with compact support, identified with a distribution $f \in \mathcal{E}'(\mathbb{R}^2)$, the Radon transform $\mathcal{R}f \in \mathcal{E}'(\mathbb{S}^1\times\mathbb{R})$ coincides with the integrable function defined by (A.1), because the duality relation (A.4) is valid for any $g \in \mathcal{E}(\mathbb{S}^1\times\mathbb{R})$. Analogously, for any function $g \in L^1_{\mathrm{loc}}(\mathbb{S}^1\times\mathbb{R})$, identified with a distribution $g \in \mathcal{D}'(\mathbb{S}^1\times\mathbb{R})$, the backprojection $\mathcal{R}^*g \in \mathcal{D}'(\mathbb{R}^2)$ coincides with the locally integrable function defined in (A.5), because (A.4) is valid for any $f \in \mathcal{D}(\mathbb{R}^2)$.

A.5 Inversion of the Radon transform

The following result provides an inversion formula to deduce a function $f \in \mathcal{D}(\mathbb{R}^2)$ from the Radon transform $\mathcal{R}f$; it is a consequence of the Fourier slice theorem. See for instance [78, Theorem 2.6] (where the formula is given for $f \in \mathcal{S}(\mathbb{R}^2)$), [87, Theorem 2.5].

Theorem A.4 (Radon inversion formula)**.** *The Radon transform can be inverted by the formula*

$$f = \mathcal{R}^* \Lambda \mathcal{R} f, \quad \textit{with} \quad \Lambda := \tfrac{1}{4\pi} \mathcal{H}_s \partial_s, \qquad f \in \mathcal{D}(\mathbb{R}^2); \tag{A.9}$$

here, $\mathcal{H}_s$ denotes the Hilbert transform with respect to s, defined by the Cauchy principal value

$$\mathcal{H} : \mathcal{D}(\mathbb{R}) \to \mathcal{E}(\mathbb{R}), \quad \mathcal{H}g(s) = \frac{1}{\pi} \, \mathrm{p.\,v.} \int_{\mathbb{R}} \frac{g(t)}{s-t} \mathrm{d}t. \tag{A.10}$$

Note that the operator Λ defines a continuous linear map $\Lambda : \mathcal{D}(\mathbb{S}^1 \times \mathbb{R}) \to \mathcal{E}(\mathbb{S}^1 \times \mathbb{R})$ such that

$$\Lambda g(\theta, s) = \frac{1}{4\pi} \mathcal{F}_1^{-1}\{|\sigma| \mathcal{F}_1[g(\theta, \cdot)](\sigma)\} = \frac{1}{8\pi^2} \int_{\mathbb{R}} \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}\sigma(t-s)} |\sigma| g(\theta, t) \, \mathrm{d}t \mathrm{d}\sigma, \quad g \in \mathcal{D}(\mathbb{S}^1 \times \mathbb{R}). \tag{A.11}$$

The inversion formula extends to distributions [58, Chap. I, Theorem 5.5].

Theorem A.5. *The inversion formula* (A.9) *is valid in $\mathcal{E}'(\mathbb{R}^2)$,* i.e.

$$f = \mathcal{R}^* \Lambda \mathcal{R} f, \quad f \in \mathcal{E}'(\mathbb{R}^2), \tag{A.12}$$

where $\Lambda : \mathcal{E}'(\mathbb{S}^1 \times \mathbb{R}) \to \mathcal{D}'(\mathbb{S}^1 \times \mathbb{R})$ is the continuous linear extension of $\Lambda : \mathcal{D}(\mathbb{S}^1 \times \mathbb{R}) \to \mathcal{E}(\mathbb{S}^1 \times \mathbb{R})$, defined by $\langle \Lambda g, \phi \rangle := \langle g, \Lambda \phi \rangle$, $g \in \mathcal{E}'(\mathbb{S}^1 \times \mathbb{R})$, $\phi \in \mathcal{D}(\mathbb{S}^1 \times \mathbb{R})$.

A.6 Filtered backprojection algorithm

The filtered backprojection algorithm is the most important algorithm in tomography [78, Chap. 5]. It is a practical implementation of the Radon formula (A.9), with a regularization performed by a low-pass filter. More precisely, the operator Λ, satisfying (9.3), is replaced by a convolution with a smooth filter ψ_Ω such that

$$\psi_\Omega(s) = \frac{1}{4\pi} \mathcal{F}_1^{-1}\{|\sigma| \hat{h}_\Omega(\sigma)\}(s) = \frac{1}{8\pi^2} \int_{-\Omega}^{\Omega} |\sigma| \hat{h}_\Omega(\sigma) \mathrm{e}^{\mathrm{i}\sigma s} \, \mathrm{d}\sigma \in \mathcal{E}(\mathbb{R}); \tag{A.13}$$

the parameter $\Omega > 0$ defines a *cutoff* pulsation, and $\hat{h}_\Omega : \mathbb{R} \to [0, 1]$ denotes an even *windowing* function with compact support $[-\Omega, \Omega]$. Therefore, the reconstruction formula becomes $\mathcal{R}^*[\mathcal{R}f \star \psi_\Omega]$, which is the so-called *filtered backprojection* (FBP). The following theorem justifies this regularization procedure. We prove this fundamental result, which is similar with [78, Eq. (5.1)], but with weaker assumptions.

Theorem A.6 (FBP)**.** *Let ψ_Ω be a filter with a windowing function $\hat{h}_\Omega$, as* (A.13)*. For any function $f \in L^1(\mathbb{R}^2)$ with compact support,*

$$\mathcal{R}^*[\mathcal{R}f \star \psi_\Omega] = f \star \Psi_\Omega \in \mathcal{E}(\mathbb{R}^2), \quad f \in L^1(\mathbb{R}^2) \cap \mathcal{E}'(\mathbb{R}^2), \tag{A.14}$$

where Ψ_Ω denotes the filter

$$\Psi_\Omega \in \mathcal{E}(\mathbb{R}^2), \quad \Psi_\Omega(x) = \mathcal{F}_2^{-1}\{\hat{h}_\Omega(|\xi|)\}(x) = \frac{1}{4\pi^2} \int_{|\xi| \leq \Omega} \hat{h}_\Omega(|\xi|) \mathrm{e}^{\mathrm{i}\xi \cdot x} \mathrm{d}\xi.$$

Proof. Fix $f \in L^1(\mathbb{R}^2) \cap \mathcal{E}'(\mathbb{R}^2)$. Firstly, consider the right-hand side of (A.14). The kernel Ψ_Ω is defined by $\Psi_\Omega = \mathcal{F}_2^{-1}(\hat{h}_\Omega \circ |\cdot|)$, with $\hat{h}_\Omega \circ |\cdot| \in \mathcal{E}'(\mathbb{R}^2) \subset \mathcal{S}'(\mathbb{R}^2)$. Therefore, $\Psi_\Omega \in \mathcal{E}(\mathbb{R}^2) \cap \mathcal{S}'(\mathbb{R}^2)$ and the convolution $f \star \Psi_\Omega \in \mathcal{E}(\mathbb{R}^2)$ satisfies

$$f \star \Psi_\Omega(x) = \mathcal{F}_2^{-1}\{\mathcal{F}_2[\Psi_\Omega]\mathcal{F}_2[f]\}(x) = \frac{1}{4\pi^2}\int_{|\xi|\leq\Omega} \hat{h}_\Omega(|\xi|)\mathcal{F}_2[f](\xi)\mathrm{e}^{\mathrm{i}x\cdot\xi}\mathrm{d}\xi.$$

Introducing polar coordinates $\xi = \sigma\theta$, once with $\sigma > 0$ and once with $\sigma < 0$, we obtain

$$f \star \Psi_\Omega(x) = \frac{1}{8\pi^2}\int_{\mathbb{S}^1}\int_{-\Omega}^{\Omega} |\sigma|\hat{h}_\Omega(\sigma)\mathcal{F}_2[f](\sigma\theta)\mathrm{e}^{\mathrm{i}\sigma x\cdot\theta}\mathrm{d}\sigma\mathrm{d}\theta. \tag{A.15}$$

Secondly, consider the convolution of the left-hand side of (A.14). The Radon transform is an integrable function $\mathcal{R}f \in L^1(\mathbb{S}^1 \times \mathbb{R})$ with compact support, since $f \in L^1(\mathbb{R}^2) \cap \mathcal{E}'(\mathbb{R}^2)$. Also, for any $\theta \in \mathbb{S}^1$, the integrable function $\mathcal{R}f(\theta,\cdot) \in L^1(\mathbb{R})$ is compactly supported. The kernel $\psi_\Omega = \frac{1}{4\pi}\mathcal{F}_1^{-1}\{|\sigma|\hat{h}_\Omega(\sigma)\}$, with $|\sigma|\hat{h}_\Omega \in \mathcal{E}'(\mathbb{R})$, is in $\mathcal{E}(\mathbb{R}) \cap \mathcal{S}'(\mathbb{R})$; therefore, for any $\theta \in \mathbb{S}^1$, the convolution $\mathcal{R}f(\theta,\cdot) \star \psi_\Omega \in \mathcal{E}(\mathbb{R}) \cap \mathcal{S}'(\mathbb{R})$ satisfies

$$\mathcal{R}f(\theta,\cdot) \star \psi_\Omega(s) = \mathcal{F}_1^{-1}\{\mathcal{F}_1[\mathcal{R}f(\theta,\cdot)]\mathcal{F}_1[\psi_\Omega]\}(s) = \frac{1}{8\pi^2}\int_{-\Omega}^{\Omega} |\sigma|\hat{h}_\Omega(\sigma)\mathcal{F}_1[\mathcal{R}f(\theta,\cdot)]\mathrm{e}^{\mathrm{i}\sigma s}\mathrm{d}\sigma, \tag{A.16}$$

where $\mathcal{F}_1[\mathcal{R}f(\theta,\cdot)](\sigma) = \mathcal{F}_2[f](\sigma\theta)$ by the Fourier slice theorem. Since $f \in \mathcal{E}'(\mathbb{R}^2)$, we obtain that $\mathcal{F}_2[f] \in \mathcal{E}(\mathbb{R}^2)$ and

$$(\theta, s) \mapsto \mathcal{R}f(\theta,\cdot) \star \psi_\Omega(s) = \frac{1}{8\pi^2}\int_{-\Omega}^{\Omega} |\sigma|\hat{h}_\Omega(\sigma)\mathcal{F}_2[f](\sigma\theta)\mathrm{e}^{\mathrm{i}\sigma s}\mathrm{d}\sigma \in \mathcal{E}(\mathbb{S}^1 \times \mathbb{R}).$$

To finish with, the backprojection $\mathcal{R}^*[\mathcal{R}f(\theta,\cdot) \star \psi_\Omega] \in \mathcal{E}(\mathbb{R}^2)$ of this smooth function coincides with the right-hand side of (A.15). □

The FBP formula (A.14) has the following practical consequences. Assume that we know line integrals $\mathcal{R}f(\theta, s)$, $(\theta, s) \in \mathbb{S}^1 \times \mathbb{R}$, of an integrable function f with compact support. At a first step, we compute a filtering $\mathcal{R}f \star \psi_\Omega$, for some cutoff pulsation $\Omega > 0$ and some windowing function $\hat{h}_\Omega$. In a second step, we compute the backprojection $\mathcal{R}^*[\mathcal{R}f \star \psi_\Omega]$. The formula (A.14) proves that this process computes $f \star \Psi_\Omega$. In particular, if $\hat{h}_\Omega(s) \to 1$ (pointwise convergence) when $\Omega \to \infty$, then $\Psi_\Omega \to \delta$ and $\mathcal{R}^*[\mathcal{R}f \star \psi_\Omega] \to f$ in $\mathcal{S}'(\mathbb{R}^2)$. Furthermore, if the function f is essentially Ω-bandlimited, *i.e.* if $|\mathcal{F}_2 f(\xi)|$ is negligible for $|\xi| > \Omega$, then the formula (A.14), with the ideal low-pass $\hat{h}_\Omega = \mathbb{1}_{[-\Omega,\Omega]}$, proves that $\mathcal{R}^*[\mathcal{R}f \star \psi_\Omega] \approx f$.

Finally, the *FBP algorithm* on the facing page implements a discretization of (A.14), for an integrable function f, with support in a disk $|x| < S$, and with essential bandwidth Ω [78, Algorithm 5.1], [88, pp. 91-92]. Various choices of the windowing function $\hat{h}_\Omega$ can be found in the literature; *e.g.* the *Ram-Lak filter* corresponds to the ideal low-pass $\hat{h}_\Omega = \mathbb{1}_{[-\Omega,\Omega]}$. The convolution is often computed by Fast Fourier Transforms, using a discrete version of (A.16). Concerning sampling conditions, the radial step must satisfy $\delta s \leq \frac{\pi}{\Omega}$ in order to get acceptable results for the convolution. Furthermore, if $\delta\theta > \frac{\pi}{\Omega S}$, then artifacts may appear outside the disk $|x| < \frac{\pi}{\Omega\delta\theta}$. We refer to [78] for more details.

A.7 Algebraic Reconstruction Technique

Tomography deals especially with a linear system, because the Radon transform is a linear operator. Unsurprisingly, some common inversion methods are iterative techniques based on linear algebra, or optimization. In this section, we present the principle of the famous Algebraic Reconstruction Technique (ART), based on the so-called Kaczmarz iterations [22, 60, 78]. We refer to [46, 56] (and

FBP algorithm.

Input. Radon transform $\mathcal{R}f$ on a regular grid, with steps $\delta\theta = \frac{2\pi}{p}$ and $\delta s = \frac{S}{q} > 0$,

$$\mathcal{R}f(\theta_j, s_l), \quad \theta_j := (\cos j\delta\theta, \sin j\delta\theta),\ s_l := l\delta s, \quad 0 \leq j \leq p-1,\ -q \leq l \leq q.$$

Step 1. For any $0 \leq j \leq p-1$, compute the discrete convolution

$$g_{j,k} = \delta s \sum_{l=-q}^{q} \psi_\Omega(s_k - s_l)\mathcal{R}f(\theta_j, s_l) \quad [\approx \mathcal{R}f(\theta_j, \cdot) \star \psi_\Omega(s_k)], \quad -q \leq k \leq q.$$

Step 2. For each reconstruction point x, compute the discrete backprojection

$$f_{\mathrm{FBP}}(x) = \delta\theta \sum_{j=0}^{p-1} (1-\omega)g_{j,k} + \omega g_{j,k+1}, \quad [\approx \mathcal{R}^*[\mathcal{R}f \star \psi_\Omega](x)],$$

where $k = k(j,x)$ and $\omega = \omega(j,x) \in [0,1)$ are chosen such that $x \cdot \theta_j = (1-\omega)s_k + \omega s_{k+1}$.

Output. FBP reconstruction $x \mapsto f_{\mathrm{FBP}}(x) \approx f \star \Psi_\Omega(x)$.

the references therein) for historical remarks, several variants of the principle, and some convergence results.

Consider a linear system decomposed into blocks of rows,

$$\begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_N \end{bmatrix} x = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}, \tag{A.17}$$

where, $x \in \mathbb{R}^n$ is the unknown, and for any $1 \leq j \leq N$, the j-th block is such that $A_j \in \mathbb{R}^{m_j \times n}$ has full row rank, and $b_j \in \mathbb{R}^{m_j}$. The Kaczmarz iterations consider the blocks of rows, block after block. The procedure is initialized with some $x_0 \in \mathbb{R}^n$. Then, for any $1 \leq j \leq N$, the iterate $x_j \in \mathbb{R}^n$ is defined as an average between x_{j-1} and its orthogonal projection onto the affine subspace $A_j x = b_j$, with respective weights $(1-\omega)$ and $\omega > 0$,

$$x_j := x_{j-1} + \omega A_j^\intercal (A_j A_j^\intercal)^{-1}(b_j - A_j x_{j-1}). \tag{A.18}$$

After a cycle of N iterations, an estimation x_N of a solution of (A.17) is obtained; the constraint of each block of rows has been used once. Then, another cycle of iterations can be performed, using x_N as initial state x_0.

In tomography, each block $A_j \in \mathbb{R}^{m_j \times n}$ corresponds typically to the matrix of the Radon transform, expressed in a collection of n fixed basis functions, and evaluated on a collection of m_j rays depending on j (one line per ray, one column per basis function). The vector $b_j \in \mathbb{R}^{m_j}$ contains tomographic projections of a function f along these rays, and x represents the decomposition of f in the selected basis. It is known that an ART is computationnally expensive, but it has the advantage of flexibility. In comparison, the FBP algorithm, deduced from the analytical formula (A.14), is very efficient but is dedicated to a particular acquisition geometry.

A.8 Microlocal analysis of the Radon transform

The microlocal analysis of the Radon transform has been extensively studied for various purposes in X-ray tomography, since the 1980s. It is a relevant framework for a description of limited data tomography, as it is detailed in numerous works of E.T. Quinto [38, 50, 69, 85–87]. We can also

Kaczmarz iterations, for the linear system (A.17).

Input. Full row rank matrices $A_j \in \mathbb{R}^{m_j \times n}$, vectors $b_j \in \mathbb{R}^{m_j}$, $1 \leq j \leq N$, initialization $x_0 \in \mathbb{R}^n$, weight $\omega > 0$.

Fix $x_0^{(0)} := x_0$.
For $k = 0, 1, 2, \ldots$, realize a cycle

$$x_j^{(k)} := x_{j-1}^{(k)} + \omega A_j^\intercal (A_j A_j^\intercal)^{-1} (b_j - A_j x_{j-1}^{(k)}), \quad 1 \leq j \leq N,$$

and fix $x_0^{(k+1)} := x_N^{(k)}$.

Output. Approximate solution $x_0^{(k+1)}$ to the linear system $A_j x = b_j$, $1 \leq j \leq N$ (A.17), obtained after $k+1$ cycles of iterations.

mention *local tomography* [47, 70, 97], [88, Chap. 5], *pseudolocal tomography* [88, Chap. 6], and *geometrical tomography* [88, Chap. 7], which deal with finding singularities of a function from the knowledge of tomographic data; these subjects are studied in depth in [88].

A crucial point is that the Radon transform defines a *Fourier integral operator* [55, Chap. VI], [84]. As a consequence, there is a *canonical relation* which implies correspondences between the singularities of a distribution and the singularities of its Radon transform [85,86], [88, Chap. 4]. We refer to standard references of microlocal analysis for a comprehensive description of Fourier integral operators, such as [54,61,99].

Theorem A.7. *The Radon transform $\mathcal{R}$ and the transpose $\mathcal{R}^*$ define Fourier integral operators with Schwartz kernel $\delta(s - x \cdot \theta) = \int_{\mathbb{R}} \frac{1}{2\pi} \mathrm{e}^{\mathrm{i}\sigma(s-\theta\cdot x)} \mathrm{d}\sigma$,*

$$\mathcal{R}f(\theta, s) = \int_{\mathbb{R}^2} \delta(s - x \cdot \theta) f(x)\, \mathrm{d}x, \qquad (\theta, s) \in \mathbb{S}^1 \times \mathbb{R},\ f \in \mathcal{D}(\mathbb{R}^2);$$
$$\mathcal{R}^* g(x) = \int_{\mathbb{S}^1 \times \mathbb{R}} \delta(s - x \cdot \theta) g(\theta, s)\, \mathrm{d}\theta \mathrm{d}s, \qquad x \in \mathbb{R}^2,\ g \in \mathcal{E}(\mathbb{S}^1 \times \mathbb{R}).$$

Theorem A.8 (Correspondence of singularities)**.** *The wavefront sets satisfy:*

$$\mathrm{WF}(\mathcal{R}f) = \big\{(\theta, s; \hat\theta, \hat s) \in \mathbb{S}^1 \times \mathbb{R} \times \mathbb{R}^2 : \hat s \neq 0 \text{ and } (s\theta + \tfrac{\hat\theta}{\hat s}\theta^\perp; \hat s\theta) \in \mathrm{WF}(f)\big\}, \quad f \in \mathcal{E}'(\mathbb{R}^2), \tag{A.19}$$
$$\mathrm{WF}(\mathcal{R}^* g) \subset \big\{(s\theta + \tfrac{\hat\theta}{\hat s}\theta^\perp; \hat s\theta), \text{ with } (\theta, s, \hat\theta, \hat s) \in \mathrm{WF}(g) \text{ and } \hat s \neq 0\big\}, \quad g \in \mathcal{D}'(\mathbb{S}^1 \times \mathbb{R}). \tag{A.20}$$

Remark A.9. If $g \in L^1_{\mathrm{loc}}(\mathbb{S}^1 \times \mathbb{R})$ satisfies the symmetry $g(-\theta, -s) = g(\theta, s)$, the inclusion (A.20) is an equality [38].

Recall that an element of a wavefront set encodes a *singularity* defined by a location and a direction. For instance, for a distribution in $\mathbb{R}^2$, the wavefront set is defined as follows.

Definition A.10. A distribution $f \in \mathcal{D}'(\mathbb{R}^2)$ is *smooth at $x_0 \in \mathbb{R}^2$ in direction $\xi_0 \in \mathbb{R}^2 \setminus \{0\}$* if, and only if, there are a cutoff $\psi \in \mathcal{D}(\mathbb{R}^2)$ with $\psi(x) = 1$ in a neighbourhood of x_0, and an open cone V containing ξ_0, such that the Fourier transform $\mathcal{F}_2[\psi f](\xi) = \langle f, \mathrm{e}^{-\mathrm{i}x\cdot\xi}\psi(x)\rangle$ is rapidly decaying at infinity for $\xi \in V$, *i.e.* $\forall m \geq 0$, $\exists c_m \in \mathbb{R}$, $\forall \xi \in V$, $|\mathcal{F}_2[\psi f](\xi)| \leq \frac{c_m}{(1+|\xi|)^m}$. The *wavefront set* of f, denoted by $\mathrm{WF}(f)$, is the set of $(x_0, \xi_0) \in \mathbb{R}^2 \times (\mathbb{R}^2 \setminus \{0\})$ such that f is not smooth at x_0 in direction ξ_0.

Theorem A.8 is a deep framework about singularities in tomography. Firstly, the relation (A.19) claims that f and $\mathcal{R}f$ have the "same" singularities; in particular, any singularity of a function f with compact support can be deduced from the singularities of the tomographic projection $\mathcal{R}f$. Secondly, the inclusion (A.20) claims that the backprojection $\mathcal{R}^*$ does not add any singularity. It

implies that for any pseudodifferential operator $\Lambda : \mathcal{E}'(\mathbb{S}^1 \times \mathbb{R}) \to \mathcal{D}'(\mathbb{S}^1 \times \mathbb{R})$, *e.g.* for $\Lambda = \frac{1}{4\pi}\mathcal{H}_s\partial_s$ as in (A.12),

$$\mathrm{WF}(\mathcal{R}^*\Lambda g) \subset \big\{(s\theta + \tfrac{\hat\theta}{\hat s}\theta^\perp; \hat s\theta), \text{ with } (\theta, s, \hat\theta, \hat s) \in \mathrm{WF}(g) \text{ and } \hat s \neq 0\big\}, \quad g \in \mathcal{E}'(\mathbb{S}^1 \times \mathbb{R}), \tag{A.21}$$

because $\mathrm{WF}\,\Lambda g \subset \mathrm{WF}\, g$. This has consequences in incomplete data tomography. In this case, the dataset is a truncated Radon transform, $g = \mathbb{1}_A\mathcal{R}f$ with $A \subsetneq \operatorname{supp}\mathcal{R}f$. The singularities of f correspond to those of $\mathcal{R}f$. Some of them may be invisible in g due to the truncation $\mathbb{1}_A$; they are also invisible in any reconstruction $\mathcal{R}^*\Lambda g$, by (A.21). Furthermore, the abrupt truncation $\mathbb{1}_A$ may introduce singularities in g which do not correspond to singularities of $\mathcal{R}f$ (or f). By (A.21), artifacts corresponding to these additional singularities are expected in $\mathcal{R}^*\Lambda g$. See, for instance, [38].

A.A Radon transform of disks

In this subappendix, we consider the specific case of the Radon transform of disks; these results are used in the proof of Theorem 2.2 in Subsection 2.4.4.

Proposition A.11 (Radon transform of a disk)**.** *Let $K = \{x \in \mathbb{R}^2 : |x - z| \leq r\}$ be a disk with radius $r > 0$ and center $z \in \mathbb{R}^2$.*
(i) The wavefront set of $\mathbb{1}_K$ represents the lines which are tangent to the circle ∂K:

$$\mathrm{WF}\,\mathbb{1}_K = \{(x; \hat x) \in \partial K \times (\mathbb{R}^2 \setminus \{0\}) : \hat x \text{ is a normal vector to } \partial K \text{ at } x \in \partial K\}.$$

(ii) The Radon transform of the disk K is given by

$$\tfrac{1}{2r}\mathcal{R}[\mathbb{1}_K](\theta, s) = [1 - (\tfrac{s - z\cdot\theta}{r})^2]^{1/2}\mathbb{1}_{|s-z\cdot\theta|\leq r}. \tag{A.22}$$

(iii) The wavefront set of $\mathcal{R}[\mathbb{1}_K]$ is given by

$$\mathrm{WF}\,\mathcal{R}\mathbb{1}_K = \big\{(\theta, s; \hat\theta, \hat s) \in \mathbb{S}^1 \times \mathbb{R} \times \mathbb{R}^2 : \hat s \neq 0 \text{ and } \theta \text{ is normal to } \partial K \text{ at } s\theta + \tfrac{\hat\theta}{\hat s}\theta^\perp \in \partial K\big\}. \tag{A.23}$$

Proof. (i) is a classical result. (ii) results from a classical computation which measures the length of the intersection of the line $x \cdot \theta = s$ and the disk K. (iii) is a consequence of (i) and (A.19). □

Proposition A.12 (Radon transform of two disks)**.** *Let $K = K_1 \cup K_2$ be a union of two disjoint disks K_1 and K_2. Let $T_K \subset \mathbb{R}^2$ denote the union of the four straight lines which are tangent to K_1 and K_2.*
(i) The wavefront sets of $\mathbb{1}_K$ and $\mathcal{R}\mathbb{1}_K$ are given by the disjoint unions

$$\mathrm{WF}\,\mathbb{1}_K = \mathrm{WF}\,\mathbb{1}_{K_1} \cup \mathrm{WF}\,\mathbb{1}_{K_2}, \quad \mathrm{WF}\,\mathcal{R}\mathbb{1}_K = \mathrm{WF}\,\mathcal{R}\mathbb{1}_{K_1} \cup \mathrm{WF}\,\mathcal{R}\mathbb{1}_{K_2}.$$

(ii) The eight couples of parameters (θ, s) associated to the four lines in T_K, are given by the intersection of the singular supports of $\mathcal{R}\mathbb{1}_{K_1}$ and $\mathcal{R}\mathbb{1}_{K_2}$,

$$\operatorname{sing\,supp}\mathcal{R}\mathbb{1}_{K_1} \cap \operatorname{sing\,supp}\mathcal{R}\mathbb{1}_{K_2} = \{(\theta, s) \in \mathbb{S}^1 \times \mathbb{R} : \{x \cdot \theta = s\} \subset T_K\}.$$

(iii) If $(\theta, s, \hat\theta_i, \hat s_i) \in \mathrm{WF}\,\mathcal{R}\mathbb{1}_{K_i}$, $i = 1, 2$, then $(\hat\theta_1, \hat s_1)$ and $(\hat\theta_2, \hat s_2)$ are linearly independent.

Proof. (i) $\mathbb{1}_K = \mathbb{1}_{K_1} + \mathbb{1}_{K_2}$ where $\mathbb{1}_{K_1}$ and $\mathbb{1}_{K_2}$ have disjoint supports, so $\mathrm{WF}\,\mathbb{1}_K = \mathrm{WF}\,\mathbb{1}_{K_1} \cup \mathrm{WF}\,\mathbb{1}_{K_2}$ and this union is disjoint; we deduce $\mathrm{WF}\,\mathcal{R}\mathbb{1}_K = \mathrm{WF}\,\mathcal{R}\mathbb{1}_{K_1} \cup \mathrm{WF}\,\mathcal{R}\mathbb{1}_{K_2}$ from (A.19). This union is disjoint due to the point (iii) hereafter. (ii-iii) are a consequence of Proposition A.11.(iii): $(\theta, s) \in \operatorname{sing\,supp}\mathcal{R}\mathbb{1}_{K_i} \Leftrightarrow$ the line $x\cdot\theta = s$ is tangent to ∂K_i, hence the intersection of the singular supports; moreover, for $(\theta, s, \hat\theta_i, \hat s_i) \in \mathrm{WF}\,\mathcal{R}\mathbb{1}_{K_i}$, $i = 1, 2$, the line $x\cdot\theta = s$ is tangent to ∂K_i at $s\theta + \frac{\hat\theta_i}{\hat s_i}\theta^\perp \in \partial K_i$, hence $\frac{\hat\theta_1}{\hat s_1} \neq \frac{\hat\theta_2}{\hat s_2}$, because $\partial K_1 \cap \partial K_2 = \emptyset$. □

Appendix B

Physical background

B.1 Electromagnetic imagery

The first part of this habilitation thesis enters in the framework of imagery based on electromagnetic waves. In this field, the goal is to produce images of a scene from records of electromagnetic radiation, in some wavelength range of the electromagnetic spectrum. Various ranges are recalled in Table B.1. Also, for illustration purposes, Figure B.1 contains several images, corresponding to several types of radiation. Image (a) has been obtained using microwaves, with a Synthetic Aperture Radar in the Ka-band (7.5-11.1 mm). Image (b) is a thermal image, in the infrared band (1-14 µm). Image (c) is a radiography, obtained with X-rays (0.01-10 nm).

(a) (b) (c)

Figure B.1: Images from various types of electromagnetic radiation.
(a) Synthetic Aperture Radar image of C-130s, in the Ka-band. Courtesy of Sandia National Laboratories, Radar ISR, `https://www.sandia.gov/radar/imagery/index.html`
(b) Thermal view of the Space Shuttle Atlantis. Source: `https://commons.wikimedia.org/wiki/File:STS-135_thermal_view.jpg`
(c) Lung radiography.

Radiation	Wavelength range [m]
Gamma Rays	$< 1 \cdot 10^{-11}$
X-Rays	$1 \cdot 10^{-11} - 1 \cdot 10^{-8}$
Ultraviolet	$1 \cdot 10^{-8} - 4 \cdot 10^{-7}$
Visible	$4 \cdot 10^{-7} - 7 \cdot 10^{-7}$
Infrared	$7 \cdot 10^{-7} - 1 \cdot 10^{-3}$
Microwaves	$1 \cdot 10^{-3} - 1 \cdot 10^{-1}$
Radio	$> 1 \cdot 10^{-1}$

Table B.1: Approximate wavelength ranges of the electromagnetic spectrum.

Concerning the acquisition, two categories of recording methods are usually distinguished. In *active* imaging, the acquisition device emits an incident electromagnetic wave, and records an interaction of this incident wave with the scene. In *passive* imaging, the acquisition device records some radiation, without emitting any wave. Concerning the use of the records, algorithms based on a suitable modeling are often applied in order to compute a reconstruction of the observed scene. The reconstruction encodes the geometry of the scene, and/or the spatial distribution of some physical parameters. Without being exhaustive, here are three examples, associated with three types of radiation.

1. Radar imaging aims at computing the reflectivity of a target, from measurements of scattered electric fields; see [41].

2. X-ray computerized tomography aims at computing the spatial attenuation of a medium, from the intensity attenuation between an X-ray source and detectors; see [60, 78].

3. In computer vision, one builds a geometric model of a scene, from several photographs in the visible band; see [57, 73].

To put this thesis in context, we are especially interested in visible to near-infrared (VIS-NIR) optics, in the visible (0.4-0.7 µm) to near-infrared (0.7-3 µm) band. This band includes, but is not limited to, photographs with current digital cameras, with a CCD (charge coupled device) or CMOS (complementary metal–oxide–semiconductor) sensor. Concerning the algorithmic part, we take benefit from algorithms from X-ray tomography, such as the inversion of the Radon transform, or the Feldkamp-Davis-Kress algorithm. That is the reason why we present a modeling of image formation in VIS-NIR optics in Section B.2, and a modeling of cone beam tomography in Section B.3.

B.2 Image formation

B.2.1 Introduction

In VIS-NIR optics, an image is generally formed according to the process displayed in Figure B.2. Light sources, such as the sun, a lamp or a laser, emit light. This light interacts with the illuminated scene; hence, some light emanates from the scene. A portion of that light passes through the optics of a camera, such as lenses, and finally reaches the camera's sensor, such as a digital sensor array.

Figure B.2: Image formation in VIS-NIR optics. Light is emitted by light sources and interacts with the illuminated scene. A portion of the light passes through the camera's optics and reaches the camera's sensor.

Since light is an electromagnetic wave, the propagation of light and its behavior at interfaces between media is governed by the Maxwell equations. In VIS-NIR optics, the wavelength is very small in comparison with some characteristic distances of the scene, while the surfaces often appear as rough surfaces. Hence, "effective" models are rather considered. In this section, we present some model commonly used in computer vision; see for instance [45, 62, 73, 98] and the references therein.

This modeling concerns the brightness of an optical image, and is related to geometrical modeling based on rays (geometrical optics). For deeper aspects, we refer to standard textbooks in optics, such as [39].

B.2.2 Pinhole camera model

A camera is often modeled as an ideal pinhole camera, which realizes a perspective projection of the observed scene [57, 73]; this provides a geometric model of an image by the means of rays of projection. We describe this ideal model.

A digital camera contains an optical system and an array of light sensors. The optical system, made of lenses, aims at directing the incident light onto the sensors, as in Figure B.3. The optical system has a rotational axis of symmetry, called the *optical axis*; the light sensors are contained in a plane orthogonal to the optical axis, called the *focal plane*. In general, a point $x \in \mathbb{R}^3$ of the scene scatters light in any direction. Therefore, the optics of the camera receives a cone of light incident from x. Ideally, the optical system focuses that light from x onto a unique point $\hat{x}$ in the focal plane; furthermore, the straight line between the point x and the image point $\hat{x}$ intersects the optical axis at a point c independent from x, called the *optical center*. In particular, the ray through the optical center c is not deviated and determines the image point $\hat{x}$, by intersection of the straight line (x, c) with the focal plane. From a geometrical point of view, the focal plane receives some perspective projection of the scene, through the optical center c. This is the model of an ideal pinhole camera.

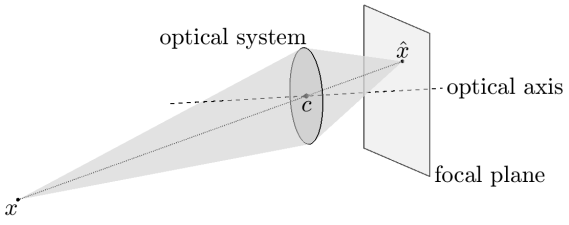


Figure B.3: Ideal camera.

For convenience, the perspective projection is often represented as in Figure B.4, where the focal plane is in front of the optical center c (the real one is obtained by central inversion with respect to c). In a world reference frame, a point x has coordinates $x = (x_1, x_2, x_3)$. We usually represent the location of the camera by the coordinates of the optical center $c = (c_1, c_2, c_3)$, whereas the orientation is represented by an orthogonal matrix $Q = [Q_1, Q_2, Q_3] \in \mathbb{R}^{3\times 3}$. The vector Q_3 is the direction of the optical axis, oriented from the camera towards the scene, while Q_1 and Q_2 represent the horizontal and vertical directions in the image. Then, in the camera frame (c, Q), the coordinates of the projection $\hat{x}$ are given by the vector

$$\tfrac{f}{\lambda} Q^\intercal (x - c), \text{ with } \lambda = Q_3^\intercal (x - c);$$

here, $Q^\intercal(x-c)$ contains the coordinates of x in the camera frame, λ represents the *depth* of x in this frame, and the *focal length* f represents the distance from the optical center c to the focal plane.

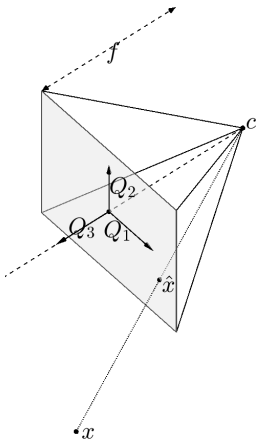


Figure B.4: Perspective projection.

The projection is collected by the array of light sensors in the focal plane, under the form of a pixelized image. Assume that the pixels are rectangles with sides aligned with Q_1, Q_2, and lengths s_1, s_2. Then, in pixel coordinates, the projection $\hat{x}$ is given by (i_1, i_2) such that

$$\lambda \begin{bmatrix} i_1 \\ i_2 \\ 1 \end{bmatrix} = K[Q^\intercal \; -Q^\intercal c] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{bmatrix}, \quad K = \begin{bmatrix} \frac{f}{s_1} & 0 & o_1 \\ 0 & \frac{f}{s_2} & o_2 \\ 0 & 0 & 1 \end{bmatrix}, \tag{B.1}$$

where (o_1, o_2) represents the pixel coordinates of the projection of the optical axis. The *extrinsic matrix* $[Q^\intercal \; -Q^\intercal c]$ depends only on the position and the orientation of the camera. The *intrinsic*

matrix K depends on the camera itself; it depends on the focal length and on the shape of a pixel on the receiver array, but does not depend on the position, nor on the orientation. In this thesis, the projections are idealized as (B.1). In particular, we assume that the eventual radial distortion have been removed, as in [73].

The relation (B.1) describes a geometrical model of image formation, from the visible scene to the light sensors of a camera. The brightness of a pixel depends on the amount of light received by the associated sensor. Modeling this brightness falls within the scope of radiometry.

B.2.3 Elements of radiometry

The field of radiometry deals with measuring or calculating electromagnetic radiation, which includes modeling image formation in VIS-NIR optics. The transfer of radiation among and between various objects, including sources and optical systems, is of particular concern. So is quantifying the mechanisms of emission, absorption, reflection and transmission of light. We refer to [106] for a tutorial text.

Radiance

The fundamental quantity is the *radiance* $[\mathtt{W.m^{-2}.sr^{-1}}]$, which represents the amount of light radiated from a surface; it is defined as the power (energy per time unit) radiated along a certain direction, per unit projected area, and per unit solid angle. More precisely, as in Figure B.5, consider an infinitesimal piece of surface, located at $x \in \mathbb{R}^3$, with area $\mathrm{d}\sigma_x$, and unit normal $\nu_x \in \mathbb{S}^2$. Consider an infinitesimal solid angle $\mathrm{d}\Omega$ around a direction $u \in \mathbb{S}^2$. The radiance at x in the direction u, denoted by $L(x,u)$, is related to the power $\mathrm{d}\Phi$ radiated from the piece of surface $\mathrm{d}\sigma_x$ through the solid angle $\mathrm{d}\Omega$ by

$$\mathrm{d}\Phi = L(x,u)\, \nu_x \cdot u \, \mathrm{d}\sigma_x \, \mathrm{d}\Omega. \tag{B.2}$$

Here, $\nu_x \cdot u$ is the cosine of the angle between the normal ν_x and the direction u; the projected area $\nu_x \cdot u \, \mathrm{d}\sigma_x$ represents the apparent area of the surface $\mathrm{d}\sigma_x$ seen from a point on the axis (x,u).

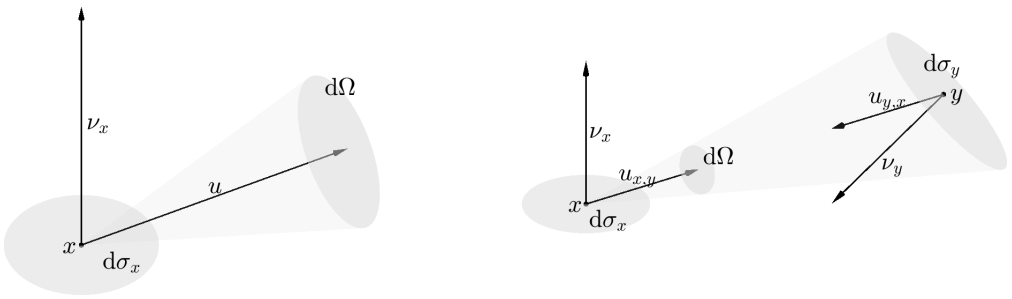


Figure B.5: Power $\mathrm{d}\Phi$ radiated by a surface element $\mathrm{d}\sigma_x$ through a solid angle $\mathrm{d}\Omega$. Left: the radiance $L(x,u)$ satisfies $\mathrm{d}\Phi = L(x,u)\, \nu_x \cdot u \, \mathrm{d}\sigma_x \, \mathrm{d}\Omega$. Right: the area $\mathrm{d}\sigma_y$ seen from x defines the solid angle $\mathrm{d}\Omega$; the corresponding radiated power is $\mathrm{d}\Phi = L(x,u_{x,y}) \frac{\nu_x \cdot u_{x,y} \mathrm{d}\sigma_x \, \nu_y \cdot u_{y,x} \mathrm{d}\sigma_y}{|x-y|^2}$.

Radiative transfer

As the right of Figure B.5, consider now a second piece of surface, located at $y \in \mathbb{R}^3$, with unit normal $\nu_y \in \mathbb{S}^2$, and infinitesimal area $\mathrm{d}\sigma_y$. Consider the direction u from x to y, and the solid angle $\mathrm{d}\Omega$ of the surface $\mathrm{d}\sigma_y$ seen from x:

$$u = u_{x,y} = \tfrac{y-x}{|y-x|} = -u_{y,x}, \quad \mathrm{d}\Omega = \tfrac{\nu_y \cdot u_{y,x}}{|x-y|^2} \mathrm{d}\sigma_y,$$

where $|x-y|$ is the euclidean distance between x and y, and the projected area $\nu_y \cdot u_{y,x} \, \mathrm{d}\sigma_y$ represents the apparent area of $\mathrm{d}\sigma_y$ seen from x. By definition (B.2) of the radiance, we obtain the differential

form of the fundamental equation of radiative transfer [106]:

$$\mathrm{d}\Phi = L(x, u_{x,y}) \frac{\nu_x \cdot u_{x,y} \mathrm{d}\sigma_x \, \nu_y \cdot u_{y,x} \mathrm{d}\sigma_y}{|x-y|^2}. \tag{B.3}$$

Assuming that the medium between the two pieces of surface is transparent, with no absorption, the radiated power $\mathrm{d}\Phi$ is transferred from $\mathrm{d}\sigma_x$ to $\mathrm{d}\sigma_y$ along the direction $u_{x,y}$; this transfer expresses energy preservation. Therefore, $\mathrm{d}\Phi$ represents also the amount of incident light on $\mathrm{d}\sigma_y$, along the direction $u_{x,y}$.

Irradiance

The fundamental quantity concerning incident light is the *irradiance* $[\mathtt{W.m}^{-2}]$, which represents the amount of light incident on a surface. The irradiance is defined as the power received by a surface, along a certain direction, and per unit area. In Figure B.5, the irradiance $\mathrm{d}I(y, u_{x,y})$ received by y, from $\mathrm{d}\sigma_x$ in the direction $u_{x,y}$, is deduced from (B.3):

$$\mathrm{d}I(y, u_{x,y}) = \frac{\mathrm{d}\Phi}{\mathrm{d}\sigma_y} = L(x, u_{x,y}) \frac{\nu_x \cdot u_{x,y} \mathrm{d}\sigma_x \, \nu_y \cdot u_{y,x}}{|x-y|^2}. \tag{B.4}$$

Radiant incidance

Concerning the detectors, the basic measurement of a detector, such as a light sensor in a camera, is the received power (energy per time unit) [106]. The *radiant incidance* $[\mathtt{W.m}^{-2}]$, defined as the power received by unit area, can be deduced. The measure is especially related to the power radiated from the visible surfaces. For an ideal camera, the optical system completely transfers the received energy to the sensor, and the incidance on a pixel is mainly proportional to the radiance of the visible point [63].

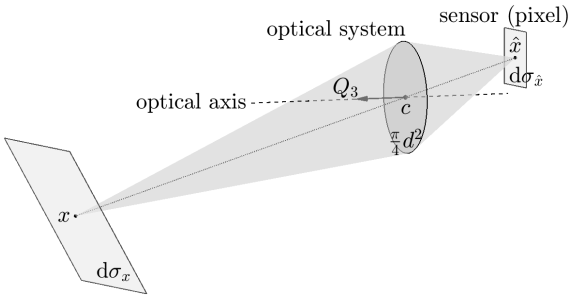


Figure B.6: Modeling of the radiant incidance on a pixel, by (B.5).

Indeed, consider a camera model as Figures B.3-B.4, and a light sensor associated to a pixel, as Figure B.6. The sensor, located at $\hat{x}$ in the focal plane, with area $\mathrm{d}\sigma_{\hat{x}}$, receives a power from a visible surface $\mathrm{d}\sigma_x$ around a point x. The surface $\mathrm{d}\sigma_x$ radiates a power $\mathrm{d}\Phi$ through the solid angle associated with the optical system seen from x. For a circular aperture with diameter $d \ll |x - c|$, assuming that the optical system completely transfers the received energy $\mathrm{d}\Phi$ to the sensor, the radiant incidance on the sensor is given after simplification by

$$\frac{\mathrm{d}\Phi}{\mathrm{d}\sigma_{\hat{x}}} = \frac{\pi}{4} \left(\frac{d}{f} \right)^2 (Q_3 \cdot u_{c,x})^4 \, L(x, u_{x,c}). \tag{B.5}$$

Up to some normalization, the measured incidance $\frac{\mathrm{d}\Phi}{\mathrm{d}\sigma_{\hat{x}}}$ coincides with the radiance $L(x, u_{x,c})$ of the visible point x, in the direction $u_{x,c}$ from x to the optical center c.

B.2.4 Reflection of light

Bidirectional Reflectance Distribution Function

VIS-NIR light often appears as being scattered from surfaces, due to the rugosity of the surfaces. This scattering is described from a macroscopic point of view by effective models. For opaque surfaces, usual models consider that light reflects off surfaces by the means of the *Bidirectional Reflectance Distribution Function* (BRDF). The BRDF is a surfacic function defined as the ratio of radiance to irradiance; it expresses how a surface reflects light toward a radiation direction when it is illuminated from an incident direction. More precisely, as in Figure B.7, assume that $\mathrm{d}I(y,u)$ is the incident irradiance received by the surfacic point $y \in \mathbb{R}^3$, along the incident direction $u \in \mathbb{S}^2$. Consider the radiance $\mathrm{d}L(y,v)$, radiated from y in the direction $v \in \mathbb{S}^2$, as a surfacic response to the incident irradiance $\mathrm{d}I(y,u)$. The BRDF $f(y,u,v)$ is such that

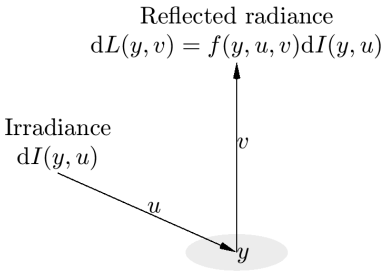


Figure B.7: BRDF $f(y,u,v)$.

$$\mathrm{d}L(y,v) = f(y,u,v)\,\mathrm{d}I(y,u). \tag{B.6}$$

The BRDF encodes the specular reflection on a perfect mirror by the means of a Dirac distribution; the corresponding direction is given by the Snell-Descartes law. On the contrary, an ideal matte surface, called a *Lambertian* surface, is encoded by a uniform BRDF, independent of the radiation angle v.

Rendering equation

The radiance satisfies an integral equation based on the BRDF. Assume that a set $S \subset \mathbb{R}^3$ represents the smooth boundary of a collection of opaque objects and consider the total radiance $L(y,v)$ at $y \in S$, in the direction $v \in \mathbb{S}^2$, as in Figure B.8. Firstly, the radiance emitted by a light source is represented by a source term $L_\epsilon(y,v)$; $L_\epsilon(y,v) = 0$ if y does not emit in the direction v. Secondly, introduce a visibility function,

$$\forall x,y \in S, \quad V(x,y) = \mathbb{1}_{]x,y[\cap S=\emptyset} = \begin{cases} 1, & \text{if } x \text{ is visible from } y, \\ 0, & \text{otherwise;} \end{cases}$$

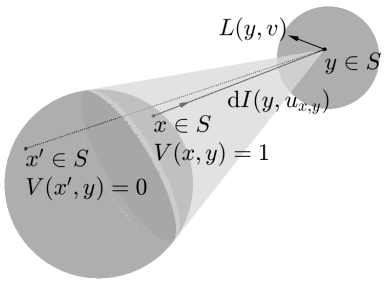


Figure B.8: A point $y \in S$ receives an irradiance $\mathrm{d}I(y,u_{x,y})$ from every visible point $x \in S$.

fix $x \in S$ visible from y, and the associated direction $u_{x,y} = \frac{y-x}{|y-x|}$. Then, y receives an irradiance $\mathrm{d}I(y,u_{x,y})$; hence, the surface reflects a radiance $f(y,u_{x,y},v)\mathrm{d}I(y,u_{x,y})$ in the direction v. Assuming linearity, the total radiance is modeled by the sum of the source term with such reflected radiances:

$$L(y,v) = L_\epsilon(y,v) + \int_{x\in S} V(x,y) f(y,u_{x,y},v)\,\mathrm{d}I(y,u_{x,y}).$$

By (B.4), we obtain finally an integral equation, the so-called *rendering equation* [65],

$$L(y,v) = L_\epsilon(y,v) + \int_{x\in S} V(x,y) f(y,u_{x,y},v) L(x,u_{x,y}) \frac{\nu_x\cdot u_{x,y}\,\nu_y\cdot u_{y,x}}{|x-y|^2}\,\mathrm{d}\sigma_x, \quad y \in S,\ v\cdot\nu_y \geq 0, \tag{B.7}$$

where $\nu_x \in \mathbb{S}^2$ denotes the exterior unit normal to S, at the point $x \in S$. In this expression, $\mathrm{d}\Omega_x = \frac{\nu_x\cdot u_{x,y}}{|x-y|^2}\mathrm{d}\sigma_x$ is a surface element on the hemisphere $u\cdot\nu_y \geq 0$ and represents the solid angle of $\mathrm{d}\sigma_x$ seen from y.

