Polarization analysis and optimization geometry

Jeanne Lefevre

financé par
IDEX Université Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE GRENOBLE ALPES

préparée dans le cadre d'une cotutelle entre l'*Université Grenoble Alpes* et *l'Université de Melbourne*

Spécialité : **SIGNAL, IMAGE, PAROLE, TELECOMS**

Arrêté ministériel : le 6 janvier 2005 – 25 mai 2016

Présentée par

Jeanne LEFEVRE

Thèse dirigée par **Nicolas LE BIHAN et** Jonathan **MANTON**

préparée au sein des Laboratoires **Gipsa-Lab** à Grenoble et **Electronical&Electrical Engineering** à Melbourne

dans l'école doctorale **EEATS**

Analyse de la polarisation et géométrie en optimisation.

Thèse soutenue publiquement le **07/12/2021** au laboratoire Gipsa-Lab à GRENOBLE
devant le jury composé de :

Mr Pierre CHAINAIS
Professeur, Ecole Centrale de Lille, CRIStAL, Examinateur
Mme Audrey GIREMUS
Professeur, Université de Bordeaux, IMS, Rapporteur
Mr Guillaume GINOLHAC
Professeur, Université Savoie Mont-Blanc, LISTIC, Rapporteur
Mr Salem SAID
Chargé de recherche HDR, Université de Bordeaux, IMS, Examinateur
Mr Olivier MICHEL
Professeur des universités , Université de Grenoble Alpes, Gipsa-Lab, Président du Jury

Abstract

In the first part of this thesis, we introduce the concept of polarization of bivariate signals by comparing several approaches present in the literature. In particular, a complex representation and a vector representation are identified. We compare these representations and highlight the fact that they induce differences in treatment. We devote a section to the definition of the instantaneous polarization, in a similar spirit to that of the definition of the instantaneous frequency. We then argue that signal representation plays a strong role in the identification of certain instantaneous parameters. As an example, we explore in more detail a representation of polarization via an embedding of complex signals in the quaternion algebra. This work is the result of a thesis which precedes us by a short time. After noting the ease of writing allowed by the quaternionic embedding, we give theoretical explanations for the success of such an enterprise. To this end, an algebraic approach to the Fourier transform is proposed, then the link between quaternion algebra and geometric algebras is established. The geometric algebras provide a new type of object: the spinors whose interest in the representation of polarization is highlighted. In the last chapter of this part, we explore through trivariate signals the extension of polarization analysis to n-variate signals and formulate an analysis based on the study of invariants. We understand polarization transformations as the action of the group U(n) on the space of possible covariance matrices for a trivariate signal. In the second part, we propose a method for solving particular optimization problems of the density estimation type. The form of the problem is known in advance but the particular problem depends on an observation vector that changes at each acquisition. Our functions are defined on differentiable manifolds, and under certain conditions on the form of the optimization problem, we show that pre-computed solutions can be used as a basis to estimate a solution in a predictable time. This part calls for knowledge in differential geometry, the gist of which is given in the first chapter. We also recall some results on the convergence of Newton's methods and give a brief overview of homotopic methods. The end of this chapter illustrates the notions introduced on the example of the Rayleigh quotient, a function whose critical points are the eigenvectors of a certain symmetric matrix.

Résumé

Dans la première partie de cette thèse, nous introduisons le concept de polarisation des signaux bivariés en comparant plusieurs approches présentes dans la littérature. En particulier, une représentation complexe et une représentation vectorielle sont identifiées. Nous comparons ces représentations et mettons en évidence le fait qu'elles induisent des différences de traitement. Nous consacrons une partie à la définition de la polarisation instantanée, dans un esprit similaire à celui de la définition de la fréquence instantanée. Nous argumentons alors que la représentation des signaux joue un rôle fort dans l'identification de certains paramètres instantanés. En guise d'exemple, nous explorons plus en détail une représentation de la polarisation via un plongement des signaux complexes dans l'algèbre des quaternions. Ce travail est le fruit d'une thèse qui nous précède de peu. Après avoir constaté les facilités d'écriture permises par le plongement quaternionique, nous donnons des explications théoriques au succès d'une telle entreprise. A cette fin, une approche algébrique de la transformée de Fourier est proposée, puis le lien entre algèbre des quaternions et algèbres géométriques est établi. Les algèbres géométriques fournissent un nouveau type d'objet: les spineurs dont l'intérêt dans la représentation de la polarisation est mis en évidence. Dans le dernier chapitre de cette partie, nous explorons à travers les signaux trivariés l'extension de l'analyse de la polarisation aux signaux n-variés et formulons une analyse basée sur l'étude d'invariants. Nous comprendrons les transformations de polarisation comme l'action du groupe U(n) sur l'espace des matrices de covariance possibles pour un signal trivarié. Dans la deuxième partie, nous proposons une méthode de résolution de problèmes d'optimisation particuliers du type estimation de densité. La forme du problème est connue en avance mais le problème particuler dépend d'un vecteur d'observation qui change à chaque acquisition. Nos fonctions sont définies sur des variétés différentiables, et sous certaines conditions sur la forme du problème d'optimisation, nous montrons que des solutions pré-calculées peuvent servir de base pour estimer une solution en un temps prévisible. Cette partie fait appel à des connaissances en géométrie différentielles dont un tableau est dressé en premier chapitre. Nous rappelons également quelques résultats sur la convergence des méthodes de Newton et donnons un bref aperçu des méthodes homotopiques. La fin de ce chapitre illustre les notions introduites sur l'exemple du quotient de Rayleigh, une fonction dont les points critiques sont les vecteurs propres d'une certaine matrice symétrique.

Contents

Part I

Representation of bivariate and trivariate signals

Introduction

> *Expression and shape mean almost more to me than knowledge itself*
>
> — **Hermann Weyl**
> (German mathematician)

Signal processing is a discipline that cares about the real world because real world provides signals. Contrarily to mathematics which are by nature recursive and study themselves, signal processing starts with quantities that have either been observed somewhere or computed to represent something of importance such as a cost function informing us on the optimal states of a system. In this chapter, we present various pieces of works related to the question of representing signals and physical quantities in mathematical language. This question of representation should be seen as a roadmap to circulate through the different parts of this chapter. It arises because of its close relationship to the theory of transformation of signals. For instance, the probably most famous transformation of a signal is the Fourier transform. It is so well-known and so inseparable from its interpretation in terms of physical properties, that it is not immediate to realise how it is in fact completely dependent on the mathematical representation of the object "time" that had to be chosen before. This representation of time as the group $(\mathbb{R}, +)$ seems so obvious, so natural to the modern reader that we do not even question the fact that it is abritrary. Indeed, we have no physical certainty about the homogenity or even continuity of time, while we certainly have mathematical certainties for the same properties on $(\mathbb{R}, +)$ with its classical topology. Hence, representations translate the properties and the assumptions we connect to a physical notion sometimes not fully consciously. The reason for this is that representations come with their own structure, a notion that is clarified in the next paragraph. It is often the case that the choice of a mathematical counterpart to a physical quantity determines the way we understand it and the role it plays in the equations. As an example: the old quantum theory was elaborated by using scalar-valued functions to represent the observables of a particle such as position, speed and energy through time. This formalism could not explain results such as the spectral lines of hydrogen. It is when Heisenberg and his colleagues proposed a matrix formulation of the observables of a quantum particle that the first logically consistent formulation of quantum mechanic was realised, giving new interpretation to the non-commutativity of things such as position and energy **[Aitchison et al., 2004]** **[Roux, 2011]**. This question of representation is very well illustrated through the historical debate around vector calculus and the use of quaternions instead of cartesian coordinates. It seems from this example that one representation tends to impose itself over all the other ones through time, so that the filter of history creates this impression that the representation we use now are the only natural counterpart for the physical object they represent. With the rise of multi-dimensional data processing, the question of representation becomes an explicit field of signal processing. In machine learning, data are represented by features which are computed to best serve the purpose previously defined **[Bishop, 2006]**. From the choice of the features depend the result

in terms of classification, or modelization. This interest for alternate mode of representations is even more obvious for data in large dimensions, as cartesian coordinates and euclidian distance are not geometrically interpretable or discriminant anymore. The idea that data lie on a subspace that is not captured by the use of cartesian coordinates but which fits better to the inner complexity of the data is the building brick for non-linear dimensionality reduction methods**[Vlachos et al., 2002]**.
For every piece of work that is presented in this chapter, we will strive to present its scientific context and to highlight where representation questions have an importance. Sections follow on in a way that the reader should gain new levels of understanding at each steps, the last sections informing the first ones. For instance we start with the usual definition of the Fourier transform, then we developp the group theory that defines it, and we apply it to the sphere using representation theory in the next section reaching a new level of abstraction at each steps. However, it is not necessary to reach the end of the chapter to understand the contributions of the first parts, and we try to adapt the complexity and abstraction level as we go on.

0.0.1 A structural view of math:

Since the work of Bourbaki in the second half of the 20th century, structures have been deemed to be the fundamental elements in the study of mathematics, fundamental in the sense that they could come before any other element was brought to the table, except maybe for a few axioms. This of course is not entirely true: a theory that could build mathemtatics entirely from a few logical bricks has not yet been found. Rather than being an objective fundation to modern mathematics, the notion of structure is an intellectual reading grid, coming with its own typology and enabling us to shed light on our thinking process. Especially in a field such as signal processing where it is necessary to be moving back and forth between the physical concrete world and its abstract translation in the mathematical language. The notion of structure questions us on the properties we assume to know about the physical world. In its article "l'architecture de la pensée mathématiques"[1] Bourbaki defines a structure as follow:

> On peut maintenant faire comprendre ce qu'il faut entendre par une *structure mathématique*. Le trait commn des diverses notions désignées sous ce nom générique, est qu'elles s'appliquent à des ensembles d'éléments dont la nature *n'est pas spécifiée*; pour définir une structure, on se donne une ou plusieurs relations où interviennent ces éléments; on postule ensuite que la ou les relations données satisfont à certaines conditions (qu'on énumère) et qui sont les *axiomes* de la structure envisagée.[2]

[1] Architecture of mathematical thinking
[2] It is now possible to explicit what should be understood by a *mathematical structure*. The common trait between the diverse notions gathered under this generic name is that they apply to a set whose element's nature *is not specified*. To define a structure, one should need one or several relations satisfying some conditions (that we enumerate) et which are the *axiomes* of the intended structure

For those who are used to working with some mathematical elements, nothing is worth examples to understand this concept of structure. A structure is the association of a set, a relation and some axioms on this relation. For instance, the group structure is defined on a non-empty set G with a relation $\cdot$ that sends two elements in G to a third one. The axioms satisfied by this relation are:

1. Associativity: $\forall x, y, z \in G : (x \cdot y) \cdot z = x \cdot (y \cdot z)$
2. Existence of a neutral element: $\exists e \in G, \forall x \in G, e \cdot x = x \cdot e = x$
3. Existence of an inverse : $\forall x \in G, \exists x', xx' = x'x = e$

Any property of a group that does not depend on the nature of its elements can be deduced from these three axioms. As it can be guessed, the most crucial element in the definition of a structure is the relation defined on a set, and it is natural that a typologie of structure is based on them. Relations that connect three elements, by sending two of them on the third are called "composition relations". A group relation is of course a "composition relation" and any structure built upon such a relation is called an "algebraic structure". For instance a field is an algebraic structure with two relations, and among the fields, a particuliar set is$(\mathbb{R}, +, \times)$ with the addition and multiplication as its composition relations. Another important type of structure are defined by binary relations. They are this time relations between two elements x and y such as the "greater than" $>$ relation in $\mathbb{R}$. The inclusion relation between sets is also a binary relation, so that a set equipped with the inclusion operation is a binary strucure. Note that not all subsets of a set are in relation! A third important type of structures are the topological structures: they provide an abstract formulation of intuitive notions such as neighborhood, limit or continuity, intuitions that derive from our perception of space. A hierarchy on structures can be advanced where the three defined type are simple structure at the basis. More complex structures can be built either by adding more axioms on the relations, for instance an abelian group is a "more complex" structure than a group in the sense that more elements are necessary to its definition. Structure can also be built by combining relations in a compatible way: a topological group is a group whose relation is continuous with respect to its topology. On the highest level of the complexity hierarchy lie sets whose elements are precisely specified, such as functions of the real variables, the sphere S^2 etc... The use of the word "complexity" in this paragraph is not related to something being complicated. Indeed, the more something is specified in mathematics, and the more hypothesis and axioms are available the less complicated it tends to be. While with very open and "non-complex" structures, the physical sense is of little use and they usually seem extremely complicated to the learner.

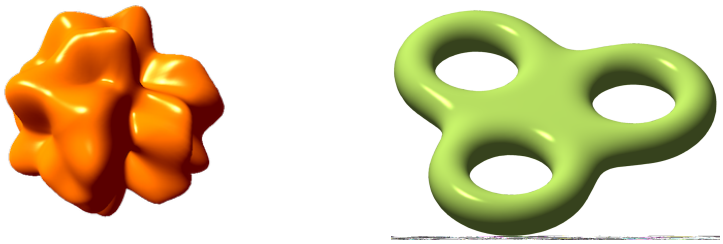
Remark 0.1.1

This opposition between complicated and complex things in mathematics explains why the teaching order of mathematics reverses the complexity hierarchy we have presented before. Indeed, at school mathematics start with notions students can relate to their physical experience of the world through very specialized objects: integers, real numbers, two-dimensonal geometrical objects etc.. It gets down as the student climbs the academic levels to the study of abstract structures. It is a known fact that, probably under the conceptual influence of Bourbaki, french politicans have tried in the seventies to align this conceptual hierarchy of mathematical notions with the order in which they were taught **[M. Criton, 2016]**. This led to the definition of sets with their operations of bijections, surjections and injections around the age of seven and the attempt to have kids count in at least three different basis. This was, predictably, a pedagogical failure as most students have no ambition or taste for the mastering of such abstract notions.

Questioning what element of structure we assume at each step is an interesting exercice. For instance, the two-dimensional sphere S^2 is an object for which everybody has a precise mental representation in which the sphere is this rotationaly invariant surface sitting in a three-dimensional space. What if we look at the sphere forgetting all elements of structures but the set? Then the sphere becomes no more than an infinite collection of elements, in bijection with any interval of $\mathbb{R}$ for instance. In other words, as the cardinality of the sphere is the same as the cardinality of any euclidian space or interval, there is no way to differentiate the sphere from an interval only with the set structure. Imagine now that we start from the same sphere, and we forget everything but its topology and set structure. Topology is roughly the study of continuous shape. The topology of the sphere is characterized by the fact that the sphere is a two-dimensional surface cutting a three-dimensional space between an "interior" and an "exterior" while having no holes (you can not drag it with a lasso, contrarily to the torus). As a consequence, any shape that can be obtained by blowing a bubble hook is a topological sphere. An interesting question is: what is the necessary layer of structure so that in that structure a sphere defines the shape I am imagining in my head (round, spherically invariant)? There are several possible answers to this question, one of them is: a metric space structure. A metric space is a specialization of a topological space, with extra axioms. In a metric space, the sphere is described as the set of points at equal distance from a central point. This necessarily draws the figure we have in mind.

0.0.2 Maps between structures

In the previous example with the sphere we put under evidence that different objects are "the same" when seen through a particular structure. This notion of being "the same" mathematically translates by the existence of a shape-preserving map called an isomorphism. The particular definition of an isomorphism depends on the particular strucure it preserves, for instance a set-isomorphism is a bijection while a vector-space isomorphism is a linear



(a) This is a topological sphere **(b)** This is not a topological sphere

Fig. 0.1. – Example and counter-example of 2-dimensional topological spheres (Source: Wikipedia)

invertible map etc.. A few examples of structures and their isomorphisms is presented in table figure 0.2. An isomorphism is a particular case of a morphism, the broader class of structure-preserving function. Two sets connected by an isomorphism can not be distinguished from the point of view of structure. A morphism between two sets however shows they share a common structure such as linearity or a topology but does not necessarily preserve quantities characterizing these structures such as dimension or genus[3]. This quote of Hermann Weyl illustrates the importance of isomorphisms:

> A guiding principle in modern mathematics is this lesson: Whenever you have to do with a structure-endowed entity X, try to determine its group of automorphisms[4], the group of those element-wise transformations which leave all structural relations undisturbed. You can expect to gain a deep insight into the constitution of X in this way.

Structures	Isomorphisms	morphisms
Group	invertible group morphisms	group morphisms
Topological space	homeomorphisms	continuous functions
Metric space	Isometries	uniformly continuous functions
Linear space	Isomorphisms	linear maps
Differentiable space	diffeomorphisms	differentiable functions

Fig. 0.2. – table of different structure morphisms and isomorphisms

[3] The genus is a topological notion that counts the number of holes inside a shape
[4] an automorphism is an isomorphism of a set into itsel

1.1 Introduction

Vector and complex representations of bivariate signals have a joint history that goes back to the very birth of the notion of vector: two-dimensional vectors were invented to propose a geometric interpretation of complex numbers **[Crowe, 1994]**. No doubt in this context that the two formalisms are in many ways equivalent, and more a question of personal or field preference than of real mathematical importance. The complex representation was more common among oceanographer **[Thomson and Emery, 2014]** than seismologists **[Kanasewich, 1981]**, and is now advocated by the signal processing community **[Schreier and Scharf, 2010]** for its capacity to jointly process the real and imaginary information. This is probably most obvious in the possibility to uniquely factorize a complex-number into an amplitude and an angle, usually called its polar form. The 2-dimensional euclidian space, not being an algebra has no such direct factorization. We show in section 3.2 how a polar form can actually be built for vector signals, but we do not use it before chapter 3. Let u and v be two jointly observed scalar time-signals and $\boldsymbol{i}$ fesignates the complex root of -1. Then the complex and vector embeddings denoted x and $\boldsymbol{x}$ are respectively:

$$\begin{aligned} x : \ \mathbb{R} &\to \mathbb{C} \\ t &\mapsto u(t) + \boldsymbol{i}v(t) \end{aligned} \qquad\qquad \begin{aligned} \boldsymbol{x} : \ \mathbb{R} &\to \mathbb{R}^2 \\ t &\mapsto \begin{pmatrix} u(t) \\ v(t) \end{pmatrix} \end{aligned}$$

Structural difference between $\mathbb{R}^2$ and $\mathbb{C}$

The transformation that sends the complex-signal $x(t)$ on the vector $\boldsymbol{x}(t)$ is a vector-space isomorphism, which tells us exactly what is equivalent in both formalisms: linear operations and the vector space structure. However, $\mathbb{C}$ possess the structure of a field which $\mathbb{R}^2$ does not have. In particular product operations in $\mathbb{C}$ can not be directly translated in $\mathbb{R}^2$. Among such operations, factorization of a complex under its polar form as $x(t) = a(t)e^{\boldsymbol{i}\theta(t)}$ has no direct counterpart in $\mathbb{R}^2$. Similarly, it is possible to encode linear operations of a vector of the plane in the shape of a complex: for any $z, z' \in \mathbb{C}$ there exists a $q \in \mathbb{C}$ such that

$$qz = z' \tag{1.1}$$

The quantity q can be interpreted as the transformation that sends z over z'. This gives elements of the complex plane a dual interpretation: they are both elements and transformations on the elements. In a euclidian space, we define elements that act linearly on vectors through the notion of group action: the linear transformation of a vector $\boldsymbol{z}$ into $\boldsymbol{z}'$ results from the action of the group $M_2(\mathbb{R})$ on $\mathbb{R}^2$ and there is a 2×2 matrix A such that

$$A\boldsymbol{z} = \boldsymbol{z}' \tag{1.2}$$

Notice that this matrix-vector equation is not strictly equivalent to the equation with complex elements (1.1): if seen as an equation in the matrix A, (1.2) yields an infinity of solution. Indeed consider for instance that

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a & 1 \\ b & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

is valid for $\{a, b \in \mathbb{R}\}$. Only one complex number q however can sasisfy (1.1) due to the invertibility of all elements in $\mathbb{C}$. We observe in the next section how the Fourier analysis is conducted in the complex or vector framework. Ultimately, we aim at comparing the representation of polarizations, which are closely related to the Fourier analysis of a signal.

1.2 Spectral analysis for stationary signals

1.2.1 Fourier transform

Spectral analysis is based on the possibility to see a signal as a sum of repeating patterns. The universal tool to conduct the spectral analysis of signals is the Fourier transform. The core of the Fourier transform does not depend much on the target-space but on the domain to which belongs the variable of the signal, this point is developped in section 2.1 of chapter 2 where the abstract framework that defines the Fourier transform on various spaces is briefly introduced. In this section, we focus on signals that vary through time, modeled as functions of the real variable $t \in \mathbb{R}$.

Definition 1. *The Fourier transform $\mathcal{F}$ applied to an absolutely summable scalar signal x indexed on the real variable t is the linear operator:*

$$\begin{aligned} \mathcal{F}\colon \quad \mathcal{L}^1(\mathbb{R}) &\to \mathcal{L}^1(\mathbb{R}) \\ x &\mapsto \mathcal{F}(x) = X : \nu \mapsto \int_{-\infty}^{+\infty} x(t) e^{-2i\pi\nu t} dt \end{aligned}$$

Note that the definition requires for the signal x to be absolutely integrable, a condition of little restriction in practice as most studied signals in physics are observed on a limited time interval, and translate by functions with finite support, obviously integrable. The force of the Fourier transform is apparent in the following propositions.

Proposition 1. *Isometry of the Fourier transform*

1. *The Fourier transform is a linear invertible transformation*
2. *Parseval-Plancherel identity: Let x, y be two square integrable signals and X, Y their Fourier transform. Then,*

$$\int_{-\infty}^{+\infty} x(t)\overline{y}(t)dt = \int_{-\infty}^{+\infty} X(\nu)\overline{Y}(\nu)d\nu$$

The second-point shows that the Fourier transform is an isometry for the set of square-integrable functions with the inner product $\langle x, y \rangle = \int_{-\infty}^{+\infty} x(t)\overline{y}(t)dt$. Together, the propositions show that $\mathcal{F}$ is a linear, bijective isometry, meaning it preserves the distance and linear relations between functions of the set of square-integrable functions $\mathcal{L}^2(\mathbb{R})$ (again, any function with finite support belongs to this set). The two representations, in the temporal and frequential world are said to be *dual*[1] and the previous property shows that they are equivalent in their ability to discriminate and compare signals. Furtermore, the Fourier transform commutes with any linear combinations of functions: summing functions in the time or spectral domains is equivalent. If two representations are too completely equivalent, it is legitimate to wonder what can the second bring to the first, and if it is needed. In the case of the Fourier transform, the answer lies in a particular type of transformations called filtering, and more particularly linear time invariant (LTI) filtering. A linear filter models a physical phenomenon where an input signal is transformed into an output signal linearly. Linearly translates as follows: if the input is the sum of two weighted signals, then the output of the filter is the weighted sum of the output of the filter for each contribution. Systems are time-invariant if the transformation they impose on the signal is independent of time. For instance, the echo of a sound produced by a place depends on the geometry of that place and does not depend on time but only on the input sound. LTI filters are completely determined by an impulse response, the theoretical response if the input signal is an infinitely concentrated impulse, mathematically a dirac centered in 0. Next proposition shows how LTI filtering blends nicely in Fourier theory.

Proposition 2. *Filtering in temporal and frequential domain*

Let x be a signal and h the impulse response of a linear filter (any complex-valued summable function indexed on time can be considered to be an impulse response of a linear filter). Then,

$$\mathcal{F}(x * h) = \mathcal{F}(x)\mathcal{F}(h)$$

The Fourier transform sends convolution products on point-wise products, it realizes a homomorphism between the two algebras: $(\mathcal{L}^2(\mathbb{R}), *)$ and $(\mathcal{L}^2(\mathbb{R}), \times)$. Here maybe lies the interest of the Fourier transform: filtering operations present themselves in a simpler form in the frequency domain than in the temporal domain. They have an interpretable physical counterpart: the filter directly increases or decreases the intensity of the signal at certain frequencies.

In the definition of the Fourier transform, it was said to apply to “scalar”-valued functions, a term that applies to real or complex numbers, and should be understood as complex unless stated otherwise. In the rest of the document, we might omit the dependency in t to alleviate

[1] The term is explained in chapter 2 section 2.1

our notations. Every signal in this chapter is time-dependent. The Fourier transform is extended to a vector signal $\boldsymbol{x}$ by:

$$\mathcal{F}(\boldsymbol{x}) = \begin{pmatrix} \mathcal{F}(x_1) \\ \vdots \\ \mathcal{F}(x_n) \end{pmatrix} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \boldsymbol{X}$$

Interestingly, the Fourier transform applied to a bivariate vector is an operation that sends $\mathbb{R}^2$ to $\mathbb{C}^2$ while the Fourier transform on a complex signal is from $\mathbb{C}$ to $\mathbb{C}$. In the first case there is an increase in dimension through the Fourier transform whereas it is not the case in the second. We already stated that the transform sending a complex bivariate signal x on its vector counterpart $\boldsymbol{x}$ is a point-wise isomorphism. The Fourier transform commutes with linear operations, hence we expect to find a point-wise isomorphism between the Fourier transforms of X and $\boldsymbol{X}$. Obviously, for dimensional reasons, it does not exist, and this highlights the fact that the Fourier transform we defined on vectors has not the same properties as the Fourier transform on complex. Typically, it does not commute with linear applications. We investigate further to try to understand why this two Fourier transforms seemingly very close behave differently. In the complex notation, the kernel of the Fourier transform $e^{-i2\pi\nu t}$ induces a rotation of the terms of the signal x, while the vector Fourier transform is equivalent to the transform

$$\mathcal{X}(\nu) = \int_{-\infty}^{+\infty} e^{-i2\pi\nu t} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} dt = \int_{-\infty}^{+\infty} e^{-i2\pi\nu t Id} \boldsymbol{x}(t) dt$$

where the matrix kernel $e^{i2\pi\nu t Id}$ does not induce a rotation on the vector $\boldsymbol{x}$. A matrix-vector formulation of the Fourier transform would be geometrically analog to the complex Fourier transform for the kernel

$$\begin{pmatrix} \cos(-2\pi\nu t) & -\sin(-2\pi\nu t) \\ \sin(-2\pi\nu t) & \cos(-2\pi\nu t) \end{pmatrix}.$$

A difference induced by this translation of the Fourier transform in the vector world is the separability of the contributions of x_1 and x_2 to the spectrum. Let $x = x_1 + \boldsymbol{i}x_2$ and $\boldsymbol{x}$ its vector counterpart. Then $X = X_1 + \boldsymbol{i}X_2$ and $\boldsymbol{X} = \begin{pmatrix} X_1 & X_2 \end{pmatrix}^T$. With the vector Fourier transform, it is possible at each ν to separate the contributions $X_1(\nu)$ and $X_2(\nu)$ while the complex Fourier transform mixes the informations in the same complex number. Another important factor is that a symmetry is preserved by the vector Fourier transform while it is not by the complex Fourier transform, it is the subject of proposition below.

Proposition 3. *Hermitian symmetry of the vector Fourier transform*

Let $\boldsymbol{x}$ be a vector-valued bivariate signal and $\boldsymbol{X}$ its Fourier transform. Then,

$$\boldsymbol{X}(-\nu) = \overline{\boldsymbol{X}(\nu)}$$

and for x the complex counterpart of $\boldsymbol{x}$ and X its Fourier transform, the following equivalence applies:

$$X(-\nu) = \overline{X(\nu)} \iff \textit{x is real}$$

This gives us a first insight on the importance of representation: depending on the representation chosen for bivariate signals, it is not the same spectral analysis that is performed and it has not the same properties. Representation influences the choice of operator. Of course, we shall be clear that even the non-isomorphism of the complex and spectral analysis of bivariate signals is relative: there exists bridges between both methods that are well known to the community, and any quantity that appears in one formalism can be calculated in the other. In terms of available information, they are in the end equivalent.
The next aspect we want to study is polarization of bivariate signal, a notion that is clearly defined for monochromatic contributions only. This is not as reductive as it may seem: any signal can be decomposed as a sum of monochromatic contributions, hence the monochromatic signal really appears to be the building block of spectral and polarization analysis. Next section discuss the interesting features of monochromatic signals in the bivariate framework.

1.2.2 Polarization

Among signals, the simplest brick from a spectral perspective are the monochromatic signals, signals that consist in the repetition of a pattern of limited time extension. This pattern can be anything, an impulse, a triangle or a complicated shape repeated over and over. A common point between them is that their frequency content is time invariant. Among the monochromatic signals, we must distinguish the monochromatic waves, for the reason that the Fourier transform is, by its definition, centered around sinusoidal signals. The wave shape is general enough as any continuous function is the limit of a trigonometric polynomial, see the remark below for a little incursion in the Stone-Weierstrass theorem.

Remark 2.2.1: The Stone-Weierstrass theorem

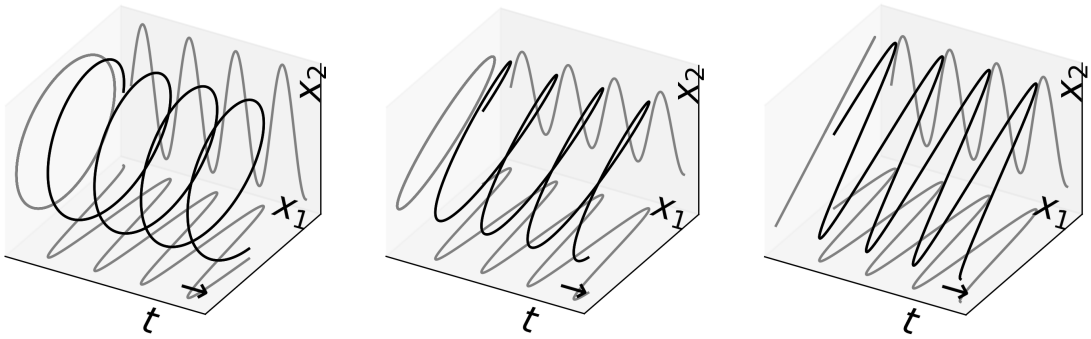
Originally, Weierstrass showed that any continuous function has an arbitrarily good polynomial approximation in 1885. In 1937, Stone seeked a generalization of the concept by asking what set of functions other than polynomials were enough to approximate arbitrarily well any continuous function. By arbitrarily good approximation we mean getting within a ε distance of the target function for any positive ε and for the distance derived from the sup-norm $\|\cdot\|_\infty$. Hence a reformulation of the question asked by Stone is, on what condition is a subalgebra in the set of continuous functions dense in this set for the sup-topology. It appeared that the crucial criteria for a subalgebra to be dense is that it separates the points of $\mathbb{R}$, i.e for two different points $a, b \in \mathbb{R}$, there must exist a function p in the algebra that satisfies $p(a) \neq p(b)$. From this theorem it

can be deduced that the subalgebra of trigonometric polynomials is dense in the set of continuous function. Note that not all functions can be used to build polynomials. Of course, it is always possible to evaluate a polynomial P in a function f, but the exponential is the only real function that satisfies $(P \times Q) \circ \exp = P \circ \exp \times Q \circ \exp$. If we wanted to use a basis of periodic triangle functions with integer periods for instance, we would find that any continuous function can indeed be approximated by an element of the algebra generated by the periodic triangle functions. However, an element of this algebra would not write as a polynomial in the triangle functions with different periods. One can therefore understand the interest of the trigonometric functions over other simple monochromatic functions.

A monochromatic wave is fully defined by the triplet: amplitude, frequency and phase. Naming (a, ν_0, ϕ_0) these quantities, a monochromatic signal writes $x(t) = a\sin(\nu_0 t + \phi_0)$. Trigonometric functions $\sin$ or $\cos$ can be used indiferently keeping in mind that they are in quadrature: $\cos(\theta - \pi/2) = \sin(\theta)$. A monochromatic bivariate wave is composed of a pair of monochromatic waves at the same frequency, hence defined by 5 parameters: two amplitudes, one frequency and two phasis written $(a_1, a_2, \nu_0, \phi_1, \phi_2)$. The elementary shape of the bivariate monochromatic wave in both formalisms is:

$$x(t) = a_1\cos(2\pi\nu_0 t + \phi_1) + \boldsymbol{i}a_2\sin(2\pi\nu_0 t + \phi_2) \quad \text{or} \quad \boldsymbol{x}(t) = \begin{pmatrix} a_1\cos(2\pi\nu_0 t + \phi_1) \\ a_2\cos(2\pi\nu_0 t + \phi_2) \end{pmatrix}$$

where $a_1, a_2 \geqslant 0$, $\nu_0 > 0$, $\phi_1, \phi_2 \in [0, 2\pi[$.



(a) Circular polarization **(b)** Elliptical polarization **(c)** Linear polarization

Fig. 1.2. – Three examples of monochromatic bivariate signals, with time along the t axis. The trace of the signal can be observed in the $x_1 - x_2$ plane. It is either a circle, an ellipse or a line depending on the state of the polarization

On figure 1.2,different monochromatic bivariate signals are represented, with time along the t-axis. Observe that for each signals, the repeating pattern draws an ellipse in the plane, known as its polarization ellipse. On figure 1.3, a non-monochromatic signal represent velocities in the ocean's depth. Ellipses are not obviously present anymore. The subject of polarization for non-stationary signals is really one of the main interest of this chapter and is first introduced

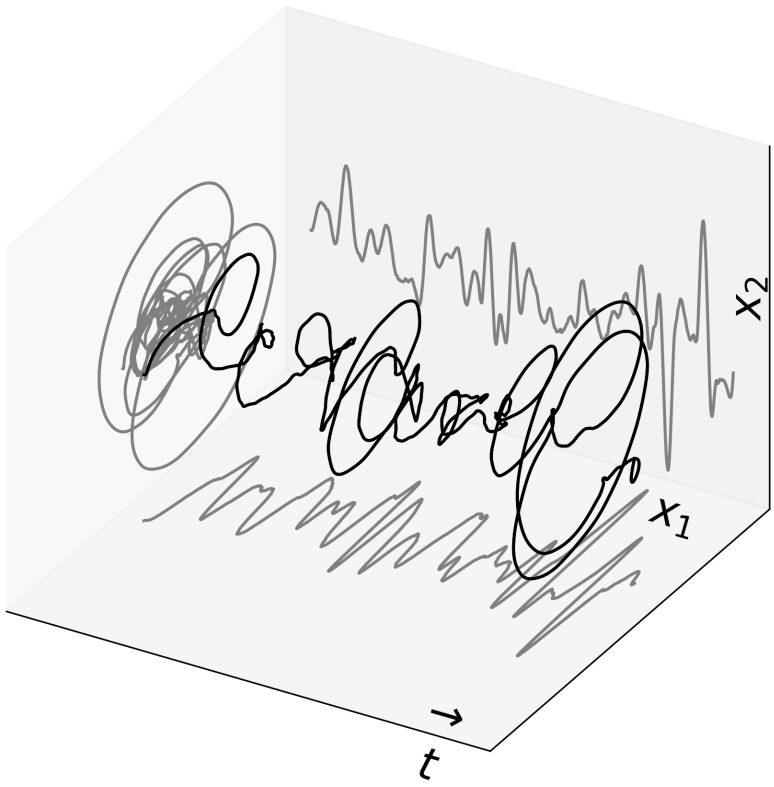


Fig. 1.3. – Plane velocities of the current at a precise depth in the ocean measured by an aquadopp device during a research cruise from St. John's, Canada to Reykjavik. In the $x_1 - x_2$ plane, the 2-dimensional velocities of the current is shown while time developps on the t-axis. The signal was smoothed to keep only tendencies and not hour-by-hour variations. A rotary behaviour is apparent, even if clearly not monochromatic. (Source of the data: **[Karstensen et al., 2016]**)

in section 1.3. The question of retrieving the polarization of a signal from its temporal representations x or $\boldsymbol{x}$ provides us with a good example of the influence of representations. The polarization ellipse is characterized by three quantities, figure 1.4 proposes geometrical

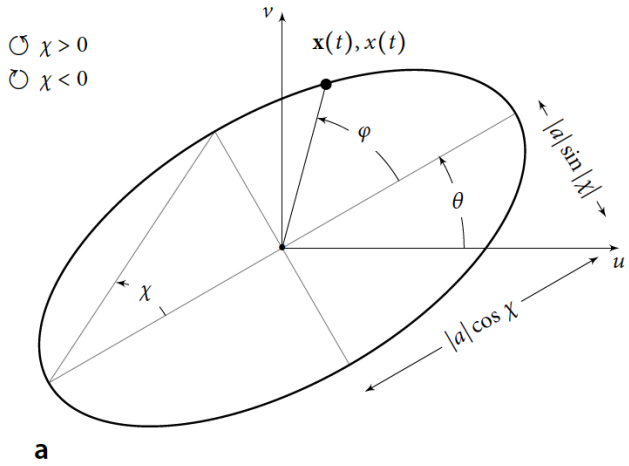


Fig. 1.4. – Parametrization of an ellipse by direct geometrical parameters. The length of the major axis is a, the inclination of the semi-major axis in the orthonormal frame is θ, and the eccentricity is represented by the angle χ such that if $\chi = 0$ the ellipse retracts to a line and if $\chi = \frac{\pi}{4}$ the Ellipse is a circle. (Source: Julien Flamant)

coordinate-free parameters for the description of the ellipse. They are preferred for their direct interpretibility in terms of the shape of the ellipse. However, computing (a, θ, χ) from the temporal representation of the signal demands access to parameters not always easy to estimate see table table 1.1. Antoher set of parameters (S_0, S_1, S_2, S_3) has been defined by the optical community. They have the advantage of being measurable even for high frequency signals such as light.

The Stokes parameters

The Stokes parameters (S_0, S_1, S_2, S_3) were brought up by Sir George Gabriel Stokes in his seminal paper of 1852 (**[Stokes, 1852]**) at a time where the study of polarization was solely driven by experiments on light. He introduced four measurable quantities, now known as the Stokes parameters, that could characterize any polarization state. The central theorem of the paper is, in its original phrasing:

> "When any number of independent polarized streams, of given refrangibility, are mixed together, the nature of the mixture is completely determined by the values of four constants, which are certain functions of the intensities of the streams, and of the azimuths and eccentricities of the ellipses by which they are respectively characterized; so that any two groups of polarized streams which furnish the same values for each of these four constants are optically equivalent."

Stokes was the first to describe polarization in terms of intensities rather than field vectors, which at optical frequencies could never be measured. From a representation perspective, the Stokes parameters introduced the idea that polarization could be represented on a 2-dimensional sphere, known as the Poincaré sphere see figure 1.5. Indeed, for a fully[2] polarized light the following relation is always satisfied:

$$S_1^2 + S_2^2 + S_3^2 = S_0^2$$

So that the triplet (S_1, S_2, S_3) identifies a point on a sphere of radius S_0.

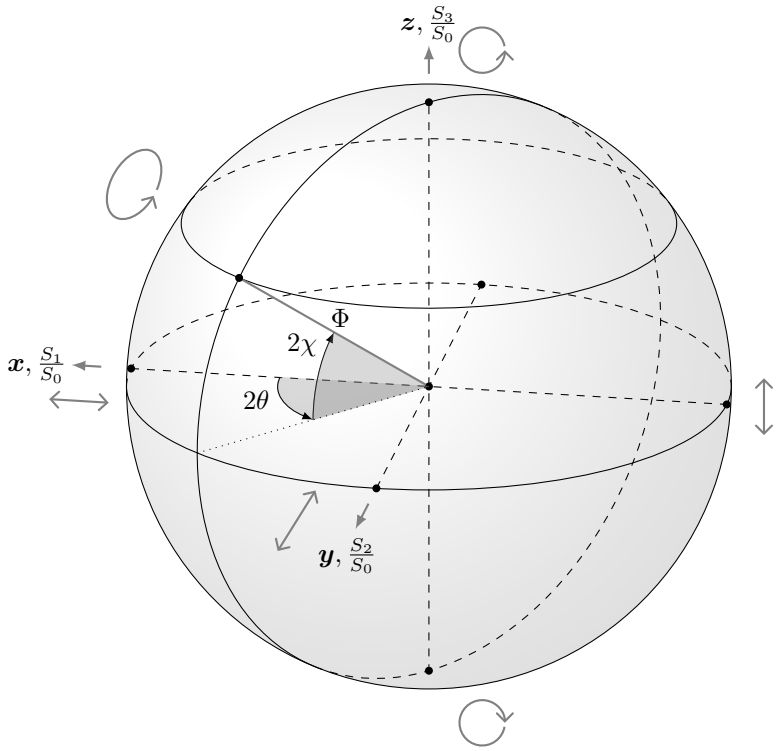


Fig. 1.5. – On the Poincare sphere, a polarization ellipse of parameters (θ, χ) is parametrized with the spherical coordinates $(2\theta, 2\chi)$. The cartesian coordinates of the polarization state are the normalized Stokes parameters. The circles denote counter-clockwise or clockwise rotation of the signal. (Source: Julien Flamant)

Two states close on the Poincaré sphere are geometrically close when drawn as ellipses. The Poincaré sphere provides a way to interpret the action of an optical device on polarized light as a rotation of states in an intuitive geometrical view **[Brosseau and Barakat, 1991]**. The

[2]For deterministic signals, the polarization is uniquely characterized by the description of the ellipse drawn in the plane. For stochastic signals however, variability in realizations yield the notion of degree of polarization: a random signal can be more or less polarized depending on the variance of the polarization ellipse between realizations. For instance, a fully polarized signal would give at each realization the same ellipse up to a scaling factor. A completely unpolarized signal would give at each realization a random ellipse, the point-wise summation of all the ellipses in the plane eventually collapsing to the origin. Polarization degree is between its minimum (unpolarized) and maximum (fully polarized) if a tendency towards one state can be observed, e.g the ellipse is more often described in the trigonometric sense than not.

Poincaré sphere is cut in two halves, the upper half for polarization that are described counter-clockwise, and the bottom half for clockwise polar states. Two polarization states antipodal on the sphere are said to be orthogonal, we will see that this is due to the nature of the transformation that sends one point on its antipodal counterpart. Polarization states described by antipodal points are similar in shape but are described in opposite directions, and the main axis of their respective ellipses are orthogonals. In the coordinate system (S_1, S_2, S_3), coordinate $(0, 0, 1)$ is the circular polarization: the ellipse is a circle. Consequently, at coordinate $(0, 0, -1)$ the polarization is also circular with an adverse rotation direction. At $(1, 0, 0)$ the polarization ellipse collapses to a horizontal line which is said to be a linear polarization. table 1.1 shows the relations between the three possible systems of parameters

Temporal description	Geometrical parameters	Stokes parameters
a_1, a_2, ϕ_1, ϕ_2	θ, χ, a	S_0, S_1, S_2, S_3
$a_1^2 + a_2^2$	a	S_0
$\frac{a_1^2 - a_2^2}{a_1^2 + a_2^2}$	$\cos 2\chi \cos 2\theta$	S_1/S_0
$\frac{2a_1a_2}{a_1^2 - a_2^2} \cos(\phi_2 - \phi_1)$	$\tan 2\theta$	S_2/S_1
$\frac{2a_1a_2}{a_1^2 + a_2^2} \sin(\phi_2 - \phi_1)$	$\sin 2\chi$	S_3/S_0

Tab. 1.1. – Relations between the different parametrizations of a polarization ellipse traced by a monochromatic bivariate signal **[Brosseau, 1998]**.

used to describe polarization.

1.2.3 Importance of the representation in the polarization analysis

It has been shown earlier in the document that polarization of bivariate signals can be described by different sets of parameters related by closed-form expressions. In this section we want to highlight the fact that polarization representation can be directly related to the choice of representation for the signal itself. A vector representation of the signal tends to develop into an algebraic description of polarization and polarizers and to the definition of the Stokes parameters while the Stokes parameters do not naturally appear in the complex approach which relies on a decomposition on circular orthogonal states.

The Jones vector and the coherency matrix

In 1942, Jones introduced a formalism in which the action of simple optical elements on a beam of light could be algebraically described using only linear operations. The Jones

calculus relies on the introduction of a Jones vector that represents light in terms of intensities and phasis. It is obtained as a complexification of the vector of intensities. If

$$\boldsymbol{x}(t) = (a_1 \cos(2\pi\nu t + \phi_1), a_2 \cos(2\pi\nu t + \phi_2))^T$$

is the monochromatic bivariate vector representing the intensity of the beam of light at each instant, its Jones vector is the complexification

$$\varepsilon = \begin{pmatrix} a_1 e^{\boldsymbol{i}(2\pi\nu t+\phi_1)} \\ a_2 e^{\boldsymbol{i}(2\pi\nu t+\phi_2)} \end{pmatrix} = \mathbf{e^{i2\pi\nu t}} \begin{pmatrix} a_1 e^{\boldsymbol{i}\phi_1} \\ a_2 e^{\boldsymbol{i}\phi_2} \end{pmatrix}. \tag{1.3}$$

From the Jones vector, a coherency matrix can be computed as:

$$C_x = \varepsilon\varepsilon^\dagger = \begin{pmatrix} a_1^2 & a_1 a_2 e^{-\boldsymbol{i}(\phi_2-\phi_1)} \\ a_1 a_2 e^{-\boldsymbol{i}(\phi_2-\phi_1)} & a_2^2 \end{pmatrix} \tag{1.4}$$

where † stands for hermitian transpose. By construction the coherency matrix belongs to the set of hermitian positive matrices, a sub-vector space of dimension 4 in $\mathrm{Mat}_2(\mathbb{C})$. Interestingly, this set admits as a basis the Pauli matrices $(\sigma_0, \sigma_1, \sigma_2, \sigma_3)$ a widely known quaternion of matrix in quantum physics:

$$\sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_2 = \begin{pmatrix} 0 & -\boldsymbol{i} \\ \boldsymbol{i} & 0 \end{pmatrix} \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{1.5}$$

The basis becomes orthonormal for the Frobenius inner product if each matrix is scaled by a factor one half. The coordinates of the coherency matrices over the orthonormal Pauli basis are exactly the Stokes parameters:

$$\begin{aligned} C_x &= \frac{a_1^2 + a_2^2}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{a_1^2 - a_2^2}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + a_1 a_2 \cos(\phi_2 - \phi_1) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + a_1 a_2 \sin(\phi_2 - \phi_1) \begin{pmatrix} 0 & -\boldsymbol{i} \\ \boldsymbol{i} & 0 \end{pmatrix} \\ C_x &= S_0 \frac{\sigma_0}{2} + S_1 \frac{\sigma_3}{2} + S_2 \frac{\sigma_1}{2} + S_3 \frac{\sigma_2}{2} \end{aligned} \tag{1.6}$$

As pointed out by **[Aiello and Woerdman, 2004]**, there has been different notational conventions in the optics community for the Pauli matrices. This explains inconsistency between the index of the Stokes parameters and index of the Pauli matrices in (1.6). In **[Gil, 2007]** for instance, the author applies a notation that verifies $C_x = \sum_{i=0}^{3} S_i \sigma_i$. We have decided however to use the convention of physicists for the Pauli matrices and to keep the Stokes parameters as they are defined in most optics papers. Inconvenient as it may seem, it makes crossing mathematical, physical, and optical points of view easier.
The relations that go from $\boldsymbol{x}$ to ε to C_x and finally to the Stokes parameters have no counterpart in $\mathbb{C}$. Indeed, the operation that built the Jones vector from the temporal vector and transformed it into the coherency matrix as in (1.3) and (1.4) used operations not available in the complex field.

Effect of an invertible transformation on the Jones vector and the coherency matrix

We underlined earlier that orthogonal polarization states are represented by antipodal states on the Poincaré sphere. See that their corresponding Jones vectors are indeed orthogonal. For instance, below are the Jones vector for respectively a horizontal linear polarized wave, a vertical linear wave, a right circular wave and a left circular wave:

$$\varepsilon^{0}_{\text{lin}} = \begin{pmatrix} ae^{i\phi} \\ 0 \end{pmatrix} \quad \varepsilon^{\frac{\pi}{2}}_{\text{lin}} = \begin{pmatrix} 0 \\ ae^{i\phi} \end{pmatrix} \quad \varepsilon^{r}_{\text{circ}} = \begin{pmatrix} ae^{i\phi} \\ -iae^{i\phi} \end{pmatrix} \quad \varepsilon^{l}_{\text{circ}} = \begin{pmatrix} ae^{i\phi} \\ iae^{i\phi} \end{pmatrix}$$

The common factor with the temporal dependency has been forgotten for clarity of presentation. The fact that antipodal points of the Poincare sphere have orthogonal Jones vectors bring to attention the fact that the Stokes vector (S_1, S_2, S_3) does not transform as could be expected under a change of frame. The special linear group $\text{GL}_2(\mathbb{R})$ represents all the possible change of frames in $\mathbb{R}^2$. Let $T \in \text{GL}_2$, under its polar decomposition, T writes as the product of a reflection $\mathcal{E}$, a rotation R_θ and a symmetric positive definite matrix P: $T = \mathcal{E}R_\theta P$. Geometrically, $\mathcal{E}$ changes the direction of some axis in the frame, T rotates it by the angle θ and the matrix P scales the all the space around the directions defined by its eigenvectors. To simplify our example, we focus on an orthogonal change of frame $T = \mathcal{E}R_\theta$. The coordinates of the signal x in this new frame is given by $x' = T^{-1}x$ and $T^{-1} = R^*_\theta \mathcal{E}^*$. Then the Jones vector and the coherency matrix in the new coordinates are:

$$\varepsilon' = T^{-1}\varepsilon \qquad\qquad C_{x'} = T^{-1}C_x(T^{-1})^* = R^*_\theta \mathcal{E}^* C_x \mathcal{E} R_\theta \tag{1.7}$$

The change of frame formula for the coherency matrix C_x has the sandwich shape $G \mapsto GC_xG^*$ inherited from the relation between the Jones vector and the coherency matrix. It is clear from here that the objects represented by the Jones vector and the coherency matrix are not affected similarly by the same transformations. Intuitively, we showed that they are elements of different natures. This idea is central to the approach of embedding a signal and its polarization analysis in one algebra, it shows that the algebra must be "big" enough to contain elements that are affected differently by linear operations. In particular, it is possible to show that it must be non-commutative. Finding such an embedding algebra is the purpose of section 1.4 where quaternions are proposed. In section 2.2 we will try to give a general understanding of the nature of the Stokes vector and the coherency matrix using the geometric algebras and introducing the notion of spinor.

Remark 2.3.2

There is not a one-to-one correspondance between Jones vectors and Stokes parameters. Considering two waves oscillating at the same frequency ν, they have the same set of Stokes parameters if the intensity of each component is the same and if the phase

difference between each component is the same. The absolute values of the phases however is not taken into account through the Stokes formalism. This is because Stokes are static, they describe a property of the signal that is invariant through time: its polarization ellipse. The instantaneous phase $\frac{\phi_1+\phi_2}{2}$ however, as its name implies, is an information on the position of the signal on its polarization ellipse at time $t = 0$. As a consequence, two Jones vectors ε and $\tilde{\varepsilon}$ related by a transformation of the type $\varepsilon = e^{i\phi} \cdot \mathrm{Id} \cdot \tilde{\varepsilon}$ give equivalent Stokes vectors. in fact, any transformation of this kind also cancels possible differences in frequencies, so that polarization is the part of the Jones vector that is invariant under the action of scalar matrices, isomorphic to $\mathbb{C}$. We will push further this definition of geometrical content through invariance in the last chapter of this part, where we aim at generalizing polarization analysis in n dimensions. We can already notice that matrices proportional to the identity that leave polarization unchanged are also the unitary matrices that commute with any matrix in $\mathrm{Mat}_2(\mathbb{C})$. It also becomes clear that representing polarization in $\mathbb{C}$, while it is defined as something invariant under product by complex numbers, can be problematic.