The rendering equation (B.7) is a model which emphasizes some mechanisms due to scattering of light by opaque objects, with extended light sources, in a transparent media. It does not take into account each possible effect of light propagation, such as transmission, subsurface diffusion, absorption, polarization, spectral dependency, and so on. For instance, we refer to [30,42,49,96] for effects concerning polarization.

Lambert's cosine law

The Lambert's cosine law

$$\rho(y)\cos\alpha \tag{B.8}$$

models uniform diffusion of light from an ideal matte opaque surface, called a *Lambertian reflector*. It is established as follows. Assume that the surface ∂D of a Lambertian reflector is illuminated by an isotropic point source located at z, as in Figure B.9. Denote by Φ the power of this source (in $\texttt{[W]}$). The associated power per unit solid angle is $\Phi/(4\pi)$ (in $\texttt{[W.sr}^{-1}\texttt{]}$). Therefore, an illuminated point $y \in \partial D$, with angle of incidence α, receives an irradiance $I(y,u) = \Phi\cos\alpha/(4\pi|z-y|^2)$ (in $\texttt{[W.m}^{-2}\texttt{]}$). By assumption, this incident irradiance is uniformly reflected off the surface: the BRDF $f(y,u,v)$ and the reflected radiance $L(y,v) = f(y,u,v)I(y,u)$ (in $\texttt{[W.m}^{-2}\texttt{.sr}^{-1}\texttt{]}$) do not depend on the radiation angle v. In this case, it can be shown that $f(y,u,v) = \frac{\rho(y,u)}{\pi}$, where $\rho(y,u) \in [0,1]$, called the *albedo*, is a dimensionless coefficient which represents the percentage of the incident irradiance which is reflected in any direction. Hence, the radiance is modeled by

$$L(y,v) = \frac{\Phi}{4\pi^2|z-y|^2}\rho(y)\cos\alpha, \quad (\cos\alpha = u_{y,z}\cdot\nu_y), \tag{B.9}$$

where we have further assumed that $\rho(y,u) = \rho(y)$ does not depend on the incident angle, as many models in computer vision. Assuming that z in far field with $|z-y| \approx R$ a large constant, we obtain that the radiance is given by (B.8), up to a constant factor.

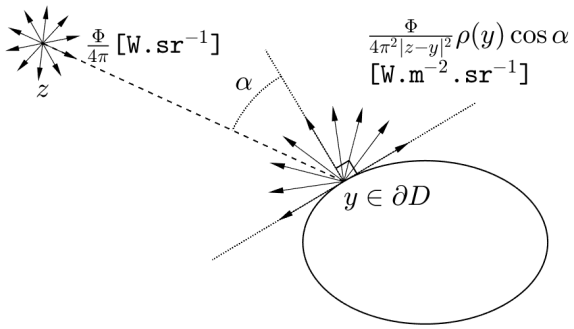


Figure B.9: Diffuse reflection by the Lambert's cosine law. Here, an isotropic point source z emits light with a power per unit solid angle $\Phi/(4\pi)$. On the surface ∂D, an illuminated point y reflects light uniformly; for an angle of incidence α, the point y reflects a radiance $\Phi/(4\pi^2|z-y|^2)\,\rho(y)\cos\alpha$ in any direction above the tangent plane. The dimensionless coefficient $\rho(y) \in [0,1]$ is the albedo, defined as the percentage of incident irradiance which is reflected.

B.3 X-ray tomography

B.3.1 Beer-Lambert's law

In transmission tomography, one usually probes a medium using X-rays [22, 78]. Assuming that the wavelengths of X-rays are very short on the scale of variation of the probed medium[1], the

[1] This is the case for a human body.

propagation can be described by geometrical optics [39]. More precisely, an X-ray is transmitted through the medium along a straight line. Along this ray, the intensity I is attenuated due to absorption. The variation $\mathrm{d}I(s)$ of the intensity, between the curvilinear abscissae s and $s+\mathrm{d}s$, is modeled by $\mathrm{d}I(s) = -a(s)I(s)\mathrm{d}s$; the inverse length $a(s)$ represents an attenuation coefficient of the traversed medium. Therefore, if an X-ray source, located at position x_0, emits towards a sensor, located at position x_1, then the received intensity I_1 is related to the emitted intensity I_0, by the Beer-Lambert's law:

$$I_1 = I_0 \exp\left(-\int_{[x_0,x_1]} a(x)\mathrm{d}\ell\right).$$

The integral $\int_{[x_0,x_1]} a(x)\mathrm{d}\ell$, where $\mathrm{d}\ell$ is the length measure, represents the total attenuation of the medium between the source x_0 and the receiver x_1. The value of this integral is considered to be the measurement associated to the ray $[x_0, x_1]$, since it is deduced from the knowledge of I_1 and I_0:

$$\int_{[x_0,x_1]} a(x)\mathrm{d}\ell = \log \tfrac{I_0}{I_1}. \tag{B.10}$$

B.3.2 Cone beam computed tomography

In practice, ray integrals such as (B.10) are collected for several rays, associated to various positions of the source and the sensors. The cone beam projection is among the most famous geometry of acquisition. As displayed in Figure B.10, it is a perspective projection, obtained for a conic beam of X-rays emanating from a fixed position; the object is located between the source and a "screen" of sensors. The resulting image is a radiography, where each pixel contains the ray integral (B.10) between the source and the sensor associated to the pixel. The contrasts in the image are due to variations in the attenuation profiles, from a ray to another one. Next, this kind of cone beam radiography can be taken under several angles of view, for instance by rotation of the device source-screen around a fixed axis, as it is performed in Figure 1.2. This is the principle of a cone beam scan.

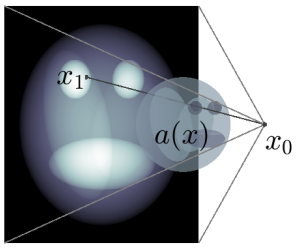


Figure B.10: Cone beam radiography.

Computed tomography aims at computing a 3D reconstruction of the attenuation $a(x)$, from the measurement of line integrals (B.10). The computed spatial attenuation is used to represent the scene; in this way, materials with different attenuations can be separated.

B.3.3 Mathematical framework

From a mathematical point of view, the *X-ray transform* of $a : x \in \mathbb{R}^3 \mapsto a(x) \in \mathbb{R}$ is defined by

$$\begin{array}{rccl} \mathcal{X}[a]: & \mathbb{R}^3 \times \mathbb{S}^2 & \to & \mathbb{R} \\ & (x_0, u) & \mapsto & \mathcal{X}[a](x_0,u) = \int_{\mathbb{R}} a(x_0+tu)\mathrm{d}t. \end{array} \tag{B.11}$$

Analogously to the ray integral (B.10), $\mathcal{X}[a](x_0,u)$ is a line integral, along the ray through x_0 and oriented by u; under obvious assumption, the ray integral (B.10) coincides exactly with $\mathcal{X}[a](x_0, \frac{x_1-x_0}{|x_1-x_0|})$. If the rays are in a plane, the X-ray transform coincides with the Radon transform (A.1); for instance, in the plane $x_3 = 0$,

$$\mathcal{R}[a](\theta, s) = \mathcal{X}[a]((s\theta, 0), (\theta^\perp, 0)), \quad \theta \in \mathbb{S}^1,\ s \in \mathbb{R}.$$

Finally, the mathematical problem of X-rays tomography deals with inverting the X-ray transform; the goal is to compute a function a from the knowledge of $\mathcal{X}[a](x_0,u)$, for a collection of rays (x_0,u). In a plane, this problem consists in inverting the Radon transform; see Appendix A.

Bibliography of Part I

[22] H. Ammari, *An Introduction to Mathematics of Emerging Biomedical Imaging*, Springer-Verlag Berlin Heidelberg, 2008.

[23] H. Andrade-Loarca, G. Kutyniok, O. Öktem, and P. Petersen, *Deep microlocal reconstruction for limited-angle tomography*, Applied and Computational Harmonic Analysis, 59 (2022), pp. 155–197. Special Issue on Harmonic Analysis and Machine Learning.

[24] H. Andrade-Loarca, G. Kutyniok, O. Öktem, and P. C. Petersen, *Extraction of Digital Wavefront Sets Using Applied Harmonic Analysis and Deep Neural Networks*, SIAM Journal on Imaging Sciences, 12 (2019), pp. 1936–1966.

[25] C.-A. Azencott, *Introduction au Machine Learning*, Dunod, 2018.

[26] B. G. Baumgart, *Geometric modeling for computer vision*, PhD thesis, Stanford Univ Ca Dept of Computer Science, 1974.

[27] J.-C. Bazin, O. Bogdan, V. Eckstein, and F. Rameau, *Camera calibration and apparatus based on deep learning.* Priority: KR10-2018-0090565, 2018–08–03. Publications: KR102227583B1, 2021. US10977831B2, 2021.

[28] I. Berechet and G. Berginc, *Advanced algorithms for identifying targets from a three-dimensional reconstruction of sparse 3D Ladar data*, in Proc. SPIE 8172, Optical Complex Systems: OCS11, G. Berginc, ed., 2011.

[29] I. Berechet, G. Berginc, and S. Berechet, *Scattering computation for 3D laser imagery and reconstruction algorithms*, in Proc. SPIE 8495, Reflection, Scattering, and Diffraction from Surfaces III, L. M. Hanssen, ed., 2012.

[30] G. Berginc, *Scattering models for range profiling and 2D-3D laser imagery*, in Proc. SPIE 9205, Reflection, Scattering, and Diffraction from Surfaces IV, L. M. Hanssen, ed., 2014.

[31] G. Berginc, I. Berechet, and S. Berechet, *Data-driving Algorithms for 3D Reconstruction from Ladar Data*, in PIERS Proceedings, Moscow, Russia, 2012, pp. 496–499.

[32] G. Berginc and M. Jouffroy, *Optronic system and method dedicated to identification for formulating three-dimensional images.* Priority: FR0905720, 2009-11-27. Publications: FR2953313B1, EP2333481B1, US8836762B2, JP5891560B2, CA2721891C.

[33] G. Berginc and M. Jouffroy, *Simulation of 3D laser systems*, in 2009 IEEE International Geoscience and Remote Sensing Symposium, vol. 2, 2009, pp. 440–443.

[34] G. Berginc and M. Jouffroy, *Simulation of 3D laser imaging*, PIERS Online, 6 (2010), pp. 415–419.

[35] ——, *3D laser imaging*, PIERS Online, 7 (2011), pp. 411–415.

[36] A. Biguri, M. Dosanjh, S. Hancock, and M. Soleimani, *TIGRE: a MATLAB-GPU toolbox for CBCT image reconstruction*, Biomedical Physics & Engineering Express, 2 (2016), p. 055010.

[37] N. Bleistein and R. A. Handelsman, *Asymptotic Expansions of Integrals*, Dover, 1986.

[38] L. Borg, J. Frikel, J. S. Jørgensen, and E. T. Quinto, *Analyzing Reconstruction Artifacts from Arbitrary Incomplete X-ray CT Data*, SIAM Journal on Imaging Sciences, 11 (2018), pp. 2786–2814.

[39] M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*, Cambridge University Press, 7th ed., 1999.

[40] J. Chen, H. Sun, Y. Zhao, and C. Shan, *Typical influencing factors analysis of laser reflection tomography imaging*, Optik, 189 (2019).

[41] M. Cheney and B. Borden, *Synthetic Aperture Radar Imaging*, in Handbook of Mathematical Methods in Imaging, O. Scherzer, ed., Springer-Verlag New York, 2011, ch. 15, pp. 655–690.

[42] R. A. Chipman, *Polarimetry*, in Handbook of optics, Vol. 2: Devices, Measurements, and Properties, M. Bass, ed., McGRAW-HILL, 1994, ch. 22.

[43] S. R. Deans, *The Radon Transform and Some of Its Applications*, Dover Publications, 2007.

[44] R. O. Duda and P. E. Hart, *Use of the Hough transformation to detect lines and curves in pictures*, Communications of the ACM, 15 (1972), pp. 11–15.

[45] J.-D. Durou, *Reconstruction 3D à partir des ombrages*, in Problèmes inverses en imagerie et en vision 2, A. Mohammad-Djafari, ed., Lavoisier, 2009, ch. 10.

[46] T. Elfving, P. C. Hansen, and T. Nikazad, *Semi-convergence properties of Kaczmarz's method*, Inverse Problems, 30 (2014), p. 055007.

[47] A. Faridani, E. L. Ritman, and K. T. Smith, *Local tomography*, SIAM Journal on Applied Mathematics, 52 (1992), pp. 459–484.

[48] L. Feldkamp, L. Davis, and J. Kress, *Practical cone-beam algorithm*, JOSA A, 1 (1984), pp. 612–619.

[49] D. S. Flynn and C. Alexander, *Polarized surface scattering expressed in terms of a bidirectional reflectance distribution function matrix*, Optical Engineering, 34 (1995), pp. 1646–1651.

[50] J. Frikel and E. T. Quinto, *Characterization and reduction of artifacts in limited angle tomography*, Inverse Problems, 29 (2013), p. 125007.

[51] R. J. Gardner, *Geometric tomography*, Notices Amer. Math. Soc. 42, 4 (1995), pp. 422–429.

[52] D. T. Gering and W. M. Wells, *Object Modeling using Tomography and Photography*, in Proc. of IEEE Workshop on Multi-View Modeling and Analysis of Visual Scenes, 1999, pp. 11–18.

[53] H. Gouraud, *Continuous Shading of Curved Surfaces*, IEEE Transactions on Computers, C-20 (1971), pp. 623–629.

[54] A. Grigis and J. Sjöstrand, *Microlocal analysis for differential operators: an introduction*, vol. 196, Cambridge University Press, 1994.

[55] V. Guillemin and S. Sternberg, *Geometric asymptotics*, vol. 14 of Mathematical Surveys and Monographs, American Mathematical Society, 1977.

[56] P. C. Hansen and M. Saxild-Hansen, *AIR Tools - A MATLAB package of algebraic iterative reconstruction methods*, Journal of Computational and Applied Mathematics, 236 (2012), pp. 2167–2178. Inverse Problems: Computation and Applications.

[57] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2nd ed., 2004.

[58] S. Helgason, *The Radon Transform*, Springer Science & Business Media, second ed., 1999.

[59] M. Henriksson, T. Olofsson, C. Grönwall, C. Bränndlund, and L. Sjöqvist, *Optical reflectance tomography using TCSPC laser radar*, in Proc. SPIE 8542, Electro-Optical Remote Sensing, Photonic Technologies, and Applications VI, 2012, pp. 105–113.

[60] G. T. Herman, *Tomography*, in Handbook of Mathematical Methods in Imaging, O. Scherzer, ed., Springer-Verlag New York, 2011, ch. 16, pp. 691–733.

[61] L. Hörmander, *Fourier integral operators. i*, Acta mathematica, 127 (1971), pp. 79–183.

[62] B. Horn, *Robot vision*, MIT press, 1986.

[63] B. K. Horn and R. W. Sjoberg, *Calculating the reflectance map*, Applied optics, 18 (1979), pp. 1770–1779.

[64] A. J. Johnson, D. L. Marks, R. A. Stack, D. J. Brady, and D. C. Munson, *Three-dimensional surface reconstruction of optical Lambertian objects using cone-beam tomography*, in Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348), vol. 2, IEEE, 1999, pp. 663–667.

[65] J. T. Kajiya, *The Rendering Equation*, in Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '86, ACM, 1986, p. 143–150.

[66] B. Kaltenbacher, T. Schuster, and A. Wald, eds., *Time-dependent Problems in Imaging and Parameter Identification*, Springer, 2021.

[67] F. Knight, D. Klick, D. Ryan-Howard, J. Theriault, B. Tussey, and A. Beckman, *Two-dimensional tomographs using range measurements*, in Laser radar III, vol. 999, SPIE, 1989, pp. 269–281.

[68] F. Knight, S. Kulkarni, R. Marino, and J. Parker, *Tomographic Techniques Applied to Laser Radar Reflective Measurements*, Lincoln Laboratory Journal, 2 (1989), pp. 143–160.

[69] V. P. Krishnan and E. T. Quinto, *Microlocal Analysis in Tomography*, in Handbook of Mathematical Methods in Imaging, O. Scherzer, ed., Springer, New York, 2015, pp. 847–902.

[70] P. Kuchment, K. Lancaster, and L. Mogilevskaya, *On local tomography*, Inverse Problems, 11 (1995), pp. 571–589.

[71] J. B. Lasche, C. L. Matson, S. D. Ford, W. L. Thweatt, K. B. Rowland, and V. N. Benham, *Reflective tomography for imaging satellites: experimental results*, in Proc. SPIE 3815, Digital Image Recovery and Synthesis IV, 1999, pp. 178–189.

[72] A. Laurentini, *The visual hull concept for silhouette-based image understanding*, IEEE Transactions on pattern analysis and machine intelligence, 16 (1994), pp. 150–162.

[73] Y. Ma, S. Soatto, J. Košecká, and S. S. Sastry, *An invitation to 3-D Vision*, vol. 26, Springer-Verlag New York, 2004.

[74] E. P. Magee, C. L. Matson, and D. Stone, *Comparison of techniques for image reconstruction using reflective tomography*, in Proc. SPIE 2302, Image Reconstruction and Restoration, T. J. Schulz and D. L. Snyder, eds., 1994, pp. 95–102.

[75] C. L. Matson, D. E. Holland, D. F. Pierrottet, D. Ruffatto, S. R. Czyzak, and D. E. Mosley, *Satellite feature reconstruction using reflective tomography: field results*, in Proc. SPIE 3219, Optics in Atmospheric Propagation and Adaptive Systems II, 1998, pp. 65–73.

[76] C. L. Matson and D. E. Mosley, *Reflective tomography reconstruction of satellite features - field results*, Applied Optics, 40 (2001), pp. 2290–2296.

[77] R. B. Melrose and G. Uhlmann, *An introduction to microlocal analysis*, Department of Mathematics, Massachusetts Institute of Technology, 2008. `http://math.mit.edu/~rbm/books/imaast.pdf`.

[78] F. Natterer and F. Wübbeling, *Mathematical methods in image reconstruction*, SIAM, 2001.

[79] M. K. Nguyen and T. T. Truong, *Inversion of a new circular-arc Radon transform for Compton scattering tomography*, Inverse Problems, 26 (2010), p. 065005.

[80] F. Odille, *Motion-corrected reconstruction*, in Magnetic Resonance Image Reconstruction: Theory, Methods, and Applications, M. Akcakaya, M. Doneva, and P. C., eds., Elsevier, 2022, ch. 13, p. 355–389.

[81] Y. Okitsu, F. Ino, and K. Hagihara, *High-performance cone beam reconstruction using cuda compatible gpus*, Parallel Computing, 36 (2010), pp. 129–141.

[82] J. K. Parker, E. Craig, D. Klick, F. Knight, S. Kulkarni, R. Marino, J. Senning, and B. Tussey, *Reflective tomography: images from range-resolved laser radar measurements*, Applied optics, 27 (1988), pp. 2642–2643.

[83] G. Peyré, *Toolbox graph.* `https://www.mathworks.com/matlabcentral/fileexchange/5355-toolbox-graph`, MATLAB Central File Exchange, 2008.

[84] E. T. Quinto, *The dependence of the generalized radon transform on defining measures*, Transactions of the American Mathematical Society, 257 (1980), pp. 331–346.

[85] ——, *Tomographic reconstructions from incomplete data-numerical inversion of the exterior Radon transform*, Inverse Problems, 4 (1988), pp. 867–876.

[86] ——, *Singularities of the X-ray transform and limited data tomography in R^2 and R^3*, SIAM Journal on Mathematical Analysis, 24 (1993), pp. 1215–1225.

[87] ——, *An introduction to X-ray tomography and Radon transforms*, in Proceedings of symposia in Applied Mathematics, vol. 63, 2006, p. 1.

[88] A. G. Ramm and A. I. Katsevich, *The Radon transform and local tomography*, CRC press, 1996.

[89] D. Riabkov, X. Xue, D. Tubbs, and A. Cheryauka, *Accelerated cone-beam backprojection using GPU-CPU hardware*, in 9th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine, Citeseer, 2007, pp. 68–71.

[90] G. Rigaud and B. N. Hahn, *3D Compton scattering imaging and contour reconstruction for a class of Radon transforms*, Inverse Problems, 34 (2018), p. 075004.

[91] S. Rit, M. V. Oliva, S. Brousmiche, R. Labarbe, D. Sarrut, and G. C. Sharp, *The Reconstruction Toolkit (RTK), an open-source cone-beam CT reconstruction toolkit based on the Insight Toolkit (ITK)*, in Journal of Physics: Conference Series, vol. 489, IOP Publishing, 2014, p. 012079.

[92] J. Sanders and E. Kandrot, *CUDA by Example: An Introduction to General-Purpose GPU Programming*, Addison-Wesley Professional, 2010.

[93] H. Scherl, B. Keck, M. Kowarschik, and J. Hornegger, *Fast GPU-based CT reconstruction using the common unified device architecture (CUDA)*, in 2007 IEEE Nuclear Science Symposium Conference Record, vol. 6, 2007, pp. 4464–4466.

[94] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, *Multi-view stereo evaluation web page*, `http://vision.middlebury.edu/mview`, (2006).

[95] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, *A comparison and evaluation of multi-view stereo reconstruction algorithms*, in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, IEEE, 2006, pp. 519–528.

[96] J. R. I. Shell, *Polarimetric remote sensing in the visible to near infrared*, PhD thesis, Rochester Institute of Technology, 2005.

[97] K. T. Smith and F. Keinert, *Mathematical foundations of computed tomography*, Applied Optics, 24 (1985), pp. 3950–3957.

[98] R. Szeliski, *Computer vision: algorithms and applications*, Springer-Verlag London, 2011.

[99] F. Trèves, *Introduction to pseudodifferential and Fourier integral operators*, vol. 2: Fourier integral operators of The University Series in Mathematics, Plenum Press, New York, 1980.

[100] G. Turk and M. Levoy, *Zippered polygon meshes from range images*, in Proceedings of the 21st annual conference on Computer graphics and interactive techniques, ACM, 1994, pp. 311–318.

[101] J. Wallis and T. Miller, *Three-Dimensional Display in Nuclear Medicine and Radiology*, The Journal of Nuclear Medicine, (1991), pp. 534–546.

[102] J.-c. Wang, S.-w. Zhou, L. Shi, Y.-H. Hu, and Y. Wang, *Image quality analysis and improvement of ladar reflective tomography for space object recognition*, Optics Communications, 359 (2016), pp. 177–183.

[103] Y. Wang, G. Yan, H. Zhu, S. Buch, Y. Wang, E. M. Haacke, J. Hua, and Z. Zhong, *JointVesselNet: Joint Volume-Projection Convolutional Embedding Networks for 3D Cerebrovascular Segmentation*, in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI, Berlin, Heidelberg, 2020, Springer-Verlag, p. 106–116.

[104] ——, *VC-Net: Deep Volume-Composition Networks for Segmentation and Visualization of Highly Sparse and Noisy Image Data*, IEEE Transactions on Visualization and Computer Graphics, 27 (2021), pp. 1301–1311.

[105] J. W. Webber and E. T. Quinto, *Microlocal Analysis of Generalized Radon Transforms from Scattering Tomography*, SIAM Journal on Imaging Sciences, 14 (2021), pp. 976–1003.

[106] W. L. Wolfe, *Introduction to radiometry*, vol. TT29, SPIE Optical Engineering Press, 1998.

Part II

Mathematical and numerical aspects of spectral computing on the Cubed Sphere

Chapter 5

Symmetry group of the Cubed Sphere

5.1 Equiangular Cubed Sphere

Various fields in computational physics, for instance climatology modelling [180], involve numerical computations on the sphere. This includes the use of spherical grids [181]. The grids obtained by radial projection of a circumscribed cube on the sphere, as in Figure 5.1, are among the most employed. These *Cubed Sphere grids* have been originally introduced in [171], and further studied, for example in [153, 158–160, 164, 165, 168]. A wide variety of numerical methods have been successfully adapted to Cubed Sphere grids, *e.g.* in [115, 116, 118, 120, 127, 140, 142, 143, 151, 154, 161, 169, 176, 179] and the references therein.

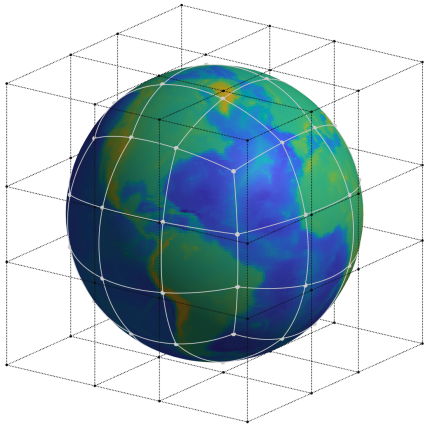


Figure 5.1: Construction of a Cubed Sphere grid, by radial projection from a circumscribed cube onto a sphere. Some cartesian grid (black dots) is defined on the faces of the cube; the Cubed Sphere grid (white dots) is defined as the radial projection of this cartesian grid, on the sphere. This construction meshes the sphere with arcs of great circles: cartesian grid lines (dotted straight lines) are projected from the cube onto arcs of great circle (white arcs). The Cubed Sphere displayed here is the equiangular Cubed Sphere CS_3 (5.1).

In this thesis, we focus on a Cubed Sphere structured by equiangular great circles: the *equiangular Cubed Sphere* $\mathrm{CS}_N \subset \mathbb{S}^2$, [115, 160], with resolution parameter $N \geq 1$, defined by

$$\mathrm{CS}_N := \Big\{ \rho(\pm 1, u, v),\ \rho(u, \pm 1, v),\ \rho(u, v, \pm 1);\ u = \tan \tfrac{i\pi}{2N},\ v = \tan \tfrac{j\pi}{2N},\ -\tfrac{N}{2} \leq i, j \leq \tfrac{N}{2} \Big\}, \qquad (5.1)$$

where $\rho(x) = \frac{x}{\|x\|}$ denotes the radial projection, from the faces of the cube $[-1, 1]^3$ onto the sphere $\mathbb{S}^2$. As suggested by Figure 5.1, where CS_3 is displayed, the Cubed Sphere CS_N is quasi-uniform, is not polarized along a specific axis, and is shaped by the cube, including discontinuities accross

"edges" (radial projection of the edges of the circumscribed cube). Furthermore, some symmetry properties, such as invariance under permutation of the cartesian coordinates, are noticed.

The success of the grid CS_N may be explained by mathematical properties, such as metric properties (quasi-uniformity) and symmetry (rotational invariance); we refer for instance to [115, 156, 157, 168], where the symmetry has been used. That is the reason why we have clearly identified the shortest geodesic arcs and the symmetry group of CS_N in [4]. Such results deepen the mathematical knowledge of CS_N and provide a valuable mathematical background for the applications; in particular, knowing the symmetry group of a grid supports the design of numerical schemes, such as interpolation methods [166], quadrature rules [175], or Discrete Fourier Transforms [155].

In the rest of this chapter, we summarize some mathematical results from [4]. We refer to the original text for some (tedious) proofs.

5.2 Symmetry group of the Cubed Sphere

We make explicit the *symmetry*, or *rotational invariance*, of the Cubed Sphere, by means of its symmetry group [112], defined as follows.

Definition 5.1. The symmetry group of a set $E \subset \mathbb{R}^3$ is the group $\mathcal{G}$ of all orthogonal matrices that leave E invariant:

$$\mathcal{G} = \left\{ Q \in \mathbb{R}^{3\times 3} : Q^\intercal Q = Q Q^\intercal = \mathrm{I}_3 \text{ and } QE = \{Qu,\, u \in E\} = E \right\}.$$

For $N = 1$, $\mathrm{CS}_1 = \{-1/\sqrt{3}, 1/\sqrt{3}\}^3$ is a scaling of $\{-1,1\}^3$. Therefore, it is clear that the symmetry group of CS_1 coincides with the symmetry group of the cube $\{-1,1\}^3$. This group is well known: it is isomorphic to the group $\mathfrak{S}_4 \times \mathbb{Z}/2\mathbb{Z}$ [112, pp. 37,38,55]; any symmetry of the cube is indeed identified with a permutation of the four principal diagonals, combined with a toggle for inversion of the cube, or not. For completeness, a matrix description is recalled hereafter.

Lemma 5.2 (Octahedral group)**.** *The symmetry group of the cube $\{-1,1\}^3$ coincides with the symmetry group the octahedron $\{(1,0,0),(0,1,0),(0,0,1),(-1,0,0),(0,-1,0),(0,0,-1)\}$, given by*

$$\mathcal{G} = \left\{ \begin{bmatrix} \epsilon_1 e_{\sigma_1} & \epsilon_2 e_{\sigma_2} & \epsilon_3 e_{\sigma_3} \end{bmatrix},\ \sigma \in \mathfrak{S}_3,\ \epsilon \in \{-1,1\}^3 \right\}, \tag{5.2}$$

where $e_1 = (1,0,0)$, $e_2 = (0,1,0)$, $e_3 = (0,0,1)$, and $\mathfrak{S}_3$ denotes the group of all permutations of $\{1,2,3\}$.

In fact, the octahedral group (5.2) determines the group of any equiangular Cubed Sphere, as stated by the main result of [4].

Theorem 5.3 (Symmetry group of the equiangular Cubed Sphere)**.** *Let $N \geq 1$. The symmetry group of the Cubed Sphere CS_N coincides with the symmetry group $\mathcal{G}$ of the cube $\{-1,1\}^3$. In other words, an orthogonal matrix Q leaves CS_N invariant if, and only if, it leaves $\{-1,1\}^3$ invariant.*

The combination of (5.1) with (5.2) shows relatively easily that any symmetry of the cube, $Q \in \mathcal{G}$, leaves the Cubed Sphere CS_N invariant. Therefore, the most difficult part of the theorem is the converse. This part can be deduced from a classification of finite subgroups of orthogonal groups, such as [112, Theorem 19.2]. Another approach consists in introducing some geometrical pattern, whose symmetry group is $\mathcal{G}$, and which is left invariant by any symmetry of CS_N. We prove in [4] that the *cuboctahedron*

$$\Omega := \{(0,\epsilon,\eta),(\epsilon,0,\eta),(\epsilon,\eta,0), \epsilon = \pm 1, \eta = \pm 1\} \tag{5.3}$$

is such a pattern. The proof is based on the shortest geodesic arcs, described in the next section.

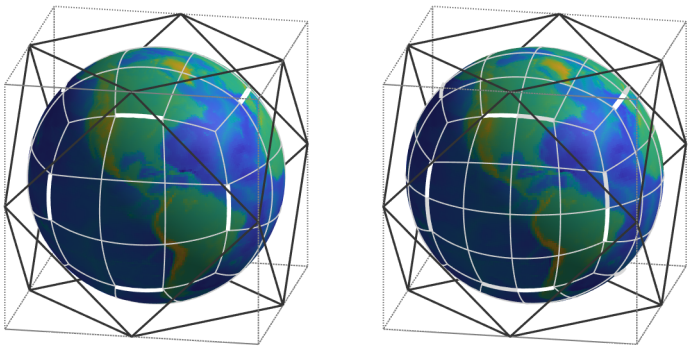


Figure 5.2: Shortest geodesic arcs on CS_N, described by Theorem 5.4. Here, minimal arcs on CS_N, solution to (5.4), are displayed in bold, around midpoints on edges. Left: N is odd ($N = 3$); right: N is even ($N = 4$). The location of the minimal arcs match with the vertices (5.3) of the cuboctahedron displayed with a black line.

5.3 Shortest geodesic distance on the Cubed Sphere

In this section, we consider the shortest geodesic arcs on the Cubed Sphere CS_N, defined as the solutions to the problem

$$\min\{\arccos u \cdot v; u \in \mathrm{CS}_N, v \in \mathrm{CS}_N, u \neq v\}. \tag{5.4}$$

This problem deals with metric properties of a widely used grid, so it has its own interest. One original feature of (5.4) is that u and v are allowed to belong to distinct grid lines, or even distinct panels; in particular, any spherical diagonal of the mesh is realizable. The main difficulty of problem (5.4) comes from this specificity, and this is somehow the lock of Theorem 5.3.

The problem (5.4) has been solved in [4], with the following theorem (see Figure 5.2).

Theorem 5.4 (Shortest geodesic arcs on the equiangular Cubed Sphere). *Let $N \geq 1$, and consider the problem* (5.4) *of the minimal arc-length between separate points on* CS_N.

(i) If N is odd, there are precisely 12 minimal arcs on CS_N *(one per edge):*

$$\{\rho(-\delta,\epsilon_1,\epsilon_2),\rho(\delta,\epsilon_1,\epsilon_2)\},\ \{\rho(\epsilon_1,-\delta,\epsilon_2),\rho(\epsilon_1,\delta,\epsilon_2)\},\ \{\rho(\epsilon_1,\epsilon_2,-\delta),\rho(\epsilon_1,\epsilon_2,\delta)\},$$
$$\delta = \tan\tfrac{\pi}{4N},\ \epsilon_1 = \pm 1,\ \epsilon_2 = \pm 1.$$

(ii) If N is even, there are precisely 24 minimal arcs on CS_N *(two per edge):*

$$\{\rho(0,\epsilon_1,\epsilon_2),\rho(\delta,\epsilon_1,\epsilon_2)\},\ \{\rho(\epsilon_1,0,\epsilon_2),\rho(\epsilon_1,\delta,\epsilon_2)\},\ \{\rho(\epsilon_1,\epsilon_2,0),\rho(\epsilon_1,\epsilon_2,\delta)\},$$
$$\delta = \pm\tan\tfrac{\pi}{2N},\ \epsilon_1 = \pm 1,\ \epsilon_2 = \pm 1.$$

Theorem 5.4 is proved in [4]; some algorithm decreases (if possible) the distance $\arccos u \cdot v$, from any inital arc $\{u, v\} \subset \mathrm{CS}_N$. The proof is tedious, because many cases must be considered (u, v can be on distinct panels, on the same panel, on a common grid line, or not, and so on). Some cases are easy; some other ones are more difficult. To mention two cases, the easiest case deals with an arc along an edge, with N even, $u = \rho(1, 0, 1)$, $v = \rho(1, Z, 1)$, and $Z > 0$, whereas one of the most difficult one deals with a diagonal, $u = \rho(1, X, Y)$, $v = \rho(1, Z, T)$, $0 \leq X < Z \leq 1$, $0 \leq Y < T \leq 1$.

To finish with, the minimal arcs are the "short arcs around the midpoints on the edges", as displayed in Figure 5.2. Therefore, their location matches with the cuboctahedron Ω defined in (5.3). Then, it can be proved that Ω is invariant under the group of CS_N, which finally implies Theorem 5.3.

5.4 Conclusion and perspectives

The symmetry group of the equiangular Cubed Sphere coincides with the symmetry group of the cube. The proposed approach to prove this result studies geodesic distances between points of the grid. Such results provide some theoretical foundation for numerical computation on the Cubed Sphere.

The symmetry group of a grid plays a central role in several contexts. It can be used to build spherical quadrature rules which are valid for as many spherical harmonics as possible [175]. Our main result shows that the group of the cube is the suitable symmetry group for the determination of quadrature weights on the Cubed Sphere. This background somehow supports the quadrature rule presented in Chapter 7. Moreover, for quadrature rules, the geometric distribution of the nodes is often examinated. Our study of the geodesic distance includes the theoretical value of the separation distance, and could serve as a tool to quantify the "uniformity" of the Cubed Sphere grid.

Another subject of interest concerns the building of a discrete Fourier analysis on the Cubed Sphere, based on the invariance under the action of the symmetry group, in the spirit of [155]. Here again, our result is a first step in this direction, since it identifies the group to be considered.

Chapter 6

Interpolation on the Cubed Sphere with spherical harmonics

6.1 Introduction

In this chapter, we consider the problem of Lagrange interpolation on the Cubed Sphere, as in [10].

Problem (Lagrange interpolation on the Cubed Sphere)**.** Let $\mathrm{CS}_N = \{x_i,\, 1 \leq i \leq \bar{N}\}$ denote the Cubed Sphere grid (5.1), where $N \geq 1$ is fixed, and $\bar{N} = 6N^2 + 2$ denotes the cardinal number. Assume that a grid function $f \in \mathbb{R}^{\mathrm{CS}_N}$ is known, which means that $f(x_i)$, $1 \leq i \leq \bar{N}$, are values given at the nodes of CS_N. The problem of Lagrange interpolation of f consists in finding a spherical function $u : \mathbb{S}^2 \to \mathbb{R}$ such that

$$u(x_i) = f(x_i),\, 1 \leq i \leq \bar{N}. \tag{6.1}$$

We define a new subspace of spherical harmonics in which such a problem has a unique solution u. This subspace is such that the solution u is minimal with respect to a reverse lexicographical order (on the degree). This implies that the components of u with high degrees are as small as possible, which avoids the high-frequency oscillations as much as possible. Our method of construction is based on an unusual factorization algorithm.

Other methods for multivariate interpolation have already been introduced to build *minimal degree interpolation spaces* enjoying various properties; we refer for instance to [122,123,166,167,172, 173] and the references therein. But the minimal property of our interpolating function, mentioned above, seems new. Also, we can mention the most standard approach in inverse problems: it would tackle (6.1) with a generalized inverse to define a solution which has the minimal norm (based on a Singular Value Decomposition - SVD), but would not constrain the high degree components.

In fact, our approach, introduced in [10], is some combination of minimal degree interpolation and generalized inversion. In our case, the method deals especially with sampling on the Cubed Sphere: "undersampled" spherical harmonics are determined and eliminated, since they cannot be reconstructed. This is achieved by a reduction of a Vandermonde matrix associated to (6.1), under some special echelon form. The reduction is based on an orthogonal factorization, deduced from the SVD of suitable matrices. It builds simultaneously an orthonormal basis of the desired interpolation space, an orthonormal basis of undersampled spaces (orthogonal to the interpolation space), and a QR factorization of the linear system associated to the Lagrange problem (in the interpolation space).

The chapter is organized as follows. In Section 6.2, we define some notation about spherical harmonics and grid functions. In Section 6.3, we propose an algebraic definition of an interpolation space, suitable for the Lagrange problem on the Cubed Sphere (Theorem 6.2). This section includes Lemma 6.1, which defines an explicit interpolating spherical harmonics with degree at most $4N - 1$ (or $4N - 2$); this new result, based on the geometrical structure of CS_N, has never been published. In Section 6.4, we formulate an algorithm to compute this space, and to solve (6.1) (Corollary 6.8),

as in [10]. We also mention a novel characterization of the solution: it is the minimal interpolation spherical harmonics with respect to a reverse lexicographical order (Corollary 6.9). In Section 6.5, we propose some numerical study. We indicate some empirical structure of the interpolation space, we examinate the distance of the Legendre basis functions to this space, and we interpolate various test functions with our scheme. In Appendix 6.A, we propose and we study an orthogonal factorization of a block matrix, under some special echelon form (Theorem 6.11). We prove that is suitable to compute the least squares approximation which is minimal for a reverse lexicographical order (Corollary 6.13). This framework is general and can be used for any block least squares problem for which some lexicographical order on the blocks is desired. In fact, similar algorithms have been developed in robotics to solve lexicographical least-squares [126].

6.2 Background and notation

6.2.1 Spherical harmonics

On the unit sphere $\mathbb{S}^2 = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_1^2 + x_2^2 + x_3^2 = 1\}$, the spherical coordinates are given by

$$x(\theta, \phi) = (\cos\theta\cos\phi, \cos\theta\sin\phi, \sin\theta) \in \mathbb{S}^2, \quad \theta \in [-\tfrac{\pi}{2}, \tfrac{\pi}{2}], \quad \phi \in \mathbb{R}, \tag{6.2}$$

where θ is the latitude and ϕ is the longitude. In these coordinates, the real Legendre spherical harmonics of degree $n \geq 0$ are defined by

$$Y_n^m(x(\theta,\phi)) = \sqrt{\tfrac{(n+1/2)(n-|m|)!}{\pi(n+|m|)!}} P_n^{(|m|)}(\sin\theta) \cdot \cos^{|m|}\theta \cdot \begin{cases} -\sin m\phi, & -n \leq m < 0, \\ \frac{1}{\sqrt{2}}, & m = 0, \\ \cos m\phi, & 0 < m \leq n, \end{cases} \tag{6.3}$$

where $P_n^{(|m|)}(t) = \frac{\mathrm{d}^{|m|}}{\mathrm{d}t^{|m|}} P_n(t)$ is the $|m|$-th derivative of the Legendre polynomial of degree n, defined by

$$P_n(t) = \tfrac{1}{2^n n!} \tfrac{\mathrm{d}^n}{\mathrm{d}t^n}(t^2 - 1)^n.$$

The infinite family $(Y_n^m)_{|m|\leq n,\, n\in\mathbb{N}}$ is a Hilbert basis of the space $L^2(\mathbb{S}^2)$, which is equipped with the usual inner product and the associated norm,

$$\langle f, g\rangle_{L^2(\mathbb{S}^2)} = \int_{\mathbb{S}^2} f(x)g(x)\mathrm{d}\sigma, \quad \|f\|_{L^2(\mathbb{S}^2)} = \langle f, f\rangle_{L^2(\mathbb{S}^2)}^{1/2}.$$

In this basis, any $f \in L^2(\mathbb{S}^2)$ admits a unique spectral expansion,

$$f = \sum_{|m|\leq n} \hat{f}_n^m Y_n^m, \quad \text{with} \quad \hat{f}_n^m = \langle f, Y_n^m\rangle_{L^2(\mathbb{S}^2)}. \tag{6.4}$$

The space

$$\mathbb{Y}_n = \mathrm{span}\{Y_n^m,\, |m| \leq n\}$$

represents the restriction to $\mathbb{S}^2$ of the homogeneous harmonic polynomials of degree n (in $\mathbb{R}^3$), whereas for every degree $D \geq 0$, the space defined by

$$\mathcal{Y}_D = \mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_D = \mathrm{span}\{Y_n^m,\, |m| \leq n,\, 0 \leq n \leq D\} \tag{6.5}$$

contains all the spherical harmonics with degree less than or equal to D.

6.2.2 Grid functions

The space of real functions defined on CS_N is denoted by

$$\mathbb{R}^{\mathrm{CS}_N} = \{f : \mathrm{CS}_N \to \mathbb{R}\}.$$

The canonical basis $(\delta_{x_i})_{1\leq i\leq \bar{N}}$ of $\mathbb{R}^{\mathrm{CS}_N}$ is defined by

$$\delta_{x_i}(x_j) = \delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases} \qquad 1 \leq i, j \leq \bar{N}.$$

In this basis, any $f \in \mathbb{R}^{\mathrm{CS}_N}$ has the decomposition

$$f = \sum_{i=1}^{\bar{N}} f(x_i)\delta_{x_i}.$$

For any real function defined on the sphere, $u : x \in \mathbb{S}^2 \mapsto u(x) \in \mathbb{R}$, the restriction of u on CS_N is the function defined by

$$u|_{\mathrm{CS}_N} := \sum_{i=1}^{\bar{N}} u(x_i)\delta_{x_i} \in \mathbb{R}^{\mathrm{CS}_N}, \quad u|_{\mathrm{CS}_N}(x_i) = u(x_i),\ 1 \leq i \leq \bar{N}. \tag{6.6}$$

In this way, u interpolates a grid function $f \in \mathbb{R}^{\mathrm{CS}_N}$ if (and only if), $u|_{\mathrm{CS}_N} = f$.

6.3 Lagrange interpolation space

We tackle the Lagrange interpolation problem on the Cubed Sphere CS_N, in an algebraic way. We define a subspace of spherical harmonics, $\mathcal{U}_N \subset L^2(\mathbb{S}^2)$, such that, in the space $\mathcal{U}_N$, the set of equations (6.1) has always a unique solution $u \in \mathcal{U}_N$. Following the procedure of [10] and the presentation of [9], the most natural way consists in eliminating any spherical harmonic of degree n which is undersampled, and in keeping only the orthogonal complement, by induction on the degree n. We summarize this procedure in this section, with a proof that is different from the one of [10]. In particular, the following lemma builds an interpolating function in a novel way.

Lemma 6.1 (Interpolating spherical harmonic with degree at most $4N-1$)**.** *Any grid function $f \in \mathbb{R}^{\mathrm{CS}_N}$ can be interpolated by a spherical harmonics $u \in \mathcal{Y}_D$ with*

$$D = \begin{cases} 4N-1, & \textit{if } N \textit{ is odd,} \\ 4N-2, & \textit{if } N \textit{ is even.} \end{cases} \tag{6.7}$$

Proof. We define functions $L_{x_i} \in \mathcal{Y}_D$ such that $L_{x_i}(x_j) = \delta_{ij}$, $1 \leq i, j \leq \bar{N}$; this implies by linearity that

$$u = \sum_{i=1}^{\bar{N}} f(x_i)L_{x_i} \in \mathcal{Y}_D$$

interpolates f, *i.e.* $u(x_i) = f(x_i)$, $1 \leq i \leq \bar{N}$.

Fix $\xi \in \mathrm{CS}_N$. Assume that $\xi = \frac{1}{r}(1, \tan\frac{i\pi}{2N}, \tan\frac{j\pi}{2N})$, with $r > 0$, $-\frac{N}{2} \leq i, j \leq \frac{N}{2}$ (similar arguments apply otherwise). We cover the Cubed Sphere CS_N by means of great circles as in Figure 6.1,

$$\mathrm{CS}_N \subset \bigcup_{\alpha\in A} \{x \in \mathbb{S}^2 : x \cdot \alpha = 0\}, \tag{6.8}$$

where the normal vector α browses the set

$$A = \{(-\sin\tfrac{k\pi}{2N}, \cos\tfrac{k\pi}{2N}, 0),\ -\tfrac{N}{2} \le k \le \tfrac{3N}{2} - 1\} \cup \{(-\sin\tfrac{l\pi}{2N}, 0, \cos\tfrac{l\pi}{2N}),\ -\tfrac{N}{2} \le l \le \tfrac{3N}{2} - 1\}. \quad (6.9)$$

The number of such circles is given by $D+1 = 4N$, $4N-1$ if N is odd, even. Indeed, the indices k and l browse $2N$ values; the corresponding circles are distinct, except if N is even and $k = l = N$. Among these $D+1$ circles, there are exactly two circles which contain $\{\xi, -\xi\}$ (one with $k = i$, and one with $l = j$). The remaining $D-1$ circles, parametrized by $\alpha \in A$ such that $\xi \cdot \alpha \neq 0$, cover $\mathrm{CS}_N \setminus \{\xi, -\xi\}$. As a result, we define the spherical function

$$L_\xi(x) = \frac{1+\xi\cdot x}{2} \prod_{\substack{\alpha\in A \\ \xi\cdot\alpha\neq 0}} \frac{x\cdot\alpha}{\xi\cdot\alpha}, \quad x \in \mathbb{S}^2.$$

In this expression, we recognize the tangent plane at $-\xi$ ($1+\xi\cdot x = 0$), and the $D-1$ great circles that do not contain $\{\xi, -\xi\}$ ($x\cdot\alpha = 0$, with $\alpha \in A$ such that $\xi\cdot\alpha \neq 0$). In particular, $L_\xi \in \mathcal{Y}_D$, $L_\xi(\xi) = 1$, $L_\xi(-\xi) = 0$, and the covering of $\mathrm{CS}_N \setminus \{\xi, -\xi\}$ implies $L_\xi(\xi') = 0$ for every $\xi' \in \mathrm{CS}_N \setminus \{\xi, -\xi\}$. □

Theorem 6.2 (Interpolation space)**.** *Let the orthogonal decomposition*

$$\begin{cases} \mathbb{Y}_n = \mathcal{W}_n \overset{\perp}{\oplus} \mathcal{W}_n^\perp, \quad n \ge 0, \\ \quad \textit{with} \quad \mathcal{W}_0 := \{0\},\ \mathcal{W}_n := \{w \in \mathbb{Y}_n : \exists v \in \mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_{n-1}, w|_{\mathrm{CS}_N} = v|_{\mathrm{CS}_N}\},\ n \ge 1. \end{cases} \quad (6.10)$$

Let T denote the evaluation operator on CS_N,

$$\begin{array}{rccc} T: & \mathscr{C}^0(\mathbb{S}^2) & \longrightarrow & \mathbb{R}^{\mathrm{CS}_N} \\ & u & \longmapsto & u|_{\mathrm{CS}_N} \end{array}. \quad (6.11)$$

Then, there exists a smallest degree $d = d(N) \ge 0$ such that the linear map $T_d := T|_{\mathcal{W}_0^\perp \oplus \cdots \oplus \mathcal{W}_d^\perp}$ is isomorphic. The space $\mathcal{W}_0^\perp \oplus \cdots \oplus \mathcal{W}_{d(N)}^\perp$ is called the interpolation space*, and is denoted by $\mathcal{U}_N$; the invert of $T_{d(N)}$ is called the* interpolation operator *and is denoted by $\mathcal{I}_N : \mathbb{R}^{\mathrm{CS}_N} \to \mathcal{U}_N$.*

Proof. Firstly, we prove, by induction on the degree $n \ge 0$, that

$$T(\mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_n) = T(\mathcal{W}_0^\perp \oplus \cdots \oplus \mathcal{W}_n^\perp). \quad (6.12)$$

For $n = 0$, this is due to $\mathbb{Y}_0 = \mathcal{W}_0^\perp$. Fix now $n \ge 1$ such that (6.12) is realized for the degree $n-1$ (induction). By definition of $\mathcal{W}_n$, $\mathbb{Y}_n = \mathcal{W}_n \oplus \mathcal{W}_n^\perp$, with $T(\mathcal{W}_n) \subset T(\mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_{n-1})$. We deduce that

$$T(\mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_n) = T(\mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_{n-1} \oplus \mathcal{W}_n^\perp) = T(\mathcal{W}_0^\perp \oplus \cdots \oplus \mathcal{W}_n^\perp),$$

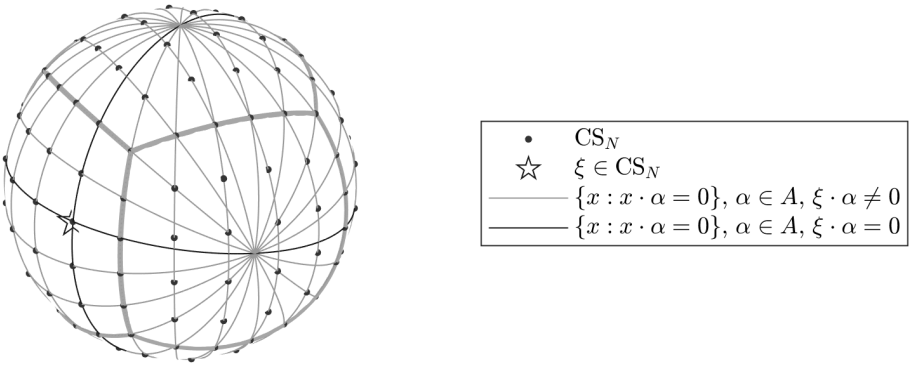


Figure 6.1: Covering of the Cubed Sphere as in (6.8). For a given $\xi \in \mathrm{CS}_N$, we cover CS_N (black dots) by $D+1$ great circles parametrized by a normal vector $\alpha \in A$, with D defined in (6.7), and A defined in (6.9). Two circles contains $\{\xi, -\xi\}$ (black circles); the $D-1$ remaining ones cover $\mathrm{CS}_N \setminus \{\xi, -\xi\}$ (gray circles). Here, $N = 5$.

which achieves the induction.

Secondly, fix

$$d = \begin{cases} 4N-1, & \text{if } N \text{ is odd} \\ 4N-2, & \text{if } N \text{ is even.} \end{cases}$$

Then, Lemma 6.1 shows that the linear map $T|_{\mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_d}$ is surjective; hence, (6.12) with $n = d$ implies that the restriction T_d is surjective too.

To conclude, we prove that T_d is also injective. Assume that there is $w \in \mathcal{W}_0^\perp \oplus \cdots \oplus \mathcal{W}_d^\perp \setminus \{0\}$ such that $Tw = 0$. Let $n \leq d$ be the degree of w. The unique constant function $u \in \mathbb{Y}_0$ such that $u|_{\mathrm{CS}_N} = 0$ is null, so $n \geq 1$. Then, there are $w_n \in \mathcal{W}_n^\perp \setminus \{0\}$ and $y \in \mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_{n-1}$ such that $w = w_n - y$. Since $Tu = 0$, $w_n|_{\mathrm{CS}_N} = y|_{\mathrm{CS}_N}$, so $w_n \in \mathcal{W}_n$, which is a contradiction. □

The subspace $\mathcal{W}_n$ represents spherical harmonic of degree n which are undersampled on CS_N, since they coincide with a spherical harmonic of smaller degree. On the contrary, any spherical harmonics in the orthogonal supplementary $\mathcal{W}_n^\perp$ is properly sampled on CS_N, since it can be reconstructed by interpolation. The resulting space $\mathcal{U}_N$ is intrinsicly defined, with an algebraic description. Unfortunately, we do not have at disposal an analytical description of $\mathcal{U}_N$, nor of the spaces $\mathcal{W}_n$, $\mathcal{W}_n^\perp$ (apart from some special cases).

Remark 6.3. Our proof shows that the optimal degree $d(N)$ satisfies $d(N) \leq 4N-1$ if N is odd, and $d(N) \leq 4N-2$ if N is even. This is an improvement of the bound proposed in the initial paper [10] ($\approx 21N$), which was based on the shortest geodesic distance of Theorem 5.4 and [146, Theorem 2.4]-[147, Lemma 3.13].

To conclude, by definition of the interpolation space $\mathcal{U}_N$ and the interpolation operator $\mathcal{I}_N$, the Lagrange interpolation problem on CS_N has a unique solution in $\mathcal{U}_N$, and for every $f \in \mathbb{R}^{\mathrm{CS}_N}$, $\mathcal{I}_N f \in \mathcal{U}_N$ denotes the unique element $u \in \mathcal{U}_N$ such that $u|_{\mathrm{CS}_N} = f$. The following result establishes that the degree of $\mathcal{I}_N f$ is minimal.

Corollary 6.4 (Minimal degree). *Let $f \in \mathbb{R}^{\mathrm{CS}_N}$ be a grid function interpolated by $\mathcal{I}_N f \in \mathcal{U}_N$. Let $u \in \mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_D$ be a spherical harmonic of degree D which interpolates f,* i.e. *$u|_{\mathrm{CS}_N} = f$. Then, the degree D of u is greater than or equal to the degree of $\mathcal{I}_N f$. In other words, the degree of the interpolation spherical harmonics $\mathcal{I}_N f$ is minimal.*

Proof. If $D > d(N)$, the result is obvious. Otherwise, $f = Tu \in T(\mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_D)$. We deduce from (6.12) that there is some $v \in \mathcal{W}_0^\perp \oplus \cdots \oplus \mathcal{W}_D^\perp$ such that $f = Tv$. Here, $D \leq d(N)$, so $v \in \mathcal{U}_N$, which implies $v = \mathcal{I}_N f$. Therefore, the degree of $\mathcal{I}_N f$ coincides with the degree of v, which is itself less than or equal to D. □

6.4 Matrix computation

In this section, we propose an algorithm to compute the interpolation space $\mathcal{U}_N$, and an interpolation function $\mathcal{I}_N f$, using numerical linear algebra. The approach deals with a numerical matrix analysis of the evaluation operator T defined in (6.11); it is based on the special echelon orthogonal factorization described in Appendix 6.A.

Definition 6.5 (Vandermonde matrix). For any $n \geq 0$, the Vandermonde matrix $\mathbf{A}_n$ is defined as the matrix of the operator $T|_{\mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_n}$,

$$\mathbf{A}_n := [Y_k^m(x_j)]_{1 \leq j \leq \bar{N},\, |m| \leq k \leq n} \in \mathbb{R}^{\bar{N} \times (n+1)^2},$$

where the column index (k, m) is sorted in lexicographical order.

By definition, the Vandermonde matrix $\mathbf{A}_n$, has a block structure, where each block corresponds to the matrix of $T|_{\mathbb{Y}_k}$, where k is a fixed degree,

$$\mathbf{A}_n = \begin{bmatrix} A_0 & A_1 & \cdots & A_n \end{bmatrix} \in \mathbb{R}^{N\times(n+1)^2}, \quad \text{with} \quad A_k := [Y_k^m(x_j)]_{1\le j\le \bar{N},\,|m|\le k} \in \mathbb{R}^{\bar{N}\times(2k+1)}. \tag{6.13}$$

As a result, Theorem 6.11 applies; there is an echelon form (6.19) such that $\mathbf{A}_n = \mathbf{V}_n\mathbf{E}_n\mathbf{U}_n^\intercal$, where

- the matrix $\mathbf{V}_n \in \mathbb{R}^{\bar{N}\times\bar{N}}$ is orthogonal,
- the matrix $\mathbf{U}_n = \operatorname{diag}(U_k, 0 \le k \le n)$ is block diagonal, with orthogonal matrices $U_k \in \mathbb{R}^{(2k+1)\times(2k+1)}$, as in (6.20),
- the matrix $\mathbf{E}_n \in \mathbb{R}^{\bar{N}\times(n+1)^2}$ is in echelon form as in (6.21), with $g_0, \dots, g_n$ for the dimensions of the blocks of rows.

Next, we recognize that (6.25) is the matrix representation of the decomposition (6.10). Therefore, the matrix $\mathbf{U}_n$ contains orthonormal bases of the spaces $\mathcal{W}_k^\perp$ and $\mathcal{W}_k$.

Definition 6.6 (Basis functions)**.** Let $n \ge 0$ be a fixed degree, and a special echelon form (6.19) of $\mathbf{A}_n$. For all $0 \le k \le n$ and $1 \le i \le 2k+1$, the *basis function* $u_k^i \in \mathbb{Y}_k$ is defined by

$$u_k^i \in \mathbb{Y}_k, \quad u_k^i(x) = [Y_k^m(x)]_{|m|\le k}^\intercal U_k(:,i),\ x \in \mathbb{S}^2,$$

so that, for any $0 \le k \le n$,

- the set $\{u_k^i, g_k + 1 \le i \le 2k+1\}$ defines an orthonormal basis of the undersampled space $\mathcal{W}_k$ (defined in (6.10)),
- the set $\{u_k^i,\ 1 \le i \le g_k\}$ defines an orthonormal basis of the space $\mathcal{W}_k^\perp$.

Remark 6.7. Here, the spaces $\mathcal{W}_k$ and $\mathcal{W}_k^\perp$ are intrinsicly defined, but it is not the case for the orthonormal bases $\{u_k^i\}$ (as it is often the case for orthonormal bases).

Therefore, the matrix $\mathbf{A}_n\tilde{\mathbf{U}}_n$, where $\tilde{\mathbf{U}}_n$ is given by (6.26), represents the opertator $T|_{\mathcal{W}_0^\perp\oplus\cdots\oplus\mathcal{W}_n^\perp}$, in the basis $\{u_k^i,\ 1 \le i \le g_k,\ 0 \le k \le n\}$. This matrix has full column rank $r_n = g_0 + \cdots + g_n$ and has the QR factorization (6.27). We deduce from Theorem 6.2 that $\mathbf{A}_d\tilde{\mathbf{U}}_d$ is invertible for the degree $d = d(N)$, but that the row rank of $\mathbf{A}_n\tilde{\mathbf{U}}_n$ is deficient if $n < d$ ($r_n < \bar{N}$). This suggests to compute incrementally the factorization (6.19), for increasing values of n, until the value of $r_n = g_0 + \cdots + g_n$ reaches the value $\bar{N}$. Following the proof of Theorem 6.11, we obtain the practical algorithm described on the next page.

The algorithm provides the optimal degree $d = d(N)$ (Theorem 6.2), a special echelon form (6.19) of $\mathbf{A}_d$. We readily get an orthonormal basis $\{u_k^i,\ 1 \le i \le g_k,\ 0 \le k \le d\}$ of the interpolation space $\mathcal{U}_N$, and a QR form of the evaluation operator $T_d : \mathcal{U}_N \to \mathbb{R}^{\mathrm{CS}_N}$,

$$\mathbf{A}_d\tilde{\mathbf{U}}_d = \mathbf{V}_d\tilde{\mathbf{E}}_d, \tag{6.14}$$

where $\mathbf{V}_d$ is orthogonal, and the upper triangular matrix $\tilde{\mathbf{E}}_d \in \mathbb{R}^{\bar{N}\times\bar{N}}$ is nonsingular. Then, any interpolation problem (6.1) can be solved with this factorization as follows.

Corollary 6.8. *Assume that the factorization* (6.14) *has been pre-computed. Let* $f \in \mathbb{R}^{\mathrm{CS}_N}$ *be a grid function on* CS_N*. Then, the unique element* $u \in \mathcal{U}_N$ *such that* $u(x_j) = f(x_j)$, $1 \le j \le \bar{N}$, *is given by*

$$\mathcal{I}_N[f](x) = [Y_n^m(x)]_{|m|\le n\le d}^\intercal\tilde{\mathbf{U}}_d\,\alpha, \quad \textit{with} \quad \alpha = (\tilde{\mathbf{E}}_d)^{-1}\mathbf{V}_d^\intercal\,[f(x_j)]_{1\le j\le\bar{N}};$$

here, the vector α *is obtained by backward substitution in the upper triangular system*

$$\tilde{\mathbf{E}}_d\,\alpha = \mathbf{V}_d^\intercal\,[f(x_j)]_{1\le j\le\bar{N}}.$$

Incremental special echelon orthogonal factorization of Vandermonde matrices

Input. Parameter N of the Cubed Sphere CS_N.

Initialization. For $n = 0$, compute the factorization $\mathbf{A}_0 = \mathbf{V}_0\mathbf{E}_0\mathbf{U}_0^\intercal$:
1. compute the matrix $\mathbf{A}_0$ defined in (6.13);
2. compute the matrices $\mathbf{V}_0$, $\mathbf{E}_0$ and $\mathbf{U}_0$ by SVD of $\mathbf{A}_0$;
3. evaluate the number of nonzero diagonal coefficients in $\mathbf{E}_0$, $r_0 = g_0$.

Iterations. For $n \geq 1$, compute the factorization $\mathbf{A}_n = \mathbf{V}_n\mathbf{E}_n\mathbf{U}_n^\intercal$:
1. compute the matrix A_n defined in (6.13);
2. compute matrices V_n, $\binom{\Lambda_n}{0}$ and U_n by SVD (6.23);
3. assemble the matrices $\mathbf{V}_n$, $\mathbf{E}_n$ and $\mathbf{U}_n$ with (6.24);
4. evaluate the number g_n of nonzero diagonal coefficients in $\binom{\Lambda_n}{0}$, and evaluate the rank of $\mathbf{A}_n$ with $r_n = r_{n-1} + g_n$.

Stopping criterion. Exit when $r_n = \bar{N}$, and set $d = n$.

Output. Smallest degree d such that the Vandermonde matrix $\mathbf{A}_d$ has full row rank, and associated factorization $\mathbf{A}_d = \mathbf{V}_d\mathbf{E}_d\mathbf{U}_d^\intercal$.

Proof. The matrix of $\mathcal{I}_N = T_d^{-1}$ is given by $(\mathbf{A}_d\tilde{\mathbf{U}}_d)^{-1} = (\tilde{\mathbf{E}}_d)^{-1}\mathbf{V}_d^\intercal$, due to (6.14). □

To finish with, Corollary 6.4 proves that the degree of the interpolation function $\mathcal{I}_N f$ is minimal. As an immediate consequence to Corollary 6.13, we have in fact a stronger result: $\mathcal{I}_N f$ is the minimal interpolation spherical harmonics, with respect to some reverse lexicographical order. Roughly speaking, this result means that the components of $\mathcal{I}_N f$ with large degrees are as small as possible.

Corollary 6.9 (Minimal interpolation spherical harmonic, for a reverse lexicographical order)**.** *Let $f \in \mathbb{R}^{\mathrm{CS}_N}$ be interpolated by $u = \mathcal{I}_N f \in \mathcal{U}_N$. Assume that $v \in \mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_d$ is another interpolation function of f,* i.e. *$u \neq v$ and $u|_{\mathrm{CS}_N} = v|_{\mathrm{CS}_N} = f$. Let $u_n, v_n \in \mathbb{Y}_n$, $0 \leq n \leq d$, be such that $u = u_0 + \cdots + u_d$, $v = v_0 + \cdots + v_d$. Then,*

$$\exists\, 0 \leq n \leq d, \quad \|u_d\| = \|v_d\|, \quad \ldots, \quad \|u_{n+1}\| = \|v_{n+1}\|, \quad \|u_n\| < \|v_n\|.$$

In other words, the function $\mathcal{I}_N f$ is the minimal spherical harmonics which interpolates f, with respect to a reverse lexicographical order on the degree.

6.5 Numerical experiments

6.5.1 Numerical dimensions

We have introduced an interpolation space $\mathcal{U}_N = \mathcal{W}_0^\perp \oplus \cdots \oplus \mathcal{W}_{d(N)}^\perp$, where $\mathcal{W}_n$ represents spherical harmonics of degree n which are undersampled on CS_N, defined in (6.10). Following our matrix analysis, the dimension of the subspace $\mathcal{W}_n^\perp$ is given by

$$g_n = \dim \mathcal{W}_n^\perp, \quad n \geq 0.$$

It corresponds to the number of nonzero singular values in the matrix $S_n = \binom{\Lambda_n}{0}$ from the SVD (6.23). In practice, the value of g_n can be estimated numerically by some thresholding of the diagonal terms of S_n. For a given threshold $0 < \tau < 1$, one considers that

$$g_n = \begin{cases} 0, \text{ if } S_n(1,1) \leq \tau, \\ \text{the number of } S_n(i,i) \text{ such that } S_n(i,i) > \tau S_n(1,1), \text{ otherwise.} \end{cases} \tag{6.15}$$

We have tabulated the numerical values of g_n, using the rule (6.15) and various thresholds τ. This has led to the following claim.

Claim 6.10. *The following assertions hold.*

(i) The matrix $\mathbf{A}_{2N-1}$ *has full column rank. Equivalently,* $r_{2N-1} = 4N^2$.

(ii) The matrix $\mathbf{A}_{3N}$ *has full row rank. Equivalently,* $r_{3N} = \bar{N}$.

This claim, whose proof is still open, has the following consequences.

1. Any spherical harmonics of degree smaller than $2N$ is properly sampled on CS_N, *i.e.*

$$\mathcal{W}_n^{\perp} = \mathbb{Y}_n, \quad n \leq 2N-1.$$

Note that the critical degree $2N$ coincides exactly with the Shannon-Nyquist angular frequency if we consider trigonometric polynomials along an equatorial grid[1] with step $\frac{\pi}{2N}$.

2. The optimal degree $d(N)$ of Theorem 6.2 satisfies

$$d(N) \leq 3N;$$

this implies that any spherical harmonics of degree larger than $3N$ is undersampled on CS_N, *i.e.*

$$\mathcal{W}_n = \mathbb{Y}_n, \quad n \geq 3N+1.$$

From now on, Claim 6.10 is assumed to perform further numerical approximations. Then, we need to select values of g_n in the range $2N \leq n \leq 3N$. When using the threshold $\tau = 10^{-4}$, with N increasing from $N = 1$ to $N = 6$, the observed values of g_n obey the rule

$$g_n = \begin{cases} 2n+1, & 0 \leq n \leq 2N-1, \\ 4(3N-n)-2, & 2N \leq n \leq 3N-2, \\ 3, & n = 3N-1, \\ 1, & n = 3N. \end{cases} \tag{6.16}$$

This suggests using (6.16) as an ansatz to infer the values of g_n in the algorithm on the preceding page, instead of thresholding singular values. In the sequel, we proceed in this way, so the ansatz (6.16) gives the values corresponding to our numerical spaces.

The ansatz (6.16) is stronger than Claim 6.10. It further implies that $d(N) = 3N$, and that for the intermediate degrees $2N \leq n \leq 3N$, less and less spherical harmonics are correctly sampled when the degree increases. We insist on the fact that contrary to Claim 6.10, the ansatz (6.16) depends on the particular choice of the selected threshold τ in (6.15). However, it has proven to be worth to be retained in the sequel to numerically evaluate a spherical harmonics Lagrange basis.

6.5.2 Least-squares approximation of Legendre spherical harmonics

To go further with the numerical analysis of the interpolation space, we consider the distance of any Legendre spherical harmonic Y_n^m to the interpolation space $\mathcal{U}_N$,

$$d(Y_n^m, \mathcal{U}_N) = \min_{u \in \mathcal{U}_N} \|Y_n^m - u\| = \left[\int_{\mathbb{S}^2} (Y_n^m(x) - u(x))^2 \, \mathrm{d}\sigma \right]^{1/2}, \quad |m| \leq n \leq 3N. \tag{6.17}$$

The distances are computed as the norms of the columns of the projector $\mathrm{I} - \tilde{\mathbf{U}}_{3N} \tilde{\mathbf{U}}_{3N}^{\intercal}$,

$$[d(Y_n^m, \mathcal{U}_N)]_{|m| \leq n \leq 3N} = \left[\|(\mathrm{I} - \tilde{\mathbf{U}}_{3N} \tilde{\mathbf{U}}_{3N}^{\intercal})(:, j)\|_2 \right]_{1 \leq j \leq (3N+1)^2};$$

[1] On the equiangular grid $\{-\frac{\pi}{4} + j\frac{\pi}{2N},\ 0 \leq j \leq 4N-1\}$, any trigonometric polynomial $\theta \mapsto \exp(\mathrm{i}k\theta)$, with $|k| < 2N$, is correctly sampled. For $k = 2N$, $\theta \mapsto \sin k\theta$ is undersampled.

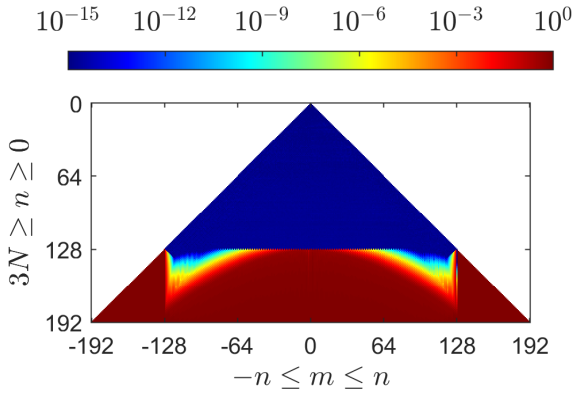


Figure 6.2: Distance (6.17) of the real Legendre spherical harmonics Y_n^m to the interpolation space of CS_N: $d(Y_n^m, \mathcal{U}_N)$, $|m| \leq n \leq 3N$, with $N = 64$.

They are displayed in Figure 6.2, for the grid CS_N with $N = 64$ ($\bar{N} = 24\,578$ nodes); similar results are displayed in [10] for other values of N. The blue color corresponds to the distance zero (up to rounding errors), which means that the function Y_n^m belongs to the interpolation space $\mathcal{U}_N$. We see again that any spherical harmonics of degree smaller than $2N$ is correctly sampled on CS_N. On the contrary, the red color corresponds to the distance 1, which means that the function Y_n^m is orthogonal to the interpolation space $\mathcal{U}_N$, and is therefore completely undersampled. We observe some pattern which is reminiscent to a *rhomboid*; roughly speaking,

- Y_n^m is accurately approximated in the space $\mathcal{U}_N$ if $M_n \leq |m| < 2N$, where $n \mapsto M_n$ is some increasing function;
- Y_n^m is orthogonal to $\mathcal{U}_N$, for $|m| > 2N$ and for $|m| < M_n$.

We refer to [10] for further results, including various statistics of the distance, and a comparison with an interpolation space built by a direct SVD[2] of the full Vandermonde matrix $\mathbf{A}_{3N}$.

6.5.3 Interpolation test cases

To finish with, we interpolate the following set of test functions on the sphere $\mathbb{S}^2$:

$$
\begin{aligned}
f_1(x,y,z) &= 1 + x + y^2 + yx^2 + x^4 + y^5 + x^2y^2z^2, \\
f_2(x,y,z) &= \tfrac{3}{4}\exp\left[-\tfrac{(9x-2)^2}{4} - \tfrac{(9y-2)^2}{4} - \tfrac{(9z-2)^2}{4}\right], \\
&\quad + \tfrac{3}{4}\exp\left[-\tfrac{(9x+1)^2}{49} - \tfrac{9y+1}{10} - \tfrac{9z+1}{10}\right], \\
&\quad + \tfrac{1}{2}\exp\left[-\tfrac{(9x-7)^2}{4} - \tfrac{(9y-3)^2}{4} - \tfrac{(9z-5)^2}{4}\right], \\
&\quad - \tfrac{1}{5}\exp\left[-(9x-4)^2 - (9y-7)^2 - (9z-5)^2\right], \\
f_3(x,y,z) &= \tfrac{1}{9}[1 + \tanh(-9x - 9y + 9z)], \\
f_4(x,y,z) &= \tfrac{1}{9}[1 + \mathrm{sign}(-9x - 9y + 9z)].
\end{aligned}
$$

The function f_1 is polynomial and $f_1 \in \oplus_{n \leq 6} \mathbb{Y}_n$. The functions f_2 and f_3 are regular and they have many Legendre spherical harmonics in their expansion. The function f_4 is discontinuous. In Figure 6.3, the interpolation errors with $N = 2$ and $N = 4$ for this set of functions is displayed. Furthermore, we display in Figure 6.4 the uniform error and the root mean squared error (RMSE)

[2] A SVD of $\mathbf{A}_{3N}$ permits to define an interpolation function whose L^2 norm is minimal, but whose degree is not minimal (in general).

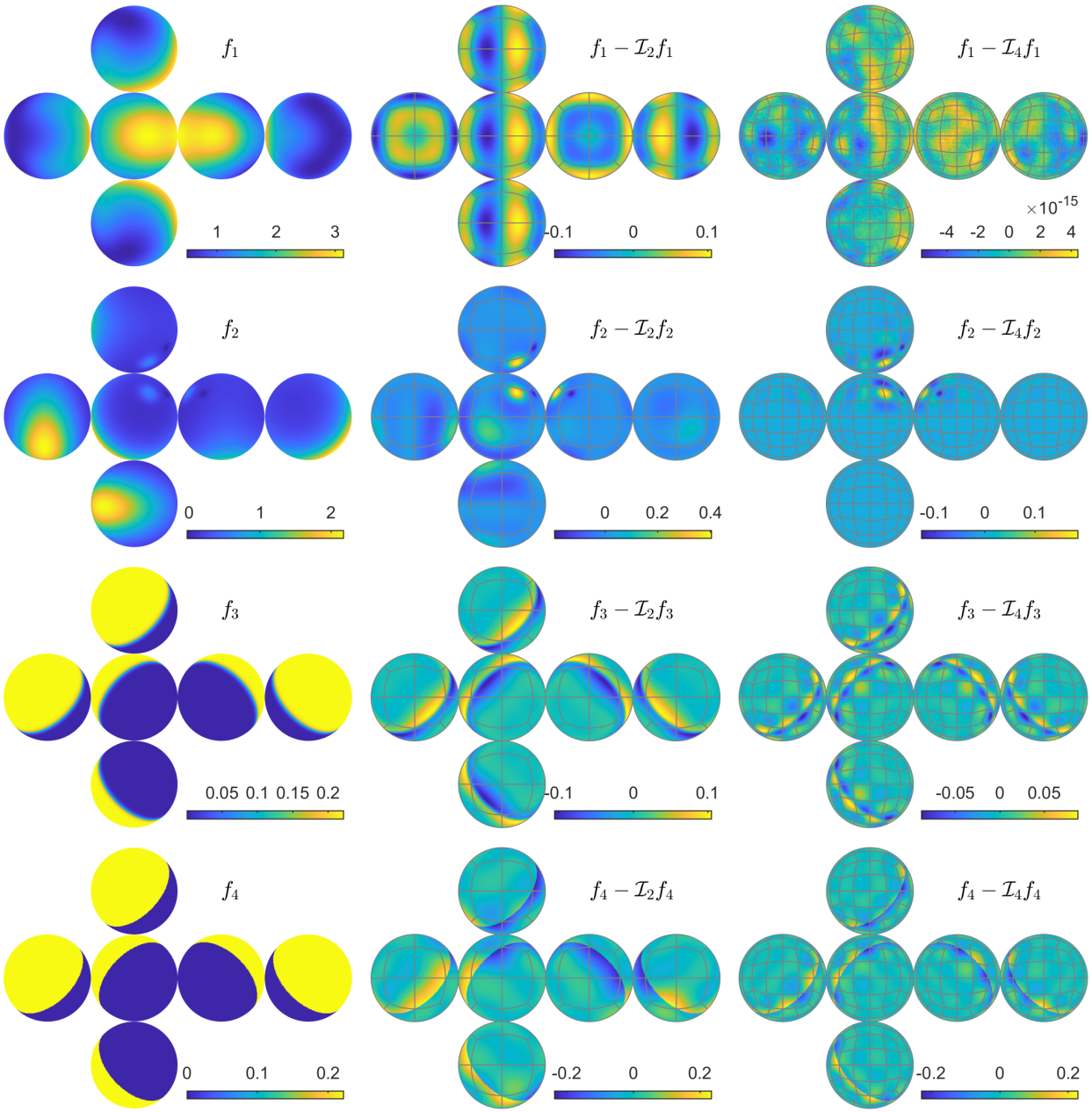


Figure 6.3: Interpolation of the test functions f_1, f_2, f_3 and f_4. Left column: the four test functions. Middle column: interpolation error on CS_2. Right column: interpolation error on CS_4.

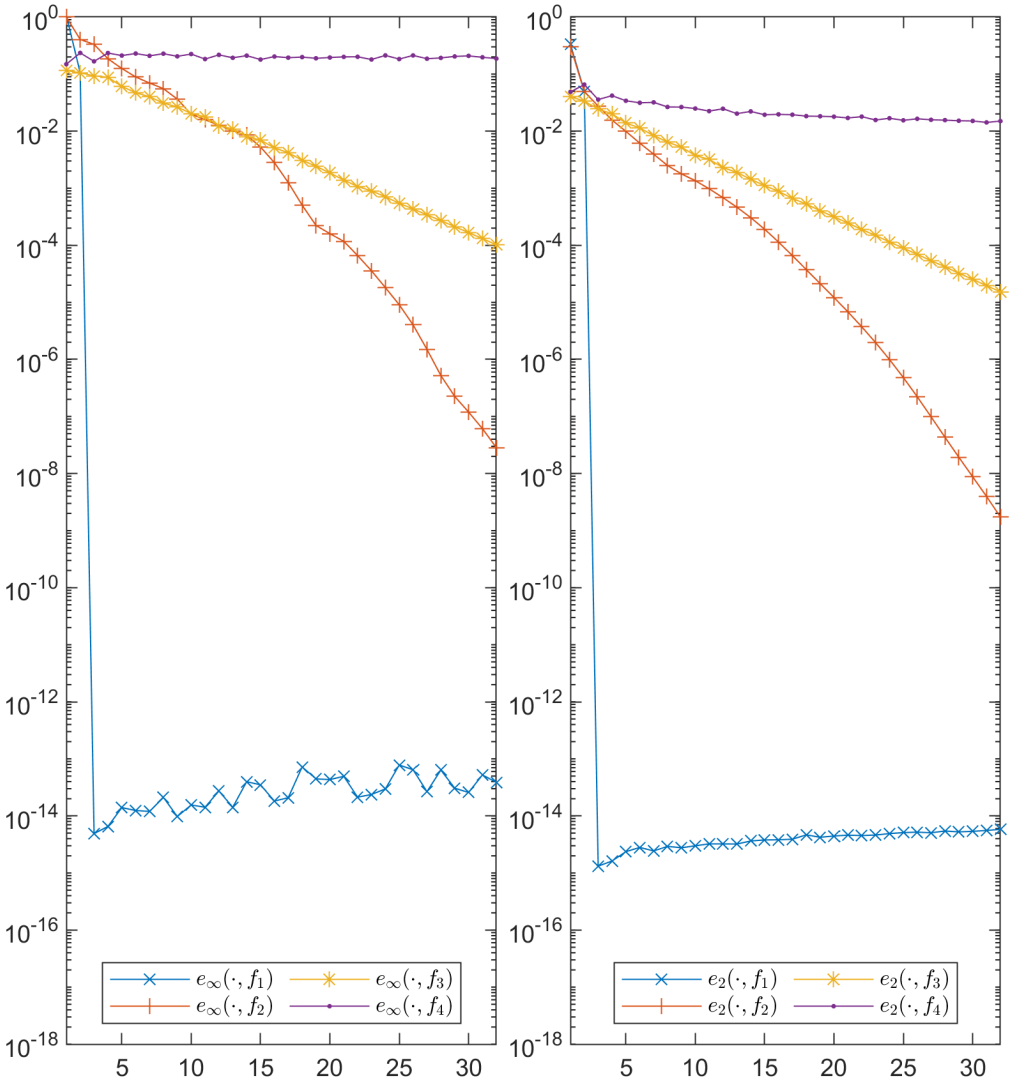


Figure 6.4: Interpolation error of test functions on CS_N, for $1 \leq N \leq 32$. Left: uniform error e_∞; right: RMSE e_2. Each error is evaluated on CS_{65}, and represented in logarithmic scale.

on CS_M, with $M = 65$; they are defined by

$$e_\infty(N, f_i) := \|f_i|_{\mathrm{CS}_M} - \mathcal{I}_N f_i|_{\mathrm{CS}_M}\|_\infty = \max_{x \in \mathrm{CS}_M} |f_i(x) - (\mathcal{I}_N f_i)(x)|,$$

$$e_2(N, f_i) := \tfrac{1}{(6M^2+2)^{1/2}} \|f_i|_{\mathrm{CS}_M} - \mathcal{I}_N f_i|_{\mathrm{CS}_M}\|_2 = \left(\tfrac{1}{6M^2+2} \textstyle\sum_{x \in \mathrm{CS}_M} |f_i(x) - (\mathcal{I}_N f_i)(x)|^2 \right)^{1/2}.$$

For N large enough, $f_1 \in \mathcal{U}_N$, which gives a null error. The smooth function f_2 is interpolated with an error decreasing with N. This is also the case for the function f_3, with a decreasing rate smaller than the one for f_2. This reflects the C^p regularity of the functions f_2 and f_3. Finally, as expected, the discontinuous function f_4 is not well interpolated. The RMSE decreases very slowly, and the uniform error does not decrease.

6.6 Conclusion

In this study, a methodology to associate a spherical harmonics subspace to the Cubed Sphere CS_N has been introduced. The particular subspace is based on a specific echelon factorization of the Vandermonde matrix. This space seems promising in terms of approximation power. It is used in Chapter 7 to design the new quadrature rule from [9].

Finally, this work took its origin in the numerical observation of the rank stated in Claim 6.10 and the ansatz (6.16). A full proof of this claim is an objective of further studies. Similarly, an analysis of the condition number of the matrix $\mathbf{A}_n$ is required as well. Partial answers are given in Chapter 8, dealing with least squares as in [11]. Also, further investigation of the symmetry properties may yield to some discrete Fourier analysis on the Cubed Sphere. Using the new interpolation procedure to various contexts is another future goal. An important goal is the application of this new framework to PDE's in meteorology.

6.A Special echelon orthogonal factorization

6.A.1 Main result

We propose some factorization, which reduces a block matrix into an echelon matrix, using orthogonal transformations only; it has the special property that the blocks of the "diagonal" are full row rank diagonal matrices.

Theorem 6.11 (Special echelon orthogonal factorization of a block matrix)**.** *Let $n \geq 0$. Let $\mathbf{A}_n$ be a matrix defined by $n+1$ blocks of columns, $A_k \in \mathbb{R}^{N \times m_k}$, $0 \leq k \leq n$,*

$$\mathbf{A}_n = \begin{bmatrix} A_0 & A_1 & \cdots & A_n \end{bmatrix} \in \mathbb{R}^{N \times M_n}, \quad \text{with} \quad M_n = m_0 + \cdots + m_n. \tag{6.18}$$

Then, $\mathbf{A}_n$ admits an echelon orthogonal factorization,

$$\mathbf{A}_n = \mathbf{V}_n \mathbf{E}_n \mathbf{U}_n^\intercal, \tag{6.19}$$

such that

- *the matrix $\mathbf{V}_n \in \mathbb{R}^{N \times N}$ is orthogonal,*
- *the matrix $\mathbf{U}_n \in \mathbb{R}^{M_n \times M_n}$ is orthogonal and block diagonal, so that*

$$\mathbf{U}_n = \begin{bmatrix} U_0 & & \\ & \ddots & \\ & & U_n \end{bmatrix} \in \mathbb{R}^{M_n \times M_n}, \quad \begin{array}{l} \text{with orthogonal matrices} \\ U_k \in \mathbb{R}^{m_k \times m_k},\ 0 \leq k \leq n, \end{array} \tag{6.20}$$

- *the matrix $\mathbf{E}_n \in \mathbb{R}^{N \times M_n}$ is in echelon form, and such that*

$$\mathbf{E}_n = \begin{bmatrix} \Lambda_0 & * & \cdots & * \\ 0 & \Lambda_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & \Lambda_n \\ 0 & \cdots & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{N \times M_n}, \begin{cases} \text{with full row rank diagonal matrices} \\ \Lambda_k = \begin{bmatrix} \ddots & 0 \end{bmatrix} \in \mathbb{R}^{g_k \times m_k} \text{ (for some } g_k \geq 0, \\ \text{and a nonincreasing positive diagonal),} \\ 0 \leq k \leq n. \end{cases} \tag{6.21}$$

Remark 6.12. Similar factorization have already been introduced in the field of robotics: see [126] for a problem of lexicographical least-squares, solved by a factorization that looks like (6.19).

Proof. The proof constructs the desired factorization (6.19), by induction on the number of blocks. For $n = 0$ (one block), it is achieved by a singular value decomposition (SVD) of the matrix $\mathbf{A}_0 = A_0$. In this case, the matrix $\mathbf{V}_0$ contains left singular vectors, the matrix $\mathbf{U}_0 = U_0$ contains right singular vectors, and the diagonal of the matrix Λ_0 contains g_0 nonincreasing positive singular values, with $g_0 = \operatorname{rank} \mathbf{A}_0 \geq 0$.

Assume now (induction step) that the result holds for some $n - 1 \geq 0$ (n blocks). Then,

$$\mathbf{A}_n = \begin{bmatrix} \mathbf{A}_{n-1} & A_n \end{bmatrix} = \begin{bmatrix} \mathbf{V}_{n-1} \mathbf{E}_{n-1} \mathbf{U}_{n-1}^\intercal & A_n \end{bmatrix} = \mathbf{V}_{n-1} \begin{bmatrix} \mathbf{E}_{n-1} & \mathbf{V}_{n-1}^\intercal A_n \end{bmatrix} \begin{bmatrix} \mathbf{U}_{n-1}^\intercal & 0 \\ 0 & \mathrm{I}_{m_n} \end{bmatrix}. \tag{6.22}$$

Here, the matrix $\mathbf{E}_{n-1}$ has a suitable echelon form, and its number of "diagonal" coefficients is given by $r_{n-1} = g_0 + \cdots + g_{n-1}$. Then, we diagonalize the last $N - r_{n-1}$ lines of the matrix $\mathbf{V}_{n-1}^\intercal A_n$. More precisely, we consider an SVD of the matrix $\mathbf{V}_{n-1}(:, r_{n-1}+1:N)^\intercal A_n$; there are orthogonal matrices $V_n \in \mathbb{R}^{(N-r_{n-1})\times(N-r_{n-1})}$, $U_n \in \mathbb{R}^{m_n \times m_n}$, and a full row rank diagonal matrix Λ_n, with g_n nonincreasing positive values on the diagonal, such that

$$\mathbf{V}_{n-1}(:, r_{n-1}+1:N)^\intercal A_n = V_n \begin{bmatrix} \Lambda_n \\ 0 \end{bmatrix} U_n^\intercal. \tag{6.23}$$

We deduce from (6.22) that

$$\begin{aligned}
\mathbf{A}_n &= \mathbf{V}_{n-1} \begin{bmatrix} \mathbf{E}_{n-1}(1:r_{n-1},:) & \mathbf{V}_{n-1}(:,1:r_{n-1})^\intercal A_n \\ 0 & V_n \begin{bmatrix} \Lambda_n \\ 0 \end{bmatrix} U_n^\intercal \end{bmatrix} \begin{bmatrix} \mathbf{U}_{n-1}^\intercal & 0 \\ 0 & \mathrm{I}_{m_n} \end{bmatrix} \\
&= \mathbf{V}_{n-1} \begin{bmatrix} \mathrm{I}_{r_{n-1}} & 0 \\ 0 & V_n \end{bmatrix} \begin{bmatrix} \mathbf{E}_{n-1}(1:r_{n-1},:) & \mathbf{V}_{n-1}(:,1:r_{n-1})^\intercal A_n U_n \\ 0 & \begin{bmatrix} \Lambda_n \\ 0 \end{bmatrix} \end{bmatrix} \begin{bmatrix} \mathbf{U}_{n-1}^\intercal & 0 \\ 0 & U_n^\intercal \end{bmatrix} \\
&= \mathbf{V}_n \mathbf{E}_n \mathbf{U}_n^\intercal,
\end{aligned}$$

with

$$\mathbf{V}_n = \mathbf{V}_{n-1} \begin{bmatrix} \mathrm{I}_{r_{n-1}} & 0 \\ 0 & V_n \end{bmatrix}, \ \mathbf{U}_n = \begin{bmatrix} \mathbf{U}_{n-1} & 0 \\ 0 & U_n \end{bmatrix}, \ \mathbf{E}_n = \begin{bmatrix} \mathbf{E}_{n-1}(1:r_{n-1},:) & \mathbf{V}_{n-1}(:,1:r_{n-1})^\intercal A_n U_n \\ 0 & \Lambda_n \\ 0 & 0 \end{bmatrix}, \tag{6.24}$$

which provides the factorization (6.19) with the desired structure. □

The proof of Theorem 6.11 provides an iterative algorithm to compute the factorization (6.19). The initialization is an SVD $\mathbf{A}_0 = \mathbf{V}_0 \mathbf{E}_0 \mathbf{U}_0^\intercal$; the number g_0 of nonzero diagonal coefficients in $\mathbf{E}_0$ gives the value $r_0 = g_0$. The iteration n computes the SVD (6.23) and deduces $\mathbf{V}_n$, $\mathbf{U}_n$ and $\mathbf{E}_n$ with (6.24). The number g_n of additionnal nonzero diagonal coefficients determines the total number of "diagonal" coefficients in $\mathbf{E}_n$, $r_n = r_{n-1} + g_n$. Note that the values of g_n represent a number of nonzero singular values; in practice, either, they are theoretically known, either they are predicted, with some thresholding for instance.

The algorithm on page 97 is an implementation of this method in the case of Vandermonde matrices, for interpolation by spherical harmonics on the Cubed Sphere. In fact, the special structure of the factorization (6.19) has been designed to implement Theorem 6.2.

6.A.2 Consequences of the special echelon factorization

The special echelon orthogonal factorization (6.19) in Theorem 6.11, and its equivalent form $\mathbf{A}_n \mathbf{U}_n = \mathbf{V}_n \mathbf{E}_n$, are very rich. We enumerate below a list of information which is directly extracted from these decompositions.

- For any $0 \le k \le n$, consider the following orthogonal decomposition of the target space $\mathbb{R}^N$,

  $$\mathbb{R}^N = \operatorname{Ran} \mathbf{A}_k \overset{\perp}{\oplus} \operatorname{Ker} \mathbf{A}_k^\intercal;$$

  the first M_k columns in $\mathbf{A}_n \mathbf{U}_n = \mathbf{V}_n \mathbf{E}_n$ give

  $$\mathbf{A}_k \mathbf{U}_n(1:M_k, 1:M_k) = \mathbf{V}_n \, \mathbf{E}_n(:, 1:M_k),$$

  where we can deduce the range $\operatorname{Ran} \mathbf{A}_k$:

- the number $r_k := g_0 + \cdots + g_k$ of nonzero "diagonal" terms in $\mathbf{E}_n(:, 1 : M_k)$ coincides with $\operatorname{rank} \mathbf{A}_k$, *i.e.* $r_k = \operatorname{rank} \mathbf{A}_k$;
- the r_k columns of $\mathbf{V}_n(:, 1 : r_k)$ represent an orthonormal basis of the range $\operatorname{Ran} \mathbf{A}_k$;
- the $N - r_k$ columns of $\mathbf{V}_n(:, r_k + 1 : N)$ represent an orthonormal basis of the null space $\operatorname{Ker} \mathbf{A}_k^\intercal$.

- For any $0 \leq k \leq n$, consider the following decomposition of the input block $\mathbb{R}^{m_k}$,

$$\mathbb{R}^{m_k} = W_k \overset{\perp}{\oplus} W_k^\perp,\ W_k := \{u \in \mathbb{R}^{m_k} : A_k u \in \operatorname{Ran} \mathbf{A}_{k-1}\},\ W_k^\perp = \{u \in \mathbb{R}^{m_k} : u \perp W_k\}, \quad (6.25)$$

  (with the convention $W_0 = \operatorname{Ker} A_0$);

  - the number g_k of diagonal terms in Λ_k coincides with the dimension of $W_k^\perp$, $g_k = \dim W_k^\perp$;
  - the g_k columns of $U_k(:, 1 : g_k)$ represent an orthonormal basis of the subspace $W_k^\perp$;
  - the $m_k - g_k$ columns of $U_k(:, g_k + 1 : m_k)$ represent an orthonormal basis of W_k.

- Consider the following matrix, extracted from $\mathbf{U}_n$,

$$\tilde{\mathbf{U}}_n = \begin{bmatrix} U_0(:, 1 : g_0) & & \\ & \ddots & \\ & & U_n(:, 1 : g_n) \end{bmatrix} \in \mathbb{R}^{M_n \times r_n}; \quad (6.26)$$

  - the matrix $\mathbf{A}_n \tilde{\mathbf{U}}_n$ has full column rank r_n and admits the QR factorization

$$\mathbf{A}_n \tilde{\mathbf{U}}_n = \mathbf{V}_n \tilde{\mathbf{E}}_n, \quad \begin{array}{l}\text{where } \tilde{\mathbf{E}}_n \in \mathbb{R}^{N \times r_n} \text{ is an upper triangular matrix,} \\ \text{with full column rank;}\end{array} \quad (6.27)$$

    here, $\tilde{\mathbf{E}}_n$ is deduced from $\mathbf{E}_n$ by removal of "redundant" columns (only the non-zeros columns of Λ_k, $0 \leq k \leq n$, are kept).

- Fix $b \in \mathbb{R}^N$ and consider the least squares problem

$$\inf_{x \in \mathbb{R}^{M_n}} \|\mathbf{A}_n x - b\|^2; \quad (6.28)$$

  - any vector $x \in \mathbb{R}^{M_n}$ is a solution to (6.28), if and only if,

$$\mathbf{E}_n(1 : r_n, :)\, \mathbf{U}_n^\intercal\, x = \mathbf{V}_n(:, 1 : r_n)^\intercal\, b; \quad (6.29)$$

  - in $\operatorname{Ran} \tilde{\mathbf{U}}_n$, the minimal value (6.28) is reached exactly once: for $x = \tilde{\mathbf{U}}_n \alpha$, where α is the unique solution to

$$\tilde{\mathbf{E}}_n(1 : r_n, :)\, \alpha = \mathbf{V}_n(:, 1 : r_n)^\intercal\, b. \quad (6.30)$$

The least squares study above, combined with the special structure of the blocks Λ_k, shows that the least squares approximation in $\operatorname{Ran} \tilde{\mathbf{U}}_n$ is the minimal one for some reverse lexicographical order.

Corollary 6.13 (Least squares approximation minimal for the reverse lexicographical order)**.** *Consider the least squares problem* (6.28)*, with the notation of Theorem 6.11. Let* $\mathbf{x} = (x_0, \ldots, x_n) \in \mathbb{R}^{m_0} \times \cdots \times \mathbb{R}^{m_n}$ *be the least squares approximation that belongs to* $\operatorname{Ran} \tilde{\mathbf{U}}_n$, i.e. $\mathbf{x} = \tilde{\mathbf{U}}_n \alpha$*, where* $\tilde{\mathbf{U}}_n$ *is defined in* (6.26)*, and* α *denotes the unique solution to* (6.30)*. Then,* $\mathbf{x}$ *is the unique minimal solution to* (6.28) *in* $\mathbb{R}^{M_n}$*, for the reverse lexicographical order in* $\mathbb{R}^{n+1}$*, which means that for any other least squares approximation* $\mathbf{x}' = (x'_0, \ldots, x'_n)$*, solution to* (6.28) *with* $\mathbf{x} \neq \mathbf{x}'$*,*

$$\exists 0 \leq k \leq n, \quad \|x_n\|_2 = \|x'_n\|_2, \quad \|x_{n-1}\|_2 = \|x'_{n-1}\|_2, \quad \ldots, \quad \|x_{k+1}\|_2 = \|x'_{k+1}\|_2, \quad \|x_k\|_2 < \|x'_k\|_2.$$

Proof. The least squares approximations $x = (x_0, \dots, x_n)$ are parametrized using the linear system (6.29). Any coefficient of $\mathbf{U}_n^\intercal x$ which is associated to a null column in a diagonal block Λ_k, $0 \leq k \leq n$, is a free parameter. The other coefficients of $\mathbf{U}_n^\intercal x$ are given in $\tilde{\mathbf{U}}_n^\intercal x$; they are associated to positive terms in the diagonal blocks, and they are uniquely determined in term of free parameters with larger indices (backward substitution).