Mueller formalism

Mueller formalism is the description of a linear transformation in the basis of the Stokes vector. Given a vector v in $\mathbb{C}^2$, the Mueller form of a transformation M gives the coordinates of M in the basis defined by $\sigma_0, \sigma_1, \sigma_2, \sigma_3$ such that if

$$vv^* = aS_0 + bS_1 + cS_2 + dS_3 = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}^T_{S_0,S_1,S_2,S_3} \quad \text{then } M(v)M(v)^* = M \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}_{S_0,S_1,S_2,S_3}$$

Example:
The Mueller matrix of the transform that rotates a vector $v \in \mathbb{C}^2$ by an angle θ is :

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(2\theta) & \sin(2\theta) & 0 \\ 0 & -\sin(2\theta) & \cos(2\theta) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{1.8}$$

Again, we observe that a rotation of a vector in $\mathbb{C}^2$ does not yield a rotation of its Stokes parameters.

Remark 2.3.3

A polarizer is an optical filter that lets light of a specific polarization pass through while blocking other rays: it filters polarizations. Simple and heavy-used polarizers are linear polarizers with a specified polarization axis, right and left circular polarizers. Phase shifters are materials that introduce a phase shift between vertical and horizontal components of the light field thus changing the polarization. Contrarily to polarizers, phase shifters preserve the intensity of the light beam. The action of these optical instruments can be represented linearly in the Jones formalism, which is one of the major interest of Jones vector. This is called Jones calculus. As vector of $\mathbb{C}^2$, Jones vectors are acted upon by matrices in $\mathrm{Mat}_2(\mathbb{C})$, a class of matrices that allows both polarizers and phase retarders to be represented. For instance, a horizontal linear polarizer can be represented by

$$P^h_{lin} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

such that the Jones vector of the output of the linear polarizer reads:

$$\varepsilon_o = P^h_{lin}\varepsilon_i = e^{\boldsymbol{i}2\pi\nu t}\begin{pmatrix} a_1 e^{\boldsymbol{i}\phi_1} \\ 0 \end{pmatrix}$$

The Mueller form of the linear polarizer is:

$$P^h_{lin(S_0,S_1,S_2,S_3)} = \frac{1}{2}\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Asymetry of the Fourier transform for complex signals

We mentioned earlier that Jones calculus can not be performed in the complex algebra. We now show how the complex representation leads naturally to a different description of polarization, in terms of a mixture of conjugate states. Let $x(t) = a_1\cos(2\pi\nu_0 t + \phi_1) + \boldsymbol{i}a_2\cos(2\pi\nu_0 t + \phi_2)$ and $\boldsymbol{x}(t)$ its vector counterpart. Then, a Fourier-analysis on this signal gives

$$\begin{aligned} \mathcal{F}(x) &= X = \tfrac{1}{2}\left(\delta_{\nu_0}(a_1 e^{\boldsymbol{i}\phi_1} + \boldsymbol{i}a_2 e^{\boldsymbol{i}\phi_2}) + \delta_{-\nu_0}(a_1 e^{-\boldsymbol{i}\phi_1} + \boldsymbol{i}a_2 e^{-\boldsymbol{i}\phi_2}\right) \\ \mathcal{F}(\boldsymbol{x}) &= \boldsymbol{X} = \tfrac{1}{2}\begin{pmatrix} a_1\delta_{\nu_0}e^{\boldsymbol{i}\phi_1} + a_1\delta_{-\nu_0}e^{-\boldsymbol{i}\phi_1} & a_2\delta_{\nu_0}e^{\boldsymbol{i}\phi_2} + a_2\delta_{-\nu_0}e^{-\boldsymbol{i}\phi_2} \end{pmatrix}^T \end{aligned} \tag{1.9}$$

Observe that there are no symmetry between the positive and negative frequencies of the complex-Fourier transform. If both positive and negative frequencies are informative, it

raises the question of the physical interpretation of negative frequencies. The answer can be found through the concept of rotary components **[Schreier and Scharf, 2010]**. It emerges by reconstructing the signal x by inverse Fourier transform of its positive frequencies in one component and negative frequencies in another component. Each component is called a "phasor":

$$x(t) = \underbrace{a_+ e^{i\theta_+} e^{-i2\pi\nu_0 t}}_{\mathcal{F}^{-1}(\delta_{\nu_0}(a_1 e^{i\phi_1} + i a_2 e^{i\phi_2}))} + \underbrace{a_- e^{i\theta^-} e^{i2\pi\nu_0 t}}_{\mathcal{F}^{-1}(\delta_{-\nu_0}(a_1 e^{-i\phi_1} + i a_2 e^{-i\phi_2}))}$$

where $a_+, a_- \geqslant 0$ and $\theta_+, \theta_- \in [0, 2\pi[$, the amplitude and phase of each *phasors* are the *rotary coefficients*. The first term has positive frequencies only, it is the counter-clockwise phasor, the second-term has negative frequencies only and it's the clockwise phasor. Clockwise and counter-clockwise refer to the direction of rotation of the phasor: $e^{i2\pi\nu_0 t}$ rotates counter-clockwise while $e^{-i2\pi\nu_0 t}$ rotates clockwise. Hence, the frequency signs have an interpretation in terms of the direction of rotation of a circularly polarized signal, while its positive value has an intepretention in terms of repetition of a pattern. This description of the signal naturally leads to a decomposition of the polarization over two orthogonal states: circular signals of opposed senses of rotation. In oceanographic studies, rotary components have been proved useful to describe the currents or inertial motions that have rotary characteristics with a preferred direction. This preference of rotary components over a linear decomposition of the polarization by oceanographers **[Thomson and Emery, 2014]** can be related to the physical caracteristics of marine signals: vortex and rotary phenomenon are common in ocean studies. In light of the previous explanation, it can also be argued that the complex representation in itself influences the study of polarization towards a description in terms of phasors, while the vector description orientates it towards a linear decomposition of polarization. The formalism retained in different discipline (oceanograpy and geophysics here) appears to closely interact with the way signals are described and considered in each discipline.

Polarization analysis for complex signals

Description of a polarization state in the complex framework can rely either on the direct temporal parameters (a_1, a_2) and (ϕ_1, ϕ_2) respectively the amplitudes and phase of the real and complex parts of the signal. Relations between these parameters and the geometry of the polarization ellipse have been described in table 1.1. Alternatively, the decomposition of the signal in terms of phasors provides four new quantities (a_+, a_-) and (θ_+, θ_-) which read:

Phasor parameters	Geometrical parameters	Geometrical description
$a_+ + a_-$	α	length of the major axis
$\lvert a_+ - a_- \rvert$	β	length of the minor axis
$\theta_+ - \theta_-$	2θ	orientation of the ellipse

This shows the simplicity in interpretating the geometrical impact of the rotary components: if amplitudes a_+, a_- are similar then the polarization tends to be linear. A non-horizontal orientation comes from a difference between θ_+ and θ_-.

1.2.4 Conclusion

We have discussed polarization for monochromatic signals, a class of limited existence in the real world. To broaden the scope of this description of polarization, we should first note that the concept of polarization is, just as the concept of frequency, ultimately only applicable to monochromatic signals. However, just as frequency analysis decomposes signals as a continuous sum of monochromatic contributions, polarization analysis can be conducted frequency-wise. In which case, at each frequency, a set of parameters describes the polarization. If components of the signal evolve smoothly with time, then polarization will be a smooth function of frequency. Another way to define and study polarization is through the notion of *instantaneous polarization* for modulated oscillations. Instantaneous polarization is related to polarization the same way *instantaneous frequencies* are related to the physical notion of frequencies **[Picinbono, 1997]**. While they are not the same object, they are seen as frequencies and polarization evolving through time. This is first counter-intuitive as their definition relies on the repetition of a pattern similar to itself, but it can be shown that the concept has a useful and well-defined physical interpretation for special class of signals, that are called "modulated oscillations" **[Lilly and Olhede, 2009]**.

1.3 Non-stationary polarization

The spectral analysis based on the Fourier transform dispatches the energy of the signal over monochromatic components, and the signal is seen as a weighted sum of the contributions at each frequency. In particular, this vision integrates information from each instant of the signal and proposes a description that is the same for each of these instants. This is particularly well-adpated to signals whose frequency content do not vary with time, called stationary signals. Although "stationary" is usually used to discuss statistical properties of stochastic time processes, it is used here in a deterministic context in the same sense as **[Picinbono, 1997]** or **[Boashash, 1992]**.

Remark 3.0.1

Note that the definitions of stationarity in a stochastic and deterministic context are not consistent. A stochastic process is stationary if all orders of a random variable do not vary with time. The building brick of stationary deterministic signals is the sine, whose instantaneous mean values equals itself and varies from 1 to -1. A deterministic sine is not a stationary random process. It is a stationary in the deterministic sense because

on whatever window of time the process is looked at, it is always best described by the same sine which is itself. This translates in the Fourier plane by a single ray above the frequency of the sine.

A deterministic stationary signal can be seen as weighted combination of sine and cosine where the weights are independent of the window of observation. In a time-frequency representation such as a spectrogram, this would give a constant frequency content over time. Of course, not all signals can be satisfactorily described with a stationary model. Consider for instance the chirp (1.10), a frequency modulated signal.

$$x(t) = a\cos\left(2\pi\frac{t}{10}^2\right) \tag{1.10}$$

Its spectral analysis would give give contributions of every frequencies at every time, while an analysis conducted in a window tightened around t, would give a spectral ray at $\frac{t}{5}$. Such signals are not well disposed to be decomposed over a stationary family of signals, and several alternatives have been proposed for the processing of this particular class. Most of it revolves around the notion of "instantaneous frequency" and of frequency and amplitude modulated signal. Modulated signals, at the peak of their popularity, were the subject of many writings because of their importance in telecomunications but also because their interpretability was subject to caution. The notion of "instantaneous frequency" is apparently paradoxal, a frequency describes the repetition of a pattern during a certain unit of time, while instantaneous eliminates all possible duration. To understand the challenges raised by extending this definition to multivariate signals, it is important to address precisely how the instantaneous frequency was built for univariate signals. This is the purpose of the next section.

1.3.1 Non-stationary univariate signal and instantaneous frequency

The concept was first approached from practical examples, a model for the modulated signal was **[Van der Pol, 1946]**

$$a(t)\cos(2\pi\phi(t)) \tag{1.11}$$

and the instantaneous frequency was defined by $\nu(t) = \frac{\partial\phi}{\partial t}$. This decomposition into an instantaneous amplitude $a(t)$ and an instantaneous phase $\phi(t)$ is problematic, as for a signal $x(t) = a(t)\cos(\phi(t))$ the pair $(a(t), \phi(t))$ is not canonical or unique. Consider for instance the signal:

$$x(t) = \cos(2\pi\phi_1(t))\cos(2\pi\phi_2(t))$$

Among the two cosine we can only decide arbitrarily which one plays the role of the phase and which one of the amplitude. Gabor in **[Gabor, 1946]** proposes a way to always define univocally a pair of phase and amplitude from a real signal. It is based on the construction of

a complex embedding of the real signal. The analytic signal associated with the real signal $x(t)$ is:

$$x_+(t) = x(t) + i\mathcal{H}[x(t)] \tag{1.12}$$

where $\mathcal{H}[\cdot]$ is the Hilbert transform defined as:

$$H[x(t)] = \text{p.v} \int_{-\infty}^{+\infty} \frac{x(t-\tau)}{\pi\tau} d\tau \tag{1.13}$$

and where p.v denotes the Cauchy principal value of integral (1.13). The spectral equivalent of (1.12) is the suppression of all the energy contained in the negative frequencies, and the doubling of positive frequencies. This gives:

$$\mathcal{F}[x_+] = 2\delta_{\nu \geqslant 0}\mathcal{F}[x]$$

The cutting of negative frequencies is interesting from a physical point of view where, for univariate signals, only positive frequencies are directly interpretable[3]. Once the complex signal $x_+(t)$ is defined, there exists a unique factorization under the shape

$$x_+(t) = a(t)e^{i2\pi\phi(t)}.$$

Thus, the analytic signal defines a unique amplitude-phase pair for any real signal x, and projecting x_+ to the real set gives:

$$x(t) = \Re[x_+(t)] = a(t)\cos(2\pi\phi(t)) \tag{1.14}$$

Remark 3.1.2

This result is for us the most significant. In a way, the definition of instantaneous amplitude and frequencies is entirely resolved by embedding the signal in an algebra that posses an injective way to attribute a pair amplitude-frequency to a real signal. Of course, not any injective definition would have yielded a useful result, and what follows of this section shows that this choice is not random and verifies some properties. We want to show in what follows that the definition of instantaneous polarization also relies on the construction of an injective application from the signal itself or an object built from the signal to a n-uplet of parameters to which we give meaning only by comparison with the monochromatic case. Hence, embedding our signal in a space where an obvious such injective transformation can be built will be one of the motives behind the work in section 1.4 and chapter 3.

[3]We already remarked that for bivariate signals negative frequencies have an interpretation in terms of clockwise rotating phasors, but univariate signals can only go one direction

The analytic signal provides a way to build a unique phase amplitude pair for any real signal x. Now, suppose a signal x is given under the shape (1.11). What are the conditions for the amplitude-phase pair to be the same as the pair defined by the analytic signal?
What if we have an a priori decomposition of a signal that makes sense for physical reason, will it match the decomposition proposed by equation (1.14)? In other terms, the question is to know when does the analytic signal matches an a priori decomposition of our signal. More precisely, as it was framed by **[Picinbono, 1997]**: what are the conditions on $[a(t), \phi(t)]$ so that the analytic signal of $x(t) = a(t)\cos(\phi(t))$ is defined by $x_+(t) = a(t)e^{i2\pi\phi(t)}$. This automatically translates to a condition on the Hilbert transform to get $\mathcal{H}[a(t)\cos(2\pi\phi(t))] = a(t)\sin(2\pi\phi(t))$ it is sufficient (but not necessary) that $\mathcal{H}[a(t)\cos(2\pi\phi(t))] = a(t)\mathcal{H}[\cos(2\pi\phi(t))]$ and $\mathcal{H}[\cos(2\pi\phi(t))] = \sin(2\pi\phi(t))$. The Bedrosian theorem gives condition to reach the first criteria.

Theorem 1 (Bedrosian theorem). *Let x, y be two signals and $z(t) = x(t)y(t)$ the signal defined as their product. If the Fourier transform $X(\nu)$ and $Y(\nu)$ are such that there exists a constant M with $|X(\nu)| = 0$ for $|\nu| > M$ and $Y(\nu) = 0$ for $|\nu| < M$ then the Hilbert transform of z reads:*

$$\mathcal{H}[z](t) = x(t)\mathcal{H}[y](t)$$

If the spectrum of a and $\cos(2\pi\phi)$ do not overlap with a occupying the low-frequency region and $\cos(2\pi\phi)$ the high frequencies, the Hilbert transform of their product simplifies. If moreover it is verified that $\mathcal{H}[cos(2\pi\phi(t))] = \sin(2\pi\phi(t)]$ then $a(t), \phi(t)$ are the instantaneous amplitude and phase of $x(t)$. However, the second condition is rarely verified in practice. For instance the chirp $t \mapsto \cos(\alpha\pi t^2)$ does not meet this requirement and its analytic signal slightly differs from $e^{i\alpha\pi t^2}$. However, under the Bedrosian conditions, it is generally accepted that the difference between the would-be instantaneous phase and the real one (defined by the Hilbert transform) is small and goes to zero as t becomes large. This is why it is usually considered that the instantaneous frequency is interpretable whenever the variations of the phase are significantly faster than the variations of the amplitude $\frac{da}{dt} << \frac{d\phi}{dt}$ and the subtelty of the definition of the instantaneous phase is rarely discussed.

Even when the Bedrosian conditions are not met, the physical interpretability of the notion of instantaneous frequency is supported by physical considerations. First, Ville noticed that the time average of the instantaneous frequency corresponds to the average "frequency" of the signal **[Ville, 1948]**. For $t \mapsto x(t)$ a signal and $t \mapsto \nu(t)$ its instantaneous frequency, the following applies:

$$\frac{\int_{-\infty}^{+\infty} \nu(t)|x_+(t)|^2 dt}{\int_{-\infty}^{+\infty} |x_+(t)|^2 dt} = \frac{\int_{-\infty}^{+\infty} \xi|X(\xi)|^2 d\xi}{\int_{-\infty}^{+\infty} |X(\xi)|^2 d\xi} \tag{1.15}$$

Equation (1.15) establishes the equality between the average frequency computed as the first moment of the instantaneous frequency or the first moment of the Fourier frequency. It implies a relation exists between the classical notion of frequency and instantaneous frequency, hence the latter can not be completely without meaning, as the first is not. The second consideration

helping with the interpretation of the instantaneous frequency was provided by the stationary phase principle. The spectrum of the analytic signal $x_+(t)$ is given by

$$\begin{aligned} X_+(\nu) &= \int_{-\infty}^{+\infty} a(t) e^{i2\pi\phi(t)} e^{-i2\pi\nu t} dt \\ &= \int_{-\infty}^{+\infty} a(t) e^{i2\pi(\phi(t)-\nu t)} \end{aligned} \tag{1.16}$$

The application of the stationary phase principle states that integral (1.16) reaches its highest values around stationary points of the phase functions $\zeta(t) = 2\pi(\phi(t) - \nu t)$. A stationary point would be a frequency ν that satisfied:

$$\frac{d\phi}{dt} - \nu = 0 \tag{1.17}$$

If the signal is not modulated and $\frac{d\phi}{dt} = \nu_0$ is independent of t, we find that the integral is only defined by what happens at frequency ν_0. This is indeed the only value of ν for which the integral does not cancel. For a modulated signal however, equation (1.17) is satisfied for $\nu(t) = \frac{d\phi}{dt}$ which depends on t. This shows the concentration of frequencies at a single frequency $\nu(t)$ around the time t.

Remark 3.1.3

As it was remarked in **[Flandrin, 2012]**, the notion of "instantaneous frequence" might first appear problematic as it seems to contradict Gabor uncertainty principle. Indeed, the principle assesses that it is impossible given a deterministic signal to access the frequency with an arbitrary precision within a time-window of an arbitrary small size. This apparent contradiction is resolved by noticing that even though the instantaneous frequency seems perfectly localized in its definition it is not the case in its computation. Indeed, the Hilbert transform used in its description is an integral operator whose kernel not only has an infinite extension, meaning it covers all instants of the signal, but also has a slow decay which confers importance to every instants of the signal, even those apparently far away from the point of interest. In light of this precision, it is necessary to separate the notions of instantaneity and locality. For this very reason, the stationary phase argument presented just before and which is used in **[Rihaczek, 1968]** and **[Boashash, 1992]** is unsatisfactory.

The main point that must be carried with us in the next section is that the possibility to define a canonical pair of instantaneous amplitude and phase comes from the possibility to associate a unique meaningful complex signal to a real one, **and** to the unicity of the polar decomposition in the complex plane. It will be demonstrated in section section 1.3.3 and section 3.2.3 that

the unicity of the decomposition is not guaranteed in higher dimensions. This leads to the necessity of further justifications to uniquely define an instantaneous frequency for modulated signals.

1.3.2 Polarization of analytic signals

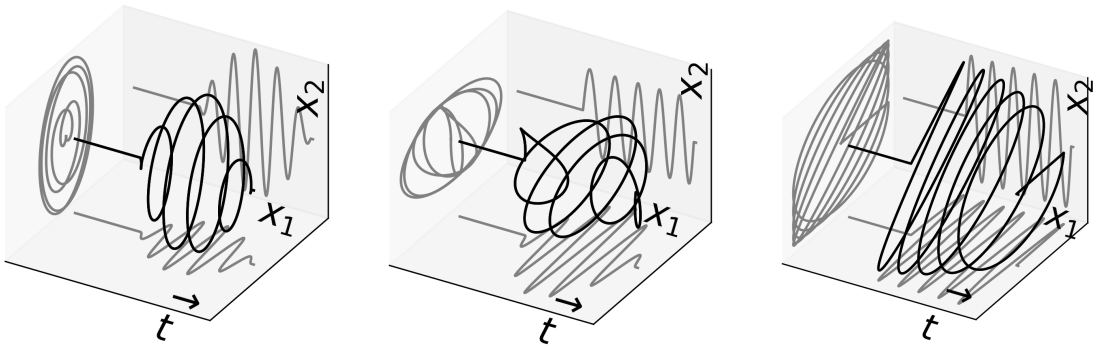
We have seen earlier how analytic signals can be defined as particular complex embeddings of real signals. They can also be defined intrinsically as signals with no negative frequencies. Then, they necessarily write as the sum of a real signal plus the imaginary unit times its Hilbert transform. For this reason, polarization properties of analytic signals are very constrained. For an analytic signal $x_+(t)$, its Fourier transform writes:

$$x_+(t) = \int_0^\infty X_+(\nu) e^{-i2\pi\nu t} d\nu$$

the contribution at a single frequency $\nu > 0$ is $X_+(\nu)e^{-i2\pi\nu t}$. It is a counter-clockwise rotary component. As the analytic signal writes as a continuous sum of such components, it can be deduced that it is circularly counter-clockwised polarized. This shows that arbitrarily polarized bivariate signals can not be written as complex analytic signals.

1.3.3 Instantaneous parameters for modulated bivariate oscillations

Constructing the analytic signal for bivariate time series relies on the construction of a pair of analytic signals. In **[Lilly and Olhede, 2009]** the authors name the cartesian pair and the rotary pair. This refers of course to the two possible representations of a bivariate signal. For a complex representation of the signal, we have seen earlier that separating the positive frequencies and negative frequencies contributions yielded a pair of rotary components. The remark in the previous paragraph showed that rotary components are a two-fold cover on analytic signals. The clockwise component is analytic and the counter-clockwise component is not analytic but an inversion of time sends it on an analytic signal. In the non-stationary case, each component contains several frequencies. Under some conditions each rotary component can be seen as an analytic signal. We do not present this work here, and focus on vector-valued signals. For a vector-valued signal, an analytic vector can be computed component-wise. The pair of components is called the cartesian pair of analytic signal. The notion of modulated ellipse, where a bivariate signal can indeed be seen as a modulated elliptical signal evolving through time is developped in **[Lilly and Olhede, 2009]**. Examples of modulated bivariate oscillations are provided in figure 1.6. The authors in **[Lilly and Olhede, 2009]** provide a way to build an instantaneous frequency that we do not reproduce here. Instead, we will start from considerations on invariance to justify rather than build the instantaneous frequency. Remember that for a monochromatic signal with a Jones vector



(a) strong amplitude modulation, slow ellipticity modulation

(b) strong ellipticity modulation

(c) slow ellipticity modulation

Fig. 1.6. – Three modulated bivariate oscillations.

$$\boldsymbol{x}_+(t) = e^{\boldsymbol{i}2\pi\nu t}\begin{pmatrix} a_1 e^{\boldsymbol{i}\phi_1} \\ a_2 e^{\boldsymbol{i}\phi_2} \end{pmatrix}$$

The coherency matrix reads:

$$C_x = \begin{pmatrix} a_1^2 & a_1 a_2 e^{-\boldsymbol{i}(\phi_2-\phi_1)} \\ a_1 a_2 e^{-\boldsymbol{i}(\phi_2-\phi_1)} & a_2^2 \end{pmatrix}$$

Note that the time-dependent term $e^{\boldsymbol{i}2\pi\nu t}$ has disappeared from the expression. Clearly, the frequency content is invariant under the action of $U(n)$. Indeed, the product of the Jones vector with a matrix of $U(n)$ is equivalent to the sandwich-product of the coherency matrix with the same matrix. The frequency content being absent from the coherency matrix, it can not be affected either in the Jones vector. Given a non-monochromatic signal $\boldsymbol{y}$, we built its Jones vector by taking component-wise analytic signals. The Jones vector and the coherency matrix read:

$$\boldsymbol{y}_+(t) = \begin{pmatrix} a_1(t) e^{\boldsymbol{i}\phi_1(t)} \\ a_2(t) e^{\boldsymbol{i}\phi_2(t)} \end{pmatrix} \quad C_y = \begin{pmatrix} a_1(t)^2 & a_1(t)a_2(t) e^{-\boldsymbol{i}(\phi_2(t)-\phi_1(t))} \\ a_1(t)a_2(t) e^{-\boldsymbol{i}(\phi_2(t)-\phi_1(t)} & a_2(t)^2 \end{pmatrix} \tag{1.18}$$

To identify an instantaneous phase, we are basically looking for a quantity that disappeared, or can not be reconstructed from C_y. Another condition is that the instantaneous phase together with the coherency matrix must enable us to reconstruct the Jones vector. One obvious candidate is:

$$\boldsymbol{\varphi}(t) = \frac{\phi_1(t) + \phi_2(t)}{2}$$

The normalization by $\frac{1}{2}$ makes this definition consistent with the monochromatic case. Then the instantaneous frequency is the derivative of the instantaneous phase normalized by 2π $\boldsymbol{\nu}(t) = \frac{1}{2\pi}\boldsymbol{\varphi}'(t)$. This clumsy argumentation is not an in-depth study of the notion of bivariate instantaneous frequency. It is more a way to have an insight into why $\boldsymbol{\varphi}(t) = \frac{\phi_1(t)+\phi_2(t)}{2}$ seems

to be a good definition of the instantaneous phase of a bivariate signal. It was shown in **[Lilly and Olhede, 2009]** that this instantaneous frequency satisfies the moment equation:

$$\frac{\int_{-\infty}^{+\infty} \boldsymbol{\nu}(t)\|\boldsymbol{y}_+(t)\|^2 dt}{\int_{-\infty}^{+\infty} \|\boldsymbol{y}_+(t)\|^2 dt} = \frac{\int_{-\infty}^{+\infty} \xi\|\boldsymbol{Y}(\xi)\| d\xi}{\int_{-\infty}^{+\infty} \|\boldsymbol{Y}(\xi)\|^2}$$

More precise developpments around the idea of building quantities from invariances can be found in section 3.3.

Time-dependent ellipse parameters

To build time-dependent ellipse-parameters, we will use the unicity of the coordinates of the coherency matrix over the Pauli basis. For each t, this leads to the definition of time-dependent Stokes parameters $S_i(t)$. Then, we show that the relations in table 1.1 are invertible and we use them to build instantaneous ellipse angles from the Stokes parameters.

The pair of analytic signals provides four real-functions that contain all the geometrical and spectral information on the signal. Instantaneous polarization can be seen as a generalization of the instantaneous amplitude for the bivariate case. The goal here is to construct the four geometrical parameters $(a(t), \chi(t), \theta(t), \phi(t))$ from $(a_1(t), a_2(t), \phi_1(t), \phi_2(t))$. We need a bijection from

$$\mathbb{R}_+ \times \mathbb{R}_+ \times [-\pi, \pi] \times]-\pi, \pi] \quad \text{to} \quad \mathbb{R}_+ \times [0, \frac{\pi}{4}[\times [0, \pi[\times [-\pi, \pi[.$$

To build this bijection with the cartesian coordinates, we take an indirect route through the Stokes parameters. Let $\boldsymbol{x}(t)$ and $\boldsymbol{x}_+(t)$ be a bivariate signal and its Jones vector as defined in (1.18). The coordinates of the time-dependent coherency matrix built in (1.18) on the Pauli basis are:

$$C_x = S_0(t)\sigma_0 + S_2(t)\sigma_1 + S_3(t)\sigma_2 + S_1(t)\sigma_3 \tag{1.19}$$

with

$$\begin{array}{lcllcl} S_0(t) & = & a_1^2(t) + a_2^2(t) & S_1(t) & = & a_1^2(t) - a_2^2(t) \\ S_2(t) & = & 2a_1(t)a_2(t)\cos(\phi_1(t) - \phi_2(t)) & S_3(t) & = & 2a_1(t)a_2(t)\sin(\phi_2(t) - \phi_1(t)) \end{array}$$

From the Stokes parameters, it is possible to reconstruct geometrical parameters $(a, \theta, \chi) \in \mathbb{R}^+ \times [0, \pi[\times [-\pi/2, \pi/2[$ using the inverse operations of 1.1.

$$\begin{array}{lcllcl} a(t) & = & S_0(t) & = & a_1^2(t) + a_2^2(t) \\ \theta(t) & = & \frac{1}{2}\arctan\frac{S_2(t)}{S_1(t)} + \frac{\pi}{2} & = & \frac{1}{2}\arctan\frac{2a_1(t)a_2(t)}{a_1^2(t) - a_2^2(t)}\cos(\phi_2(t) - \phi_1(t)) + \frac{\pi}{2} \\ \chi(t) & = & \frac{1}{2}\arcsin\frac{S_3(t)}{S_0(t)} & = & \frac{1}{2}\arcsin\frac{2a_1(t)a_2(t)}{a_1^2(t) + a_2^2(t)}\sin(\phi_2(t) - \phi_1(t)) \end{array}$$

This definition of instantaneous geometrical parameters is very indirect and not quite satisfying. In section 1.4 and section 3.2.3 we propose two different ways to get to the same definition by a more direct and intuitive way. Next section explains how embedding bivariate signals in the quaternion algebra leads to a direct definition of geometrical parameters. Later in section section 3.2 considering the action of invertible transformations over the space of $\mathbb{C}$-valued vectors will lead to a similar result.

1.4 Quaternions: a unifying frame

Embedding bivariate signals in a bigger algebra provides an elegant way to build instantaneous geometrical parameters. This approach, based an a polar factorization of a quaternion number, is not without recalling the definition of the analytic pair of a signal from the Euler form in the complex plane. The beauty of this framework resides in a large part in that it prolongates the special relation real and complex signals share in signal analysis to a relation between complex and quaternionic signals. The quaternion and complex fields display many symetries and similarities, so that the relation between complex and quaternionic signals can often be intuitively derived from knowledge about the real/complex case.

1.4.1 Hamilton and its "Elements of Quaternions" **[Hamilton, 1866]**

Nothing can enlighten us more on the use and purpose of quaternions than the original writings[4] of their inventor [5] William R. Hamilton. The "Elements of Quaternions" was actually published by his son, after Hamilton's death but was in a final shape enough at this time to be a prime source material. When Hamilton started his work with vectors in the three-dimensional space, the concept of two-dimensonal vectors was quite recent, even though cartesian coordinates had more than 200 years. Their algebra, additive properties, scalar multiplication, Chasles-relations were all well-understood for the plane or the space. It is worth noting that bidimensional vector had appeared as a geometrical interpretation of the complex algebra. It is then naturally that Hamilton was trying to find the algebra congruent with three-dimensional vectors. After failing to find any three-dimensional algebra satisfying his conditions, he had the insight that a fourth dimension was necessary, through his concept of the ratio of vectors. It was clear to him that the ratio of a vector u by a vector v had to be interpreted as the action that performed on u should give v as he says page 106 of his Elements:

"*It is evident that the supposed operation of division (whatever its full geometrical import may afterwards be found to be), by which we here conceive ourselves to pass from a given divisor line alpha, and from a given dividend-line beta to what we have called provisionally) their geometric quotient q, may (or rather must) be conceived to correspond to some converse act (as yet not*

[4] The book can be consulted freely at this link.
[5] Or discoverer, we are not opposed to a Heideggerian conception of truth.

fully known) of geometrical multiplication : in which new act the former quotient q becomes a FACTOR and operates on the line alpha, so as to produce (or generate) the line beta. We shall therefore write, as in algebra,

$$\beta = q \cdot \alpha \text{ or simply, } \beta = q\alpha \text{ when } \frac{\beta}{\alpha} = q$$

"

Remark 4.1.1

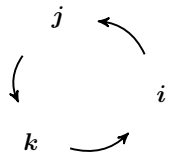
In this quote Hamilton is describing a sort of relation that we would today call a group action, but the abstract notion of group was in progress, and anachronical at the time. The modern reader would probably be tempted to solve this geometrical problem with matrices, and in particular the special linear group $GL(\mathbb{R}^3)$ which encodes all the possible transformations of vectors in $\mathbb{R}^3$. Even though matrices were known and used to solve linear systems of equations by the time of Hamilton, their interpretation in terms of linear map was not popular, and he came with an original solution: the quaternions. Similarly to complex numbers for bi-dimensional vectors, the quaternions have the capacity to represent both three-dimensional vectors and transformations of these vectors.

Because the description of the transformation of a three-dimensional vector required at least three angles and a scaling factor, Hamilton eventually defined the Quaternion algebra with four distinct dimensions. It became clear after its definition that its special properties such as the invertibility of all elements, associativity of its laws etc.. made it a trail worth developping.

1.4.2 Quaternion algebra

Named $\mathbb{H}$ in tribute to Hamilton, the field of quaternion can be understood as an extension of the complex field, as it follows similar rules. It is the only four dimensional division ring over the real numbers, and is the next natural "step" after the complex field as no such three-dimensional algebra can be found. Just as $\mathbb{C}$ is a vector space for the directions 1 and $\boldsymbol{i}$, $\mathbb{H}$ has the four directions $1, \boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ where the two additional "axes" $\boldsymbol{j}$ and $\boldsymbol{k}$ satisfy $\boldsymbol{j}^2 = -1, \boldsymbol{k}^2 = -1$ and $\mathbb{H} = \{a + b\boldsymbol{i} + c\boldsymbol{j} + d\boldsymbol{k} \mid a, b, c, d \in \mathbb{R}\}$. Quaternions are non-commutative for the multiplication, an element that surprised Hamilton, and probably slowed him down in his research. They obey the following rules:

$$\begin{aligned} \boldsymbol{ij} &= -\boldsymbol{ji} &&= \boldsymbol{k} \\ \boldsymbol{jk} &= -\boldsymbol{kj} &&= \boldsymbol{i} \\ \boldsymbol{ki} &= -\boldsymbol{ik} &&= \boldsymbol{j} \\ \boldsymbol{ijk} &= &&-1 \end{aligned}$$



The cartesian coordinates of a quaternion are the real numbers (a, b, c, d) such that

$$z = a + \boldsymbol{i}b + \boldsymbol{j}c + \boldsymbol{k}d$$

and the norm of z is:

$$\|z\| = \sqrt{a^2 + b^2 + c^2 + d^2} \tag{1.20}$$

A pure quaternion $\boldsymbol{\mu}$ is a quaternion with a vanishing real part: $\boldsymbol{\mu} = b\boldsymbol{i} + c\boldsymbol{j} + q\boldsymbol{k}$. Pure quaternions form a subspace of dimension 3, and vectors of $\mathbb{R}^3$ can be embedded in the quaternion vector space through the isomorphism:

$$\begin{aligned} \text{emb}^{\mathbb{H}}_{\mathbb{R}^3} \quad : \mathbb{R}^3 \quad &\to \mathbb{H} \\ \begin{pmatrix} v_1 & v_2 & v_2 \end{pmatrix}^T \quad &\mapsto v_1\boldsymbol{i} + v_2\boldsymbol{j} + v_3\boldsymbol{k} \end{aligned} \tag{1.21}$$

This embedding also induces the embedding of some linear transformations of vectors into $\mathbb{H}$, see (1.24). Some elementary involutions **[Ell and Sangwine, 2005]** are particularly useful when working with quaternions, they are defined with respect to a pure unitary quaternion i.e a pure quaternion that has a norm 1. Given $z = a + b\boldsymbol{i} + c\boldsymbol{j} + d\boldsymbol{k}$

$$\begin{aligned} \bar{z} &= a - b\boldsymbol{i} - c\boldsymbol{j} - d\boldsymbol{k} && \textit{is the conjugate of } z \\ z^{*\boldsymbol{j}} &= a + b\boldsymbol{i} - c\boldsymbol{j} + d\boldsymbol{k} && \textit{is the involution of axis } \boldsymbol{j} \textit{ of } z \\ z^{*\boldsymbol{\mu}} &= -\boldsymbol{\mu}\bar{z}\boldsymbol{\mu} && \textit{is the involution of axis } \boldsymbol{\mu} \textit{ where } \boldsymbol{\mu} \textit{ is a pure unitary quaternion} \end{aligned}$$

Pure imaginary quaternions are sometimes referred to as "axis" as they can be visualized living on a two-dimensional sphere in a three-dimensional space, they define a direction in the three dimensional space.

Remark 4.2.2

The 2-dimensional sphere is not equivalent to the set of all directions in $\mathbb{R}^3$, as every direction in space can be represented by two antipodal points on the sphere. The set of all directions is the projective space $\mathbb{P}_2(\mathbb{R})$, it has the topology of the sphere quotiented by the relation that puts in the same class antipodal points. It can be visualized in first approximation as a half-sphere, but this representation is misleading as the topology is wrong: the real projective plane has no boundaries. The pure imaginary quaternions are homeomorphic to the sphere, not the real projective space.

The rules of the algebra being laid down, functions defined by power series can (under guarantees of convergence), be defined on the quaternion algebra. It is therefore possible to extend the trigonometric functions such as sine, cosine and exponential to the quaternions. A simple calculation shows that the exponential of a pure quaternion satisfies the relation:

$$e^q = \cos(\|q\|) + q\sin(\|q\|)$$

The following proposition follows:

Proposition 4. *Polar decomposition There is a unique polar decomposition of a quaternion $\boldsymbol{z}$ as:*

$$\boldsymbol{z} = |z|e^{\boldsymbol{\mu}\lambda}$$

where $\boldsymbol{\mu}$ is a pure unitary quaternion, and $\lambda \in [0, 2\pi[$. The Euler-polar form involves specific axes and reads ***[Bulow and Sommer, 2001]****:*

$$\boldsymbol{z} = |z|e^{\boldsymbol{i}\alpha}e^{-\boldsymbol{k}\beta}e^{\boldsymbol{j}\gamma} \tag{1.22}$$

where $\alpha \in [-\pi/2, \pi/2[$, $\beta \in [0, \pi/2]$, $\gamma \in [-\pi, \pi[$ are uniquely determined.

Quaternions have the ability to represent 3-dimensional rotations as follow: let $v \in \mathbb{R}^3$ be embedded in $\mathbb{H}$ by

$$\begin{pmatrix} v_1 \\ v_2 \\ v_2 \end{pmatrix} \leftrightarrow v_1\boldsymbol{i} + v_2\boldsymbol{j} + v_3\boldsymbol{k} \tag{1.23}$$

and q be a unitary quaternion with polar form $q = e^{\boldsymbol{\mu}\lambda}$ then the transformation

$$T_q \colon v \mapsto qvq^{-1} \tag{1.24}$$

is a rotation of v around the axis $\boldsymbol{\mu}$ (which is a pure quaternion) by the angle 2λ. Note that the analogy is possible because if v is a pure quaternions then $T_q(v)$ is also a pure quaternion, due to the particular shape of the "sandwich" product in its expression. As a pure quaternion, it can be embedded in $\mathbb{R}^3$ following (1.23) **[Vicci, 2001]**. The set of quaternions is said to "embedd" $\mathbb{C}$, as for any pure imaginary quaternion $\boldsymbol{\mu}$, the subspace $\mathbb{C}_{\boldsymbol{\mu}} = \{a + \boldsymbol{\mu} \mid a, b \in \mathbb{R}\}$ is similar from the point of view of any structure (linear, algebraic, topological...) to $\mathbb{C}$.

1.4.3 Embedding complex signals in the Quaternion Algebra

Representation of complex signals

In Hamilton's mind, the power of quaternions resided in the ability of a single algebra to contain both vectors and their transformations. This is certainly their most striking feature and they are still used today in control to represent systems involving calcultations of

three-dimensional orientations, such as robotics **[Pervin and Webb, 1982]** or flight control **[Wie and Barba, 1985] [Cooke et al., 1992]**. A more original contribution of quaternions can be found in the recent work of **[Flamant et al., 2016]** where quaternions are used to analyze the polarization of bivariate signals. The approach is original in the sense that quaternions have always historically been related to trivariate quantities, while bivariate signals were handled with complex numbers. Furthermore, the embedding of vectors in the quaternion algebra always uses the vector space structure of pure quaternions as in (1.21). In their work however, **[Flamant et al., 2016]** embed bivariate signals in the subspace $\mathbb{C}_{\boldsymbol{i}}$ of $\mathbb{H}$. The signal $\boldsymbol{x} = (x_1 \ x_2)^T \in \mathbb{R}^2$ gives $x = x_1 + \boldsymbol{i}x_2 \in \mathbb{H}$.

Quaternion-Fourier transform

The extra-dimensions of the quaternion algebra are used for the spectral analysis:

Definition 2 (Quaternion Fourier transform). *The Quaternion Fourier Transform (QFT) of axis $\boldsymbol{j}$ of a function $x : \mathbb{R} \mapsto \mathbb{H}$ is **[Flamant et al., 2016]**:*

$$\mathcal{F}^q\{x\}(\nu) = \int_{-\infty}^{+\infty} x(t)e^{-\boldsymbol{j}2\pi\nu t}dt$$

Due to the non-commutativity of $\mathbb{H}$, the position of the exponential on the right must be preserved for consistency reasons. A definition with the exponential on the left would give equivalent results with some inversion of signs. We name $X^q(\nu)$ the quaternion-Fourier transform of a signal x. Direct computations for $x = x_1 + \boldsymbol{i}x_2$ yield:

$$\begin{aligned}\mathcal{F}^q\{x\}(\nu) &= \int_{-\infty}^{+\infty} (x_1(t) + \boldsymbol{i}x_2(t))[\cos(2\pi\nu t) + \boldsymbol{j}\sin(2\pi\nu t)]dt \\ X^q(\nu) &= \int_{-\infty}^{+\infty} x_1(t)\cos(2\pi\nu t) + \boldsymbol{j}x_1(t)\sin(2\pi\nu t) + \boldsymbol{i}x_2(t)\cos(2\pi\nu t) + \boldsymbol{k}x_2(t)\sin(2\pi\nu t)dt\end{aligned} \tag{1.25}$$

First observation: the quaternion-Fourier transform computes the same quantities as the classical Fourier transform, it's based on cosine and sine transformation of a signal. Contrarily to the complex Fourier transform however, it separates the contributions from each component: the spectral information of x_1 is on span$\{1, \boldsymbol{j}\}$ while the spectral information of x_2 is on span$\{\boldsymbol{i}, \boldsymbol{k}\}$. In the rest, we write $\mathcal{F}(x)$ or X to designate the Fourier transform of a signal x carried on the axis $(1, \boldsymbol{j})$ so that (1.25) writes

$$X^q(\nu) = X_1(\nu) + \boldsymbol{i}X_2(\nu)$$

Second observation, the quaternion-Fourier transform satisfies the symmetry:

Proposition 5. *Let $x = x_1 + \boldsymbol{i}x_2 \in \mathbb{H}$ be a quaternion-valued signal on in the subspace $\mathbb{C}_{\boldsymbol{i}}$. Then, its quaternion-Fourier transform satisfies:*

$$X^q(-\nu) = \overline{X^q(\nu)}^{\boldsymbol{i}}$$

and

$$x(t) = Proj_{\mathbb{C}_{\boldsymbol{i}}}\left\{2\int_0^\infty X^q(\nu)d\nu\right\} \tag{1.26}$$

The quaternion-Fourier transforms reunite properties from the complex and vector Fourier transform. Its kernel is a rotation operator, and the result belongs to an algebra where it can be factorized with a module and a phase as with the complex-Fourier transform. However, it separates contributions from the real and imaginary parts of the signal and exhibits a symmetry between positive and negative frequencies.

Quaternion-embedding

As stated in (1.26), only positive frequencies are needed to reconstruct the signal from its quaternion-Fourier transform. This motivates the definition of a quaternion embedding for complex signal:

Definition 3 (quaternion embedding). *Let $x \in \mathbb{C}_{\boldsymbol{i}}$ be an absolutely integrable signal. Then its quaternion embedding $x_+(t) \in \mathbb{H}$ is the signal*

$$\begin{aligned} x_+(t) &= x(t) + \boldsymbol{j}\mathcal{H}[x](t) \\ &= x_{1+}(t) + \boldsymbol{i}x_{2+}(t) \end{aligned} \tag{1.27}$$

where $x_{k+}(t)$ is the analytic signal expressed on axis $(1, \boldsymbol{j})$ of the univariate signal x_k. The quaternion embedding satisfies

$$x_+(t) = 2\int_0^\infty X^q(\nu)d\nu$$

it is related to the original signal by:

$$x(t) = Proj_{\mathbb{C}_{\boldsymbol{i}}}(x_+(t)) \tag{1.28}$$

Interest of the quaternion embedding can be illustrated by bivariate monochromatic signals. The quaternion-Fourier transform of $x = a_1\cos(2\pi\nu_0 t + \phi_1) + \boldsymbol{i}a_2\cos(2\pi\nu_0 t + \phi_2)$ is

$$X^q(\nu) = \delta_{\nu_0}(a_1 e^{\boldsymbol{j}\phi_1} + \boldsymbol{i}a_2 e^{\boldsymbol{j}\phi_2}) + \delta_{-\nu_0}(a_1 e^{-\boldsymbol{j}\phi_1} + \boldsymbol{i}a_2 e^{-\boldsymbol{j}\phi_2})X^q(\nu) \quad = \lambda\delta_{\nu_0} + \overline{\lambda}^{\boldsymbol{i}}\delta_{-\nu_0} \tag{1.29}$$

This expression is not without reminding the expression of the Fourier transform of a complex monochromatic signal x' in (1.9)

$$X'(\nu) = a_+ e^{i\theta_+}\delta_{\nu_0} + a_- e^{-i\theta_-}\delta_{\nu_0}$$

except that were the spectral and geometrical information is contained in two different complex numbers, the couple $(a_+e^{i\theta_+}, a_-e^{i\theta_-})$, the same information is in only one quaternion number $\lambda = a_1 e^{j\phi_1} + \boldsymbol{i}a_2 e^{j\phi_2}$. As a consequence, only one rotary component will be necessary in the reconstruction of x. The quaternion λ can be factorized under its polar form (1.22)

$$\lambda = |\lambda| e^{i\theta'} e^{-k\chi'} e^{j\phi}$$

The quaternion embedding is obtained by inverse Fourier transform of the positive frequencies only of (1.29) and reads:

$$x_+(t) = |\lambda| e^{i\theta'} e^{-k\chi'} e^{j(2\pi\nu_0 t + \phi)}$$

Identifying $x(t)$ with the projection on the complex plane of $x_+(t)$ as in (1.28) gives

$$a_1\cos(2\pi\nu_0 t + \phi_1) + \boldsymbol{i}a_2\cos(2\pi\nu_0 t + \phi_2) = |\lambda| e^{i\theta'}[\cos(\chi')\cos(2\pi\nu_0 t + \phi) + \boldsymbol{i}\sin(\chi')\sin(2\pi\nu_0 t + \phi)]$$

After a bit of calculation work, it is possible to identify the angles of the quaternion factorization with the geometrical angles of the polarization ellipse as defined in table table 1.1[6]. Calling (a, θ, χ) the ellipse parameters of x they verify:

$$|\lambda|^2 = a_1^2 + a_2^2 = a^2 \quad \phi = \phi_2 + \phi_1 \quad \chi = \chi' \quad \theta = \theta'$$

The phase ϕ is a common phase factor, defining the position of the signal on its ellipse at time $t = 0$. The quaternion embedding of the signal finally writes:

$$x_+(t) = a e^{i\theta} e^{-k\chi} e^{j(2\pi\nu_0 t + \phi)}$$

Non-stationary signals

We saw in section section 1.3.3 that an instantaneous-frequency description of non-stationary bivariate signal necessarily relied on a pair of analytic signal. This is still verified in the quaternion algebra where the quaternion embedding (1.27) incorporates informations from two analytic signals. However, the quaternion algebra offers the possibility of a canonical factorization of the quaternion embedding that effectively articulates informations from the two signals. It results that for any signal x the quaternion embedding provides a unique quadruplet of time functions $(a(t), \theta(t), \chi(t), \phi(t))$ where a is the amplitude modulation, θ, χ are geometrical modulations and ϕ is the instantaneous phase such that the instantaneous

6

Temporal description:	$\sqrt{a_1^2 + a_2^2}$	$\frac{a_1^2 - a_2^2}{a_1^2 + a_2^2}$	$\frac{2a_1a_2}{a_1^2 - a_2^2}\cos(\phi_2 - \phi_1)$	$\frac{2a_1a_2}{a_1^2 + a_2^2}\sin(\phi_2 - \phi_1)$
Geometrical parameters:	a	$\cos 2\chi \cos 2\theta$	$\tan 2\theta$	$\sin 2\chi$

frequency is $\frac{d\phi}{dt}$. Concretely, following the exact same steps as in the previous section, for a complex signal $x(t) = x_1(t) + \boldsymbol{i}x_2(t)$ its quaternion embedding is

$$x_+(t) = a(t)e^{\boldsymbol{i}\theta(t)}e^{-\boldsymbol{k}\chi(t)}e^{\boldsymbol{i}2\pi\phi(t)} \tag{1.30}$$

The instantaneous frequency $\nu(t) = \frac{d\phi(t)}{dt}$ averages to the mean frequency of the signal with respect to the quaternion Fourier transform:

$$\frac{\int_{-\infty}^{+\infty}\nu(t)a^2(t)dt}{\int_{-\infty}^{+\infty}a(t)^2} = \frac{\int_{-\infty}^{+\infty}\nu|X^q(\nu)|d\nu}{\int_{-\infty}^{+\infty}|X^q(\nu)|^2d\nu}$$

Note that the unambiguous definition of a common instantaneous frequency between the two real components of a bivariate signal was made possible by the structure of the quaternion algebra that allows for a unique polar factorization once the order of the axis $(\boldsymbol{i}, -\boldsymbol{k}, \boldsymbol{j})$ is given. All the instantaneous parameters computed in this framework equal the instantaneous parameters we exposed in section 1.3.3.

1.4.4 Stokes parameters

The quaternion embedding of a signal gives access to the time-dependent Stokes parameters through the relation:

$$x_+(t)\boldsymbol{j}\overline{x_+(t)} = S_0(t) + \boldsymbol{i}S_3(t) + \boldsymbol{j}S_1(t) + \boldsymbol{k}S_2(t)$$

Note that this shape recalls the expression of the coherency matrix over the Pauli basis, so that we can identify $C_x \simeq x_+\boldsymbol{j}\overline{x_+}$ and $(\sigma_0, \sigma_1, \sigma_2, \sigma_3) \simeq (1, \boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k})$. This seems too curious to be random, and indeed we expose the relation between the Pauli matrices and the quaternion algebra in section 2.2. Again, the Stokes parameters defined here are equals to the Stokes parameters we would obtain by representing x as a signal. We compute exactly the same quantities, just the relation and the way to obtain these quantities differ. Note that in the quaternion algebra, we had the possibility to find a factorization of the analytic signal that exposed directly instantaneous geometric parameters, it is not necessary to compute first the instantaneous Stokes parameters. This convinces us that the quaternion algebra is a very nice embedding for the processing of bivariate signals and their polarization. The natural question is wether this approach of embedding the signal in an algebra big enough to contain state of polarizations can be generalized.

1.4.5 Limit of this formalism: the impossibility to extend it to higher dimensions

Representing bivariate signals in $\mathbb{H}$ instead of $\mathbb{C}$ or $\mathbb{C}^2$ has been a fruitful approach, that easily generalizes all desired features of spectral analysis and unify them in an elegant framework. The key to this success was the ability to work in an algebra with enough dimensions to treat

all geometrical parameters separately. Working in $\mathbb{C}$ requires to decompose polarization over orthogonal states, and working in $\mathbb{C}^2$ will at some point require matrix-vector factorizations. It was natural from this observation to ask wether this scheme could be reproduced for any dimensions. It appears that bivariate signals are a very special case. section 2.2 of next chapter tries to give elements to explain why everything works so well in the quaternion space. We do not believe n-variate signals should be processed through algebras, first, higer dimensional fields will lose the invertibility of all elements, and will not be as nice to work with as quaternions. Then, we believe the natural generalization for the processing of polarization is achieved through the notion of group action, and in particular the action of the group $U(n)$ on $\mathbb{C}^n$ and on the space of hermitian matrices. This is the approach we developp in chapter 3.