The n'th block of lines determine x_n; the special structure of Λ_n implies that the coefficients of $U_n(:, g_n + 1 : m_n)^\intercal x_n$ are free parameters, but that the coefficients of $U_n(:, 1 : g_n)^\intercal x_n$ are uniquely determined. Write

$$\|x_n\|^2 = \|U_n(:, 1 : g_n)^\intercal x_n\|^2 + \|U_n(:, g_n + 1 : m_n)^\intercal x_n\|^2;$$

therefore, building a minimal norm $\|x_n\|$ for the n'th block of equations means canceling the free parameters, *i.e.* setting $U_n(:, g_n + 1 : m_n)^\intercal x_n = 0$.

More generally, as soon as $x_{k+1}, \dots, x_n$ are fixed, the k'th block of lines determine x_k. The special structure of Λ_k implies that the coefficients of $U_k(:, g_k + 1 : m_k)^\intercal x_k$ are free, but that the coefficients of $U_k(:, 1 : g_k)^\intercal x_k$ are uniquely determined in term of $x_{k+1}, \dots, x_n$. Here again, the free parameters $U_k(:, g_k + 1 : m_k)^\intercal x_k$ must be zero to minimize $\|x_k\|$, due to

$$\|x_k\|^2 = \|U_k(:, 1 : g_k)^\intercal x_k\|^2 + \|U_k(:, g_k + 1 : m_k)^\intercal x_k\|^2.$$

Lastly, if all of the conditions $U_k(:, g_k + 1 : m_k)^\intercal x_k = 0$, $0 \leq k \leq n$, are fulfilled, then $x = \tilde{\mathbf{U}}_n \alpha \in \operatorname{Ran} \tilde{\mathbf{U}}_n$ and is the unique solution determined with (6.30). □

Chapter 7

Octahedral quadrature rule on the Cubed Sphere

7.1 Introduction

This chapter deals with a recent spherical quadrature rule, defined on the equiangular Cubed Sphere. This rule, originally introduced in [9], is based on Lagrange interpolation described in Chapter 6. The study takes benefit from the symmetry group given in Chapter 5.

Numerical integration on the sphere has been considered by several authors. We refer to the review [137]. The set of nodes and/or the associated weights are commonly identified by requiring exactness for a set of spherical harmonics, such as the spherical harmonics of degree smaller than a given value. For some optimal methods similar to Gauss quadratures, the nodes and the weights are both unknown; the associated set of equations is difficult. Some modern methods deal with the theory of t-designs, whose main purpose is to optimize the distribution of nodes, so that the quadrature rule with equal weights has degree of precision t. This theory has a mathematical and physical interest in itself. We refer to [117] for a review. For another class of methods, the weights are unknown, but the nodes are prescribed; this paper falls into this category.

Here the Cubed Sphere nodes are selected from the beginning as "good" quadrature nodes, and therefore, only the weights have to be identified. In [156], two examples of weights have been suggested. The first one was based on some extended trapezoidal rule, attributing some area to each node. The second one was thought as a perturbation of the first one with a design based on some optimization principle. See also [115] for another rule, including a Simpson like formula. Here we come back to the general question of the "best choice" of weights associated to the Cubed Sphere nodes. As in the general approach, we require exactness of the quadrature for a particular set of spherical harmonics. Using the space $\mathcal{U}_N$ defined in Theorem 6.2 immediately delivers a quadrature rule. This quadrature is different from the ones mentioned above. The space $\mathcal{U}_N$ remarkably enjoys invariance under the action of the group of the cube. This is somehow expected, since the group of CS_N is in fact the group of the cube, (or of the octahedron), as stated in Theorem 5.3. As will be shown below, the new quadrature rule inherits this invariance. This property is highly desirable. It is well known that group invariance is the backbone for the design of highly accurate spherical quadratures, [107,149,150,175]. Refer for this to the review [137].

The chapter is organized as follows. Section 7.2 establishes several rotational invariance properties of the interpolation space $\mathcal{U}_N$. In Section 7.3, the new quadrature is introduced. By construction, this quadrature is exact on the space $\mathcal{U}_N$. In addition, it is invariant under the octahedral group. This in particular implies that it is exact for a proportion of 15/16 of all (real) Legendre spherical harmonics. In Section 7.4, we display numerical results for a large series of test cases. It is observed that the new rule is only slightly suboptimal, when compared to the optimal Lebedev rules. This somehow supports the main Ansatz of this study, namely that the Cubed Sphere nodes are good quadrature nodes on the sphere. Lastly, Appendix 7.A reports URLs with some available quadrature

rules, including an open archiv for the new rule.

7.2 Rotational invariance of the interpolation space

In this section, we study the invariance of the interpolation space $\mathcal{U}_N$ under the symmetry group $\mathcal{G}$ of the Cubed Sphere CS_N. (Recall that $\mathcal{U}_N$ is defined in Theorem 6.2, and that $\mathcal{G}$ is the octahedral group (5.2), due to Theorem 5.3). We call "rotated" a function defined as follows.

Definition 7.1 ("Rotated" function)**.** Assume that $Q \in \mathcal{G}$ leaves a set E invariant, *i.e.* $Q^\intercal E = E$. Let $f : E \to \mathbb{R}$ be a function defined on E. The "rotated" function, denoted by $f(Q^\intercal \cdot)$, is the function

$$f(Q^\intercal \cdot) : x \in E \mapsto f(Q^\intercal x) \in \mathbb{R}.$$

Our main invariance result is the following theorem.

Theorem 7.2 (Invariance of the interpolation space)**.** *Let $n \geq 0$.*
(i) The undersampled subspace $\mathcal{W}_n$ defined in (6.10) *is invariant under $\mathcal{G}$,* i.e.

$$\forall Q \in \mathcal{G},\ \forall u \in \mathcal{W}_n,\ u(Q^\intercal \cdot) \in \mathcal{W}_n.$$

(ii) The subspace $\mathcal{W}_n^\perp$ is invariant under $\mathcal{G}$, i.e.

$$\forall Q \in \mathcal{G},\ \forall u \in \mathcal{W}_n^\perp,\ u(Q^\intercal \cdot) \in \mathcal{W}_n^\perp.$$

(iii) The interpolation space $\mathcal{U}_N$ is invariant under $\mathcal{G}$, i.e.

$$\forall Q \in \mathcal{G},\ \forall u \in \mathcal{U}_N,\ u(Q^\intercal \cdot) \in \mathcal{U}_N.$$

Proof. Fix $Q \in \mathcal{G}$, *i.e.* $Q \in \mathbb{R}^{3\times 3}$ is an orthogonal matrix such that $Q^\intercal \mathrm{CS}_N = \mathrm{CS}_N$.
(i) If $n = 0$, $\mathcal{W}_0 = \{0\}$ is invariant under $\mathcal{G}$. Fix now $u \in \mathcal{W}_n \subset \mathbb{Y}_n$ with $n \geq 1$. There exists $v \in \mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_{n-1}$ such that $u|_{\mathrm{CS}_N} = v|_{\mathrm{CS}_N}$, or equivalently, $(u-v)|_{\mathrm{CS}_N} = 0$. Firstly, $u(Q^\intercal \cdot) \in \mathbb{Y}_n$ and $v(Q^\intercal \cdot) \in \mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_{n-1}$. Secondly, $(u(Q^\intercal \cdot) - v(Q^\intercal \cdot))\,|_{\mathrm{CS}_N} = (u-v)|_{\mathrm{CS}_N}(Q^\intercal \cdot) = 0$, and therefore $u(Q^\intercal \cdot)|_{\mathrm{CS}_N} = v(Q^\intercal \cdot)|_{\mathrm{CS}_N}$; here, the commutation between rotation and restriction is justified by the following lemma.

Lemma 7.3 (Rotation commutes with restriction)**.** *For all $Q \in \mathcal{G}$, $n \geq 0$, and $u \in \mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_n$,*

$$u(Q^\intercal \cdot)|_{\mathrm{CS}_N} = u|_{\mathrm{CS}_N}(Q^\intercal \cdot) \in \mathbb{R}^{\mathrm{CS}_N}.$$

We postpone the proof of the lemma until the end of this section.
(ii) The result is a consequence of (i). Indeed, fix $u \in \mathcal{W}_n^\perp \subset \mathbb{Y}_n$ with $n \geq 0$. Then $u(Q^\intercal \cdot) \in \mathbb{Y}_n$. Furthermore, for every $v \in \mathcal{W}_n$,

$$\langle u(Q^\intercal \cdot), v\rangle_{L^2(\mathbb{S}^2)} = \int_{\mathbb{S}^2} u(Q^\intercal x)v(x)\mathrm{d}\sigma = \int_{\mathbb{S}^2} u(y)v(Qy)\mathrm{d}\sigma = \langle u, v(Q\cdot)\rangle_{L^2(\mathbb{S}^2)}\,; \quad (y := Q^\intercal x).$$

$\mathcal{W}_n$ is invariant under $\mathcal{G}$, so $v(Q\cdot) \in \mathcal{W}_n$. Then $v(Q\cdot)$ is orthogonal to u because $u \in \mathcal{W}_n^\perp$. We obtain $\langle u(Q^\intercal \cdot), v\rangle = \langle u, v(Q\cdot)\rangle = 0$, which proves $u(Q^\intercal \cdot) \in \mathcal{W}_n^\perp$.
(iii) The space $\mathcal{U}_N = \mathcal{W}_0^\perp \oplus \cdots \oplus \mathcal{W}_d^\perp$ is a sum of invariant subspaces due to (ii). □

Corollary 7.4 (Interpolation and symmetry)**.** *(i) The interpolation operator commutes with any symmetry of the group $\mathcal{G}$:*

$$\forall f \in \mathbb{R}^{\mathrm{CS}_N},\ \forall Q \in \mathcal{G},\ [\mathcal{I}_N f](Q^\intercal \cdot) = \mathcal{I}_N[f(Q^\intercal \cdot)].$$

(ii) The interpolation operator preserves the invariance property; in other words, if $f \in \mathbb{R}^{\mathrm{CS}_N}$ is invariant under $\mathcal{G}$, i.e. $\forall Q \in \mathcal{G}$, $f(Q^\intercal \cdot) = f$, *then $\mathcal{I}_N f$ is invariant under $\mathcal{G}$,* i.e. $\forall Q \in \mathcal{G}$, $[\mathcal{I}_N f](Q^\intercal \cdot) = \mathcal{I}_N f$.

Proof. (i) Firstly, $f(Q^\intercal\cdot) \in \mathbb{R}^{\mathrm{CS}_N}$ and $u = \mathcal{I}_N[f(Q^\intercal\cdot)] \in \mathcal{U}_N$ is the unique element of $\mathcal{U}_N$ such that $u|_{\mathrm{CS}_N} = f(Q^\intercal\cdot)$. Secondly, $v = \mathcal{I}_N f \in \mathcal{U}_N$ is the unique element of $\mathcal{U}_N$ such that $v|_{\mathrm{CS}_N} = f$. Due to Theorem 7.2.(iii), $v(Q^\intercal\cdot) \in \mathcal{U}_N$. By Lemma 7.3, $v(Q^\intercal\cdot)|_{\mathrm{CS}_N} = v|_{\mathrm{CS}_N}(Q^\intercal\cdot) = f(Q^\intercal\cdot)$, which proves $u = v(Q^\intercal\cdot)$. (ii) is an immediate consequence of (i). □

Proof of Lemma 7.3. Firstly, $\mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_n$ is invariant under the action of Q. Therefore $[u(Q^\intercal\cdot) : x \in \mathbb{S}^2 \mapsto u(Q^\intercal x)] \in \mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_n$, and $u(Q^\intercal\cdot)|_{\mathrm{CS}_N}$ is defined by

$$u(Q^\intercal\cdot)|_{\mathrm{CS}_N} = \sum_{i=1}^{\bar{N}} u(Q^\intercal x_i)\delta_{x_i} \in \mathbb{R}^{\mathrm{CS}_N}.$$

On the other hand, $[u|_{\mathrm{CS}_N} : x \in \mathrm{CS}_N \mapsto u(x)] \in \mathbb{R}^{\mathrm{CS}_N}$, with CS_N left invariant by Q. Then the function $u|_{\mathrm{CS}_N}(Q^\intercal\cdot)$ is well-defined and is given by

$$u|_{\mathrm{CS}_N}(Q^\intercal\cdot) : x \in \mathrm{CS}_N \mapsto u(Q^\intercal x) \in \mathbb{R}^{\mathrm{CS}_N}.$$

At every $x = x_i \in \mathrm{CS}_N$, the two functions have the same value, $u(Q^\intercal x_i)$. □

7.3 A new quadrature on the Cubed Sphere

In this section, we study a new quadrature rule on CS_N; it is defined by interpolation as follows.

Theorem 7.5 (Quadrature rule)**.** *Let $u : \mathbb{S}^2 \to \mathbb{R}$ be a given function. The quadrature rule $\mathcal{Q}_N$ is defined by*

$$\mathcal{Q}_N u := \int_{\mathbb{S}^2} \mathcal{I}_N[u|_{\mathrm{CS}_N}](x)\mathrm{d}\sigma,$$

where $\mathcal{I}_N : \mathbb{R}^{\mathrm{CS}_N} \to \mathcal{U}_N$ is the interpolation operator defined in Theorem 6.2.
(i) Without loss of generality, assume that the first basis function in $\mathcal{U}_N$ is $u_0^1 = \frac{1}{\sqrt{4\pi}}$ (see Definition 6.6). Then, the formula $\mathcal{Q}_N$ can be expressed as follows:

$$\mathcal{Q}_N u = \sum_{j=1}^{\bar{N}} \omega_N(x_j)u(x_j), \text{ with } \omega_N \in \mathbb{R}^{\mathrm{CS}_N} \text{ such that } [\omega_N(x_j)] = \mathbf{V}_d(\tilde{\mathbf{E}}_d^\intercal)^{-1}[\sqrt{4\pi}\ 0\ \cdots\ 0]^\intercal; \quad (7.1)$$

here, the lower triangular matrix $\tilde{\mathbf{E}}_d^\intercal$ and the orthogonal matrix $\mathbf{V}_d$ are given by the QR factorization (6.14).
(ii) The formula $\mathcal{Q}_N$ is exact on $\mathcal{U}_N$, i.e.

$$\forall u \in \mathcal{U}_N, \quad \mathcal{Q}_N u = \int_{\mathbb{S}^2} u(x)\mathrm{d}\sigma.$$

(iii) The rule $\mathcal{Q}_N$ and the weight ω_N are invariant under $\mathcal{G}$, i.e.

$$\forall Q \in \mathcal{G}, \quad \forall u \in \mathcal{U}_N,\ \mathcal{Q}_N(u(Q^\intercal\cdot)) = \mathcal{Q}_N(u), \quad \text{and} \quad \omega_N(Q^\intercal\cdot) = \omega_N.$$

Proof. (i-ii) Firstly, if $u \in \mathcal{U}_N$, $\mathcal{Q}_N$ exactly integrates u, since u coincides with $\mathcal{I}_N u$. In particular, for each basis function, denoted here by $u_j \in \mathcal{U}_N$ with $1 \le j \le \bar{N}$, $\mathcal{Q}_N u_j = \int_{\mathbb{S}^2} u_j(x)\mathrm{d}\sigma$. For $u_1 = \frac{1}{\sqrt{4\pi}}$, $\int_{\mathbb{S}^2} u_1(x)\mathrm{d}\sigma = \sqrt{4\pi}$. For every $2 \le j \le \bar{N}$, $u_j \perp u_1$, which means $\int_{\mathbb{S}^2} u_j(x)\mathrm{d}\sigma = 0$. Then, $[\mathcal{Q}_N u_j]_{1\le j\le \bar{N}} = [\sqrt{4\pi}\ 0\ \cdots\ 0]^\intercal$. Secondly, fix $\omega_N \in \mathbb{R}^{\mathrm{CS}_N}$ such that

$$\omega_N(x_i) = \int_{\mathbb{S}^2} \mathcal{I}_N[\delta_{x_i}]\mathrm{d}\sigma, \quad 1 \le i \le \bar{N}. \quad (7.2)$$

By linearity, we deduce from (6.6) that

$$\mathcal{Q}_N u = \sum_{i=1}^{\bar{N}} \omega_N(x_i) u(x_i) = [u(x_i)]^\intercal [\omega_N(x_i)].$$

Using the basis functions, we obtain

$$(\mathbf{A}_d \tilde{\mathbf{U}}_d)^\intercal [\omega_N(x_i)]_{1\le i\le \bar{N}} = [\mathcal{Q}_N u_j]_{1\le j\le \bar{N}} = [\sqrt{4\pi}\ 0\ \cdots\ 0]^\intercal,$$

where the matrix $\mathbf{A}_d\tilde{\mathbf{U}}_d \in \mathbb{R}^{\bar{N}\times\bar{N}}$, defined in Section 6.4, is non singular, and admits the QR factorization (6.14).
(iii) Fix $Q \in \mathcal{G}$ and $u \in \mathcal{U}_N$. By Theorem 7.2, $u(Q^\intercal\cdot) \in \mathcal{U}_N$. Thus, by (ii) and change of variable,

$$\mathcal{Q}_N(u(Q^\intercal\cdot)) = \int_{\mathbb{S}^2} u(Q^\intercal x)\mathrm{d}\sigma = \int_{\mathbb{S}^2} u(x)\mathrm{d}\sigma = \mathcal{Q}_N(u) \quad (x := Q^\intercal x).$$

Fix now $1 \le i \le \bar{N}$ and $u = \mathcal{I}_N\delta_{x_i} \in \mathcal{U}_N$. By Corollary 7.4, $u(Q^\intercal\cdot) = \mathcal{I}_N[\delta_{x_i}(Q^\intercal\cdot)] \in \mathcal{U}_N$, with $\delta_{x_i}(Q^\intercal\cdot) = \delta_{Qx_i}$. Therefore by (7.2), $\mathcal{Q}_N(u) = \omega_N(x_i)$ and $\mathcal{Q}_N(u(Q^\intercal\cdot)) = \omega_N(Qx_i)$. Then, by invariance of $\mathcal{Q}_N$,

$$\omega_N(x_i) = \mathcal{Q}_N(u) = \mathcal{Q}_N(u(Q^\intercal\cdot)) = \omega_N(Qx_i). \qquad \square$$

The quadrature rule exactly integrates the $\bar{N}$ spherical harmonics of $\mathcal{U}_N$. Taking benefit from the rotational invariance, we prove furthermore that it exactly integrates an infinite number of spherical harmonics.

Corollary 7.6. *The quadrature rule $\mathcal{Q}_N$ exactly integrates $\frac{15}{16}$ of **all** real Legendre spherical harmonics. More precisely, for all $|m| \le n$,*

$$\mathcal{Q}_N(Y_n^m) = \int_{\mathbb{S}^2} Y_n^m(x)\mathrm{d}\sigma, \quad \textit{if} \quad \begin{cases} n \equiv 1\,(2), \\ \textit{or, } m < 0, \\ \textit{or, } m \ge 0 \textit{ and } m \equiv 1,2,3\,(4); \end{cases}$$

equivalently, $\mathcal{Q}_N(Y_n^m) \ne \int_{\mathbb{S}^2} Y_n^m(x)\mathrm{d}\sigma \Rightarrow n \equiv 0\,(2)$, $m \ge 0$ and $m \equiv 0\,(4)$.

Proof. Fix $n \ge 1$ and $|m| \le n$. Then $\int_{\mathbb{S}^2} Y_n^m(x)\mathrm{d}\sigma = 0$. For well chosen n, m, we build a symmetry $Q \in \mathcal{G}$ such that $Y_n^m(Q^\intercal\cdot) = -Y_n^m$. In such cases, we obtain $\mathcal{Q}_N(Y_n^m) = \mathcal{Q}_N(Y_n^m(Q^\intercal\cdot)) = -\mathcal{Q}_N(Y_n^m)$, which proves $\mathcal{Q}_N(Y_n^m) = 0 = \int_{\mathbb{S}^2} Y_n^m(x)\mathrm{d}\sigma$. Recall the spherical coordinates $x(\theta,\phi) = (\cos\theta\cos\phi, \cos\theta\sin\phi, \sin\theta)$, $\phi \in [-\pi,\pi]$, $\theta \in [-\frac{\pi}{2},\frac{\pi}{2}]$, and $Y_n^m(x(\theta,\phi)) := Y_n^m(\theta,\phi)$.

Case 1: $n \equiv 1\,(2)$ and $m \equiv 0\,(4)$. Then $\theta \mapsto P_n^{|m|}(\sin\theta)$ is odd, so is $\theta \mapsto Y_n^m(x(\theta,\phi))$; hence,

$$Y_n^m(Q^\intercal x(\theta,\phi)) = Y_n^m(x(-\theta,\phi)) = -Y_n^m(x(\theta,\phi)), \quad \text{for} \quad Q := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

Case 2: $m < 0$. Then $\phi \mapsto Y_n^m(x(\theta,\phi))$ is odd, so,

$$Y_n^m(Q^\intercal x(\theta,\phi)) = Y_n^m(x(\theta,-\phi)) = -Y_n^m(x(\theta,\phi)), \quad \text{for} \quad Q := \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Case 3: $m \equiv 1,3\,(4)$. Then $m(\phi+\pi) \equiv m\phi + \pi\,(2\pi)$, and

$$Y_n^m(Q^\intercal x(\theta,\phi)) = Y_n^m(x(\theta,\phi+\pi)) = -Y_n^m(x(\theta,\phi)), \quad \text{for} \quad Q := \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Case 4: $m \equiv 2\,(4)$. Then $m(\phi + \frac{\pi}{2}) \equiv m\phi + \pi\,(2\pi)$, and

$$Y_n^m(Q^\intercal x(\theta,\phi)) = Y_n^m(x(\theta,\phi+\tfrac{\pi}{2})) = -Y_n^m(x(\theta,\phi)), \quad \text{for} \quad Q := \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \qquad \square$$

Remark 7.7. In Corollary 7.6 (and its proof), the quadrature rule $\mathcal{Q}_N$ can be replaced by any linear form $\mathcal{Q} : L^2(\mathbb{S}^2) \to \mathbb{R}$ which is invariant under the octahedral group $\mathcal{G}$. In particular, the 15/16-property of the corollary holds for any spherical quadrature with octahedral symmetry. Therefore Corollary 7.6 also holds for the Lebedev rules [149].

Remark 7.8. The ratio 15/16 of the real Legendre basis is obtained asymptotically. In [115,156], a similar approach based on invariance properties reported an asymptotic ratio of 7/8 of the complex Legendre basis exactly integrated. Here, in the proof of Corollary 7.6, the real Legendre basis is used instead. Using this basis allows to prove that exact quadrature actually holds up to 15/16 of all spherical harmonics.

7.4 Numerical results

7.4.1 Symmetry invariance assessment

We begin by two numerical assessments related to interpolation in $\mathcal{U}_N$.

First, we illustrate that the interpolation operator $\mathcal{I}_N$ preserves the invariance property, as stated in Corollary 7.4.(ii). Fix $N = 6$ and consider the series of symmetric functions $g_i \in \mathbb{R}^{\mathrm{CS}_N}$, described in Table 7.1. By construction, each function g_i, $1 \le i \le 6$, is constant along any orbit, *i.e.* $\forall Q \in \mathcal{G}$, $g_i(Q^\intercal \cdot) = g_i$, and is supported by a set of symmetric nodes. For $i \le 5$, g_i takes the value 1 along the orbit of $a_i \in \mathrm{CS}_N$, and the value 0 otherwise. The orbit of a_1 contains the vertices of an octahedron. The orbit of a_2 contains the vertices of a cube. The orbit of a_3 contains the vertices of a cubaoctahedron. The orbit of a_4 is included in the edges of an octahedron. The orbit of a_5 is "generic", with cardinal number 48. In Figure 7.1, we visualize how the symmetry is reflected in the interpolating functions $\mathcal{I}_N g_i \in \mathcal{U}_N$, $1 \le i \le 6$. The octahedral symmetry predicted by Corollary 7.4.(ii) can be observed; the functions $\mathcal{I}_N g_i$ are constant along any orbit.

i	g_i	a_i	$\lvert\operatorname{supp} g_i\rvert$
1	$\frac{1}{8}\sum_{Q\in\mathcal{G}} \delta_{Qa_1}$	$[1\,0\,0]^\intercal$	6
2	$\frac{1}{6}\sum_{Q\in\mathcal{G}} \delta_{Qa_2}$	$\frac{1}{\sqrt{3}}[1\,1\,1]^\intercal$	8
3	$\frac{1}{4}\sum_{Q\in\mathcal{G}} \delta_{Qa_3}$	$\frac{1}{\sqrt{2}}[1\,0\,1]^\intercal$	12
4	$\frac{1}{2}\sum_{Q\in\mathcal{G}} \delta_{Qa_4}$	$(1+\tan^2\frac{\pi}{6})^{-1/2}[1\,0\,\tan\frac{\pi}{6}]^\intercal$	24
5	$\sum_{Q\in\mathcal{G}} \delta_{Qa_5}$	$(1+\tan^2\frac{\pi}{12}+\tan^2\frac{\pi}{6})^{-1/2}[1\,\tan\frac{\pi}{12}\,\tan\frac{\pi}{6}]^\intercal$	48
6	$g_1+g_2+g_3+g_4+g_5$		98

Table 7.1: Grid functions with octahedral symmetry. The function g_i, $1 \le i \le 6$, takes the value 1 on its support and is invariant under $\mathcal{G}$.

Second, we assess the invariance of the interpolation space, stated in Theorem 7.2, and the commutation between interpolation and rotation, stated in Corollary 7.4.(i). For that purpose, for each basis function $u_j \in \mathcal{U}_N$, we compare $u_j(Q^\intercal \cdot) = [\mathcal{I}_N u_j](Q^\intercal \cdot)$ with $\mathcal{I}_N[u_j(Q^\intercal \cdot)]$. Indeed, by linearity, Corollary 7.4.(i) is equivalent to:

$$\forall 1 \le j \le \bar{N},\ \forall Q \in \mathcal{G},\ [\mathcal{I}_N u_j](Q^\intercal \cdot) = \mathcal{I}_N[u_j(Q^\intercal \cdot)].$$

If this condition is achieved, then for each basis function u_j, $u_j(Q^\intercal \cdot) = [\mathcal{I}_N u_j](Q^\intercal \cdot) \in \operatorname{Ran}\mathcal{I}_N = \mathcal{U}_N$, which implies Theorem 7.2.(iii) by linearity. This also implies Theorem 7.2.(ii), due to $u_j(Q^\intercal \cdot) \in$

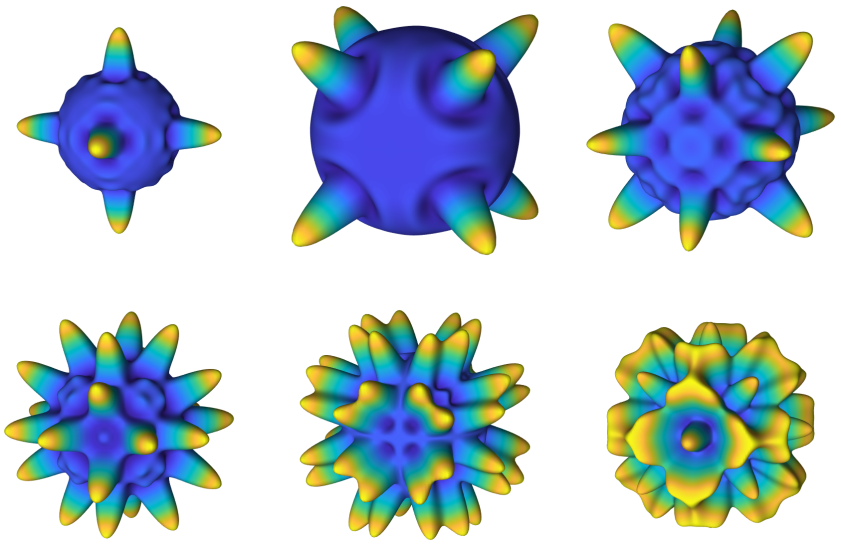


Figure 7.1: Interpolation with octahedral symmetry. For every $1 \leq i \leq 6$, the symmetric function $\mathcal{I}_N g_i$ is represented by the surface $(1.5 + \mathcal{I}_N g_i(x))x$, $x \in \mathbb{S}^2$.

$\mathbb{Y}_n \cap \mathcal{U}_N = \mathcal{V}_n$. We compare the functions on a fine grid CS_M ($M = 33$), by computing the relative error

$$\epsilon_{N,j}(Q) := \frac{\max_{x \in \mathrm{CS}_M} |u_j(Q^\intercal x) - \mathcal{I}_N[u_j(Q^\intercal \cdot)](x)|}{\max_{x \in \mathrm{CS}_M} |u_j(x)|}.$$

Then we compute the maximal error ϵ_N, and we repeat the procedure for several values of N:

$$\epsilon_N := \max\{\epsilon_{N,j}(Q),\, Q \in \mathcal{G},\, 1 \leq j \leq \bar{N}\}, \quad 1 \leq N \leq 16. \tag{7.3}$$

The results reported in Table 7.2 are in agreement with the invariance stated in Theorem 7.2 and Corollary 7.4.(i).

N	1	2	3	4	5	6	7	8
ϵ_N	2.5e-15	3.4e-15	7.8e-15	1.4e-14	9.7e-15	9.3e-15	1.3e-14	1.1e-14
N	9	10	11	12	13	14	15	16
ϵ_N	1.4e-14	1.5e-14	1.9e-14	1.7e-14	3.4e-14	2.2e-14	1.8e-14	2.9e-14

Table 7.2: Numerical invariance: $u_j(Q^\intercal \cdot) = \mathcal{I}_N[u_j(Q^\intercal \cdot)]$, $Q \in \mathcal{G}$, up to relative error ϵ_N (7.3).

7.4.2 Quadrature weight

We have computed the quadrature weight $\omega_N \in \mathbb{R}^{\mathrm{CS}_N}$ for $1 \leq N \leq 32$, and $N = 64$. Some of them are displayed in Figure 7.2. As can been observed, the weight is positive, $\omega_N > 0$, and the maximum value is reached at the center of a panel. Moreover, some statistics of the weights are given in Figure 7.3. It reveals that the distribution of the weights ω_N is quasi-uniform. In particular,

$$\frac{\max \omega_N}{\min \omega_N} \approx \sqrt{2}.$$

We recognize the ratio between the surface element at the center of a panel of CS_N, and the smallest surface element of a panel, as it is derived in [168, Eq. (20)].

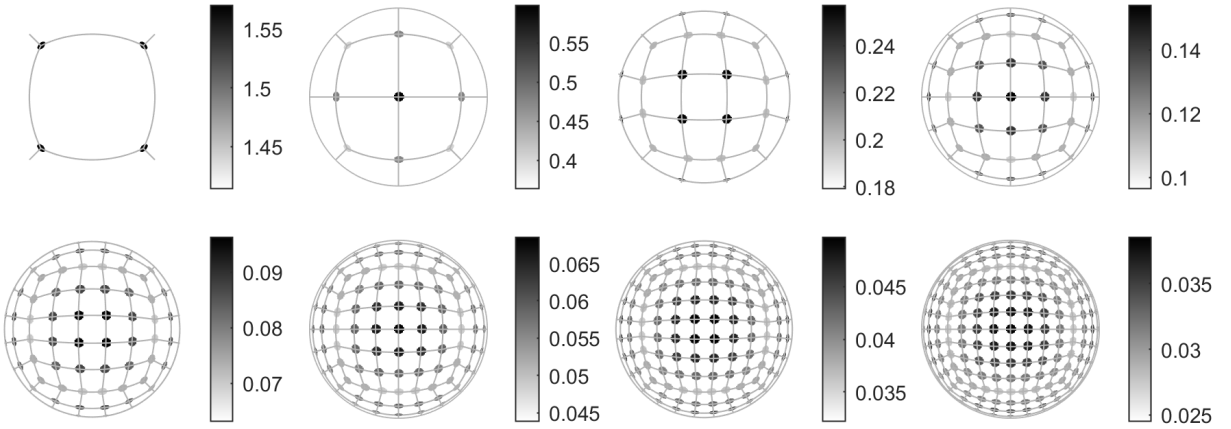


Figure 7.2: Representation of the weight values ω_N, for the eight Cubed Spheres with $1 \leq N \leq 8$.

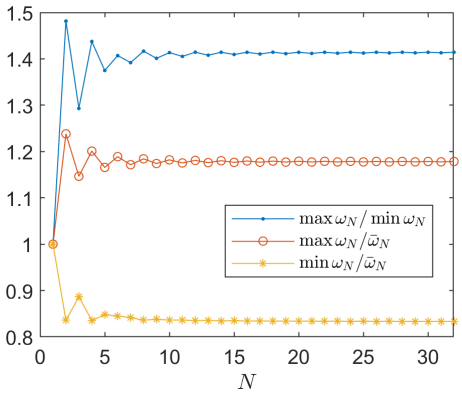


Figure 7.3: Statistical distribution of the weight values ω_N in (7.1), $1 \leq N \leq 32$. The maximum, minimum, and mean values satisfy $\max \omega_N \approx 1.41 \min \omega_N$, $\max \omega_N \approx 1.18\, \bar{\omega}_N$, and $\min \omega_N \approx 0.83\, \bar{\omega}_N$.

7.4.3 Quadrature of test functions

We test the accuracy of the quadrature formula $\mathcal{Q}_N$ on the series of functions reported in Table 7.3. They are displayed in Figure 7.4. These functions serve as testing functions for quadrature assessment. References are indicated in Table 7.3. The exponential function f_1 is a smooth, non trivial function. The Franke function f_2 is a standard test case. The function f_3 is smooth, except near the South pole, where it has an infinite spike. The cosine cap function f_4 is continuous but is not differentiable on the circle $z = \frac{\sqrt{3}}{2}$. The function f_5 is the characteristic function of a spherical cap; it is not continuous. Similarly, the discontinuous function f_6 represents an hemisphere; it is a standard test function.

We report in Table 7.4 the quadrature error

$$\eta_N(f_i) = \left| \int_{\mathbb{S}^2} f_i \mathrm{d}\sigma - \mathcal{Q}_N f_i \right|, \quad N = 1,\, 2,\, 4,\, 8,\, 16,\, 32,\, 64, \quad 1 \leq i \leq 6.$$

Moreover, Table 9.1 reports a rate of convergence $r_N(f_i)$, defined by the equation

$$\eta_{2N}(f_i) = \frac{\eta_N(f_i)}{2^{r_N(f_i)}}.$$

Note that the computations have been performed with `Matlab`, in double precision. In particular the machine epsilon is approximately 2.2×10^{-16}; we do not compute the rate when the relative error is close to this value. For the smooth function f_1, the error rapidly reaches a value which is about 10^{-14}. For the Franke function f_2, a thinner grid is required to reach such values, but a very fast

i	$f_i(x,y,z)$	$\int_{\mathbb{S}^2} f_i(x,y,z)\mathrm{d}\sigma$	Ref.
1	$\exp(x)$	$14.7680137457653\cdots$	[113, 128]
2	$\frac{3}{4}\exp[-\frac{(9x-2)^2}{4} - \frac{(9y-2)^2}{4} - \frac{(9z-2)^2}{4}]$ $+\frac{3}{4}\exp[-\frac{(9x+1)^2}{49} - \frac{9y+1}{10} - \frac{9z+1}{10}]$ $+\frac{1}{2}\exp[-\frac{(9x-7)^2}{4} - \frac{(9y-3)^2}{4} - \frac{(9z-5)^2}{4}]$ $-\frac{1}{5}\exp[-(9x-4)^2 - (9y-7)^2 - (9z-5)^2]$	$6.6961822200736179523\cdots$	[109, 114, 115, 129, 156]
3	$\frac{1}{10}\frac{\exp(x+2y+3z)}{(x^2+y^2+(z+1)^2)^{1/2}}\mathbf{1}(z > -1)$	$4.090220018862976\cdots$	[109]
4	$\cos(3\arccos z)\mathbf{1}(3\arccos z \leq \frac{\pi}{2})$	$\frac{\pi}{8}$	inspired from [109]
5	$\mathbf{1}(z \geq \frac{1}{2})$	π	
6	$\frac{1}{9}[1 + \mathrm{sign}(-9x - 9y + 9z)]$	$\frac{4\pi}{9}$	[114, 115, 129, 156]

Table 7.3: Test functions and exact integration values.

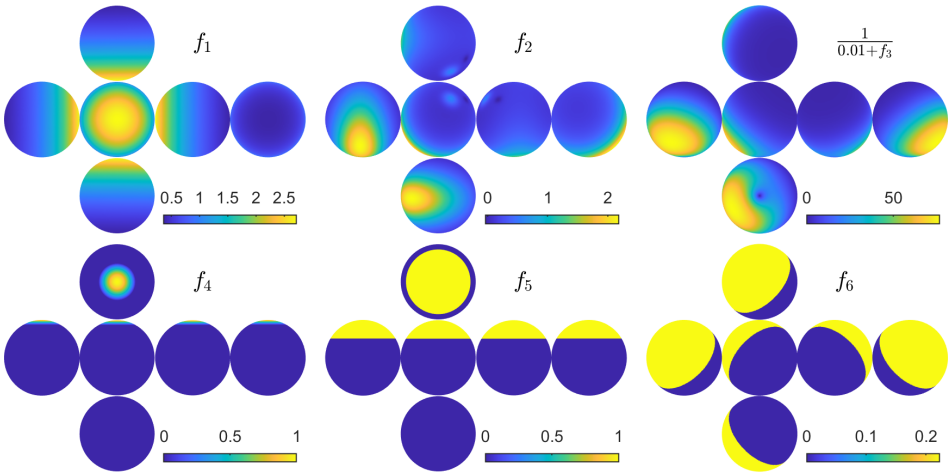


Figure 7.4: Test functions of Table 7.3.

convergence is still observed. For the spike function f_3, the convergence rate is $r_N(f_3) \approx 1$. For the continous cap function f_4, and the discontinuous one f_5, the error slowly decrease, at a convergence rate which depends on the grid size. For the cap function f_6, which is discontinous and "symmetric" (supported by a hemisphere), the error is close to the machine epsilon, independently of the grid size.

N	$\bar{N}$	$\eta_N(f_1)$	$\eta_N(f_2)$	$\eta_N(f_3)$	$\eta_N(f_4)$	$\eta_N(f_5)$	$\eta_N(f_6)$
1	8	4.8e-02	8.2e-01	2.4e-01	3.9e-01	3.1e+00	6.7e-16
2	26	2.0e-06	1.5e-02	1.7e-02	2.1e-01	9.9e-01	0.0e+00
4	98	1.2e-14	2.2e-03	7.8e-03	2.0e-02	6.7e-02	2.2e-16
8	386	1.8e-15	9.0e-06	3.8e-03	4.8e-03	6.4e-02	6.7e-16
16	1538	7.1e-15	5.5e-09	1.9e-03	3.0e-04	1.5e-02	6.7e-16
32	6146	5.3e-15	1.8e-15	9.5e-04	3.1e-04	7.9e-03	6.7e-16
64	24578	3.6e-15	1.8e-15	4.8e-04	1.9e-07	2.6e-03	4.4e-16

Table 7.4: Quadrature error $\eta_N(f_i) = |\int_{\mathbb{S}^2} f_i \mathrm{d}\sigma - \mathcal{Q}_N f_i|$.

N	$r_N(f_1)$	$r_N(f_2)$	$r_N(f_3)$	$r_N(f_4)$	$r_N(f_5)$	$\bar{r}_N(f_1)$	$\bar{r}_N(f_2)$	$\bar{r}_N(f_3)$	$\bar{r}_N(f_4)$	$\bar{r}_N(f_5)$
1	15	5.8	3.9	0.93	1.7	15	2.8	5.4	2.5	-0.0033
2	27	2.7	1.1	3.4	3.9	26	3.3	1.1	3.3	1.8
4	2.8	7.9	1	2.1	0.077	2.1	4.1	1	1.9	1.3
8		11	1	4	2.1		13	1.2	2.8	1.7
16		22	1	-0.047	0.92		24	0.92	2.2	1.4
32			1	11	1.6			0.93	2.7	1.6

Table 7.5: Convergence rate $r_N(f_i)$ of the error $\eta_N(f_i)$, and convergence rate $\bar{r}_N(f_i)$ of the average error $\bar{\epsilon}_N(f_i)$, over 1000 random orthogonal transformations of the grid.

7.4.4 Sensitivity to the grid orientation

Here we consider more closely the accuracy of the rule $\mathcal{Q}_N$: we modify randomly the orientation of the grid, [129, 156]. We compute

$$\epsilon_N(f_i, Q) = \left| \int_{\mathbb{S}^2} f_i \mathrm{d}\sigma - \mathcal{Q}_N f_i(Q^\intercal \cdot) \right|,$$

where Q browses a set of 1000 randomly selected orthogonal matrices (uniform law in $[0, 2\pi]$ for the Euler angles, and uniform law in $\{-1, 1\}$ for the orientation). The worst error, the average error and their ratio are defined by

$$\varepsilon_N(f_i) = \max_Q \epsilon_N(f_i, Q), \quad \bar{\epsilon}_N(f_i) = \tfrac{1}{1000} \sum_Q \epsilon_N(f_i, Q), \quad \rho_N(f_i) = \frac{\bar{\epsilon}_N(f_i)}{\varepsilon_N(f_i)}.$$

The worst error $\varepsilon_N(f_i)$ and the ratio $\rho_N(f_i)$ are displayed[1] in Figure 7.5. We report in Table 9.1 a convergence rate $\bar{r}_N(f_i)$ of the average error $\bar{\epsilon}_N(f_i)$, defined by

$$\bar{\epsilon}_{2N}(f_i) = \frac{\bar{\epsilon}_N(f_i)}{2^{\bar{r}_N(f_i)}}.$$

The worst errors $\varepsilon_N(f_1)$ and $\varepsilon_N(f_2)$ fastly decrease, and $\varepsilon_N(f_6)$ is zero, up to rounding errors. This indicates that the quadrature rule $\mathcal{Q}_N$ efficiently integrates the smooth functions f_1, f_2, and the symmetric cap function f_6, independently of the grid orientation. For the function f_4, which is continuous and non differentiable, the worst error $\varepsilon_N(f_4)$ decreases at constant rate. The decrease of the worst error $\varepsilon_N(f_5)$ of the "generic" cap function f_5, which is discontinuous, is slower. And for the spike function f_3, the worst error $\varepsilon_N(f_3)$ slowly decreases, with oscillations.

[1] In order to clarify the figure, we have elimated the following ratios: $\rho_N(f_1), N > 3$, and $\rho_N(f_6)$. Indeed, these ratios are "large", because the associated errors are almost zero.

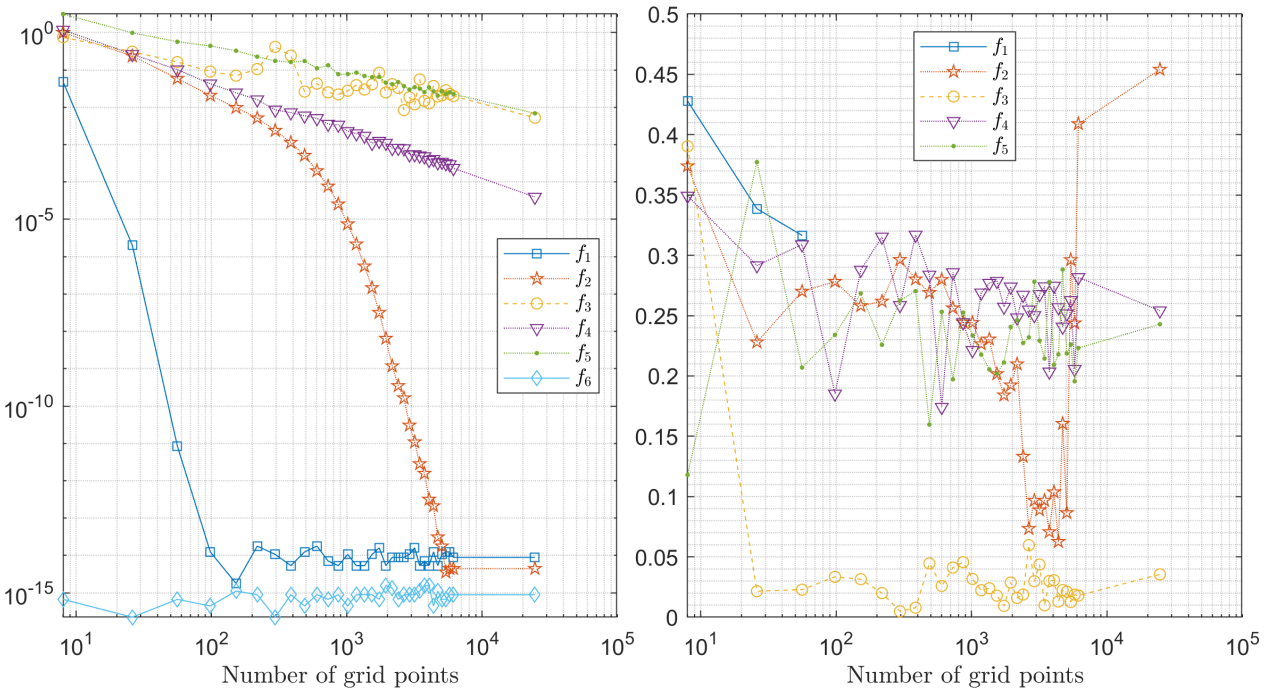


Figure 7.5: Statistics of the quadrature error $\epsilon_N(f_i, Q) = |\int_{\mathbb{S}^2} f_i \mathrm{d}\sigma - \mathcal{Q}_N f_i(Q^\intercal \cdot)|$, where Q scans a set of 1000 random orthogonal matrices. Left: worst error $\varepsilon_N(f_i)$. Right: ratio $\rho_N(f_i) = \bar{\epsilon}_N(f_i)/\varepsilon_N(f_i)$ of the average error divided by the worst one.

Roughly speaking, Figure 7.5 indicates that

$$\bar{\epsilon}_N(f_i) \approx 0.25\varepsilon_N(f_i),\ i \neq 3, \quad \bar{\epsilon}_N(f_3) \approx 0.025\varepsilon_N(f_3).$$

Except for f_3, the worst error is not very large in comparison with the average error (factor 4). This indicates that the result is almost insensitive to the grid orientation. For the function f_3 with a spike, the situation is different (factor 40); the error is sensitive to the grid orientation. Concerning the speed of convergence, we note $\bar{r}_N(f_3) \approx 1$ for the spike function, $\bar{r}_N(f_5) \approx 1.4$ for the discontinuous cap function, $\bar{r}_N(f_4) \approx 2.6$ for the continuous one. The average errors for f_1, f_2 and f_6 converge fastly, since it was already the case for the worst errors.

7.4.5 Comparison with other quadrature rules

We compare our quadrature rule $\mathcal{Q}_N$ with some spherical quadrature rules of the literature, summarized in Table 7.6.

Abbr.	Description	Ref.
CS-BBC21	Interpolation on the Cubed Sphere by spherical harmonics	This work ($\mathcal{Q}_N$)
CS-CP18	Octahedral quadrature on the Cubed Sphere, by least-square	[156]
CS-BC18	Corrected bivariate trapezoidal on the Cubed Sphere	[115]
Lebedev	Gauss quadrature, invariant under the octahedral group	[149, 150]
t-design	Spherical t-design	[182, 183]

Table 7.6: Quadrature of the literature used for comparison.

Rules on the Cubed Sphere We use two other rules on the Cubed Sphere; CS-BC18 is a correction of some bivariate trapezoidal rule, CS-CP18 is an octahedral rule which minimizes some least-square error concerning the integration of Legendre spherical harmonics.

Optimal quadrature rules We also use "optimal" quadrature rules, whose distribution of nodes is "optimized". Firstly, our rule is invariant under the octahedral group $\mathcal{G}$, so we compare with the Lebedev rule, which is an optimal octahedral rule. Indeed, the optimal grids/weights of Lebedev maximize the degree of precision, under the constraint of invariance under $\mathcal{G}$. Secondly, our weights are quasi-uniform, so we compare with spherical t-designs. This rules have equal weights and degree of precision t; the associated spherical grids have $\sim \frac{t^2}{2}$ nodes, which is the optimal order.

Results are given on Figure 7.6. The worst error after 1000 random orthogonal matrices is plotted related to the number of grid points using different quadrature rules.