2 A theory-driven perspective on the quaternion embedding

> *No one fully understands spinors. Their algebra is formally understood, but their geometrical significance is mysterious. In some sense they describe the "square root" of geometry and, just as understanding the concept of $\sqrt{-1}$ took centuries, the same might be true of spinors.*
>
> **— Sir Michael Atiyah**
> (British mathematician)

Contents

In this chapter we propose a two-fold review of the quaternion Fourier transform. First, we will be interested by an abstract point of view on the Fourier transform and will define it on an abstract discrete group. The transition to continuous group will not be thoroughly explained, but we believe discrete groups while remaining tractable give all the keys to understand the extension of representation theory to continuous group. We will therefore focus the study on $(\mathbb{R}, +)$, showing at which steps how can quaternions be involved and how it fits in this theoretical definition of the Fourier transform. The second chapter directs its attention to the algebraic operations performed in $\mathbb{H}$ for the bivariate signal analysis. It shows that they can be understood in the context of Geometric algebra, a structure that embeds the quaternions. These considerations will then be confronted in the optic of extending the approach to trivariate signal analysis. It is concluded that bivariate signal analysis is a special case, both from a mathematical and physical point of view.

2.1 An algebraic construction of the Fourier transform

The Fourier transform is a tool that sends a signal defined over a given domain, it can be time or space or less usual structures, to a function over a dual variable. The two functions are related by the metric: the Fourier transform preserves the energy and the inner product of two signals, and linear operations. However, what does the notion of "dual variable" recover? We know from experience that the dual of time is frequency but it would be more accurate to say that the dual of time together with the operation of adding time intervals are frequencies. Indeed, it will appear in this section that spectral analysis is strongly dependent on a group structure, that is to say a composition law between elements of the group. As an example, we will show that a spectral analysis on $(\mathbb{R}, \times)$ is given by the Mellin transform and not by the Fourier transform. Finally, this abstract perspective on the ideas of spectral analysis is a mandatory step for any attempt to generalize the notion of spectral analysis to functions on domains that are not time. Examples include, functions on the sphere, functions over graphs, functions on Lie groups etc... We adopt in this chapter a formal language, apt at rendering with mathematical rigor the notions we aim to convey. Most of the material can be found in **[Peyré, 2004]**.

2.1.1 Groups and characters:

We consider a finite discrete group $(G, \times)$ with N distinct elements $G = \{g_1, \dots, g_N\}$. The space of $\mathbb{C}$-valued function on G is called $\mathbb{C}[G]$ and inherits a vector-space structure from $\mathbb{C}$. Generators of $\mathbb{C}[G]$ are:

$$(\delta_{g_i})_{0 \leqslant i < N} : g \mapsto \begin{cases} 1 \text{ if } g = g_i \\ 0 \text{ else} \end{cases} \tag{2.1}$$

It is immediate that N elements are sufficient to decompose any function of $\mathbb{C}[G]$, and so its dimension is N or less. For instance, in $\mathbb{Z}/2\mathbb{Z}$, the function $f : \begin{cases} 0 \mapsto 1 \\ 1 \mapsto \boldsymbol{i} \end{cases}$ can be decomposed as $f = \delta_0 + \boldsymbol{i}\delta_1$. The inner product on $\mathbb{C}[G]$ is:

$$\forall f, h \in \mathbb{C}[G], \ \langle f, h \rangle = \frac{1}{|G|} \sum_{g_i \in G} f(g_i)\overline{h(g_i)} \quad \text{where } \overline{h(g_i)} \text{ is the complex conjugate of } h(g_i) \tag{2.2}$$

Under this inner product, the family in (2.1) is orthonormal, it is a basis of $\mathbb{C}[G]$ whose dimension is therefore exactly N. On top of its vector-space structure, $\mathbb{C}[G]$ can be equiped with the structure of an algebra using a product operation $*$. We define $*$ with respect to the basis (2.1) by:

Definition 4 (convolution product). $\forall (g, h) \in G^2, \delta_g * \delta_h = \delta_{gh}$

Rather than depending on the product in the complex algebra, the convolution product uses the group relation and relies on the group structure. To that extent, it can be considered

a more natural product for a function space than the term-wise product where $(f.h)(g) = f(g).h(g) \ \ \forall g$. It is extended to any element of $\mathbb{C}[G]$ using the distributivity rule. For instance, using the function f of $\mathbb{C}[\mathbb{Z}/2\mathbb{Z}]$ defined in the previous example:

$$f * \delta_1 = (\delta_0 + \boldsymbol{i}\delta_1) * \delta_1 = \delta_{0+1} + \boldsymbol{i}\delta_{1+1} = \delta_1 + \boldsymbol{i}\delta_0$$

Dual and characters of a group:

Another basis of $\mathbb{C}[G]$ will be found by considering the dual G^* of G. The dual G^* is the set of all the continuous unitary morphisms from G to $\mathbb{T}$, where $\mathbb{T}$ is the unitary torus of an abelian field $\mathbb{K}$. In most cases, the field is $\mathbb{C}$, and $\mathbb{T} = \mathbb{U} = \{e^{i\theta}, \theta \in [0, 2\pi[\}$. Considering $\mathbb{K} = \mathbb{C}$ in the rest of the paragraph, a morphism is an application χ in $\mathbb{C}[G]$ verifying linearity conditions:

$$\chi(1_G) = 1_{\mathbb{C}} \qquad \text{and, for all } (g_i, g_j) \in G^2 \qquad \chi(g_i \times g_j) = \chi(g_i)\chi(g_j)$$

The elements of G^* are called the characters of G. In $(\mathbb{Z}/2\mathbb{Z}, +)$ where $+$ denotes the modulo addition in this context the characters satisfy:

$$\chi(\overline{0}) = 1 \qquad \chi(\overline{1})^2 = \chi(\overline{1} + \overline{1}) = \chi(0)$$

Only two applications denoted χ_1 and χ_2 satisfy these conditions. They value 1 at $\mathbf{0}$ and respectively 1 and -1 at $\mathbf{1}$. The unit function, constant equal to one, is always in the dual. Note also that χ_1 and χ_2 are unitary and orthogonal to each other, with respect to the inner product (2.2). Knowing that the dimension of $\mathbb{C}[\mathbb{Z}/2\mathbb{Z}]$ is two, we deduce that the characters of $\mathbb{Z}/2\mathbb{Z}$ are an orthonormal basis of $\mathbb{C}[\mathbb{Z}/2\mathbb{Z}]$. This is a general property of the dual as stated in proposition 6.

Proposition 6. *In an abelian[1] finite group $(G, \times)$, the elements of the dual G^* are an orthonormal basis of $\mathbb{C}[G]$ and the dual is isomorphic (not canonically) to G* ***[Simon, 1996]****.*

Example: the group $\mathbb{Z}/n\mathbb{Z}$

We can illustrate proposition 6 on the cyclic group $\mathbb{Z}/n\mathbb{Z}$. For $0 \leqslant i \leqslant n-1$ we note $\bar{i}$ the class of i in $\mathbb{Z}\ n\mathbb{Z}$. A character on this group would be completely determined by its value in $\overline{1}$. Indeed, if $\chi(\overline{1}) = u$, then $\chi(\bar{i}) = u^i$ as $\bar{i} = i \times \overline{1}$. For $0 \leqslant j \leqslant n-1$ let χ_j denote the following function:

$$\chi_j : \begin{cases} 1 \mapsto e^{\frac{i2\pi j}{n}} \\ \bar{i} \mapsto e^{\frac{i2\pi ji}{n}} \end{cases}$$

It is immediate to check that χ_j is a morphism and that it is unitary. Hence it is a character of $\mathbb{Z}/n\mathbb{Z}$. Conversely, a character would verify $\chi(1)^n = \chi(0) = 1$. Hence $\chi(1)$ is a nth-root

[1] A group $(G, \times)$ is abelian if its law satisfie: $\forall x, y \in G, x \times y = y \times x$. The real and complex fields are abelian, while the quaternion field or the set of $n \times n$ real matrices are not.

of the unity: there is a $j \in [|0, n-1|]$ such that $\chi(1) = e^{\frac{i2\pi j}{n}} = \chi_j(1)$. All the characters of $\mathbb{Z}/n\mathbb{Z}$ are therefore accounted for in the $(\chi_j)_{0 \leqslant j \leqslant n-1}$. Furthermore, the application $\overline{j} \mapsto \chi_j$ realizes an isomorphism between $\mathbb{Z}/n\mathbb{Z}$ and $\mathbb{Z}/n\mathbb{Z}^*$. Orthonormality derives from the fact that $\sum_{k=0}^{n-1} e^{\frac{i2\pi k}{n}} = \frac{1-e^{\frac{i2\pi}{n}n}}{1-e^{\frac{i2\pi}{n}}} = 0$.

Note that the unitary conditions on characters is not invoked in a cyclic group. It derives automatically from the nilpotence of each element in this group. This remark is true for any abelian finite group.

2.1.2 From characters to the Fourier transform:

The dual G^* of G provides an orthonormal basis of $\mathbb{C}[G]$. It is tempting to use this basis to compute coordinates of arbitrary functions, and indeed, coordinates along this basis have a special significance for the signal processing community. For a character χ and a function f, we call "Fourier coefficient" $c_f(\chi)$ the coordinate of f along χ in the basis made by G^*.

$$c_f(\chi) = \langle f, \chi \rangle = \frac{1}{|G|} \sum_{g \in G} f(g)\overline{\chi}(g)$$

In $\mathbb{Z}/n\mathbb{Z}$:

$$c_g(\chi_j) = \frac{1}{n} \sum_{0 \leqslant k \leqslant n-1} f(\overline{k}) e^{-\frac{i2\pi jk}{n}}$$

Name $x_k = f(k)$ a sample from a discrete signal, and the above expression becomes the expression of the coefficient at frequency j of the discrete Fourier transform of the signal $(x_k)_k$. The previous analysis shows how the discrete Fourier transform implicitely relies on a cyclic structure of the sample. It is this cyclic structure that allows for a very strong efficiency from a computation point of view. The Fourier transform is the function that send a character on the Fourier coefficient of the conjugate character:

Definition 5. *The Fourier transform named $\mathcal{F}$ is the application:*

$$\begin{aligned} \mathcal{F} \colon \mathbb{C}[G] &\to \mathbb{C}[G^*] \\ f &\mapsto \widehat{f} \end{aligned}$$

where $\widehat{f}$ is defined by:

$$\forall \chi \in G^*, \quad \widehat{f}(\chi) = \sum_{x \in G} f(x)\chi(x) = \|G\| c_g(\overline{\chi}) = \|G\| \langle f, \overline{\chi} \rangle \tag{2.3}$$

It has a special relationship with the convolution product defined in 4. Let δ_g, δ_h be two delta functions as defined in (2.1). The Fourier transform of the convoluted product $\delta_g * \delta_h$ is:

$$\mathcal{F}(\delta_g * \delta_h) = \mathcal{F}(\delta_h)\mathcal{F}(\delta_g)$$

Proof.

$$\mathcal{F}(\delta_g * \delta_h) = \sum_{k\in G} \delta_{gh}(k)\chi(k) = \chi(hg) = \chi(h)\chi(g)$$
$$\mathcal{F}(\delta_g * \delta_h) = \left(\sum_{k\in G} \delta_g(k)\chi(k)\right)\left(\sum_{k\in G} \delta_h(k)\chi(k)\right)$$

□

It follows that for any two functions f_1, f_2 the Fourier transform sends the convolution product over the point-wise product:

$$\mathcal{F}(f_1 * f_2) = \mathcal{F}(f_1)\mathcal{F}(f_2)$$

In light of what precedes, we believe the reason the conjugate $\overline{\chi}$ is used in the definition of the Fourier transform rather than χ is because it enables the relation between the Fourier transform and the convolution product to be extended to non-abelian fields where $\overline{\chi(g)\chi(h)} = \overline{\chi(g)\chi(h)}$ might not be verified. Typically, if the characters are quaternion-valued we have: $\overline{\chi}(g)\overline{\chi}(h) = \overline{\chi}(h)\overline{\chi}(g)$. One of the most remarkable property of the Fourier transform is recalled below:

Proposition 7. *For any abelian group $(G, \times)$, the Fourier transform defined in definition 5 is a vector-space isomorphism between $\mathbb{C}[G]$ and $\mathbb{C}[G^*]$*

This property induces the well-known Parseval and Plancherel theorems for classical continuous case, and the inversion formula. This introduction does not delve into the numerous applications of the Fourier transform in signal processing. It aims at succintly recalling the theoretical background behind the construction of the Fourier transform in a general context (not as general as possible however, as a definition through representation theory deals with non-abelian group as well).

2.1.3 The Fourier transform on $\mathbb{R}$

Previous sections showed the construction of the Fourier transform on discrete groups and without going into details, we will assume that all generalizes to the infinite abelian group $(\mathbb{R}, +)$ with no complication. In a continuous group such as $\mathbb{R}$, the inner product would involve an integral rather than a sum. We consider:

$$\langle f, h\rangle = \int_{-\infty}^{+\infty} f(t)\overline{h(t)}dt \tag{2.4}$$

for functions $(f, h) \in \mathbb{C}[\mathbb{R}]$ such that the integral (2.4) is convergent.

The only continuous morphisms on $(\mathbb{R}, +)$ belong to the exponential family[2] and the unitary condition restrains the characters of $\mathbb{R}$ to the family $(e^{i\mu t})_{\mu\in\mathbb{R}}$. The relation between the characters of $\mathbb{R}$, and the Fourier transform on $\mathbb{C}[\mathbb{R}]$ becomes visible. Following definition 5:

Definition 6. *The Fourier transform on $\mathbb{R}$ is defined for $f \in \mathbb{C}[\mathbb{R}]$ such that the following integral converges by:*

$$\mathcal{F}(f)(t \mapsto e^{i\nu t}) = \nu \mapsto \int_{-\infty}^{+\infty} f(t)e^{i\nu t}dt \tag{2.5}$$

Note that the domain variable is a character rather than an element of $\mathbb{R}$ itself. The fact $t \mapsto e^{-i\nu t}$ and ν can be used interchangeably in the definition comes from the self-duality of $\mathbb{R}$. This is aknowledged in the classical framework where $\widehat{f}(\nu) = \int_{-\infty}^{+\infty} f(t)e^{-2i\pi\nu t}dt$ by talking of the frequency ν as a "dual variable". But given the definition above, the true argument of the Fourier transform $\widehat{f}$ is the application $t \mapsto e^{-2i\pi\nu t}$.
Interestingly, changing the group $(\mathbb{R}, +)$ for $(\mathbb{R}^*, \times)$ the characters become $(e^{i\gamma ln(t)})_{\gamma\in\mathbb{R}}$ and the decomposition of a function on this basis almost defines the Mellin transform on the imaginary axis **[Bertrand et al., 1994]**. Indeed, following definition 5:

$$\mathcal{F}^{\times}(f)(\gamma) = \int_{-\infty}^{+\infty} f(t)t^{i\gamma}dt = \mathcal{M}(f)(i\gamma + 1)$$

The next steps show how the Fourier transform on $\mathbb{H}$ defined in definition 8 relates to this algebraic aspect of the definition of the Fourier transform.

2.1.4 Extension to the Field of quaternions

The most important step for the construction of the Fourier transform is the identification of a basis of $\mathbb{K}[G]$ made of morphisms. When $\mathbb{K} = \mathbb{C}$, this basis is found in G^*, the dual of G. For $\mathbb{K} = \mathbb{H}$ however, we will see that the set of unitary morphisms from G to $\mathbb{H}$ generates $\mathbb{H}[G]$ but is not a basis due to redundancy. We will relate this property to properties of polynomial in $\mathbb{C}$ and $\mathbb{H}$. Then, we will show how quotienting enables us to identify a basis of $\mathbb{H}[G]$ in the set of unitary morphisms that we call $(G^*)_{\mathbb{H}}$. This example will show how the framework above can be adapted to the construction of a quaternion Fourier transform for quaternion-valued functions.

[2] This can be shown by first noticing that a continuous morphism on $\mathbb{R}$ is differentiable and then showing it is solution to a differential equation solved only by the exponential.

Relation between polynomial roots and self-duality

Proposition 8. *Let G be a discrete abelian group. The dimension of $\mathbb{K}[G]$ as a $\mathbb{K}$ vector-space is independent of the field $\mathbb{K}$ and*

$$dim_{\mathbb{K}}(\mathbb{K}[G]) = |G|$$

Proof. Let δ_g be the delta function on G that values 0 if $h \neq g$ and $1_{\mathbb{K}}$ if $h = g$. The set $(\delta_g)_{g\in G}$ obviously generates $\mathbb{K}[G]$ by $\mathbb{K}$-linear combinations. Furthermore, evaluating a null linear combinations of (δ_{g_i}) in g_i shows that every coefficients of the linear combination is zero. The family is free and and $\dim(\mathbb{K}[G]) = |G|$. □

The fundamental theorem of algebra states that in $\mathbb{C}$, a non-constant polynomial of degree n has exactly n roots. This is not true if we replace the field $\mathbb{C}$ by $\mathbb{R}$ for instance. Polynomials may have less roots than their degree, for instance the equation $X^2 + 1$ has no real roots. It is not true either in the quaternion field $\mathbb{H}$ where polynomials can have more roots than their degrees, and even an infinity of roots. Now, the number of unitary morphisms of a discrete group G in a field K is related to the number of roots of the unit in the field K. For instance, in the group $\mathbb{Z}\ n\mathbb{Z}$ a morphism χ is completely defined by its value at $\overline{1}$. Considering $n\overline{1} = \overline{0}$, the morphism χ satisfies:

$$\chi(\overline{1})^n = 1$$

Thus, $\chi(\overline{1})$ is a root of the polynomial $P(X) = X^n - 1$ in the field $\mathbb{K}$. In $\mathbb{C}$ the n complex roots of the unity provided exactly n distinct characters. In the quaternion field, the set of roots of P grows to $\{e^{\frac{2\pi\boldsymbol{\mu}kt}{n}} | k \in \mathbb{N}\ ,\ \boldsymbol{\mu}$ a pure unitary quaternion $\}$. This group contains an infinite number of element. Each element characterizes a unique morphism from G to $\mathbb{K}$. Thus, the number of characters over the field $\mathbb{K}$ is equal to the number of roots of $X^{|G|} - 1$ in this field. The dimension of $\mathbb{K}[G]$ however is independent of $\mathbb{K}$. Consequently, if $|G| > 2$, the isomorphism between $\mathbb{K}[G]$ and $(G^*)_{\mathbb{K}}$ holds only for $\mathbb{K} = \mathbb{C}$. In $\mathbb{R}$, unitary morphisms are not enough to build a basis of $\mathbb{R}[G]$ while in $\mathbb{H}$ they are too many. However, if $\mathbb{C} \subset \mathbb{K}$ a direct reasoning on the dimension brings:

Proposition 9. *For a field $\mathbb{K}$ such that $\mathbb{C} \subseteq \mathbb{K}$, a basis of $\mathbb{K}[G]$ is $(G^*)_{\mathbb{C}}$*

Proof. We have $(G^*)_{\mathbb{C}} \subseteq (\mathbb{K}[G])$ and $\dim((\vec{G^*})_{\mathbb{C}}) = \dim(\mathbb{C}[G]) = \dim(\mathbb{K}[G])$. □

The fact that complex-valued functions are enough to generate the set of quaternion-valued functions associated with a few simple results we prove below will enable us to replace the imaginary unit $\boldsymbol{i}$ by any pure unitary quaternion $\boldsymbol{\mu}$ in the expression of a basis of $\mathbb{K}[G]$. Recall that a quaternion $\boldsymbol{\mu}$ is pure unitary if and only if

$$\boldsymbol{\mu}^2 = -1$$

Definition 7. *For a pure imaginary quaternion $\boldsymbol{\mu}$, let $\pi_{\boldsymbol{\mu}}$ be the field isomorphism in $\mathbb{H}$ that sends the quaternion $z = a + \boldsymbol{i}b + \boldsymbol{j}c + \boldsymbol{k}d$ on the quaternion whose imaginary part has been multiplied by $\boldsymbol{\mu}$*

$$\pi_{\boldsymbol{\mu}}(z) = a + \boldsymbol{\mu}(\boldsymbol{i}b + \boldsymbol{j}c + \boldsymbol{k}d)$$

Then $\pi_{\boldsymbol{\mu}}$ describes a vector-space isomorphism in $\mathbb{H}[G]$ through the relation:

$$\pi_{\boldsymbol{\mu}}(f) \colon x \mapsto \pi_{\boldsymbol{\mu}}(f(x))$$

Proposition 10. *The map $\pi_{\boldsymbol{\mu}}$ realizes an isometry in $\mathbb{H}[G]$.*

Proof. Let f be a function in $\mathbb{C}_{\boldsymbol{i}}[G]$. Then

$$\begin{aligned} \|\pi(f)\|^2 &= \sum_{x \in G} \pi(f(x))\overline{\pi(f(x))} \\ &= \sum_{x \in G} \|f(x)\|^2 \end{aligned}$$

as it is clear that $\|\pi_{\boldsymbol{\mu}}(z)\| = \|z\|$. Finally

$$\|\pi(f)\|^2 = \|f\|^2$$

□

In particular, an isometry preserves the inner-product, which enables us to state the following result.

Proposition 11. *If $(f_1, \dots, f_n)$ is an orthonormal basis of $\mathbb{H}[G]$ then the family $(\pi(f_1), \dots, \pi(f_n))$ is also an orthonormal basis of $\mathbb{H}[G]$.*

In $\mathbb{Z}/n\mathbb{Z}$ we have already shown that a basis of the set of $\mathbb{H}$-valued functions is $(\chi_1, \dots, \chi_n)$ where

$$\chi_i \colon \overline{k} \mapsto e^{\boldsymbol{i}\frac{ik}{2\pi}}$$

Applying proposition 11 shows that the family $(\tilde{\chi_1}, \dots, \tilde{\chi_n})$ with

$$\tilde{\chi_i} \colon \overline{k} \mapsto e^{\boldsymbol{\mu}\frac{ik}{2\pi}}$$

for a pure imaginary quaternion $\boldsymbol{\mu}$ is also an orthonormal basis of $\mathbb{H}[\mathbb{Z}/n\mathbb{Z}]$. Finally, we have proved a result that might seem perfectly obvious, but it helps us show that the likeness

between the complex Fourier transform and the quaternion Fourier transform goes deeper than just a likeness in form. It has algebraic roots.

Quaternion Fourier transform on $\mathbb{R}$

We built the Fourier transform on $(\mathbb{R}, +)$ by analogy with the the Fourier transform on a discrete group by applying the same formula. The Fourier transform of $f \in \mathcal{L}^1(\mathbb{R})$ in χ is

$$\mathcal{F}(f)(\chi) = \int_{-\infty}^{+\infty} f(t)\overline{\chi(t)dt}$$

where χ is a morphism from $(\mathbb{R}, +) \to (\mathbb{C}, +)$. We do note use the concept of basis in the vector space of integrable functions as it is of infinite dimensions. However, the idea that the set of characters of $(\mathbb{R}, +)$ is enough to represent a function f is present and on certain conditions on f, the Fourier transform is invertible. This shows that the characters capture the complexity of the function space. We have shown that changing the kernel in the exponential characters of the group $\mathbb{Z}/n\mathbb{Z}$ did not change their ability to decompose any function in $\mathbb{H}[\mathbb{Z}/n\mathbb{Z}]$. We define the quaternion Fourier transform of axis $\boldsymbol{\mu}$ as in **[Flamant et al., 2016]**.

Definition 8. *For a pure unitary quaternion $\boldsymbol{\mu}$, the Quaternion Fourier Transform (QFT) of axis $\boldsymbol{\mu}$ of a function $f : \mathbb{R} \mapsto \mathbb{H}$ is:*

$$\mathcal{F}_q\{f\}(\omega) = \int_{-\infty}^{+\infty} f(t)e^{-\boldsymbol{\mu}\omega t}dt$$

The classical Fourier transform is a particular case of the quaternion Fourier transform obtained for $\boldsymbol{\mu} = \boldsymbol{i}$ and applied on a class of signals in $\mathbb{C}$. It was shown in **[Flamant et al., 2016]** that the quaternion-Fourier transform is invertible for any kernel $\boldsymbol{\mu}$ under the same conditions than the classical Fourier transform. In this part, we proved this latter result on discrete groups, which are the perfect settings to understand the algebraic implications of the Fourier transform. The non-commutativity of quaternions does not interfere with the most fundamental properties of the Fourier transform, because most of them result from an isomorphism between a group G and a set of morphisms used to compute the Fourier transform. We showed that this isomorphism still holds, even if not directly between G and the groups of characters of G. Finally, the extension of the concept of duality to the field $\mathbb{H}$ is compatible with the definition of a Fourier transform for functions from $\mathbb{R} \in \mathbb{H}$. One drawback of this introduction however, is that the Fourier transform is fundamentally based on the notion of characters, being defined as functions from a group into a field $\mathbb{K}$. This is the reason why an extension of the Fourier transform under this form to n-variate signals where $n > 4$ is compromised. There are no fields of dimension more than 4 that embeds the complex

numbers and are similar enough to the complex number. This can be seen as a consequence of the following theorem.

Theorem 2 (Frobenius 1877). *If E is an algebra on the field of scalars $\mathbb{R}$ of finite dimensions such that all the elements of E are invertible, then $E \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$.*

There are no spaces "above" the quaternions, unless you are ready to loose some important properties such as associativity or invertibility of elements. Clifford algebras, that we introduce in next section are a typical examples of finite dimensional algebras over $\mathbb{R}$ where all elements are not invertible.

2.2 The quaternion-embedding: a special case of Geometric algebra embedding

Although vectors have proved to be extremely useful quantities to describe all sorts of data, it has not gone unnoticed that they are without a product structure. A product is however useful in some geometrical descriptions, where vectors can be seen both as objects and operators: this is typically what drives the difference between the complex field and $\mathbb{R}^2$ and the quaternions and $\mathbb{R}^3$ as we saw in the previous sections. In this regard, the Clifford algebras are one proposition to build an algebraic structure around a vector space that embeds vectors and their product. Given a vector space V over a field $\mathbb{K}$ and a quadratic form $q\colon V \to \mathbb{K}$ the Clifford algebra $\mathcal{C}(V,q)$ embeds V and its linear structure as well as products of vectors of V. For instance, squared vectors are defined by:

$$\forall v \in V, v^2 = q(v)1$$

where V is seen as a subset of $\mathcal{C}(V,q)$. It is closely related to exterior algebras, generated around a vector space by the "wedge" product $\wedge$, if $q = 0$ then $\mathcal{C}(V,0) = \wedge V$. Among such algebras, the geometric algebras is a special kind where the quadratic form q is consistent with the metric of V such that in the geometric algebra $v^2 = \|v\|^2 1$ where $\|\cdot\|$ is the norm in V.

2.2.1 the Geometric algebra $\mathcal{C}\ell_3$

The clifford algebra $\mathcal{C}\ell_3$ is the geometric algebra built around $\mathbb{R}^3$, with the quadratic form compatible with the euclidian metric. It is an 8-vector space on the field $\mathbb{R}$. It contains scalars, vectors, bivectors (product of vectors) and pseudo-scalars. The basis elements of $\mathcal{C}\ell_3$ are presented in table 2.1. The particularity of this vector space is to be equipped with an intern product that should not be confused with the inner product or the wedge product (It can be shown that the clifford product is actually a linear combination of both).

scalar	vectors	bivectors	pseudo-scalar
1	e_1, e_2, e_3	e_{12}, e_{31}, e_{23}	$e_{123} = \mathbf{I}$
$1^2 = 1$	$e_i^2 = 1$	$e_{ij}^2 = -1$	$e_{123}^2 = -1$

Tab. 2.1. – Basis element of the Clifford algebra $\mathcal{C}\ell_3$ and their square

The following product rules hold:

$$e_i e_j = -e_j e_i = e_{ij} \text{ and } e_1 e_2 e_3 = e_{123} = \mathbf{I} \tag{2.6}$$

Every element can be obtained as the product of the three vectors e_1, e_2, e_3 with each other. There is a natural vector-space isomorphism between $\mathbb{R}^3$ and $e_1, \vec{e_2}, e_3$ in $\mathcal{C}\ell_3$. In this isomorphism, the triplet (e_1, e_2, e_3) is sent on an orthonormal basis for $\mathbb{R}^3$. For this reason, $\mathcal{C}\ell_3$ is

seen as en embedding algebra for $\mathbb{R}^3$. To grasp intuition about the product rules and relations in the algebra, it can be useful to look at it through the isomorphisms existing between $\mathcal{C}\ell_3$ and $Mat(2,\mathbb{C}), \mathbb{H}$ and $\mathbb{C}$ given in table 2.2.

$\mathcal{C}\ell_3$	$\mathrm{Mat}(2,\mathbb{C})$	$\mathbb{H}$	$\mathbb{C}$	$\mathbb{R}^n$
1	Id	1	1	$\mathbb{R}$
e_1, e_2, e_3	$\sigma_1, \sigma_2, \sigma_3$	$\varnothing$	$\varnothing$	$\mathbb{R}^3$
e_{12}, e_{13}, e_{23}	$\sigma_1\sigma_2, \sigma_1\sigma_3, \sigma_2\sigma_3$	$-\boldsymbol{k}, \boldsymbol{j}, -\boldsymbol{i}$		$\mathbb{R}^3$
$\mathbf{I}$	$\sigma_1\sigma_2\sigma_3$	$\varnothing$	$\boldsymbol{i}$	$\boldsymbol{i}\mathbb{R}$

Tab. 2.2. – Isomorphisms between subspaces of Clifford algebras and other vector spaces

Pauli matrices, used in table 2.2, are defined as in (1.5) by:

$$\sigma_1 = \begin{pmatrix} 0 & \boldsymbol{i} \\ \boldsymbol{i} & 0 \end{pmatrix} \quad \sigma_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{2.7}$$

Just like in a matrix space, the product is anticommutative in general, and the addition is distributive for the product. Products can be decomposed along a commutative and anti-commutative part: for Clifford elements z_1, z_2

$$z_1 z_2 = z_1 \cdot z_2 + z_1 \wedge z_2 \quad , \quad z_1 \cdot z_2 = \frac{z_1 z_2 + z_2 z_1}{2} \quad , \quad z_1 \wedge z_2 = \frac{z_1 - z_2}{2}$$

In many ways (see for instance (2.13)) $(\cdot, \wedge)$ behave like an inner and a wedge product in a vector space and are called respectively the inner and the outer product on $\mathcal{C}\ell_3$. Working in $\mathcal{C}\ell_3$ can be preferred over working in $\mathrm{Mat}(2,\mathbb{C})$ whenever its inner structure separating vectors from scalars and bivectors is relevant. The notation $\mathbf{I}$ for e_{123} stresses the algebraic and linear equivalence existing between the subspaces $Z(\mathcal{C}\ell_3) = \mathrm{vect}\{1, \mathbf{I}\}$ and $\mathbb{C}$. The subspace $Z(\mathcal{C}\ell_3)$ is called the center of $\mathcal{C}\ell_3$ because its elements are exactly those that commute with any z in $\mathcal{C}\ell_3$. Commutation is important, in particular for the definition of an inverse. Naturally[3], not all elements of $\mathcal{C}\ell_3$ have an inverse for the product rule, but when they have, it is both a *right* and a *left* inverse. The following decomposition of a Clifford element on the basis defined in table 2.1 exhibits eight independent scalar coordinates for any element z of $\mathcal{C}\ell_3$:

$$z = \alpha + a_1 e_1 + a_2 e_2 + a_3 e_3 + b_1 e_{12} + b_2 e_{31} + b_3 e_{23} + \beta \mathbf{I}$$

The norm is defined respectively to these coordinates by:

$$\|z\| = \sqrt{\alpha^2 + a_1^2 + a_2^2 + a_3^2 + b_1^2 + b_2^2 + b_3^2 + \beta^2} \tag{2.8}$$

A more compact and structurally informative way to write z is:

$$z = \alpha + \boldsymbol{v} + \mathbf{I}\boldsymbol{w} + \beta \mathbf{I} \tag{2.9}$$

[3]Otherwise there would be endless possibilities to define fields of any dimensions over $\mathbb{R}$, which as states the Frobenius theorem is not possible

where the vector $\boldsymbol{v} = ae_1 + be_2 + ce_3$ is the vector part and the vector $\boldsymbol{w} = b_1e_3 + b_2e_2 + b_3e_1$ when multplied by $\mathbf{I}$ gives the bivector part. Indeed, using (2.6) it is easy to check that $\mathbf{I}e_1 = e_{23}, \mathbf{I}e_2 = e_{13}, \mathbf{I}e_3 = e_{12}$. As $\mathbf{I}$ commutes with any element, its position on the left or the right of $\boldsymbol{w}$ is equivalent. We refer to blocks of this decomposition by

- $\langle z \rangle_0 = \alpha$ The projection of z on its scalar part
- $\langle z \rangle_1 = \boldsymbol{v}$ The projection of z on its vector part
- $\langle z \rangle_2 = \mathbf{I}\boldsymbol{w}$ The projection of z on its bivector part
- $\langle z \rangle_3 = \beta\mathbf{I}$ The projection of z on its pseudo-vector part

The index i in $\langle z \rangle_i$ is refered to as the "grade" of $\langle z \rangle_i$ and denotes the maximal number of vectors involved in a product in one of the terms of $\langle z \rangle_i$. For instance, the grade of $1 + \mathbf{I}$ is three, because $\mathbf{I}$ is the product of three vectors while 1 is the product of zero vector. A first approach to grade is to compare it with the degree of a polynom. Like degrees, grades tend to sum when elements are multiplied but with exceptions: in $e_1^2 = 1$ the multiplication of two one graded elements brings a 0-graded number. The rules on grade can be formalized as follows: the grade of the product of one-graded elements is either 0 or 2 in a way that the grade can never be more than 3. Applying this result, it follows that mutliplying an even number of vectors together can only give an even-graded element. The stability of even-graded elements is formalized in the following property

Proposition 12. *The set of all even-graded elements of $\mathcal{C}\ell_3$ form a subalgebra called $\mathcal{C}\ell_{3+}$.*

$$\mathcal{C}\ell_{3+} = \{a + be_{12} + ce_{13} + de_{14} | (a, b, c, d) \in \mathbb{R}^4\}$$

There is a canonical isomorphism $\mathbb{H} \simeq \mathcal{C}\ell_{3+}$ following table 2.2. The space of even-graded elements is used in section section 2.2.1 where it is identified with the spinor space.

Conjugations in $\mathcal{C}\ell_3$ There are several useful automorphisms, similar to the conjugation in a complex field, in the Clifford algebra. Given a Clifford number $z = \alpha + \boldsymbol{v} + \mathbf{I}\boldsymbol{w} + \beta\mathbf{I}$ we define:

$$\begin{aligned} \hat{z} &= \alpha - \boldsymbol{v} + \mathbf{I}\boldsymbol{w} - \beta\mathbf{I} && \textit{Grade involution} \\ \tilde{z} &= \alpha + \boldsymbol{v} - \mathbf{I}\boldsymbol{w} - \beta\mathbf{I} && \textit{Reversion} \\ \overline{z} &= \alpha - \boldsymbol{v} - \mathbf{I}\boldsymbol{w} + \beta\mathbf{I} && \textit{Clifford conjugation} \end{aligned}$$

The Grade involution produces the norm and the inner product of Clifford numbers by:

$$\langle y, z \rangle = \langle \hat{y}z \rangle_0 = \langle y\hat{z} \rangle_0 \quad \text{and} \quad \|z\|^2 = \langle \hat{z}z \rangle_0$$

when applied to a vector $\boldsymbol{v}$, they yield the euclidian norm and inner product.
The conjugation produces what is called the amplitude $|z|$ of a clifford multivector z:

$$|z| = \sqrt{z\overline{z}} \tag{2.10}$$

It is a complex-like number (i.e $|z| = \alpha' + \mathbf{I}\beta'$) and belongs to the centrum of the algebra $Z(\mathcal{C}\ell_3)$, hence commuting with any number. The amplitude defines the invertibility and the inverse of a Clifford number through the equivalence:

$$\text{The clifford number } z \text{ is invertible} \iff z\overline{z} \neq 0$$
$$\text{Its inverse is then } z^{-1} = \frac{\overline{z}}{z\overline{z}}$$

The division is made unambiguous due to the commutativity of $z\overline{z} = \overline{z}z$ with any multivector. The inverse does not always exist, for instance the multivector $z = 1 + e_1$ is not invertible, its amplitude is $(1 + e_1)(1 - e_1) = 0$.
The absence of inverse for some elements must of course have some consequences on the possibilities to represent operations as Clifford elements. Many operators needed in signal processing are reversible, a property that must translate by invertibility in the chosen algebra.

Exponential in Clifford algebra The best way to generalize the exponential to a vector space is to rely on its power series. As a reminder, for any $z \in \mathbb{C}$:

$$e^z = \sum_{n=0}^{+\infty} \frac{z^n}{n!} \tag{2.11}$$

The same power series converges for any z in $\mathcal{C}\ell_3$, and for any element x such that $x^2 = -\|x\|^2$, it sums to $\cos\|x\| + \frac{x}{\|x\|}\sin\|x\|$ while for elements such that $y^2 = \|y\|^2$ it gives $\cosh\|y\| + \frac{y}{\|y\|}\sinh\|y\|$. The pseudo-scalar $\mathbf{I}$ behaves like the imaginary unit $\boldsymbol{i} \in \mathbb{C}$, and following the previous remark:

$$e^{\mathbf{I}\theta} = \cos(\theta) + \mathbf{I}\sin(\theta)$$

The same holds for bivectors: $e^{e_{ij}\theta} = \cos(\theta) + e_{ij}\sin(\theta)$. However, for a vector e_i the exponential developps in:

$$e^{e_i\theta} = \cosh(\theta) + e_i\sinh(\theta)$$

Transformation groups in $\mathcal{C}\ell_3$

We show in this section how $\mathcal{C}\ell_3$ provides two classes of objects, the vectors and the spinors, that are affected differently by linear transformations. This appears as a necessary condition to express quantities as inhomogeneous as a Jones vector and its coherency matrix in the same algebra. We show how the quaternion algebra inherits some of this structure, which explains its success at representing polarization for bivariate signals.

Each invertible element x in $\mathcal{C}\ell_3$ is mapped to an action $T_x\colon \mathcal{C}\ell_3 \to \mathcal{C}\ell_3$ defined by [Gallier, 2008]

$$T_x\colon z \mapsto \widehat{x}zx^{-1} \tag{2.12}$$

It is convenient to demand for transformations that stabilize the vector space $V = e_1, \vec{e_2}, e_3$. They are given by the Clifford group Γ : for v a vector of $\mathcal{C}\ell_3$ and $\gamma \in \Gamma$ the transform $\widehat{\gamma}v\gamma^{-1}$ is a vector.

Example:

The vectors themselves belong to the Clifford group and they are mapped to reflections [Gallier, 2008]. To show that, we need to introduce the colinear and orthogonal part of vector. For (x, v) two vectors, x writes as the sum of its orthogonal projection over v and a rest:

$$x = x_{\|v} + x_{\perp v} \text{ with } x_{\|v} \wedge v = 0, x_{\perp v} \cdot v = 0 \tag{2.13}$$

Then computing $T_x(v) = \widehat{x}vx^{-1}$ yields:

$$T_x(v) = -xvx^{-1} = -\left[x \cdot (v_{\|x} + x \wedge v_{\perp x})\right] x^{-1}$$

Recall that the inner product between two vectors is commutative, and the outer product is anticommutative, so we can write:

$$\begin{aligned} T_x(v) &= [v_{\|x} \cdot x - v_{\perp x} \wedge x]x^{-1} = [v_{\|x} - v_{\perp x}]xx^{-1} \\ T_x(v) &= v_{\|x} - v_{\perp x} \end{aligned}$$

T_x realizes a reflection with respect to the hyperplane orthogonal to x.
We identify two subgroups of $\Gamma(G)$ of interest for the rest of our study.

Definition 9. *The* **Pin** *group of $\mathcal{C}\ell_3$ is the subgroup of elements of G that satisfy $x\overline{x} = 1$. Note that as $x^{-1} = \frac{\overline{x}}{x\overline{x}}$ for an element of the* ***Pin*** *group we have $x^{-1} = \overline{x}$*

$$\mathbf{Pin}(3) = \{x \in \mathcal{C}\ell_3 | \widehat{x}v\overline{x} \in V \quad \forall v \in V, x\overline{x} = 1\}$$

The **Spin** *group is the subgroup of* **Pin(3)** *of even-graded elements.*

$$\mathbf{Spin}(3) = \{x \in \mathcal{C}\ell_{3+} | \widehat{x}v\overline{x} \in V \quad \forall v \in V, x\overline{x} = 1\}$$

The group **Pin**(3) plays the same role as $O(3)$ over trivariate vectors, while **Spin**(3) relates to $SO(3)$. An even-graded element writes $x = \alpha + \beta \mathbf{I}\boldsymbol{w}$ where $\boldsymbol{w}$ is a unitary vector. This factorizes as

$$x = \frac{1}{\sqrt{\alpha^2+\beta^2}}\left(\frac{\alpha}{\sqrt{\alpha^2+\beta^2}} + \frac{\beta}{\sqrt{\alpha^2+\beta^2}}\mathbf{I}\boldsymbol{w}\right)$$

$$x = \frac{1}{\sqrt{\alpha^2+\beta^2}} e^{\mathbf{I}\boldsymbol{w}\theta} \qquad \text{where } 0 \leqslant \theta < 2\pi, (\cos\theta, \sin\theta) = \left(\frac{\alpha}{\sqrt{\alpha^2+\beta^2}}, \frac{\beta}{\sqrt{\alpha^2+\beta^2}}\right).$$

Next proposition deduces from this result that elements of the group **Spin**(3) are mapped to rotations on V by (2.12).

Proposition 13. *Let $x \in$ **Spin**(3), it writes under exponential form as $x = e^{\mathbf{I}\boldsymbol{\mu}\theta}$ for a unitary vector $\boldsymbol{\mu}$. the action T_x on V writes:*

$$T_x \colon v \mapsto e^{\mathbf{I}\boldsymbol{\mu}\theta} v e^{-\mathbf{I}\boldsymbol{\mu}\theta}$$

And $T_x(v)$ describes a rotation of an angle 2θ around the axis $\boldsymbol{\mu}$ of the vector v.

Note that due to the "sandwich" form of the action T_x, an element and its opposite in the **Spin** group are sent to the same action on V. Mathematically, **Spin**(3) realizes a double-cover of the rotation space $SO(3)$ through the morphism

$$\psi \colon \pm e^{\mathbf{I}\boldsymbol{\mu}\theta} \mapsto R_{(\boldsymbol{\mu},2\theta)}$$

where $R_{(\boldsymbol{\mu},2\theta)}$ is uniquely defined in $SO(3)$ as being the rotation of axis $\boldsymbol{\mu}$ and angle 2θ. This remark can be used to understand the difference between vectors and spinors. While vectors are "insensitive" to the sign of an element in the **Spin** group, spinors will be elements transformed by the **Spin** group in a way that two elements of opposite signs are not sent over the same transform **[Wikipedia contributors, 2021b]**. A simple way to realize that is to define the **Spin** action over the spinor space to be a one-sided[4] product. Then **Spin**(3) would act on the hypothetic spinor space S by

$$\mathbf{Spin}(3) \to \mathcal{L}(S) \qquad S \to S \tag{2.14}$$

$$x \mapsto U_x \qquad U_x \colon s \mapsto xs \tag{2.15}$$

An obvious requirement for the spinor space is that it must be stable under the left action of **Spin**(3). A suitable candidate is $\mathcal{C}\ell_{3+}$ **[Matthew R. and Kosowsky, 2006]**, the set of even graded elements defined in proposition 12.

Definition 10. *The spinor space **Spinor**(3) is the subalgebra of even-graded elements in $\mathcal{C}\ell_{3+}$ acted upon by the **Spin**(3) group as in* (2.15).

[4] we choose left-sided action in the present document each time the choice is proposed

Given a vector v, a spinor s and an element $x = e^{\mathbf{I}\mu\theta}$ from the **Spin** group, x induces a rotation with an angle 2θ while it only rotates s by an angle θ. This behaviour is exactly what was displayed by the Jones vector / coherency matrix duo, where a rotation of an angle θ on the Jones vector repercuted as a rotation of an angle 2θ on some coordinates of the coherency matrix (1.8). We can now transition smoothly to next section that shows how a bivariate signal and its polarization analysis can be embedded in the Clifford algebra and how it relates to the quaternion algebra.

2.2.2 Embedding of bivariate signal analyis in $\mathcal{C}\ell_3$

The spinor structure of the Jones vector

The Jones vector is obtained in the vector framework by taking the analytic signal of each component. The resulting vector belongs to $\mathbb{C}^2$, a set that is isomorphic in the Clifford algebra to $\mathcal{C}\ell_{3+}$. The relation between the Jones vector and the coherency matrix depends on the algebra in which it is defined. In a complex frame, it is not possible to define these quantities. In the vector frame, the coherency matrix C_x is obtained from the jones vector ε by $C_x = \varepsilon\varepsilon^*$ (see section 1.2.3), while in the quaternion representation the relation reads: $C_x = x_+\boldsymbol{j}\overline{x_+}$ (see section 1.4.4) where x_+ is the quaternion analytic signal of x. The affirmation that the quaternion couple (x_+, C_x) and vector-matrix couple (ε, C_x) represent the same quantity is supported by the fact that the same scalar elements can be separated inside of each expression. Further, both representations yield to a linear combination of the Stokes parameters in their respective basis. We argued earlier that an interesting angle on the relation between ε and c is to observe how they are affected by a same transformation. If a signal x is rotated by the transformation $R(\theta)$, then so is its Jones vector ε. The coherency matrix c however undergoes the transformation $R(\theta)cR(-\theta)$. Using the structure of the Clifford algebra, the fact that a change of frame does not yield the same transform over ε or c highlights that both objects are not of a common nature. One of them, the matrix c behaves like a vector and the other one, the Jones vector behaves like a spinor. The difference in nature is obvious in the vector representation as one is a column vector and the other a square matrix, but the information is lost in the quaternion algebra. Observe that between the vector and the Clifford frame, the matrix becomes the vector while the vector becomes the spinor. So that interestingly, the signal is a spinor relatively to its polarization state represented as a vector. Next paragraph gives the concrete embeddings and relations between the Jones vector and the coherency matrix in the Clifford algebra.

Embedding of the Jones vector, the coherency matrix and the Stokes parameters in $\mathcal{C}\ell_3$

We wrote just before that the coherency matrix behaves like a vector, which is not completely correct in our context. Vectors, in the Clifford algebra, are elements which are acted upon by the **Spin** group for instance, such that for a given element of the spin group, the vector

undergoes a precise rotation. One part of the coherency matrix however, is not affected by any transformation of the **Spin** or **Pin** group, as it is precisely the part invariant to unitary transforms: the trace. The coherency matrix was decomposed over the Pauli basis as

$$c = S_0 I + S_1\sigma_2 + S_2\sigma_3 + S_3\sigma_1$$

It is obvious with this writing that any transform with a sandwich shape leaves the first term invariant due to the commutativity of I with every matrix:

$$RcR^{-1} = S_0 I + R(S_1\sigma_2 + S_2\sigma_3 + S_3\Sigma_1)R^{-1}$$

The vector part is precisely given by $S_1\sigma_2 + S_2\sigma_3 + S_3\Sigma_1$ and can be visualized on the Poincare sphere as a vector of $\mathbb{R}^3$. We can now precise the embedding of each object in the Clifford algebra. The Jones vector, as a spinor must be embedded in $\mathcal{C}\ell_{3+}$. It is sent through a transform ξ left to determine on the sum of a vector that contains all the polarization information and an element of the center of $\mathcal{C}\ell_3$ that contains power information. Whatever the embedding we choose for one of the quantities, it would be possible to build consistent embeddings for the other so we need to fix one of them to propose a precise basis for our embedding. Remember that the Jones vector of $y(t) = (y_1(t), y_2(t))^T$ is composed of the analytic signals $y_{1+}(t), y_{2+}(t) = a_1(t)e^{i\phi_1(t)}, a_2(t)e^{i\phi_2(t)}$. We drop the dependency in t in the rest of our calculations as it is of no importance here. Let

$${y_+}^{\mathcal{C}\ell_3} = a_1 e^{e_{12}\phi_1} + \mathbf{I}a_2 e^{e_{12}\phi_2}$$

Then we must determine an application $\xi\colon \mathcal{C}\ell_{3+} \to V$ such that for $u \in \mathcal{C}\ell_{3+}, x \in \mathbf{Spin}(3)\xi(xu) = \widehat{x}\xi(u)x^{-1}$, in clear, ξ must send the Jones vector on polarization vector such that the image of the Jones vector transformed by x is the polarization vector transformed by x. Applying this condition to $u = 1$ yields:

$$\forall x \in \mathbf{Spin}(3) \subset \mathcal{C}\ell_{3+}, \xi(x) = \widehat{x}\xi(1)x^{-1}$$

Therefore we set $\xi\colon x \mapsto \widehat{x}\xi(1)x^{-1}$ and to determine $\xi(1)$ we use the condition that ξ must be vector-valued. This leaves $\xi(1) = e_i$ for $i = 1, 2, 3$ but given the choice we made for the Jones vector, it is more convenient to choose $\xi(1) = e_3$. Then,

$$\xi({y_+}^{\mathcal{C}\ell_3}) = S_1 e_+ S_2 + S_3$$

which as we see, is exactly what could be expected. To build a scalar-valued quantity from ${y_+}^{\mathcal{C}\ell_3}$ the simplest is to take the norm which in this case writes

$$\|{y_+}^{\mathcal{C}\ell_3}\|^2 = {y_+}^{\mathcal{C}\ell_3}\overline{{y_+}^{\mathcal{C}\ell_3}} = S_0$$

Relation with the quaternion-analysis

The relation between the quaternion $\mathbb{H}$ space and the Clifford algebra $\mathcal{C}\ell_3$ is particularly rich as they are related by several isomorphisms. First, there is an isomorphism of algebra between the even-graded elements of $\mathcal{C}\ell_3$ and $\mathbb{H}$ realized by:

$$\chi\colon \mathcal{C}\ell_{3+} \to \mathbb{H} \begin{cases} 1 \mapsto 1 \\ e_{12} \mapsto -\boldsymbol{k} \\ e_{23} \mapsto \boldsymbol{j} \\ e_{13} \mapsto -\boldsymbol{i} \end{cases}$$

Note that χ preserves the highest level of structure possible: all linear as well as group relations are faithfully transmitted by χ such that $\forall \lambda \in \mathbb{R}, u, v, w \in \mathcal{C}\ell_{3+} \chi(\lambda u + vw) = \lambda\chi(u) + \chi(v)\chi(w)$. On the other hand, the isomorphism that sends the vector space V from the Clifford algebra to the pure imaginary quaternions is a linear-space isomorphism: it does not render the product relation between terms. This second isomorphism is defined by:

$$\rho\colon \mathcal{C}\ell_{3+} \to \mathbb{H} \begin{cases} e_1 \mapsto \boldsymbol{i} \\ e_2 \mapsto \boldsymbol{j} \\ e_3 \mapsto \boldsymbol{k} \end{cases}$$

Note that to define the action of the Clifford group over the vector subspace V, only linear operations in V are necessary. As elements of the **Spin** group are in $\mathcal{C}\ell_{3+}$, they can also be represented through χ in the quaternion space. In conclusion, through these two isomorphisms, the quaternion space can embed vectors, spinors and the **Spin** group and its actions on vectors and spinors from the Clifford algebra.

2.2.3 conclusion

Another way to read this section is to see how the quaternion field is unique from multiple point of views. From the Frobenius theorem, it is the bigger field over $\mathbb{R}$ with nice properties, and from what precedes, when embedded in a geometric algebra, it is isomorphic to several subspaces in that algebra. Not all spaces can be algebra that embed vectors and spinors in the same structure. Actually, we believe that the Quaternion algebra might be the only one with this property. This is why we did not push further the idea to represent signals and their polarization state as homogeneous quantities in one algebra. Of course, geometric algebras could be used, taking advantage of their natural division in different subspaces where elements are not acted upon similarly be the same transformations. If geometric algebras have been useful to put under light the spinor role played by quaternions in some equations and their vector role in others, we believe they are not the right space to conduct the analysis of polarization in general. Indeed, it appears that what really matters to study polarization is to be able to quantify how some transformations act on some objects. Therefore, we need a

structure where actions of elements on others can be represented as agreeably as possible. This leads us quite naturally to a matrix-vector and matrix-matrix formalism that we developp in next chapter.

3 Towards a geometrical representation of trivariate signals

> *Numbers measure size, groups measure symmetry.*
>
> — **M.A. Armstrong**
>
> Groups and Symmetry **[Armstrong, 2013]**

Contents

Between the bivariate and the n-variate signal, trivariate signals and their polarization have been treated as a particular case due to their physical significance in the optics **[Gil, 2007]** and oceanographic **[Lilly, 2011]** communities. The usual analysis relies on a matrix-vector formalism and the definition of scalar quantities called Stokes parameters. We believe however that trivariate signals analysis rise specific questions and problematics compared to bivariate analysis that are worth being explored. Trivariate analysis leads the way to the generalization of polarization study in n dimensions. We propose a new definition of polarization quantities through invariance rather than a decomposition over an arbitrary basis. All the work proposed here generalizes smoothly to n dimensions.

3.1 Introduction

Remember that we started this general reflection on polarization through one leading idea: how can the representation space of the signal help us process polarization. We showed in the last two chapters that a good representation for polarization analysis of monochromatic waves in two dimensions takes into account the particular behavior that the Jones vector has towards the coherency matrix and the Stokes parameters. We also showed that to process instantaneous polarization, being able to factorize the signal under a shape parametrized by angles is a real plus. In this section where the goal is to broaden the scope of our analysis to n-variate signals, we start a general reflexion upon the nature of polarization and its definition through invariance. The question to represent signals in algebras that embed vectors and spinors is dropped for the reason that dealing with the geometric algebra of $\mathbb{R}^n$ is all but practical. We prefer the more tractable vector-matrix approach. We exemplify our approach on trivariate signals, but almost every result generalizes smoothly from trivariate to n-variate signal processing. In comparison, bivariate processing appears to be an outlier. We believe the reason to be the commutativity between rotations in 2 dimensions, a property that does not generalize to more dimensions. Even if we have not yet formalized this idea, we think it has drastic consequences from the point of view of polarization and embedding of signals that consequently do not generalize to higher dimensions. On the other hand, as we will argue at the end of this chapter, it seems to us that polarization is ultimately a bivariate notion, illustrated through the fact that there is no higher-dimensional Poincaré sphere.
The first part of this chapter introduces the modulated-wave model for trivariate signals and a polar factorization based on the work by **[Lilly, 2011]**. It is shown how this polar factorization defines instantaneous orientation and geometrical parameters. We give in this part a first definition of instantaneous frequency through invariance. In a second time, we drop the time-dependency and we focus on the structural behavior of the coherency matrix. We show that from the coherency matrix, a quantity invariant to the action of $SO(n)$ can be extracted, and we call this quantity of "polarization", in the sense that it describes the shape of the signal and is independent of the choice of a coordinate frame.

3.2 A polar factorization for trivariate signals

Trivariate signals are embedded in the 3-dimensional euclidian vector space as:

$$\boldsymbol{x}(t) = (x_1(t), x_2(t), x_3(t))^T$$

Supposing this vector results in the recording of a trivariate physical quantity rather than three independent quantities recorded separately, common information is shared between the components. For instance, a spatial oscillation is a signal where all components oscillate at the same pulsation and the resulting signal $\boldsymbol{y}$ writes like:

$$\boldsymbol{y}(t) = (a_1 \cos(\omega t + \phi_1), a_2 \cos(\omega t + \phi_2), a_3 \cos(\omega_t + \phi_3))^T \tag{3.1}$$

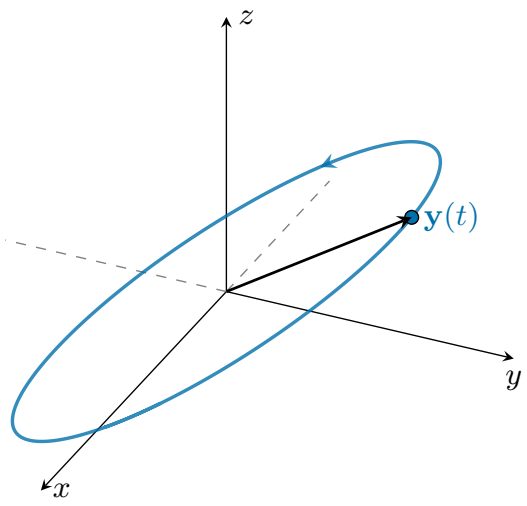


Fig. 3.1. – A monochromatic trivariate signal draws an ellipse in the $x-y-z$ space. The sense of rotation is indicated with an arrow.

It is contained in a plane where it describes an ellipse with period $\frac{2\pi}{\omega}$, see figure 3.1. The complexification of the monochromatic signal $\boldsymbol{y}$ is called the Jones vector and defined by:

$$\boldsymbol{y}_+(t) = (y_{1+}(t), y_{2+}(t), y_{3+}(t)) \tag{3.2}$$

It factorizes in:

$$\boldsymbol{y}_+(t) = \|y_+(t)\| e^{\boldsymbol{i}(\omega t+\phi_1)}(\alpha_1, \alpha_2 e^{\boldsymbol{i}(\phi_2-\phi_1)}, \alpha_3 e^{\boldsymbol{i}(\phi_3-\phi_1)})^T \tag{3.3}$$

with $0 \leqslant \phi_1 < 2\pi$ and for $1 \leqslant i \leqslant 3$: $\alpha_i = \dfrac{a_i}{\sqrt{{a_1}^2 + {a_2}^2 + {a_3}^2}}$. The first scalar-term in the expression contains the pulsation ω, a phase factor of no significance, and the norm of the signal. It describes a univariate monochromatic wave. The vector part contains terms specific to multivariate signals. Two phase differences and two independent amplitudes (the squared α_i sum to one) specify a polarization state. We can infer from the shape of the signal (see figure 3.1) that it must contain information on the polarization plane and the shape of the polarization ellipse. A plane can be parametrized by its normal vectors, hence two supplementary parameters compared with the identification of polarization for bivariate vectors.

3.2.1 Parametrization of an ellipse in three dimensions

The parametric equation of an ellipse in three dimensions is of the shape (3.1). An elliptic signal in a plane can be easily parametrized in relation with the geometry of the ellipse if the coordinate system is aligned with the axis of the ellipse. For instance, a plane elliptic signal

with major axis of size a aligned with the x-coordinate and minor axis of size b aligned with the y-coordinate of an orthonormal coordinate system writes in this system:

$$\boldsymbol{y}'(t) = (a\cos(\omega t), b\sin(\omega t), 0)^T \tag{3.4}$$

which goes with the analytic representation:

$$\boldsymbol{y}'_+(t) = \sqrt{a^2+b^2}e^{\boldsymbol{i}\omega t}(\cos(\chi), -\boldsymbol{i}\sin(\chi), 0)^T \tag{3.5}$$

where the ellipticity $\chi \in [-\frac{\pi}{4}, \frac{\pi}{4}]$ measures the disparity between a and b such that $|\tan(\chi)| = |\frac{b}{a}|$. The sign of χ accounts for the direction in which the ellipse is described, positivity stands for trigonometric direction and negativity for the opposite. Clearly, the transformation that aligns the axis of the coordinate frame with the axis of the ellipse is special orthogonal: it goes from an orthonormal basis to another without inverting the direction of one axis. This means there exists a matrix O in $SO(3)$ that sends $(\boldsymbol{y}', \boldsymbol{y}'_+)$ on $(\boldsymbol{y}, \boldsymbol{y}_+)$. Special orthogonal transformations describe rotations that can be factorized as rotations around the basis axis[1]. The transformation O gives the orientation of the plane, and the tilting of the ellipse, in its original coordinate system. The authors in **[Lilly, 2011]** identify the transformation O as follows:

$$O(\theta, \alpha, \beta) = J_3(\alpha)J_1(\beta)J_3(\theta) \tag{3.6}$$

where the rotation matrices J_7, J_2 respectively rotate around the x-axis and the z-axis and write explicitely as:

$$J_7(\beta) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\beta & -\sin\beta \\ 0 & \sin\beta & \cos\beta \end{pmatrix} \qquad J_2(\alpha) = \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Under the transformation O, an ellipse aligned on the $x-y$ axes is tilted in its plane with a precession angle $\theta \in [-\pi, \pi]$. The plane containing the ellipse is then tilted with the zenith angle $\beta \in [0, \frac{\pi}{2}]$ around the x-axis and a last rotation with the azimuth angle $\alpha \in [-\pi, \pi]$ around the z-axis sends the plane to its final position see section 3.2.1. Note that J_7, J_2 do not commute in expression (3.6).

Parameters θ and χ have the same interpretation as the angles of the same name in the bivariate framework see section 1.2.2. Parameters α and β position the plane in the trivariate space. A normal vector to the plane containing the ellipse is $(\sin\alpha\sin\beta, -\cos\alpha\sin\beta, \cos\beta)^T$. For a fixed order of the matrix J_3 and J_1 and the intervals $\alpha, \beta \in [0, 2\pi[$, $\theta \in [0, \pi[$, $\chi \in [-\frac{pi}{2}, \frac{\pi}{2}[$, there is a one to one correspondance between the set of parameters $(\alpha, \beta, \theta, \chi)$ and the set of four free parameters in the vector $(\alpha_1, \alpha_2 e^{\boldsymbol{i}(\phi_2-\phi_1)}, \alpha_3 e^{\boldsymbol{i}(\phi_3-\phi_1)})^T$ **[Lilly, 2011]**. We summarize this result in proposition 14.

[1] This is the Euler parametrization of a rotation

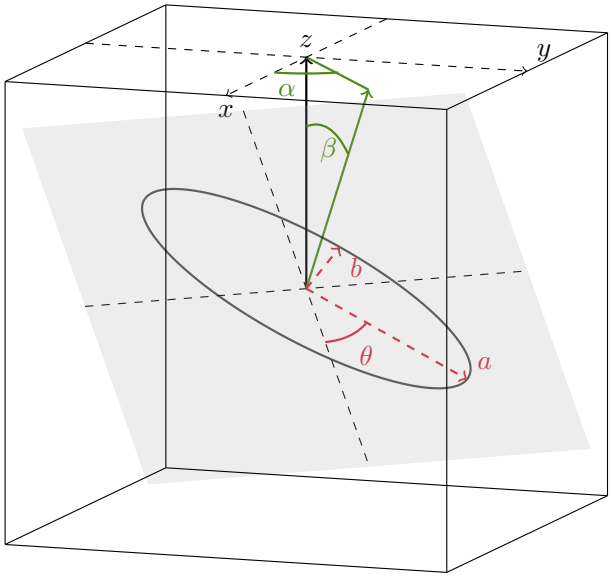


Fig. 3.2. – schematic of an ellipse in three dimensions with semi-major axis of length a and semi-minor axis of size b, precession angle θ, zenith angle β and azimuth angle α. The plane containing the ellipse is darkened, and the $x-y$ axis in this plane are represented in dashed lines. Parameters characterizing the ellipse in the plane are in red, and parameters characterizing the plane are in green. The green axis represents the normal to the plane containing the ellipse, while the black axis is the z axis of the coordinate system. On the upper face of the cube, the dashed lines represent the $x-y$ axis of the coordinate system.

Proposition 14. *Given a monochromatic trivariate signal $\boldsymbol{y}$ with pulsation ω and its Jones vector $\boldsymbol{y}_+$, there is a unique set of parameters $(a, \theta, \alpha, \beta, \chi)$ such that*

$$a > 0 \qquad \alpha, \beta \in [0, 2\pi[\qquad \theta \in [0, \pi[\qquad \chi \in \left[0, \frac{\pi}{4}\right]$$

and

$$\boldsymbol{y}_+(t) = a(t)e^{i\omega t} J_7(\alpha) J_2(\beta) J_7(\theta) \begin{pmatrix} \cos\chi \\ -\boldsymbol{i}\sin\chi \\ 0 \end{pmatrix} \tag{3.7}$$

Proof. See [Lilly, 2011] for the proof of this result. □

The expressions of $(\alpha, \beta, \theta, \chi)$ as functions of the cartesian coordinates of $\boldsymbol{y}$ can be found in [Lilly, 2011] but will not be used here. Interestingly, the vector $(\cos\chi, -\boldsymbol{i}\sin\chi, 0)^T$ can be seen as the product

$$\begin{pmatrix} \cos\chi & \boldsymbol{i}\sin\chi & 0 \\ -\boldsymbol{i}\sin\chi & \cos\chi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos\chi \\ -\boldsymbol{i}\sin\chi \\ 0 \end{pmatrix}$$

Calling $J_1(-\chi)$ the matrix identified in the above expression, the expression (3.7) can be further factorized as:

$$\boldsymbol{y}_+(t) = a(t)e^{i\omega t}J_7(\alpha)J_2(\beta)J_7(\theta)J_1(-\chi)\begin{pmatrix}1\\0\\0\end{pmatrix} \tag{3.8}$$

and the product $J_7(\alpha)J_2(\beta)J_7(\theta)J_1(-\chi)$ belongs to the set of unitary matrices $SU(3)$. There is an injection from the set of Jones vector to the set of unitary matrices $U(3)$, meaning that any polarization state can be visualized as a matrix of $SU(3)$. This is in a way dual to the representation of a polarization state through the coherency matrix, which embeds the Jones vector in $\mathfrak{su}(3)$, the Lie algebra associated with $SU(3)$ that we will define later.

3.2.2 Trivariate modulated oscillation

In a stationary context, the association of the signal with a polarization ellipse is straightforward if operated frequency-wise, because pure oscillation in three dimensions are exactly the signals that parametrize ellipses in space. Four parameters are necessary to describe the polarization ellipse and they can all be advantageously embedded in a matrix of $U(3)$. This would enable for instance the study of evolving polarization as a trajectory in $U(3)$. The problematic of defining polarization for non-stationary signals remains the same for bivariate and trivariate signals. If the instantaneous frequency can be obtained as the function that satisfies a first-order moment equation, polarization satisfies no such equation[2]. Thus, we relied entirely, in the bivariate case, on a given decomposition of the coherency matrix, see section 1.3.3. The instantaneous Stokes parameters are defined as the coordinates of the coherency matrix at time t on the Pauli basis. Alternatively, when embedding the signal in the quaternion space, we used the Euler polar form whose coordinates at time t defined instantaneous angles that parametrized the polarization ellipse in proposition 20. In both cases, the definition of instantaneous geometrical parameters was made possible thanks to a decomposition a priori of an object built from the signal. Therefore, we argue that in the trivariate case, instantaneous geometrical parameters can be obtained through (3.8).

[2]Actually, the Stokes parameters do satisfy first-order moment equations, see **[Flamant et al., 2017]**

3.2.3 Instantaneous parameters for multivariate modulated oscillations

Before introducing the main result of this subsection, we introduce the Gell-man matrices [Gil, 2007] noted $(\lambda_i)_{1\leqslant i\leqslant 8}$. The Gell-man matrices are an extension to $U_3(\mathbb{R})$ of the Pauli matrices, they write:

$$\lambda_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \lambda_2 = \begin{pmatrix} 0 & -\boldsymbol{i} & 0 \\ \boldsymbol{i} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\lambda_4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad \lambda_5 = \begin{pmatrix} 0 & 0 & -\boldsymbol{i} \\ 0 & 0 & 0 \\ \boldsymbol{i} & 0 & 0 \end{pmatrix}$$

$$\lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad \lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\boldsymbol{i} \\ 0 & \boldsymbol{i} & 0 \end{pmatrix} \quad \lambda_8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}$$

With the inner-product $\langle A, B\rangle = \mathrm{tr}(AB^\dagger)$ they form an orthogonal basis of the space of traceless hermitian matrices. These matrices are related to the group $SU(3)$ through exponientation.

Proposition 15. *given a traceless hermitian matrix H, the matrix $U_\theta = e^{\boldsymbol{i}\theta H}$ belongs to $SU(3)$. In particular:*

$$e^{\boldsymbol{i}\chi\lambda_1} = J_1(\chi) \qquad e^{\boldsymbol{i}\theta\lambda_7} = J_7(\theta) \qquad e^{\boldsymbol{i}\beta\lambda_2} = J_2(\beta) \tag{3.9}$$

Proof. Let X be a traceless hermitian matrix. Then $\boldsymbol{i}X$ is skew-hermitian and satisfies

$$X + X^\dagger = 0$$

As X and $X^\dagger$ commute we can compose this relation with the exponential and write

$$e^{X+X^\dagger} = e^X e^{X^\dagger} = Id$$

Given that $e^{X^\dagger} = (e^X)^\dagger$, the matrix $U = e^X$ satisfies $UU^\dagger = \mathrm{Id}$ and belongs to $U(n)$. Then, note that $\det(U) = e^{\mathrm{tr}(X)}$ to conclude. Concerning the computation of $e^{i\lambda_k}_{k\in\{1,2,7\}}$, it is useful to note that $\boldsymbol{i}\lambda_k$ for $k \in \{1, 2, 7\}$ is "almost" a square root of $-\mathrm{Id}$. Indeed $\lambda_2^2, \lambda_7^2, \lambda_1^2$ have the same diagonal structure with two -1 and one 0 on the diagonal. Just the position of the 0 varies. We then use the result that for A such that $A^2 = -\mathrm{Id}$ then $e^{\theta A} = \cos(\theta)\mathrm{Id} + \sin\theta A$. □

For a proof that the exponential map sends the Lie algebra $\mathfrak{su}(3)$ on $SU(3)$, we refer to ***[Fegan, 1991]****.*

The exponential shape of the unitary matrices will be used in the next proposition.

Proposition 16. *Given a non-stationary signal $\boldsymbol{x}$, and its three-variate analytic signal $\boldsymbol{x}_+ = (a_1(t)e^{i\varphi_1(t)}, a_2(t)e^{i\varphi_2(t)}, a_3(t)e^{i\varphi_3(t)})^T$, at each t there is a unique set of parameters $\varphi(t), \alpha(t), \beta(t), \theta(t), \chi(t)$ contained in the intervals defined in proposition 14 such that:*

$$\boldsymbol{x}_+(t) = \boldsymbol{a}(t)e^{i\boldsymbol{\varphi}(t)Id}e^{i\alpha(t)\lambda_7}e^{i\beta(t)\lambda_2}e^{i\theta(t)\lambda_7}e^{-\chi(t)\lambda_1}\begin{pmatrix}1\\0\\0\end{pmatrix} \tag{3.10}$$

Then, $\boldsymbol{a}(t) = \sqrt{a_1^2(t) + a_2^2(t) + a_3^2(t)}$ and $\boldsymbol{\varphi}(t) = \frac{\varphi_1(t)+\varphi_2(t)+\varphi_3(t)}{3}$.

proposition 16 basically identifies a unique polar form for three-variate analytic signals. The new formulation with exponentials has the interest to present a homogeneous writing in which each angle rotates around a matricial axis. Unicity of the parameters is derived from a similar result in **[Lilly and Olhede, 2012]**. Modulated oscillations are not always interpretable. For instance, a vector made of n independent white noise should not be described in terms of instantaneous ellipse, as it implies joint structure in small interval of time, which has no reason to exist in this case. A modulation condition can be found in **[Lilly, 2011]**, roughly if the derivatives of the angles $(\theta, \alpha, \beta, \chi)$ and of the amplitude $a(t)$ are small enough compared to $|\varphi'(t)|$, the modulated oscillation model holds. With these conditions, polarization angles and amplitudes are smooth functions of time. Examples of modulated trivariate signals are given in figure (figure 3.3).

3.2.4 Stokes parameters

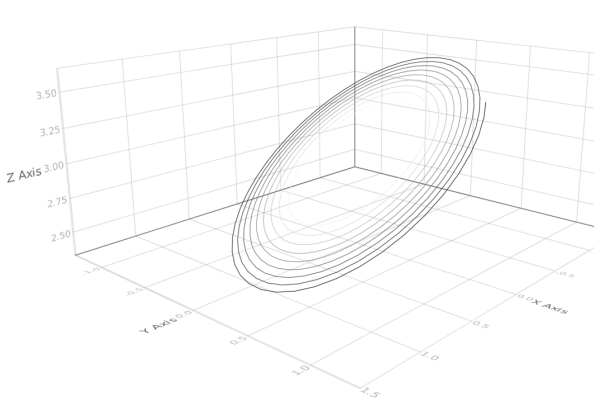
The coherency matrix for a trivariate signal $\boldsymbol{x}$ with Jones vector $\boldsymbol{x}_+$ is :

$$C_x = \boldsymbol{x}_+\boldsymbol{x}_+{}^\dagger = \mathrm{tr}(C_x)Id + C_x$$

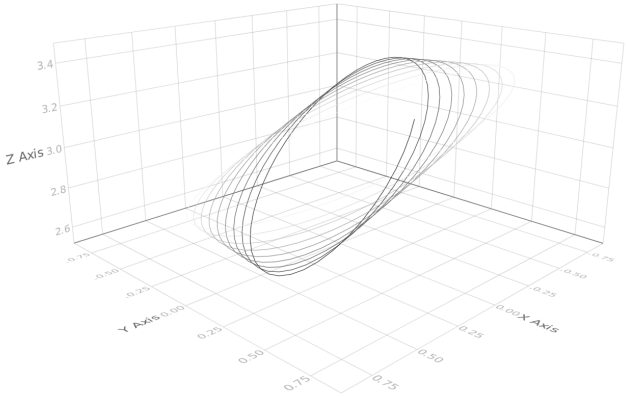
where C_x is a traceless, hermitian (by construction) matrix in $\mathfrak{su}_3(\mathbb{C})$. Following the remark in section 3.2.3 about the Gell-man basis, the coherency matrix C_x has a unique set of coordinates on the Gell-man basis that we call its Stokes parameters, noted $(\Gamma_i)_{0\leqslant i\leqslant 8}$ **[Gil, 2007]**. The Stokes parameter Γ_0 is such that:

$$C_x = \Gamma_0\mathrm{Id} + \Gamma_1\lambda_1 + \Gamma_2\lambda_2 + \ldots\Gamma_8\lambda_8$$

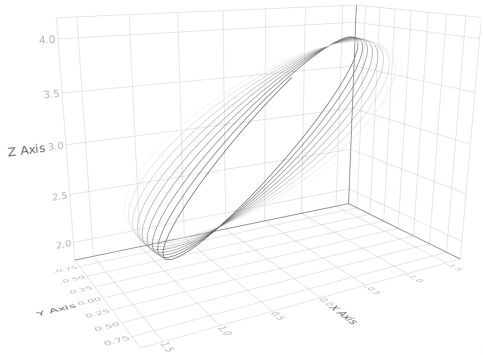
The first parameter Γ_0 carries information on the total energy of the signal, contained in the trace of C_x. All the other Stokes parameters depend on a mixture of the components



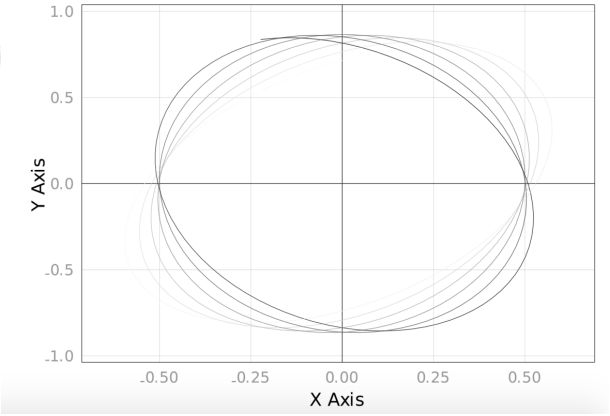
(a) Amplitude modulation



(b) Ellipticity modulation



(c) Azimuth modulation



(d) Tilting modulation

Fig. 3.3. – Four examples of non-stationary modulated trivariate signals where only one geometrical parameter varies with time. The instantaneous frequency ϕ' ranges linearly from 12π to 18π for all figures. On (figure 3.3a), the amplitude $\mathbf{a}(\mathbf{t})$ varies in $[0.5, 1]$ while all other parameters are held constants. In figure (figure 3.3b), the ellipticity χ varies resulting in a deformation of the ellipse in a constant plane. Third figure (figure 3.3c) shows a signal where the azimuth angle varies from 0 to $\pi/5$. The plane containing the ellipse follows a rotation around the x-axis, while the ellipse itself remains unchanged. Figure (figure 3.3d) is the evolution of Θ, the tilting of the ellipse in the plane from 0 to $\pi/4$. Time is visualized as a gray scale on the curve with black for $t = 0$.

of the Jones vector and are therefore called “Geometric parameters". Their interpretation is straightforward in the 2D case, where for instance the signal lives in the $x-y$ plane. Interpretation of the Stokes parameters in the 3D case was addressed for instance in **[Gil, 2014]**. Interestingly, the Gell-man matrices are also related through the exponential mapping (see proposition 15) to rotation matrices over $\mathbb{C}^3$. The exponentiation $e^{i\theta\lambda_i}$ of a Gell-man matrix gives a unitary matrix in $SU(3)$ representing a rotation of an angle θ around a particular axis of $\mathbb{C}^3$ **[Stillwell, 2008]**. This result was exploited in the factorization of the trivariate analytic signal. Rather than relating the Stokes parameters to a parametrization of the polarization ellipse through direct expressions, we propose an analysis of the information contained in the coherency matrix through invariances.

3.3 Invariance analysis of the coherency matrix

Before introducing the goal of this section, we need a very small theoretical background that we expose hereafter. In all the section $\mathfrak{su}(n)$ is the vector space of all skew-hermitian traceless matrices, and $\mathfrak{u}(n)$ is the vector space of skew-hermitian matrices. Let $x \in \mathbb{C}^n$ be a n-variate signal and x_+ its corresponding Jones vector. Then the coherency matrix C_x multiplied by the imaginary unit belongs to $\mathfrak{u}(n)$. Indeed, $(\boldsymbol{i}y_+y_+^\dagger)^\dagger = \bar{\boldsymbol{i}}y_+y_+^\dagger = -\boldsymbol{i}C_x$. We have already noted in section 2.2 that if x_+ undergoes a linear transformation represented by the matrix U, its coherency matrix C_x undergoes a linear transformation related to U. We call the action of U over the coherency matrix the Adjoint action and denote it $Ad(U)$.

Definition 11. *(Adjoint action) Let $(U(n), \times)$ denote the set of unitary matrices and $\mathfrak{u}(n)$ the set of skew-hermitian matrices. Every matrix U in $U(n)$ can be mapped to an automorphism of $\mathfrak{u}(n)$ as follows:*

$$Ad\colon U(n) \to \mathcal{L}(\mathfrak{u}(n)) \qquad \textit{with} \qquad Ad(U)\colon \mathfrak{u}(n) \to \mathfrak{u}(n)$$
$$U \mapsto Ad(U) \qquad\qquad\qquad\qquad X \mapsto UXU^\dagger$$

Then,

$$y_+ = Ux_+ \Leftrightarrow C_y = UC_xU^\dagger = Ad(U)\cdot C_x$$

We can check that for all $U \in U(n)$ and all $X \in \mathfrak{u}(n)$ (resp. $\mathfrak{su}(n)$)

$$Ad(U)\cdot X \in \mathfrak{u}(n) \text{ (resp. } \mathfrak{su}(\mathfrak{n}))$$

which is obtained simply by computing $[Ad(U)\cdot X]^\dagger = -Ad(U)\cdot X$ and $\mathrm{tr}(Ad(U)\cdot X) = \mathrm{tr}(X)$. This result also shows that hermitian matrices and traceless hermitian matrices are stabilized by the adjoint action of $U(n)$. Generally, everything we prove on skew-hermitian matrices in this part is also true for hermitian matrices, hence for coherency matrices. This is due to the relation $\boldsymbol{i}\times$ skew-hermitian $=$ hermitian.

Definition 12. *We say that a coherency matrix C_x is invariant under the action of U if:*

$$Ad(U).C_x = C_x$$

Next result is immediate.

Proposition 17. *The matrix C_x is invariant under the action of U if and only C_x and U commute.*

The next two definitions introduce the notion of invariance for a subspace, central to this part.

Definition 13. *A subspace $W \in \mathfrak{u}_n$, is said to be invariant under the action of U if*

$$Ad(U) \cdot W \subseteq W$$

Definition 14. *A subspace $W \in \mathfrak{u}_n$ is said to be invariant under the action of a subgroup $H \subseteq U(n)$ if*

$$\forall U \in H, Ad(U) \cdot W \subseteq W.$$

A subspace is invariant under the action of a subgroup if and only if it is invariant under the action of each element of the subgroup.

Definition 15. *A subspace W is said to be irreducible under the action of a subgroup H if W contains no invariant subspace other than the null subspace and itself under the action of H.*

Given a basis $\mathcal{B}$ of $\mathfrak{u}(n)$, and given that $Ad(U)$ is a linear application on $\mathfrak{u}(n)$ one can write $Ad(U)$ under its matricial form in basis $\mathcal{B}$. For each U this highlights the existence of invariant subspaces under $Ad(U)$. Below we compute explicitely the coordinates of the Adjoint actions of $U(2)$ on $\mathfrak{u}(2)$ in the Pauli basis.

3.3.1 Coordinates of the Adjoint action of $U(2)$ over $\mathfrak{u}(2)$

We assume the following result which empirically is true in 2 and 3 dimensions:

Proposition 18. *For U in $U(2)$, there are $r, t, u, s \in \mathbb{R}$ such that*

$$U = e^{ir\sigma_0} e^{is\sigma_1} e^{it\sigma 2} e^{iu\sigma_3} \tag{3.11}$$

where $(\sigma_0, \sigma_1, \sigma_2, \sigma_3)$ are the Pauli matrices such that $\mathcal{B}_p = (i\sigma_0, i\sigma_1, i\sigma_2, i\sigma_3)$ defines an orthogonal basis of $\mathfrak{u}(2)$.

Notice that for U as defined in (3.11) the Adjoint application satisfies

$$Ad(U) = Ad(e^{ir\sigma_0}) \circ Ad(e^{it\sigma_1}) \circ Ad(e^{iu\sigma 2}) \circ Ad(e^{is\sigma_3})$$

Hence, any adjoint action on $\mathfrak{u}(2)$ can be derived from the action of the four matrices identified in (3.11). The following property provides coordinates for each element. We recall that

$$\sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

then,

$$e^{ir\sigma_0} = \begin{pmatrix} e^{ir} & 0 \\ 0 & e^{ir} \end{pmatrix} \quad e^{is\sigma_1} = \begin{pmatrix} \cos(s) & i\sin(s) \\ i\sin(s) & \cos(s) \end{pmatrix} \quad e^{it\sigma_2} = \begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix} \quad e^{iu\sigma_3} = \begin{pmatrix} e^{iu} & 0 \\ 0 & e^{-iu} \end{pmatrix}$$

Proposition 19. *The matrices of the linear application $Ad(e^{ir\sigma_j})_{j\in 0..3}$ in the basis $\mathcal{B}_p$ are respectively:*

$$Ad(e^{ir\sigma_0}) \simeq \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}_{\mathcal{B}_p} \qquad Ad(e^{is\sigma_1}) \simeq \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos(2s) & -\sin(2s) \\ 0 & 0 & \sin(2s) & \cos(2s) \end{pmatrix}_{\mathcal{B}_p}$$

$$Ad(e^{it\sigma_2}) \simeq \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(2t) & 0 & \sin(2t) \\ 0 & 0 & 1 & 0 \\ 0 & -\sin(2t) & 0 & \cos(2t) \end{pmatrix}_{\mathcal{B}_p} \qquad Ad(e^{iu\sigma_3}) \simeq \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(2u) & -\sin(2u) & 0 \\ 0 & \sin(2u) & \cos(2u) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}_{\mathcal{B}_p}$$

From the matricial expressions in proposition 19 it is clear that the coordinate on σ_0 of a signal (which is its power) is unaffected by the transforms $Ad(e^{i\sigma_k})_{0\leqslant k\leqslant 3}$. However, all the other Adjoint actions fix one coordinate and rotate the other two. They act like rotations around successively $\sigma_1, \sigma_2, \sigma_3$. In terms of invariance, the subspace $\vec{\sigma_0}$ is invariant under the action of $U(2)$, while the subspace $\vec{\sigma_1}$ is invariant under the Adjoint action of the subgroup $\{e^{is\sigma_1} \mid s \in \mathbb{R}\}$.

3.3.2 Invariance theory

The coherency matrix contains all the information on the signal except its frequency. We separate this information in power, orientation and polarization. Orientation refers to information on the plane in which the signal is contained, while polarization refers to the ellipse it traces in this plane. In this section we aim to show that these quantities are not

homogeneous and can be separated by identifying some set of transforms to which they are invariant. We try to identify subgroups of transforms in $U(n)$ that separate these quantities. Any linear transformation of a matrix in $\mathfrak{u}(n)$ can be represented by the adjoint action of a matrix of $U(n)$. Hence we need to identify in $U(n)$ subgroups of matrices corresponding to: power change, orientation change and ellipticity change. In the same way, we need to identify elements in $\mathfrak{u}(\mathfrak{n})$ that are invariant by the action of two among three of these groups of transformation. These elements can then be said to carry "only" polarization, orientation or power information. However, one must realize that the information we propose to separate is not linearly independent, hence the objective to define a basis for the coherency matrix compatible with the segmentation of information we propose is not realistic.

Proposition 20. *There are only two subspaces of $\mathfrak{u}(n)$ invariant under the adjoint action of $U(n)$. They are*

$$\Gamma_n = \{\lambda Id \mid \lambda \in \mathbb{C}\} \qquad\qquad \mathfrak{su}(n) = \{X \in \mathfrak{u}(n) \mid tr(X) = 0\}$$

Proof. For all matrices U in $U(n)$, the adjoint action of U restricted to scalar matrices is the identity because scalar matrices commute with any matrix. We proved a stronger result than stated in the proposition, every matrix in W is actually invariant under the action of any matrix in $U(n)$. We already showed the invariance of $\mathfrak{su}(n)$. The fact that there are the only two invariant subspaces can be shown using **[Fegan, 1991]** □

In physical terms, a consequence of proposition 20 is that two signals y_1, y_2 are related by a unitary transformation $U \in U(n)$ if and only if they have the same amplitude or power. Of course, U might be indexed on time such that forall $t \in T$

$$y_2(t) = U(t)y_1(t) \iff \exists a(t), y_i(t) = a(t)u_i(t) \text{ with } \|u_i(t)\| = 1, i \in \{1, 2\}$$

We can derive the following definition from the previous remark.

Definition 16. *We say a quantity computed from a signal x is a power-quantity if it is invariant under the action of $U(n)$.*

The idea of what follows is to identify subgroups that have a particular interpretation in terms of transformation of the signal. We call the data of all Adjoint actions defined from elements of $U(n)$ over $\mathfrak{u}(n)$ a representation of $U(n)$. Computing coordinates for the representation of $U(n)$ over $\mathfrak{u}(n)$ is not practical in the general case. Another way to identify potentially interesting subgroups in $U(n)$ is to identify subspaces in $\mathfrak{u}(n)$ and consider the subgroups they are mapped to in $U(n)$. This is done in proposition 21.

Proposition 21. *Let $X \in \mathfrak{u}(n)$ then X decomposes as:*

$$\underbrace{\frac{1}{n}tr(X)I}_{\in\Gamma_n} + \underbrace{Re\left[X - \frac{1}{n}tr(X)\right]}_{\Sigma_1} + \underbrace{i\,Im\left[X - \frac{1}{n}tr(X)\right]}_{\Sigma_2} \tag{3.12}$$

where $\Sigma_1 \in \mathcal{A}_n^0$ belongs to the set of skew-symmetric real matrices with zero trace, $\Sigma_2 \in S_n^0$ is a real symmetric matrix with zero trace. The decomposition (3.12) is unique and derives from the following decomposition of $\mathcal{M}_n(\mathbb{R})$ as a direct sum for the Frobenius scalar product:

$$\mathcal{M}_n(\mathbb{R}) = \Gamma \oplus \mathcal{A}_n^0 \oplus S_n^0 \tag{3.13}$$

Through the exponential mapping, decomposition (3.13) restricted to $\mathfrak{u}(n)$ identifies three subgroups in $U(n)$. We have:

$$\exp\colon \begin{cases} \Gamma_n \cap \mathfrak{u}(n) & \to \Gamma_n \cap U(n) \\ \mathcal{A}_n^0 & \to O(n) \\ S_n^0 & \to \Sigma \subseteq \mathcal{H}_n^{++} \end{cases}$$

where $O(n)$ is the set of $n \times n$ real orthogonal matrices and $\mathcal{H}_n^{++}$ are the positive definite hermitian matrices. Transformations of a signal can be directly reported as transformations of a coherency matrix, thus as a transformation over $U(n)$. We propose the following definition

Definition 17. *We call "polarization content" of a signal x quantities computed from x, invariant under the action of $O(n)$ but not under the action of $U(n)$.*

It is clear that information about the shape of an object must be invariant under an orthonormal change of frame. The action of $O(n)$ on $\mathfrak{u}(n)$ is not irreducible and contains three invariant subgroups that are exactly the subgroups identified earlier in (3.13). Furthermore, the representation of $O(n)$ on each of these subgroups is irreducible. Interestingly, we have the following property:

Proposition 22. *The three subspaces of $\mathfrak{u}(n)$ identified in proposition 21 are invariant under the adjoint action of $O(n)$.*

Proof. Scalar matrices commute with all matrices so the action of a matrix O over Γ_n is the identity. Symmetric matrices are stabilized by the adjoint action of $O(n)$ (check that the new obtained matrix is diagonal in a orthonormal basis, hence symmetric), invariance of skew-symmetric matrices is deduced from the first two thanks to the direct sum (3.13). □

Polynomial invariants under the action of $O(n)$

We have identified subspaces of $\mathfrak{u}(n)$ invariant under the Adjoint action of $O(n)$. To characterize polarization, we are looking after scalar quantities that can be obtained as polynomials in the entries of the coherency matrix C_x. We consider a quantity to describe polarization if it is invariant under the action of $O(n)$ but not under the action of $U(n)$. We note Σ_1 and Σ_2 the orthogonal projections on $\mathcal{A}_n^0$ and S_n^0 respectively. Next propositions will lead to the identification of polynomials in the entries of the matrices Σ_1, Σ_2 invariant under the action of $O(n)$. We introduce a bit of vocabulary related to group actions (the Adjoint action is a group action).

Definition 18. *The orbit $\mathcal{O}(X)$ of an element $X \in \mathfrak{u}(n)$ under the action of $U(n)$ is the set:*

$$\mathcal{O}(X) = \{Ad(U) \cdot X \mid U \in U(n)\}$$

Two elements X, Y are on the same orbit if and only if there is a $U \in U(n)$ such that $Y = Ad(U) \cdot X$. Orbits are related to invariance. Given a subspace W of $\mathfrak{u}(n)$, and a subgroup $H \subseteq U(n)$, W is irreducible under the action of H if and only if there is only one orbit under the action of H in W. Orbits are a practical way to slice a space into invariant subspaces for the action of a group.

Proposition 23. *Two matrices X, Y in S_n^0 are on the same $O(n)$-orbit if and only if they have the same eigenvalues.*

Proof. Let $(X, Y) \in S_n^0$, they are real symmetric matrices and are therefore diagonalizable in an orthogonal basis. Henceforth there exists $P, Q \in O(n)$ and diagonal matrices D_1, D_2 such that $X = P^T D_1 P$, $Y = Q^T D_2 Q$. If X, Y are on the same orbit, there is $O \in O(n)$ such that

$$X = OYO^T \iff X = OQ^T D_2 O^T Q.$$

The diagonalization of a matrix is unique up to permutation, hence $\text{sp}(X) = \text{sp}(D_2) = \text{sp}(Y)$. Conversely, if $\text{sp}(D_1) = \text{sp}(D_2)$ then there is a permutation matrix $\mathcal{E}$ such that $D_2 = \mathcal{E}^T D_1 \mathcal{E}$. Then,

$$Y = Q^T \mathcal{E}^T D_1 \mathcal{E} Q = (Q(E)P)^T P D_1 P^T (P\mathcal{E}Q)$$

Calling $O = (Q\mathcal{E}P)^T$ we have

$$Y = OXO^T$$

The orthogonal group is stable by product and transposition therefore O belongs to $O(n)$ and X, Y belong to the same orbit in S_n^0 under the action of $O(n)$. □

Next proposition characterizes $O(n)-$invariance in the subgroup $\mathcal{A}_n^0$.

Proposition 24. *Two matrices in $\mathcal{A}_n^0$ are on the same orbit under the action of $O(n)$ if and only if they are on the same orbit under the action of $U(n)$.*

Proof. A skew-symmetric matrix is conjugated under the action of $O(n)$ to a block diagonal matrix **[Fegan, 1991]**. Let $A \in \mathcal{A}_n^0$, if n is even, there are $\lambda_{i1 \leqslant i \leqslant \frac{n}{2}}$ and the eigenvalues of A are $(\boldsymbol{i}\lambda_1, -\boldsymbol{i}\lambda_1, \boldsymbol{i}\lambda_2, -\boldsymbol{i}\lambda_2, \ldots, \boldsymbol{i}\lambda_{\frac{n}{2}}, -\boldsymbol{i}\lambda_{\frac{n}{2}})$. There is an orthogonal matrix O such that

$$A = O \begin{pmatrix} 0 & \lambda_1 & & & & \ldots & 0 \\ -\lambda_1 & 0 & & & & & \\ & & 0 & \lambda_2 & & & \\ \vdots & & -\lambda_2 & 0 & & & \vdots \\ & & & & \ddots & & \\ & & & & & 0 & \lambda_n \\ 0 & & & & \ldots & -\lambda_n & 0 \end{pmatrix} O^T$$

and if n is odd, then A has only $\frac{n-1}{2}$ pairs of non-cancelling eigenvalues and factorizes with the same block-diagonal shape with a 0 at the n-th position on the diagonal. Using permutations and reflections, it is always possible to change the position of each block and the sign of the λ_i, hence all the orbit of a matrix A under the action of $O(n)$ contains all the real skew-symmetric matrices with the same spectrum. Furthermore, the orbit of A under $O(n)$ is necessarily included in the intersection of the orbit of A under $U(n)$ with the set of real matrices, since $O(n) \subset U(n)$. The intersection of $\mathcal{O}_{U(n)}(A)$ with $M_n(\mathbb{R})$ contains exactly all the real matrices that share the spectrum of A, hence

$$\mathcal{O}_{O(n)}(A) = \mathcal{O}_{U(n)}(A) \cap M_n(\mathbb{R}) = \{A' \in M_n(\mathbb{R}), \mathrm{sp}(A') = \mathrm{sp}(A)\}.$$

□

Proposition 25. *A polynomial in $\Sigma_2 \in S_n^0$ is invariant if and only if it is a symmetric polynomial in the eigenvalues of Σ_1.*

Proof. This results directly from the equivalence proven in proposition 23. □

Proposition 26. *A polynomial in the eigenvalues of a skew-symmetric matrix $\Sigma_1 \in \mathcal{A}_n^0$ is $O(n)$-invariant.*

Proof. This results from the fact that two matrices on the same $O(n)$-orbit have the same eigenvalues, see proposition 25. □

These propositions provide us with a practical way to build all invariant quantities under the action of $O(n)$, hence all the polarization quantities it is possible to build from the coherency matrix. Several approaches are possible, we propose two basis for the symmetric polynomials:

the kth power-sum basis and the elementary symmetric basis related together by the Newton identities.

Definition 19. *The elementary symmetric polynomials in* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *are:*

$$e_0(\boldsymbol{X}) = 1 \qquad e_k(\boldsymbol{X}) = \sum_{1 \leqslant j_1 < j_2 < \ldots < j_k} X_{j_1} X_{j_2} \ldots X_{j_k}$$
$$e_1(\boldsymbol{X}) = \sum_i X_i \qquad e_n(\boldsymbol{X}) = X_1 \ldots X_n$$

There is one elementary polynomial of degree d *for* $d \leqslant n$ *and it is the sum of all possible products of* d *monomials* $X_{j_1}, \ldots, X_{j_d}$*. The elementary symmetric polynomials are the building block of symmetric polynomials in* $(X_1, \ldots, X_n)$ *in the sense that any symmetric polynomials can be seen as a polynomial in* $(e_1, \ldots, e_n)$*.*

Definition 20. *The power-sum symmetric polynomials in* $\boldsymbol{X} = X_1, \ldots, X_n$ *are:*

$$p_0(\boldsymbol{X}) = n \qquad p_k(\boldsymbol{X}) = \sum_i {X_i}^k$$
$$p_1(\boldsymbol{X}) = \sum_i X_i \qquad p_n(\boldsymbol{X}) = \sum_i {X_i}^n$$

Just like the elementary symmetric polynomials, the family of power sum polynomials contains one element for each degree $d \leqslant n$ *made of the sum of all* d*th power of the variables. It also generates the ring of all symmetric polynomials in* $(X_1, \ldots, X_n)$*.*

The characteristic polynomial can be related to the elementary symmetric polynomials by noting that for $(\lambda_1, \ldots, \lambda_n) = \mathrm{Sp}(A)$

$$\chi_A(X) = (X - \lambda_1) \ldots (X - \lambda_n)$$

developps into:

$$\chi_A(X) = X^n - (\lambda_1 + \ldots \lambda_n) X^{n-1} + \ldots + (-1)^{n-1} \sum_{i_1 \leqslant i_2 \leqslant i_{n-1}} (\lambda_{i_1} \ldots \lambda_{i_{n-1}}) X + (-1)^n \lambda_1 \ldots \lambda_n$$

$$\chi_A(\chi) = \sum_{i=0}^{n} (-1)^i e_i(\lambda_1, \ldots, \lambda_n) X^{n-i}$$

The ring of polynomial invariant in the eigenvalues of a matrix A is generated by either the elementary symmetric polynomials in the eigenvalues or the power-sum polynomial in the eigenvalues. The following algorithm provides a way to compute the elementary symmetric polynomials in the eigenvalues without computing explicitely the eigenvalues. Given a matrix

$M = (m_{ij})$ it gives expressions of the symmetric polynomials in the eigenvalues of M as functions of the (m_{ij}).

The Faddeev-Leverrier algorithm **[Wikipédia, 2019]** is a way to evaluate recursively the elementary symmetric polynomials in the eigenvalues of a matrix.

Definition 21. *The Faddeev-Leverrier algorithm applied to a matrix M proceeds as follows:*

$$\textit{Let } M_0 = M \textit{ and } \forall\ 1 \leqslant k \leqslant n : M_k = M\left(M_{k-1} - \frac{1}{k}\textit{tr}(M_{k-1})I\right)$$

Then the following identities ensue:

$$\forall 1 \leqslant k \leqslant n \quad e_k(\lambda_1, \ldots, \lambda_n) = (-1)^{k+1}\frac{1}{k}\textit{tr}(M_{k-1}) \tag{3.14}$$

Note that

$$\frac{1}{k+1}\textit{tr}(M_k) = \frac{1}{k+1}\textit{tr}(MM_{k-1}) - \frac{1}{k(k+1)}\textit{tr}(M)\textit{tr}(M_{k-1})$$

Applying the Faddeev-Leverrier algorithm to low-dimensional cases yields:
For $n = 2, \boldsymbol{\lambda} = (\lambda_1, \lambda_2)$

$$e_1(\boldsymbol{\lambda}) = \text{tr}(M) \qquad e_2(\boldsymbol{\lambda}) = \frac{1}{2}\left(\text{tr}(M)^2 - \text{tr}(M^2)\right)$$

For $n = 3, \boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$

$$e_1(\boldsymbol{\lambda}) = \text{tr}(M) \quad e_2(\boldsymbol{\lambda}) = \frac{1}{2}\left(\text{tr}(M)^2 - \text{tr}(M^2)\right) \quad e_3(\boldsymbol{\lambda}) = \frac{1}{3}\text{tr}(M^3) + \frac{1}{6}\text{tr}(M)^3 - \frac{1}{2}\text{tr}(M)\text{tr}(M^2)$$

Evaluating the power-sum polynomials in the eigenvalues $\boldsymbol{\lambda}$ of M is more direct and yields:

$$\text{For } 1 \leqslant k \leqslant n \ \ p_k(\boldsymbol{\lambda}) = \text{tr}(A^k)$$

In the matrix space $\mathcal{M}_n(\mathbb{C})$, each symmetric polynomial $e_i(\boldsymbol{\lambda})$ or $p_i(\boldsymbol{\lambda})$ potentially has a real and an imaginary part, yielding two elementary invariant quantities. We use to our advantage the decomposition of the space as a direct sum in equation (3.13) to compute only real quantities. Let M write as

$$M = \text{tr}(M)\text{Id} + \underbrace{\Sigma_1}_{\in \mathcal{A}_n^0} + \boldsymbol{i}\underbrace{\Sigma_2}_{S_n^0} \tag{3.15}$$

Then, for $1 \leqslant i \leqslant n$ and e_i, p_i the orthogonality of the decomposition (3.15) brings[3]:

$$e_i(M) = e_i(\text{tr}(M)\text{Id}) + e_i(\Sigma_1) + \boldsymbol{i}e_i(\Sigma_2)$$
$$p_i(M) = p_i(tr(M)\text{Id}) + p_i(\Sigma_1) + \boldsymbol{i}p_i\,(\Sigma_2)$$

But, due to the structure of $\Gamma_n, \mathcal{A}_n^0, S_n^0$, many of the terms defined above will cancel. This is summarized in following proposition:

Proposition 27. *Simplified invariants for* $n = 2, 3$

• *For* $n = 2$ *and* $M = tr(M)Id + \Sigma_1 + \Sigma_2$*, the only non-cancelling terms when computing the elementary symmetric polynomials and power-sum symmetric polynomials are: terms are:*

$$e_1(tr(M)Id) = tr(M) \qquad e_2(\Sigma_1) = \frac{1}{2}\|\Sigma_1\|^2 \qquad e_2(\Sigma_2) = -\frac{1}{2}\|\Sigma_2\|^2$$
$$p_k(tr(M)Id) = tr(M)^k \qquad p_2(\Sigma_1) = -\|\Sigma_1\|^2 \qquad p_2(\Sigma_2) = \|\Sigma_2\|^2$$

Observe that without suprise, the two basis for the symmetric polynomials yield the same quantities up to a constant. The only $O(n)$*-invariant quantities in dimension* 2 *are the trace and the norm of the components in the different* $O(n)$*-invariant subspaces.*

• *For* $n = 3$ *and the same decomposition of* M *the non-cancelling terms are:*

$$e_1(tr(M)Id) = 3tr(M) \qquad e_2(\Sigma_1) = \frac{1}{2}\|\Sigma_1\|^2 \qquad e_2(\Sigma_2) = -\frac{1}{2}\|\Sigma_2\|^2$$
$$e_3(\Sigma_1) = \frac{1}{3}tr({\Sigma_2}^3) \qquad e_3(\Sigma_2) = \frac{1}{3}tr({\Sigma_2}^3)$$
$$p_k(tr(M)Id) = 2tr(M)^k \qquad p_2(\Sigma_1) = -\|\Sigma_1\|^2 \qquad p_2(\Sigma_2) = \|\Sigma_2\|^2$$
$$p_3(\Sigma_1) = tr({\Sigma_1}^3) \qquad p_3(\Sigma_2) = \|tr({\Sigma_2}^3)\|$$

Clearly, one of the consequence of the decomposition over the direct sum (3.13) is that the elementary symmetric and the power-sum polynomials yield the same quantities (the remark can be extended to the dimension n *painlessly). This means that computations can not be simplified by preferring one basis over the other. In the* 2 *and* 3 *dimensional case, we find the same terms that are the norms of the projection over each subspaces identified in (3.13). But two additional quantities are present in the three-dimensonal case, they are in terms of the trace of a matrix to the cube.*

[3] To alleviate notations, we note from here $e_i(M)$ (resp. $p_i(M)$) for the elementary symmetric (resp. power sum symmetric) polynomial of degree i evaluated in the eigenvalues of the matrix M

Proof. The nullity of the trace $\text{tr}(\Sigma_1) = \text{tr}(\Sigma_2) = 0$ and the (skew)-symmetric properties of the matrix: $\Sigma_1^T = -\Sigma_1$ and $\Sigma_2^T = \Sigma_2^T$ together with the expression of the norm : $\|M\|^2 = \text{tr}(MM^T)$ bring the results above. □

Computation of polarization quantities for bivariate and signals

We propose here to evaluate the symmetric polynomials in the eigenvalues of the coherency matrix of a bivariate signal. They will be functions of the entries of the coherency matrix, and a bit of work yields expressions in terms of the angles and amplitudes parametrizing the signal. We fall back on predictable results, the invariance of the ellipticity angle χ under rotation of the coordinate frame for instance. The expression of the coherency matrix of a bivariate signal in the Pauli basis is consistent with its decomposition on the direct sum $\mathcal{M}_2(\mathbb{R})$. It writes:

$$C_x = \underbrace{S_0 \frac{i\sigma_0}{2}}_{\in \Gamma_n} + \underbrace{S_3 \frac{i\sigma_2}{2}}_{= \Sigma_1 \in \mathcal{A}_n^0} + \underbrace{S_2 \frac{i\sigma_1}{3} + S_1 \frac{i\sigma_3}{2}}_{= \Sigma_2 \in S_n^0}$$

The Stokes parameters (S_0, S_1, S_2, S_3) can be expressed in terms of amplitude and angles parametrizing the signal and its polarization by (see table 1.1):

$$S_0 = a^2 \quad \left| \quad \frac{S_1}{S_0} = \cos(2\chi)\cos(2\theta) \right. \quad \left| \quad \frac{S_2}{S_0} = \sin(2\theta)\cos(2\chi) \right. \quad \left| \quad \frac{S_3}{S_0} = \sin(2\chi) \right.$$

The Pauli-basis being an orthonormal basis, the invariants identified in proposition 27 have simple expressions in terms of the Stokes parameters:

$$\begin{array}{rcccl} \text{tr}(C_x) & = & S_0 & = & a \\ \|\Sigma_1^2\|^2 & = & S_3^2 & = & a^2 \sin(2\chi) \\ \|\Sigma_2^2\|^2 & = & S_1^2 + S_2^2 & = & a^2 \cos^2(2\chi)\cos^2(2\theta) + a^2 \cos^2(2\chi)\sin^2(2\theta) \\ & & & = & a^2 \cos^2(2\chi) \end{array}$$

Without surprise, parameters invariant to the action of the rotation group describe the power and the ellipticity of the signal. It follows that the only parameter describing a purely geometrical information for bivariate signals is χ : the ellipticity angle for the polarization ellipse.