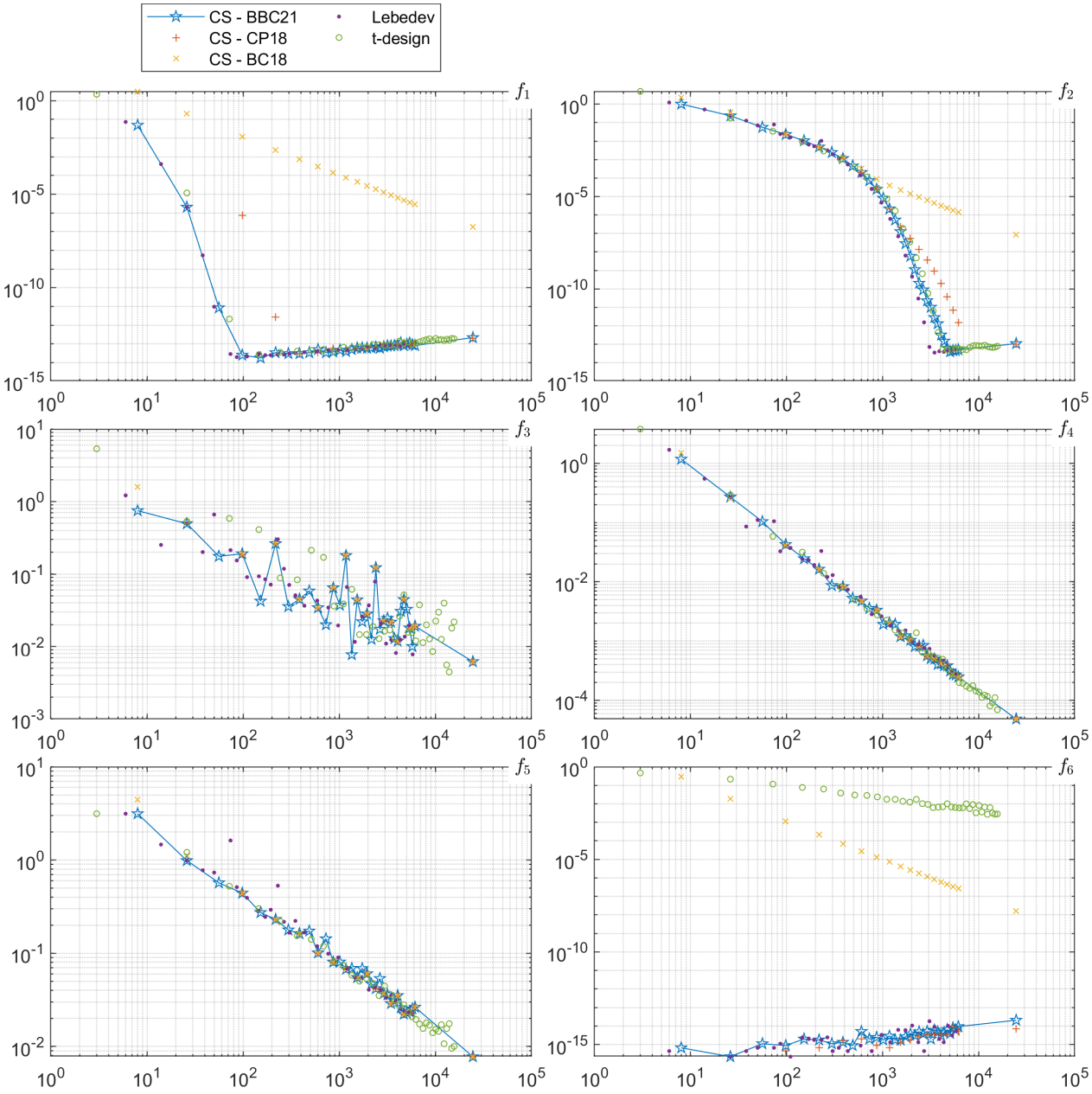


Figure 7.6: Worst quadrature error (for 1000 random orthogonal transformations of the grid), versus the number of grid points.

Comparison on the Cubed Sphere Among the quadrature rules on CS_N, the new rule $\mathcal{Q}_N$ outperforms CS-BC18 for the smooth functions f_1 and f_2, and for the symmetric cap function

f_6. The rules $\mathcal{Q}_N$ and CS-CP18 give similar accuracy for most of the cases, with the following exceptions. The rule $\mathcal{Q}_N$ integrates f_1 more accurately than CS-CP18 before convergence, and $\mathcal{Q}_N$ converges slightly faster than CS-CP18 for f_2.

Comparison with optimal rules For the smooth functions f_1 and f_2, the rules $\mathcal{Q}_N$ and t-design have similar accuracy, whereas the Lebedev rule converges slightly faster. For the function f_3 with a spike, the worst errors are almost similar; they decay slowly with oscillations. For the cap functions f_4 and f_5, the methods converge slowly with similar accuracy. For the "symmetric" cap function f_6, $\mathcal{Q}_N$ and the Lebedev rule are exact (up to rounding errors) and give better accuracy than the t-design rule.

Overall, the rule $\mathcal{Q}_N$ on a fixed grid CS_N displays remarkable accuracy, compared to "optimal" quadrature methods, which require "optimal" grids (Lebedev and t-design rules).

7.4.6 Accuracy of the new quadrature rule

The quadrature rule $\mathcal{Q}_N$ is designed to integrate exactly any spherical harmonics belonging to the space $\mathcal{U}_N$. In addition, it integrates 15/16 of *all* Legendre spherical harmonics (see Corollary 7.6). Here, we numerically display detailed accuracy properties of the rule $\mathcal{Q}_N$.

First, for a selected set of tolerances $\epsilon = 10^{-p}$, we give the degree of precision $d_N(\epsilon)$, defined as the largest integer such that

$$\forall |m| \leq n \leq d_N(\epsilon), \left| \int_{\mathbb{S}^2} Y_n^m \mathrm{d}\sigma - \mathcal{Q}_N Y_n^m \right| \leq \epsilon.$$

The results are reported in Table 7.7. It is observed that except for $N = 3, 4, 64$, the degree $d_N(10^{-14})$ is $2N+1$ if N is odd, and $2N+3$ if N is even. For the three exceptions, the degree is found higher than the generic one. We have $d_N(10^{-14}) = 4N-1$ for $N = 3$, 4, and $d_{64}(10^{-14}) = 2\cdot 64+11$. Furthermore, the Table 7.7 implicitly displays an accuracy information obtained for some of the Legendre spherical harmonics that are not exactly integrated. For example in the case $N = 8$, the first error above the threshold 10^{-14} belongs to the interval $(10^{-6}, 10^{-4}]$. This error is obtained for the degree $n = 20$ (since the rule is exact for odd degrees).

Second, we focus on the quadrature errors

$$\eta(Y_n^m) = \left| \int_{\mathbb{S}^2} Y_n^m \mathrm{d}\sigma - \mathcal{Q}_N Y_n^m \right|, \quad n \equiv 0\,(2),\ m \geq 0 \text{ and } m \equiv 0\,(4). \tag{7.4}$$

Here, we consider the series of the 1/16 of all the Legendre spherical harmonics which are possibly non exactly integrated by $\mathcal{Q}_N$, (see Corollary 7.6). We have computed the quadrature error for this series for $n \leq 1024$, and for the two grids CS_N with $N = 8$ (386 nodes), and $N = 31$ (5768 nodes). The computed errors are displayed in Figure 7.7 using both an histogram form, and a cumulative distribution function. As observed, some errors are zero (up to rounding errors). This is consistent with the data in Table 7.7 (if $n \leq d_N(10^{-14})$, $\eta(Y_n^m) \leq 10^{-14}$). Note also that the largest obverved errors belong to the interval $(1, 10)$.

Finally, we further develop these observations by comparing the rule $\mathcal{Q}_N$ with the Lebedev rules. As noted in Remark 7.7 above, errors with the Lebedev quadrature can occur only with the same set of Legendre functions (referred to as the "1/16 serie"). Therefore we numerically compare the accuracy of the Lebedev rules with the rule $\mathcal{Q}_N$ on this series. Figure 7.7 reports a comparison between the two Lebedev grids with 434 nodes and 5810 nodes and the Cubed Sphere rule $\mathcal{Q}_N$ with 386 nodes ($N = 8$) and 5810 nodes ($N = 31$), respectively. It is observed that the Lebedev rules exactly integrate a larger set of spherical harmonics; this was somehow expected, since the Lebedev rules are defined to maximize the degree of precision over octahedral grids. But surprisingly, the distribution of the largest errors of the Lebedev rule is very similar with the one of the rule $\mathcal{Q}_N$. In particular, the largest errors of $\mathcal{Q}_N$, defined on the fixed octahedral grid CS_N, are not above the

largest errors of the Lebedev's optimal grid. We even notice on the cumulative density function plots that the number of errors below a moderate tolerance ϵ can be slightly larger with the rule $\mathcal{Q}_N$; this is observed in particular with $N = 31$ and $\epsilon = 10^{-7}$. These observations indicate the interest of the rule $\mathcal{Q}_N$ when compared with an optimal rule.

N	N	$d_N(10^{-14})$	$d_N(10^{-12})$	$d_N(10^{-10})$	$d_N(10^{-8})$	$d_N(10^{-6})$	$d_N(10^{-4})$
1	8	3	3	3	3	3	3
2	26	7	7	7	7	7	7
3	56	11	11	11	11	11	11
4	98	15	15	15	15	15	15
5	152	11	11	11	11	11	11
6	218	15	15	15	15	15	17
7	296	15	15	15	15	15	17
8	386	19	19	19	19	19	21
9	488	19	19	19	19	19	23
10	602	23	23	23	23	23	27
11	728	23	23	23	23	23	27
12	866	27	27	27	27	29	33
13	1016	27	27	27	27	29	33
14	1178	31	31	31	31	33	39
15	1352	31	31	31	31	33	41
16	1538	35	35	35	35	39	45
17	1736	35	35	35	35	39	47
18	1946	39	39	39	41	43	51
19	2168	39	39	39	41	45	55
20	2402	43	43	43	45	49	59
21	2648	43	43	43	45	49	65
22	2906	47	47	47	49	53	65
23	3176	47	47	47	51	55	71
24	3458	51	51	51	55	59	75
25	3752	51	51	51	55	61	79
26	4058	55	55	57	59	65	85
27	4376	55	55	57	59	65	89
28	4706	59	59	61	65	69	95
29	5048	59	59	61	65	71	97
30	5402	63	63	65	69	75	101
31	5768	63	63	65	69	77	105
32	6146	67	67	71	73	81	111
64	24578	139	143	147	155	183	255

Table 7.7: Quadrature rule $\mathcal{Q}_N$: observed degree of precision $d_N(\epsilon)$ for various tolerances ϵ.

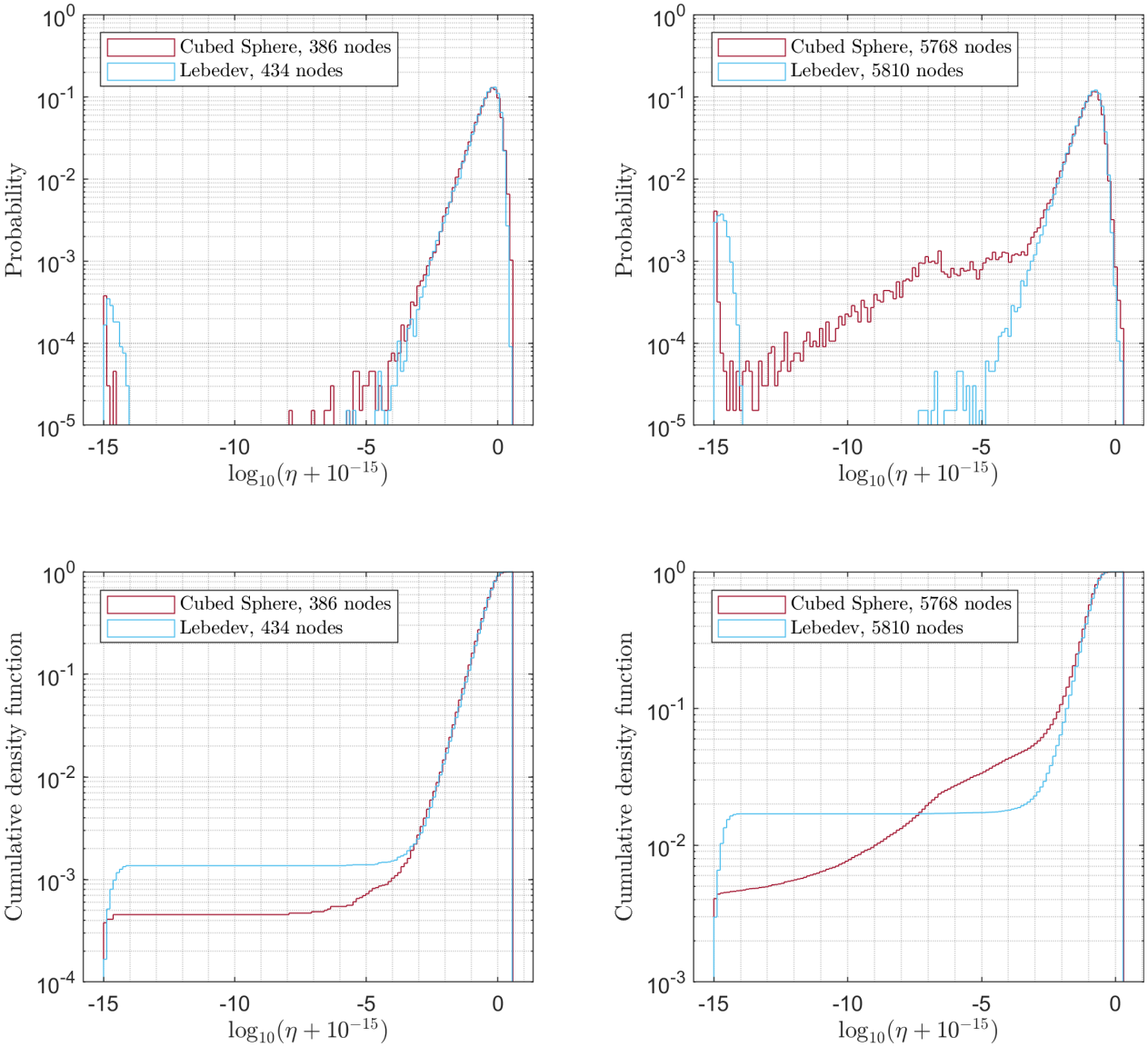


Figure 7.7: Comparison of the errors on the 1/16 series of spherical harmonics (see Corollary 7.6) between the rule $\mathcal{Q}_N$ and Lebedev's rules for two pairs of grids. Left column: CS (386 nodes)/Lebedev (434 nodes). Right column: CS (5768 nodes)/Lebedev (5810 nodes). The quadrature error η is reported for the spherical harmonics Y_n^m, with $n \equiv 0\,(2)$, $m \equiv 0\,(4)$, $0 \leq m \leq n \leq 1024$. Top: an histogram with logarithmic rescaling of the errors η (7.4) is displayed for both rules. Bottom: the cumulative density function (cdf) in logarithmic scale for both rules is reported. On these plots, the range of the logarithmic error $\log(\eta + 10^{-15})$ has been uniformly divided into 128 classes; for any class $[c_1, c_2)$, the probability (top line) represents the percentage of errors η such that $10^{c_1} \leq \eta + 10^{-15} < 10^{c_2}$, whereas the cumulative density (bottom line) represents the percentage of errors η such that $\eta + 10^{-15} < 10^{c_2}$. As a conclusion, the Lebedev's rules exactly integrate more spherical harmonics, but the distributions of the largest errors are similar; moreover, for a large grid, the percentage of errors below a moderate threshold ($\epsilon = 10^{c_2}$) is larger for the rule $\mathcal{Q}_N$ (bottom-right with $\epsilon > 10^{-7}$).

7.5 Conclusion

We have designed a new quadrature rule $\mathcal{Q}_N$ on the Cubed Sphere CS_N. It has the property to be quasi-uniform with positive weights. The numerical results can be compared in accuracy with optimal rules such as t-designs and Lebedev rules. This supports the claim of the "approximation power" of the Cubed Sphere. Among the questions open, a better convergence analysis must be performed. Proving the positivity of the weights is also an important goal. Overall, the symmetry properties of the Cubed Sphere as a support for quadrature seems a promising topic.

7.A Quadrature rules data

The data for the various rules used in this study can be found as follows:

- The new rule associated to the Cubed Sphere nodes is available on the open archiv `https://hal.archives-ouvertes.fr/hal-03223150/file/xyzwCSN.zip`
- For the Lebedev rules, we have used the `Matlab` function `getLebedevSphere` (by R.M. Parrish). The code is available on `https://fr.mathworks.com/matlabcentral/fileexchange/27097-getlebedevsphere`
- The t-designs have been found on R.S. Womersley webpage `https://web.maths.unsw.edu.au/~rsw/Sphere/EffSphDes/sf.html`

Chapter 8

Least squares approximation on the Cubed Sphere

8.1 Introduction

This chapter, extracted from [11], studies least squares fitting by a spherical harmonic on the Cubed Sphere grid.

We consider the approximation of functions defined on the Cubed Sphere by means of spherical harmonics. Assume that a grid function $y \in \mathbb{R}^{\mathrm{CS}_N}$ is known, which means that $y(\xi)$, $\xi \in \mathrm{CS}_N$, are values given at the nodes ξ of CS_N. We approximate these data by a spherical harmonic $f \in \mathcal{Y}_D$, where $\mathcal{Y}_D = \mathbb{Y}_0 \oplus \cdots \oplus \mathbb{Y}_D$ is the space of spherical harmonics with degree at most D. The standard least squares approximation problem is

$$\inf_{f \in \mathcal{Y}_D} \sum_{\xi \in \mathrm{CS}_N} |f(\xi) - y(\xi)|^2. \tag{LS}$$

Our main observation is that the choice $D = 2N - 1$ leads to a well posed and well conditioned problem. In addition, the resulting spherical harmonic possesses interesting properties for approximating a function given at the nodes of CS_N only. These facts are assessed theoretically and numerically hereafter.

In [10], we have introduced a spherical harmonics subspace dedicated to Lagrange interpolation on the Cubed Sphere, as presented in Chapter 6. In practice, this space is a direct sum $\mathcal{Y}_{2N-1} \oplus \mathcal{Y}'$. The second subspace $\mathcal{Y}'$ complements $\mathcal{Y}_{2N-1}$; it is such that $\mathcal{Y}' \subsetneq \mathbb{Y}_{2N} \oplus \cdots \oplus \mathbb{Y}_{3N}$. This interpolation framework has been used in [9] to define new spherical quadrature rules of accuracy comparable to optimal ones (Lebedev rules), as presented in Chapter 7. Here, we show that the first subspace $\mathcal{Y}_{2N-1}$ is a suitable choice if one wants a least squares approximant instead of an interpolant.

Approximating and interpolating data on the sphere by spherical harmonics is an old and important topic. It is still widely used nowadays in many areas in physics such as quantum chemistry, numerical climatology, cosmology, gravitation, neutronic, etc. It is also central in harmonic analysis on spheres and balls since it is the three dimensional counterpart of trigonometric approximation. For fundamental and applied aspects of spherical harmonics analysis, refer to the two recent monographs [113,121] (theory and applications). Concerning applications in geomathematics, many chapters in the reference [130] are concerned with spherical harmonics. Regarding specifically least squares, recent works include [111,136].

The outline is as follows. In Section 8.2 the setup of the problem is given. A general positive weight function is included to define the least squares functional. Theoretical results are given in Section 8.3. These results in particular concern estimates of the condition number of the collocation matrix. Section 8.4 is devoted to the structure of the matrix involved in the least squares problem (LS). In particular, the attractive block structure of this matrix is described in case of a rotationally invariant weight function. Finally, various numerical results are reported in Section 8.5.

8.2 Setup of the Least Squares problem

Our notation is as follows. For any $N \geq 1$, the equiangular Cubed Sphere CS_N is a set of $\bar{N} = 6N^2 + 2$ nodes $\xi \in \mathbb{S}^2$, defined in (5.1). The space of grid functions $\mathbb{R}^{\mathrm{CS}_N}$ has been introduced in Subsection 6.2.2; for any function $g : x \in \mathbb{S}^2 \mapsto g(x) \in \mathbb{R}$, the grid function $g|_{\mathrm{CS}_N} \in \mathbb{R}^{\mathrm{CS}_N}$ is defined by $g|_{\mathrm{CS}_N}(\xi) = g(\xi)$, $\xi \in \mathrm{CS}_N$.

The spherical harmonic Y_n^m with index (n, m) is defined in spherical coordinates (θ, ϕ) by (6.3); it is such that

$$Y_n^m(x(\theta,\phi)) = q_n^m(\sin\theta) \cdot (\cos\theta)^{|m|} \cdot \begin{cases} \sin m\phi, & m < 0, \\ \cos m\phi, & m \geq 0, \end{cases} \tag{8.1}$$

where $x(\theta, \phi)$ is defined in (6.2), and q_n^m is the polynomial of degree $n - |m|$, with the parity of $n + |m|$, defined by

$$q_n^m(t) = \sqrt{\tfrac{(n+1/2)(n-|m|)!}{\pi(n+|m|)!}} \cdot \left(\tfrac{\mathrm{d}^{|m|+n}}{\mathrm{d}t^{|m|+n}} \tfrac{1}{2^n n!} (t^2-1)^n \right) \cdot \begin{cases} -1, & m < 0, \\ \frac{1}{\sqrt{2}}, & m = 0, \\ 1, & m > 0. \end{cases} \tag{8.2}$$

For any $D \geq 0$, the space $\mathcal{Y}_D$ of spherical harmonics with degree at most D, given by (6.5), has dimension $\dim \mathcal{Y}_D = (D+1)^2$.

Let $\omega(\xi) > 0$, $\xi \in \mathrm{CS}_N$, be a given positive weight function. Let $y(\xi)$, $\xi \in \mathrm{CS}_N$, be a set of data given at the nodes of the CS_N. We define a functional $\mathcal{L}$ by

$$f \mapsto \mathcal{L}(f) = \sum_{\xi \in \mathrm{CS}_N} \omega(\xi) |f(\xi) - y(\xi)|^2. \tag{8.3}$$

We consider the least squares problem: find $f \in \mathcal{Y}_D$ solution of

$$\inf_{f \in \mathcal{Y}_D} \mathcal{L}(f). \tag{WLS}$$

We also use the quadrature rule $\mathcal{Q}$ associated to ω: for $f : \mathbb{S}^2 \to \mathbb{R}$,

$$\begin{aligned} \mathcal{Q}(f) &= \sum_{\xi \in \mathrm{CS}_N} \omega(\xi) f(\xi) \\ &= \int_{\mathbb{S}^2} f(x) \mathrm{d}\sigma - e_N(f), \end{aligned} \tag{8.4}$$

where e_N denotes the quadrature error. In the particular case where the data y are such that $y = g|_{\mathrm{CS}_N}$ for a given function g, we have

$$\mathcal{L}(f) = \|f - g\|^2_{L^2(\mathbb{S}^2)} - e_N(|f-g|^2). \tag{8.5}$$

For fixed values of N and D, we call the *Vandermonde matrix* of the problem the rectangular matrix A_N^D defined by

$$A_N^D = \left[Y_n^m(\xi) \right]_{\substack{\xi \in \mathrm{CS}_N \\ |m| \leq n \leq D}} \in \mathbb{R}^{\bar{N} \times (D+1)^2}. \tag{8.6}$$

We define the diagonal matrix $\Omega_N \in \mathbb{R}^{\bar{N} \times \bar{N}}$ by

$$\Omega_N = \mathrm{diag}(\omega(\xi))_{\xi \in \mathrm{CS}_N} \in \mathbb{R}^{\bar{N} \times \bar{N}}. \tag{8.7}$$

In vector form, the problem (WLS) is expressed as

$$\inf_{\hat{f} \in \mathbb{R}^{(D+1)^2}} \|\Omega_N^{1/2} (A_N^D \hat{f} - \boldsymbol{y})\|^2, \tag{8.8}$$

where $\boldsymbol{y} = [y(\xi)]_{\xi \in \mathrm{CS}_N} \in \mathbb{R}^{\bar{N}}$. Uniqueness for (8.8), or equivalently for (LS) or (WLS), is equivalent to the injectivity of A_N^D. In this case, (8.8) is equivalent to the linear system

$$A_N^{D\intercal} \Omega_N A_N^D \, \hat{f} = A_N^{D\intercal} \Omega_N \, \boldsymbol{y}. \tag{8.9}$$

A natural interpretation of (8.9) is as follows. Consider the following analog of the Discrete Fourier Transform (DFT) of the data $y = g|_{\mathrm{CS}_N}$, located at the nodes of CS_N instead of at the $\theta_j = 2j\pi/N \in [0, 2\pi), j = 0, \ldots, N$, in the standard DFT. The Fourier-like coefficients are the components of the vector

$$\begin{aligned} \mathrm{DFT}(\boldsymbol{y}) &= \Big[\sum_{\xi \in \mathrm{CS}_N} \omega(\xi) Y_n^m(\xi) y(\xi) \Big]_{(n,m)} \\ &= A_N^{D\intercal} \Omega_N \, \boldsymbol{y}. \end{aligned} \tag{8.10}$$

On the other hand, the analog of the Inverse Discrete Fourier Transform (IDFT) of a set of data $\hat{f} = [\hat{f}_n^m]_{0 \le |m| \le n \le D}$ is the grid function

$$\begin{aligned} \mathrm{IDFT}[\hat{f}](\xi) &= \sum_{0 \le |m| \le n \le D} \hat{f}_n^m Y_n^m(\xi), \quad \xi \in \mathrm{CS}_N, \\ &= \Big[(A_N^D) \hat{f} \Big](\xi). \end{aligned}$$

This means that in matrix form, A_N^D coincides with the IDFT operator. Therefore in terms of DFT/IDFT transforms, the solution $f \in \mathcal{Y}_D$ of (8.9) has coefficients $\hat{f} = [\hat{f}_n^m]$ solution of

$$\mathrm{DFT}\Big(\mathrm{IDFT}[\hat{f}] - \boldsymbol{y} \Big) = 0.$$

For any N, there is a maximal degree D such that the matrix A_N^D is injective (full column rank), thus guaranteeing that (WLS) has a unique solution. The proof consists in observing that such degrees D form a nonempty set of integers. We have $\operatorname{rank} A_N^D \le \min\big((\bar{N}, (D+1)^2\big)$. Therefore assuming that A_N^D has full column rank implies that $(D+1)^2 \le \bar{N}$ which means

$$D \le \bar{N}^{1/2} - 1 \approx 2.45N - 1.$$

Fix a value N, and consider the Cubed Sphere CS_N. How to select $N \mapsto D_N$ in order for the two following conditions to hold?

$$\left\{ \begin{array}{l} \text{(i) For every degree } D \le D_N, \text{ the Vandermonde matrix } A_N^D \text{ is injective so that} \\ \text{the least squares problem (LS) has a unique solution.} \\ \text{(ii) The condition number } \mathrm{cond}(A_N^{D_N}) \text{ is bounded above for } N \to +\infty. \end{array} \right. \tag{P}$$

In other words, the matrix $A_N^{D_N}$ is required to satisfy both injectivity and asymptotic stability. In what follows, we assess that

$$D_N = 2N - 1 \tag{8.11}$$

is a natural candidate for (P) to be fullfilled.

Remark 8.1. Note that the value $D = 2N-1$ corresponds to the Nyquist cutoff angular frequency of a signal sampled with stepsize $\pi/(2N)$ (one dimensional problem). Note also that for a given integer N, it may exist $D > 2N-1$ such that A_N^D is injective. However, we are interested in a generic value of D, expressed in function of N such that the property (P) holds.

In Section 8.3, several theoretical results are proved, supporting (8.11). And Section 8.5 reports numerical results further supporting this claim.

8.3 Theoretical results

In this section we prove several facts supporting that $D = 2N - 1$ is a truncation value fullfilling the property (P). Specifically we will show that

1. For $1 \leq N \leq 4$, the condition $D \leq 2N - 1$ is equivalent to the injectivity of the matrix A_N^D, (Proposition 8.4).

2. For $N \geq 5$, we show that $D \leq N + 2$ implies injectivity of the matrix A_N^D (Proposition 8.5). Combined with (8.2), this gives that a condition on the largest D for A_N^D to be injective is that

$$N + 2 \leq D \leq \sqrt{6N^2 + 2} - 1$$

3. Finally, Theorem 8.7 shows that $D = 2N$ gives that the condition number of A_N^{2N} is asymptotically unbounded. Thus $D \geq 2N$ does not satisfy (ii) in (P) and therefore one must select $D \leq 2N - 1$.

Remark 8.2. A full proof of the fact that the matrix A_N^{2N-1} is injective for all N is not yet available.

8.3.1 The case $1 \leq N \leq 4$

The case $1 \leq N \leq 4$ corresponds to a small Cubed Sphere grid ranging from $\bar{N} = 8, (N = 1)$ nodes to $\bar{N} = 98, (N = 4)$ nodes. Consider the spherical harmonic $Y_{2N}^{-2N} \in \mathbb{Y}_{2N}$, given by, see (8.1)

$$Y_{2N}^{-2N}(x(\theta, \phi)) = q_{2N}^{-2N}(\sin\theta) \cdot \cos^{2N}\theta \cdot \sin(2N\phi).$$

By shifting the angle ϕ by $\pi/4$ we obtain $f_N \in \mathbb{Y}_{2N}$ defined by

$$f_N(x(\theta, \phi)) = Y_{2N}^{-2N}(x(\theta, \phi - \tfrac{\pi}{4})). \tag{8.12}$$

Lemma 8.3 (Function f_N restricted to CS_N for $N \leq 4$)**.** *For $1 \leq N \leq 4$, the function $f_N \in \mathbb{Y}_{2N}$ vanishes at all nodes of CS_N ($f|_{\mathrm{CS}_N} \equiv 0$). This implies that A_N^D has not full column rank if $N \leq 4$ and $D \geq 2N$.*

Proof. The spherical harmonic f_N is deduced from Y_{2N}^{-2N} by a rotation of $\pi/4$ around the pole axis. By invariance of $\mathbb{Y}_{2N}$ by rotation, we have $f_N \in \mathbb{Y}_{2N}$. In addition, for any $N \leq 4$, it turns out that CS_N is contained in the set M_N of meridians defined by

$$M_N := \Big\{ x(\theta, \phi) : \theta \in [-\tfrac{\pi}{2}, \tfrac{\pi}{2}],\ \phi \equiv \tfrac{\pi}{4}\,(\tfrac{\pi}{2N}) \Big\}. \tag{8.13}$$

Along these meridians, the longitude angle ϕ is such that $2N(\phi - \frac{\pi}{4}) \equiv 0\,(\pi)$, hence

$$f_N(x(\theta, \phi)) = q_{2N}^{-2N}(\sin\theta) \cdot \cos^{2N}\theta \cdot \sin\big(2N(\phi - \tfrac{\pi}{4})\big) = 0. \tag{8.14}$$

This implies that $f(\xi) = 0$ for all $\xi \in \mathrm{CS}_N$. In particular, for any $D \geq 2N$, the linear map $f \in \mathcal{Y}_{2N} \mapsto f|_{\mathrm{CS}_N}$ is not injective. Therefore for $D \geq 2N$, the matrix A_N^D is not injective. □

Proposition 8.4 (Full column rank in the case $1 \leq N \leq 4$)**.** *Let A_N^D be the Vandermonde matrix* (8.6)*. If $1 \leq N \leq 4$, then*

$$A_N^D \text{ has full column rank } \Leftrightarrow D \leq 2N - 1;$$

In particular, the largest degree D_N such that $A_N^{D_N}$ has full column rank is $D_N = 2N - 1$.

Proof. Fix $N \leq 4$. Lemma 8.3 proves that if A_N^D has full column rank then $D \leq 2N-1$. For the converse, we fix $D \leq 2N-1$ and we prove that the linear map $f \in \mathcal{Y}_D \mapsto [f(\xi)]_{\xi \in \mathrm{CS}_N} \in \mathbb{R}^{6N^2+2}$ is injective. So we fix $f \in \mathcal{Y}_D$ such that $f(\xi) = 0$ for every $\xi \in \mathrm{CS}_N$ and we prove that $f = 0$. First, we introduce $2N$ meridian circles associated to the longitudes $\phi \equiv \frac{\pi}{4} \, (\frac{\pi}{2N})$,

$$\mathscr{C}(\psi) = \{x(\theta, \psi),\, \theta \in [-\tfrac{\pi}{2}, \tfrac{\pi}{2}]\} \cup \{x(\theta, \psi + \pi),\, \theta \in [-\tfrac{\pi}{2}, \tfrac{\pi}{2}]\}, \quad \psi \equiv \tfrac{\pi}{4} \, (\tfrac{\pi}{2N}),$$

and we prove that f is null on these great circles, *i.e.*

$$f|_{\mathscr{C}(\psi)} = 0, \quad \psi \equiv \tfrac{\pi}{4} \, (\tfrac{\pi}{2N}), \tag{8.15}$$

where $f|_{\mathscr{C}(\psi)}$ denotes the restriction of f to $\mathscr{C}(\psi)$. The assumption $N \leq 4$ implies that each great circle $\mathscr{C}(\psi)$ contains $4N$ points of CS_N. Since f vanishes on CS_N, these points give $4N$ zeros for $f|_{\mathscr{C}(\psi)}$. Since $f|_{\mathscr{C}(\psi)}$ represents a trigonometric polynomial with degree at most D, with $4N$ zeros, and $4N \geq 2D+1$, we obtain $f|_{\mathscr{C}(\psi)} = 0$.

Second, any great circle $\mathscr{C}$ not containing the pole $(0,0,1)$, contains $4N$ points in the set $\cup_{\psi \equiv \frac{\pi}{4} \, (\frac{\pi}{2N})} \mathscr{C}(\psi)$. Therefore (8.15) implies that f has $4N$ zeros on $\mathscr{C}$. It results that $f|_{\mathscr{C}}$ represents a trigonometric polynomial with degree at most D and $4N$ zeros; since $4N \geq 2D+1$, we obtain

$$f|_{\mathscr{C}} = 0, \quad (0,0,1) \notin \mathscr{C}. \tag{8.16}$$

Since the great circles in (8.15) and (8.16) cover the sphere, we have $f = 0$ on $\mathbb{S}^2$. □

8.3.2 The case $N \geq 5$

Lemma 8.3 and Proposition 8.4 give a full answer to the injectivity of A_{2N}^{2N-1} in the case $1 \leq N \leq 4$. Consider now the case $N \geq 5$. We have

Proposition 8.5 (Case $N \geq 5$)**.** *Suppose $N \geq 5$. We have*

$$D \leq N+2 \Rightarrow A_N^D \text{ has full column rank,}$$

Therefore, the largest degree D such that A_N^D has full column rank satisfies

$$N+2 \leq D \leq \bar{N}^{1/2} - 1 \, (\approx 2.45N - 1).$$

Proof. Fix $N \geq 5$ and $D \leq N+2$. Consider $f \in \mathcal{Y}_D$ such that $f(\xi) = 0$ for every $\xi \in \mathrm{CS}_N$. For $\psi = -\frac{\pi}{4}, \frac{\pi}{4}$, $\mathscr{C}(\psi)$ contains $4N$ points from CS_N, which implies

$$f|_{\mathscr{C}(\psi)=0}, \quad \psi = \pm\tfrac{\pi}{4}.$$

The two considered circles intersect at the poles $(0,0,\pm 1)$. Therefore any tangential derivative of f is zero at the poles, and each pole is a zero of order at least 2. Next, for any other angle $\psi \equiv \frac{\pi}{4} \, (\frac{\pi}{2N})$, $\mathscr{C}(\psi)$ contains at least $2N+2$ zeros of f on CS_N, and the poles as two additional zeros of order 2. Rolle's Theorem implies that the derivative of $f|_{\mathscr{C}(\psi)}$ (identified with a trigonometric polynomial) has $2N+6$ zeros. Since it is a trigonometric polynomial with degree at most D and $2N+6 \geq 2D+1$, we obtain (8.15). And we conclude as in the proof of Lemma 8.3. □

Remark 8.6. In the proof with $N \geq 5$, the bottleneck on the degree comes from the meridian circles that do not contain $4N$ points of CS_N. If there were $4N$ points per meridian circle, one would obtain the degree $2N-1$.

As in the proof of Lemma 8.3, the function f_N defined in (8.12) vanishes on the set M_N defined in (8.13). This implies that f_N vanishes at all nodes of the four equatorial panels (I)–(IV) of CS_N. Regarding panels (V) and (VI), f_N satisfies the estimate

$$|f_N(x(\theta,\phi))| \leq \gamma_N \cdot \cos^{2N}\theta, \quad \theta \in [-\tfrac{\pi}{2}, \tfrac{\pi}{2}],\, \phi \in \mathbb{R}. \tag{8.17}$$

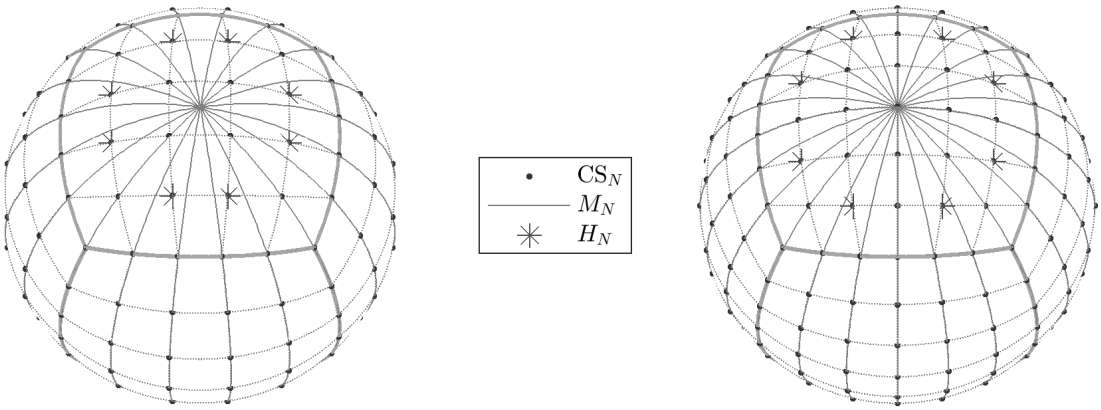


Figure 8.1: Equiangular Cubed Sphere and equiangular meridians. The Cubed Sphere CS_N (black dots) meshes $\mathbb{S}^2$ with equiangular arcs of great circles (dotted lines), including the radial projection of the edges of $[-1,1]^3$ (bold gray lines). The set M_N of equiangular meridians with longitude $\phi \equiv \frac{\pi}{4}\,(\frac{\pi}{2N})$ (gray lines) contains "many" points of CS_N; the remaining points of CS_N belong to the set H_N (indicated with star symbols) defined in (8.19). The size of H_N is given in (8.21), and is estimated by $|H_N| \sim \frac{1}{3}\bar{N}$. Left panel: N is odd ($N = 5$), right panel: N is even ($N = 6$).

The constant γ_N is

$$\gamma_N = \sqrt{\tfrac{4N+1}{2\pi}} \cdot \tfrac{\sqrt{(4N)!}}{2^{2N}(2N)!} \sim \tfrac{1}{\pi^{1/2}} \left(\tfrac{2N}{\pi}\right)^{1/4} (\approx 0.504\, N^{1/4}). \tag{8.18}$$

(The Stirling formula has been used). The behaviour of f_N on the north panel (V) and south panel (VI) is obtained by inspecting the nodes located outside the set M_N in (8.13), where the estimate (8.17) holds. Let $H_N \subset \mathrm{CS}_N$ be the set of nodes defined by

$$H_N := \big\{ \tfrac{1}{r}(u, v, \pm 1) : r = (1 + u^2 + v^2)^{1/2},\ u = \tan \tfrac{i\pi}{2N},\ v = \tan \tfrac{j\pi}{2N}, \\ -\tfrac{N}{2} < i, j < \tfrac{N}{2},\ |i| \neq |j| \text{ and } i \neq 0 \text{ and } j \neq 0 \big\}. \tag{8.19}$$

It turns out (see Figure 8.1) that

$$\mathrm{CS}_N \setminus M_N \subset H_N. \tag{8.20}$$

Furthermore, the number of nodes in the set H_N is given by

$$|H_N| = \begin{cases} 2(N-1)(N-3), & \text{if } N \text{ is odd,} \\ 2(N-2)(N-4), & \text{if } N \text{ is even.} \end{cases} \tag{8.21}$$

In the next theorem it is proved that $f_N|_{\mathrm{CS}_N}$ "almost vanishes" at all the Cubed Sphere nodes $\xi \in \mathrm{CS}_N$. This will show that when taking $D = 2N$, the Vandermonde matrix A_N^{2N} cannot have full column rank (injective) while keeping a bounded condition number.

Theorem 8.7 (Asymptotics for the condition number of A_N^D)**.** *Fix $N \geq 1$ and $D \geq 2N$.*

(i) The smallest singular value of the matrix A_N^D, denoted by $\sigma_{\min}(A_N^D)$, satisfies

$$\sigma_{\min}(A_N^D)^2 \leq \gamma_N^2 \cdot |H_N| \cdot \left(\frac{2}{3}\right)^{2N} \underset{N\to+\infty}{\sim} N \left(\frac{2N}{\pi}\right)^{3/2} \left(\frac{2}{3}\right)^{2N} \underset{N\to+\infty}{\to} 0,$$

where γ_N is given by (8.18)*, and $|H_N|$ is the estimation* (8.21) *of the size of* $\mathrm{CS}_N \setminus M_N$*. In particular*

$$\lim_{N\to+\infty} \sigma_{\min}\Big((A_N^{2N})^\intercal A_N^{2N}\Big) = 0.$$

(ii) In the case where A_N^D is injective, the condition number of A_N^D, denoted by $\mathrm{cond}(A_N^D)$*, satisfies*

$$\mathrm{cond}(A_N^D)^2 \geq \frac{\bar{N}}{|H_N|}\cdot\frac{1}{4\pi\gamma_N^2}\cdot\left(\frac{3}{2}\right)^{2N} \underset{N\to+\infty}{\sim} \frac{1}{4}\Big(\frac{\pi}{2N}\Big)^{1/2}\Big(\frac{3}{2}\Big)^{2N+1} \underset{N\to+\infty}{\to} +\infty,$$

where $\bar{N} = 6N^2+2$. In particular,

$$\lim_{N\to+\infty} \mathrm{cond}\left((A_N^{2N})^\intercal A_N^{2N}\right) = +\infty.$$

Proof. (i) Let $D \geq 2N$ be fixed. Consider first the case $\bar{N} < (D+1)^2$, the matrix A_N^D cannot have full column rank. In this case $\sigma_{\min}(A_N^D) = 0$ and the result is obvious. Next consider the case $\bar{N} \geq (D+1)^2$. Then $\sigma_{\min}(A_N^D)^2$ is the smallest eigenvalue of the symmetric matrix $A_N^{D\intercal}A_N^D$. It is expressed as the minimum Rayleigh quotient

$$\sigma_{\min}(A_N^D)^2 = \inf_{\substack{\hat{f}\in\mathbb{R}^{(D+1)^2}\\ \|\hat{f}\|=1}} \Big(\hat{f}^\intercal A_N^{D\intercal}A_N^D\hat{f}\Big).$$

With the Fourier-like expansion (6.4), we obtain $\hat{f}^\intercal A_N^{D\intercal}A_N^D\hat{f} = \|A_N^D\hat{f}\|^2 = \sum_{\xi\in\mathrm{CS}_N} f(\xi)^2$, so that

$$\sigma_{\min}(A_N^D)^2 = \inf_{\substack{f\in\mathcal{Y}_D\\ \|f\|_{L^2(\mathbb{S}^2)}=1}} \left(\sum_{\xi\in\mathrm{CS}_N} f(\xi)^2\right).$$

Let f_N be the function defined in (8.12); then f_N is a rotation of the unitary function $Y_{2N}^{-2N} \in \mathcal{Y}_D$, so $f_N \in \mathcal{Y}_D$ with $\|f\| = 1$, which proves that

$$\sigma_{\min}(A_N^D)^2 \leq \sum_{\xi\in\mathrm{CS}_N} f(\xi)^2.$$

Using $\mathrm{CS}_N = (\mathrm{CS}_N \cap M_N) \cup (\mathrm{CS}_N \cap H_N)$ and that $f_N \equiv 0$ on M_N, we deduce

$$\sigma_{\min}(A_N^D)^2 \leq \sum_{\xi\in\mathrm{CS}_N\cap H_N} f(\xi)^2.$$

If $H_N = \emptyset$, we have $\sigma_{\min}(A_N^D)^2 = 0$ and (i) is proved. Otherwise,

$$\sigma_{\min}(A_N^D)^2 \leq |H_N| \max_{\xi\in H_N} f(\xi)^2,$$

where $|H_N|$ is given by (8.21). Using (8.17), we have

$$\max_{\xi\in H_N} f(\xi)^2 \leq \gamma_N^2 c^{2N}, \quad \text{with} \quad c = \max_{\xi\in H_N} \cos^2\theta(\xi).$$

For any $\xi = \frac{1}{(1+u^2+v^2)^{1/2}}(u,v,\pm1) \in H_N$, with $|u|,|v| < 1$, the latitude angle $\theta(\xi)$ is such that $\cos^2\theta(\xi) = 1 - \sin^2\theta(\xi) = 1 - \frac{1}{1+u^2+v^2} < \frac{2}{3}$, which proves that $c < \frac{2}{3}$.

(ii) If A_N^D is injective, the condition number is the ratio

$$\mathrm{cond}(A_N^D) = \frac{\sigma_{\max}(A_N^D)}{\sigma_{\min}(A_N^D)},$$

where $\sigma_{\min}(A_N^D)$ has been bounded from below in (i), and $\sigma_{\max}(A_N^D)$ denotes the largest singular value of A_N^D. The square $\sigma_{\max}(A_N^D)^2$ is the largest eigenvalue of $A_N^{D\intercal}A_N^D$ and it is the maximum Rayleigh ratio

$$\sigma_{\max}(A_N^D)^2 = \sup_{\substack{\hat{f}\in\mathbb{R}^{(D+1)^2}\\ \|\hat{f}\|=1}} \left(\hat{f}^\intercal A_N^{D\intercal} A_N^D \hat{f}\right) = \sup_{\substack{f\in\mathcal{Y}_D\\ \|f\|_{L^2(\mathbb{S}^2)}=1}} \sum_{\xi\in \mathrm{CS}_N} f(\xi)^2.$$

With the particular choice $f(x) = Y_0^0(x) = \frac{1}{\sqrt{4\pi}}$ we obtain the lower bound $\sigma_{\max}(A_N^D)^2 \geq \frac{\bar{N}}{4\pi}$. □

8.4 Structure of the normal matrix

In this section, we consider the problem (WLS) and the matching quadrature rule (8.4). Recall that the matrix attached to (WLS) is the matrix $A_N^{D\intercal}\Omega_N A_N^D$ in (8.9). We show in Theorem 8.8 below how close to an orthonormal system the set of functions (Y_n^m) is, for D a fixed integer. Next, Section 8.4.2 considers the particular case where the weight function $\omega(\xi)$ has the cubic symmetry. In this case, a suitable ordering of the indices n and m leads to a particular block diagonal structure of the matrix $A_N^{D\intercal}\Omega_N A_N^D$, which is fully specified.

8.4.1 Least squares and quadrature rule accuracy

Suppose the integer D fixed and consider the least squares problem (WLS) in Section 8.2. Proving the well posedness of (WLS) amounts to establish bounds for the condition number of the matrix $A_N^{D\intercal}\Omega_N A_N^D$. We have

$$A_N^{D\intercal}\Omega_N A_N^D = \left[\sum_{\xi\in\mathrm{CS}_N} \omega(\xi) Y_n^m(\xi) Y_{n'}^{m'}(\xi)\right]_{\substack{|m|\leq n\leq D\\ |m'|\leq n'\leq D}} \in \mathbb{R}^{(D+1)^2\times(D+1)^2}. \tag{8.22}$$

This matrix contains inner products involving the gridfunctions $(Y_n^m)|_{\mathrm{CS}_N}$, for the discrete weighted inner product defined by

$$(y_1, y_2)_\omega := \sum_{\xi\in\mathrm{CS}_N} \omega_N(\xi) y_1(\xi) y_2(\xi), \quad y_1, y_2 : \mathrm{CS}_N \to \mathbb{R}. \tag{8.23}$$

The functions Y_n^m are orthonormal for the inner product $\langle\cdot,\cdot\rangle_{L^2(\mathbb{S}^2)}$. However, for the discrete product $(.,.)_\omega$, we only have $(Y_n^m, Y_{n'}^{m'})_\omega \approx \delta_{nn'}\delta_{mm'}$. Let E_N^D be the symmetric matrix defined by

$$E_N^D = \left[e_N(Y_n^m Y_{n'}^{m'})\right]_{\substack{|m|\leq n\leq D\\ |m'|\leq n'\leq D}} \in \mathbb{R}^{(D+1)^2\times(D+1)^2}. \tag{8.24}$$

The entries of the matrix E_N^D are the quadrature errors of the products $Y_n^m Y_{n'}^{m'}$.

Theorem 8.8. *Fix $N \geq 1$ and $D \geq 0$. Let A_N^D be the Vandermonde matrix with degree D on CS_N, defined in (8.6). Let $\omega : \mathrm{CS}_N \to (0,\infty)$ be the weight of a spherical quadrature rule on CS_N, with error e_N; let Ω_N be the associated diagonal matrix, defined in (8.7). Then*

$$A_N^{D\intercal}\Omega_N A_N^D = \mathrm{I}_{(D+1)^2} - E_N^D. \tag{8.25}$$

In particular, assume that $(\omega_N)_{N\geq 1}$ is a sequence of weight functions defining a convergent quadrature rule on $\mathcal{Y}_{2D}$, i.e. *$\forall f \in \mathcal{Y}_{2D}$, $e_N(f) \xrightarrow[N\to\infty]{} 0$, then*

$$A_N^{D\intercal}\Omega_N A_N^D \xrightarrow[N\to\infty]{} \mathrm{I}_{(D+1)^2}.$$

Moreover, if the rule converges with order $p > 0$, i.e. $\forall f \in \mathcal{Y}_{2D}, \exists C_f \geq 0, \forall N \geq 1, |e_N(f)| \leq C_f N^{-p}$, *then*

$$A_N^{D\intercal}\Omega_N A_N^D = \mathrm{I}_{(D+1)^2} + \mathcal{O}\left(\frac{1}{N^p}\right).$$

The relation (8.25) expresses the fact that the matrix $A_N^{D\intercal}\Omega_N A_N^D$ is close to the identity, assuming that the error matrix entries are small. This is in particular the case when ω defines an accurate quadrature rule on the space $\mathcal{Y}_{2D}$. This will require that D is not too large compared to N. On the contrary, for large values of D, the entries in E_N^D are not a priori small.