Trivariate signals: Again, the decomposition of the space $M_3(\mathbb{C})$ as the direct sum (3.13) is consistent with the Gell-Man basis. For a trivariate coherency matrix C_x this writes:

$$C_x = \underbrace{\Gamma_0\lambda_0}_{\in \Gamma_3} + \underbrace{\Gamma_2\lambda_2 + \Gamma_5\lambda_5 + \Gamma_7\lambda_7}_{= \Sigma_1 \in S_n^0} + \underbrace{\Gamma_1\lambda_1 + \Gamma_3\lambda_3 + \Gamma_4\lambda_4 + \Gamma_6\lambda_6 + \Gamma_8\lambda_8}_{= \Sigma_2 \in \mathcal{A}_n^0}$$

Naming $(a, \alpha, \beta, \theta, \chi)$ all the parameters of the trivariate signals, the first invariants are:

$$\|\Gamma_0 \lambda_0\|^2 = a^2 \qquad \|\Sigma_1\|^2 = a^4(1 - \cos(4\chi)) \qquad \|\Sigma_1\|^3 = 0$$

Symbolic computations have been obtained using a symbolic library in python. Computing the invariants of Σ_2 proved more challenging. If symbolic evaluations are not available, numerical evaluations are underway. Again, for the invariants we were able to compute, only a and χ, the amplitude and the ellipticity appear in the expressions.

3.3.3 Conclusion

Identifying sets of transformations of a signal and their invariants sheds light on the physical or mathematical meaning of the parameters involved. We believe it could even lead to new parametrizations precisely fitted to a description of the signal in terms of how it is acted upon by a set of transforms. Trivariate signals and their polarization for instance, are decomposed linearly over the Pauli basis or factorized over elements in $U(3)$. We have only dealt here with an introduction to the subject, and believe that promising avenues exist. In particular, we are curious to determine the group of transformations that allows us to modify the geometry, i.e. the ellipticity of a signal, without affecting the other parameters. We would then expect to see as invariants of this group of transforms the orientation parameters of the signal, i.e. for trivariate signals the angles α, β, θ. Even the frequency of a signal can be interpreted as the parameter involved through a particular group of transform, the group of scalar matrix with a complex exponential kernel on the diagonal. Clearly, being interested in signals from the point of view of invariances opens many possibilities.

Conclusion

In all this part, the study of the polarization of bivariate and trivariate signals has been the support of a reflection on the representation of information in mathematical form in signal processing. We have seen that when two scalar quantities respecting certain linearity rules are measured, they can be either plunged in the body of complexes, or seen as the two components of a vector. We have shown that depending on the choice of embedding, the properties of the spaces in question allow to define new equivalent representations which make new quantities appear. For example, the polar factorization in the complexe field gives rise to an amplitude and a phase which must be interpreted. The original idea that we have tried to develop, without however fully exploiting its potential, is that the right approach to understand the quantities constructed from a signal via the operations available in the representation space is via the invariances. Indeed, it seems to us that it is simpler to understand and interpret transformation groups rather than scalar or vector quantities. Seeing these quantities as invariants of a certain transformation group on the other hand, allows us to understand both the way they are constructed within their space, and their geometric interpretation. For example, studying the effect of unitary transformations on the Stokes parameters allowed us to realize that these quantities were not homogeneous to the usual descriptors of the signal, i.e. two amplitudes. Thanks to this element, we were able to understand the operation by which the Stokes parameters are obtained in the quaternion algebra and no longer see it as an arbitrary operation or as a "copy" of the vector-matrix formalism. Finally, in the case of trivariate signals, we argue that the Stokes parameters are not necessarily the ideal descriptors of the signal geometry, considering they are adapted to a linear decomposition of the covariance matrix. We would prefer a set of descriptors consistent with a definition of a set of transformation groups, for example, transformations affecting only orientation, those affecting only geometry, and those affecting frequency. Independent quantities would then be defined as invariant to the action of all but one of these subsgroups. This study could not be completed due to lack of time, and we stopped at the easiest case, i.e. the quantities invariant by changes of orientation. We think that the interest of such an approach is that it also allows to approach the real complexity of the problem. For example, in the trivariate case, 9 Stokes parameters are needed to describe the signal entirely (without its frequency content) whereas in its polar form the same signal requires only 5 parameters: 4 angles and one amplitude. There is thus a redundancy of information contained in the Stokes parameters, but this one being clearly non-linear, it is not easy to capture.

Part II

optimization geometry

Introduction

In this part, we deal with a problem apparently independent of the one in the previous part. Indeed, we focus on an optimization problem for functions defined on varieties. However, as in the first part, our focus will be more on understanding the implications of a certain choice of representation for our optimization problem than on simulating it. It should therefore not be surprising that we propose a formal algorithm that has not yet been tested. The development and testing of the algorithm is a work in progress between the University of Melbourne and the lab INS2I in France. Since we are dealing with functions defined on differential varieties, we recall in a first chapter some notions on differential varieties, and the tools we will use. It is worth noting that several approaches are possible in differential geometry. We discuss a little about intrinsic or extrinsic coordinates for example. Beyond the coordinates, it seems to us that the two usual approaches found in the literature are the topological approach **[Lee, 2000]** and the metric approach **[Absil et al., 2009]**. Again, the question of the structure with which the space is studied is paramount. In the topological approach, a preponderant place is left to the coordinate patches which are the main tool to study the manifolds. We thus come back permanently to the study of Euclidean spaces. In the metric approach, on the contrary, the differentiable manifold acquires new tools (a metric tensor, an inner-product...), thanks to which it is rarely necessary to refer to a Euclidean space. This is what we will call the Riemannian framework. It is worth noting that on a differentiable manifold, it is always possible to define Riemannian tools, so the choice to work only with coordinate patches is not linked to a lack of hypothesis on the studied set. We have chosen a rather topological formulation of manifolds, for several reasons. First of all, this is how differentiable manifolds were presented to us during our studies, and this results in a personal affinity with the topological formulation. Secondly, it seemed to us that the Riemannian notations, although they allow experienced users to gain in speed of reasoning and conciseness, require an effort of acquisition which is only justified for a large-scale work or a particularly subtle work. Finally, the topological approach, with all its possible heaviness, is nevertheless the approach that best allows one to "understand what is going on" on a manifold whose "good properties" actually come from its relation with a Euclidean space or with its tangent spaces, which amounts to the same thing. We have to admit however, that should this work be continued, we would recommend to switch to the Riemannian framework. Differential geometry is not the only new topic dealt with in this part, we also introduce some results on two extremely classical optimization methods: the Newton method and the Gradient descent as well as some general remarks on optimization algorithms. By chance, we place ourselves in the easiest framework in optimization, that of locally convex functions around their critical points. This allows us to use well known convergence theorems and to extend them to differential manifolds easily. A very brief introduction is given on homotopy methods, about which we recommend the book from **[Allgower and Georg, 2012]** which was a real pleasure to read. Eventually, we took our distance from the ideas developped in

this book but is was extremely valuable in understanding what goal optimization geometry was after.

Litterature review

4

Contents

4.1 Manifolds and differentiation

4.1.1 Differential in normed vector space

One way to understand differential geometry is to see it as the extension of calculus from vector spaces to the broader class of differentiable spaces, the latter term needing a formal definition. Calculus is the study of continuous changes. It uses intensively the physically intuitive concept of rate of change, formalizing it mathematically into the notion of derivatives, then differentials. For functions of the real variable, differentials can be nicely constructed using geometrical objects such as tangents to curves and are unambiguous. For multivariate analysis, the definition is less easy as two fundamental elements of the definition of differentials have to be true at the same time: a differential is the best linear approximation of a function at a point, and when applied to a direction v, it measures an instantaneous rate of change in that direction. The gist of the second notion is preserved through the following definition.

Definition 22. *Gâteau derivative The directional derivative of a function $f : U \subset \mathbb{R}^n \to V \subset \mathbb{R}^m$ is defined by the following quantity (when the limit exists):*

$$df_x.v = \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t} \tag{4.1}$$

The term $df_x.v$ reads "The directional derivative of f at x in the direction v". The function f admits a directional derivative or a "Gâteau" derivative at x in the direction v if the previous limit exists and is finite. We insist on the fact that the existence of a directional derivative and the property of differentiability are two distinct concepts in multivariate calculus. The following definition highlights this difference.

Definition 23. *Fréchet differential A function on a vector space is differentiable at x if there exists a linear operator L_x that satisfies:*

$$f(x + h) = f(x) + L_x \cdot h + o(h) \tag{4.2}$$

The linear application L_x is called the differential of f at x and the notation $L_x \cdot h$ instead of $L(h)$ is used to emphasize the linearity in h. This definition is valid because L_x is unique, from now we write $Df(x)$ for "the differential of f at x". Note the difference with the previous definition, the directional derivative is a vector that gives the intensity and direction of the rate of change in a precise direction, while the differential is a linear application that best approximates f around x. We emphasize the importance of vocabulary: in multivariate calculus, the word "derivative" always refer to a directional derivative, while differentials is

the generalization of the univariate derivative. Only for functions defined on the real line $\mathbb{R}$ will we use derivative or differentials interchangably.

Remark 1.1.1

A topology, and more precisely a norm is necessary in all the developpements we just presented. Its necessity has been hidden in the notation $o(h)$. When writing $f(x)+o(h)$ with $f\colon E \to F$, the notation $o(h)$ refers to a function $o\colon E \to F$ that satisfies $\lim_{\|h\|_E \to 0} \frac{\|o(h)\|_F}{\|h\|_E} = 0$. The limit of the ratio does not depend on the norm chosen as long as they are equivalent (in finite dimension, all the norms are equivalent).

Example. A curve is an application γ from $\mathbb{R}$ to a normed vector space E. A linear application from $\mathbb{R}$ to E necessarily writes as $h \mapsto hv$ where v is a vector of E. Equation (4.2) in definition 23 can be divided by the real number h. The differentiability in a equals to the existence of a vector $v \in E$ such that:

$$\frac{\gamma(t+h)-\gamma(t)}{h} = v + o(1)$$

We observe that a curve is differentiable at x if and only if it admits a directional derivative at x. The equivalence between both definitions is not true in general. It is achieved here because in $\mathbb{R}$, only one "direction" h is available.

Let $\gamma\colon \theta \mapsto (\cos\theta, \sin\theta)$ be the parametrization of a circle of unit 1 centered on 0 in $\mathbb{R}^2$.

$$\begin{aligned}\frac{\gamma(t+h)-\gamma(t)}{h} &= \left(\frac{\cos(\theta+h)-\cos(\theta)}{h}, \frac{\sin(\theta+h)-\sin(\theta)}{h}\right)\\ &= (-\sin\theta, \cos\theta) + o(1)\end{aligned}$$

Relation between differentiability and directional derivative Both concepts are not strangers to each other as the example of curves showed us. Precisely, when a function f is differentiable, it admits directional derivatives and $df_x \cdot v = Df(x) \cdot v$. The latter notation is preferentially used in this document, and we try to maintain consistency as we believe the use of different and inconsistent notations is one of the aspect that makes multivariate analysis so technical. However, if a function admits directional derivatives in all directions, then it does not follow that it is differentiable. It yet needs to be shown that the directional derivatives fit together in a linear way. Simple examples can be derived from complex analysis where, for instance, the function $z \mapsto \overline{z}$ is not $\mathbb{C}$-differentiable while admitting directional derivatives. For real-valued functions, a particularly useful set of directional derivatives are the partial derivatives, which play an important role when expressing the differential $Df(x)$ into a matrix form.

Definition 24. *Given a function $f : E \to \mathbb{R}$ and a basis $(e_i)_i$ of E the partial derivative $\partial_i f$ is:*

$$\partial_i f(x) = Df(x) \cdot e_i$$

Partial derivatives are directional derivatives in a direction picked inside of a basis of the domain of f. Note that because f is real-valued, $\partial_i f(x)$ is a real number. By linearity of $Df(x)$, it follows that if $h \in E$ has coordinates $(h_i)_i$ in the basis $(e_i)_i$ then

$$Df(x).h = \sum_i h_i \partial_i f(x)$$

This motivates the definition of the gradient below.

Definition 25. *The gradient of $f : E \to \mathbb{R}$ at x is the vector $\nabla f(x) = (\partial_1 f(x), \dots, \partial_{\dim E} f(x))^T \in E$. If f is differentiable, the differential and the gradient are related by:*

$$\forall v \in E \quad Df(x) \cdot v = \langle \nabla f(x), v \rangle$$

Example. Consider the function $f = \|\cdot\|^2$ that sends a vector on its squared norm.

$$\begin{aligned} f : \mathbb{R}^2 &\to \mathbb{R} \\ (x, y)^T &\mapsto x^2 + y^2 \end{aligned}$$

Let (e_1, e_2) be the natural orthonormal basis, $e_1 = (1, 0)^T$ and $e_2 = (0, 1)^T$. Then,

$$\partial_1 f(x, y) = 2x \qquad \partial_2 f(x, y) = 2y$$

Hence,

$$\nabla f(z) = (2x, 2y)^T$$

and

$$Df(x, y) \cdot (x', y') = \left\langle (2x, 2y)^T, (x', y') \right\rangle = 2(xx' + yy').$$

In particular, the derivatives of the squared norm at 0 are 0 in every direction. The derivative of the norm at a point $(x, 0)$ in a direction (dx, dy) depends only on dx.

In the same spirit, one can describe the differential of a function $f : \mathbb{R}^n \to \mathbb{R}^m$ as a $m \times n$ matrix called the Jacobian matrix. Write

$$\forall x \in \mathbb{R}^n, f(x) = (f_1(x), f_2(x), \dots, f_m(x))$$

Then each f_j admits partial derivatives with respect to a basis $(e_1, \dots, e_n)$ of $\mathbb{R}^n$.

Definition 26. *The Jacobian of a differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ satisfying $f = (f_1, \dots, f_m)$ at x is the matrix*

$$J_f(x) = \begin{pmatrix} \partial_1 f_1(x) & \partial_2 f_1(x) & \dots & \partial_n f_1(x) \\ \partial_1 f_2(x) & \partial_2 f_2(x) & \dots & \partial_n f_2(x) \\ \vdots & & & \\ \partial_1 f_m(x) & \partial_2 f_m(x) & \dots & \partial_n f_m(x) \end{pmatrix}$$

and, for a vector $v \in \mathbb{R}^n$ the derivative of f at x in the direction v satsifies:

$$Df(x) \cdot v = J_f(x)v$$

Example. Consider the function f that sends a complex number $z = x + \boldsymbol{i}y$ on z^2. We see f as a function of $\mathbb{R}^2$.

$$\begin{aligned} f\colon \mathbb{R}^2 &\to \mathbb{R}^2 \\ (x, y) &\mapsto (x^2 - y^2, 2xy) \end{aligned}$$

Then $f_1(x, y) = x^2 - y^2$ and $f_2(x, y) = 2xy$ and

$$J_f(x, y) = \begin{pmatrix} 2x & -2y \\ 2y & 2x \end{pmatrix}$$

Finally, if $f\colon \mathbb{R}^n \to \mathbb{R}$ is real-valued, and twice differentiable, then its gradient is vector-valued and one can define the Jacobian of the gradient of f. We call this object the Hessian of f. It is a squared matrix defined by:

$$\operatorname{Hess} f(x) = \begin{pmatrix} \partial_1\partial_1 f(x) & \dots & \partial_n\partial_1 f(x) \\ \partial_1\partial_2 f(x) & & \partial_n\partial_2 f(x) \\ \vdots & & \vdots \\ \partial_1\partial_n f(x) & \dots & \partial_n\partial_n f(x) \end{pmatrix}$$

What precedes has shown how to define point-wise differentiability in a normed vector-space, and how to compute derivatives. Linearity of the domain and target space have been instrumental to defining differentials and derivatives, which is why extending calculus to non-linear space might seem at first problematic. We carefully build the notion of non-linear differentiable spaces in the next section, and provide definition for differentials of functions on such spaces as well as a practical way to compute them.

4.1.2 Differentiable spaces

The definition of a differentiable space passes through the definition of differentiable maps on this space. It is a common fact in mathematics that maps between sets play an important

role in the definition and classification of the sets themselves. Topological spaces can be compared through homeomorphisms, linear space through isomorphisms and differential spaces through diffeomorphisms for instance.

Example. Let S^2 be the two-dimensional sphere, notably not a linear space, and let f be a map on S^2. A directional derivative of f on S^2 would be the measure of the rate of change of f in a given direction on S^2. Furthermore, the differential of f would be the linear application that fits together all the directional derivatives in a linear way. One can immediatly understand that a first complication will be to define a linear set of directions on a non-linear set. For this reason, not every spaces are differentiable.

There are two ways to propose a solution to the problem we just raised. Basically, one can go with an intrinsic view of the sphere, such as there is no "outside" of the sphere, and the definition of every object, directions included, must be contained within the sphere. Or we can go with an "extrinsic" view of the sphere, where we imagine it sitting in a 3-dimensional space. Directions on the sphere at a point x become vectors of $\mathbb{R}^3$ tangent to the sphere at x. The latter approach is less satisfying from a theoretical point of view, but enables us to build intuition more easily about differentiation in non-linear spaces. In particular, it enables us to build the differential of curves on a non-differentiable space. We illustrate this with a curve on the sphere.

Derivative of a curve in extrinsic coordinates

Let $\gamma\colon \mathbb{R} \to S^2$ be a curve on the sphere, where the sphere is described as the set of points in $\mathbb{R}^3$ at distance 1 of the origin: this is an extrinsic description of the sphere. Then γ writes:

$$\gamma(t) = \begin{pmatrix} \gamma_1(t) & \gamma_2(t) & \gamma_3(t) \end{pmatrix}^T \quad \text{with} \quad \gamma_1^2(t) + \gamma_2^2(t) + \gamma_3^2(t) = 1$$

The differential of γ is

$$D\gamma(t) = \begin{pmatrix} \gamma_1'(t) & \gamma_2'(t) & \gamma_3'(t) \end{pmatrix}^T$$

and differentiating the second relation in () yields

$$\big\langle \gamma(t), \gamma'(t) \big\rangle = 0$$

Once derivatives of curves are defined, it is possible to use them to compute the differential of another object: scalar-valued functions on a manifold. This leads to the definition of the object with a central role in differentiation in non-linear spaces: the tangent space.

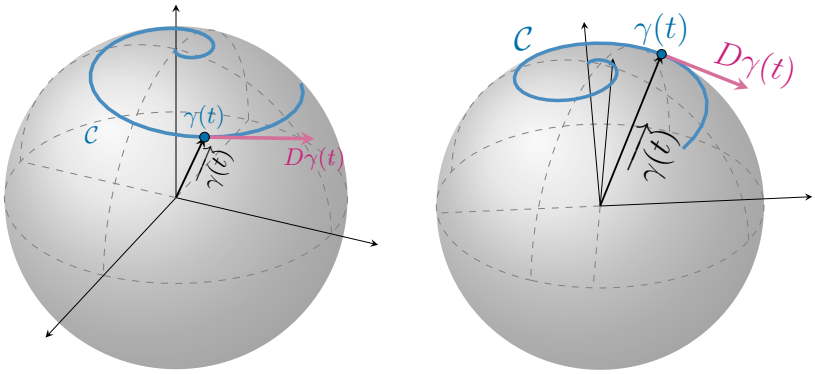


Fig. 4.1. – On the left, a curve $\mathcal{C} = \{\gamma(t),\ t \in I \subseteq \mathbb{R}\}$ is traced in blue on the 2-dimensional sphere S^2 embedded in $\mathbb{R}^3$. An orthonormal frame of $\mathbb{R}^3$ is shown. The point $\gamma(t)$ is represented in blue, and the vector $\vec{\gamma}(t)$ is shown as a black arrow coming from the origin of the frame. The derivative $D\gamma(t)$ is in magenta. It is clearly tangent to the curve $\mathcal{C}$, and the figure on the right highlights its orthogonality to $\vec{\gamma}(t)$.

Tangent space

Let $f\colon S^2 \to \mathbb{R}$ be a continuous map on the sphere. The 2-sphere has no linear structure, it is thus impossible to write $f(u+tv)$ for u, v points on the sphere. Instead, we consider a curve $\gamma\colon]-\varepsilon, \varepsilon[= U \to S^2$ which is differentiable and satisfies $\gamma(0) = u$. We have shown that the differentiability of γ is well defined through classical calculus and

$$D\gamma(0) = \lim_{t\to 0} \frac{\gamma(t) - \gamma(0)}{t} = v \in \mathbb{R}^3$$

is a vector tangent to the sphere S^2 at $\gamma(0)$ (see figure 4.1). The composition $f \circ \gamma$ is a function from an interval of $U \subset \mathbb{R}$ to $\mathbb{R}$. If $f \circ \gamma$ is differentiable in 0 as a function of $\mathbb{R}$ to $\mathbb{R}$, then we can define

$$Df(u).v := D(f \circ \gamma)(0) \tag{4.3}$$

This derivative is well-defined only if we can prove that the quantity $D(f \circ \gamma)(0)$ does not depend on the choice of the curve γ satisfying $\gamma(0) = u$ and $\gamma'(0) = v$. The spaces where the derivative does not depend on the curve chosen are differentiable, we call them manifolds or differentiable manifolds in the sequel. In conclusion, a direction in a manifold can be defined through a curve in that manifold and more precisely through the equivalence class of all the smooth curves having the same derivative at one point. Measuring the rate of change of the function f on a curve $[\gamma]$ is equivalent to measuring the rate of change in the direction given by $[\gamma'(0)]$ (square brackets designate an equivalence class). Now, taking all the possible smooth curves through a point x in the sphere will give many possible directions for the derivatives. The set of all these directions have a vector space structure **[Lee, 2013]** that is the same at each x on S^2. We name this vector space of possible derivatives for a

curve through x the tangent space at x of S^2, usually noted T_xS^2 (see figure 4.2). It is a 2-dimensional vector-space and for $x, y \in S^2$, $T_xS^2 \simeq T_yS^2$.

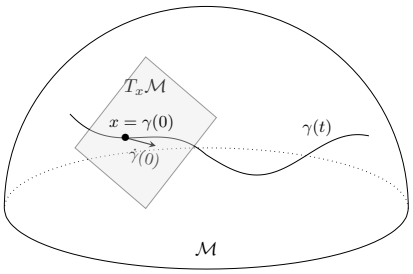


Fig. 4.2. – The curve $\gamma \in \mathbb{R}^3$ lies on the sphere so that a vector tangent to γ is tangent to the sphere. The set of tangent vectors at x is the set of vectors tangent to any possible smooth curve γ and is thus the tangent space at x of S^2

We now have a new criteria for the non-differentiability of spaces. Imagine a cross drawn in the plane. Clearly, on the arms of the cross, the tangent space has dimension 1, it is the set of vectors tangent to the line. On the crossing point however, two different directions intersect: the set of tangent vectors is not a vector space, it contains two orthogonal vectors but we can not build a curve on the cross tangent to an axis cutting the cross without being parallel to one of the arms. When the tangent spaces do not have a linear structure or when their dimensions are not constant on the space, the space is not differentiable.

The main points outlied in the previous paragraph are:

1. the differential of a function is a linear map that can be computed by fitting together directional derivatives.

2. To extend calculus to non-linear spaces (such as the sphere), it is necessary to define a vector space of directions at every point.

The second point is achieved through smooth curves passing at that point. The curve itself gives a direction in the manifold, the derivatives of all such curves has a linear structure. The definition of directional derivative in that case can be obtained using linear calculus by composition of functions. Because the set of all possible curves gives a tangent vector space of direction, it is then possible to define differentiability through the existence of a linear approximation of the function f at x, whose domain is the tangent space at x. It was possible to use vector-space calculus to define calculus on a manifold because this manifold was embedded in a Euclidian space (the sphere was embedded in $\mathbb{R}^3$). This approach is not as restrictive as it seems to study arbitrary manifolds, because the Withney embedding (or Whitney immersion) theorem **[Lee, 2013]** states that any Hausdorff and secound-countable[1] manifold can be embedded in $\mathbb{R}^n$ for a large enough n. However, we will see in the next paragraph that it is not necessary to use an embedding to build differentiation on a manifold.

[1] Hausdorff and second-countable are two properties that refer to the topology of the set. A set is Hausdorff if any two distinct points can be separated, i.e, can be in two distinct open sets, while the second-countable conditions states that the set has a countable basis of open subsets

It is certainly easier to understand manifolds through their embeddings, but it is sometimes necessary to have a standalone definition given in next section.

4.1.3 Finite-dimensional differentiable manifolds

Definition 27 (finite-dimensional differentiable manifold). *A finite n-dimensional manifold is a second-countable Hausdorff topological space M together with a countable atlas[2] $(U_i, \varphi_i)_{i\in I}$ such that: for all $i \in I$, the map φ_i is a homeomorphism[3] from U_i to its image in $\mathbb{R}^n$. The manifold M is said to be differentiable if the φ_i overlap in a differentiable way. That is, for $U_i \cap U_j \neq \varnothing$, the transition map defined by:*

$$\varphi_j \circ \varphi_j^{-1} : \varphi_i(U_i \cap U_j) \to \mathbb{R}^n$$

is differentiable.

section 4.1.3 illustrates the notion of coordinate charts.

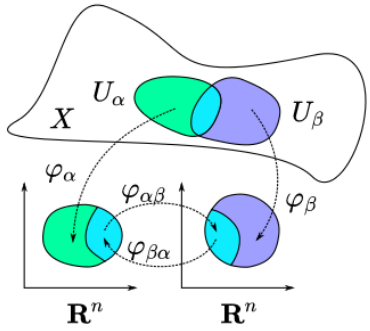


Fig. 4.3. – On the manifold X two coordinate patches are represented in colors. Below they are mapped to the euclidian space $\mathbb{R}^n$. The coordinate patches overlapp, which is shown by a blue region. The diffeomorphism realized by the transition maps is represented by arrows. (Source: Wikipedia)

In general, any adjective as C^k, *differentiable* or *smooth* that can be combined with manifold refers to the transition maps of that manifold. Hence, a manifold is C^k if all its transition maps are C^k. Because transition maps are applications on open-sets of $\mathbb{R}^n$, calculus on manifolds is founded on vector-space calculus. All manifolds in the rest of the document are supposed smooth.

[2] An atlas is a basis of open sets. The idea is to cover your set with open sets as you would if you had to cover a surface with a collection of various towels. Every point in the set belongs to at least one of the open sets in the Atlas, it can belong to more than one (the towels can overlapp, as long as they cover all the surface. The topological condition on the set ensures us that it is possible to build such a basis using only a countable infinite number of open sets.)

[3] a homeomorphism is a continuous invertible application whose inverse is also continuous. If two spaces are homeomorphic, they are topologically equivalent.

Definition 28. *Given $\mathcal{M}$ a n-dimensional manifold, we call a coordinate patch on $\mathcal{M}$ the data of an open-set $U \subset \mathcal{M}$ and a map φ such that*

$$\varphi \colon U \to \varphi(U) \subset \mathbb{R}^n$$

is a homeomorphism. The map φ is a coordinate chart on U.

Remark 1.3.2

The definition of a n-dimensional manifold might look obscure to a novice to differential geometry. We give here a step by step explanation of it. First, not every sets are eligible to become manifolds, they need the right topology (Hausdorff and second-countable) which is not something we worry about in the present document because it is true in any normed space. Then, once a set with the right topology has been identified, we must check that it is locally similar enough to a Euclidian space. A property is local in mathematics if it is verified on neighborhoods, here the open sets of the atlas. The concept of similarity is achieved through homeomorphisms, the tool that is used to compare topologies. If two topological sets have an homeomorphism between them, they are topologically equivalent. The expression "locally similar to Euclidian space" translates in: around every point in the manifold, there is an open subset that can be mapped through an homeomorphism to a Euclidian space. The regularity of the maps can not be directly specified as we have not yet defined regularity on a manifold, which is why we use the transition maps to specify a structure on the coordinate charts. Because the transition maps are functions on vector space, classical calculus can be applied to them.

Remark 1.3.3

Typically, the number of charts necessary to cover a n-dimensional manifold in subset homeomorphic to $\mathbb{R}^n$ is not big. The 2-dimensional sphere for instance requires only two charts. You can imagine the charts as the number of parts you need to cut out so that each part can be flattened out. There is actually a general result stating that a n-dimensional manifold can be covered with $n + 1$ charts. **[Kohan, 2014]**.

4.1.4 Tangent vectors and Differentiation in a coordinate chart

We admit a way to build the tangent space on a general manifold $\mathcal{M}$ using the transition maps. The tangent space at a point $p \in \mathcal{M}$ is denoted $T_p\mathcal{M}$, it is a linearization of the manifold at this point. Each vector in the tangent space can be understood as a direction on the manifold. The dimensions of the tangent spaces are constant and equal to the dimension of the manifold. All these results can be found in **[Lee, 2000]**. Just like in Euclidian geometry, the differential of a function is the best linear approximation of this function at a given point. In the case of a manifold, the differential is the best linear approximation of the function in the space of the best linear approximation of the manifold, a double linearization is necessary to define the differential. Let $f\colon M \to \mathbb{R}^n$ a function from a manifold to a euclidian space. The differential of f at x is an application from the tangent space $T_x\mathcal{M}$ to $\mathbb{R}^n$ and we note:

$$\begin{aligned} Df(x)\colon T_xM &\to \mathbb{R}^n \\ v &\mapsto Df(x).v \end{aligned}$$

The vector v belongs to the tangent space at x, but can be understood as a direction in the manifold itself, through a curve tangent to that direction. The former presentation of tangent vectors through curves is still valid, and we will relate it to the coordinate charts. Let $x \in \mathcal{M}$ and $\varphi\colon U_x \to \mathbb{R}^n$ a coordinate chart defined on $U_x \ni x$. Let $\mathcal{B}$ be a basis in $\mathbb{R}^n$ Then, there are n functions $\varphi_1 \dots, \varphi_n$ such that

$$\forall x \in U_x,\ \varphi(x) = (\varphi_1(x), \dots, \varphi_n(x))_{\mathcal{B}} \tag{4.4}$$

Definition 29. *The numbers $(\varphi_1(x), \dots, \varphi_n(x))$ defined in (4.4) are the intrinsic coordinates of the point $x \in \mathcal{M}$. Intrinsic coordinates are not unique, they depend on the choice of a coordinate chart at the point x and on the basis of $\mathbb{R}^n$.*

An elementary direction in $\mathcal{M}$ with respect to the basis $\mathcal{B}$ is a direction that can be covered by points with only one non-zero coordinate. For instance, a curve tangent to the direction $e_i \in \mathcal{B}$ at x writes $c\varphi^{-1}t \mapsto (0, \dots, 0, \underbrace{\langle {\varphi_i}^{-1}(t), e_i \rangle}_{ithposition}, 0, \dots, 0)$ for a constant c. A vector tangent to this curve in the tangent space is colinear to e_i. Using intrinsic coordinates, we can construct a basis for the tangent spaces. This is a very important construction as it enables us to give vector or matrix form to differentials of functions on manifolds see proposition 28. This is used extensively in 5.

Proposition 28. *Let $x \in \mathcal{M}$ and $(\varphi_1, \dots, \varphi_n)$ a system of intrinsic coordinates around x. Then, the differential $\partial\varphi_i(x)$ is a tangent vector of $\mathcal{M}$ at x. The set of derivatives of the intrinsic coordinates forms a basis on the tangent space $T_x\mathcal{M}$ noted $(\partial\varphi_1(x), \dots, \partial\varphi_n(x))$ and called the coordinate-induced basis or coordinate basis **[Wikipedia contributors, 2018]** on*

$T_x\mathcal{M}$. Coordinates provide a way to compute derivatives and express explicitely tangent vectors. The derivative of f at x can be expressed in terms of the derivative of a function on a Euclidian space, see:

$$Df(x) = \left[D(f \circ \varphi^{-1})(\varphi(x))\right]_{(\partial\varphi_1(x),\dots,\partial\varphi_n(x))} \tag{4.5}$$

where $f \circ \varphi^{-1}$ is a function from $\mathbb{R}^n$ to $\mathbb{R}$ such that its derivative is the usual derivative in $\mathbb{R}^n$. To apply the derivative at x in a direction v, note that the coordinates of v in the right-hand term of the equation must be taken in the basis induced by φ on the tangent space. It is this relation that gives a practical way to compute derivatives of functions on manifolds and present them under a vector or matrix form. It is necessary to understand that parametrization is arbitrary on a manifold, but once it is defined, it ricochets on the tangent spaces. Keep in mind that if a tangent vector is explicitely given as a vector with coordinates, these coordinates refer to the choice of a parametrization.

Second-order derivation

Once first-order derivation for functions $f\colon \mathcal{M} \to \mathbb{R}^m$ have been defined, second-order derivation follows smoothly. The application Df sends $x \in \mathcal{M}$ to $Df(x) \in \mathcal{L}(T_x\mathcal{M}, \mathbb{R}^n)$. Because T_x is isomorphic to $\mathbb{R}^n$ we can reinterpret the last space as $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. Hence, Df is a function from a manifold to a euclidian space and falls into the category of functions for which we have defined the derivative. We note D^2f the second-derivative of f. It sends a point x in the manifold $\mathcal{M}$ to a linear application in $\mathcal{L}(\mathbb{R}^{\backslash}, \mathbb{R}^{\updownarrow}), \mathcal{L}(\mathbb{R}^{\backslash}, \mathbb{R}^{\updownarrow})$.

Gradient and Hessian

In proposition 28, the derivative of $f\colon \mathcal{M} \to \mathbb{R}^m$ has been directly related to the derivative of a function $\tilde{f}\colon \mathbb{R}^n \to \mathbb{R}^m$. We can therefore define a gradient and a Hessian of f through the Gradient and the Hessian of $\tilde{f}$. Remember that given a coordinate patch (U, φ) on $\mathcal{M}$ the function $\tilde{f}$ is defined on $\varphi(U) \subseteq \mathbb{R}^n$ and on $\varphi(U)$ the functions $\tilde{f}$ and f are related by:

$$\tilde{f} = f \circ \varphi^{-1}$$

Then, for $x \in U$, we call the Jacobian of f at x the matrix

$$J_f(x) = J_{\tilde{f}}(\varphi(x))$$

The derivative of f and its Jacobian are related by:

$$Df(x) \cdot v_x = J_f(x) \cdot (v_x)_{(\partial\varphi_1(x),\dots,\partial\varphi_n(x))}$$

where $(v_x)_{\partial\varphi_1(x),\dots,\partial\varphi_n(x)}$ are the coordinates of the tangent vector v_x in the basis $(\partial\varphi_1(x),\dots,\partial\varphi_n(x))$. If f is a $\mathbb{R}$-valued function on $\mathcal{M}$ then its gradient and its Hessian are defined similarly, as the transposed of its Jacobian i.e. for $f\colon \mathcal{M} \to \mathbb{R}$

$$\nabla f(x) = \nabla \tilde{f}(\varphi(x)) \qquad\qquad \operatorname{Hess} f(x) = \operatorname{Hess} \tilde{f}(\varphi(x))$$

The Jacobian, gradient and the Hessian of a matrix are always defined through a coordinate patch. This can be problematic. It means the notation can only be used if all the variables considered are contained in the same coordinate patch. We can partially contourn this difficulty by noticing that when writing

$$\nabla f(x) \cdot v$$

the fact that the coordinates of v must be valid in the tangent space $T_x\mathcal{M}$ is made unambiguous by the dependency in x in the differential. Hence, one can always write $\nabla f(x) \cdot v$ and let the reader make the mental effort to remember there must be a coordinate patch around x such that the coordinates of v are related to the coordinate chart chosen. The confusion can arise because when we write

$$\nabla f(x) \cdot v \qquad\qquad \nabla f(y) \cdot v$$

we use the same vector v (they both have the same coordinates in their respective coordinate-induced tangent basis). However, if we could draw $v \in T_x\mathcal{M}$ and $v \in T_y\mathcal{M}$, they would not be parallel vectors in the embedding space!

Notations

We introduce here some notations we will use in the rest of the document and in particular in chapter 5. Partial differentiation will be noted

$$D_x f(x', y') = D[x \mapsto f(x, y')](x')$$

The index x refers to the variable of differentiation and not to the the variable where the differential is evaluated. Similarly, $D_{x_1} f(x)$ would be the short way to write the differential of f with respect to the first variable of a given coordinate chart where $\varphi(x) = (x_1, \dots, x_n)$. To alleviate possible confusions, we remind that "partial derivatives" are defined for functions from a manifold $\mathcal{M}$ to $\mathbb{R}$ by $\partial_i f(x) = Df(x).(1, 0, :, 0)$, they are a completely different objects than partial differentiation. Vocabulary and notations are indeed one of the most tedious aspect of differential geometry. Interestingly, the addition of a new concepts, instead of complexifying the whole framework has a tendency to simplify it. Notations in Riemannian

manifolds [Absil et al., 2009] look usually more agreeable than notations in an abstract topological manifold for instance.

4.1.5 Geodesics and exponential mapping

Geodesics are the curves on a manifold that generalize the notion of straight lines, and the exponential mapping provides a way to parametrize geodesics. Given two points p, q on a manifold, a geodesic between p and q is a curve γ that passes through p and q and such that the length of γ between p and q is the minimal length that can be obtained for a curve joining these two points. The notion of the length of a curve between p and q is defined as follows

$$L(\gamma) = \int_{\gamma^{-1}(p)}^{\gamma^{-1}(q)} \|D\gamma(t)\| dt$$

where the norm is defined in the tangent space $T_{\gamma(t)}\mathcal{M}$. We admit that we know how to compute norms in tangent spaces and that the norm evolves smoothly with tangent spaces. This is a cheap hypothesis given all manifolds can be embedded in a euclidian space, and the euclidian norm can be used on the tangent spaces. On the sphere for instance, geodesics between two points are circles joining these two points. The exponential mapping can be defined using a property of the geodesic: given $p \in \mathcal{M}$ and v a tangent vector in $T_p\mathcal{M}$, there exists a unique geodesic γ_v such that $\gamma_v(0) = p$ and $D\gamma_v(0) = v$. Then, we define

$$\exp_p : v \mapsto \gamma_v(1)$$

From the definition of γ we can see that $e_p^{tv} = \gamma_v(t)$. The parameter t in the exponential is a way to move along the geodesic. The exponential is locally a diffeomorphism mapping between the tangent space at p and the manifold around p [Absil et al., 2009]. Geodesics also provide an interpretation of tangent vectors represented by the same coordinates in two different tangent spaces. When we write

$$v \in T_x\mathcal{M} \text{ and } v \in T_y\mathcal{M}$$

we are using the same notations for two different vectors that have the same coordinates in their respective basis. However, one can define the tangent basis in $T_x\mathcal{M}$ and in $T_y\mathcal{M}$ in a consistent way. That is, for a tangent vector of coordinates $v \in T_x\mathcal{M}$ and for all points p on the geodesic γ passing through x and tangent to $v \in T_x\mathcal{M}$ at x, let the vector $v \in T_p\mathcal{M}$ denotes the tangent vector that is tangent to γ at p.

Relation between the exponential mapping and the coordinate charts

We have already showned that the data of a coordinate chart can be related to a coordinate-induced basis by the mean of derivating each coordinate of the coordinate chart. Conversely, the exponential mapping gives us a way to define a coordinate chart given a basis of tangent

vectors in the tangent space. We use the result (that we do not demonstrate) that the exponential mapping defines locally a smooth diffeomorphism on $\mathcal{M}$.

Proposition 29. *Given $(v_1, \dots, v_n)$ a basis of tangent vectors in $T_x\mathcal{M}$, there is an open space $U_x \in \mathcal{M}$ around x such that*

$$\varphi \colon \mathcal{M} \to \mathbb{R}^n \tag{4.6}$$

$$p \mapsto \left(\exp_x^{-1}(p)\right)_{(v_1,\dots,v_n)} \tag{4.7}$$

where $\left(\exp_x^{-1}(p)\right)_{(v_1,\dots,v_n)}$ are the coordinates of $\exp_x^{-1}(p)$ in the basis $(v_1, \dots, v_n)$. Furthermore, if $(v_1, \dots, v_n)$ is the coordinate-basis induced by a coordinate chart ψ, then the coordinate chart built by (4.7) satisfies

$$\exp_p^{-1}(x) = \psi(x) - \psi(p)$$

Proof. The first part of the proposition is a direct consequence of $\exp$ being locally a smooth diffeomorphism. To prove the second point, write for $(x, p) \in U$ the geodesic $\gamma(t) = \exp_p(tw)$ between x and p has a constant derivative w. Note $b = \exp_{p,w}^{-1}(x)$

$$\begin{aligned} \exp_{p,w} \colon [0, b] &\to \gamma \\ t &\mapsto \exp_p(tw) \end{aligned}$$

Therefore

$$\begin{aligned} \psi(x) - \psi(p) &= \int_0^b D(\psi \circ \exp_{p,w}(t) dt \\ &= \int_0^b w dt \\ \psi(x) - \psi(p) &= w \exp_{p,w}^{-1}(x) = \exp_p^{-1}(x) \end{aligned}$$

□

This one-to-one correspondance between the set of tangent basis for a tangent space at a point x and the set of coordinate charts around this point is used in section 4.2 and part II.

4.1.6 Inverse function theorem and its consequences

The differential being the best linear local approximation of a function, it locally preserves some very strong properties such as invertibility. In this section, we state this result and show how it evolves into building parametrized manifolds from smooth maps, in particular using level sets of smooth maps.

Theorem 3 (Inverse function theorem). *Let f be an application of class $C^k (k \geqslant 1)$ from a manifold $\mathcal{M}$ to $\mathcal{N}$ and $a \in \mathcal{M}$ a point such that $Df(a)$ is invertible. Then there exists an open set U containing a such that $f : U \to f(U)$ is a C^k diffeomorphism.*

In other words, if the differential of f is an isomorphism of vector space around a point a, then f is also an isomorphism of differentiable space around a. The theorem can be readapted to the case where $Df(a)$ is injective and not bijective. Note that the conditions of the theorem can only be satisfied if $T_a\mathcal{M}$ and $T_{f(a)}\mathcal{N}$ have the same dimension, hence if $\mathcal{M}$ and $\mathcal{N}$ are manifolds with the same dimension. However, strong results still hold when the differential is an injective application.

The next theorem characterizes the level sets of some particular points for the function f which are called the regular values of f. In this theorem it becomes clear that the differential at one point contains local information about the function.

Theorem 4 (Regular value theorem). *Let $f : M \to N$ be a smooth map between manifolds, and let $p \in M$ be such that $Df(p)$ is surjective at each point in $f^{-1}(p)$. Then p is called a regular value of f and $f^{-1}(p)$ is a smooth submanifold in M with dimension $\dim M - \dim N$. Further, the tangent space at x is the kernel of $Df(x)$: $T_x f^{-1}(p) = \ker(Df(x))$ at each point $x \in f^{-1}(p)$.*

By this theorem, the level sets of regular values by smooth maps are smooth submanifolds embedded in the domain space. The next theorem is a local reinterpretation of this theorem.

Theorem 5 (Implicit function theorem). *Let m, n be positive integers, let $T = U \times V$ be an open subset of $\mathbb{R}^n \times \mathbb{R}^m$ and $f : T \to \mathbb{R}^m$ be a continuously differentiable function on T. We name D_1 the differential with respect to the n coordinates in U, and D_2 the differential along the m coordinates in V [Lang, 2012]. Let $(x_0, y_0) \in T$ be such that $f(x_0, y_0) = 0$.*
If $D_2 f(x_0, y_0)$ is invertible, then there are open sets $U' \subseteq U, V' \subseteq V$ such that $x_0 \in U', y_0 \in V'$ and there is a function $g : U' \to V'$ differentiable at x_0 such that $(x, g(x)) \in T$ and

$$\forall x \in U' \quad f(x, g(x)) = 0 \tag{4.8}$$

Moreover, the differential of g is not implicit and equals:

$$Dg(x_0) = -\left[D_2 f(x_0, y_0)\right]^{-1} D_1 f(x_0, y_0) \tag{4.9}$$

This theorem can only be applied locally inside of a coordinate patch in a manifold, as it requires open subset of a Euclidian space. The implicit function of the implicit function theorem is naturally the function g. This theorem gives condition under which the solution to a particular equation is a smooth function in the parameters of the equation.

4.1.7 Conclusion

In this section we have introduced tools from differential geometry in Euclidian spaces and manifolds that will be used in chapter 5. We emphasized the difference between extrinsic and intrinsic coordinates, derivatives and differentials and showed that both partial derivatives and intrinsic coordinates play a major role in the practical computation of differentials in multivariate calculus. We have given a definition of straight lines on manifolds and proposed a way to parametrize them through the exponential mapping. This will be necessary to generalize optimization algorithms to manifolds in section 4.2. In a second part, we have recalled the inverse function theorem and some of its corrolaries. They show conditions under which it is possible to define differentiable implicit functions. Again, this will be an instrumental tool in chapter 5 and section 4.3.

4.2 Optimization algorithms

In this section, we want to understand a few results on the convergence of the Newton and gradient methods, both in terms of "where" should the method start, and "how fast" can it converge towards a critical point of a function f. We consider a function f from a normed vector space $(E, \|\cdot\|)$ to $\mathbb{R}$, and if not stated otherwise, we assume the goal of our algorithm is to find a solution to the equation $f(x) = 0$. When dealing with $\mathcal{L}(E)$ the set of endomorphisms in E, we use the same notation for the matrix norm induced by the vector norm, and the vector norm itself.

4.2.1 Presentation

In all that follows, there is a dependency in the function f that we keep implicit.

Definition 30. *The Gradient map $\mathcal{G}$ of parameter L and the Newton map $\mathcal{N}$ are respectively*

$$\begin{array}{rcl|rcl} \mathcal{G} & : E & \to E & \mathcal{N} & : E & \to E \\ & x & \mapsto x - \frac{1}{L} f(x) & & x & \mapsto x - [Df(x)]^{-1} f(x) \end{array}$$

For $x \in E$, the Gradient sequence (resp. the Newton sequence) are, when they are defined:

$$x_0 = x \quad \forall k > 0 \;\; x_{k+1} = \mathcal{G}^k(x_k) \text{(resp. } \mathcal{N}(x_k)\text{)}.$$

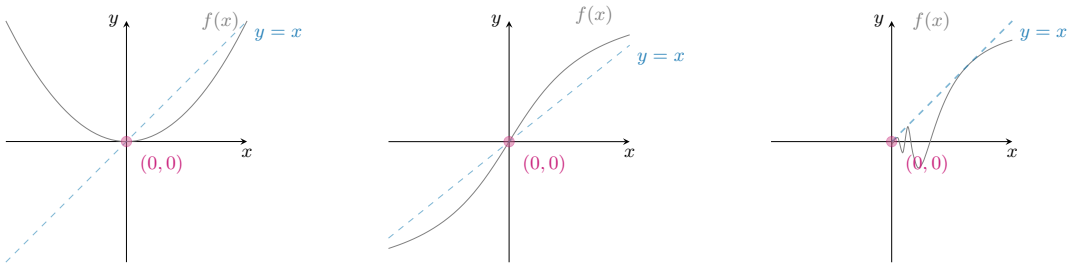
Alternatively, $(x_k -_{k \geqslant 0}$ can be defined by : $x_k = \mathcal{G}^k(x)$ (resp. $\mathcal{N}^k(x)$).

Let x^* be a point such that $f(x^*) = 0$ and $[Df(x^*)]^{-1}$ exists, then x^* is a fixed point of both $\mathcal{G}$ and $\mathcal{N}$. The convergence of the sequence $\mathcal{D}^k(x_0)$ for $\mathcal{D} \in \{\mathcal{G}, \mathcal{N}\}$ towards x^* can be related to the nature of x^* as a fixed point of $\mathcal{D}$. All types of behaviour for a continuous function around its fixed points are possible: a fixed point can be "attracting" the image by the functions of points in the vicinity of x^* are sent closer to x^*, "repulsive" if the opposite happens, or neutral if the behaviour of the function in its vicinity fits neither of the preceding categories **[Ford, 2005]** see figure 4.4.

The fixed-point theorem identifies one situation that guarantees convergence of the Gradient and the Newton sequence with a geometric rate.

Theorem 6. *(Banach-Picard fixed point theorem) Let $X \subseteq E$ and $f : X \to E$ be a contraction, i.e a map $f : X \to X$ with constant $0 \leqslant k < 1$ such that*

$$\|f(x) - f(y)\| < k\|x - y\| \qquad \forall x, y \in X$$



(a) Attracting fixed-point. **(b)** Repulsive fixed-point. **(c)** Neutral fixed-point.

Fig. 4.4. – The point $(0,0)$ is a fixed point for all three functions represented above. The graph of the bissectrice $y = x$ is given as a reference to compare empirically with the tangents on the graph of f at $(0,0)$. The position of the tangents relatively to the bissectrice gives the nature of the fixed point. If the tangents of the graph of $|f|$ are always below the graph of $x \mapsto |x|$ then $(0,0)$ is attractive, if they are always above it is repulsive and if their relative position varies it is unpredictable hence neutral.

Then f has a unique fixed point $x^* \in X$. Furthermore, for $x_0 \in X$, the sequence $(x_n = f^n(x_0))_{n\geqslant 0}$ converges geometrically to x^*. The rate of convergence of the sequence x_k is geometrical:

$$d(x^*, x_n) \leqslant \frac{k^n}{1-k} d(x_0, f(x_0))$$

It is clear from previous theorem that if there exists $k < 1$ and a region V_{x^*} around x^* where

$$\left\| x - y + \frac{1}{L}(f(y) - f(x)) \right\| \leqslant k\|x-y\| \quad \text{resp.} \quad \|x - y - [Df(x)]^{-1} f(x) - Df(y)]^{-1} f(y)\| < k\|x-y\|$$

then the Gradient sequence (resp. Newton sequence) intialized at any point in V_x converges geometrically toward x^*. This is typically the ideal setting in which to apply optimization algorithms. In the rest of this section, we give conditions and results on the convergence of the Newton and Gradient methods.

4.2.2 Convergence and Convexity

Given a descent method $\mathcal{D}$, and an initial point x_0, the critical question is to know wether or not the sequence $(x_k)_k$ converges. The answer is rarely trivial. For instance, the sequence $(x_k)_k$ generated by the Newton method or the gradient descent does not always converge. The convergence depend on the initialization point x_0, the behaviour of f around x_0 and metaparameters such as the step-size for the Gradient descent. For certain class of functions however, the convergence is guaranteed. The class of convex functions for instance is particularly well-behaved for optimization algorithms.

Rate of Convergence

We suppose in this paragraph that the sequence (x_k) converges towards a zero x^* of f. The rate of convergence is the speed at which the sequence $(x_k)_k$ converges towards its limit x^*. It is characterized by the ratio of the distance of two successive iterations with x^*. This ratio measures how much closer to the solution we get in one iteration. The quantity of interest is $\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^\alpha}$ where α is the largest real number for which the ratio is bounded. We already assumed that the series converges, we furthermore make the assumption that it never equals its limit so that the denominator $\|x_k - x^*\|$ never cancels. This hypothesis is naturally respected when x_k is generated by an iterative method as such presented before, because x^* being a fixed point of the iterative map, if the sequence x_k achieves x^* it becomes stationary and we can limitate our study to what happens before. Under this hypothesis, D'Alembert criteria shows that the ratio $\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|}$ must always be less or equal to 1.

Definition 31. *Let $(x_k)_k$ be a convergent sequence with limit x^*. The convergence is said to be:*

- *linear if there is a constant $0 < \tau < 1$ such that $\forall k > k_1, \|x_{k+1} - x^*\| \leqslant \tau \|x_k - x^*\|$*
- *superlinear if $\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \to 0$*
- *quadratic if there is a constant C such that $\forall k > k_1, \|x_{k+1} - x^*\| \leqslant C\|x_k - x^*\|^2$*

higher order of convergence can be defined, but we will not try in this document to reach for better than quadratic convergence which is sufficiently fast for most possible applications.

Basin of convergence and approximate zeros

Definition 32. *For a function f with a zero x^*, the basin of convergence of an iterative root-finding method $\mathcal{D}$ around x^* is the real number ρ such that there exists $i \in \{1, 2, \dots, \}$, $C < 1$ and*

$$\forall x \in \mathcal{B}(x^*, \rho),\ \|D^k(x) - x^*\| < C\|D^{k-1}(x) - x^*\|^i$$

The convergence radius defines the radius of the ball centered in x^* where the iterative method $\mathcal{D}$ converges at least linearly towards x^*.

Definition 33. *An approximate zero for a root-finding iterative method $\mathcal{D}$ is an element x such that the sequence*

$$x_0 = x \qquad\qquad x_k = D^k(x)\ \ \forall k \geqslant 1$$

satisfies

$$\forall k > k_1, \|x_{k+1} - x^*\| \leqslant C\|x_k - x^*\|^2$$

for some x^* such that $f(x^*)$ and some $C \in \mathbb{R}_+^*$.

The root-finding method initialized at an approximate zero converges quadratically to a root of f. Typically, approximate zeroes offer the guarantee to approximate a zero of f with any precision in a low number of iterations **[Blum et al., 2012]**.

Convexity

Convexity is a property that describes the graph of a function $f : \mathbb{R}^n \to \mathbb{R}$. In optimization, it guarantees the convergence of most iterative processes.

Definition 34. *A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex on a region U if for all $x, y \in U$*

$$\forall 0 \leqslant t \leqslant 1 \quad f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y)$$

For functions that are twice differentiable, we have a characterization of convexity through the Hessian matrix:

Proposition 30. *A twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex on a region U if and only if the eigenvalues of the Hessian matrix $\operatorname{Hess} f(x)$ are positive for all $x \in U$.*

We define α-strong convexity as the following notion:

Definition 35. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is α-strong convex on a region U if it is twice differentiable and for $(\lambda_1(x), \ldots, \lambda_n(x))$ the eigenvalues of $\operatorname{Hess} f(x)$,*

$$\inf_{x \in U} \{\lambda_1(x), \ldots, \lambda_n(x)\} \geqslant \alpha > 0$$

4.2.3 Gradient descent

Presentation

The Gradient descent or steepest descent algorithm follows a very simple principle: it follows the direction towards which the function locally decreases. Given a function $f : \mathbb{R}^n \to \mathbb{R}$ and its gradient ∇f, the gradient map is the function

$$\mathcal{G} : x \mapsto x - \frac{1}{L} \nabla f(x)$$

where L is a parameter called the "step-size" that must be specified beforehand. Hence, each iterate from the gradient descent is obtained as $x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$.

Convergence analysis

The convergence of the gradient descent is at best linear, and the only kind of equations that can be solved in only on iteration using the gradient descent are linear equations. It is generally known to not be a very efficient optimization method and needs strong guarantees to ensure convergence. However, when these conditions can be met, the gradient descent is a cheap algorithm that can improve at each iteration the approximation of a solution by a factor 2 without consuming a lot of resources. The following theorem gives conditions on the convergence and efficiency of the Gradient descent.

Theorem 7. *Let x^* be a critical point of f. If there is a ball $\mathcal{B}(x^*, \varepsilon)$ such that for any $x \in \mathcal{B}$ the eigenvalues of $\mathrm{Hess}\, f(x)$ are strictly positive and the ratio between the biggest eigenvalue $L(x)$ and the lowest eigenvalue $K(x)$ satisfies $\frac{K(x)}{L(x)} \geqslant \frac{7}{8}$ then the gradient descent iteration*

$$x_{k+1} = x_k - \frac{1}{L}\nabla(f)(x_k) \tag{4.10}$$

Where L is an upper bound on $\|Hf(x_k)\|$ satisfies:

1. $\|x_{k+1} - x^*\| \leqslant \frac{1}{2}\|x_k - x^*\|$

Proof.

$$\begin{aligned}
x_{k+1} - x^* &= x_k - x^* - \frac{1}{L}[\nabla f(x_k) - \nabla f(x^*)] \\
&= x_k - x^* - \frac{1}{L}\int_0^1 H(x_k + t(x^* - x_k))(x^* - x_k)dt \\
\|x_{k+1} - x^*\| &\leqslant \|x^* - x_k\| \int_0^1 \left\| I_d - \frac{1}{L}H(x_k + t(x^* - x_k)) \right\| dt \\
&\leqslant \|x^* - x_k\| \mathrm{sup}_{x\in\mathcal{B}} \left\| I_d - \frac{1}{L}H(x) \right\|
\end{aligned}$$

To bound the norm $\|Id - \frac{1}{L}H(x)\|$ we developp it using the trace:

$$\begin{aligned}
\left\| Id - \frac{1}{L}H(x) \right\|^2 &= \frac{1}{n}Tr([Id - \frac{1}{L}H(x)][Id - \frac{1}{L}H(x)]^T) \\
&= \frac{1}{n}[Tr(Id) - \frac{2}{L}Tr(H(x)) + \frac{1}{L^2}Tr(H(x)H(x)^T)] \\
&= 1 - \frac{2}{nL}\sum_{i=1}^{n}\lambda_i + \frac{1}{nL^2}\sum_{i=1}^{n}\lambda_i^2
\end{aligned}$$

given that $-1 \leqslant -\frac{2}{nL}\sum_{i=1}^{n}\lambda_i + \frac{1}{nL^2}\sum_{i=1}^{n}\lambda_i^2 \leqslant 0$, and naming $K = \min_{i,x}\lambda_i(x)$, we have

$$\begin{aligned}\left\|Id - \frac{1}{L}H(x)\right\|^2 &\leqslant 1 - min_{x\in\mathcal{B}}\left(\frac{2}{nL}\sum_{i=1}^{n}\lambda_i(x)\right) + max_{x\in\mathcal{B}}\left(\frac{1}{nL^2}\sum_{i=1}^{n}\lambda_i^2(x)\right)\\ &\leqslant 1 - \left(2\frac{K}{L} - 1\right)\\ &\leqslant 2(1 - \frac{K}{L})\end{aligned}$$

If $\frac{K}{L} \geqslant \frac{7}{8}$, then

$$\left\|Id - \frac{1}{L}H(x)\right\|^2 \leqslant \frac{1}{4}$$

and the result follows. □

Note that the hypothesis used in the theorem implie that f is α-strong convex for $\alpha = \frac{7}{8}L$. This theorem identifies a basin of convergence around x^* where convergence happens at least linearly with a coefficient $\frac{1}{2}$. Observe that the constant of the linear convergence rate is directly related to the maximal ratio of eigenvalues of $\mathrm{Hess}\, f(x)$ in a region around x^*. It is well-known that the Gradient descent, even when it converges, is very sensitive to the conditioning of the problem. Indeed, assume the ratio $\frac{K}{L}$ to be close to $\frac{1}{2}$, then the rate of the linear convergence becomes close to 1, meaning each step gets barely closer to x^* than the previous. The authors in **[Karimi et al., 2020]** have identified another set of conditions under which good convergence properties of the Gradient descent can be achieved.

Definition 36. *(Polyak-Lojiasewicz Inequality) A function satisfies the PL inequality on an open set U containing a local minimum x^* if $\nabla f(x^*) = 0$ and for some $\mu > 0$ the following holds:*

$$\frac{1}{2}\|\nabla f(x)\|^2 \geqslant \mu(f(x) - f(x^*)), \forall x \in U.$$

This inequality requires that the gradient grows faster than the the function moves away than its minimal value. Note that if the PL inequality holds on an open set U, then x^* is necessarily the unique minima of f on U.

Theorem 8. *Let f be a function with a L-Lipschitz continuous gradient, and at least one local minima x^* satisfying $\nabla f(x^*) = 0$ on its domain of definition. Assume f satisfies the PL inequality, then the gradient method with step-size $\frac{1}{L}$*

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

has a global linear convergence rate

$$f(x_k) - f(x^*) \leqslant \left(1 - \frac{\mu}{L}\right)^k (f(x^0) - f(x^*))$$

Interestingly in this theorem, the convergence rate applies to how fast the value of f gets closer to its minimal value, not on the actual convergence of the sequence (x_k).

4.2.4 Newton method

Presentation

For univariate functions, the Newton method can be seen as an improvement of the Gradient method where an adaptative strategy for the choice of the step-size L has been found. For multivariate functions however, it appears that the Newton method differs from the gradient descent by the choice of the descent direction, and not only by its norm. The Newton map for a function f is:

$$\mathcal{N} : x \mapsto x - \mathcal{H}f(x)^{-1} \cdot \nabla f(x)$$

The graphical idea behind the Newton method is that it fits a parabola to the graph of f and finds its minimum in one step.

Convergence analysis

The convergence of the Newton method is easier than for the gradient descent. Namely, if a function has convex regions, i.e. domains on which the function is locally convex, the Newton method initialized in one of these regions converges quadratically to the unique minima in that region. However, when not in a convex zone, the Newton sequence can be periodic, divergent or converge to the wrong point. This algorithm is therefore extremely sensitive to initialization, but very efficient once correctly initialized. This sensitivity to initialization is exemplified in figure 4.5.

Next theorem gives a result on the convergence of the Newton sequence initialized in a convex region of the graph of f. This result will be used in chapter 5 where each critical point of a function f is contained in a convex region of the function.

Theorem 9. *Given x^* a critical point of $f : \mathbb{R}^n \to \mathbb{R}$, suppose there exists $r, L, \alpha > 0$ such that:*

1. *The function f is α-strong convex on $\mathcal{B}(x^*, r)$*
2. *$\forall x, y \in \mathcal{B}(x^*, r),\ \|\text{Hess}\, f(x) - \text{Hess}\, f(y)\| \leqslant L\|x - y\|$*

Then, $\forall x_0 \in \mathcal{B}(x^*, r),\ \ \forall k \in \mathbb{N}$ the Newton sequence $(x_k)_{k\geqslant 0}$ initialized at x_0 satisfies:

$$\left\| \frac{x_k - x^*}{\gamma} \right\| \leqslant \left\| \frac{x^* - x^0}{\gamma} \right\|^{2^k}$$

where $\gamma = \frac{2\alpha}{L}$.

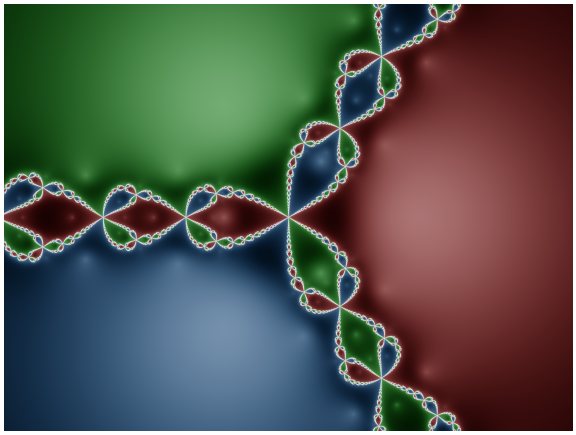


Fig. 4.5. – The Newton method is applied to the function on $\mathbb{C}$ that to z maps $|z^3 - 1|$. Each of the three colour represents the basin of convergence of one of the three roots of $z^3 - 1$. Initializing the Newton method in the green region yields the root z_1, in the blue region it yields z_2 and the red region yields z_3. The colored region draw a fractal division of the complex plane. In the central region of the figure, a very small error in the initialization can lead to the convergence towards the wrong root. This shows how sensitive to initialization the Newton method can be even for simple polynomial functions. (Source: Wikipedia)

Under the hypothesis of the theorem, if $r < \frac{2\alpha}{L}$ any $x \in \mathcal{B}(x^*, r)$ is an approximate zero of ∇f. Note that α-strong convexity were not enough and that a bound on $D^3 f$ was also required. Typically, the Newton method is extremely efficient when it is correctly initialized, here the bounds become worthless if $\|x^* - x^0\| \geqslant \gamma$ which shows that even inside of the convex region the convergence is not quadratic everywhere.

4.2.5 Newton method on Manifolds

In all that preceded, functions were defined on Banach spaces provided with a linear structures. To extend optimization algorithm to non-linear spaces with a differentiable structure, one has basically two options. Either you consider an optimization algorithm is characterized by what it does, or by how it does it. The Newton method for instance can be characterized by the operations that produce each iterate, or by the idea of approximating a curve by a second-order polynomial at each step to find its zeros. This two different views yield two different ways to generalize optimization algoritms to non-linear spaces. The one we will present is based on the work of **[Absil et al., 2009]** and transposes the algebraic operations that produce Newton iterates to manifolds. Another approach can be found in **[Manton, 2015]**.

Reformulation

Let $g\colon \mathcal{M} \to \mathbb{R}$ be a function defined on a manifold, and our objective is to minimize g, that is to find the zeroes of the derivative (or the gradient) of g. We can not define the sequence

$$x_0 \in \mathcal{M}, \quad x_{k+1} = x_k - [\operatorname{Hess} g(x)]^{-1} \cdot \nabla g(x)$$

in $\mathcal{M}$. Remember that in virtue of the Riesz theorem, the linear form $\nabla g(x)$ on the tangent space at x of $\mathcal{M}$ can be interpreted as a tangent vector in $T_x\mathcal{M}$. $[\mathrm{Hess}\, g(x)] \cdot \nabla g(x) \in T_x\mathcal{M} \notin \mathcal{M}$. Secondly, even if the latter expression belonged to $\mathcal{M}$, there are no ways to "deduct" a point from another point in $\mathcal{M}$. Hence, these two steps must be adapted to the context of manifolds. The idea of the Newton method is to move on the manifold in a direction given by the inverted Hessian applied to the gradient. Moving on the manifold in a precise direction is a notion that is conveyed by curves, as we have seen in section 4.1. Hence, the point x_{k+1} can be obtained from the point x_k by following a curve for a certain distance. The curve can be determined by its tangent vector at x_k, we have seen in section 4.1.5 that there is a unique geodesic passing through x_k tangent to a certain direction. Using the exponential mapping defined in (4.1.5) the iterate x_{k+1} can be defined as:

$$x_{k+1} = exp_{x_k}\left(-[\mathrm{Hess}\, g(x)]^{-1} \cdot \nabla g(x)\right)$$

Convergence analysis

Very conveniently, all the convergence analysis we conducted in Euclidian spaces can be reused to study the convergence on the Newton method on manifolds. Indeed, assume that for $k \geqslant K \in \mathbb{N}$, all the iterates (x_k) of the Newton method can be contained in the same coordinate patch (U_K, φ_K) in $\mathcal{M}$, and that $\mathrm{Hess}\, g(x_k)$ and $\nabla g(x_k)$ are written in the induced coordinate basis of the tangent space. Let $(\tilde{x}_k)_{k\geqslant 0}$ be the Newton sequence generated by $f \circ \varphi - 1 \colon \mathbb{R}^n \to \mathbb{R}$. Then, it derives from proposition 29 that

$$\varphi(x_{k+1}) = \varphi(\exp_{x_k}\left(-[\mathrm{Hess}\, g(x)]^{-1} \cdot \nabla g(x)\right) = \varphi(x_k) - [\mathrm{Hess}\, g(x)]^{-1} \cdot \nabla g(x)$$
$$\implies \varphi(x_{k+1}) - \varphi(x_k) = \tilde{x}_{k+1} - \tilde{x}_k$$

In particular, if $(\tilde{x}_k)_k$ converges quadratically to $\tilde{x}^*$, then by injectivity of φ, $(x_k)_k$ converges to x^*. To qualify the nature of the convergence in the manifold, we equip $\mathcal{M}$ with a norm $\|\cdot\|_{\mathcal{M}}$. Note that $\|\varphi(\cdot)\|$ is also a norm on $\mathcal{M}$ and that when composed with φ^{-1}, they both yield a norm on $\mathbb{R}^n$. All norms are equivalent in $\mathbb{R}^n$ therefore there are $c \leqslant d \in \mathbb{R}_+^*$ such that

$$\forall x \in \mathcal{M}, \quad c\|\varphi(x)\| \leqslant \|x\|_{\mathcal{M}} \leqslant d\|\varphi(x)\| \tag{4.11}$$

Consequently, if $\psi(x_k)$ converges quadratically to x^* then x_k also converges quadratically to x^* for the norm considered.

Convexity on a manifold: Riemannian geometry provides tools to define convexity of $\mathbb{R}$-valued functions on a manifold independently from the coordinate charts **[Ahmad et al., 2019]**. As we have made the choice to not developp Riemannian tools in this thesis, we will simply use the following consideration: if f is twice differentiable and $\mathrm{Hess}\, f(x)$ is the Hessian of f at x in the coordinate basis induced by φ, then if $\lambda_{\min}(\mathrm{Hess}\, f(x)) = \alpha$, the function $f \circ \varphi^{-1}$ is α-convex. In particular an extension of theorem 9 can be derived on manifolds.

Theorem 10. *Let $f\colon \mathcal{M} \to \mathbb{R}$ a function on a manifold and x^* a critical point of f such that $Df(x^*) = 0$ (this equality is independant of coordinate charts). Suppose there exists $r, L, \alpha > 0$ such that:*

1. $\inf_{x \in \mathcal{B}(x^*, r)} (\lambda_1(\operatorname{Hess} f(x)), \dots, \lambda_n \operatorname{Hess} f(x)) \geqslant \alpha$
2. $\forall x, y \in \mathcal{B}(x^*, r),\ \|\operatorname{Hess} f(x) - \operatorname{Hess} f(y)\| \leqslant L\|\varphi(x) - \varphi(y)\|_{\mathcal{M}}$

Then, $\forall x_0 \in \mathcal{B}(x^*, r),\ \ \forall k \in \mathbb{N}$, there exists $\varepsilon > 0$ such that the Newton sequence $(x_k)_{k \geqslant 0}$ initialized at x_0 satisfies:

$$\frac{\|x_k - x^*\|_{\mathcal{M}}}{\gamma} \leqslant \varepsilon \frac{\|x^* - x^0\|_{\mathcal{M}}^{2^k}}{\gamma^{2k}}$$

where $\gamma = \frac{2\alpha}{L}$.

Proof. From the equivalence of norms written in (4.11) we have:

$$\|x_k - x\|_{\mathcal{M}} \leqslant d\|\tilde{x}_k - \tilde{x}^*\|$$

and theorem 9 applied to the function $f \circ \varphi^{-1}$ yields

$$\tilde{x}_k - \tilde{x}^* \leqslant \frac{\|\tilde{x}_0 - \tilde{x}^*\|^{2^k}}{\gamma}$$

Applying the equivalence of norm once again brings:

$$\|\tilde{x}_0 - \tilde{x}^*\| \leqslant \frac{1}{c}\|x_k - x^*\|_{\mathcal{M}} \quad \text{and finally} \quad \|x_k - x\|_{\mathcal{M}} \leqslant \frac{d}{c}\frac{\|x_k - x^*\|_{\mathcal{M}}^{2^k}}{\gamma^{2^k}}$$

□

Remark 2.5.1

There is something unsatisfying in the way we prove the convergence of the Newton sequence generated for a function f on a manifold. Because we reduced it to the convergence of the Newton sequence of $f \circ \varphi - 1$, it depends on the choice of parametrization. In **[Manton, 2015]**, the authors explore the idea of using this dependency at their advantage, to lower the complexity of the optimization problem directly in the parametrization.

4.3 Homotopy methods

Finding the zeros of a function has long been a preoccupying problem and with reasons since solving any equation resumes ultimately to a root finding problem. Most optimization problems can be reformulated so that a solution is the zero of a special function, usually a gradient and hence are solved by root-finding methods. These methods are usually iterative processus that hopefully converge to a root when the number of iteration becomes large. If the procedure converges, an arbitrarily good approximation of the convergence point can be obtained by stopping the procedure after a number of iteration. The difficulty is that for a given function, there is no equivalence between the divergence of the numerical method and the absence of roots, there are subsequently no guarantees that all the roots of a function will be found. Iterative procedures like the Newton method or the gradient descent for instance have well-known local properties, but potentially chaotic global behaviour. The sequence generated by a Newton method on a smooth function is not always a convergent sequence **[Wikipedia contributors, 2021a]**, it can be periodic, or chaotic even in the presence of roots. The importance of locality is such that the behaviour of the sequence generated by an iterative method usually depends strongly and sensitively on the first term which we call the initialization point. In this paragraph, we propose to present a group of root finding methods that address the problem of finding an initialization.

4.3.1 Context and concepts

Work on homotopy methods can be dated back as far as Poincaré in 1881, who came accross it while working on the 3-body problem[4]. The homotopy methods were designed to solve non-linear systems of equation $f(x) = 0$ by continuously deforming them into a simpler one **[Allgower and Georg, 2012]**. Thus it introduces a deformation or homotopy function H and a new parameter t that embedds the original problem:

$$\begin{aligned} H\colon [0,1] \times \mathbb{R}^n &\to \mathbb{R}^n \\ (t,x) &\mapsto H(t,x) \end{aligned}$$

such that $H(0,x) = g(x)$ is a simple function with known roots and $H(1,x) = f(x)$. The embedding is defined so that when t moves from 0 to 1, the solutions to $H(t,x) = 0$ move from the roots of g to the roots of f. An exemple for polynomial functions is shown on figure figure 4.6.

The solutions of the equation $H(t,x) = 0$ take the special form of disconnected smooth paths, we will show later why this is this way. We will refer to a particular disconnected component as a "solution curve" or "solution component" for the equation $H(t,x) = 0$. Such a component is fully determined by one of its point, so that we can refer to the component of

[4]The three-bodies problem in celestial mechanics is the problem of describing the motions of three bodies that attract each other with gravitational forces. It is trivial for two bodies, but shows particularly difficult from three to more bodies **[Wikipedia contributors, 2021c]**

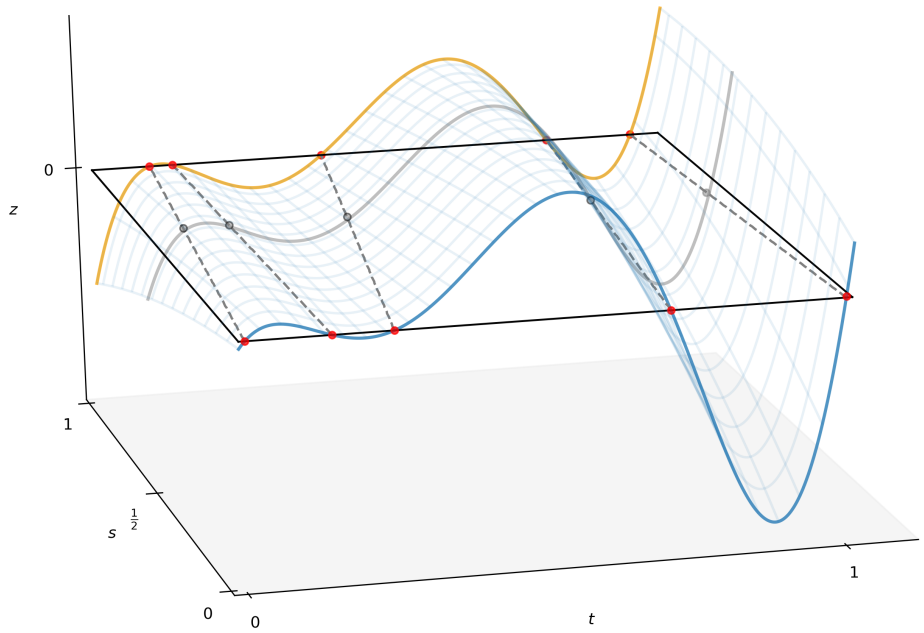


Fig. 4.6. – The wireframe represents a continuous deformation of the blue polynomial curve into the orange polynomial curve. This deformation is parametrized on the y-axis by parameter s. On the plane of altitude $z = 0$, paths between the zeros of the orange and blue curves are visible in dotted lines.

(t_i, x_i) unequivocally. The fact that each point belongs to at most one solution component will be argumented in the rest of this section.

Remark 3.1.1

Homotopy methods such as we understand them here differ from the homotopy analysis method (HAM) introduced by Liu [He, 2003] which relies on a Taylor expansion of one of the solution curve leading to a closed-form expression of the solution. The HAM is well-suited to non-linear differential system of equations, whereas the more general homotopy method deals with problems such as finding the roots of a polynom P given the roots of a polynom Q of the same degree.

The homotopy method in itself describes the idea of building a suitable map H that continuously sends a function g on a function f. The resolution of the problem of finding the roots of f requires a numerical method that allows to derive numerically the solutions of $f = 0$ from the solutions of $g = 0$. This numerical method is called a continuation method, it traces paths in $H^{-1}(0)$ between the roots of g and the roots of f. It is important to understand that the homotopy method does not provide a closed-form expression of the roots of f, but a smooth deformation between g and f. It is from the properties of this transformation that we can deduce (or not) the existence of paths elligible to a continuation method between all the

roots of g and all the roots of f. Homotopy methods are always coupled with continuation methods that iteratively sample the solution curves and produce estimates of the desired roots. The necessity to link together these two methods provides a first understanding of what good properties should be expected from the embedding function H. In the case of polynomials, supposing H sends Q continuously on P for instance, we can name three hypothesis **[Li, 1997]**:

1. triviality: the roots of Q are trivial to find.
2. smoothness: no singularities along the solution curves occur.
3. accessibility: all isolated roots of P can be reached.

To respect the second and third hypothesis, H must guarantee the existence of smooth paths connecting every root of Q to every root of P. The absence of singularity should guarantee in particular the absence of turning points or crossings of solution curves, and the third condition implies that two paths can not merge. These conditions are actually not specific to homotopy methods applied to polynomial, they are necessary to the numerical resolution of any smooth homotopy problem by a continuation method, as we will see later.

About the existence of the function H

The main difficulty of homotopy methods resides in the definition of a homotopy map satisfying the above properties. The problem is even more complex if several roots of a function are considered through the same homotopy. There are no constructive way to guarantee the existence of a suitable homotopy map between two functions g and f, even if the convex combination $H(t,x) = tg(x) + (1-t)f(x)$ is very often used as a baseline. In this chapter, we build a homotopy map in a way that guarantees all the properties listed above using the structure of a family of function, as it will be seen in section

The definition of a homotopy map H is always complementary to a given numerical method to solve the original root-finding problem, and it is the nature of this method that gives the particular conditions on H to ensure convergence. The next section will present the continuation-prediction method, and we will deduce what conditions H should satisfy for this method to converge.

4.3.2 Numerical methods to solve a homotopy problem

There are at least two different sets of method to solve numerically a homotopy problem. The predictor-corrector (PC) methods which will be our focus, and the piecewise-linear (PL) methods **[Allgower and Georg, 2012]**. They both start from the assumption that there is a path $s(t)$ defined implicitely by $H(t, s(t)) = 0$ and a starting point (t_0, s_0) such that $H(t_0, s_0) = 0$. The prediction-correction method starts from a sample $(t_0, s(t_0))$ and iteratively traces samples $(t_1, s(t_1)), \ldots, (t_n, s(t_n))$ until a traversing criteria, usually a condition on t_n, is

reached. Each sample is found in two steps: first a prediction step uses the prior information $(t_{n-1}, s(t_{n-1}))$ and predicts a new point $y_n = (t'_{n-1}, s'_{n-1})$ using the condition $H(t, s(t)) = 0$ (we will see more precisely how this condition can be used). Then, a correction-step projects the point y_n on a point $\omega_n = (t_n, s(t_n))$ that satisfies $H(w_n) = 0$. The core structure of a prediction-correction algorithm is shown below:

Algorithm 1 Generic continuation method

Ensure: A point $u_n = (t_n, x_n)$ satisfying $H(t_n, x_n) = 0$ within a given precision criteria

Require: $u_0 = (t_0, x_0) \in \mathbb{R} \times \mathbb{R}^N s.t. H(u_0) = 0$ and a stepsize $\tau_0 > 0$

1: **while** $|1 - t_n| > \varepsilon$ **do**

2: **Prediction:** Choose a steplength $\tau_k > 0$. Predict a point y_k such that $H(y_k) \simeq 0$ and $\|u_{k-1} - y_k\| \simeq \tau_k$.

3: **Correction:** Let ω be the solution to the minimization problem:

$$\arg\min_{\omega \in \mathbb{R} \times \mathbb{R}^n} \{\|y_k - \omega\| \mid H(\omega) = 0\} \tag{4.12}$$

4: set $u_k := \omega$.

5: Return $u_n = (t_n, x_n)$.

6: **end while**

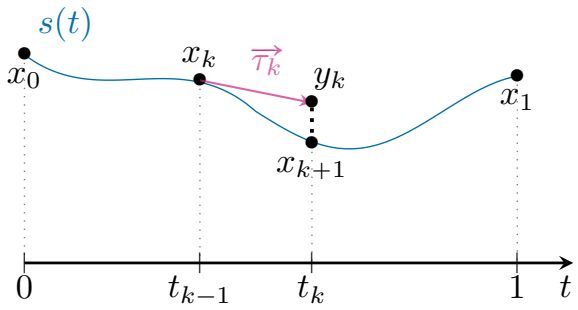


Fig. 4.7. – The different steps of algorithm 1 are represented on a theoretical curve $\{s(t) \mid t \in [0,1]\}$ represented in blue. At step k, the point x_k on the curve is used to compute the point y_k. This writes $u_k + \vec{\tau_k} = y_k$. The predicted point y_k is projected on $\{s(t) \mid t \in [0,1]$ using a correction step. The black dots on the curve represent the points effectively sampled by the continuation algorithm.

We will not unravel here all the subtleties of continuation methods, the interested reader can consult **[Allgower and Georg, 2012]** for a comprehensive study of homotopy methods and their numerical solutions. Rather, we want to give a quick idea of the convergence of such methods, and how the prediction and correction step can be performed.

About the step-size

The algorithm converges if the stopping criteria is reached, that is if for some n, $|1 - t_n| < \varepsilon$. From the structure of the algorithm

$$t_n \leqslant \sum_{k=0}^{n} \tau_k \tag{4.13}$$

so that the choice of the step-size at the start of the prediction step can determine the convergence of the algorithm. If the terms in $(\tau_k)_{k\geqslant 0}$ are small, the correction step is made easier but the convergence requires a large number of loops. If the $(\tau_k)_{k\geqslant 0}$ are so small that the series in (4.13) converges to a value less than one, then the algorithm has no chance to ever converge. We do not detail the choice of the step-size here, it depends on the particular expression of H in general. In chapter 5 where a continuation algorithm is built, we use a constant step-size. It is important to understand that what limitates the step-size is the correction step. The idea is that for a small step-size, the predicted point stay close to the curve see figure 5.5, and the correction step is faster to compute. For a big step-size, the distance between the curve and the predicted point becomes less predictable and so does the convergence of the correction step.

Prediction step

The classical setting for the prediction step is to work with functions H and s differentiable. Then, differentiating the relation $H(t, s(t)) = 0$ yields

$$[D_t H(t, s(t)) \;\; D_s H(t, s(t))] \begin{pmatrix} 1 \\ D_t s(t) \end{pmatrix} = 0 \tag{4.14}$$

We will use the following result:

Proposition 31. *Let $H : [0, 1] \times \mathbb{R}^n \to \mathbb{R}^n$ be a homotopy map and $DH(t, x)$ its derivative at the point (t, x). If $DH(t, x)$ has full rank, then there is a unique vector (α, β)*

$$\alpha > 0 \qquad \|(\alpha, \beta)\| = 1 \qquad DH(t, x) \cdot (\alpha, \beta) = 0.$$

Proof. This is a direct application of the rank-nullity theorem applied to the linear application $DH(t, x) : \mathbb{R}^{n+1} \to \mathbb{R}$. □

Therefore, if $DH(t, s(t))$ is of full rank, a vector colinear to $D_t s(t)$ can be found unequivocally in the kernel of $DH(t, s(t))$. Note that $D_t s(t)$ is graphically the vector tangent to the curve s at the point $s(t)$ so that a linear approximation of s around $s(t)$ would be:

$$s(t') \simeq s(t) + (t' - t) D_t(s(t)) \tag{4.15}$$

the prediction step is a first-order approximation of the curve at the point x_{k-1} so that

$$y_k = u_{k-1} + \tau_k \frac{1}{\left\| \begin{pmatrix} 1 & D_t s(t_{k-1}) \end{pmatrix}^T \right\|} \begin{pmatrix} 1 & D_t s(t_{k-1}) \end{pmatrix}^T \tag{4.16}$$

It must be noted that though the vector $[1 \;\; D_t s(t_{k-1})]$ could be used without normalization and would provide the strict equality $\sum_k^n \tau_k = t_n$, the litterature usually prefers to use a normalized version of this vector so that $\|y_k - u_{k-1}\| = \tau_k$. Therefore, the step-size τ_k is the upper bound on the distance between the predicted point and the solution curve. This is important in particular to have guarantees on the correction step. The step-size can be interpreted either as the amount of trust we have in the linear approximation of the curve by its tangent at u_k, or as an estimate of the convergence radius for the correction step. The two assumptions needed to perform a prediction step through a linearization of the solution curve is that the homotopy map is continuously differentiable with a full-rank derivative on the curve, and that the curve itself is differentiable. The following theorem shows that one assumption is enough.

Theorem 11. *Let $H\colon [0,1] \times \mathbb{R}^n \to \mathbb{R}^n$ be a continuously differentiable homotopy map. Let (t_0, x_0) a point where $H(t_0, x_0) = 0$ and suppose that $DH(t_0, x_0)$ has full-rank. Then, there exists $\delta > 0$, U_x an open set containing x and a curve $s\colon [t_0 - \delta, t_0 + \delta] \cap [0,1] \to U_x$ such that $H(t, s(t)) = 0$ and s is continuously differentiable.*

Proof. This is a direct consequence of the implicit function theorem applied to H, see section 4.1.6. □

Without a correction step, sequential iterations of linear approximations can lead to a huge difference between the discrete points computed and the real solution curve, as it is represented on the figure 4.8. This approximation is also knwon as the Euler discretization of a curve, and it has been shown that the global error (i.e the difference between the last point computed by the Euler approximation and the same abscissa point on the curve) made by a Euler discretization on a smooth curve is in $O(\tau)$ **[Taras I., 2021]**. To gain in precision, a correction step that projects each point back on the curve is necessary.

Correction step

The correction step is applied on a predicted point y_k and should match y_k to the appropriate point u_k on the curve $s(t)$. The appropriate point is usually defined by the closest point to y_k on the curve s. This step is realized through an iterative root-finding method, starting at y_k and applied to H. The curve s is a natural attractor for the root finding method because it is in the kernel of H. To make convergence certain in a region around this curve, it should be isolated: the points in the neigborhood of the curve should not be in the kernel of H. It can be shown that when iterative methods like the Gradient descent and the Newton

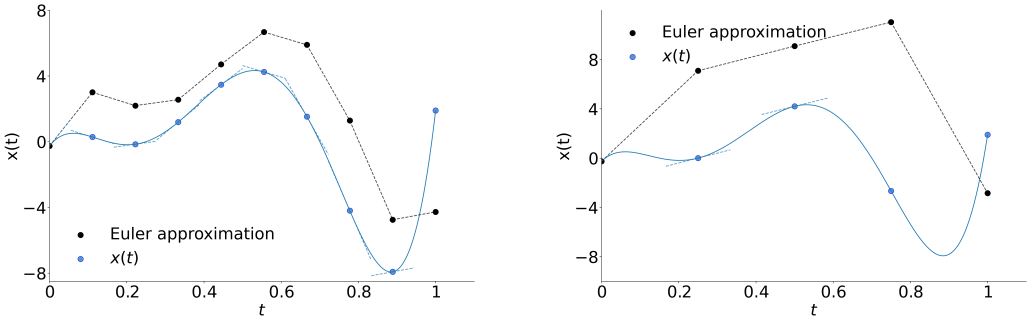


Fig. 4.8. – Examples of Euler discretizations of a polynomial curve for two different constant step-sizes: on the left $\tau = 0.1$ on the right $\tau = 0.25$.

method converge to a point on the curve, this point is approximately the minimizer of $d(y_k, \{s(t) \mid t \in [0,1]\})$.

Optimization Geometry

5

Contents

5.1 Introduction

A classic formulation for an optimization problem is to minimize a function $f : x \mapsto f(x)$ over a domain space $x \in X$, subject to constraints A. If the function f has some features such as global or local convexity, then numerical method such as the Newton descent will have sufficiently good convergence properties to consider the problem solvable, this was detailed for the Newton and Gradient methods in section 4.2. Otherwise, an approach consists in reformulating the minimization problem so that it displays good properties. In **[Delage and Ye, 2010]** the authors use strong duality to that effect. Alternatively, **[Zhou and So, 2017]** developp new criteria inside of which optimization method can be used reliably such as the error bounds models. All these methods need two fundamental elements to behave properly: a function f with nice properties, and a good starting point for the iterative method supposed to solve the optimization problem. The difficulty of the second point has been argued already in section 4.2.

Real-Time optimization In this thesis, real-time optimization describes the challenge of providing an estimate of $\arg\min_{x \in X} f(x; \theta)$ for an input value $\theta \in \Theta$ in a short, or at least predictable time. The parametrized function $f(\cdot, \theta)$ is known in advance. Such problems are faced for instance in the bayesian estimation of a probability law. Assume there is a prior that the probability law has a density of the shape $f(\boldsymbol{x})$ where $\boldsymbol{x}$ is a vector of parameters. Then, for observations θ, the best estimate $\boldsymbol{x}^*$ minimizes the log-likelihood $\mathcal{L}(x; \theta)$, whose shape is known in advance. If observations θ vary slowly through time, then under smoothness assumptions on the family $\{f_\theta; \theta \in \Theta\}$ it can be derived that solutions $\{\arg\min_{x \in X}\} f(x; \theta)$[1] also vary slowly (considering a distance between subsets of X). This classifies as an adaptative problem, naturally leading towards tracking methods. Essentially, tracking methods when confronted to an adaptative optimization problem, take advantage of the proximity of two succesive solutions to find costlessly a good starting point for iterative optimization methods. If x^* is a solution at time t, then a $t + \Delta t$ the solution writes $y^* = x^* + \Delta x^*$ and given that Δx^* is small enough, an iterative procedure for the optimization problem at $t + \Delta t$ starting at x^* converges super-linearly to $x^* + \Delta x^*$. If however two successive observations are not correlated, real-time optimization problems tend to be treated as once-off optimization problems, and knowledge on the family of functions f_θ is of no use. In this chapter, we propose a method for real-time uncorrelated optimization problems, where a compromise between tracking methods and once-off optimization is found. The two major contributions of the method we propose is that it guarantees to find all the local minima of a function $f_\theta : X \to \mathbb{R}$ and it guarantees a higher-bound on the complexity of the algorithm.

Complexity If a real-time optimization problem defined by $f : X \times \Theta \to \mathbb{R}$ is treated as once-off optimization, the complexity of the problem at each time t depends on the functional $x \mapsto f(x; \theta)$, and if the function is "complicated" with regards to optimization algorithm, then

[1] Note that this set has a number of elements that can vary from one θ to another, as f_{θ_1} and f_{θ_2} can have a different number of minimizer.

at each time t a complicated problem must be solved. Even when tracking is impossible due to lack of correlation between observations, we believe as in **[Manton, 2012]** it is possible to reduce the overall complexity of the problem by using pre-computed informations, such as solutions found for previous observations. Supposing $f(\cdot;\theta)$ has a unique global minimum x^*_θ for each θ, it was shown in **[Manton, 2012]** that there exists locally an implicit function $g\colon \theta \to x^*_\theta$, and we are looking for an algorithm whose complexity would depend mainly on the features of g which we consider to be the inner complexity of the problem. In the example of log-likelihood presented earlier, it can be noted that if the densities belong to the Gaussian family, then there is a closed expression providing for each set of observations θ the best parameters x^*. The function g in this case is both explicit and global. The perspective is shifted from a focus on x to a focus on θ parametrizing the family of function $\{f(\cdot;\theta)\}_{\theta\in\Theta}$. See figure figure 5.1.

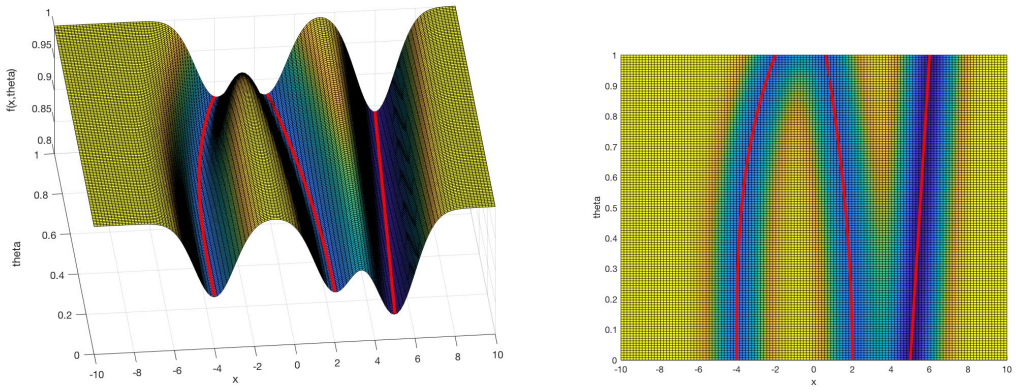


Fig. 5.1. – For each θ, the function $f_{\boldsymbol{\theta}}(x)$ is a mixture of gaussian. The local minima of the function are the means of the three gaussians in the mixture. The means vary smoothly with θ and are easy to track. On the right, the red curves trace the local minima in the (x,θ) plane, and would be the graph of three implicit g functions that we introduced earlier. It can be observed that the graph defined by the red curves is much simpler than the graph of a single function f_θ

We call "optimization geometry" this framework attached to consider families of function as a whole rather than separate functions. Consider the following example where f is defined by $f(x;\theta) = (x-\theta)^2$ then it is obvious that the minimum of this funtion is reached when $x = \theta$. In this context, the function g would be $g(\theta) = \theta$. If this function g is explicitely available, which is generaly not the case, it is obvious that the shape of f does not matter. Suppose h is any complicated non-convex function with a root in zero, then if the real-time optimization problem amounts to minimizing $f(x;\theta) = h(x-\theta)$, we consider the intrinsic complexity of this problem to be the same as our first trivial example. Our idea as in **[Manton, 2012]** is that it is really the complexity of g that defines the intrinsic complexity of the optimization problem rather than the complexity of $f(.,\theta)$ for a given θ.

The goal of optimization geometry is to use precomputations on a family of functions to simplify the quest for the local or global minima of one element in this family. A first step in doing so is to show that for nice enough functions, the critical points fit together in a nice

way meaning it is possible to define a smooth path of critical points from one critical point to another.

5.2 Main ideas

5.2.1 Smooth path between critical points

First, we define a nice class of functions on which the method we will propose has guaranteed performances. This is done in definition 37.

Definition 37. *Let X, Θ be two smooth compact manifolds without boundaries, with dimensions k and n respectively, and $M = X \times \Theta$ their Cartesian product of dimension $k + n$. Let $f : X \times \Theta \to \mathbb{R}$ be a smooth cost function on M. f is called* fibre-wise Morse *on M if the restriction*

$$\begin{aligned} f_\theta : \quad X &\to \mathbb{R} \\ x &\mapsto f(x;\theta) \end{aligned}$$

is a Morse function for every $\theta \in \Theta$. We recall that a Morse function is a smooth real-valued function that has no degenerate critical points ***[Guillemin and Pollack, 2010]****.*

A direct consequence of the definition of a Morse function is that its critical points are isolated. A less direct but still close consequence is that Morse functions on compact sets can only have a finite number of critical points. These two elements are of importance in the definition of an homotopy method in section 4.3.

Remark 2.1.1

However strong the fibre-wise Morse assumption might look, it is actually quite cheap. Considering Morse functions are dense in the set of smooth functions (see **[Guillemin and Pollack, 2010]**), it is reasonable to believe that the general case is that f_θ is Morse for almost every $\theta \in \Theta$. If it is clear that there exists one θ where f_θ is not Morse, our analysis can be conducted on the submanifold (with boundary) $\Theta \backslash V_\theta$ where V_θ is an open set containing θ.

Definition 38. *A point (x^*, θ) is called* fibre-wise critical *if x^* is a critical point of f_θ i.e*

$$Df_{\boldsymbol{\theta}}(x^*) = 0$$

The next theorem shows how critical points of a smooth family of Morse functions fit together in a nice way, and is the first step in showing why the family of functions rather than just the individual cost function should be considered.

Theorem 12. *The set $\tilde{N}$ of fibre-wise critical points of the fibre-wise Morse function $f\colon M = X \times \Theta \to \mathbb{R}$ is a n-dimensional smooth submanifold of M in the sense of [Guillemin and Pollack, 2010]. Furthermore, it is topologically closed and has no boundaries.*

Proof. First, $\tilde{N}$ is not empty as we do not consider the degenerate case where M, F are 0-dimensional manifolds and a Morse function can not be constant. By working locally, one can show that $\tilde{N}$ is actually made of the level sets of 0 for particular functions. Let (U, ϕ) be a coordinate patch on X. The function $f_\theta \circ \phi^{-1}\colon \mathbb{R}^k \to \mathbb{R}$ satisfies the following equivalences for all x in U and θ in Θ:

$$D[f_\theta \circ \phi^{-1}](\phi(x)) = 0_{\mathbb{R}^k} \iff D f_{\boldsymbol{\theta}}(x) = 0 \tag{5.1}$$

$$\operatorname{rank}\left(D^2[f_\theta \circ \phi^{-1}](\phi(x))\right) = \operatorname{rank}\left(D^2 f_{\boldsymbol{\theta}(x)}\right) \tag{5.2}$$

The notation 0 on the right-hand side of the equation stands for the linear application in $\mathcal{L}(T_x X, \mathbb{R})$ that sends every vector in the tangent space $T_x X$ on 0. These equivalences derive from the notion that the derivative of a function composed with a coordinate chart coincides with the derivative of the function on the manifold expressed in the basis of tangent vectors defined by the coordinates. Let $g\colon U \times \Theta \to \mathbb{R}^k$ be defined by:

$$g\colon (x, \theta) \mapsto D[f_\theta \circ \phi^{-1}](\phi(x)) \tag{5.3}$$

then, $\tilde{N} \cap U \times \Theta = g^{-1}(\{0_{\mathbb{R}^k}\})$. Indeed, let $(x, \theta) \in g^{-1}(\{0\})$, then, by the equivalence in equation (5.1), $D f_{\boldsymbol{\theta}(x)} = 0$ so that (x, θ) is a fibre-wise critical point of f. Conversely, if $(x, \theta) \in \tilde{N}$ and $x \in U$, then $D f_{\boldsymbol{\theta}(x)} = 0$ and by equation (5.1) $(x, \theta) \in g^{-1}(\{0\})$. To apply the regular value theorem to g, it is necessary to show that 0 is a regular value. First g is differentiable because f is smooth. The derivative of g is a linear application that can be written as a matrix in a basis of tangent vectors of M at (x, θ). Ordering the tangent vectors as $(\partial x_1, \dots, \partial x_k, \partial \theta_1, \dots, \partial \theta_n)$ where $(x_1, \dots, x_k)$ are coordinates on X and $(\theta_1, \dots, \theta_n)$ are coordinates on Θ, the matrix expression of Dg has the shape:

$$Dg(x, \theta) = \Big(\underbrace{D^2[f_\theta \circ \phi^{-1}](\phi(x))}_{k \times k} \quad \underbrace{D_\theta[D f_\theta \circ \phi^{-1}](\phi(x))}_{k \times n} \Big) \tag{5.4}$$

Let $(x, \theta) \in g^{-1}(0_{\mathbb{R}^k})$, then $D^2 f_{\boldsymbol{\theta}(x)}$ is invertible because x is a critical point of the Morse function f_θ. By equation (5.2) one deduces that $Dg(x, \theta)$ is of full rank k. As this is true for any antecendent of 0 by g, 0 is a regular value of g and its preimage is a smooth submanifold

of $U \times \Theta$ of dimension n by the regular value theorem. By skimming through an atlas of X, we can show that any $\tilde{N} \cap U_i$ is a smooth submanifold of M where the $(U_i)_i$ are a countable cover of X. We can conclude that $\tilde{N}$ is a smooth submanifold of M with dimension n. Furthermore, if M has no boundaries, then $\tilde{N}$ has no boundaries either. Finally, $\tilde{N}$ is the countable union of preimages of a closed subset by smooth functions and is therefore closed. □

We refer the reader to **[Manton, 2012]** for a proof of this theorem based on locally smooth vector-fields forming a basis of tangent spaces.

Remark 2.1.2

The fact that $g\colon \theta \mapsto x_\theta{}^*$ is only defined locally is a direct consequence of the fact that f_θ might have several critical points. If for each θ, f_θ had a unique critical point which is a global minimum, then the function g could be globally defined and would send θ to the corresponding global minimum of f_θ. Notice that the smoothness of N proved in the previous theorem implies the smoothness of $g\colon \Theta \to X$ (the local parametrization of $\tilde{N}$). Indeed, one of the definitions of the smoothness of a manifold is that it has a smooth parametrization.