Proof. In the matrix $A_N^{D\intercal}\Omega_N A_N^D$, the entry with row index (n,m) and column index (n',m') contains the discrete inner product $\left(Y_n^m|_{\mathrm{CS}_N}, Y_{n'}^{m'}|_{\mathrm{CS}_N}\right)_{\omega_N}$, as described in (8.22) and (8.23). Using the quadrature rule (8.4) with $g = Y_n^m Y_{n'}^{m'}$ shows that this element is expressed as

$$\left(Y_n^m|_{\mathrm{CS}_N}, Y_{n'}^{m'}|_{\mathrm{CS}_N}\right)_{\omega_N} = \int_{\mathbb{S}^2} Y_n^m(x) Y_{n'}^{m'}(x)\,\mathrm{d}\sigma - e_N(Y_n^m Y_{n'}^{m'}).$$

Since the family $(Y_n^m)_{0\leq|m|\leq n\leq D}$ is orthonormal in $L^2(\mathbb{S}^2)$, we have

$$\int_{\mathbb{S}^2} Y_n^m(x) Y_{n'}^{m'}(x)\,\mathrm{d}\sigma = \left\langle Y_n^m, Y_{n'}^{m'}\right\rangle_{L^2(\mathbb{S}^2)} = \begin{cases} 1, & \text{if } (n,m) = (n',m'), \\ 0, & \text{otherwise.} \end{cases}$$

This proves (8.25). The symmetry of E_N^D is obvious.

Finally for a convergent rule, for all $|m| \leq n \leq D$ and $|m'| \leq n' \leq D$, the entry of E_N^D with indices (n,m) and (n',m') is related to $f = Y_n^m Y_{n'}^{m'} \in \mathcal{Y}_{2D}$, so that by hypothesis $e_N(Y_n^m Y_{n'}^{m'}) \to 0$. For a convergence of order $p > 0$, there is furthermore a constant $C_{n,m}^{n',m'}$ such that $|e_N(Y_n^m Y_{n'}^{m'})| \leq C_{n,m}^{n',m'} N^{-p}$. □

8.4.2 Block structure of $(A_N^D)^\intercal \Omega_N A_N^D$ for a symmetric weight function

The weight function $\xi \in \mathrm{CS}_N \mapsto \omega(\xi)$ plays the role of a parameter in the problem (WLS). Here we consider the particular case where $\omega(\xi)$ has the cubic symmetry. This property has been considered in [156], [9].

Theorem 8.9. *Assume that $\omega : \mathrm{CS}_N \to (0,\infty)$ is invariant under the symmetry group $\mathcal{G}$ of the cube $\{-1,1\}^3$. Consider a nonzero entry $e_N(Y_n^m Y_{n'}^{m'})$ in the matrix E_N^D defined in (8.24), with row index (n,m), and column index (n',m'). Then the following conditions hold*

(i) $n \equiv n'\,(2)$ *(same parity for the degrees);*

(ii) $m, m' \geq 0$, *or* $m, m' < 0$ *(same sign for the orders);*

(iii) $m \equiv m'\,(2)$ *(same parity for the orders);*

(iv) if $m, m' \equiv 0\,(2)$, *then* $m \equiv m'\,(4)$.

Proof. The principle of the proof is close to the one of Corollary 7.6 ([9, Corollary 10]). By Theorem 5.3, the group of the Cubed Sphere coincides with the group $\mathcal{G}$ of the cube, given by (5.2). Therefore, the quadrature error defines a linear form

$$e_N : \mathcal{Y}_{2D} \to \mathbb{R}, \quad e_N(g) = \int_{\mathbb{S}^2} g(x)\,\mathrm{d}\sigma - \sum_{\xi\in\mathrm{CS}_N} \omega(\xi) g(\xi),$$

which is invariant under $\mathcal{G}$, *i.e.* $\forall Q \in \mathcal{G}$, $e_N(g(Q^\intercal\cdot)) = e_N(g)$. In the sequel, for all (n,m) and (n',m') violating at least one of the conditions *(i)-(iv)* in Theorem 8.9, we consider $g = Y_n^m Y_{n'}^{m'} \in \mathcal{Y}_{2D}$, and

we exhibit a matrix $Q \in \mathcal{G}$ satisfying $g(Q^\intercal x) = -g(x)$. This is a sufficient condition to ensure that $e_N(g) = 0$, due to $e_N(g) = e_N(g(Q^\intercal \cdot)) = e_N(-g) = -e_N(g)$. The proof is a calculation in spherical coordinates, based on the expression

$$g(x(\theta,\phi)) = (q_n^m q_{n'}^{m'})(\sin\theta)\cdot\cos^{|m|}\theta\cos^{|m'|}\theta \\ \cdot(\sin(m\phi)\mathbf{1}_{m<0}+\cos(m\phi)\mathbf{1}_{m\geq 0})(\sin(m'\phi)\mathbf{1}_{m'<0}+\cos(m'\phi)\mathbf{1}_{m'\geq 0}).$$

Case 1: (ii) is violated. Assume $m < 0$ and $m' \geq 0$ (without loss of generality), then

$$g(Q^\intercal x(\theta,\phi)) = g(x(\theta,-\phi)) = -g(x(\theta,\phi)), \quad \text{for} \quad Q := \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Case 2: (iii) is violated. Assume that $m \equiv 1\,(2)$ and $m' \equiv 0\,(2)$ (without loss of generality). Then $m(\phi+\pi) \equiv m\phi + \pi\,(2\pi)$, $m'(\phi+\pi) \equiv m'\phi\,(2\pi)$, and

$$g(Q^\intercal x(\theta,\phi)) = g(x(\theta,\phi+\pi)) = -g(x(\theta,\phi)), \quad \text{for} \quad Q := \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Case 3: (iii) is satisfied but (i) is violated. Assume that $n + |m| \equiv 1\,(2)$ and $n' + |m'| \equiv 0\,(2)$ (without loss of generality). Then $\theta \mapsto (q_n^m q_{n'}^{m'})(\sin\theta)$ is odd, hence

$$g(Q^\intercal x(\theta,\phi)) = g(x(-\theta,\phi)) = -g(x(\theta,\phi)), \quad \text{for} \quad Q := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

Case 4: (iv) is violated. Assume that $m \equiv 2\,(4)$ and $m' \equiv 0\,(4)$ (without loss of generality). Then $m(\phi+\frac{\pi}{2}) \equiv m\phi + \pi\,(2\pi)$, $m'(\phi+\frac{\pi}{2}) \equiv m'\phi\,(2\pi)$, and

$$g(Q^\intercal x(\theta,\phi)) = g(x(\theta,\phi+\tfrac{\pi}{2})) = -g(x(\theta,\phi)), \quad \text{for} \quad Q := \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \qquad \square$$

Roughly speaking, if the weight function ω is symmetric, then at most a percentage of

$$100 \times \frac{3}{32} (= 9.375\%) \tag{8.26}$$

of all the entries in E_N^D are nonzero. Indeed, Case *(i)* divides by 2 the number of possible nonzero entries. Then Case *(ii)* further divides by 2 this number. And finally Cases *(iii-iv)* multiply this number by $\frac{3}{8}$. At this point, two facts suggest the approximation

$$A_N^{D\intercal}\Omega_N A_N^D \approx \mathrm{I}_{(D+1)^2}:$$

- an approximate ratio of $\frac{29}{32}$ of all entries are zero if the weight function ω is assumed symmetric (Theorem 8.9);
- the remaining entries (approximate ratio of $\frac{3}{32}$) are small assuming that ω defines an accurate spherical quadrature rule in $\mathcal{Y}_{2D}$, (see Theorem 8.8).

In particular, the condition number of the matrix $A_N^{D\intercal}\Omega_N A_N^D$ is expected to be close to 1, so that (WLS) is expected to be well-posed. This point is further investigated numerically in SubSection 8.5.3.

Next, we go one step further in the analysis taking benefit from the orthogonality relations in Theorem 8.9. Indeed, Theorem 8.9 suggests to sort the indices (n,m) using the following criteria, ordered by decreasing priority:

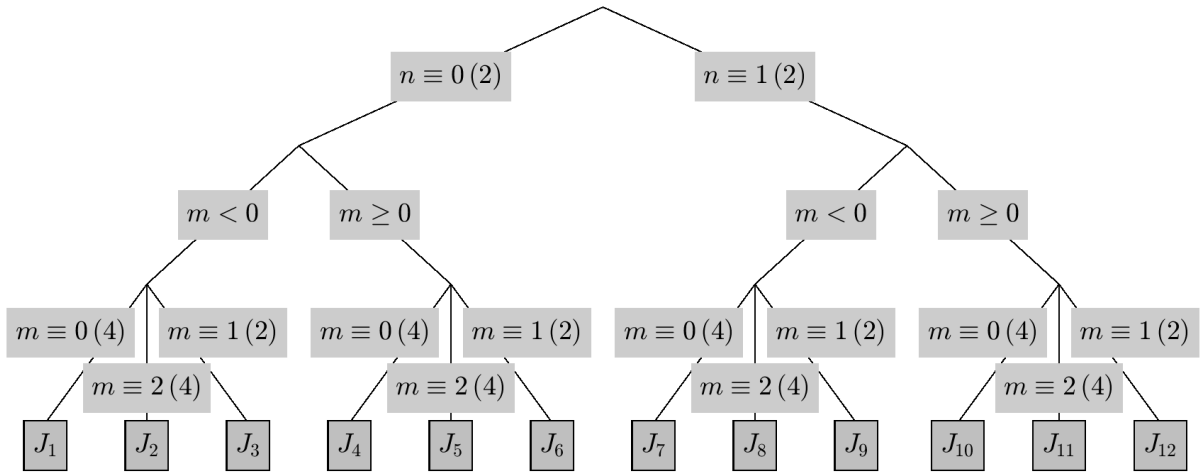


Figure 8.2: Classification tree for partioning the set of indices $\{(n,m) : |m| \leq n \leq D\}$ as a disjoint union $J_1 \cup \ldots \cup J_{12}$; for instance, $J_4 = \{|m| \leq n \leq D,\ n \equiv 0\,(2),\ m \geq 0,\ m \equiv 0\,(4)\}$.

- case $n \equiv 0\,(2)$ and case $n \equiv 1\,(2)$;
- case $m < 0$ and case $m \geq 0$;
- case $m \equiv 0\,(4)$, then case $m \equiv 2\,(4)$, and finally case $m \equiv 1\,(2)$.

This particular ordering expresses the set of indices as a disjoint union of the twelve sets J_k, $1 \leq k \leq 12$. Figure 8.2 displays the resulting classification tree. It is an expression of the orthogonality relations in Theorem 8.9.

Corollary 8.10. *Fix $N \geq 1$, $D \geq 0$. Fix a weight function $\omega : \mathrm{CS}_N \to (0,\infty)$ invariant under the group $\mathcal{G}$ of $\{-1,1\}^3$ in (5.2). Let J_k, $1 \leq k \leq 12$, denotes a partioning of the set of indices $|m| \leq n \leq D$, displayed in Figure 8.2.*

(i) Assume that the indices $(n,m) \in J_k$, $1 \leq k \leq 12$ in the Vandermonde matrix A_N^D are sorted along increasing k (for the rows and for the columns). Then ${A_N^D}^{\intercal}\Omega_N A_N^D$ is block diagonal, as shown in Figure 8.3.

(ii) The following orthogonal decomposition holds for the discrete inner product (8.23),

$$\big\{f|_{\mathrm{CS}_N},\ f \in \mathcal{Y}_D\big\} = \bigoplus_{k=1}^{12} \mathrm{Span}\{Y_n^m|_{\mathrm{CS}_N},\ (n,m) \in J_k\}.$$

Assuming a symmetric weight function ω, Corollary 8.10 reveals that the matrix ${A_N^D}^{\intercal}\Omega_N A_N^D$, associated to (WLS), is block diagonal for a particular ordering of the indices. This has the following consequence to solve the system (8.9). Instead of solving a linear system with $(D+1)^2$ unknowns, the system is solved by blocks. It consists in solving 8 square linear systems with approximately $\frac{1}{16}(D+1)^2$ unknowns, and 4 square linear systems with approximately $\frac{1}{8}(D+1)^2$ unknowns. These resolutions can be obviously performed in parallel.

Remark 8.11. In Corollary 7.6 , the "15/16" property is coined as meaning exactness of a symmetric quadrature rule for a certain proportion of 15/16 of all spherical harmonics. This property can be deduced from Corollary 8.10. Indeed, consider a symmetric weight function ω, a row index $(n,m) \notin J_4$, and the column index $(n',m') = (0,0) \in J_4$; then $Y_{n'}^{m'} \equiv (4\pi)^{-1/2}$, and we deduce from Corollary 8.10.(i) that

$$\mathcal{Q}(Y_n^m) = \sum_{\xi \in \mathrm{CS}_N} \omega(\xi) Y_n^m(\xi) = 0 = \int_{\mathbb{S}^2} Y_n^m(x)\,\mathrm{d}\sigma.$$

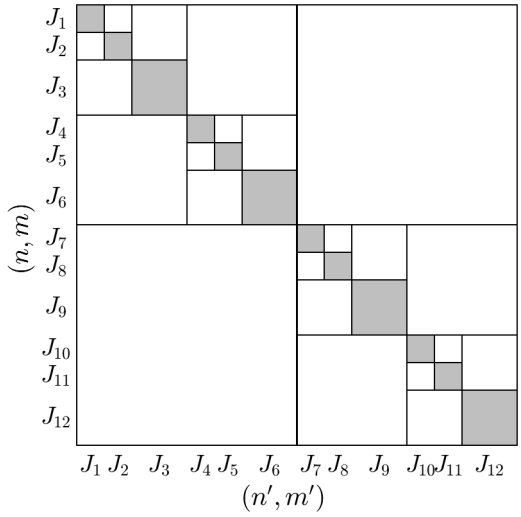


Figure 8.3: Block diagonal structure of the matrix $A_N^{D\intercal}\Omega_N A_N^D$, assuming that ω is invariant under $\mathcal{G}$; the sets of indices J_k are defined in Figure 8.2. The white cells contains only null coefficients; they represent an approximate ratio of $\frac{29}{32}$ of the entries.

Since J_4 contains about 1/16 of the indices, we see that the quadrature rule $\mathcal{Q}$ associated to ω exactly integrates an approximate proportion of 15/16 of all the Y_n^m.

8.5 Numerical results

8.5.1 Condition number of the Vandermonde matrix

In this section, we assess numerically that the problem (LS) is well-posed for the degree $D = 2N-1$, but not for $D = 2N$. We proceed as follows. For all $1 \leq N \leq 32$, with $D = 2N-1$ and $D = 2N$, we first compute a singular value decomposition of the Vandermonde matrix A_N^D in (8.6). Second, we extract the minimal singular value $\sigma_{\min}(A_N^D)$ and the maximal one $\sigma_{\max}(A_N^D)$. Then the condition number $\mathrm{cond}(A_N^D) = \sigma_{\max}(A_N^D)/\sigma_{\min}(A_N^D)$ is evaluated. The computation has been performed in double precision in `Matlab`, using the `svd` function. The results in Figure 8.4 are as follows

1. For $D = 2N-1$ (left panel in Figure 8.4), we observe that the minimal singular value is "far" from 0, and that $\mathrm{cond}(A_N^D) \approx 1.19$ is close to 1. This is a numerical indication that the matrix A_N^{2N-1} is injective, that the problem (LS) is well-posed for $D = 2N-1$, which implies that the critical degree D_N in (P) is such that $D_N \geq 2N-1$.

2. For $D = 2N$ (right panel in Figure 8.4), $\sigma_{\min}(A_N^D)$ is observed to be close to 0 for $N \in \{1, 2, 3, 4, 5, 7, 9\}$ (the machine epsilon is about $2.2 \cdot 10^{-16}$); for $N \geq 10$, it is positive and decays to 0 when N increases. Hence, for $N \in \{1, 2, 3, 4, 5, 7, 9\}$, A_N^{2N} is not injective. This suggests $D_N \leq 2N-1$. This is consistent with Proposition 8.4 which proves the result for $N \leq 4$. This numerical observation, combined with the discussion above, supports the fact that

$$D_N = 2N-1, \quad N \in \{1, 2, 3, 4, 5, 7, 9\}.$$

   For the other values of N, it is numerically apparent that A_N^{2N} is injective. Nevertheless, for these values of N, $\mathrm{cond}(A_N^{2N}) > 10^4$, and blows-up when N increases. This implies that $\mathrm{cond}(A_N^{2N\intercal} A_N^{2N}) > 10^8$ and blows-up as well. Therefore, for $D = 2N$, these numerical results are in agreement with the theoretical result in Theorem 8.7 and indicates that the ill-posedness of (LS) is true in all cases. In addition, the ill-posedness level increases with N.

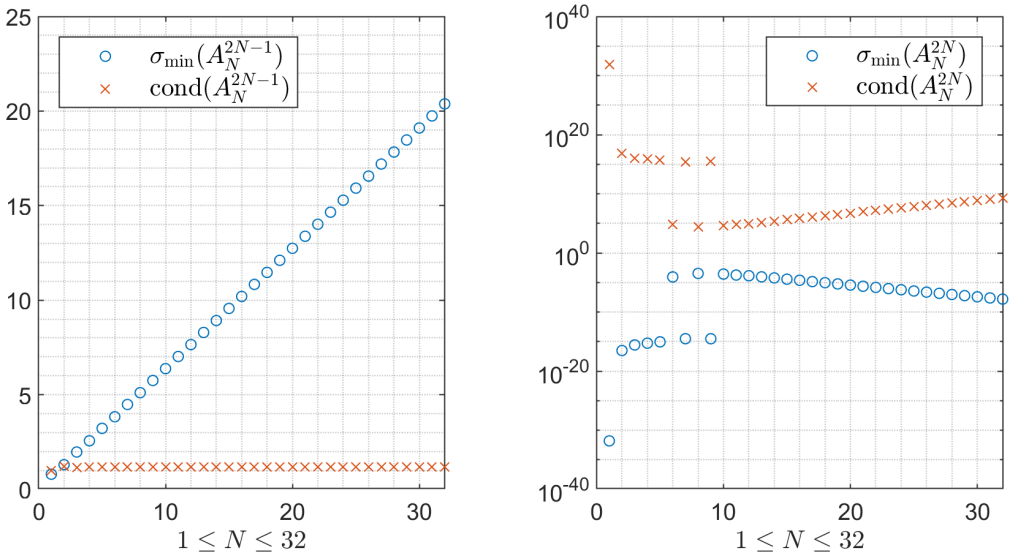


Figure 8.4: Smallest singular value ($\sigma_{\min}$) and condition number (cond) of the Vandermonde matrices A_N^D, $1 \le N \le 32$, with $D = 2N-1$ (left panel), and $D = 2N$ (right panel, in log-scale). Left panel: A_N^{2N-1} is injective ($\sigma_{\min} \gg 0$) and well-conditioned ($\mathrm{cond} \approx 1.19$). Right panel: A_N^{2N} is observed to be not numerically injective if N is small ($\sigma_{\min} \approx 0$), and is ill-conditioned otherwise ($\mathrm{cond} > 10^4$).

i	$f_i(x,y,z)$	Comment
1	$\exp(x)$	Very smooth
2	$\frac{3}{4}\exp[-\frac{(9x-2)^2}{4} - \frac{(9y-2)^2}{4} - \frac{(9z-2)^2}{4}]$ $+\frac{3}{4}\exp[-\frac{(9x+1)^2}{49} - \frac{9y+1}{10} - \frac{9z+1}{10}]$ $+\frac{1}{2}\exp[-\frac{(9x-7)^2}{4} - \frac{(9y-3)^2}{4} - \frac{(9z-5)^2}{4}]$ $-\frac{1}{5}\exp[-(9x-4)^2 - (9y-7)^2 - (9z-5)^2]$	Smooth
3	$\frac{1}{10}\frac{\exp(x+2y+3z)}{(x^2+y^2+(z+1)^2)^{1/2}}\mathbf{1}(z > -1)$	Infinite spike at the south pole ($z = -1$)
4	$\cos(3\arccos z)\mathbf{1}(3\arccos z \le \frac{\pi}{2})$	Continuous, not differentiable ($z = \frac{\sqrt{3}}{2}$)
5	$\mathbf{1}(z \ge \frac{1}{2})$	Discontinuous spherical cap ($z = \frac{1}{2}$)

Table 8.1: A series of test functions representative of various regularity properties.

The numerical study above suggests that for any $N \ge 1$, the value of D_N in Property (P) is $D_N = 2N-1$. This in particular means that any $f \in \mathcal{Y}_{2N-1}$ is correctly sampled on the Cubed Sphere CS_N with angular step $\frac{\pi}{2N}$, since (LS) can reconstruct f from $f|_{\mathrm{CS}_N}$ in a stable way. If $f \in \mathcal{Y}_D$ with $D \ge 2N$, this property is not guaranteed.

8.5.2 Accuracy of the least squares approximation

The results in Section 8.3 assess the fact that $\mathcal{Y}_{2N-1}$ is the largest spherical harmonics subspace leading to well-posedness and well-conditioning of the problem (LS). Here we further assess this property by evaluating the accuracy of least-squares approximations of a series of test functions.

First, we report in Table 8.1 five functions extracted from Table 7.3 (and plotted in Figure 7.4). This series of functions is representative of various regularity properties. For each $1 \le N \le 32$ and for each test function f_i, $1 \le i \le 5$, we compute the least-squares approximation $\tilde{f}_i \in \mathcal{Y}_{2N-1}$ of f_i from the grid function $f_i|_{\mathrm{CS}_N}$: $\tilde{f}_i$ is evaluated as the unique solution to (LS), for $D = 2N-1$ and $y = f_i|_{\mathrm{CS}_N}$. The accuracy is measured by the relative discrete error, on a fixed fine grid CS_M. We

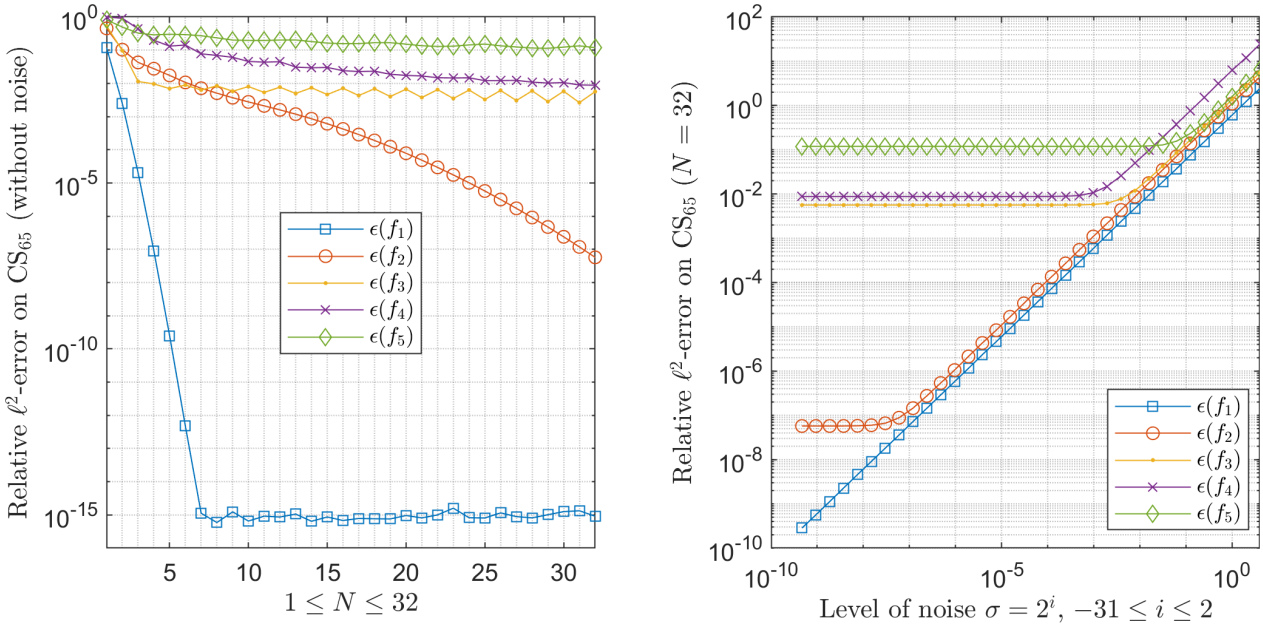


Figure 8.5: Least-squares approximation (LS) of the test functions f_i in Table 8.1. Left panel: for any $1 \le N \le 32$, the approximation $\tilde{f}_i \in \mathcal{Y}_{2N-1}$ is computed from $f_i|_{\mathrm{CS}_N}$, and the relative ℓ^2-error $\epsilon(f_i)$ defined in (8.27) is plotted. Right panel: for any level of noise $\sigma = 2^j, -31 \le j \le 2$, the approximation $\tilde{f}_i \in \mathcal{Y}_{63}$ is computed from a noisy dataset $f_i|_{\mathrm{CS}_N} + \sigma\mathcal{N}(0,1)$ with $N = 32$, and $\epsilon(f_i)$ is plotted.

have chosen $M = 65$. The relative error is defined by

$$\epsilon(f_i) := \left(\frac{\sum_{\xi \in \mathrm{CS}_M} |f_i(\xi) - \tilde{f}_i(\xi)|^2}{\sum_{\xi \in \mathrm{CS}_M} |f_i(\xi)|^2} \right)^{1/2}, \quad M = 65. \tag{8.27}$$

The errors $\epsilon(f_i)$ are displayed in Figure 8.5 (left panel). For the smooth functions f_1 and f_2, the error rapidly converges to 0 when N increases; this is especially true for f_1. For the continuous but not differentiable function f_4, the convergence is slow. For the spike function f_3 and the discontinuous function f_5, the convergence cannot be claimed from the plot. These observations are not surprising: it is expected that the convergence rate depends on the decay of the Fourier coefficients, which is related to the smoothness.

Second, fix the grid resolution to $N = 32$. For each test function f_i, $1 \le i \le 5$, for any $\sigma = 2^j$, $-31 \le j \le 2$, we corrupt the grid function $f_i|_{\mathrm{CS}_N}$ with a gaussian noise with zero mean, and standard deviation σ. We compute an approximation $\tilde{f}_i \in \mathcal{Y}_{63}$ of f_i as the unique solution to (LS), for $D = 2N-1$ and $y(\xi) = f_i(\xi) + \sigma u(\xi)$, $\xi \in \mathrm{CS}_{32}$, where $[u(\xi)]$ contains independent realizations of the normal law $\mathcal{N}(0,1)$. Here again, we evaluate the accuracy of this approximation by the relative error (8.27); this error depends on σ (and on the experiment), and we denote it by $\epsilon(f_i)(\sigma)$. These errors are displayed in Figure 8.5 (right panel). One observes that

$$\epsilon(f_i)(\sigma) \approx \epsilon(f_i)(0) + \sigma,$$

where $\epsilon(f_i)(0)$ is the error without noise for $N = 32$ (displayed on the left panel). In other words, a level of noise σ in the dataset increases the error by σ. This reveals that approximating a function by least-squares on CS_N in the space $\mathcal{Y}_{2N-1}$ is very stable.

Third, we show numerically that differentiating the least-squares approximation (LS) in $\mathcal{Y}_{2N-1}$ ($D = 2N-1$) permits to approximate derivatives. Assume that f is a differentiable function on $\mathbb{S}^2$, known by the grid function $f|_{\mathrm{CS}_N}$. The least squares approximation (LS) of f with $D = 2N-1$ is

$$\tilde{f} = \sum_{|m| \le n \le 2N-1} \tilde{f}_n^m Y_n^m \in \mathcal{Y}_{2N-1}, \tag{8.28}$$

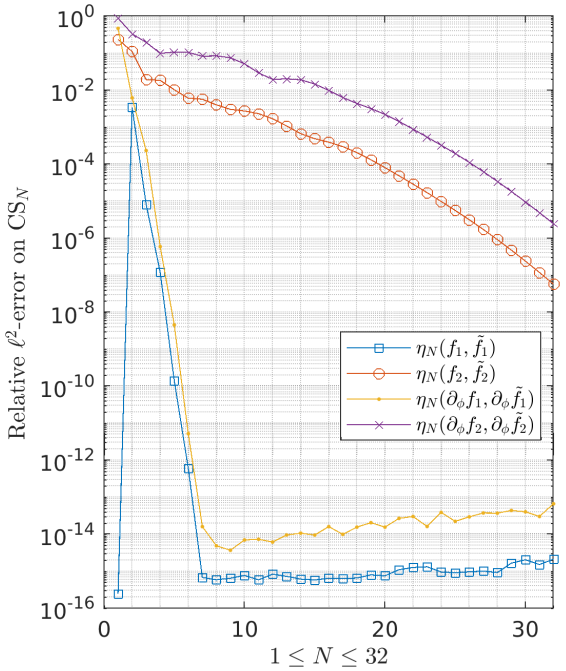


Figure 8.6: Spectral differentiation on CS_N with respect to the longitude angle ϕ. For $f = f_1, f_2$ from Table 8.1, for any $1 \le N \le 32$, the approximate derivative $\partial_\phi \tilde{f}$ is computed from the least-squares approximation $\tilde{f} \in \mathcal{Y}_{2N-1}$. Relative ℓ^2-errors defined in (8.30) are plotted.

and $y = f|_{\mathrm{CS}_N}$. Consider for instance the derivative with respect to the longitude ϕ,

$$\partial_\phi \tilde{f}(x(\theta,\phi)) = \sum_{|m| \le n \le 2N-1} m \cdot \tilde{f}_n^{-m} Y_n^m(x(\theta,\phi)) \in \mathcal{Y}_{2N-1}. \tag{8.29}$$

We test this principle on the smooth functions defined in Table 8.1: $f = f_1, f_2$. For each value of $1 \le N \le 32$, we approximate $\partial_\phi f$ by $\partial_\phi \tilde{f}$ satisfying (8.29), and we calculate the relative ℓ^2-errors on the grid CS_N:

$$\eta_N(f,\tilde{f}) = \left(\frac{\sum_{\xi \in \mathrm{CS}_N} |f(\xi) - \tilde{f}(\xi)|^2}{\sum_{\xi \in \mathrm{CS}_N} |f(\xi)|^2} \right)^{1/2}, \; \eta_N(\partial_\phi f, \partial_\phi \tilde{f}) = \left(\frac{\sum_{\xi \in \mathrm{CS}_N} |\partial_\phi f(\xi) - \partial_\phi \tilde{f}(\xi)|^2}{\sum_{\xi \in \mathrm{CS}_N} |\partial_\phi f(\xi)|^2} \right)^{1/2}; \tag{8.30}$$

here, the exact derivative is given by

$$\partial_\phi f(x(\theta,\phi)) = -x_2(\theta,\phi)\partial_{x_1} f(x(\theta,\phi)) + x_1(\theta,\phi)\partial_{x_2} f(x(\theta,\phi),$$

where $x_1(\theta,\phi)$ and $x_2(\theta,\phi)$ denote the horizontal coordinates of $x(\theta,\phi)$. As can be observed in Figure 8.6, the error for the derivative and the error for the function itself have a similar behavior in function of N. The least squares approximation converges to the exact function and the spectral derivative converges to the exact derivative; the observed convergence rates are similar.

8.5.3 Pseudo-orthogonality for the discrete inner product

We evaluate numerically the relation (8.25); it represents some "pseudo-orthogonality" of the Legendre basis, for the discrete inner product (8.23).

First, we consider the uniform quadrature rule on CS_N, defined by $\omega(x) = 4\pi/\bar{N}$. In this case, the matrix $A_N^{D\intercal} \Omega_N A_N^D$ in (8.9) is expressed as

$$A_N^{D\intercal} \Omega_N A_N^D = \frac{4\pi}{\bar{N}} A_N^{D\intercal} A_N^D.$$

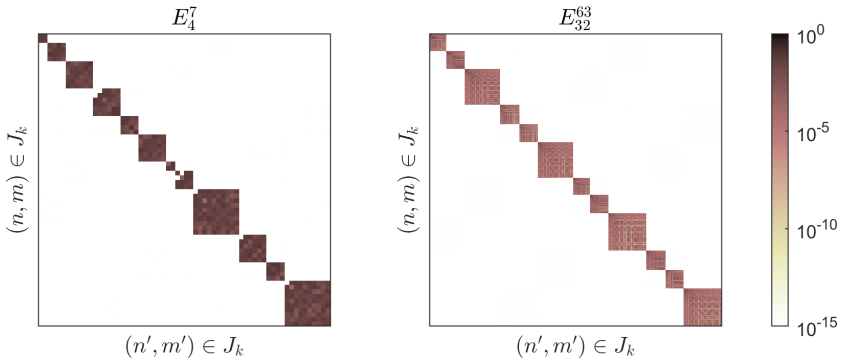


Figure 8.7: Matrix $E_N^D = [e_N(Y_n^m Y_{n'}^{m'})]$ from (8.25)-(8.24), with the uniform weight $\omega = 4\pi/\bar{N}$, $D = 2N-1$, $N = 4$ (left panel) and $N = 32$ (right panel). The indices are arranged by the classification tree of Figure 8.2. The displayed value is $10^{-15} + |e_N(Y_n^m Y_{n'}^{m'})|$, in logarithmic scale. The observed structure is the block diagonal structure predicted by Figure 8.3 (Corollary 8.10). The sparsity score is 9.961% (left panel), resp. 9.387% (right panel), which is close to a ratio of 3/32.

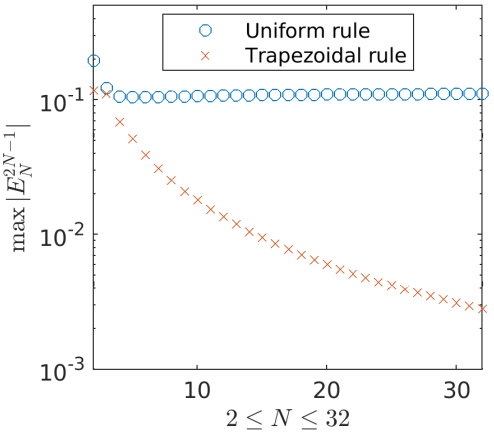


Figure 8.8: Maximal entry $\max |e_N(Y_n^m Y_{n'}^{m'})|$ of the matrix E_N^{2N-1}, $2 \le N \le 32$. The result depends on the accuracy of the quadrature rule ω. It is smaller for the trapezoidal weight than for the uniform weight.

The uniform weight is invariant under $\mathcal{G}$. Thus Theorem 8.9 predicting a sparse structure of $E_N^D = \mathrm{I}_{(D+1)^2} - A_N^{D\intercal} \Omega_N A_N^D$ can be applied. More precisely, Corollary 8.10 predicts that E_N^D has the block diagonal structure in Figure 8.3, for a suitable ordering of the indices. Figure 8.7 reports this structure, where E_N^{2N-1} is displayed for $N = 4$ and $N = 32$. In these matrices, the percentage of coefficients above 10^{-14} is respectively 9.961% and 9.387%, which is close to the ratio (8.26). Furthermore, we compute the largest entry of E_N^{2N-1} for $1 \le N \le 32$. As displayed in Figure 8.8, this value is about 0.1 (except for $N = 1$, for which the observed value is the machine epsilon). Therefore, the matrix corresponding to (LS) (without weight) is close to be proportional to the identity matrix.

$$A_N^{2N-1\intercal} A_N^{2N-1} \approx \frac{\bar{N}}{4\pi} (\mathrm{I}_{4N^2} \pm 0.1).$$

Second, we consider the weight ω of the trapezoidal rule in [156, Definition 3.1]. It is invariant under $\mathcal{G}$, so E_N^D has a sparse structure as before. This rule is second order accurate; so it is more accurate than the uniform one, and the entries of E_N^D are expected to be smaller, due to Theorem 8.8. This is confirmed in Figure 8.8; the maximal entry of E_N^{2N-1} is below 0.1, and it decays to zero when N increases.

8.6 Conclusion

This chapter considers weighted least-squares approximation by spherical harmonics on the equiangular Cubed Sphere CS_N. From a theoretical point of view, the symmetric positive semi-definite matrix of the normal equations is expected to be a pertubation of the identity matrix; the magnitude of the pertubation depends on the accuracy of the quadrature rule associated to the weight. This indicates that the Legendre spherical harmonics should be almost orthogonal for some discrete inner product on CS_N. In the case of a symmetrical weight, the matrix is block diagonal; this structure directly provides subspaces of spherical harmonics which are exactly orthogonal for the discrete inner product, disregarding the magnitude of the pertubation.

From a numerical point of view, the matrix has a condition number close to 1 if the cutoff (angular) frequency is fixed to $2N-1$, whereas it is not anymore the case if higher frequencies are considered. Numerical results indicate that $\mathcal{Y}_{2N-1}$ is a suitable approximation space for fitting or differentiating a smooth function from values on CS_N.

Future work also includes further mathematical analysis on the one hand. On the other hand, the block structure and the well conditioning of the matrix shown in Section 8.4.2 opens the way to a parallel Conjugate Gradient solver. This is a preliminary step before to investigate a genuine fast solver.

Chapter 9

A discrete Funk-Radon transform on the Cubed Sphere

9.1 Introduction

The Funk transform from [131], also called the Funk-Minkowski transform, the Funk-Radon transform, or the spherical Radon transform, is an integral transform which averages a function along great circles on the unit sphere $\mathbb{S}^2$. This transform, similar integral transforms, and associated inverse problems, are the subject of many mathematical studies, such as [138,144,152,162,170] and the references therein. These transforms play an important role in various applications, including photoacoustic tomography [139,185], Synthetic Aperture Radar [184] and diffusion Magnetic Resonance Imaging (dMRI) [141,178].

To specify one successful example from medicine, *Q-Ball Imaging* images the orientation of fibers in biological tissues [178]. The key step of this method computes the Funk transform of dMRI signals recorded on discrete spherical grids. The original computation [178] is a trapezoidal quadrature rule, applied on an interpolating function. The numerical scheme has been improved in [125, 135], using a spectral method on a regularized least squares approximation. The success[1] of the articles [125, 135, 178] attests that it is crucial to master the Funk transform in discrete configurations.

This chapter, which has been extracted from the paper [5], is devoted to a mathematical study of a discrete Funk transform, in order to provide theoretical and numerical guarantees. The studied transform is a particular case of the approaches introduced in [125, 135]; it is based on a spectral method combined with a least squares fitting. The main feature of this work is that we restrict our attention to least squares fitting without any regularization, so as to get a mathematical framework which is as clear as possible. The least squares functional comprises a fitting term, but it does not contain any artificial penalty. In particular, no regularization functional nor regularization weight have to be tuned in our approach.

The least squares problem fits values given on a spherical grid by a spherical harmonics with prescribed degree. The grid and the degree must be carefully chosen to insure that the problem is well-conditioned, which means that the corresponding matrix must have full column rank and a suitable condition number. This matrix, called a Vandermonde matrix as in [146, 147], [10], or an alternant matrix as in [108, p. 112], contains spherical harmonics restricted to the grid. In general, finding theoretically the rank and the condition number of such a matrix enters into the framework of harmonic analysis and is not an easy task. Geometrical and metric properties of the grid, as defined in [134, 137], come into play. For example, [146, Theorem 2.4]– [147, Lemma 3.13] give a lower bound on the degree to insure a full row rank property; this bound is inversely proportional to the separation distance. Another example is [110, Theorem 3.5], which proves a full column rank

[1] The website of the journal *Magnetic Resonance in Medicine* mentions 1430 citations for [178], 283 citations for [135], and 559 citations for [125], on January 05, 2023.

property, assuming that the mesh norm is smaller than the inverse of the degree.

Choosing or defining a spherical grid with suitable properties is itself an important subject. We refer to [181] for a historical presentation of several grids, and to [134] for a comparison of many popular grids, such as spiral grids, polyhedral grids, random grids, and so on. Some approaches compute an "optimal" grid as the numerical solution to an optimization problem; see for instance [119] for various optimization criteria, including the conditioning of a least squares problem. Some other approaches define grids in an elementary explicit way. Among these simple grids, the equiangular Cubed Sphere [168, 171] is obtained by radial projection of a circumscribed cube, from cartesian lines on the faces of the cube towards great circles on the sphere.

This Cubed Sphere (and some variants) is very popular and is widely studied in numerical climatology and meteorology; see for instance [120, 140, 142, 143, 151, 153, 154, 159, 161, 164, 176]. We have recently studied various approximation schemes on this grid using spherical harmonics. Lagrange interpolation has been considered in [10], a spherical quadrature rule in [9], and least squares approximation in [11], as presented in the previous chapters. Among the results, [11] gives the largest degree which numerically guarantees a condition number that is uniformly bounded. In this chapter, we propose a further study concerning spectral computing on the Cubed Sphere. We investigate the use of this grid for computing Funk transforms.

Our methodology contains two steps. In a first step, we define a family of discrete Funk transforms which act between spaces of grid functions, for a general grid. They are obtained as in [125,135], but without any regularization, and with an evaluation of the (continuous) transform on the initial grid. We prove new properties satisfied by these transforms, in order to give some mathematical background. In particular, we show that the pseudoinverse of such a transform represents an inverse discrete Funk transform very analogous to the direct one. We also provide a theoretical estimation of stability, which mainly depends on the conditioning of the least squares problem. It implies that stability is guaranteed as soon as the least squares problem is well-conditioned.

Then, in a second step, we focus on a framework which guarantees this condition of stability. We select the equiangular Cubed Sphere for the grid and we introduce a rule on the degree such that the conditioning is kept under control. The study is similar with [11], but the dimensions of the least squares problem have been reduced due to our specific problem. Indeed, the null space of the Funk transform contains any odd function, so we assume from the beginning that the approximation space contains only even spherical harmonics. Also, symmetry consideration allows to halve the grid, so we restrict the Cubed Sphere to an hemisphere.

The paper is organized as follows. In Section 9.2, we summarize some notation and background concerning spherical computation. In Section 9.3, we study a discrete Funk transform, based on a spectral method applied on a least squares fitting. In Section 9.4, we focus on the case where the grid is the equiangular Cubed Sphere. In Section 9.5, the relevance of the approach is shown by various numerical tests, such as test of accuracy and stability on synthetic dMRI signals.

9.2 Background and notation

9.2.1 Even spherical harmonics

We keep the notation of Chapter 6 for the spherical harmonics. For every $D \geq 0$, the functions $(Y_n^m)_{|m|\leq n\leq D}$, given by (6.3), define an orthonormal basis of the space $\mathcal{Y}_D$ of the spherical harmonics with degree at most D. We introduce the subspace of the even functions in $\mathcal{Y}_D$, denoted by $\mathcal{Y}_D^{\text{ev}}$; it is spanned by the even degrees, *i.e.*

$$\mathcal{Y}_D^{\text{ev}} = \text{span}\{Y_{2n}^m,\, 0 \leq n \leq \tfrac{D}{2},\, |m| \leq 2n\}.$$

In the sequel, we always assume that the degree D is even when considering $\mathcal{Y}_D^{\text{ev}}$ (because $\mathcal{Y}_D^{\text{ev}} = \mathcal{Y}_{D-1}^{\text{ev}}$ otherwise); under this assumption, the dimension of $\mathcal{Y}_D^{\text{ev}}$ is given by

$$d_D = \tfrac{1}{2}(D+1)(D+2).$$

9.2.2 Funk transform

The Funk transform, denoted by $\mathcal{F}$, maps a spherical function $f : \mathbb{S}^2 \to \mathbb{R}$ to a spherical function $\mathcal{F}f : \mathbb{S}^2 \to \mathbb{R}$ as follows. For any unit vector $\alpha \in \mathbb{S}^2$, $\mathcal{F}f(\alpha)$ is defined as the average of f along the great circle that is orthogonal to α, *i.e.*

$$\mathcal{F}f(\alpha) = \frac{1}{2\pi} \int_{\{x \in \mathbb{S}^2 : x \cdot \alpha = 0\}} f \, \mathrm{d}s, \quad \alpha \in \mathbb{S}^2, \quad f : \mathbb{S}^2 \to \mathbb{R}, \tag{9.1}$$

where s denotes the length measure on the circle $\{x \in \mathbb{S}^2 : x \cdot \alpha = 0\}$; in this definition, the function f is required to be integrable along any great circle (with respect to the length measure), so that the integrals are defined.

The Funk transform $\mathcal{F}f$ is an even function, *i.e.* $\mathcal{F}f(-\alpha) = \mathcal{F}f(\alpha)$, $\alpha \in \mathbb{S}^2$. If f is odd, *i.e.* $f(-x) = -f(x)$, $x \in \mathbb{S}^2$, then $\mathcal{F}f = 0$. In any case, $\mathcal{F}f = \mathcal{F}f^{\mathrm{ev}}$, where $f^{\mathrm{ev}}(x) = \frac{1}{2}(f(x) + f(-x))$ denotes the even part of f. For these reasons, the Funk transform can be considered between spaces of even functions, without loss of generality. We follow this convention throughout the article. Hence, in the sequel, we consider even functions only.

Spherical harmonics are eigenfunctions of the Funk transform $\mathcal{F}$ [131], so that it defines an isomorphism on $\mathcal{Y}_D^{\mathrm{ev}}$,

$$\mathcal{F} : \mathcal{Y}_D^{\mathrm{ev}} \to \mathcal{Y}_D^{\mathrm{ev}}, \quad \mathcal{F}Y_{2n}^m = P_{2n}(0)\, Y_{2n}^m, \quad \text{with} \quad P_{2n}(0) = (-1)^n \tfrac{1\cdot 3\cdot 5\cdots(2n-1)}{2\cdot 4\cdot 6\cdots(2n)}, \quad |m| \le 2n,\ 0 \le n \le \tfrac{D}{2}. \tag{9.2}$$

The associated nonsingular matrix is the block diagonal matrix

$$\Lambda = \mathrm{diag}\left[(-1)^n \tfrac{1\cdot 3\cdot 5\cdots(2n-1)}{2\cdot 4\cdot 6\cdots(2n)} \mathrm{I}_{4n+1}, 0 \le n \le \tfrac{D}{2}\right] \in \mathbb{R}^{d_D \times d_D}. \tag{9.3}$$

This structure suggests the spectral method for computing Funk transforms, as it has been introduced in [125, 135].

9.2.3 Grid functions

In Subsection 6.2.2, we have introduced grid functions in the case of the Cubed Sphere grid. We extend the notation to other grids as follows. A spherical grid is defined as a finite subset of the unit sphere, $G \subset \mathbb{S}^2$. A grid function on G is a function $b : G \to \mathbb{R}$ defined on G. The space of such functions is denoted by

$$\mathbb{R}^G = \{b : G \to \mathbb{R}\}.$$

Numbering the elements of G by $\xi_1, \ldots, \xi_M$, where M denotes the cardinal number, the canonical basis $(\delta_{\xi_i})_{1 \le i \le M}$ of $\mathbb{R}^G$ is defined by

$$\delta_{\xi_i}(\xi_j) = \begin{cases} 1, \text{ if } i = j, \\ 0, \text{ otherwise,} \end{cases} \qquad 1 \le i, j \le M.$$

In this basis, any $b \in \mathbb{R}^G$ is represented by the column vector $\boldsymbol{b} = [b(\xi_i)]_{1 \le i \le M} \in \mathbb{R}^M$, due to

$$b = \sum_{i=1}^{M} b(\xi_i) \delta_{\xi_i}.$$

For any real function defined on the sphere, $f : \mathbb{S}^2 \to \mathbb{R}$, the restriction of f on the grid G is the grid function $f|_G \in \mathbb{R}^G$ defined by

$$f|_G := \sum_{i=1}^{M} f(\xi_i) \delta_{\xi_i}, \quad f|_G(\xi_i) = f(\xi_i),\ 1 \le i \le M.$$

9.3 Discrete Funk transform on a spherical grid

In this section, we study a discrete Funk transform on a general grid. We assume that

- $G = \{\xi_1, \dots, \xi_M\} \subset \mathbb{S}^2$ is a spherical grid with cardinal number M,
- $b \in \mathbb{R}^G$ is a given grid function on G,
- $D \geq 0$ is a fixed even degree.

9.3.1 Least squares fitting

One looks for an even spherical harmonics $f \in \mathcal{Y}_D^{\mathrm{ev}}$ which fits the grid function b. The least squares problem minimizes a fitting error as follows,

$$\inf_{f \in \mathcal{Y}_D^{\mathrm{ev}}} \sum_{i=1}^{M} (f(\xi_i) - b(\xi_i))^2. \tag{LS$^{\mathrm{ev}}$}$$

Remark 9.1. For the particular case $G = \mathrm{CS}_N$, the problem (LS$^{\mathrm{ev}}$) is similar to (LS), with the specificity that the approximation space is restricted to even spherical harmonics.