Remark 2.1.3

We treated the case with manifolds X, Θ compact but without boundaries. The previous proof in case of boundaries will be very similar with $\tilde{N}$ inheriting boundaries from M so that the boundary of $\tilde{N}$ is a submanifold in the boundarie of M. Then, in all the development that follows, boundaries should be treated separetely. We will not develop the subject of boundaries in this document, but we believe it can be dealt without major difficulties.

In the following proposition, the shape of $\tilde{N}$ is further specified. It sits in $\mathcal{M}$ over Θ, and can be locally parametrized by $\theta \in \Theta$.

Proposition 32. *The set $\tilde{N}$ locally defines a smooth section over Θ. For p_0 in $\tilde{N}$, there is a neighborhood W_0 of p_0 in $\tilde{N}$ such that the map*

$$\begin{aligned} \pi_0\colon \tilde{N} \cap W_0 &\to \Theta \\ (x, \theta) &\mapsto \theta \end{aligned}$$

is a smooth diffeomorphism.

Proof. The idea is to show that locally, $\tilde{N}$ can be parametrized by Θ. To show that, we will show that a basis of the tangent space of $\tilde{N}$ at a point can be built from a basis of the tangent space of Θ. We work locally around $p_0 = (x_0, \theta_0) \in \tilde{N}$. Let $((U_{x_0}, \varphi) \times (V_{\theta_0}, \zeta))$ be a convex coordinate patch in $\tilde{N}$ such that $(x_0, \theta_0) \in U_{x_0} \times V_{x_0}$. We write $\zeta = (\theta_1, \dots, \theta_n)$ the coordinates on V_{θ_0}. Note that for all $\theta \in V_{\theta_0}$, a basis for the tangent space $T_\theta\Theta$ is $(\partial\theta_1, \dots, \partial\theta_n)$. The function g defined by:

$$g \colon (x, \theta) \mapsto \nabla f_\theta(x) \tag{5.5}$$

is well-defined on (U_{x_0}, θ_{x_0}). The level set of 0 by g is exactly

$$\mathcal{G}_0 = g^{-1}(0) = \tilde{N} \cap (U_{x_0} \times V_{x_0}) \subset \tilde{N}$$

In particular, the tangent space at $p \in \mathcal{G}_0$ is equal to the tangent space at $p \in \tilde{N}$. The regular value theorem applied to g in the regular value 0 tells us the tangent space to $\mathcal{G}_0$ at (x_0, θ_0) is the kernel of $Dg(x_0, \theta_0)$. the Jacobian of g relatively to the coordinate charts (φ, ζ) writes:

$$J_g(x_0, \theta_0) = \Big(\underbrace{\operatorname{Hess} f_{\theta_0}(x_0)}_{k \times k} \quad \underbrace{D_\theta D_x f(x_0, \theta_0)}_{k \times n} \Big) \tag{5.6}$$

The function f_{θ_0} is Morse, and x_0 is one of its critical points which implies that Hessian $\operatorname{Hess} f_{\theta_0}(x_0)$ is invertible, and its kernel is reduced to 0. Hence, tangent vectors of $\tilde{N}$ at (x_0, θ_0) can be written $[0, \dots, 0, \lambda_1\partial\theta_1, \dots, \lambda_n partial\theta_n]$ for some coordinates $(\lambda_1, \dots, \lambda_n)$. The dimension of the tangent space is exactly n, hence a basis for the tangent space at (x, θ) is $(0, \partial\theta_1, \dots, (0, \partial\theta_n))$. To conclude, we use the fact that the map that sends a coordinate chart on a basis of tangent vectors is locally invertible, (see for instance normal coordinates **[Datchev, 2021]**). The basis identified for the tangent space of $\tilde{N}$ at p_0 implies there exists a neigborhood W_0 of p_0 of $\tilde{N}$ where

$$\psi \colon W_0 \to \mathbb{R}^k \times \mathbb{R}^n (x, \theta) \mapsto (0_{\mathbb{R}^k}, \theta_1(\theta), \dots, \theta_n(\theta)) \tag{5.7}$$

is a coordinate chart. The striking feature of (5.7) is that the coordinates of the point (x, θ) do not depend on x. Besides, ψ is a local diffeomorphism, hence injective. This shows that locally, the data of θ identifies a unique x. In particular, W_0 only intersects one connected component of $\tilde{N}$. Noting $\pi_{\mathbb{R}^n}$ the canonical surjection of $\mathbb{R}^k \times \mathbb{R}^n$ over $\mathbb{R}^n$ the map

$$\begin{aligned} \pi_0 = \zeta^{-1} \circ \pi_{\mathbb{R}^n} \circ \psi \colon W_0 \cap \tilde{N} &\mapsto \Theta \\ (x, \theta) &\mapsto \theta \end{aligned}$$

is a diffeomorphism. □

Combining theorem 12 and proposition 32 shows that the shape of the submanifold of fibre-wise critical points of a fibre-wise Morse function is deeply connected and constrained by the topology of M itself. Indeed, $\tilde{N}$ is closed and has no boundaries, and because it is locally parametrizable by θ, it can have no turning point with regard to Θ. An illustrative way to see it is that the submanifold $\tilde{N}$ sits over the base space Θ in M, like a topological copy of Θ. Hence, there can be no boundaries or limit points in θ. A direct consequence is the following corollary.

Corrolary 1. *Let $f : X \times \Theta \to \mathbb{R}$ be a fibre-wise Morse function as in definition 37. Then, there exists an integer $K \geqslant 2$ such that for every $\theta \in \Theta$, the function f_θ has exactly K critical points.*

Proof. Let $\theta \in \Theta$ and K be the number of critical points of f_θ. Then, because f_θ is Morse, it is not constant and has at least a global minima and a global maxima. Hence, $K \geqslant 2$. Furthermore, we have already mentioned that the compactness of X implies that a Morse function on X has a finite number of critical point. Then $K < +\infty$. Clearly, K is the number of intersections of $\tilde{N}$ with the fibre $\{(x, \theta) \in X\}$. Remember that $\tilde{N}$ is locally a smooth section over Θ (directly from proposition 32). For each $x^* \in \tilde{N} \cap \{(x, \theta)\ ,\ x \in X\}$ there is an open neighborhood V_{x^*} of (x^*, θ) diffeomorphic to an open neighborhood in Θ. Let $r > 0$ be such that

$$\forall x^* \in \tilde{N} \cap \{(x, \theta), x \in X\}\, , \mathcal{B}((x^*, \theta), r) \subseteq V_{x^*}$$

We choose $r > 0$ such that the ball of radius r centered on (x^*, θ) is included for each x^* in the region where $\tilde{N}$ is diffeomorphic to an open neighborhood of Θ. Then for any θ' at a distance less that r of θ, the equality:

$$\text{card}\left(\tilde{N} \cap \{(x, \theta)\ ,\ x \in X\}\right) = \text{card}\left(\tilde{N} \cap \left\{(x, \theta')\ ,\ x \in X\right\}\right) = K$$

is verified. We showed that for any θ, there is a neighborhood of θ where all the neighbors yield the same number of fibre-wise critical points of f. This being valid anywhere on the set Θ, a proof by contradiction shows the result. □

A Torus example Consider $M = S^1 \times S^1$ the two-dimensional torus parametrized by (x, θ), where x describes a rotation along horizontal circles (see figure 5.2) and θ is a rotation along vertical axis. Let f be a fibre-wise Morse function over the torus. The set of fibre-wise critical points of f is a one-dimensional closed differentiable submanifold of M with no boundaries. Hence, it is made of one or several loops (closeness condition) winding their way around the torus and never intersecting each other (submanifold condition). Furthermore, It is locally parametrizable by the vertical circles θ (vertical circle in figure 5.2) which means in particular it can have no turning point with respect to this coordinate. This condition excludes for instance a horizontal circle around the Torus. If all the functions of the family f_θ are the same however, then $\tilde{N}$ is a vertical circle. The only way to cross the same angle θ twice is by turning around the torus, with the angle θ only increasing or only decreasing modulo 2π.

Then, $\tilde{N}$ can only be made of one or several circles winding their ways around the torus the same number of times.

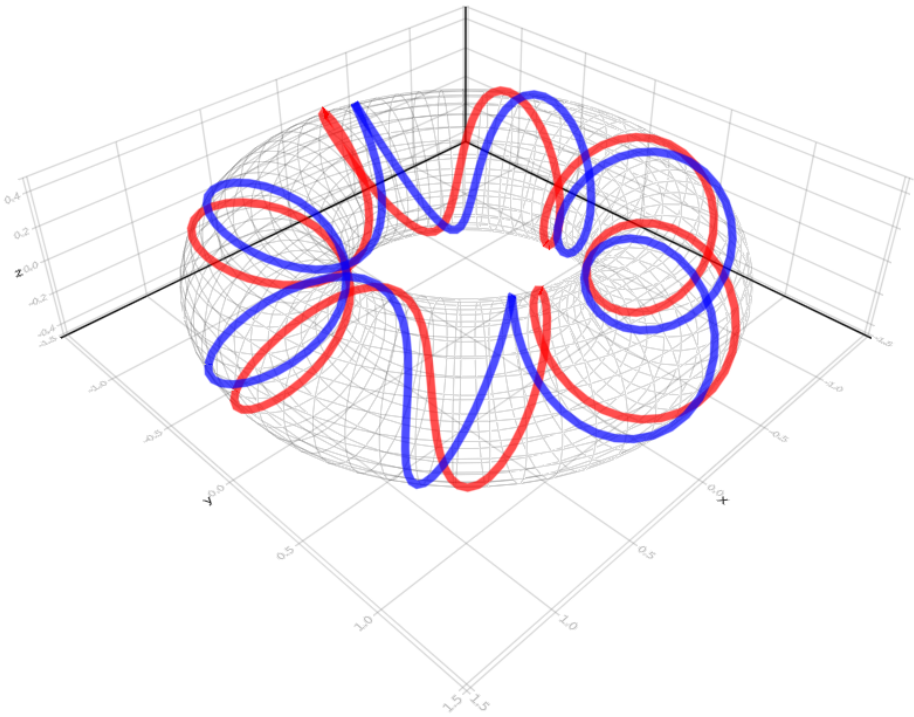


Fig. 5.2. – A possible shape for the fibre-wise critical points of a fibre-wise Morse function on the torus. The submanifold $\tilde{N}$ is represented by the red and blue lines and is made of two connected components winding their way seven times around the Torus. The number K of critical points in this example is 14, it is the number of intersection of $\tilde{N}$ with any horizontal circle at θ constant.

The following corollary shows that if $\tilde{N}$ is made of finitely many connected components, then just some of them contain local minima. It ensures in particular that the index of a critical point along a connected component is constant.

Corrolary 2. *There is a submanifold $N \subset \tilde{N}$ made of finitely many connected components and containing only fibre-wise local minima.*

Proof. Given two points $(\theta_0, x_0), (\theta_1, x_1)$ on the same connected component, there exists a smooth path γ that joins them. The Hessian of f is a smooth function from $X \times \Theta$ to $T_X \otimes T\Theta$. Let π_{T_X} be the canonical surjection of $T_X \otimes T\Theta$ over T_X.

$$\begin{aligned} \pi_{T_X} \circ \operatorname{Hess} f \colon X \times \Theta &\to T_X \\ (x, \theta) &\mapsto \operatorname{Hess} f_\theta(x) \end{aligned}$$

The index of a critical point x^* in X is given by the sign of the eigenvalues of $\operatorname{Hess} f_\theta$. The projection π is smooth so that the eigenvalues of $\operatorname{Hess} f_\theta$ can not change sign without passing

by 0. As $\mathrm{Hess}\, f_\theta$ is never critical by hypothesis, eigenvalues of $\mathrm{Hess}\, f_\theta$ can not change sign when θ ranges γ. Hence, x_0 and x_1 have the same index. □

Assume a real time context and at instant $t = t_1$, a vector of data θ_1 is observed. The optimization problem is on f_{θ_1} while we have precomputed informations on f_{θ_0} where θ_0 is another vector of data, and we have no a priori that θ_1 and θ_0 are close. If we build a path between θ_0 and θ_1, can we deduce a path between a local minimum of f_{θ_0} and a local minimum of f_{θ_1} so that every element on the path is a local minimum for some f_θ ? The next corollary shows that the answer is yes, and even proves that there exist a smooth path realizing this condition.

Corrolary 3. *Let θ_0 and θ_1 be two elements from the observation set Θ. Let $\theta\colon [0,1] \to \Theta$ be a smooth diffeomorphism joining θ_0 and θ_1,* i.e. *$\boldsymbol{\theta}(0) = \theta_0$ and $\boldsymbol{\theta}(1) = \theta_1$. Let x_0 be a local minimum of f_{θ_0}. and let $N_0 \subset N$ be the connected component containing (θ_0, x_0). Then there is a unique continuous path $x^*\colon [0,1] \to X$ satisfying $(\boldsymbol{\theta}(t), x^*(t)) \in N_0, \forall t \in [0,1]$. The endpoint $x^*(1) = x_1$ is a local minimum of f_{θ_1}. Furthermore, x^* is continuously differentiable.*

Proof. Because $\theta\colon [0,1] \to \Theta$ is a diffeomorphism, there is a chart (V, ζ) on Θ such that $\boldsymbol{\theta}([0,1]) \subseteq V$. Hence, from proposition proposition 32, there is an open set invertible section $\pi_0\colon X \times \boldsymbol{\theta}([0,1]) \cap N_0 \to \Theta$. We can set $\pi_0^{-1}(\theta_0) = x^*(0)$ and set $(x^*(t), \boldsymbol{\theta}(t)) = \pi^{-1}(\boldsymbol{\theta}(t))$. This uniquely defines x^*. Again from proposition 32, the inverse section π^{-1} is continuously differentiable which induces the same property for x^*. Graphically, $x^*(t)$ is the vertical lift of $\boldsymbol{\theta}([0,1])$ in N_0 starting at x_0. This is essentially a consequence of the implicit function theorem, but our function being not defined between vector spaces, we found easier to use work with coordinates. □

figure 5.3 shows an example of how a path in θ induces a lift in N. Note that the lift is not unique as long as x_0 is not defined, and this is true even if N has only one connected component. Indeed, in the Torus example where N winds its ways several times around the Torus, a connected component can intersect several points of the shape (θ, x_i) where θ is fixed.

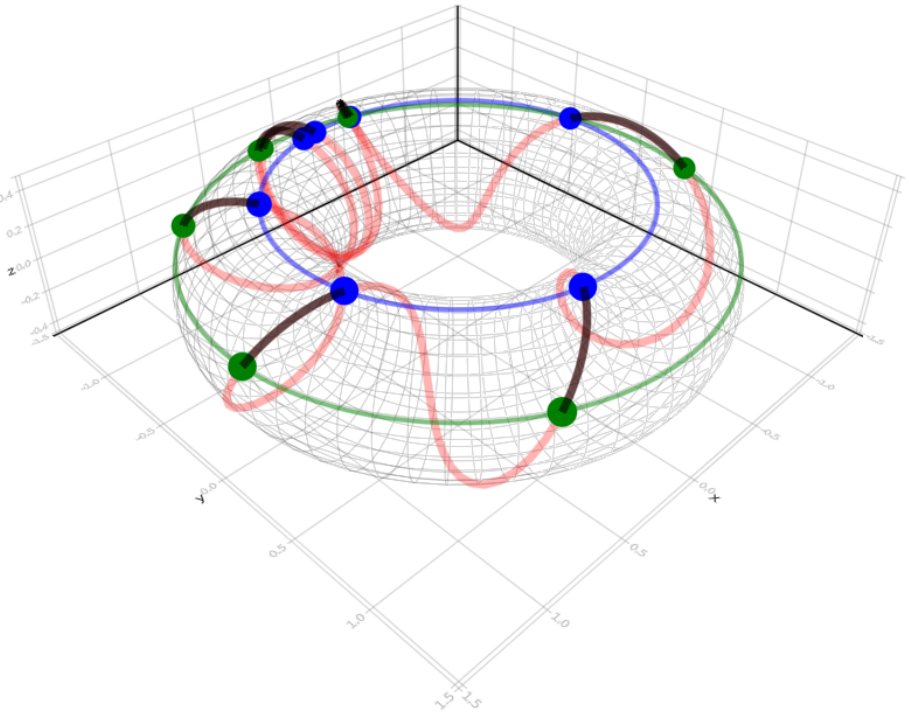


Fig. 5.3. – Illustration of corrolary 3 on the torus where an admissible submanifold of fibre-wise critical points is represented in red. The torus is parameterized by (x, θ) where x is the angle along the horizontal circles and θ the angle along the vertical circles. The green circle is the set of points with a fixed $\theta = \theta_0$ and the blue circle has a fixed $\theta = \theta_1$. Green dots are fibre-wise critical points of f_{θ_0} while blue dots are fibre-wise critical points of f_{θ_1}. The dark lines are the lifts in N of a curve $\boldsymbol{\theta}(t)$ that goes from θ_0 to θ_1. Note that the lift generates as many possible $x^*(t)$ as the number of local minima of f_{θ_0}. However, fixing a starting point (one of the green dots) uniquely determines a section of dark line, i.e a function $x^*(t)$.

5.3 A Homotopy method for optimization geometry

Previous section has shown the existence of a smooth path made of local minima, joining the two points of interest in our problem. A homotopy method can take advantage of this path as long as it is a natural attractor for some recursive algorithm. We will show how this is the case for the path x^* defined in corrolary 3. Our work differs from previous work on homotopy methods, a summary of which can be found in **[Allgower and Georg, 2012]**, which usually focus on ordinary differential equations or level sets of 0 by the homotopy map, the latter approach was briefly presented in section 4.3. Being on the tangent bundle TX, we consider the level set of the submanifold made of the zeros of the different tangent spaces in TX. Let $f : X \times \Theta \to \mathbb{R}$ be a fibre-wise Morse function as in definition 37. Let θ_0, $\theta_1 \in \Theta$ and let $x_0 \in X$ be a local minimum of f_{θ_0}. We aim at finding a local minimum x_1 of f_{θ_1} by exploiting the known local minimum x_0 of f_{θ_0}. To do so, we first need to define a diffeomorphic curve $\boldsymbol{\theta} : [0,1] \to \Theta$ such that $\boldsymbol{\theta}(0) = \theta_0$ and $\boldsymbol{\theta}(1) = \theta_1$. For example, if Θ is a metric manifold, one can choose for $\boldsymbol{\theta}$ the geodesic joining θ_0 and θ_1. From corrolary 3, we know that there exists a curve $x^* : [0,1] \to X$ of local minima of the functions $f_{\boldsymbol{\theta}(t)}$. Thus, starting from x_0, we propose to follow the curve $x^* : [0,1] \to X$ corresponding to the chosen curve $\theta : [0,1] \to \Theta$ in order to find a local minimum $x^*(1) = x_1$ of f_{θ_1}.

The curve $x^* : [0,1] \to X$ is implicitly defined, hence one cannot expect to obtain it in close form in general and an iterative method is needed to estimate it. To construct $x^* : [0,1] \to X$, we exploit its graph, which is defined as

$$G(x^*) = \{(t, x^*(t)) : t \in [0,1]\} \subset [0,1] \times X.$$

To do so, we rely on the characterisation of $G(x^*)$ provided in proposition 33. It showed that locally, the curve x^* exactly contains all and only the antecedent of points of the shape $0_x \in TX$ for H. The function H thus introduced is the homotopy map we will use in this section.

Proposition 33. *Let the mapping*

$$\begin{array}{rccl} H: & [0,1] \times X & \to & TX \\ & (t,x) & \mapsto & \nabla f_{\boldsymbol{\theta}(t)}(x) \in T_x X. \end{array}$$

Then, every point $(t, x^*(t))$ satisfies the relation

$$H(t, x) = 0_x$$

Furthermore, for every point $(t, x^*(t))$, there is a neighbourhood V around this point such that $\{(t,x) \in [0,1] \times X : H(t,x) = 0_x\} \cap V$ contains only points in the graph of x^*, i.e, only points of the shape $(t, x^*(t))$ for some $t \in [0,1]$.

Proof. Let $t \in [0,1]$ and remember that by definition, $(\boldsymbol{\theta}(t), x^*(t)) \in N$. Then $H(t, x^*(t)) = \nabla f_{\boldsymbol{\theta}(t)}(x^*(t)) = 0_x$ and $(t, x^*(t))$ is indeed a zero of H. Let N_j be the connected component in N containing $((\boldsymbol{\theta}(t), x^*(t)))$. Let d be a continuous distance on the manifold X. Because critical points of a Morse function are isolated, there is a strictly positive constant ε which is the minimal d-distance between two critical points of f_θ in X for θ varying in Θ. Take the ball in M made of the cross-product of an open ball of radius $\varepsilon/2$ centered in $x^*(t)$ and an open ball of any size centered in $\boldsymbol{\theta}(t)$. Then, $f_{\boldsymbol{\theta}(t)}$ has only one critical point in $\mathcal{B}\left(x^*(t), \frac{\varepsilon}{2}\right)$. Take the open set $U \subset M$ made of the cross-product of $\mathcal{B}\left(x^*(t), \frac{\varepsilon}{2}\right)$ with an open set W around $\boldsymbol{\theta}(t)$. The intersection $N \cap U = N_j \cap U$ and obviously for every $\theta \in W$, there is a unique $x \in X$ such that (θ, x) belongs to U. Then, by construction of x^*, if $\theta \in \boldsymbol{\theta}([0,1])$, the unique x matching it in U is $x^*(\theta^{-1}(\theta))$. Let $V = \theta^{-1}((\boldsymbol{\theta}([0,1]) \cap W) \cap \mathcal{B}\left(x^*(t), \frac{\varepsilon}{2}\right)$. V is open and contains $(t, x^*(t))$. Furthermore, every point in V can be written $(s, x^*(s))$ for some $s \in [0,1]$. □

Remark 3.0.1

The constant ε graphically represent the smallest distance between two intersections of N with the fibre $\{(x, \theta) \mid x \in X\}$.

An equation for the curve $x^* : [0,1] \to X$ is not known in closed form, so that in order to be able to follow this curve we need its tangent. Given $(t, x^*(t)) \in G(x^*)$, we are able to define the tangent space $T_{(t,x^*(t)}G(x^*)$, which is a one-dimensional subspace of $\mathbb{R} \times T_{x^*(t)}X$. To do so, we need to differentiate the function H defined in proposition 33. Differentiating H with respect to $t \in [0,1]$ in direction $dt \in \mathbb{R}$ yields $\frac{\partial}{\partial t} \nabla f_{\boldsymbol{\theta}(t)}(x) dt$, while differentiating it with respect to $x \in X$ in direction $dx \in T_x X$ yields $\operatorname{Hess} f_{\boldsymbol{\theta}(t)}(x) dx$. This writes,

$$DH(t,x)(dt, dx) = \frac{\partial}{\partial t} \nabla f_{\boldsymbol{\theta}(t)}(x) dt + \operatorname{Hess} f_{\boldsymbol{\theta}(t)}(x) dx.$$

It is readily checked that, as expected, $DH(t,x)$ is a mapping from $\mathbb{R} \times T_x X$ onto $T_x X$. lemma 1 provides a closed form expression of the tangent space $T_{(t,x^*(t))}G(x^*)$ of $G(x^*)$ at $(t, x^*(t))$ and an illustration is proposed in figure 5.4.

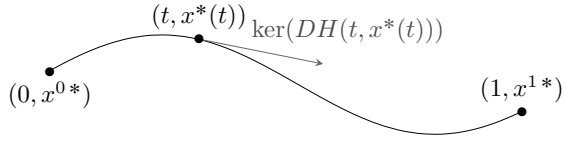


Fig. 5.4. – Schematic illustration of the graph $G(x^*)$ of the curve $x^* : [0,1] \to X$ of local minima corresponding to the curve $\theta : [0,1] \to \Theta$. At $(t, x^*(t)) \in G(x^*)$, the tangent space of the graph, which is a one-dimensional subspace of $\mathbb{R} \times T_{x^*(t)}X$, is given by $\ker(DH(t, x^*(t)))$.

Lemma 1. *The tangent space $T_{(t,x^*(t))}G(x^*)$ of the graph $G(x^*)$ at $(t, x^*(t))$ is*

$$T_{(t,x^*(t))}G(x^*) = \ker(DH(t, x^*(t))).$$

Furthermore, the kernel of the mapping $DH(t, x^*(t))$ is

$$\ker(DH(t, x^*(t))) = \{\lambda(1, \xi) : \lambda \in \mathbb{R}\},$$

where $\xi \in T_{x^*(t)}X$ is the unique solution to

$$\text{Hess}\, f_{\boldsymbol{\theta}(t)}(x^*(t))\xi = -\frac{\partial}{\partial t} \nabla f_{\boldsymbol{\theta}(t)}(x^*(t)). \tag{5.8}$$

Proof. From proposition 33, the graph $G(x^*)$ is the connected component of $\{(t, x) \in [0, 1] \times X : H(t, x) = 0_x\}$ that contains $(0, x_0)$. One can check that the tangent space of $\{(t, x) \in [0, 1] \times X : H(t, x) = 0_x\}$ at (t, x) is

$$\{(dt, dx) \in \mathbb{R} \times T_xX : DH(t, x)(dt, dx) = 0_x\},$$

which corresponds to the kernel of $DH(t, x)$. Thus, we obtain $T_{(t,x^*(t))}G(x^*) = \ker(DH(t, x^*(t)))$. The space $\ker(DH(t, x^*(t)))$ is a one-dimensional subspace of $\mathbb{R} \times T_{x^*(t)}X$. Therefore, it can be written as $\{\lambda(1, \xi) : \lambda \in \mathbb{R}\}$, where $\xi \in T_{x^*(t)}X$. To find ξ, we need to solve

$$DH(t, x^*(t))(1, \xi) = 0_x,$$

which yields the equation given above. The solution to this equation exists and is unique because, for all $t \in [0, 1]$, the Hessian of $f_{\boldsymbol{\theta}(t)}$ is positive definite at $x^*(t)$. Indeed, this property follows from the fact that f is fibre-wise Morse on M. □

These results are sufficient to allow the developpment of an iterative algorithm that estimates the curve $x^* : [0, 1] \to X$ of local minima. Given a predetermined sequence $\{t_k\}_{0 \leqslant k \leqslant K}$ arranged in increasing order such that $t_0 = 0$ and $t_K = 1$, it returns the sequence $\{x_k^*\}$ of local minima of the functions f_{θ_k}, where $\theta_k = \boldsymbol{\theta}(t_k)$. The proposed method is described in algorithm 2 and a schematic illustration is provided in figure 5.5. Every iteration can be decomposed into two steps. The first one is the prediction step (lines 2-4 in algorithm 2), which consists in following the direction dx_k^* in $T_{x_k^*}X$ provided by the tangent space of $G(x^*)$ at (t_k, x_k^*). It yields $y_k \in X$, which is obtained by taking the exponential mapping defined in section 4.1.5 of dx_k^* at x_k^*. The second one is the correction step (line 5 in algorithm 2), where x_{k+1}^* is obtained by projecting y_k on the curve $x^* : [0, 1] \to X$. This is achieved by minimising $f_{\theta_{k+1}}$ with the Newton method initialised at y_k.

Algorithm 2 optimization geometry algorithm

Require: $\{t_k\}_{0\leqslant k\leqslant K}$ arranged in increasing order, with $t_0 = 0$ and $t_K = 1$; $\{\theta_k\}_{0\leqslant k\leqslant K}$ such that $\theta_k = \boldsymbol{\theta}(t_k)$; local minimum x_0 corresponding to θ_0.
Ensure: Sequence $\{x_k^*\}_{0\leqslant k\leqslant K}$ of local minima corresponding to each θ_k.
1: **for** $k = 0$ **to** K **do**
2: Solve $\mathrm{Hess}\, f_{\theta_k}(x_k^*)\xi_k = -\frac{\partial}{\partial t}\nabla f_{\theta_k}(x_k^*)$ for ξ_k.
3: Compute $dx_k^* = dt_k\xi_k$, where $dt_k = t_{k+1} - t_k$.
4: Compute $y_k = \exp^X_{x_k^*}(dx_k^*)$.
5: Compute x_{k+1}^* by solving $\mathrm{argmin}_{x\in X} f_{\theta_{k+1}}(x)$ with the Newton method initialised at y_k.
6: **end for**

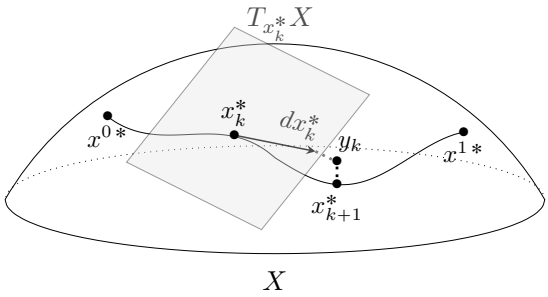


Fig. 5.5. – Schematic illustration of the procedure of algorithm 2. Given iterate $x_k^* \in X$, a prediction step is achieved by computing $y_k \in X$ through the exponential of $dx_k^* \in T_{x_k^*}X$, which is provided by the tangent space of the graph $G(x^*)$ at (t_k, x_k^*). A correction step is then performed in order to obtain the next iterate x_{k+1}^* by projecting the predicted point y_k onto $x^* : [0,1] \to X$.

5.3.1 Convergence analysis

In this part we give the sketch of a proof for the convergence of the homotopy method provided in algorithm 2. A more rigorous and complete proof is left for future work, as well as a discussion about the rate of convergence. Given a fibre-wise Morse function $f : X \times \Theta \to \mathbb{R}$, a fibre-wise local minimum (θ_0, x_0) of f and a path $\theta : [0,1] \to \Theta$ satisfying conditions of corrolary 3 such that $\boldsymbol{\theta}(0) = \theta_0$ and $\boldsymbol{\theta}(1) = \theta_1$, the aim is to show that there exists an integer K and a sequence $\{t_k\}_{0\leqslant k\leqslant K}$ such that algorithm 2 converges to the local minimum x_1.

The idea is to show that there exists $\delta_0 > 0$ such that, for all k, we can choose $dt_k = t_{k+1} - t_k \geqslant \delta_0$ allowing to predict a point $y_k = \exp^X_{x_k^*}(dt_k\xi_k) \in X$ sufficiently close to the local minimum x_{k+1}^* of $f_{\theta_{k+1}}$. By sufficiently close, we mean that there exists $\varepsilon_0 > 0$ such that $d(y_k, x_{k+1}^*) < \varepsilon_0$, where d is the distance on X, and that the Newton method minimising $f_{\theta_{k+1}}$ converges to x_{k+1}^* for any starting point y satisfying $d(y, x_{k+1}^*) < \varepsilon_0$. It follows that for algorithm 2 to return x_1, a sequence $\{t_k\}$ of at most $\left\lfloor \frac{1}{\delta_0} \right\rfloor + 1$ elements is required, where $\lfloor\cdot\rfloor$ is the floor function. To show this convergence result, the following points are to be proven:

1. The convergence to a point on the curve $x^* : [0,1] \to X$ of the Newton method on line 5 of algorithm 2 needs to be guaranteed when initialised with the predicted point y_k. We can show that there exists $\varepsilon_0 > 0$ such that the Newton method converges to

x^*_{k+1} if the initial point y satisfies $d(y, x^*_{k+1}) < \varepsilon_0$. As ε_0 must not depend on k, this step basically requires f to admit basins of attractions of constant size across X and Θ.

2. As $y_k = \exp^X_{x_k}(dt_k \xi_k)$, the only parameter we can pilot to ensure $d(y_k, x^*_{k+1}) < \varepsilon_0$ is the step-size dt_k. We need to show that $d(y_k, x^*_{k+1})$ can be bounded by an expression involving only constants and dt_k.

3. The latter expression is then used to find a step-size dt_k that puts y_k in the convergence radius ε_0 of x^*_{k+1}.

4. Finally, we need to check that the sequence $\{t_k\}$ does not converges to $l < 1$, *i.e.* that x_1 can be reached in a finite number of steps. Hence, it is needed to prove that there exists $\delta_0 > 0$ such that, for all k, we have $d(y_k, x^*_{k+1}) < \varepsilon_0$ for $dt_k \geqslant \delta_0$.

The remaining of this section details how the four previous points can be proven.

Point 1: By definition of fibre-wise Morse functions, f_θ is locally convex around each critical point x^*. Thus, the Newton method has a basin of convergence around each critical point x^*. Let ε_k be the radius of the basin around x^*_k. From the Morse condition, $\varepsilon_k > 0$. It is then needed to show that the infimum ε_0 of ε_k over all possible x^*_k is strictly greater than zero. This can be achieved by (*i*) proving that ε_k is greater than a strictly positive continuous function in $X \times \Theta$ and (*ii*) using the fact that a continuous function on a compact set always reaches its bounds. Thus, we have $0 < \varepsilon_0 < \varepsilon_k$ for all possible ε_k. The minoring function in (*i*) can be built by exploiting implicit functions sending θ over a zero of the eigenvalue $\lambda_i(\mathrm{Hess}\, f_{\boldsymbol{\theta}(x)})$.

Point 2: Bounding the distance $d(y_k, x_{k+1})$ can be done in two steps. First, we bound the distance $d(x^*_k, x^*_{k+1})$ between points on $x^* : [0,1] \to X$ and the distance $d(x^*_k, y_k)$. Second, the triangular inequality is used to obtain the wished bound. The curve $x^* : [0,1] \to X$ admits a Lipschitz constant L with respect to the distance d on X. It comes from the fact that Dx^* is continuous and therefore bounded on $[0,1]$. We have:

$$d(x^*(t_{k+1}), x^*(t_k)) \leqslant dt_k L. \tag{5.9}$$

Moreover, as $y_k = \exp^X_{x^*_k}(dx^*_k)$, we have $d(x^*_k, y_k) = \|dx^*_k\|_x = dt_k \|\xi_k\|_x$. Notice that $\xi_k = \nabla x^*(x_k) \in T_x X$, thus

$$d(x^*_k, y_k) = dt_k \| \nabla x^*(x_k)\|_{x_k} \tag{5.10}$$

The operator norm on $\mathcal{L}(T_{x_k} X)$ and the vector norm on $T_{x_k} X$ both derive from the metric and are compatible. It follows that $\| \nabla x^*(x_k)\|_{x_k} \leqslant L$. Combining (5.10) and (5.9) hence gives

$$d(y_k, x_{k+1}) \leqslant 2 dt_k L. \tag{5.11}$$

Point 3: With Equation (5.11), we can set $dt_k = \dfrac{\varepsilon_0}{2L}$ in order to get $(y_k, x^*_{k+1}) \leqslant \varepsilon_0$

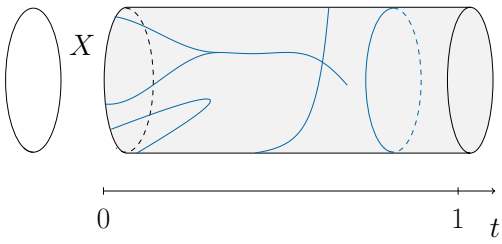


Fig. 5.6. – The domain X is a circle (represented on the left) and the t axis is at the bottom. Every section of the cylinder represents the space X. For a fixed t, every point on a blue line above t is a critical point of the function $f_{\theta(t)}$. In this example from left to right: components of $\tilde{N}$ merge, have a turning point with respect to t, cross each other,have a parametrization whose derivative goes to zero and do not admit a parametrization with respect to t (the blue circle). All this situation are prevented if f_θ is fibre-wise Morse.

Point 4: In point 2, we could have bounded Dx^* only locally, which would have given a bigger (hence better) step-size in point 3. However, taking the supremum of the derivative on $[0,1]$ brings that the step-size $dt_k = \frac{\varepsilon_0}{2L}$ is independent of k. Hence, we can simply choose $\delta_0 = \frac{\varepsilon_0}{2L}$.

Non-convergence On figure 5.6 we illustrate some conditions on $\tilde{N}$ under which the homotopy algorithm does not converge. All the cases that happen in the figure are prevented by the fibre-wise Morse condition on f.

5.4 Illustration

5.4.1 A trivial example of the optimization geometry algorithm

In this section, an illustration of the proposed optimization geometry method is provided. Even though this example is trivial, it illustrates the interest of the method in practice. Let $X = [0,1]$, $\Theta = [0,1]$ and

$$\begin{array}{rccc} f: & X \times \Theta & \to & \mathbb{R} \\ & (x,\theta) & \mapsto & \left(x - \theta^2\right)^2. \end{array} \tag{5.12}$$

Given $\theta \in \Theta$, the global minimum x^* of $f_\theta : x \mapsto f(x,\theta)$ is simply $x^* = \theta^2$. For the sake of the example, we will however apply the optimization geometry method to obtain it.

First, we need to verify that f is fibre-wise Morse on $M = [0,1] \times [0,1]$. As it is polynomial in x and θ, it is smooth on M. Furthermore, for all $\theta \in \Theta$,

$$\frac{d^2 f_\theta}{dx^2} = 2.$$

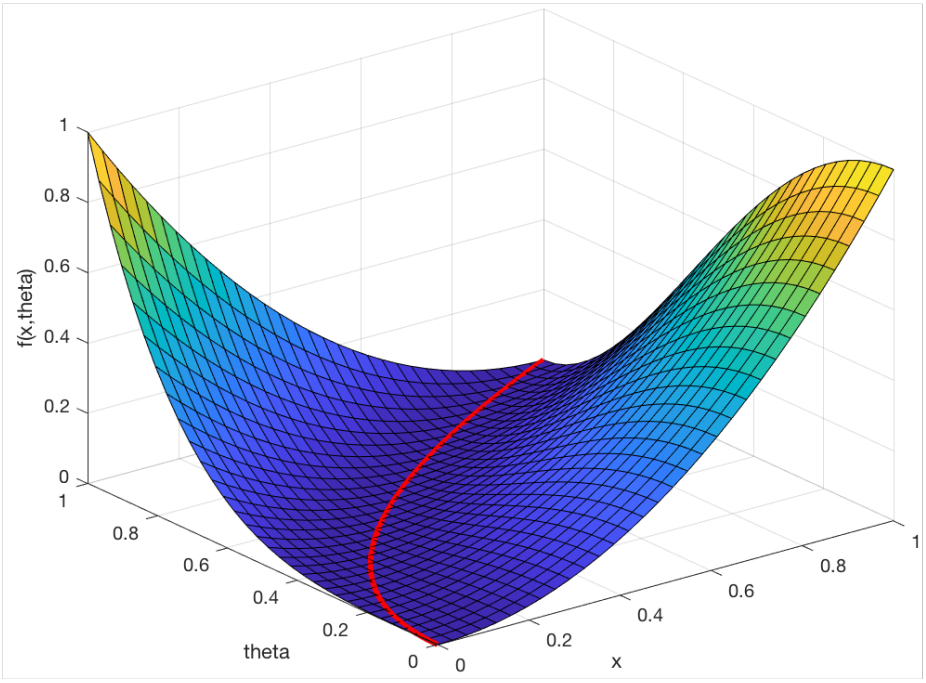


Fig. 5.7. – Graph of the function $f : X \times \Theta \to \mathbb{R}$ defined in (5.12) with $X = [0,1]$ and $\Theta = [0,1]$. The red curve corresponds to the submanifold of fibre-wise critical points of f embedded **[Guillemin and Pollack, 2010]** in the graph of f. Given θ_0 for which the minimum x_0 is known, the optimization geometry method aims at following this curve in order to find the solution x_1 of the optimization problem for θ_1.

Hence, for all $x \in X$, the second order derivative of f_θ is positive definite. It is in particular true at the critical point of f_θ. It follows that f is indeed fibre-wise Morse on M and our optimization geometry method can be employed.

Let $\theta_0 \in \Theta$, for which the global minimum of f_{θ_0} is $x_0 = \theta^{0\,2}$, and $\theta_1 \in \Theta$, for which the corresponding minimum x_1 is assumed to be unknown. Let the curve

$$\begin{array}{rccl} \theta: & [0,1] & \to & \Theta \\ & t & \mapsto & t\theta_1 + (1-t)\theta_0. \end{array}$$

The first order derivative of $f_{\boldsymbol{\theta}(t)}$ is

$$df = 2\left(x - \boldsymbol{\theta}(t)^2\right) dx.$$

It follows that the function $H : [0,1] \times X \to T_x X \simeq \mathbb{R}$ defined in proposition 33 is

$$H(t,x) = 2\left(x - \boldsymbol{\theta}(t)^2\right).$$

Its first derivative is

$$DH(t,x)(dt,dx) = 2dx - 4\boldsymbol{\theta}(t)\dot{\theta}(t)dt,$$

where $\dot{\theta}(t) = \theta_1 - \theta_0$. Thus, solving equation in lemma 1, one can check that $\ker(DH(t,x)) = \{\lambda(1,\xi) : \lambda \in \mathbb{R}\}$, where

$$\xi = 2\boldsymbol{\theta}(t)\dot{\theta}(t) = 2(\theta_1 - \theta_0)(\theta_0 + (\theta_1 - \theta_0)t).$$

Let $\theta_0 = 0$ ($x_0 = 0$), $\theta_1 = 1$ and $\{t_k\} = \{0, 0.2, \ldots, 0.8, 1\}$ (*i.e.*, $dt_k = 0.2$). Within these settings, $\theta_k = \boldsymbol{\theta}(t_k) = t_k$ for all k. algorithm 2 proceeds as follows at the k^{th} iteration:

- the direction ξ_k is given by $\xi_k = 2t_k$;
- the resulting predicted point is $y_k = x_k^* + 2t_k dt_k$;
- the corrected point $x_{k+1}^* = \theta_{k+1}^2$ is obtained with one iteration of the Newton method[2].

An illustration of the procedure can be found in section 5.4.1.

[2] Note that the important point here is not the Newton method but the homotopy procedure.

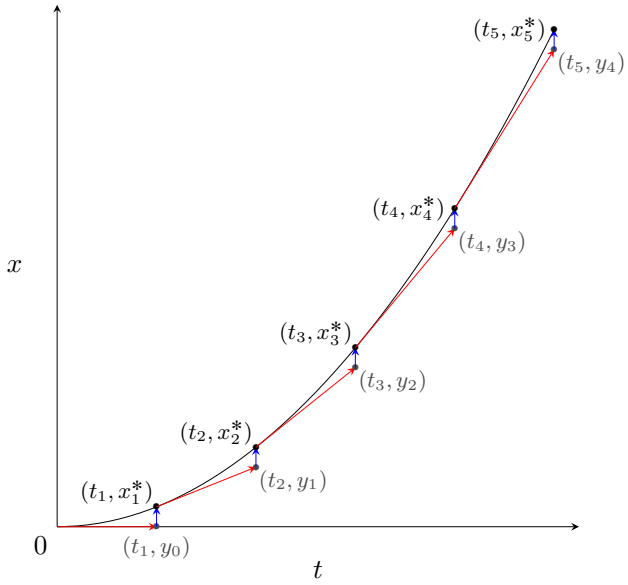


Fig. 5.8. – On this figure, all the steps of the homotopy algorithm are represented. The black curve is the graph of x^*, which usually has no closed form. The red arrow show the prediction step, in the direction of the tangent to the black curve. The blue arrow shows the projection back to the curve achieved by a descent method such as the Newton method. This is the correction step. Size of dt was taken constant equal to 0.2.

5.4.2 Optimization geometry analysis of the Rayleigh quotient

Eigenpairs of matrices are commonly approximated through iterative methods such as power and inverse iteration, QR algorithms or others **[Demmel, 1997]**. For a general matrix with no structure, these methods provide a linear rate of convergence. Algorithms for structured matrices however can reach cubic convergence **[Demmel, 1997]**. For Hermitian matrices, an efficient method is the Rayleigh quotient iteration, which possesses cubic convergence rate to one of the eigenpairs **[Helmke and Moore, 2012]**. This section recalls the definition and interest of the Rayleigh quotient and shows how applying optimization geometry to solving this problem leads to manifolds as the natural domain space for computations.

Three definitions of the Rayleigh quotient

Let A be a symmetric real matrix with eigenvalues $(\lambda_1, \dots, \lambda_n)$, and $S_n(\mathbb{R})$ be the set of real symmetric matrices. The Rayleigh quotient with respect to A, denoted $R(\cdot; A)$, is the smooth function built from:

$$\begin{array}{rccccc} R\colon & \mathbb{R}^n\backslash\{0\} & \times & S_n(\mathbb{R}) & \to & \mathbb{R} \\ & x & ; & A & \mapsto & \frac{x^T A x}{x^T x} \end{array}$$

The following theorem is a special case of the Courant-Fischer minimax theorem characterizing the eigenvalues of a real symmetric $n \times n$ matrix **[Golub and Van Loan, 2012]**.

Theorem 13. *The critical points (resp. critical values) for $R(\cdot; A)$ are the eigenvectors (resp. eigenvalues) of A.*

Note that the Rayleigh quotient is scale-invariant, that is:

$$\forall \gamma \in \mathbb{R}, \quad R(x; A) = R(\gamma x; A)$$

In practice, this property raises a problem for numerical methods. For instance, the Newton iteration at x_k on a scale-invariant function yields the Newton iterate $\mathcal{N}(x_k) = 2x_k$ [Absil et al., 2009]. As a consequence, the Newton method applied to the Rayleigh quotient does not converge to the eigenvectors of the matrix A, unless its initial point is actually an eigenvector. Because of its homogeneity, R is also ill-suited to an optimization geometry framework. Let A a symmetric matrix be the data, and x the vector-parameter on which we optimize. It is obvious that for any A, $R(\cdot; A)$ is not a Morse function because if x^* is a critical point, then the whole line defined by γx^*, $\gamma \in \mathbb{R}$ is also made of critical points. Hence critical points of the Rayleigh quotient for any matrix A are not isolated, they are necessarily degenerate. In a sense, it is the excess of symmetry in R that generates a problem. To reduce the redundant information, one can change the domain space of the Rayleigh quotient. There are two ways to do so that are briefly introduced hereafter.

Rayleigh quotient on the sphere

Rather than defining R over $\mathbb{R}$, where an infinity of points are sent to the same number, one can define R on the sphere S^{n-1}. By abuse of notation, we consider:

$$\begin{array}{rccccc} R^s & : S^{n-1} & \times & S_n(\mathbb{R}) & \to & \mathbb{R} \\ & (x & , & A) & \mapsto & x^T A x. \end{array}$$

The product $x^T A x$ is not defined if x is a point on the sphere, but this expression should be understood as: $\phi(x)^T A \phi(x)$ where ϕ is an embedding from the sphere to the set of vector with norm 1 in $\mathbb{R}^n$. The eigenvectors of A can be found by $\phi(x^*)$ where x^* is a critical point of R. Similarly, we define $-x$ the antipodal element to x on the sphere. Then, if $\phi(x^*)$ is an eigenvector of A, x^* and $-x^*$ are critical points of R^s. Furthermore, R^s and R coincide on points where they are both defined: for all x in the image of the embedding ϕ, $R^s(\phi^{-1}(x)) = R(x)$.

Rayleigh quotient on the projective space

Similarly, because R sends every line in $\mathbb{R}^n$ to the same number, it is tempting to redefine the domain of R by quotienting $\mathbb{R}^n$ by the relation $x \sim \gamma x$, $\forall \gamma \in \mathbb{R}$. This gives the real projective space RP^{n-1} known as the set of directions in $\mathbb{R}^n$. It is a compact smooth manifold which, contrarily to the sphere, is not embedded in $\mathbb{R}^n$. Imagine building the projective

space by gluing every point on a sphere with its antipodal. This operation in $\mathbb{R}^n$ leads to a self-intersection, showing it can not be a faithful representation (an embedding) of RP^{n-1}. We define:

$$\begin{array}{rccccc} R^p: & RP^{n-1} & \times & S_n(\mathbb{R}) & \to & \mathbb{R} \\ & \overline{x} & , & A & \mapsto & \overline{x}^T A \overline{x} \end{array}$$

where $\overline{x}$ denotes an equivalence class in $\mathbb{R}^n$: $\overline{x} = \{\lambda x \ ; \ \lambda \in \mathbb{R}\}$. Again, note that R^p and R evaluated on the same element (modulo the equivalence relation) coincide. Furthermore, if x^* and $-x^*$ are both critical points of R^s_A, $\overline{x}^*$ is the only critical point of R^p_A. These restrictions of the domain of R now enable to give the conditions under which R^s_A or R^p_A are Morse functions.

Theorem 14. *Let A be a symmetric matrix in $S_n(\mathbb{R})$. The functions R^s_A and R^p_A are Morse if and only if all the eigenvalues of A are distincts.*

Proof. For η a tangent vector to the sphere at x, the second-derivative of R^s_A is [Absil et al., 2009]:

$$D^2(R^s_A)(x).\eta_x = 2(Id - xx^T)(A\eta_x - \eta_x x^T Ax)$$

if v is an eigenvector of A with eigenvalue λ:

$$D^2(R^s_A)(v).\eta_v = 2(Id - vv^T)(A\eta_v - \lambda_v \eta_v)$$

the last expression cancels if and only if η_v is an eigenvector of A for the eigenvalue λ_v. Because all eigenvalues are distinct, it cancels only for vectors proportional to v. As the tangent plane at v on S^{n-1} contains only vectors orthogonal to v in $\mathbb{R}^n$, the second derivative $D^2(R^s_A)(v)$ is strictly injective. This is true for any eigenvector of A, hence the critical points of R^s_A are not singular.
Similarly, for $\eta_{\overline{x}}$ a tangent vector of PR^{n-1} at $\overline{x}$, we have

$$D^2 R^p_A(\overline{x}).\eta_{\overline{x}} = 2P^h_{\overline{x}} \left(A\eta_{\overline{x}} - \lambda_i \eta_{\overline{x}}\right)$$

where $P^h_{\overline{x}} = I - \frac{1}{\|\overline{x}\|^2} \overline{x}\overline{x}^T$ is a non-vanishing matrix. Hence, $D^2 R^p_A(\overline{x}).\eta_{\overline{x}} = 0 \iff \eta_{\overline{x}} = \overline{x}$ which is not possible, as the tangent space $T_{\overline{x}} RP^{n-1} = \{y \in \mathbb{R}^n, x^T y = 0\}$ does not contain $\overline{x}$. □

The previous theorem shows that the Rayleigh quotient fits into our optimization geometry framework as long as there exists a path $\Gamma(t)$ between A and B where all the matrices $\Gamma(t)$ have distinct eigenvalues. In that case only, the function $R^{s,p}_\Gamma$ is fibre-wise Morse. We discuss in next section about the possibility to find such a path.

A path in the set of symmetric matrices

As mentioned earlier, homotopy methods are essentially path-following method, they rely on the definition of a homotopy map [Allgower and Georg, 2012] parametrised by a variable t such that $H(0,\cdot)$ is a function whose zeros are known and $H(1,\cdot)$ is the function whose zeros are seeked. For instance, a classical map that is built to solve the problem of finding the eigenpairs of a matrix A is [Li and Rhee, 1989],[Chu, 1988]:

$$H(t,x,\lambda) = \left(\left((1-t)D + tA\right)x - \lambda x, \frac{x^T x - 1}{2}\right)$$

Observe that

$$H\left(t,(x,\lambda)\right) = 0 \iff (\lambda, x) \text{ is an eigenpair of } (1-t)D + tA \text{ with } \|x\| = 1$$

In particular

$$(1,(x,\lambda)) = \left(Ax, \frac{x^T x - 1}{2}\right)$$

so that the set $H\left(1,(\cdot,\cdot)\right)^{-1}\left(\{0\}\right)$ is exactly made of the eigenpairs of A for eigenvectors of norm 1. Similarly, for $t = 0$, $H\left(0,(x,\lambda)\right) = \left(Dx, \frac{x^T x - 1}{2}\right)$ so that the antecedent of 0 are the eigenpairs of the matrix D. The eigenpairs of D are the starting points of the continuation method used to solve this homotopy problem, they should therefore be known in advance. This can be achieved through precomputations or more simply by choosing D diagonal which is the approach adopted in [Li and Rhee, 1989],[Chu, 1988]. Note that the second term of the function is to control the norm of x, so that if H is defined on the sphere or the projective space, it is not needed anymore. It was already noticed in [Li and Rhee, 1989] that the homotopy map is well suited to continuation method if and only if all the matrices in the set $\{(1-t)A + tD \mid t \in [0,1]\}$ have distinct eigenvalues. We find the same conditions inside the optimization geometry framework, which is logical as the numerical method for homotopy method and optimization geometry are interchangeable. The following result is shown in [Chu, 1988].