We introduce the basis (Y_{2n}^m) of $\mathcal{Y}_D^{\mathrm{ev}}$. Then any $f \in \mathcal{Y}_D^{\mathrm{ev}}$ admits a spectral expansion (6.4),

$$f = \sum_{0 \leq n \leq D/2,\, |m| \leq 2n} \hat{f}_{2n}^m Y_{2n}^m \in \mathcal{Y}_D^{\mathrm{ev}}, \quad \text{with} \quad \hat{f} = [\hat{f}_{2n}^m]_{0 \leq n \leq D/2,\, |m| \leq 2n} \in \mathbb{R}^{d_D};$$

the matrix of the linear map $f \in \mathcal{Y}_D^{\mathrm{ev}} \mapsto [f(\xi_i)]_{1 \leq i \leq M} \in \mathbb{R}^M$ is given by the Vandermonde matrix

$$A = [Y_{2n}^m(\xi_i)]_{\substack{1 \leq i \leq M \\ 0 \leq n \leq D/2,\, |m| \leq 2n}} \in \mathbb{R}^{M \times d_D}. \tag{9.4}$$

Here, the row index is i, and the column index is the couple (n, m). Assuming a lexicographic ordering for (n, m), an expanded form of A is given by

$$A = \begin{bmatrix} Y_0^0(\xi_1) & \cdots & Y_{2n}^{-2n}(\xi_1) & \cdots & Y_{2n}^m(\xi_1) & \cdots & Y_{2n}^{2n}(\xi_1) & \cdots & Y_D^D(\xi_1) \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ Y_0^0(\xi_i) & \cdots & Y_{2n}^{-2n}(\xi_i) & \cdots & Y_{2n}^m(\xi_i) & \cdots & Y_{2n}^{2n}(\xi_i) & \cdots & Y_D^D(\xi_i) \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ Y_0^0(\xi_M) & \cdots & Y_{2n}^{-2n}(\xi_M) & \cdots & Y_{2n}^m(\xi_M) & \cdots & Y_{2n}^{2n}(\xi_M) & \cdots & Y_D^D(\xi_M) \end{bmatrix}.$$

Then, the problem (LS$^{\mathrm{ev}}$) can be written in matrix form as

$$\inf_{\hat{f} \in \mathbb{R}^{d_D}} \|A\hat{f} - \boldsymbol{b}\|^2,$$

where $\|\cdot\|$ denotes the euclidean norm in $\mathbb{R}^M$.

In this chapter, we assume that the grid G and the degree D are such that the Vandermonde matrix A has full column rank. Then the problem (LS$^{\mathrm{ev}}$) admits a unique solution. This solution, denoted[2] by

$$\ell[b] \in \mathcal{Y}_D^{\mathrm{ev}}, \quad \ell[b] = \arg \inf_{f \in \mathcal{Y}_D^{\mathrm{ev}}} \sum_{i=1}^{M} (f(\xi_i) - b(\xi_i))^2, \tag{9.5}$$

is given by

$$\ell[b] = [Y_{2n}^m(\cdot)]_{0 \leq n \leq D/2,\, |m| \leq 2n}^{\intercal} \widehat{\ell[b]}, \quad \text{with} \quad \widehat{\ell[b]} = (A^{\intercal}A)^{-1}A^{\intercal}\boldsymbol{b} \in \mathbb{R}^{d_D}. \tag{9.6}$$

[2] ℓ as the first letter of "least squares".

Here, the vector $\widehat{\ell[b]}$ of the spectral coefficients satisfies a linear system, whose matrix is symmetric and positive-definite:

$$A^\intercal A\,\widehat{\ell[b]} = A^\intercal \boldsymbol{b}.$$

For the matrix norm induced by the euclidean norm, the condition number of this linear system is given by

$$\operatorname{cond}(A^\intercal A) = \operatorname{cond}(A)^2, \quad \text{with} \quad \operatorname{cond}(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)},$$

where $\sigma_{\max}$, resp. $\sigma_{\min}$, denote the maximum, resp. minimum, singular value. In Section 9.4, we propose a choice of G and D which guarantees (at least numerically) that the condition number of the Vandermonde matrix is close to 1 ($\operatorname{cond} A \approx 1$).

Remark 9.2. If the condition number is large ($\operatorname{cond} A \gg 1$), or if A has not full column rank, then regularization is needed. This is outside the scope of this work, so we refer in this case to [133] for a general reference about ill-posed problems, [111] for various regularization operators dealing with spherical harmonics on the sphere, and [125, 135] for regularization in a framework of Funk transforms.

We conclude this subsection by a simple result which permits to halve the grid in the case of a central symmetry, provided that the grid function is replaced by its even part.

Proposition 9.3. *Assume that G in invariant with respect to the central symmetry, so that M is even, and $\xi_{M/2+i} = -\xi_i$, $1 \le i \le M/2$ (up to a reordering). Then the problem* ($\mathrm{LS}^{\mathrm{ev}}$) *is equivalent to*

$$\inf_{f \in \mathcal{Y}_D^{\mathrm{ev}}} \sum_{i=1}^{M/2} (f(\xi_i) - b^{\mathrm{ev}}(\xi_i))^2, \quad \textit{with} \quad b^{\mathrm{ev}}(\xi_i) = \tfrac{1}{2}(b(\xi_i) + b(-\xi_i)),\ 1 \le i \le M/2.$$

Proof. The grid is invariant under the central symmetry $-\mathrm{I}_3$ ($\xi \leftrightarrow -\xi$), so we split the grid function b into $b = b^{\mathrm{ev}} + b^{\mathrm{odd}}$, where $b^{\mathrm{ev}}(\xi) = \frac{1}{2}(b(\xi) + b(-\xi))$ is an even grid function ($b^{\mathrm{ev}}(-\xi) = b^{\mathrm{ev}}(\xi)$), and $b^{\mathrm{odd}}(\xi) = \frac{1}{2}(b(\xi) - b(-\xi))$ is odd ($b^{\mathrm{odd}}(-\xi) = -b^{\mathrm{odd}}(\xi)$). Therefore, for any $f \in \mathcal{Y}_D^{\mathrm{ev}}$,

$$\sum_{i=1}^{M} (f(\xi_i) - b(\xi_i))^2 = \sum_{i=1}^{M} (f(\xi_i) - b^{\mathrm{ev}}(\xi_i))^2 + \sum_{i=1}^{M} b^{\mathrm{odd}}(\xi_i)^2 - 2\sum_{i=1}^{M} (f(\xi_i) - b^{\mathrm{ev}}(\xi_i)) b^{\mathrm{odd}}(\xi_i).$$

In the right hand side, the first term is twice the sum indexed by $1 \le i \le M/2$, because $(f - b^{\mathrm{ev}})^2$ is an even grid function. The second term is a constant C which does not depend on f. The third term is null, because $(f - b^{\mathrm{ev}})b^{\mathrm{odd}}$ is an odd grid function. Therefore,

$$\sum_{i=1}^{M} (f(\xi_i) - b(\xi_i))^2 = 2\sum_{i=1}^{M/2} (f(\xi_i) - b^{\mathrm{ev}}(\xi_i))^2 + C,$$

which proves the result. □

9.3.2 Discrete Funk transform

We study various mathematical properties of a discrete Funk transform defined as follows.

Definition 9.4 (Discrete Funk transform)**.** Let $G = \{\xi_1, \dots, \xi_M\} \subset \mathbb{S}^2$ be a spherical grid and $D \ge 0$ be an even degree, such that the Vandermonde matrix A in (9.4) has full column rank. The discrete Funk transform F is defined as a linear mapping between spaces of grid functions, by

$$\begin{array}{rccc} \mathrm{F}: & \mathbb{R}^G & \longrightarrow & \mathbb{R}^G \\ & b & \longmapsto & \mathrm{F}[b] = \Big(\mathcal{F}(\ell[b])\Big)\Big|_G, \end{array} \tag{9.7}$$

where $\ell[b]$ is the least squares fitting in (9.5), and $\mathcal{F}$ is the Funk transform in (9.1). In other words, the discrete Funk transform of a grid function is the Funk transform applied on the least squares fitting, then restricted to the initial grid.

Property 9.5. *In the basis $(\delta_{\xi_i})_{1\le i\le M}$ of $\mathbb{R}^G$, the matrix of the discrete Funk transform* F *is given by*

$$\mathbf{F} = A\,\Lambda\,(A^\intercal A)^{-1}A^\intercal \in \mathbb{R}^{M\times M}. \tag{9.8}$$

Proof. The matrix of the least squares operator, $\ell : \mathbb{R}^G \to \mathcal{Y}_D^{\mathrm{ev}}$, in the bases (δ_{ξ_i}) and (Y_{2n}^m), is given by $(A^\intercal A)^{-1}A^\intercal$, due to (9.6). The matrix of the Funk transform $\mathcal{F} : \mathcal{Y}_D^{\mathrm{ev}} \to \mathcal{Y}_D^{\mathrm{ev}}$, in the basis (Y_{2n}^m), is the matrix Λ in (9.3). And the matrix of $f \in \mathcal{Y}_D^{\mathrm{ev}} \mapsto f|_G \in \mathbb{R}^G$, in the bases (Y_{2n}^m) and (δ_{ξ_i}), is given by the Vandermonde matrix A. The discrete Funk transform F is the composition of these linear maps, so its matrix $\mathbf{F}$ is given by the product of the matrices. □

The following result establishes that the spherical function $\mathcal{F}(\ell[b])$ can be exactly recovered from its restriction Fb on G (so that the restriction is "lossly").

Proposition 9.6. *The discrete Funk transform* $\mathrm{F} : \mathbb{R}^G \to \mathbb{R}^G$ *and the Funk transform* $\mathcal{F} : \mathcal{Y}_D^{\mathrm{ev}} \to \mathcal{Y}_D^{\mathrm{ev}}$ *are related by*

$$\ell \circ \mathrm{F} = \mathcal{F} \circ \ell.$$

In other words, the least squares fitting of the discrete Funk transform coincides with the Funk transform of the least squares fitting,

$$\ell[\mathrm{F}b](\alpha) = \mathcal{F}(\ell[b])(\alpha), \quad b \in \mathbb{R}^G,\ \alpha \in \mathbb{S}^2.$$

Proof. Similarly as the proof of Property 9.5, the matrix of $\ell \circ \mathrm{F}$ is given by

$$(A^\intercal A)^{-1}A^\intercal \cdot A\Lambda\,(A^\intercal A)^{-1}A^\intercal = \Lambda\,(A^\intercal A)^{-1}A^\intercal,$$

where we recognize the matrix of $\mathcal{F} \circ \ell$ on the right hand side. □

In practice, $f : \mathbb{S}^2 \to \mathbb{R}$ is a spherical function that is sampled on the grid G, so that the given data is $b = f|_G$. One uses the discrete Funk transform $\mathrm{F}[f|_G]$, or equivalently $\mathcal{F}(\ell[f|_G])$, in order to approximate some values $\mathcal{F}f(\alpha)$ of the Funk transform $\mathcal{F}f$. The following result shows that this method is exact if $f \in \mathcal{Y}_D^{\mathrm{ev}}$.

Theorem 9.7 (Exactness on $\mathcal{Y}_D^{\mathrm{ev}}$)**.** *The discrete Funk transform is exact on* $\mathcal{Y}_D^{\mathrm{ev}}$*, which means that*

$$\mathrm{F}[f|_G] = (\mathcal{F}f)|_G, \quad f \in \mathcal{Y}_D^{\mathrm{ev}}. \tag{9.9}$$

More generally, for every $f \in \mathcal{Y}_D^{\mathrm{ev}}$, the Funk transform $\mathcal{F}f$ can be computed exactly from the grid function $f|_G$, with

$$\mathcal{F}f(\alpha) = [Y_{2n}^m(\alpha)]_{0\le n\le D/2,\,|m|\le 2n}^\intercal \Lambda\,(A^\intercal A)^{-1}\,A^\intercal\,[f(\xi_i)]_{1\le i\le M}, \quad \alpha \in \mathbb{S}^2, \quad f \in \mathcal{Y}_D^{\mathrm{ev}}. \tag{9.10}$$

Proof. Any function $f \in \mathcal{Y}_D^{\mathrm{ev}}$ fits exactly the grid values $[f(\xi_i)]_{1\le i\le M}$, so that the unique solution of $(\mathrm{LS}^{\mathrm{ev}})$ with $b = f|_G$ is the initial function f itself,

$$\ell[f|_G] = f, \quad f \in \mathcal{Y}_D^{\mathrm{ev}}.$$

Injecting this equality into the definition of $\mathrm{F}[f|_G]$ proves (9.9). Also, we obtain $\mathcal{F}f = \mathcal{F}(\ell[f|_G])$; hence, we have (9.10) due to the matrix of $\mathcal{F} \circ \ell$ (see the proof of Proposition 9.6). □

Now, we investigate the inversion of the discrete Funk transform F. We introduce the Moore-Penrose pseudoinverse $\mathbf{F}^\dagger$ of the matrix $\mathbf{F}$, since it is not expected to be nonsingular. We refer to [132, pp. 257-258] for usual consideration about such a pseudoinverse. In our case, the pseudoinverse $\mathbf{F}^\dagger$ maps any $\boldsymbol{c} \in \mathbb{R}^M$ to the minimum norm solution $\boldsymbol{b} = \mathbf{F}^\dagger\boldsymbol{c} \in \mathbb{R}^M$ of the least squares problem $\inf_{\boldsymbol{b}\in\mathbb{R}^M} \|\mathbf{F}\boldsymbol{b} - \boldsymbol{c}\|^2$. We prove that the pseudoinverse $\mathbf{F}^\dagger$ represents an *inverse discrete Funk transform* that is analogous to the direct transform F.

Theorem 9.8 (Pseudoinversion). *The Moore-Penrose pseudoinverse of* $\mathbf{F}$ *is given by*

$$\mathbf{F}^\dagger = A\,\Lambda^{-1}\,(A^\intercal A)^{-1}A^\intercal \in \mathbb{R}^{M\times M}. \tag{9.11}$$

Therefore, the pseudoinverse $\mathbf{F}^\dagger$ *represents the* inverse discrete Funk transform $\mathrm{F}^\dagger$*, defined by*

$$\begin{array}{rccl} \mathrm{F}^\dagger: & \mathbb{R}^G & \longrightarrow & \mathbb{R}^G \\ & c & \longmapsto & \mathrm{F}^\dagger[c] = \Big(\mathcal{F}^{-1}(\ell[c])\Big)\Big|_G. \end{array} \tag{9.12}$$

Proof. The matrix $\mathbf{F}^\dagger$ is the Moore-Penrose pseudoinverse of $\mathbf{F}$ (and conversely), because (9.8) and (9.11) imply that the four Moore-Penrose conditions [132, p. 257] are satisfied:

$$\mathbf{F}\mathbf{F}^\dagger\mathbf{F} = \mathbf{F}, \quad \mathbf{F}^\dagger\mathbf{F}\mathbf{F}^\dagger = \mathbf{F}^\dagger, \quad (\mathbf{F}\mathbf{F}^\dagger)^\intercal = \mathbf{F}\mathbf{F}^\dagger, \quad (\mathbf{F}^\dagger\mathbf{F})^\intercal = \mathbf{F}^\dagger\mathbf{F}.$$

Furthermore, the matrix of the transform $\mathrm{F}^\dagger$ defined in (9.12) is given by $\mathbf{F}^\dagger$. This result and its proof are analogous to Property 9.5. The difference is that the diagonal matrix Λ of the isomorphic Funk transform $\mathcal{F}: \mathcal{Y}_D^{\mathrm{ev}} \to \mathcal{Y}_D^{\mathrm{ev}}$ is replaced by the inverse diagonal matrix Λ^{-1}, since it represents the inverse transform $\mathcal{F}^{-1}$. □

The relations (9.7) and (9.12) are very similar, so are (9.8) and (9.11). More generally, as soon as some result is established for one of the transforms F and $\mathrm{F}^\dagger$, some counterpart is expected for the other one. For instance, the counterpart of Proposition 9.6 is given hereafter.

Proposition 9.9. *The inverse discrete Funk transform* $\mathrm{F}^\dagger: \mathbb{R}^G \to \mathbb{R}^G$ *and the inverse Funk transform* $\mathcal{F}^{-1}: \mathcal{Y}_D^{\mathrm{ev}} \to \mathcal{Y}_D^{\mathrm{ev}}$ *satisfy*

$$\ell \circ \mathrm{F}^\dagger = \mathcal{F}^{-1} \circ \ell.$$

In other words, the least squares fitting of the inverse discrete Funk transform coincides with the inverse Funk transform of the least squares fitting,

$$\ell[\mathrm{F}^\dagger c](\xi) = \mathcal{F}^{-1}(\ell[c])(\xi), \quad c \in \mathbb{R}^G,\ \xi \in \mathbb{S}^2.$$

Proof. Analogous to the proof of Proposition 9.6. □

Now, we express mapping properties of $\mathbf{F}$ and $\mathbf{F}^\dagger$ in term of the Vandermonde matrix A.

Proposition 9.10. *The following assertions hold.*

(i) The composition of $\mathbf{F}$ *and* $\mathbf{F}^\dagger$ *coincides with the orthogonal projection on* $\operatorname{Ran} A$*,*

$$\mathbf{F}\mathbf{F}^\dagger = \mathbf{F}^\dagger\mathbf{F} = A(A^\intercal A)^{-1}A^\intercal. \tag{9.13}$$

In particular,

$$\forall \boldsymbol{b} \in \operatorname{Ran} A, \quad \mathbf{F}^\dagger\mathbf{F}\,\boldsymbol{b} = \mathbf{F}\mathbf{F}^\dagger\,\boldsymbol{b} = \boldsymbol{b}. \tag{9.14}$$

(ii) The null space and the range of $\mathbf{F}$ *satisfy*

$$\operatorname{Ker}\mathbf{F} = \operatorname{Ker} A^\intercal = (\operatorname{Ran} A)^\perp, \quad \operatorname{Ran}\mathbf{F} = \operatorname{Ran} A, \quad \mathbb{R}^M = \operatorname{Ker}\mathbf{F} \overset{\perp}{\oplus} \operatorname{Ran}\mathbf{F}.$$

(iii) The null space and the range of $\mathbf{F}^\dagger$ *satisfy* $\operatorname{Ker}\mathbf{F}^\dagger = \operatorname{Ker}\mathbf{F}$*,* $\operatorname{Ran}\mathbf{F}^\dagger = \operatorname{Ran}\mathbf{F}$*.*

Proof. (i) Since A has full column rank, the orthogonal projection on $\operatorname{Ran} A$ is given by the matrix $\Pi = A(A^\intercal A)^{-1}A^\intercal$. Then $\mathbf{F}\mathbf{F}^\dagger = \mathbf{F}^\dagger\mathbf{F} = \Pi$ can be easily checked with (9.8) and (9.11). And this implies (9.14) due to $\Pi\boldsymbol{b} = \boldsymbol{b}$, for any $\boldsymbol{b} \in \operatorname{Ran} A$.
(ii) The orthogonal decomposition $\mathbb{R}^M = \operatorname{Ker} A^\intercal \oplus \operatorname{Ran} A$ is a consequence of classical linear algebra. Secondly, $\operatorname{Ker} A^\intercal \subset \operatorname{Ker}\mathbf{F}$ is easily seen in (9.8), and $\operatorname{Ker}\mathbf{F} \subset \operatorname{Ker}\mathbf{F}^\dagger\mathbf{F} = \operatorname{Ker}\Pi$, with $\operatorname{Ker}\Pi = (\operatorname{Ran} A)^\perp = \operatorname{Ker} A^\intercal$. Thirdly, $\operatorname{Ran}\mathbf{F} \subset \operatorname{Ran} A$ is easily seen in (9.8); furthermore, (9.14) proves that $\operatorname{Ran} A \subset \operatorname{Ran}\mathbf{F}\mathbf{F}^\dagger$, with $\operatorname{Ran}\mathbf{F}\mathbf{F}^\dagger \subset \operatorname{Ran}\mathbf{F}$. The last equality is a consequence of the first two ones.
(iii) The null space and the range of $\mathbf{F}^\dagger$ are obtained analogously as those of $\mathbf{F}$. □

Translating this proposition to grid functions results in the following corollary.

Corollary 9.11. *The following asssertions hold.*

(i) The composition of F *and* $\mathrm{F}^\dagger$ *coincides with the restriction of the least squares fitting,*

$$\mathrm{F}\circ\mathrm{F}^\dagger = \mathrm{F}^\dagger\circ\mathrm{F}: \begin{array}{ccc}\mathbb{R}^G & \longrightarrow & \mathbb{R}^G \\ b & \longmapsto & \ell[b]|_G\,.\end{array} \tag{9.15}$$

(ii) The transform $\mathrm{F}^\dagger$ *is the usual inverse transform of* F*, if the spaces are restricted to the subspace* $\mathcal{Y}_D^{\mathrm{ev}}|_G := \{f|_G,\, f\in\mathcal{Y}_D^{\mathrm{ev}}\}$, i.e.*, the linear mappings*

$$b\in\mathcal{Y}_D^{\mathrm{ev}}|_G \mapsto \mathrm{F}b\in\mathcal{Y}_D^{\mathrm{ev}}|_G, \quad c\in\mathcal{Y}_D^{\mathrm{ev}}|_G \mapsto \mathrm{F}^\dagger c\in\mathcal{Y}_D^{\mathrm{ev}}|_G,$$

are two isomorphisms which are inverses of each other.

Proof. (i) The relation (9.15) is the translation of (9.13), from matrices to their linear maps.
(ii) The subspace $\mathcal{Y}_D^{\mathrm{ev}}|_G\subset\mathbb{R}^G$ is the translation to grid functions of the space $\operatorname{Ran}A$. Translating (9.14) shows that

$$\forall b\in\mathcal{Y}_D^{\mathrm{ev}}|_G, \quad (\mathrm{F}\circ\mathrm{F}^\dagger)(b) = (\mathrm{F}^\dagger\circ\mathrm{F})(b) = b.$$

The combination of this result with Proposition 9.10.(ii-iii) shows the result. □

To finish with, we provide estimations of stability. They show that stability is expected if the condition number of the Vandermonde matrix A is suitable.

Theorem 9.12 (Stability)**.** *The maximum singular value of* $\mathbf{F}$*, resp.* $\mathbf{F}^\dagger$*, satisfies*

$$\sigma_{\max}(\mathbf{F}) \le \operatorname{cond}A, \quad \sigma_{\max}(\mathbf{F}^\dagger) \le \frac{\operatorname{cond}A}{|P_{2N-2}(0)|} \underset{N\to\infty}{\sim} \sqrt{\pi N}\operatorname{cond}A. \tag{9.16}$$

Remark 9.13. The largest singular values represent stability constants, since perturbing a vector $\boldsymbol{b}\in\mathbb{R}^M$ by $\boldsymbol{\epsilon}\in\mathbb{R}^M$ induces a perturbation on the transform $\mathbf{F}\boldsymbol{b}$, resp. $\mathbf{F}^\dagger\boldsymbol{b}$, which satisfies $\|\mathbf{F}(\boldsymbol{b}+\boldsymbol{\epsilon})-\mathbf{F}\boldsymbol{b}\|\le\sigma_{\max}(\mathbf{F})\|\boldsymbol{\epsilon}\|$, $\|\mathbf{F}^\dagger(\boldsymbol{b}+\boldsymbol{\epsilon})-\mathbf{F}^\dagger\boldsymbol{b}\|\le\sigma_{\max}(\mathbf{F}^\dagger)\|\boldsymbol{\epsilon}\|$.

Proof. The maximum singular value $\sigma_{\max}$ coincides with the matrix norm induced by the euclidean norm and is therefore sub-multiplicative. Hence, we deduce from (9.8) that

$$\sigma_{\max}(\mathbf{F}) \le \sigma_{\max}(A)\,\sigma_{\max}(\Lambda)\,\sigma_{\max}(A^\dagger),$$

where $A^\dagger = (A^\intercal A)^{-1}A^\intercal$ is the Moore-Penrose pseudoinverse of the injective matrix A. On the right hand-side, $\sigma_{\max}(\Lambda) = P_0(0) = 1$, and $\sigma_{\max}(A^\dagger) = \frac{1}{\sigma_{\min}(A)}$ is the inverse of the minimum singular value of A. Therefore,

$$\sigma_{\max}(\mathbf{F}) \le \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} = \operatorname{cond}(A).$$

For similar reasons, we see with (9.11) that

$$\sigma_{\max}(\mathrm{F}^\dagger) \le \sigma_{\max}(A)\,\sigma_{\max}(\Lambda^{-1})\,\sigma_{\max}(A^\dagger) = \frac{\operatorname{cond}(A)}{|P_{2N-2}(0)|}.$$

Here, $\sigma_{\max}(\Lambda^{-1}) = \frac{1}{|P_{2N-2}(0)|}$, with $P_{2N-2}(0)$ given by (9.2); the asymptotics $\frac{1}{|P_{2N-2}(0)|}\sim\sqrt{\pi N}$ can checked with the Stirling formula $n!\sim\sqrt{2\pi n}\exp(-n)n^n$. □

9.4 Discrete Funk transform on the Cubed Hemisphere

In this section, we investigate the discrete Funk transform in the case of the equiangular Cubed Sphere CS_N.

9.4.1 Cubed Hemisphere

To begin with, for every $N \geq 1$, the grid CS_N is invariant under the central symmetry (see Theorem 5.3). Hence, Proposition 9.3 shows that any least squares problem ($\mathrm{LS}^{\mathrm{ev}}$) on this grid can be reduced to a problem on a half-grid, without changing the solution. Therefore, we restrict the grid CS_N to the Northern hemisphere and a half of the equator circle, without loss of generality. The resulting grid is displayed in Figure 9.1 and is defined below.

Definition 9.14. Let $N \geq 1$. The *Cubed Hemisphere* CH_N is defined by

$$\mathrm{CH}_N = \mathrm{CS}_N \cap \{x(\theta,\phi) \in \mathbb{S}^2, \quad (\theta > 0 \text{ and } 0 \leq \phi < 2\pi) \text{ or } (\theta = 0 \text{ and } 0 \leq \phi < \pi)\}, \tag{9.17}$$

so that CS_N splits into $\mathrm{CS}_N = \mathrm{CH}_N \cup (-\mathrm{CH}_N)$, and CH_N has the cardinal number $3N^2+1$.

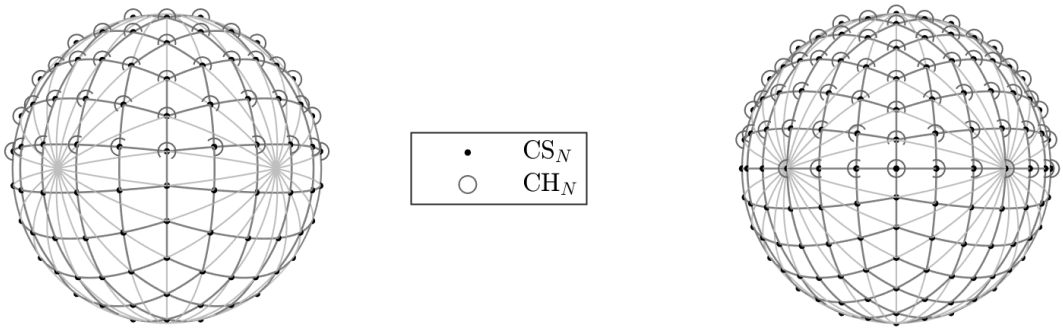


Figure 9.1: Cubed Sphere and Cubed Hemisphere. The Cubed Sphere CS_N (black dots) defined in (5.1) is obtained by intersecting equiangular meridian circles (gray lines). The Cubed Hemisphere CH_N (gray circles) defined in (9.17) is located in the Northern hemisphere; it contains half of the points from CS_N. Left: N is odd ($N = 5$). Right: N is even ($N = 6$).

In the remainder of this section, we consider the grid

$$G = \mathrm{CH}_N = \{\xi_i,\ 1 \leq i \leq M\}, \quad M = 3N^2+1,$$

where $N \geq 1$ is fixed.

9.4.2 Degree

We tune the degree D in term of the parameter N, so that the problem ($\mathrm{LS}^{\mathrm{ev}}$) is well-conditioned. We argue that the value $D = 2N-2$ is a suitable choice.

The main motivation is the following claim, which is related to the study in Chapter 8.

Claim 9.15. *Let $N \geq 1$, $G = \mathrm{CH}_N$, and $D = 2N-2$. Then, the corresponding Vandermonde matrix $A \in \mathbb{R}^{(3N^2+1)\times(2N-1)N}$ defined in* (9.4) *has full column rank and is well-conditioned, with a condition number uniformly bounded with N.*

Unfortunately, a complete proof of Claim 9.15 is not available yet. The most convincing argument that we have at disposal is the numerical evidence displayed in Figure 9.2. These numerical results indicate that

$$\operatorname{cond} A \leq 2^{1/4} \leq 1.2, \quad 1 \leq N \leq 64,$$

and they suggest that $\operatorname{cond} A$ grows to $2^{1/4}$ when $N \to \infty$. We also have a partial proof, concerning the full column rank property in the case $N \leq 4$.

Proposition 9.16. *Assume that $1 \leq N \leq 4$, $G = \mathrm{CH}_N$, and consider the degree $D = 2N-2$. Then, the Vandermonde matrix A in* (9.4) *has full column rank.*

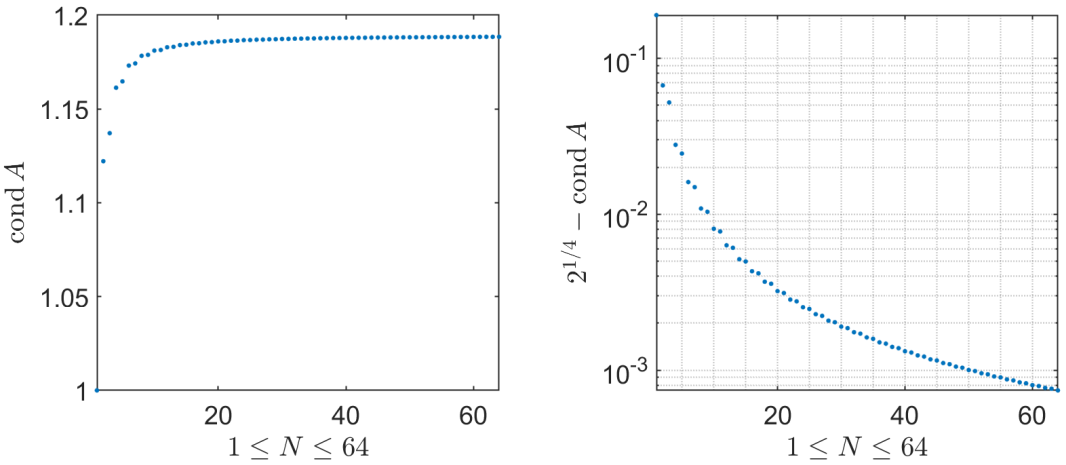


Figure 9.2: Numerical evidence of Claim 9.15: the condition number of the Vandermonde matrix A is plotted for $G = \mathrm{CH}_N$, $D = 2N-2$, and $1 \leq N \leq 64$. Left: $\mathrm{cond}\, A$ is bounded from above by 1.2. Right: $2^{1/4} - \mathrm{cond}\, A$ decays to zero (plot in log-scale).

Proof. Equivalently, we prove the injectivity of the linear map $f \in \mathcal{Y}_D^{\mathrm{ev}} \mapsto [f(\xi_i)]_{1\leq i\leq M} \in \mathbb{R}^M$. This result is an immediate consequence of Proposition 8.4. Indeed, if $f \in \mathcal{Y}_D^{\mathrm{ev}}$ is such that $f|_{\mathrm{CH}_N} = 0$, then $f \in \mathcal{Y}_D$ satisfie $f|_{\mathrm{CS}_N} = 0$, because the symmetry of f implies that $f(-\xi_i) = f(\xi_i) = 0$, for every $1 \leq i \leq M$. Since $N \leq 4$ and $D \leq 2N-1$, Proposition 8.4 implies that $f = 0$. □

Also, the following theorem proves that degrees larger than $2N-2$ must be proscribed in general; hence, the degree $D = 2N-2$ in Claim 9.15 is the largest "recommended" one.

Theorem 9.17. *For the grid $G = \mathrm{CH}_N$, and an even degree $D \geq 2N$, let A_N^D be the Vandermonde matrix* (9.4)*. Let $\sigma_{\min}(A_N^D)$, $\mathrm{cond}(A_N^D)$, denote its smallest singular value, resp. condition number.*

(i) For all $N \leq 4$ and $D \geq 2N$, the matrix A_N^D has not full column rank (hence, $\sigma_{\min}(A_N^D) = 0$, and $\mathrm{cond}(A_N^D) = +\infty$).

(ii) There exists a sequence $(\epsilon_N)_{N\geq 1}$ with asymptotics $\epsilon_N \underset{N\to+\infty}{\sim} \frac{N}{2}\left(\frac{2N}{\pi}\right)^{3/2}\left(\frac{2}{3}\right)^{2N} \underset{N\to+\infty}{\to} 0$, such that

$$\forall N \geq 1,\ \forall D \geq 2N, \quad \sigma_{\min}(A_N^D)^2 \leq \epsilon_N.$$

(iii) There is a sequence $(K_N)_{N\geq 1}$ with asymptotics $K_N \underset{N\to+\infty}{\sim} \frac{1}{4}\left(\frac{\pi}{2N}\right)^{1/2}\left(\frac{3}{2}\right)^{2N+1} \underset{N\to+\infty}{\to} +\infty$, such that, for all $N \geq 1$ and $D \geq 2N$ such that A_N^D has full column rank,

$$\mathrm{cond}(A_N^D)^2 \geq K_N.$$

Proof. We refer to the proofs of Proposition 8.4 and Theorem 8.7 which establish similar results for the matrix

$$[Y_n^m(\xi)]_{\substack{\xi\in\mathrm{CS}_N \\ |m|\leq n\leq D}} \in \mathbb{R}^{(6N^2+2)\times(D+1)^2}.$$

The proofs are based on the examples $f_N = Y_{2N}^{-2N}(x(\theta, \phi - \frac{\pi}{4})) \in \mathcal{Y}_{2N}$, and $Y_0 \in \mathcal{Y}_{2N}$. The same strategy applies for A_N^D, since $f_N, Y_0 \in \mathcal{Y}_{2N}^{\mathrm{ev}}$, so we get almost the same estimations. The slight difference is a factor $\frac{1}{2}$ in (ii), due to the restriction of CS_N to CH_N. This factor disappears in the estimation of the condition number in (iii), since it is a ratio. □

Remark 9.18. The critical degree $2N$ corresponds usually to oscillations at the Nyquist's frequency for a uniform one-dimensional grid with step $\frac{\pi}{2N}$. Here, the critical example f_N oscillates at this frequency along the (equatorial) grid $\phi \equiv \frac{\pi}{4}\,(\frac{\pi}{2N})$, so that f_N is undersampled along the equator.

In the sequel, if the grid is $G = \mathrm{CH}_N$, then we select the degree $D = 2N - 2$, as a consequence of Claim 9.15 and Theorem 9.17.

9.4.3 Discrete Funk transform

Assuming that Claim 9.15 is true, we consider the discrete Funk transform F, in the case $G = \mathrm{CH}_N$, $D = 2N - 2$, where $N \geq 1$ is fixed. Of course, any result of Subsection 9.3.2 applies. In particular, Theorem 9.12 guarantees estimations of stability based on the condition number plotted in Figure 9.2. We check this point in Figure 9.3, where we observe that the maximum singular values satisfy

$$\sigma_{\max}(\mathbf{F}) \approx 1.00218, \quad \sigma_{\max}(\mathbf{F}^\dagger) \approx \sqrt{2(2N-2)}, \qquad 1 \leq N \leq 32;$$

this is in agreement with the theoretical bounds (9.16).

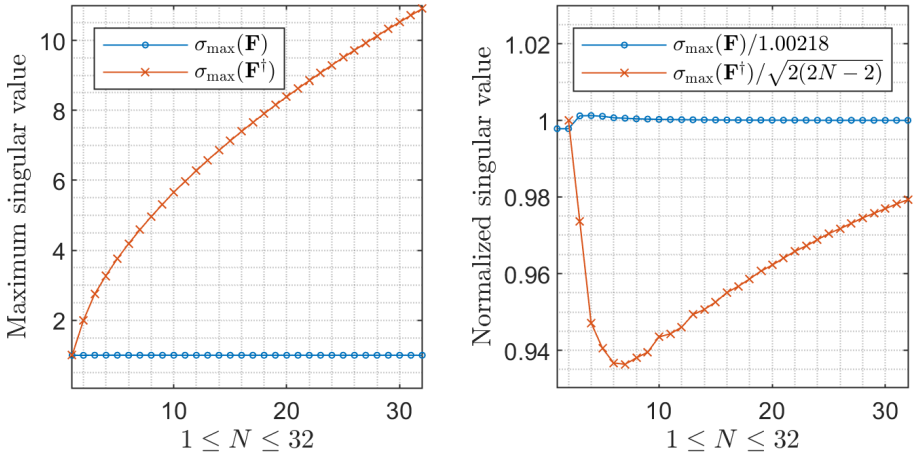


Figure 9.3: Stability constants associated to the discrete transforms F and $\mathrm{F}^\dagger$, in the case $G = \mathrm{CH}_N$, $D = 2N - 2$. Left: the maximum singular values $\sigma_{\max}(\mathbf{F})$ and $\sigma_{\max}(\mathbf{F}^\dagger)$ are plotted in term of N. Right: the same singular values are plotted, but with "normalization" factors.

9.5 Numerical results

We perform various numerical experiments, in order to assess the quality and the efficiency of the discrete Funk transform on the Cubed Hemisphere.

9.5.1 Accuracy and convergence of the discrete Funk transform

We evaluate the accuracy of the discrete Funk transform when it is used to approximate Funk transforms from values on CH_N.

For that purpose, we introduce test functions,

$$g^{(k)} := \sum_{\substack{0 \leq n \leq 100 \\ n \equiv 0\,(2)}} \sum_{-n \leq m \leq n} \hat{g}_n^{(k)} (2 + 0.5\cos(m) + 0.25\sin(m))\, Y_n^m \in \mathcal{Y}_{100}^{\mathrm{ev}}, \tag{9.18}$$

$$\hat{g}_n^{(-\infty)} := \tfrac{1}{n!}, \quad \hat{g}_n^{(k)} := (n+1)^k,\ k = -6, -4, -2, -1, 0, \tag{9.19}$$

where the various rates of decay of the spectral coefficients encode various "smoothness" properties. Here, $g^{(k)} \in \mathcal{Y}_{100}^{\mathrm{ev}}$, so we compute the Funk transform $\mathcal{F}g^{(k)}$ by Theorem 9.7, with $G = \mathrm{CH}_N$, $N = 51$ and $D = 2N - 2$. For any $\alpha \in \mathbb{S}^2$, we use the relation (9.10) to compute $\mathcal{F}g^{(k)}(\alpha)$ from the values of $g^{(k)}$ on the grid CH_{51}. This computation is exact, up to rounding errors.

Consider now the discrete Funk transform F, associated to the grid $G = \mathrm{CH}_N = \{\xi_i, 1 \leq i \leq M\}$ and the degree $D = 2N - 2$. For any function g, we approximate the vector $[(\mathcal{F}g)(\xi_i)]_{1\leq i\leq M}$ by $\mathbf{F}[g(\xi_i)]_{1\leq i\leq M}$ with a relative error $\eta_N[g]$ defined by

$$\eta_N[g] := \frac{\|\mathbf{F}[g(\xi_i)]_{1\leq i\leq M} - [(\mathcal{F}g)(\xi_i)]_{1\leq i\leq M}\|}{\|[(\mathcal{F}g)(\xi_i)]_{1\leq i\leq M}\|}; \tag{9.20}$$

here, $\|\cdot\|$ denotes the euclidean norm in $\mathbb{R}^M$. For $g = g^{(k)}$, the reference vector $[(\mathcal{F}g^{(k)})(\xi_i)]_{1\leq i\leq M}$ is computed as mentioned in the previous paragraph (relation (9.10) with CH_{51} for the data grid, and $\alpha \in \mathrm{CH}_N$ for the evaluation). In particular, $\eta_{51}[g^{(k)}]$ is zero (up to rounding errors).

We have plotted the errors $\eta_N[g^{(k)}]$ in Figure 9.4 (left panel), for $1 \leq N \leq 32$. Overall, the behavior of the observed error depends on the rate of decay of the spectral coefficients; the error converges fastly to zero for rapidly decaying coefficients. We quantify this phenomenon in Table 9.1, where we report numerical convergence rates $r_N[g]$ such that

$$\eta_{2N}[g] = \eta_N[g]\, 2^{-r_N[g]}, \quad \text{with} \quad r_N[g] = \log_2 \eta_N[g] - \log_2 \eta_{2N}[g]. \tag{9.21}$$

N	$r_N[g^{(-\infty)}]$	$r_N[g^{(-6)}]$	$r_N[g^{(-4)}]$	$r_N[g^{(-2)}]$	$r_N[g^{(-1)}]$	$r_N[g^{(0)}]$
1	3.7	4.5	2.9	0.28	-1.4	1.8
2	11	4.9	3.1	1.3	0.3	-0.19
4	29	5.4	3.4	1.5	0.78	-0.11
8	5.3	5.4	3.4	1.8	1.2	0.093
16	0.83	5.5	3.6	2.2	1.8	1

Table 9.1: Convergence rates (9.21) of the errors (9.20), for the test functions from (9.18-9.19).

For $g^{(-\infty)}$, with a factorial decay of the coefficients, $\hat{g}_n^{(-\infty)} = 1/n!$, the very fast convergence appears as a blow up of the rate $r_N[g^{(-\infty)}]$. For the functions $g^{(-k)}$, $k = 6, 4, 2$, with a decay $\hat{g}_n^{(-k)} = 1/(n+1)^k$, the rate looks like $r_N[g^{(-k)}] \approx k - 0.5$. For $g^{(-1)}$, with the slow decay $\hat{g}_n^{(-1)} = 1/(n+1)$, and $g^{(0)}$ with constant values $\hat{g}_n^{(0)} = 1$, the convergence analysis is not so clear.

In a word, the discrete Funk transform F (or its matrix $\mathbf{F}$) approximates the Funk transform from values on a Cubed Hemisphere. It converges fastly for smooth functions, for which the spectral coefficients decay rapidly to zero. The observed rates of convergence suggests to analyze theoretically the speed of convergence in Sobolev spaces. We defer this point to further studies.

9.5.2 Funk transform of Gaussian models

We study the accuracy of the discrete Funk transform on Gaussian models from dMRI, for the grid $G = \mathrm{CH}_N$ and the degree $D = 2N - 2$.

We consider Gaussian models in the following form,

$$S(x) = \exp(-b\, x^{\intercal}\mathbf{D}x), \quad x \in \mathbb{S}^2, \tag{9.22}$$

where $b \geq 0$, and $\mathbf{D} \in \mathbb{R}^{3\times 3}$ is a symmetric positive definite matrix. Such models describe the dMRI signal S in Diffusion Tensor Imaging. The so-called *diffusion tensor* $\mathbf{D}$ models intrinsic diffusion properties of biological tissues. The parameter b is the so-called *b-value*, and is a parameter of the acquisition. The unit vector x, represents a *gradient direction*, and browses a hemispherical grid during the acquisition. Gaussian models such as (9.22) appear also in High Angular Resolution Diffusion Imaging [124]. In this field, a weighted average of several Gaussian models can be introduced to model the signal from crossing fibers. The orientation of the fibers can be imaged using an Orientation Distribution Function, which is computed approximately as the Funk transform of the

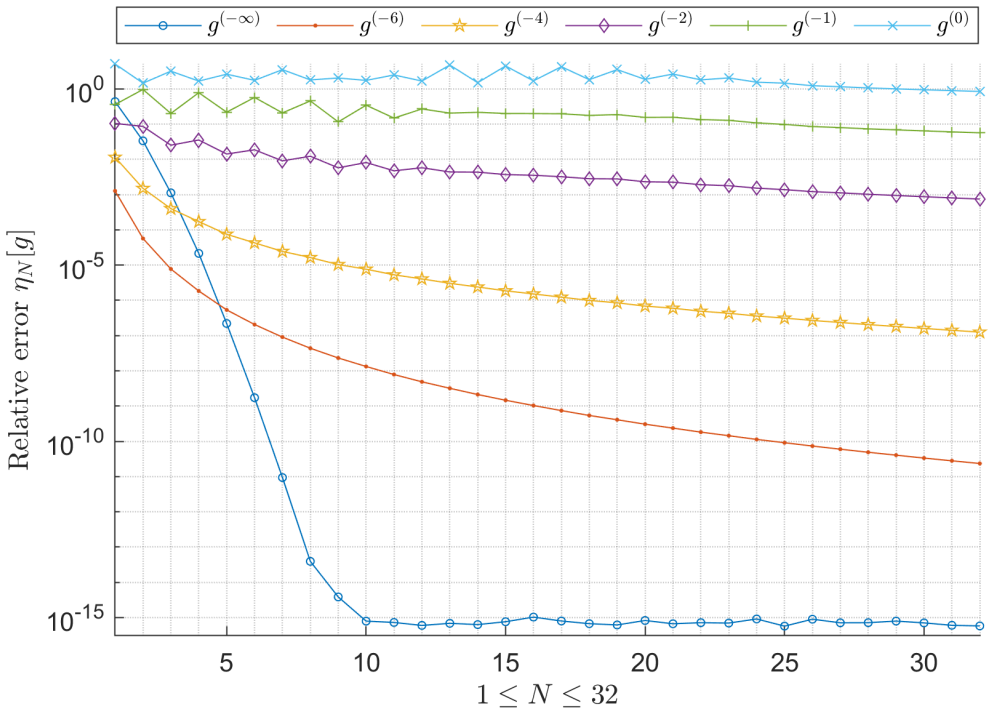


Figure 9.4: Approximation of the Funk transform $[(\mathcal{F}g)(\xi_i)]_{1\leq i\leq M}$ by the discrete transform $\mathbf{F}[g(\xi_i)]_{1\leq i\leq M}$, for the grid CH_N, and the degree $2N-2$. The relative error $\eta_N[g]$ in (9.20) is plotted, for each test function $g = g^{(k)}$ from (9.18-9.19).

recorded signal S [125,178]. Hence, it is crucial to be able to compute accurately the Funk transform of a Gaussian model, from a discrete set of values.

In this paper, we consider Gaussian signals

$$S_j(x) = \exp(-b_j\, x^{\intercal}\mathbf{D}_j x), \quad x \in \mathbb{S}^2, \quad 1 \leq j \leq 6. \tag{9.23}$$

The b-values b_j and the diffusion tensors $\mathbf{D}_j$ are defined in Table 9.2. Our values are inspired by the values from [125]. The b-value $b = 1000$ [s/mm^2] is an usual clinical value, whereas $b = 3000$ [s/mm^2] is considered as relatively high. For the diffusion tensors, we have chosen diagonal matrices $\mathbf{D}_i$, defined by the eigenvalues μ_1, μ_2, $\mu_3 > 0$. The matrix $\mathbf{D}_3$ has been found in the synthetic data generation in [125]. The other matrices have been defined as "variations" of this matrix, in order to obtain more or less anisotropy; see the last column of Table 9.2, where anisotropy is measured by means of the *fractional anisotropy* (FA),

$$\mathrm{FA} = \frac{1}{\sqrt{2}}\sqrt{\frac{(\mu_1-\mu_2)^2+(\mu_1-\mu_3)^2+(\mu_2-\mu_3)^2}{\mu_1^2+\mu_2^2+\mu_3^2}} \in [0,1]. \tag{9.24}$$

Firstly, we assume that the (even) signal S_j is recorded on $G = \mathrm{CH}_N$, with $N \geq 1$ and $1 \leq j \leq 6$. We compute reference values $[(\mathcal{F}S_j)(\xi_i)]_{1\leq i\leq M}$ using trapezoidal rules[3]. Then, we compute the discrete Funk transform $\mathbf{F}[S_j(\xi_i)]_{1\leq i\leq M}$. It approximates $[(\mathcal{F}S_j)(\xi_i)]_{1\leq i\leq M}$, with a relative error $\eta_N[S_j]$, where η_N is defined in (9.20). We have plotted these errors in Figure 9.5 (left panel) to evaluate the accuracy of the procedure. Overall, a fast convergence due to the smoothness of the Gaussian signals is observed. For the isotropic functions S_1 and S_4, the error is always zero (up

[3]Here, an integral along a great circle $x \cdot \alpha = 0$ is an integral of a smooth 2π-periodic function over a period, so the trapezoidal rule converges exponentially to the true integral [177]. Therefore, we apply successive trapezoidal rules as follows. We start with an angular step $\frac{\pi}{8}$. We evaluate the associated trapezoidal rule; then, we divide the angular step by two, and we iterate. The iterations are stopped as soon as the relative increase of the value between two successive iterations is below the tolerance 10^{-13}.

j	b_j [s/mm^2]	$\mathbf{D}_j$ [mm^2/s]	FA
1	1000	$10^{-6}\,\mathrm{diag}(300, 300, 300)$	0
2	1000	$10^{-6}\,\mathrm{diag}(300, 600, 900)$	0.46
3	1000	$10^{-6}\,\mathrm{diag}(300, 300, 1700)$	0.80
4	3000	$10^{-6}\,\mathrm{diag}(300, 300, 300)$	0
5	3000	$10^{-6}\,\mathrm{diag}(300, 600, 900)$	0.46
6	3000	$10^{-6}\,\mathrm{diag}(300, 300, 1700)$	0.80

Table 9.2: Parameters of the Gaussian signals (9.23). The anisotropy is measured by the fractional anisotropy (FA), defined in (9.24).

to rounding errors), because S_1, $S_4 \in \mathcal{Y}_0 \subset \mathcal{Y}^{\mathrm{ev}}_{2N-2}$, so (9.9) applies. With S_2, S_5, and S_3, S_6, we observe that increasing the b-value induces a loss in accuracy; this is because a Gaussian becomes sharper with high b-values.

Secondly, we show that the orientation of the grid does not matter. For that purpose, we consider "rotations" of the signals S_j:

$$S_j(Q^\intercal \cdot) : x \mapsto S_j(Q^\intercal x) = \exp(-b\, x^\intercal\, Q\mathbf{D}_j Q^\intercal\, x),$$

where $Q \in \mathbb{R}^{3\times 3}$ is a random orthogonal matrix. The relative error of approximation of the Funk transform becomes $\eta_N[S_j(Q^\intercal \cdot)]$, and can be computed as before. For each function S_j, we repeat this procedure for 30 random orthogonal matrices Q, and we plot the maximum error

$$\max_Q \eta_N[S_j(Q^\intercal \cdot)] \tag{9.25}$$

in Figure 9.5 (right panel). We obtain a similar conclusion than before, so that the conclusion does not depend on the orientation of the grid.

Thirdly, we investigate the effect of noise. We corrupt the signals as follows. We fix a value of N. For any $1 \le j \le 6$, for any $\sigma = 2^{-p}$, with $2 \le p \le 31$, we corrupt S_j, by a "speckle" noise and an additive noise with level σ:

$$S_j^\sigma(\xi_i) = |S_j(\xi_i)(1 + \sigma u_i) + \sigma v_i|, \quad 1 \le i \le M,$$

where the u_i, v_i, are $2M = 6N^2 + 2$ independent realizations of the normal law $\mathcal{N}(0, 1)$. In this case, the relative error on the signal is given by

$$\frac{\|[S_j^\sigma(\xi_i) - S_j(\xi_i)]_{1\le i\le M}\|}{\|[S_j(\xi_i)]_{1\le i\le M}\|}. \tag{9.26}$$

We compute the discrete Funk transform $\mathbf{F}[S_j^\sigma(\xi_i)]_{1\le i\le M}$, which approximates $[(\mathcal{F}S_j)(\xi_i)]_{1\le i\le M}$ with a relative error

$$\frac{\|\mathbf{F}[S_j^\sigma(\xi_i)]_{1\le i\le M} - [(\mathcal{F}S_j)(\xi_i)]_{1\le i\le M}\|}{\|[(\mathcal{F}S_j)(\xi_i)]_{1\le i\le M}\|}. \tag{9.27}$$

In Figure 9.6, we have plotted the relative error (9.27) on the transform against the relative error (9.26) on the signal. Two values of N are considered. On the left, $N = 5$, so that the grid CH_N contains 76 points, and the approximation space is $\mathcal{Y}^{\mathrm{ev}}_8$. On the right, $N = 10$, so that the grid CH_N contains 301 points, and the approximation space is $\mathcal{Y}^{\mathrm{ev}}_{18}$. Roughly speaking, we observe that the relative error on the transform is the maximum between the relative error on the signal, and the relative error on the transform from the noise-free case (displayed in Figure 9.5). This result is in agreement with the stability constant of the transform $\mathbf{F}$, $\sigma_{\max}(\mathbf{F}) \approx 1$ in Figure 9.3.

To conclude, the Funk transform of Gaussian models can be accurately evaluated by the discrete transform on the Cubed Hemisphere, and in a very stable way.

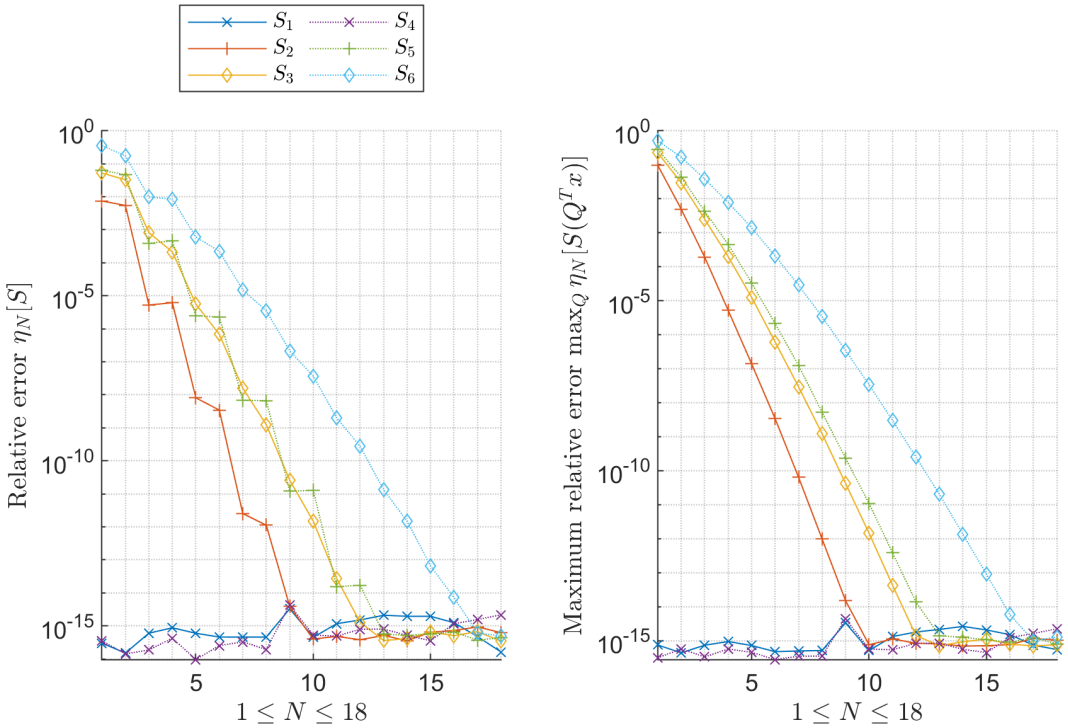


Figure 9.5: Accuracy of the discrete Funk transform on CH_N with degree $2N-2$, for the Gaussian signals S_j in (9.23) and Table 9.5. Left: $\mathbf{F}[S_j(\xi_i)]_{1\le i\le M}$ approximates $[(\mathcal{F}S_j)(\xi_i)]_{1\le i\le M}$ with relative error $\eta_N[S_j]$ in (9.20); we plot $\eta_N[S_j]$ with $1 \le N \le 32$. Right: for any orthogonal matrix $Q \in \mathbb{R}^{3\times 3}$, the same procedure applied to the "rotated" Gaussian $S_j(Q^\intercal \cdot)$ results in a relative error $\eta_N[S_j(Q^\intercal \cdot)]$; we plot the maximum error (9.25), where Q scans a set of 30 random orthogonal matrices.

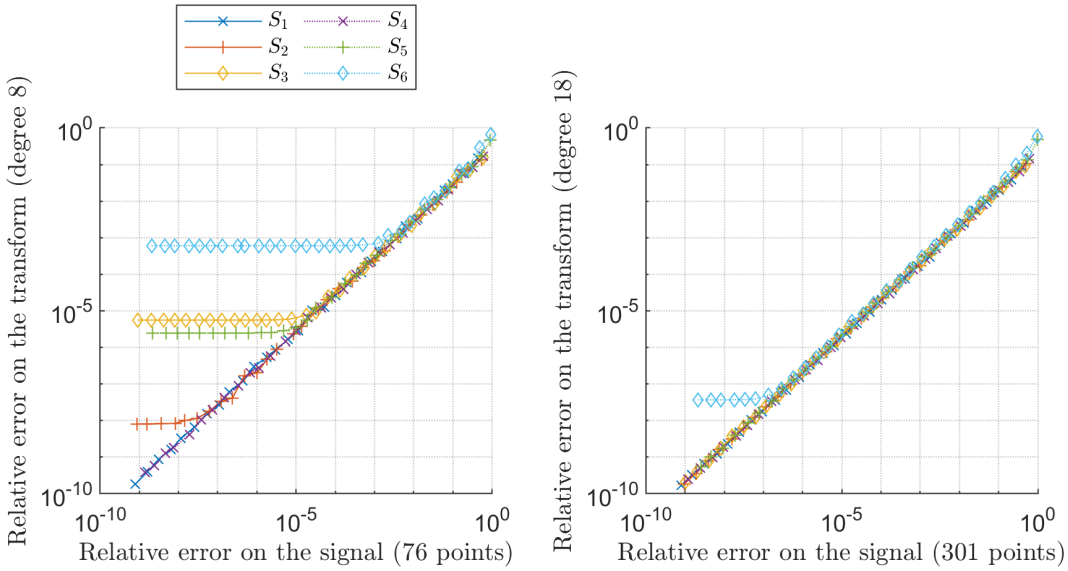


Figure 9.6: Accuracy of the discrete Funk transform on CH_N (degree $2N-2$), for Gaussian signals S_j corrupted by noise. The relative error (9.27) on the transform is plotted against the relative error (9.27) on the signal (logarithmic scale). Left: $N = 5$; right: $N = 10$.

9.5.3 Comparison of the Cubed Sphere and the icosahedral grid

We compare discrete Funk transforms on Cubed Hemispheres with discrete Funk transforms on icosahedral grids.

For the Cubed Sphere, we consider the Cubed Hemisphere CH_N with cardinal number $M = 3N^2 + 1$. As an alternative grid, we consider an *icosahedral grid*. It is based on a regular triangular lattice onto each face of an icosahedron inscribed in $\mathbb{S}^2$. The icosahedral grid is defined as the projection of the vertices of this lattice onto $\mathbb{S}^2$. We further halve the grid by symmetry consideration (Proposition 9.3), in the same way as CS_N has been halved. Assuming that each edge of the original icosahedron has been divided into N parts, the resulting half-grid contains $M = 5N^2 + 1$ points; we still call this grid an *icosahedral grid*, and we denote it by Ico_N. Such grids have already been used for computing Funk transforms in [125] with $M = 81, 321$ (which correspond to the parameters $N = 2, 8$).

Firstly, in order to obtain a stable discrete Funk transform, the degree D must be carefully tuned. For CH_N, we use the rule $D = 2N - 2$, as it has been introduced in Section 9.4. For the icosahedral grid Ico_N, we do not know such a rule on the degree. To overcome this disadvantage, we compute numerically D as the largest degree D such that $\operatorname{cond} A \leq 2$, where A denotes the Vandermonde matrix (9.4) with $G = \mathrm{Ico}_N$. In Figure 9.7, we plot the obtained degree D against the number of points of the grid, for the grid CH_N with $1 \leq N \leq 32$, and the grid Ico_N with $1 \leq N \leq 25$. We observe that for equivalent number of grid points, the degree associated to the icosahedral grid is larger than the degree associated to the Cubed Hemisphere. Therefore, the icosahedral grid permits to work in a larger approximation space $\mathcal{Y}_D^{\mathrm{ev}}$, while keeping a very small condition number ($\operatorname{cond} A \leq 2$).

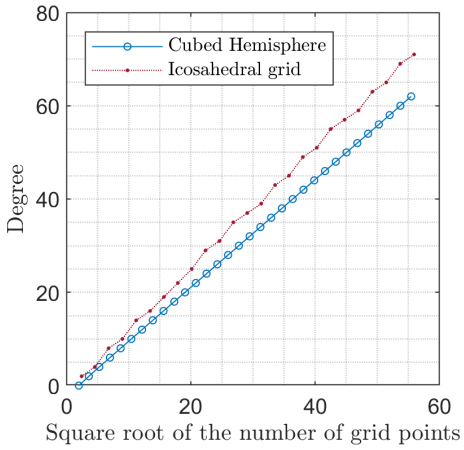


Figure 9.7: Degree D for least squares fitting on the Cubed Hemisphere CH_N, resp. the Icosahedral grid Ico_N. For CH_N, the number of grid points is $M = 3N^2 + 1$, the degree is $D = 2N - 2$, and $1 \leq N \leq 32$. For Ico_N, $M = 5N^2 + 1$, D is the largest degree such that $\operatorname{cond} A \leq 2$, and $1 \leq N \leq 25$.

Secondly, we consider successively the discrete Funk transform associated to the grids

$$G = \mathrm{CH}_N,\ 1 \leq N \leq 32, \quad G = \mathrm{Ico}_N,\ 1 \leq N \leq 25,$$

with the degree D discussed above. We evaluate the accuracy on the test function $g = g^{(k)}$ defined in (9.18-9.19), for $k = -\infty, -6, -4, -2$, by means of the relative error

$$\eta[g] = \max_Q \frac{\|\mathbf{F}[g(Q^\intercal \xi_i)]_{1\leq i\leq M} - [(\mathcal{F}g)(Q^\intercal \xi_i)]_{1\leq i\leq M}\|}{\|[(\mathcal{F}g)(Q^\intercal \xi_i)]_{1\leq i\leq M}\|}, \quad (G = \{\xi_i, 1 \leq i \leq M\}); \tag{9.28}$$

here the "orientation" Q browses a set of 30 random orthogonal matrices. The computed errors are displayed in Figure 9.8. Overall, the two grids define transforms with similar accuracy, and similar

properties of convergence. This test reveals also that for very smooth functions and very small grids, the isosahedral grid defines a more accurate transform.

Lastly, we repeat the same procedure, but we corrupt the data with a level of noise $\sigma = 10^{-6}$. We compute the relative error

$$\eta_{\text{noise}}[g] = \max_{(Q,u,v)} \frac{\|\mathbf{F}[g(Q^\intercal \xi_i)(1+\sigma u_i) + \sigma v_i]_{1\leq i\leq M} - [(\mathcal{F}g)(Q^\intercal \xi_i)]_{1\leq i\leq M}\|}{\|[(\mathcal{F}g)(Q^\intercal \xi_i)]_{1\leq i\leq M}\|}, \quad \sigma = 10^{-6}, \tag{9.29}$$

where the maximum is taken over 30 experiments; each experiment fixes Q as a random orthogonal matrix, and the u_i, v_i as $2M$ independent realizations of the normal law $\mathcal{N}(0,1)$. The obtained errors are depicted in Figure 9.9. The observations of the noise-free case still apply. We further observe that the errors are almost the same as soon as the noise becomes dominant.

To conclude, considering the Cubed Sphere is simpler than considering an icosahedral grid, for which further studies (or computation) are needed to keep the conditioning under control. Moreover, the resulting accuracy is almost the same, except for very smooth functions on very small grids; in this case, the icosahedral grid is more advantageous if the noise is small enough.

9.5.4 Computation time

As a further indicator of efficiency, we measure the computation time of discrete Funk transforms, for the grid $G = \text{CH}_N$, and the degree $D = 2N - 2$. For each value of N, we fix a random vector $\boldsymbol{b}$, and we measure the time dedicated to the assembly of the matrix A and the computation of $\mathbf{F}\boldsymbol{b}$. The code is written in `Matlab`, and $\ell[b]$ is computed with a simple command such as `(A'*A)\(A'*b)`. The program is executed on a laptop Dell Precision 7540; the processor is an Intel i9-9880H@2.30 GHz. The experiment is repeated six times, and we report the average values of the running times in Table 9.3.

Parameter N	1	2	4	8	16	32	64
Number of grid points M	4	13	49	193	769	3073	12289
Degree D	0	2	6	14	30	62	126
CPU time (s)	1.8e-04	1.2e-04	3.7e-04	2.3e-03	2.2e-02	5.0e-01	1.5e+01

Table 9.3: Running time dedicated to the computation of a discrete Funk transform $\mathbf{F}\boldsymbol{b}$, for the grid $G = \text{CH}_N$, and the degree $D = 2N - 2$.

These preliminary results show that our transform is computed relatively fastly for small grids, despite our "brute force" implementation has not been optimized. Further studies are required to decrease these times. Combining symmetry consideration and an iterative solver such as Conjugate Gradient Least Squares (CGLS) is an option to consider in the future.

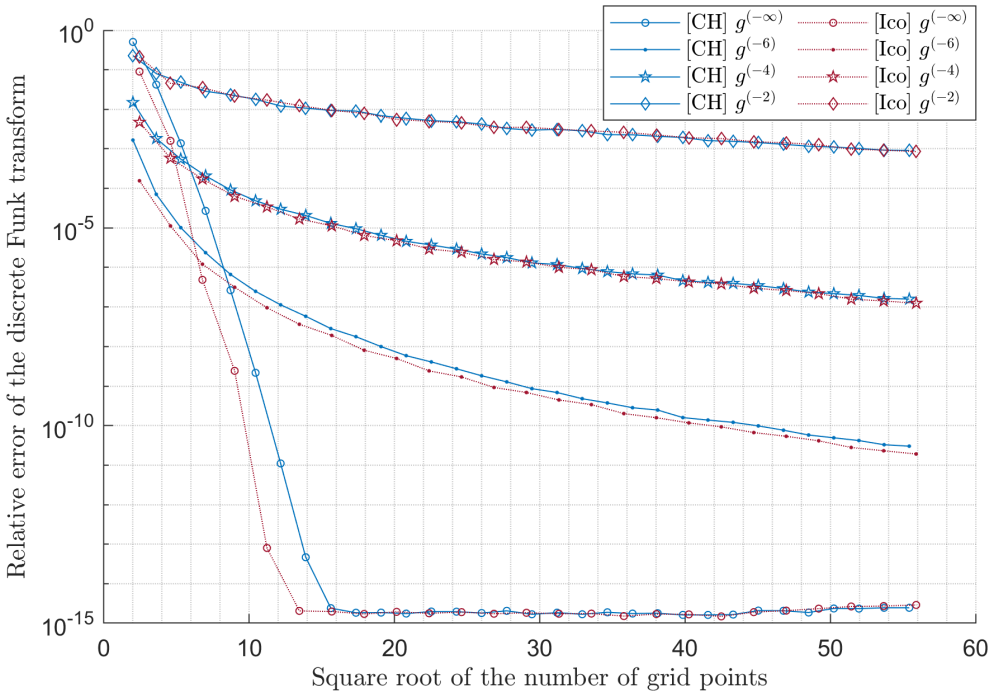


Figure 9.8: Accuracy of the Funk transform associated to the Cubed Hemisphere (CH), resp. the Icosahedral grid (Ico). The relative error $\eta[g]$ in (9.28) is plotted against $\sqrt{M}$, with M the number of grid points. The degree D is plotted in Figure 9.7.

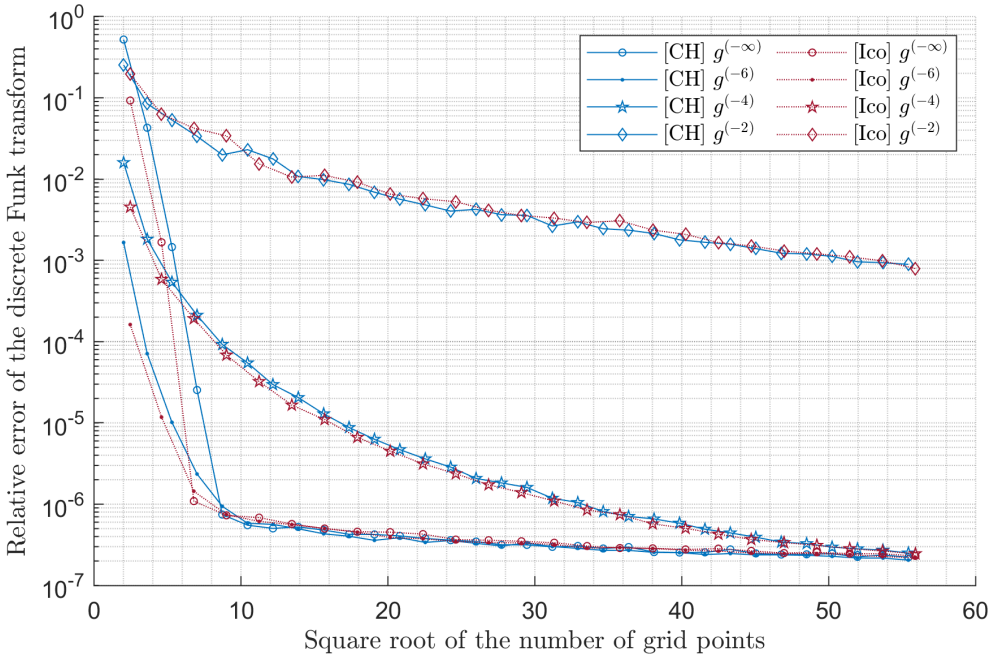


Figure 9.9: Accuracy of the Funk transform associated to the Cubed Hemisphere (CH), resp. the Icosahedral grid (Ico), with a level of noise $\sigma = 10^{-6}$. The relative error $\eta_{\text{noise}}[g]$ in (9.29) is plotted against $\sqrt{M}$, with M the number of grid points. The degree D is plotted in Figure 9.7.

9.6 Conclusion and perspectives

This chapter deals with mathematical and numerical properties of some discrete Funk transforms, including their (pseudo)inversion. As a special case, the study includes a simple framework based on the Cubed Sphere. Our theoretical and numerical results indicate that stability and suitable convergence properties are expected in this context, despite regularization has not been applied. This mathematical background about discrete Funk transforms could potentially have applications in any field where integrals along great circles on a sphere are considered.

This work opens problems to be addressed in the future. Finding the "best" spherical grid and the "best" degree is an open question. For the case of the Cubed Hemisphere CH_N, proving that our rule on the degree ($D = 2N - 2$) results in a small condition number is still open. Another point concerns the speed of convergence, which should be quantified, for instance in Sobolev spaces. Concerning implementation aspects, writing a "fast" algorithm has still to be done. A first step in this direction could be an efficient solver for the least squares problem, taking into account further symmetry consideration. To finish with, comparing our transform with time-tested transforms on real experiments is a goal for further studies; in particular, testing the Cubed Hemisphere in Q-ball imaging may be instructive.

Chapter 10

Conclusion and perspectives

10.1 Sampling and spectral computing on the Cubed Sphere

10.1.1 Metric properties of the Cubed Sphere

Geometrical and metric properties, as defined in [134, 137], are usually introduced to quantify the quality of a mesh. For the Cubed Sphere, we can mention the mesh norm and the separation distance, defined by

$$h_N = \sup_{y\in\mathbb{S}^2} \min_{x\in \mathrm{CS}_N} \arccos y\cdot x, \quad \delta_N = \min_{y\neq x\in \mathrm{CS}_N} \arccos y\cdot x.$$

Our results permit to give an analytical expression of the separation distance δ_N; it implies the asymptotics $\delta_N \sim \frac{\pi}{2\sqrt{2}N}$.

Concerning the mesh norm, a preliminary numerical study has revealed that it is achieved at the center of a panel, for

$$\begin{cases} y = (1,0,0),\ x = \rho(1, \pm\tan\frac{\pi}{4N}, \pm\tan\frac{\pi}{4N}), & \text{if } N \text{ is odd,} \\ \begin{cases} y = \rho(t, (1+2t^2)^{1/2} - (1+t^2)^{1/2}, -1 + (1+t^2)^{1/2}), \\ x \in \{(1,0,0), \rho(1,0,t), \rho(1,t,t)\}, \\ \text{with } t := \tan\frac{\pi}{2N}, \end{cases} & \text{if } N \text{ is even,} \end{cases}$$

where ρ denotes the radial projection. To the author's knowledge, this result is new and has never been proved. Proving such a result, or more generally studying metric properties, is an important subject since it could provide a valuable background for the analysis of some approximation problems on the Cubed Sphere.

Another open problem related to areas deals with covering $\mathbb{S}^2$ by cells T_j, with areas $|T_j|$, $1 \leq j \leq \bar{N}$, such that each cell T_j contains exactly one point $x_j \in \mathrm{CS}_N$, and such that the spherical quadrature rule $\mathcal{Q}f = \sum_{j=1}^{\bar{N}} |T_j| f(x_j)$ is optimal, for some criterion to be defined.

10.1.2 Approximation and Vandermonde matrices on the Cubed Sphere

Data approximation on the Cubed Sphere CS_N by a spherical harmonic can be formulated as the least squares problem (LS). From a matrix point of view, the problem especially depends on the Vandermonde matrice A_N^D defined in (8.6),

$$A_N^D = [Y_n^m(x)]_{\substack{x\in \mathrm{CS}_N \\ |m|\leq n\leq D}}.$$

Taking benefit from the geometrical structure of the Cubed Sphere, based on great circles, we have obtained some theoretical results about this matrix and the fitting problem (LS):

(i) for $N \leq 4$, A_N^D has full column rank if, and only if, $D \leq 2N-1$, so (LS) has a unique solution only if $D \leq 2N-1$;

(ii) for $N \geq 5$, and $D = N+2$, A_N^D has full column rank, so (LS) has a unique solution;

(iii) for $N \to \infty$, and $D = 2N$, $\operatorname{cond} A_N^D \to \infty$, so (LS) becomes ill-conditioned if the degree corresponds to the Shannon-Nyquist's frequency along the equator;

(iv) for $D = 4N-1$, or $D = 4N-2$ if N is even, A_N^D has full row rank, so every solution of (LS) is an interpolating function.

Moreover, numerical results show that:

(i) for $D = 2N-1$, A_N^D has full column rank with a small condition number, which implies that (LS) is well-posed and well-conditioned if the Shannon-Nyquist's condition is strictly respected along the equator;

(ii) for $D = 3N$, A_N^D has full row rank, which implies the existence of interpolating functions.

Proving these last two results from a theoretical point of view is still open; there is indeed some gap between the numerical observations and the known theoretical results.

10.1.3 Special functions dedicated to the Cubed Sphere

We have considered the following space, dedicated to Lagrange interpolation by a spherical harmonic on the Cubed Sphere,

$$\mathcal{U}_N = \oplus_{n\geq 0}\mathcal{W}_n^{\perp}, \quad \mathcal{W}_n := \{f \in \mathbb{Y}_n : \exists g \in \mathcal{Y}_{n-1},\, f|_{\mathrm{CS}_N} = g|_{\mathrm{CS}_N}\},\, n \geq 1, \quad \mathcal{W}_0 := \{0\}.$$

Interpolating in this space is analogous to the usual trigonometric interpolation on $[0, 2\pi]$. But here, the practical computation of such a space is based on numerical linear algebra, and is biased by truncation of singular values of some matrices.

To go further, we would like to find special functions on the sphere in order to describe analytically this space, or a variation of it. This would be an improvement towards a suitable sampling theory on the Cubed Sphere, and new approximation algorithms may emerge.

10.1.4 Quadrature rules on the Cubed Sphere

Various quadrature rules on the Cubed Sphere, such as

$$\int_{\mathbb{S}^2} f(x)\,\mathrm{d}\sigma \approx \sum_{x\in\mathrm{CS}_N} \omega(x) f(x), \quad \omega : \mathrm{CS}_N \to \mathbb{R}, \quad f : \mathbb{S}^2 \to \mathbb{R},$$

have already been introduced; in particular, an octahedral rule which is almost as accurate as optimal octahedral rules has been deduced from interpolation. There are still directions to be explored, for instance to get accurate quadrature rules that can be computed fastly. Moreover, studying quadrature errors is related to the study of the singular values of the Vandermonde matrices A_N^D, which is a further motivation to deepen the subject.

Several works are in progress. As an application of least square fitting by a spherical harmonic, we can define a new quadrature rule by integration of our least square approximation. Preliminary numerical results indicate that this approach is an interesting option to be strengthened.

Another method, introduced in [156], can be summarized by

$$\omega(x) = \delta^2 \cdot \frac{(1+\tan^2\alpha)(1+\tan^2\beta)(1+\tan^2\gamma)}{2(\tan^2\alpha+\tan^2\beta+\tan^2\gamma)^{3/2}}, \text{ with } x = \frac{1}{(\tan^2\alpha+\tan^2\beta+\tan^2\gamma)^{1/2}}(\tan\alpha, \tan\beta, \tan\gamma); \quad (10.1)$$

here, x browses CS_N when (α, β, γ) browses a uniform grid on the faces of the cube $[-\frac{\pi}{4}, \frac{\pi}{4}]^3$, with angular step $\delta = \frac{\pi}{2N}$. This rule can be understood as a bivariate trapezoidal rule on the faces of $[-\frac{\pi}{4}, \frac{\pi}{4}]^3$, with a correction on the eight corners in order to get a "uniform formula"; the expression of the weight is especially due to a change of variable from the sphere to the faces of the cube. Remarkably, the simple correction on the corners permits to reach a fourth-order numerical accuracy with respect to δ [156]. But proving theoretically this result is still open.

Moreover, I have found a further correction which is sixth-order accurate in numerical experiments. This rule is the octahedral rule given by (10.1), except for the corners, where the weight is defined to exactly integrate the constant function $x \in \mathbb{S}^2 \mapsto 1$, *i.e.*

$$\omega(3^{-1/2}(\pm 1, \pm 1, \pm 1)) = \frac{1}{8}\left[4\pi - \textstyle\sum_{x \in \mathrm{CS}_N \setminus \{3^{-1/2}(\pm 1, \pm 1, \pm 1)\}} \omega(x)\right].$$

The resulting rule has never been published so far, and proving the sixth-order accuracy is still open.

10.1.5 Fast algorithms on the Cubed Sphere

There is still a need of data approximation based on a fast algorithm that is analogous to the Fast Fourier Transform, in the specific framework of the Cubed Sphere. The inclusion

$$\mathrm{CS}_N \subset \mathrm{CS}_{2N}, \quad N \equiv 0\,(2),$$

may be trick to achieve such a goal.

In the absence of such an algorithm and in a first step, one may develop an efficient solver in a framework of (weighted) least squares fitting (WLS). An option consists in solving the set of normal equations by an iterative solver, such as the so-called CGLS (Conjugate Gradient Least Squares). The problem can be solved within a few iterations, due to the very small condition number. Also, this solver can be accelerated taking benefit from symmetry consideration, since it implies a block diagonal structure. Preliminary numerical results confirm the relevance of this approach.

A further option to accelerate the iterations of CGLS would consist in computing the products matrix-vector with fast spherical Fourier algorithms from [145, 148]. At the end, a few fast spherical Fourier transforms should result in an accurate least-squares fitting, which should be itself evaluated by means of a fast spherical Fourier transform. This strategy, which may require some tuning, has not been tested yet.

10.1.6 New representations of the octahedral group

The symmetry group of the Cubed Sphere is given by the octahedral group $\mathcal{G}$ in (5.2). This implies various invariance and orthogonality properties related to data approximation on CS_N. Among them, we can define octahedral quadrature rules on CS_N, for which 15/16 of the real spherical harmonics are automatically integrated. Also, the normal matrix associated to the (octahedral) weighted least squares (WLS) has the block diagonal structure displayed in Figure 8.3.

To go further, one can introduce the Cubed Sphere in a framework of representation theory. Here, the space of grid functions $\mathbb{R}^{\mathrm{CS}_N}$ is invariant under the octahedral group $\mathcal{G}$, so is the interpolation space $\mathcal{U}_N$. These spaces define two new (equivalent) representations of $\mathcal{G}$:

$$\mu(Q)(f) := f(Q^{\mathsf{T}}\cdot),\ f \in \mathbb{R}^{\mathrm{CS}_N}, \quad \tilde{\mu}(Q)(u) := u(Q^{\mathsf{T}}\cdot),\ u \in \mathcal{U}_N, \quad Q \in \mathcal{G}.$$

Our interpolation scheme preserves symmetry, as defined in [166, 167]; numerical linear algebra permits to compute some orthonormal basis of $\mathcal{U}_N$ which contains bases of the isotypic components, and whose evaluation on CS_N provides an orthogonal basis of $\mathbb{R}^{\mathrm{CS}_N}$ and its isotypic components. But defining analytically this basis, or a similar one, is still open. Lastly, another remark to be explored is that representation theory gives a framework to define Fourier-kind transforms [155].

10.1.7 Comparison with other grids

The presented works especially focus on the Cubed Sphere grid, but some approaches can be extended to other spherical grids. For instance, the same interpolation procedure, based on the echelon factorization of the Vandermonde matrix, still applies for other grids. For the discrete Funk transform based on least square fitting, our numerical results show that the Icosahedral grid slightly outperforms the Cubed Sphere in the case of very smooth functions and small grids. Therefore, it seems interesting to continue the comparison between the Cubed Sphere and the Icosahedral grid.

10.1.8 Spherical computation in image and vision

Spectral computation on the Cubed Sphere can be related to the first part of this thesis, and more generally to image and vision.

In medical imaging, and more specifically in diffusion Magnetic Resonance Imaging, *Q-ball imaging* consists in imaging the orientation of fibers in biological tissues by computation of some discrete Funk-Radon transform. We have tested the principle of such a transform, in the case of the Cubed Sphere. To go further, our transform should be tested on real data; such a work would be the very first study in medical imaging with the Cubed Sphere used for the acquisition grid.

In computer graphics, radiative transfer models, such as the rendering equation (B.7), contain angular parameters among the variables. This suggests sampling on spherical grids or using spherical harmonics, as in spherical harmonic lighting [163,174]. One may wonder if the Cubed Sphere could have an interest in this kind of framework, or more generally in any radiative models with scattering.

10.2 Special echelon factorization and lexicographical least-squares

10.2.1 Numerical analysis of the special echelon factorization

Our interpolating spherical harmonic on the Cubed Sphere is computed by means of the matrix factorization (6.19), applied on a suitable matrix. This factorization solves various optimization problems which enter into the framework of lexicographical least-squares, where a sequence of least-squares fitting are performed. This approach can be understood as some generalization of Moore-Penrose inversion based on the Singular Value Decomposition (SVD), and it is similar to some methods from robotics [126].

Therefore, there is some interest in studying such methods from a numerical analysis point of view. Studying the stability and the accuracy of our algorithm is still open. Also, writing an efficient algorithm that is similar to the iterative methods dedicated to the SVD is open. Ideally, we would like to find suitable matrices, $\mathbf{V}_j$, $\mathbf{U}_j$ and $\mathbf{E}_j$, $j \geq 0$, such that (6.19) is obtained at the limit $j \to \infty$, in a fast stable accurate way.

10.2.2 Application of lexicographical least-squares in learning

In this work, lexicographical optimization has been introduced to build an interpolation space on the Cubed Sphere, based on spherical harmonics ordered by increasing degree. Similar approaches can be introduced for other interpolation problems, such as multivariate polynomial interpolation in $\mathbb{R}^d$. Also, the usual trigonometric interpolation can be understood as an implementation of our approach, considering one block per (increasing) frequency.

The proposed formalism relies especially on linear algebra, which paves the way towards various applications. Overall, one can build models where there is some ranking between several blocks of variables, or where successive linear misfits are minimized. This kind of subject has already appeared in robotics [126]. One may wonder if there are other fields where the approach has an interest. Developing new learning algorithms based on this principle has still to be explored. In particular, writing an efficient algorithm which builds a lexicographical linear model from a huge amount of data is still a challenge.

Bibliography of Part II

[107] C. Ahrens and G. Beylkin, *Rotationally invariant quadratures for the sphere*, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 465 (2009), pp. 3103–3125.

[108] A. C. Aitken, *Determinants and matrices*, New York: interscience publishers, third ed., 1944.

[109] C. An and S. Chen, *Numerical Integration over the Unit Sphere by using spherical t-design*, `arXiv:1611.02785v1`, (2016).

[110] C. An, X. Chen, I. H. Sloan, and R. S. Womersley, *Well Conditioned Spherical Designs for Integration and Interpolation on the Two-Sphere*, SIAM journal on numerical analysis, 48 (2010), pp. 2135–2157.

[111] ——, *Regularized least squares approximations on the sphere using spherical designs*, SIAM Journal on numerical analysis, 50 (2012), pp. 1513–1534.

[112] M. Armstrong, *Groups and symmetry*, Springer, 1988.

[113] K. Atkinson and W. Han, *Spherical harmonics and approximations on the unit sphere: an introduction*, vol. 2044, Springer Science & Business Media, 2012.

[114] C. H. Beentjes, *Quadrature on a spherical surface*, Technical report, Oxford University - `https://cbeentjes.github.io/notes/2015-Quadrature-Sphere`, (2015).

[115] M. Brachet, *Schémas compacts hermitiens sur la Sphère: applications en climatologie et océanographie numérique*, PhD thesis, Université de Lorraine, 2018 (in French).

[116] M. Brachet and J.-P. Croisille, *Spherical Shallow Water simulation by a cubed sphere finite difference solver*, Quarterly Journal of the Royal Meteorological Society, 147 (2021), pp. 786–800.

[117] J. S. Brauchart and P. J. Grabner, *Distributing many points on spheres: Minimal energy and designs*, Journal of Complexity, 31 (2015), pp. 293–326. Oberwolfach 2013.

[118] C. Chen and F. Xiao, *Shallow water model on cubed-sphere by multi-moment finite volume method*, Journal of Computational Physics, 227 (2008), pp. 5019–5044.

[119] X. Chen, R. S. Womersley, and J. J. Ye, *Minimizing the condition number of a Gram matrix*, SIAM Journal on optimization, 21 (2011), pp. 127–148.

[120] S. Chevrot, R. Martin, and D. Komatitsch, *Optimized discrete wavelet transforms in the cubed sphere with the lifting scheme—implications for global finite-frequency tomography*, Geophysical Journal International, 191 (2012), pp. 1391–1402.

[121] F. Dai and Y. Xu, *Approximation theory and harmonic analysis on spheres and balls*, Springer Mongraphs in Mathematics, Springer-Verlag, 2013.

[122] C. De Boor and A. Ron, *On multivariate polynomial interpolation*, Constructive Approximation, 6 (1990), pp. 287–302.

[123] ——, *Computational aspects of polynomial interpolation in several variables*, Mathematics of Computation, 58 (1992), pp. 705–727.

[124] M. Descoteaux, *High Angular Resolution Diffusion Imaging (HARDI)*, John Wiley & Sons, Ltd, 2015, pp. 1–25.

[125] M. Descoteaux, E. Angelino, S. Fitzgibbons, and R. Deriche, *Regularized, Fast, and Robust Analytical Q-Ball Imaging*, Magnetic Resonance in Medicine, 58 (2007), pp. 497–510.

[126] A. Escande, N. Mansard, and P.-B. Wieber, *Hierarchical quadratic programming: Fast online humanoid-robot motion generation*, The International Journal of Robotics Research, 33 (2014), pp. 1006–1028.

[127] M. Faham and H. Nasir, *Weakly Orthogonal Spherical Harmonics in a Non-Polar Spherical Coordinates and its Application to Functions on Cubed-Sphere*, Sultan Qaboos University Journal for Science, 17 (2012), pp. 200–213.

[128] J. Fliege and U. Maier, *The distribution of points on the sphere and corresponding cubature formulae*, IMA Journal of Numerical Analysis, 19 (1999), pp. 317–334.

[129] B. Fornberg and J. M. Martel, *On spherical harmonics based numerical quadrature over the surface of a sphere*, Advances in Computational Mathematics, 40 (2014), pp. 1169–1184.

[130] W. Freeden, M. Z. Nashed, and T. Sonar, *Handbook of Geomathematics*, Springer, 2010.

[131] P. Funk, *Über Flächen mit lauter geschlossenen geodätischen Linien*, Mathematische Annalen, (1913).

[132] G. H. Golub and C. F. Van Loan, *Matrix computations*, The Johns Hopkins University Press, third ed., 1996.

[133] P. C. Hansen, *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*, vol. 4, Siam, 2005.

[134] D. P. Hardin, T. Michaels, and E. B. Saff, *A Comparison of Popular Point Configurations on S^2*, Dolomites Research Notes on Approximation, 9 (2016), pp. 16–49.

[135] C. P. Hess, P. Mukherjee, E. T. Han, D. Xu, and D. B. Vigneron, *Q-Ball Reconstruction of Multimodal Fiber Orientations Using The Spherical Harmonic Basis*, Magnetic Resonance in Medicine, 56 (2006), pp. 104–117.

[136] K. Hesse and Q. T. L. Gia, *L^2 error estimates for polynomial discrete penalized least-squares approximation on the sphere from noisy data*, Journal of Computational and Applied Mathematics, 408 (2022).

[137] K. Hesse, I. H. Sloan, and R. S. Womersley, *Numerical integration on the sphere*, in Handbook of Geomathematics, W. Freeden, Z. M. Nashed, and T. Sonar, eds., Springer, 2010.

[138] R. Hielscher and M. Quellmalz, *Reconstructing a function on the sphere from its means along vertical slices*, Inverse Problems and Imaging, 10 (2016), pp. 711–739.

[139] Y. Hristova, S. Moon, and D. Steinhauer, *A radon-type transform arising in photoacoustic tomography with circular detectors: spherical geometry*, Inverse Problems in Science and Engineering, 24 (2016), pp. 974–989.

[140] L. Ivan, H. De Sterck, S. A. Northrup, and C. P. T. Groth, *Multi-dimensional finite-volume scheme for hyperbolic conservation laws on three-dimensional solution-adaptive cubed-sphere grids*, Journal of Computational Physics, 255 (2013), pp. 205–227.

[141] J. H. Jensen, G. R. Glenn, and J. A. Helpern, *Fiber Ball Imaging*, NeuroImage, 124 (2016), pp. 824–833.

[142] B. A. Jones, G. H. Born, and G. Beylkin, *Comparison of the Cubed-Sphere Gravity Model with the Spherical Harmonics*, Journal of Guidance, Control, and Dynamics, 33 (2010), pp. 415–425.

[143] H.-G. Kang and H.-B. Cheong, *An efficient implementation of a high-order filter for a cubed-sphere spectral element model*, Journal of Computational Physics, 332 (2017), pp. 66–82.

[144] S. Kazantsev, *Funk–Minkowski transform and spherical convolution of Hilbert type in reconstructing functions on the sphere*, Siberian Electronic Mathematical Reports, 15 (2018), pp. 1630–1650.

[145] J. Keiner and D. Potts, *Fast evaluation of quadrature formulae on the sphere*, Mathematics of computation, 77 (2008), pp. 397–419.

[146] S. Kunis, *A note on stability results for scattered data interpolation on euclidean spheres*, Advances in Computational Mathematics, 30 (2009), pp. 303–314.

[147] S. Kunis, H. M. Möller, and U. von der Ohe, *Prony's method on the sphere*, The SMAI Journal of Computational Mathematics, 5 (2019), pp. 87–97.

[148] S. Kunis and D. Potts, *Fast spherical Fourier algorithms*, Journal of Computational and Applied Mathematics, 161 (2003), pp. 75–98.

[149] V. I. Lebedev, *Quadratures on a sphere*, USSR Computational Mathematics and Mathematical Physics, 16 (1976), pp. 10–24.

[150] V. I. Lebedev and D. Laikov, *A quadrature formula for the sphere of the 131st algebraic order of accuracy*, Doklady Mathematics, 59 (1999), pp. 477–481.

[151] D. Lee and A. Palha, *A mixed mimetic spectral element model of the rotating shallow water equations on the cubed sphere*, Journal of Computational Physics, 375 (2018), pp. 240–262.

[152] A. K. Louis, M. Riplinger, M. Spiess, and E. Spodarev, *Inversion algorithms for the spherical radon and cosine transform*, Inverse Problems, 27 (2011), p. 035015.

[153] J. L. McGregor, *Semi-Lagrangian Advection on Conformal-Cubic Grids*, Monthly Weather Review, 124 (1996), pp. 1311–1322.

[154] R. D. Nair, S. J. Thomas, and R. D. Loft, *A Discontinuous Galerkin Transport Scheme on the Cubed Sphere*, Monthly Weather Review, 133 (2005), pp. 814–828.

[155] G. Peyré, *L'algèbre discrète de la transformée de Fourier*, Ellipses, 2004.

[156] B. Portelenelle and J.-P. Croisille, *An efficient quadrature rule on the Cubed Sphere*, Journal of Computational and Applied Mathematics, 328 (2018), pp. 59–74.

[157] R. J. Purser, *Sets of optimally diversified polyhedral orientations*, Office note, National Centers for Environmental Prediction (U.S.), 489 (2017).

[158] ——, *Möbius Net Cubed-Sphere Gnomonic Grids*, Office note, National Meteorological Center (U.S.), 496 (2018).

[159] R. J. Purser and M. Rančić, *Smooth quasi-homogeneous gridding of the sphere*, Quarterly Journal of the Royal Meteorological Society, 124 (1998), pp. 637–647.

[160] R. J. Purser and M. Tong, *A minor modification of the gnomonic cubed-shaped sphere grid that offers advantages in the context of implementing moving hurricane nests*, Office note, National Centers for Environmental Prediction (U.S.), 486 (2017).

[161] W. M. Putman, *Development of the finite-volume dynamical core on the cubed-sphere*, PhD thesis, The Florida State University, 2007.

[162] M. Quellmalz, *A generalization of the Funk–Radon transform*, Inverse Problems, 33 (2017), p. 035016.

[163] R. Ramamoorthi and P. Hanrahan, *An efficient representation for irradiance environment maps*, in Proceedings of the 28th annual conference on Computer graphics and interactive techniques, 2001, pp. 497–500.

[164] M. Rančić, R. J. Purser, and F. Mesinger, *A global shallow-water model using an expanded spherical cube: Gnomonic versus conformal coordinates*, Quarterly Journal of the Royal Meteorological Society, 122 (1996), pp. 959–982.

[165] M. Rančić, R. J. Purser, D. Jović, R. Vasic, and T. Black, *A Nonhydrostatic Multiscale Model on the Uniform Jacobian Cubed Sphere*, Monthly Weather Review, 145 (2017), pp. 1083–1105.

[166] E. Rodriguez Bazan and E. Hubert, *Multivariate interpolation: Preserving and exploiting symmetry*, Journal of Symbolic Computation, 107 (2021), pp. 1–22.

[167] ——, *Symmetry in multivariate ideal interpolation*, Journal of Symbolic Computation, 115 (2023), pp. 174–200.

[168] C. Ronchi, R. Iacono, and P. S. Paolucci, *The "cubed sphere": a new method for the solution of partial differential equations in spherical geometry*, Journal of Computational Physics, 124 (1996), pp. 93–114.

[169] J. A. Rossmanith, *A wave propagation method for hyperbolic systems on the sphere*, Journal of Computational Physics, 213 (2006), pp. 629–658.

[170] B. Rubin, *Inversion formulas for the spherical Radon transform and the generalized cosine transform*, Advances in Applied Mathematics, 29 (2002), pp. 471–497.

[171] R. Sadourny, *Conservative finite-difference approximations of the primitive equations on quasi-uniform spherical grids*, Monthly Weather Review, 100 (1972), pp. 136–144.

[172] T. Sauer, *Polynomial interpolation of minimal degree*, Numerische Mathematik, 78 (1997), pp. 59–85.

[173] T. Sauer and Y. Xu, *On multivariate Lagrange interpolation*, Mathematics of computation, 64 (1995), pp. 1147–1170.

[174] A. Schneider, S. Schönborn, B. Egger, L. Frobeen, and T. Vetter, *Efficient Global Illumination for Morphable Models*, in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3885–3893.

[175] S. L. Sobolev, *Cubature formulas on the sphere invariant under finite groups of rotations*, Dokl. Akad. Nauk SSSR, 146 (1962), pp. 310–313.

[176] S. J. Thomas, J. M. Dennis, H. M. Tufo, and P. F. Fischer, *A Schwarz preconditioner for the cubed-sphere*, SIAM Journal on Scientific Computing, 25 (2003), pp. 442–453.

[177] L. N. Trefethen and J. A. C. Weideman, *The Exponentially Convergent Trapezoidal Rule*, SIAM Review, 56 (2014), pp. 385–458.

[178] D. S. Tuch, *Q-Ball Imaging*, Magnetic Resonance in Medicine, 52 (2004), pp. 1358–1372.

[179] P. A. Ullrich, P. H. Lauritzen, and C. Jablonowski, *Some considerations for high-order 'incremental remap'-based transport schemes: edges, reconstructions, and area integration*, International Journal for Numerical Methods in Fluids, 71 (2013), pp. 1131–1151.

[180] W. M. Washington, L. Buja, and A. Craig, *The computational future for climate and earth system models: on the path to petaflop and beyond*, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367 (2009), pp. 833–846.

[181] D. L. Williamson, *The evolution of dynamical cores for global atmospheric models*, Journal of the Meteorological Society of Japan. Ser. II, 85B (2007), pp. 241–269.

[182] R. S. Womersley, *Spherical designs with close to the minimal number of points*, Applied Mathematics Report AMR09/26, University of New South Wales, Sydney, Australia, (2009).

[183] ——, *Efficient spherical designs with good geometric properties*, in Contemporary computational mathematics-A celebration of the 80th birthday of Ian Sloan, Springer, 2018, pp. 1243–1285.

[184] C. E. Yarman and B. Yazici, *Inversion of Circular Averages using the Funk Transform*, in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, vol. 1, 2007, pp. I–541–I–544.

[185] G. Zangerl and O. Scherzer, *Exact reconstruction in photoacoustic tomography with circular integrating detectors II: Spherical geometry*, Mathematical Methods in the Applied Sciences, 33 (2010), pp. 1771–1782.