Proposition 34. *Let E be the set of n-tuples in $\mathbb{C}^n$ such that $D = diag(d_1,\ldots,d_n)$ satisfies: $(1-t)A + tD$ have distinct eigenvalues for all $t \in [0,1]\}$. Then, E is a dense open subset of $\mathbb{C}^n$ of full Lebesgue measure in $\mathbb{C}^n$.*

From a practical point of view, the previous theorem says that the homotopy method can be initialized with a random diagonal matrix in $\mathrm{diag}(\mathbb{C}^n)$ and it gives a probability one that the homotopy map from D to A built from this matrix is well-behaved. By well-behaved, we mean that the continuation method used to numerically solve the problem is proved to converge in a finite number of steps. Interestingly, looking into the proof of this theorem shows that it is crucial to choose randomly D in $\mathbb{C}^n$. The set E' of real n-tuples satisfying the condition of proposition 34 has Lebesgue measure 0 in $\mathbb{C}^n$. In optimization geometry, we wish

to use the pre-computed points from previous iteration of an optimization algorithm, meaning that the starting points are not picked at random. But as we saw, if the starting point is real, it should be chosen carefully to maintain the eigenvalue separation throughout the path. Given the poor reliability of the convex combination in the real case, we propose to examine a path following method based on the geodesic between two matrices. Let A and B be two invertible positive symmetric matrices. Then, the geodesic between A and B is parametrised by: $\Gamma(t) = A^{\frac{1}{2}}(A^{-\frac{1}{2}}BA^{-\frac{1}{2}})^t A^{\frac{1}{2}}$. Note that positiveness was not a criteria until there, but because the set of invertible symmetric matrices is not connected it is convenient to choose A and B positive from here. figure 5.9 compares in the 2-dimensional case $A, B \in S_2^+(\mathbb{R})$ the Riemmannian geodesic and the convex combination. figure 5.10 displays the resulting paths that would be followed by a continuation method in the cylinder $S^1 \times [0, 1]$. It is visible on the figures that the choice of the path in $S_2^+(\mathbb{R})$ greatly influences the path in the set of eigenvectors. This has its importance, as convergence of continuation methods is more or less rapid given elements such as the variation of the derivatives on this path, and the absolute value of the derivative. In this precise example, a high-value for the derivative results in a curve with almost vertical sections. Hence our prediction is that for the two matrices chosen in figure figure 5.9, the straight line gives better numerical results than the geodesic, as we can see the variations on the grey curves are smoother than the variations on the red curve. This result however can not be generalized to every possible couple of matrices.

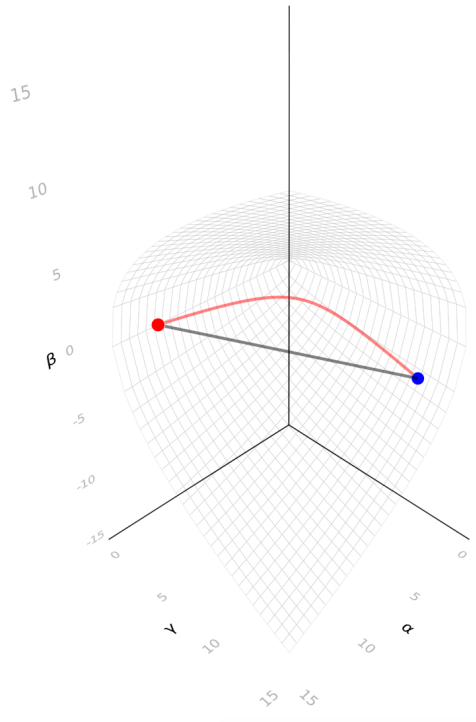


Fig. 5.9. – Three dimensional illustration of paths in the set of positive semi-definite matrices: paths go from A (red dot) at $t = 0$ to B (blue dot) at $t = 1$. Black line is a straight line between the two while red line is the Riemmannian geodesic. The wireframe in the background is the limit set of positive semi-definite matrices.

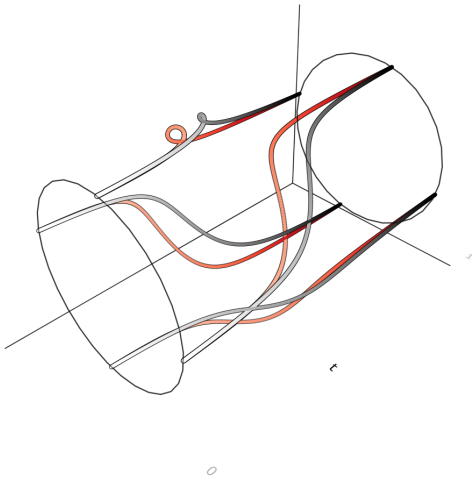


Fig. 5.10. – Paths in the set of one-dimensional eigenvectors: black and red lines are the critical point along respectively the straight line and the Riemmannian geodesic between A and B of the Rayleigh quotient. Parametrization of the path is shown both as a color gradient and on the x-axis: left-white for $t = 0$, right-darker for $t = 1$.

5.5 conclusion

In this chapter we presented the idea of optimization geometry, and showed how it compares with the work of the homotopy community on the special example of the Rayleigh quotient. We highlighted prioritary targets for future work, such as determining properties on the separate eigenvalue condition along a geodesic in the set of symmetric positive matrices. Future work will consist in bringing more insight and theoretical results on the relation between the choice of the paths in a particular instance of optimization geometry, and the convergence properties of optimization methods. Rayleigh quotient is a simple example of how knowledge on the structure of the function, and in particular the data space, is necessary to adapt the path strategy. If probabilistic results are available, deterministic approaches are still an open problem, for the Rayleigh quotient as for many other functions.

Conclusion

In this part we have seen that considering the context of an optimization problem, i.e. the general form of the optimization problem rather than its particular form relative to a given time or situation, allows us to propose original solutions based on curve sampling. The fundamental element to remember here, which allows us to conduct the study and to guarantee the convergence of the method we propose, is the inverse function theorem, reformulated as the implicit function theorem. It is really the cornerstone on which the ideas of the optimization geometry framework are based. Thus, a possible extension of this framework would be to look for conditions on the function modelling the optimization problem other than the fiber-wise Morse condition which would guarantee the application of the inverse function theorem.

List of publications

Conference

[ResearchGate] J. Lefevre, N. Le Bihan, P-O. Amblard. "*A Geometrical Study of the Bivariate Fractional Gaussian Noise*". In: *2018 IEEE Statistical Signal Processing Workshop (SSP)*, 2018, Fribourg (Germany)

DOI:10.1109/SSP.2018.8450737

[Eurasip] J. Lefevre, J. Manton, N. Le Bihan. "*An optimisation geometry framework for the Rayleigh quotient*". In: *2020 28th European Signal Processing Conference (EUSIPCO)*, 2020

DOI:10.23919/Eusipco47968.2020.9287217

[Research Gate] J. Lefevre, F. Bouchard, J. Manton, S. Said and N. Le Bihan. "*On Riemannian and non-Riemannian Optimisation, and Optimisation Geometry*". In: *IFAC-PapersOnLine Elsevier*, 2020

DOI:10.1016/j.ifacol.2021.06.119

appendix

A

The following appendix contains unpublished work, unnecessary to understand the work we presented in this thesis although not completely unconnected. The second and third sections, on the convolution product on the sphere and functions defined on the Hopf fibrations were trails not followed to their end during the PhD but we believe they offer promising prospects for future work.

A few results on the convergence of the Newton method

How to choose the initial point and what conditions other than convexity can produce a convergent sequence of Newton iterate? We will adress these questions by confronting two approaches that prove similar results using different regularity hypothesis on the function. In the Kantorovitch theory, the Newton method is applied to a function of class $\mathcal{C}^2(E, \mathbb{R})$ while the alpha-theory of Smale applies to analytic functions. In both theories, two kind of results are derived. First, authors focus on a zero x^* of f and give a measure of the basin of convergence of the Newton method around this point. More precisely, they compute the radius of a ball such that the Newton method initialized inside of this ball converges to x^* and doubles its precision with each iteration. This first result is interesting if we know the approximate position of a critical point and want guarantees on how precisely the Newton method should be initialized. It can be used in continuation methods, see next section, to guarantee convergence. It can also be useful to have a theorem that doesn't prio information on the position of the zeros of f. The second kind of result we present compute wether a point x_0 is "good" starting point for the Newton method, independently from any knowledge on the zeros of f. The context for both theories is as follows: E, F are real Banach spaces, U is an open set in E, $\mathcal{B}_E(x, r)$ is the ball in E centered in x with positive radius r. Mentions to the space E might be omitted when it is clear what space the ball belongs to.

Kantorovitch criteria This paragraph will show two results on the convergence of the Newton method for minimizing a function $f\colon U \to F$ of class $\mathcal{C}^2$.

Theorem 15. *Let $x^* \in U$ be such that $f(x^*) = 0$ and $Df(x^*)$ is invertible. We define the constant $K(f, x^*, r)$ as*

$$K(f, x^*, r) = \sup_{x \in \mathcal{B}(x^*, r)} \left\| [Df(x^*)]^{-1} D^2 f(x) \right\|$$

If $\|x_0 - x^*\| \leqslant min(\frac{1}{K}, r)$ *, the Newton sequence* $(\mathcal{N}^k(x_0))_{k\geqslant 0}$ *is defined and converges towards* x^**. Furthermore,*

$$\|\mathcal{N}^k(x_0) - x^*\| \leqslant \frac{1}{2^{2^k-1}}\|x_0 - x^*\|$$

When the conditions of the theorems are verified, the point x_0 is an approximate zero of f and the Newton sequence converges to x^*. The constant K can not be computed point-wise, it is a supremum over the region $\mathcal{B}(x^*, r)$. Furthermore, it implies the knowledge of one root x^* and gives a basin of convergence around this root. The theorem requires more hypothesis, but it does not use the knowledge of a zero of x. Given a point x_0, it determines if x_0 is an approximate zero (and even better) of the function f.

Definition 39. *The following quantity will be used in next theorem:*

$$\beta(f, x_0) = \|Df^{-1}(x_0)f(x_0)\|$$

if $Df(x_0)$ *is invertible and* $\beta(f, x_0) = +\infty$ *otherwise.*

Theorem 16. *Let* $x_0 \in U$ *and* $r > 0$ *such that* $\overline{\mathcal{B}}(x_0, r) \subset U$*. If the following conditions are satisfied:*

1. $2\beta(f, x_0) \leqslant r$
2. $2\beta(f, x_0)K(f, x_0, r) \leqslant 1$

then there exists a unique $x^* \in \overline{\mathcal{B}}(x_0, r)$ *such that:*

1. $f(x^*) = 0$.
2. $Df(x^*)$ *is an isomorphism.*
3. $\|x_0 - x^*\| \leqslant \alpha\beta(f, x_0)$

where $1.6 \leqslant \alpha = \sum_{k=0}^{\infty} \frac{1}{2^{2^k-1}} \leqslant 1.7$*. Furthermore, the Newton sequence* $(\mathcal{N}^k(x_0))_{k\geqslant 0}$ *converges towards* x^* *and*

$$\|\mathcal{N}^k(x_0) - x^*\| \leqslant \alpha\left(\frac{1}{2}\right)^{2^k-1}\beta(f, x_0)$$

Smale criteria In Smale theory, functions are of class $\mathcal{C}^\infty$ and given Taylor's formula in **??** it is the sum of a power series.

Definition 40. *For* $x \in U$ *if* $Df(x)$ *let:*

$$\gamma(f, x) = sup_{k\geqslant 2}\left\|Df(x)^{-1}\frac{D^k f(x)}{k!}\right\|^{\frac{1}{k-1}}$$

and $\gamma(f,x) = \infty$ *otherwise.*

Note the similarity with the definition of $K(f,x,r) = \sup_{y\in\mathcal{B}(\S,\nabla)}\|Df(x)^{-1}D^2f(y)\|$ in Kantorovitch theory, and that $2K(f,x,r) \leqslant \gamma(f,x)$, the equality being verified typically for quadratic functions. The following result can be derived [Blum et al., 2012]

Theorem 17. *Suppose that* $f(x^*) = 0$ *and that* $Df(x^*)$ *is an isomorphism. Then, if*

$$\|x - x^*\| \leqslant \frac{5-\sqrt{17}}{4\gamma(f,x^*)}$$

the iterate of the Newton method

$$x_k = x_{k-1} + f'^{-1}(x_{k-1})f(x_{k-1}) \tag{A.1}$$

satisfies

$$\|x_k - x^*\| \left(\frac{\gamma(x^*)}{\psi(u)}\right)^{2^k-1} |x - x^*|^{2^k}$$

where $\psi(x) = 2x^2 - 4x + 1$ *and* $u = (x - x^*)\gamma(x^*)$. *Furthermore, if* $\|x - x^*\| \leqslant \frac{3-\sqrt{7}}{2\gamma(x^*)}$ *then*

$$\|x_k - x^*\| \left(\frac{1}{2}\right)^{2^k-1} |x - x^*|$$

Next theorem characterizes approximate zeros

Definition 41. *Let*

$$\beta(f,x) = \|Df(x)^{-1}f(x)\|$$

and

$$\alpha(f,x) = \beta(f,x)\gamma(f,x) = \|Df(x)^{-1}f(x)\| \sup \left\|Df(x)^{-1}\frac{D^kf(x)}{k!}\right\|^{\frac{1}{k-1}}$$

Next theorem was obtained by **[Xing-hua and Dan-fu, 1990]** based on the work of **[Blum et al., 2012]**. We prefer its formulation as it will be easier to compare it with the equivalent theorem in Kantorovitch theory.

Theorem 18. *For all* $x \in U$ *such that* $\alpha(f,x) < \frac{13-3\sqrt{17}}{4}$, *the Newton sequence* $(N^k(x))_k$ *converges towards a zero* x^* *of* f *and*

$$\left\|x^* - N^k(x)\right\| \leqslant \frac{5-\sqrt{17}}{\gamma(f,x)}\left(\frac{1}{2}\right)^{2^k-1}$$

Remark 0.0.1

Given that an analytic is also a smooth function, it is natural and interesting to compare the rate and bounds obtained by theorem 15 to those obtained for the same function and the same points in theorem theorem 18 . As we will show in the remark, the comparison is not straightforward and can not be generalized to all analytic functions. Note that for an analytic function, the constant $\gamma(f,x) = \sup_{k\geqslant 2}\left\|\frac{f'(x)^{-1}f^{(k)}(x)}{k!}\right\|^{\frac{1}{k-1}}$ must always be greater than $\|\frac{Df(x)^{-1}D^2f(x)}{2}\| = \gamma(f,x)$ which shows the same terms as theorem 15. If the sup in the definition of γ is reached for $k=2$ then $\gamma \leqslant \frac{\|Df(x)^{-1}\|\|D^2f(x)}{2}$ by submiltiplicativity of the norm. Using the same notations as in theorem 2.7.2 yields $\gamma \leqslant \frac{L}{2h}$. The condition in theorem 15 is that $\|x-x^*\| \leqslant \frac{2h}{3L}$ whereas the condition in theorem 2.7.1 is that $\|x-x^*\| \leqslant \frac{3-\sqrt{7}}{2\gamma}$, and we have $\frac{2h}{3L} \leqslant \frac{1}{3\gamma}$. But because $\frac{1}{3} \geqslant \frac{3-\sqrt{7}}{2}$, it is not possible to conclude from here that the fast convergence radius for theorem 18 is bigger or smaller as for theorem theorem 15.

Signal on the sphere and convolution product: a critical review

generalities about groups and homogeneous spaces

Group action

Let G be a group and X a topological set. A Group action of G on X is the data of a map $G\times X \to X$. For instance, for $X=G$, the group action defined by the natural product on G is

$$\begin{aligned} G\times G &\to G \\ (g_1,g_2) &\mapsto g_1g_2 \end{aligned}$$

The group action defines for each $g\in G$ a map on X commonly called g, and $g(x)$ is usually replaced by gx. For instance, if G is S_n the symmetric group and $X=[|1,\dots,n|]$, the action defined by $\tau_{12}\in G$ is the map

$$\begin{aligned} X &\to X \\ i &\to 2 \qquad \text{if } i=1 \\ i &\to i \qquad \text{if } i\neq 1 \end{aligned}$$

Definition 42. *[Orbit] For a fixed group action of G on X, the orbit of an element x in X is the set $\{gx, g \in G\}$.*

Definition 43. *[Transitivity] A group action $G \times X \mapsto X$ is transitive if for every pair of elements $x, y \in X$ there is an element $g \in G$ such that $gx = y$. This element needs not be unique. If a group action is transitive, every element x has the same orbit X.*

Let H be a subgroup of G, for any $g \in G$, the set $g^{-1}Hg = \{g^{-1}hg, h \in H\}$ is also a subgroup of G and is said to be the conjugate of H.

Definition 44. *[Isotropy group] The isotropy group of an element $x \in X$ is the set of stabilizers of x in G for the group action defined. It is a subgroup of G and we note*

$$I_x = \{g \in G, gx = x\}$$

For two point x, y on the same orbit say $y = gx$, the isotropy groups I_x and I_y are conjugate subgroups. More precisely, $I_x = g^{-1}I_y g$

Example. The group $SO(3)$ of all rotations in $\mathbb{R}^3$ acts on the two sphere $S^2 = \{x \in \mathbb{R}^3, \|x\| = 1\}$ through the maps

$$\begin{aligned} S^2 &\to S^2 \\ x &\mapsto Rx \end{aligned}$$

where Rx is like a matrix-vector product for the usual matricial representation of rotations. The action of $SO(3)$ on S^2 is transitive. Indeed, for any x and y on the sphere, there is a rotation that sends x on y. The isotropy group of a point x on the sphere is the set of all rotations not moving x i.e all the rotations of axis x. Because isotropy groups in $SO(3)$ are sets of matrices with the same axis of rotations, they are isomorphic to $SO(2)$, the rotations on the unit circle. Thus Isotropy group are abelian subgroups of $SO(3)$ (they are even maximal abelian subgroups). Any two points on the sphere being on the same orbit (by transitivity), all isotropy groups are conjugate.

Cosets

Let H be a subgroup of G and $g \in G$. The left coset of H in G with respect to g is the set $\{gh\ ,\ h \in H\}$. The right coset is identically defined as $Hg = \{hg, h \in H\}$. Right and left cosets are not generally subgroups.There is a one to one correspondance between right and left cosets. Indeed, any right coset Hg is a left coset of the conjugate subgroup $g^{-1}Hg$. If the right and left cosets of H coincide for g running through all elements of G, then H is a normal

subgroup (i.e elements of H commute with any element in G). The set of all left cosets of H is a partition of G. It is the set of equivalency classes for following equivalence relation:

$$x \sim y \text{ if and only if } xy^{-1} \in H$$

From this relation, it is clear that the quotient G/H has a transitive action on H.

Notations:

- G/H is the set of left cosets
- $H\backslash G$ is the set of right cosets
- $K\backslash G/H$ is the set of double cosets $\{KgH, g \in G\}$

Definition 45. *[Representative] A representative is an element in one of the left (or right) cosets of a subgroup H. For instance, g is always a representative of the left coset gH. A set of exactly one representative for each left (or right) cosets of H is called a transversal or a cross section of G for G/H.*

Some properties

The subgroup H itself is a left and a right coset of itself. If G is a discrete group, all left and right cosets have the same order equal to the order of H and the number of left (or right) cosets is known as the index of H in G. The set of left cosets G/H is a group if and only if H is a normal subgroup. If G is a connected Lie group acting transitively on X, then there exists a unique homeomorphism between G/I_x and X where I_x is one of the isotropy groups.

Example The group $SO(3)$ has a representation in terms of all the rotations on the 2-sphere $S^2 = \{x \in \mathbb{R}^3 \| x\|^2 = 1\}$. It is the subgroup of auto-adjoint matrices in $Gln(\mathbb{R}^3)$ with determinant 1. The group $SO(3)$ acts on the sphere S^2 through the transformations

$$\phi_R : x \in S^2 \mapsto Rx \in S^2, R \in SO(3) \tag{A.2}$$

Each non-trivial rotation R in $SO(3)$ has exactly two invariants in S^2. Explicitely, for each non-trivial R, there is a unique couple $(\mu, -\mu)$ in S^2 such that:

$$R\mu = \mu \text{ and } R(-\mu) = -\mu \tag{A.3}$$

The set of all matrices admitting μ as an invariant form a maximal abelian subgroup of $SO(3)$, called a maximal torus $\mathbb{T}_\mu$, or the isotropy group of μ. Obviously, $\mathbb{T}_\mu = \mathbb{T}_{-\mu}$, but for $\nu \neq \mu \neq -\mu$, the two tori $\mathbb{T}_\mu$ and $\mathbb{T}_\nu$ have a trivial intersection. Except for the identity, each element of $SO(3)$ belongs to exactly one maximal torus: the set of all maximal tori form a

partition of $SO(3)$.

Left cosets of a torus: Because each torus in $SO(3)$ is isomorphic to $SO(2)$ and because the choice of the axis doesn't really matter, the cosets of any torus $\mathbb{T}_x$ are often called the cosets of $SO(2)$ in $SO(3)$. The cosets are defined by the following equivalency relation:

$$R \sim R' \iff RR'^{-1} \in \mathbb{T}_x \tag{A.4}$$

and $RR'^{-1} \in \mathbb{T}_x \iff RR'^{-1}x = x$. In other words, the equivalency classes are made of all rotations around the axis x on the sphere S^2. This is true for any point μ on S^2 can be mapped so that the left coset space $SO(3)/SO(2) \simeq S^2$. This result generalizes to dimension n and the left coset space $SO(n)/SO(n-1)$ is isomorphic to the $(n-1)$-sphere. It is useful to recall that in dimension n the groups of rotations in $\mathbb{R}^n$ sharing a fix point are no longer abelian, but still isomorphic to $SO(n-1)$.

Remark 0.0.2

The group $SO(3)$ has no non-trivial normal subgroups unlike $SU(2)$.
If H is a maximal torus in $SO(3)$, then all the H-double cosets are maximal tori (correpsonding to circles on the sphere) due to the fact that all maxima tori are conjugate with each others. It is in bijection with the projective space $\mathbb{R}P^2$
The set of left cosets however can be shown to be in bijection with the 2-sphere S^2. The following section will give us a precise framework to understand the transition from functions on a group G and its algebra for the convolution product to function on a group X where G acts transitively on X. Even though it is based on a measure-theory framework, results naturally specializes to functions when needed.

Invariant Markov Processes under Lie Groups Action

In the following, we consider only compact subgroups K or H of a group G. We suppose G admits left or right invariant Haar measure. This measure restricts to a left and right invariant Haar measure on a compact subgroup, and we call k the measure of the subgroup K. All sets and functions in this document are measurable unless specified otherwise. Like in the first section, we name g the map induced by the action of the element $g \in G$ on a set X. The convolution for two measures μ, ν on the group G is:

$$\mu * \nu(f) = \int_G f(xy)\mu(dx)\nu(dy) \tag{A.5}$$

It extends into the convolution product for function f, g on G with respect to the measure ρ:

$$f * g(x) = \int_G f(xy^{-1})g(y)d\rho(y) \tag{A.6}$$

Definition 46. *Invariance A function f or a measure μ on X is invariant under a measurable map $g\colon X \to X$ if*

$$f \circ g = f$$

The function or measure is said to be invariant under the action of a group G if it is g-invariant for any $g \in G$.

Example. Lévy processes The Lévy processes are all the Markov processes whose statistic are invariant under the action of $\mathbb{R}$ seen as an additive group on itself. As a result, the law is identical at each instant t and the increments are stationary and independant.

Let X be a topological space and G acts transitively on X. Fix a o in X and define $K = \{g \in G, go = o\}$. This defines a subgroup in G called the isotropy group. It is well known that the quotient space G/K where K is an isotropy group is homeomorphic to X. We define the projection map

$$\pi\colon G \mapsto G/K \tag{A.7}$$

$$g \mapsto gK \tag{A.8}$$

This map is continuous and open for the quotient topology on G/K. Under this topology, the action of $G \to G/K$, $xK \mapsto gxK$ is continuous. We can identify G/K and X under the diffeomorphism that sends $\varphi\colon gK \mapsto go$. We identify π and $\varphi \circ \pi$ in the following. A measurable map $S : X \mapsto G$ is a section map on X if $\pi \circ S = id_X$. In general a section map is not continuous on X, but for any $x \in X$, there is a section S continuous on a neighborood of x (the section needs not be the same for each x of course). Naming k the invariant Haar measure on K, and μ a left Haar measure on X, the convolution between two measures μ, ν on X is

$$\mu * \nu(f) = \int_{X\times X}\int_K f(S(x)ky)dk\mu(dx)\nu(dy) \tag{A.9}$$

for any f measurable. This definition is compatible with the definition of the convolution product on G when $K = e$ and **does not depend on the choice for the section map**. We will prove this later comment.

Proof. Any two section map satisfy $\pi \circ S = Id_X$. This means that for any $x \in X$, for any S section map, $S(x)$ belongs to the left coset in G that sends o onto x. For two section maps S, S' there is g, k_1, k_2 such that $S(x) = gk_1$, $S'(x) = gk_2$. Hence, $S(x) = S'(x)k_2^{-1}k_1$. Using the group structure of K, we can say that for any two sections S, S', for any $x \in X$ we have $S(x) = S'(x)k_x$ where $k_x \in K$. Hence $S(x)ky = S'(x)ky$ for any two sections S, S'. □

Remark 0.0.3

The image of S in G defines a cross-section of the left cosets in G. However, a map from X to a cross-section Γ can not always be written $\Gamma = gS(X)$ for some S and some $g \in G$.

Proof. Let γ be a map that send X onto a cross-section Γ of G. Let's name $\gamma(x) = G_x$ where $G_x o = a(x)$. The map $a(x)$ can be any bijection on X. Take the example where a sends each point on its antipodal. Thus, a has no fixed point. It follows that there is no g such that $a(x) = gx$. □

It follows that a map $\gamma X \mapsto \Gamma$ defined like above through a bijection a with no fix-point does not satisfy $\pi \circ \gamma = Id$.

The convolution product for measures on X (and G) is associative:

$$(\mu * \nu) * \gamma = \mu * (\nu * \gamma) \tag{A.10}$$

hence the n-fold convolution $\mu_1 * \mu_2 * \ldots * \mu_n$ is well defined. If ν is K-invariant the integral simplifies in

$$\mu * \nu(f) = \int_{X \times X} f(S(x)y)\mu(dx)\nu(dy) \tag{A.11}$$

and if μ is also K-invariant, so is $\mu * \nu$. An equivalent of this result for functions will be seen in more details in the Gelfand pairs section. The following properties help to understand the transition from functions on G with a K-right or left invariant propertie to functions defined on X.

Proposition 35. *(a) The map $\nu \mapsto \pi\nu$ is a bijection from the set of K-right invariant measures on G onto the set of measures on X. It is also a bijection from the set of K-bi-invariant measures on G onto K-invariant measures on X.*
(b) If ν is a measure on X, its corresponding K-right invariant measure on G is $\pi^{-1}(\nu)(f) = \int_X \int_K f(S(x)k)dk\nu(dx)$.
(c) The map π preserves the convolution product

$$\pi(\mu_1 * \mu_2) = (\pi\mu_1) * (\pi\mu_2) \tag{A.12}$$

provided μ_1 is K-right invariant or μ_2 is K-left invariant or μ_2 is K-conjugate invariant.

The convolution product on measures propagates to a convolution product on functions on X through the following relation. If μ_1 and μ_2 are measures on X with density functions f_1 and

f_2, then $\mu_1 * \mu_2$ has density $f_1 * f_2$. The definition of the convolution product respecting this property is:

$$f_1 * f_2(gK) = \int_G f_1(hK) f_2(h^{-1}gK)\rho(dh) \tag{A.13}$$

This integration can not be reduced to an integration over X.

The following section will give a result about commutative subalgebras of $\mathcal{L}^1(G)$ for a non-commutative group G. It will also detail a property seen in this section: if a measure on G is right and left invariant for the action of an isotropy group K, then the convolution product of this measure with another measure on G is also right and left invariant for this action.

Gelfand pairs

Given a group $(G, \times)$, with a uniform measure $dm_G(G)$ let $\mathcal{L}^1(G)$ be the set of absolutely integrable functions from G to $\mathbb{C}$. In the following G is compact and unimodular. $\mathcal{L}^1(G)$ with the convolution product $*$ is an algebra. For f, g in $\mathcal{L}^1(G)$ the convolution product of f and g writes

$$(f * g)(x) = \int_G f(xt^{-1})g(t)dm_G(t) \tag{A.14}$$

The convolution product on a non-commutative group is not commutative. It is however possible to exhibit a commutative subalgebra of $(\mathcal{L}^1(G), *)$ by the mean of Gelfand pairs. A Gelfand pair is precisely a pair (G, K) where K is a subgroup of G, such that $\mathcal{L}^1(K\backslash G/K)$ is commutative. The space $\mathcal{L}^1(K\backslash G/K)$ is a subspace of $\mathcal{L}^1(G)$ of functions that are K-left and right invariant. This set is in direct homeomorphism with functions of the double coset space $K\backslash G/K$ hence the notation. Notice that K needs not be an isotropy group at this step, and no group action has been defined. For instance, if G is abelian, $(G, \{e_G\})$ where e_G is the neutral element of G, is a Gelfand pair. A functions f on G is K-left and right invariant if it satisfies:

$$\forall x \in G, f(tx) = f(xt) = f(x)\ \forall t \in K \tag{A.15}$$

It is natural to ask how we can find subgroups K such that (G, K) is a Gelfand pair. The case is particularly simple if G is a linear semi-simple connected compact Lie groups. Given any involution σ on G with invariant subgroup K, (G, K) is a gelfand pair.

Example. Once again we developp the case of $SO(3)$ the group of rotations in $\mathbb{R}^3$. This group is a compact simply connected linear Lie group, which is one of the simplest possible case. Rotations in $SO(3)$ do not commute in general. More precisely, two rotations around different axis do not commute and the set $\mathcal{L}^1\{SO(3)\}$ is a non-commutative algebra. To find a commutative subalgebra, we need to find a subgroup K of $SO(3)$ such that there is an involution admitting K as an invariant. Consider the application

$$\sigma : R \mapsto WRW^{-1} \tag{A.16}$$

where $W^2 = I$ and $W \neq I$ (i.e, W is a non-trivial square root of the identity). It is easy to check that $\sigma \circ \sigma = Id$. The invariants of σ are in the subgroup of matrices commuting with W, conversely, any such matrix is an invariant. In other words the invariant subgroup of σ is the isotropy group containing W. If ν is an element on the sphere invariant under W, we name I_ν the isotropy group of ν. From the previous comments, $(SO(3), I_\nu)$ is a Gelfand pair (a result sometimes stated as "$(SO(n), SO(n-1))$ is a gelfand pair"). As a consequence, the functions in $\mathcal{L}^1(I_\nu \backslash G / I_\nu)$ form a commutative algebra for the convolution product. From definition, the set $I_\nu \backslash G / I_\nu$ is also $I_\nu \backslash (G / I_\nu)$, where the coset space G/I_ν is homeomorphic to S^2. It yields that functions I_ν-right and left invariant on $SO(3)$ are homeomorphic to functions I_ν-left invariant on S^2. Such functions are constant on each orbit of S^2 under the action of I_ν. We already said that I_ν contains rotations around the same axis and can be seen as a copy of $SO(2)$ inside $SO(3)$. Thus, the orbit of $x \in S^2$ is a circle going through x orthogonal to ν. Hence, functions of $\mathcal{L}^1(I_\nu \backslash G / I_\nu)$ are homeomorphic to functions on the sphere constant on all the circles perpendicular to ν, they only depend on one parameter $\theta = \nu^T x$ where x is the position on the sphere.

Review and analysis of the proposed convolution products on the sphere

Because a class of natural signal on the sphere exist, a need for signal processing tools for signal on the sphere arise. Among which the convolution product holds a significant place, due to its relation to filtering. Just like the definition of the convolution for functions on a graph, an issue is raised by the absence of group structure on the sphere. However, through the action of the group of rotation on the sphere, the problem can be lifted by considering functions on the sphere to be particular functions on the group of rotations We will present in a short way different definitions found in the litterature, and will focus more particularly on the one given in **[Sadeghi et al., 2012]**.

Natural convolution product

The first definition is maybe the most natural. It is found in **[Driscoll and Healy, 1994]**, in a more general shape in **[Liao, 2009]** and in another but equivalent shape in **[Le Bihan et al., 2016]**. It consists in seeing functions on the sphere as functions on $SO(3)$ constant on left cosets of a given isotropy group (the choice of which has no significance). We also say that such functions are I-right invariant where I is the isotropy group in question. Next definitions will precise this notion. For the isotropy group I of e_0, we define π, the open continuous function

$$\begin{aligned} \pi \colon SO(3) &\to SO(3)/I_x \\ r &\mapsto rI \end{aligned} \tag{A.17}$$

for the quotient topology, and ϕ the homeomorphism

$$\begin{aligned} \phi\colon SO(3)/I &\to S^2 \\ \dot{R} &\mapsto \dot{R}e_0 \end{aligned} \tag{A.18}$$

where $\dot{R}$ symbolizes the equivalency class of R. The map $\phi \circ \pi$ is continuous and open. We will slightly abuse notations and use π for π or $\pi \circ \phi$. The map π defines a homeomorphism from functions on $SO(3)$ constant on left cosets of I to functions on the sphere by:

$$\begin{aligned} \mathcal{L}^1\{SO(3)\} &\to \mathcal{L}^1\{S^2\} \\ f &\mapsto \tilde{f} = f \circ \pi \end{aligned} \tag{A.19}$$

The definition of the convolution product follows:

$$\begin{aligned} f *_1 g(x) &= \int_G f(xt^{-1})g(t)dt \\ f *_1 g(x) &= \int_G \tilde{f}(xt^{-1}e_0)\tilde{g}(te_0)dt \end{aligned}$$

This convolution yields a function $f *_1 g$ constant on the left cosets of I. To see that, just check that for $k \in I$

$$f *_1 g(xk) = \int_G \tilde{f}(xkt^{-1}e_0)\tilde{g}(te_0)dt = f *_1 g(x) \tag{A.20}$$

Apply the substitution $t = uk$ in the middle integral. Note that the measure on $SO(3)$ is invariant, so that $d(uk) = du$. Then the integral becomes:

$$f *_1 g(xk) = \int_G \tilde{f}(xu^{-1}e_0)\tilde{g}(uke_0)du$$

and $ke_0 = e_0$ by definition. Hence, equality (A.20) is true.

Convolution product as an integration on the sphere

Some convolution products on the sphere found in the signal processing literature **[Yeo et al., 2008]**, **[Wandelt and Gorski, 2001]**, **[Sadeghi et al., 2012]** derive from the transform on $SO(3)$

$$T(f,g)(x) = \int_G f(xy)g(y)dy \tag{A.21}$$

similar but not identical to the convolution product. This bilinear transform extends to functions on the sphere by

$$T(\tilde{f},\tilde{g})(x) = \int_G \tilde{f}(xye_0)\tilde{g}(ye_0)dy \tag{A.22}$$

note that the resulting function is not in general I-right invariant. Indeed we have for $k \in I$:

$$\begin{aligned} T(\tilde{f},\tilde{g})(xk) &= \int_G \tilde{f}(xkt.e_0)\tilde{g}(t.e_0)dt \\ &= \int_G \tilde{f}(xu.e_0)\tilde{g}(k^{-1}u.e_0)du \end{aligned}$$

take for instance k^{-1} the rotation of axis e_0 that sends e_1 to its antipodal on the sphere $-e_1$, and let $\tilde{g}$ be the function that is $+1$ on the half-sphere centered on e_1 and is 0 on the half sphere centered on $-e_1$. Then, $T(\tilde{g},\tilde{g})(I) = 1$ whereas $T(\tilde{g},\tilde{g})(Ik) = 0$. One may wonder why authors in signal processing define a transform on the sphere based on this transform on $SO(3)$. The reason is (I believe) in the following result:

Proposition 36. *The integral in definition (A.21) reduces to an integration over S^2 for functions I-right invariant functions. Let f, g be two functions I-right invariant on $SO(3)$ and $\tilde{f}, \tilde{g}$ their image in $\mathcal{L}^1\{S^2\}$ by π (**??**). The transform T of f and g writes:*

$$T(f,g)(x) = \int_{S^2} \tilde{f}(xu)\tilde{f}(u) \tag{A.23}$$

Proof. Let f, g be two functions I-right invariant on $SO(3)$ and $\tilde{f}, \tilde{g}$ their image in $\mathcal{L}^1\{S^2\}$by π (**??**).Define a section map S satisfying $\pi \circ S = Id_{S^2}$. The transform writes:

$$\begin{aligned} T(f,g)(x) &= \int_G \tilde{f}(xt.e_0)\tilde{g}(t.e_0)dt \\ &= \int_{G/I}\int_I \tilde{f}(x\dot{t}k.e_0)\tilde{g}(\dot{t}k.e_0)dkd\dot{t} && \text{(A.24)} \\ &= \int_{G/I}\int_I \tilde{f}(x\dot{t}e_0)\tilde{g}(\dot{t}e_0)dkd\dot{t} && \text{(A.25)} \\ &= \int_{G/I} \tilde{f}(x\dot{t}e_0)\tilde{g}(\dot{t}e_0)d\dot{t} && \text{(A.26)} \\ &= \int_{S^2} \tilde{f}(xS(u).e_0)\tilde{g}(S(u).e_0)du && \text{(A.27)} \\ T(f,g)(x) &= \int_{S^2} \tilde{f}(xv)\tilde{g}(v)dv && \text{(A.28)} \end{aligned}$$

In (A.24), $\dot{t}$ is the notation for the equivalency class of t in G/I. The measure $\dot{t}$ is the measure $\pi \circ dt$ where dt is the Haar measure on $SO(3)$. The measure dk is the normalized Haar measure induced by dt on I. The equation is valid because of the invariance of the Haar measure on $SO(3)$ see **[Garrett, 2014]**. In equation (A.25) just notice that for any $k \in I$,

$k.e_0 = e_0$. To go from equation (A.26) to equation (A.27) we use the substitution $\dot{t} = S(u)$ and $d\dot{t} = du$ by the canonical homeomorphism defined in (A.18) and by invariance of the measure on S^2. In equation (A.28) we apply the change of variables $S(u).e_0 = v$ by noticing that the action of $\{S(u), u \in S^2\}$ is transitive on S^2. Moreover, by invariance $du = dv$. □

Even though functions defined through this transform can not be seen as function on the spheres and is probably not associative, the authors in **[Sadeghi et al., 2012]** consider it more appropriate (for some reason) and have proposed a developpment that I will quickly present in the next section.

Convolution product by Sadeghi et. al.

The authors in **[Sadeghi et al., 2012]** have proposed a modification of the transform defined in (A.23) whose spirit is to lift the inconvenience of the non-stability of this transform. They define it in the following way (we adapted notations). For two functions f and g on the sphere, their convolution product (as defined by the authors) is a function on the sphere defined by:

$$f * g(x) = \int_{S^2} f(\alpha(x).u)g(u)du \tag{A.29}$$

where α is a map from S^2 to a transversal Γ of $SO(3)/I$ in $SO(3)$. It is more precisely defined by:

$$\alpha(y) = (T_{e_1}(\phi)T_{e_2}(\nu)T_{e_1}(\pi - \phi))^{-1} \tag{A.30}$$

where y is the point on the sphere with coordinates $(\cos\phi \sin\nu, \cos\phi\cos\nu, \cos\phi)$ in the orthonormal base (e_1, e_2, e_0), and $T_{e_1}(\phi)$ is a rotation around the axis e_1 of an angle ϕ. The image of α called Γ is a transversal of $SO(3)$ and is in the intersection of $SO(3)$ with symmetric matrices. This last condition ensures the commutativity of the defined convolution product. One may notice that this convolution product is the restriction to Γ of the bilinear transform in (A.23), composed with γ so that the domain stays in S^2.

Signals on the Hopf fibration: an introduction

We call Hopf fibration the description of the sphere S^3 in terms of a 2-sphere S^2 and circles S^1. Formally, the Hopf fibration is given by S^3 together with a projection $\pi : S^3 \to S^2$ such that the preimage of every point in S^2 by π is homeomorphic to a circle: $\forall p \in S^2,\ \pi^{-1}(p) \simeq S^1$. The first image that might come to mind with this description is the sphere S^3 as a cartesian product $S^2 \times S^1$. This image however is not accurate as the former cartesian product doesn't have the same topological feature as the 3-sphere. The difference will be of the same order as the difference between the Torus $S^1 \times S^1$ and the sphere S^2. Using cohomology, it is actually possible to show that no sphere can be expressed as a cartesian product $X \times Y$ unless one

of X or Y is a point (https://math.stackexchange.com/questions/77175/decomposing-the-sphere-as-a-product).

Parametrization: Extrinsic and Intrinsic coordinates To accurately describe the projection map π and to precisely describe different points on the 3-sphere, one needs to define a coordinate system on this sphere: a parametrization. One set of parameter should define a unique position on the sphere, and it is required that close parameters should give close points in the manifold (under a previously defined distance). Precisely, we require a parametrization to be a homeomorphism from a subset of S^3 to an open set of $\mathbb{R}^3$. It means the mapping is invertible and both the mapping and its inverse are continuous. There are two ways it can be done, by using extrinsic coordinates. A set of extrinsic coordinates for the sphere are given by coordina

As with any manifold, one of the question that arises is the parametrization. How can we represent a point p on S^3 with coordinates such that small variations of these coordinates result in a small variation of the position of p on S^3. More precisely, we are looking for invertible maps $\phi : U\text{open} \subset \mathbb{R}^3 \to S^3$ such that ϕ and ϕ^{-1} are continuous. It should be noted that the usual representation of a point on the sphere as a vector of $\mathbb{R}^4$ is not a parametrization. Indeed, let $p = (a, x, y, z)$ such that $a^2 + x^2 + y^2 + z^2 = 1$. For any open subset of $\mathbb{R}^4$, a small variation of the proposed coordinates (a, x, y, z) potentially result in p leaving the sphere. So the condition that the coordinate function should be defined from an open-space to S^3 is actually important. As a consequence, the number of coordinates must be the intrinsic dimension of the manifold. Thus, a valid parametrization of S^3 is not the vectors of norm 1 in $\mathbb{R}^4$ as this is not a parametrization, but it is the mapping

$$\varphi\colon (\psi, \theta, \phi) \mapsto (\cos\psi, \sin\psi\cos\theta, \sin\psi\sin\theta\cos\phi, \sin\psi\sin\theta\sin\psi) \tag{A.31}$$

- $\psi \in]0, \pi[$
- $\theta \in]0, \pi[$

One usual parametrization of the 3-sphere using the embedding of S^3 in the Euclidian space $\mathbb{R}^4$. One usual parametrization

Application in signal processing

One particular type of signals in physic can be naturally represented on the Hopf fibration, even if it does not use all the degrees of freedom allowed. We will discuss the properties of the polarization of the cosmic microwave background (CMB). Supposing a fixed observation point has been determined, a sensor at that position can be directed in every direction and record a measure of the polarization of the CMB. This polarization is assumed to be purely linear because it is supposed to be generated by Thomson scattering. A linear polarization can be represented by two scalars, the Stokes parameters S_1, S_2 or in the physic litterature U, V.

These two scalars give the coordinate of a vector in the plane, indicating both the direction of polarization and the intensity of the polarization. Of course, these coordinates suppose a coordinate frame has been determined beforehand. This means that to fully characterize a polarization state in a given direction, one needs a coordinate frame, and the coordinates of the vector of polarization in this frame. Because we will limit ourselves to choosing orthonormal coordinate frame, only one unitary vector is enough to set the whole frame, completion being done through right hand rule. So, the observed function has domain in the 2-sphere together with the tangent bundle on the two sphere. However, because the polarization has one value per direction, if we change the coordinate frame, the value of the Stokes parameters must change accordingly. Naming $f(p, y) = S_1 + \boldsymbol{i}S_2$ where p is a point on the sphere and y is a vector in T_pS^2, the following relation must be satisfied:

$$f(p, R(\alpha)y) = e^{-\boldsymbol{i}\alpha} f(p, y) \tag{A.32}$$

where $R(\alpha)y$ is the rotation of y by an angle alpha around the axis orthogonal to the tangent plane pointing outwards (the rotation goes in the direction of a right-hand oriented frame). Rotating the coordinate frame must rotate oppositely the value of the Stokes parameter. One can therefore see that evaluating the Stokes parameters for a single coordinate frame is enough to get all the information needed about the fuction. So instead of definition our function from the tangent bundle on the sphere S^2, we could define it over a unitary section of this tangent bundle. By the hairy-ball theorem, such a continuous section does not exist. Using a non-continuous section as the domain of f is not necessarily a major problem, but it yields uncomfortable situation, as the continuity of f can not be determined at least at one point. Imagine the situation where an arbitrary point on the sphere is used as a reference, and in each tangent plane, we consider the coordinate frame issued from the vector tangent to the geodesic between the reference point and the present point. This defines a smooth vector field over the sphere everywhere but at the point chosen as a reference. Furthermore, as we approach the reference point, the vector field varies faster and faster, which can result in high varying values of f around this region. The variations however have nothing to do with the physical nature of the polarization at this point, but are completely related to choice of the domain. Another issue with this approach is that two different choices of vector field will yield two completely different function f, a difference that can not be expressed in terms of a unitary change of coordinates for instance. Hence, the choice of a reference point distorts the information, which is not a desirable behaviour. This is why a section of the tangent bundle on the sphere, though the most intuitive choice is not the best. Defining the function on the whole tangent bundle and restricting it using relation A.32 is a clumsy option as it requires to artificially define values for f on vanishig coordinate frames. A better explanation for that can be found in **[?]**.

Considering again earlier options, we notice that only an angle is necessary to determine the position of a unitary vector in the plane. If this vector is rotated by 2π then it must come back to its initial position. In other words, the space of possible orthonormal frame is topologically equivalent to a circle, so that we are really looking after a circle bundle over

S^2 rather than a line bundle. The cricle bundle over S^2 has a well-known topology: as a total space it is the sphere S^3. The bundle structure is present through a projection map $\pi\colon S^3 \to S^2$. Defining the polarization of the CMB as a function over S^3 results in embedding the problem in a higher dimensional space than necessary. This implies the presence of redundancy characterized through a relation of the same type as A.32. Two arguments in our view justify this embeding approach.

Bibliography

[Absil et al., 2009] Absil, P.-A., Mahony, R., and Sepulchre, R. (2009). *Optimization algorithms on matrix manifolds*. Princeton University Press.

[Ahmad et al., 2019] Ahmad, I., Khan, M. A., and A Ishan, A. (2019). Generalized geodesic convexity on riemannian manifolds. *Mathematics*, 7(6):547.

[Aiello and Woerdman, 2004] Aiello, A. and Woerdman, J. (2004). Linear algebra for mueller calculus. *arXiv preprint math-ph/0412061*.

[Aitchison et al., 2004] Aitchison, I. J., MacManus, D. A., and Snyder, T. M. (2004). Understanding heisenberg's "magical" paper of july 1925: A new look at the calculational details. *American Journal of Physics*, 72(11):1370–1379.

[Allgower and Georg, 2012] Allgower, E. L. and Georg, K. (2012). *Numerical continuation methods: an introduction*, volume 13. Springer Science & Business Media.

[Armstrong, 2013] Armstrong, M. A. (2013). *Groups and symmetry*. Springer Science & Business Media.

[Barakat, 1963] Barakat, R. (1963). Theory of the coherency matrix for light of arbitrary spectral bandwidth. *JOSA*, 53(3):317–323.

[Bertrand et al., 1994] Bertrand, J., Bertrand, P., and Ovarlez, J.-P. (1994). The mellin transform. *ONERA, TP no. 1994-98*.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

[Blum et al., 2012] Blum, L., Cucker, F., Shub, M., and Smale, S. (2012). *Complexity and real computation*. Springer Science & Business Media.

[Boashash, 1992] Boashash, B. (1992). Estimating and interpreting the instantaneous frequency of a signal. i. fundamentals. *Proceedings of the IEEE*, 80(4):520–538.

[Brosseau, 1998] Brosseau, C. (1998). *Fundamentals of polarized light: a statistical optics approach*. Wiley-Interscience.

[Brosseau and Barakat, 1991] Brosseau, C. and Barakat, R. (1991). Jones and mueller polarization matrices for random media. *Optics communications*, 84(3-4):127–132.

[Bulow and Sommer, 2001] Bulow, T. and Sommer, G. (2001). Hypercomplex signals-a novel extension of the analytic signal to the multidimensional case. *IEEE Transactions on signal processing*, 49(11):2844–2852.

[Chu, 1988] Chu, M. T. (1988). A note on the homotopy method for linear algebraic eigenvalue problems. *Linear Algebra and its Applications*, 105:225–236.

[Cooke et al., 1992] Cooke, J. M., Zyda, M. J., Pratt, D. R., and McGhee, R. B. (1992). Npsnet: Flight simulation dynamic modeling using quaternions. *Presence: Teleoperators & Virtual Environments*, 1(4):404–420.

[Crowe, 1994] Crowe, M. J. (1994). *A history of vector analysis: The evolution of the idea of a vectorial system*. Courier Corporation.

[Datchev, 2021] Datchev, K. (2021). Geodesics and their minimization properties. consulted on 16-09-2021.

[Delage and Ye, 2010] Delage, E. and Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612.

[Demmel, 1997] Demmel, J. W. (1997). *Applied numerical linear algebra*, volume 56. Siam.

[Driscoll and Healy, 1994] Driscoll, J. R. and Healy, D. M. (1994). Computing fourier transforms and convolutions on the 2-sphere. *Advances in applied mathematics*, 15(2):202–250.

[Ell and Sangwine, 2005] Ell, T. and Sangwine, S. (2005). Quaternion involutions.

[Fegan, 1991] Fegan, H. D. (1991). *Introduction to compact Lie groups*, volume 13. World Scientific Publishing Company.

[Flamant et al., 2016] Flamant, J., Bihan, N. L., and Chainais, P. (2016). Time Frequency Analysis of Bivariate Signals.

[Flamant et al., 2017] Flamant, J., Chainais, P., and Le Bihan, N. (2017). Polarization spectrogram of bivariate signals. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3989–3993. IEEE.

[Flandrin, 2012] Flandrin, P. (2012). Une fréquence peut-elle être instantanée?

[Ford, 2005] Ford, G. A. (2005). Graphical analysis of orbits & fixed points of functions. retrived on 15-04-2021.

[Gabor, 1946] Gabor, D. (1946). Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(26):429–441.

[Gallier, 2008] Gallier, J. (2008). Clifford algebras, clifford groups, and a generalization of the quaternions. *arXiv preprint arXiv:0805.0311*.

[Garrett, 2014] Garrett, P. (2014). Unwinding and integration on quotients.

[Gil, 2007] Gil, J. J. (2007). Polarimetric characterization of light and media-physical quantities involved in polarimetric phenomena. *The European Physical Journal Applied Physics*, 40(1):1–47.

[Gil, 2014] Gil, J. J. (2014). Interpretation of the coherency matrix for three-dimensional polarization states. *Physical Review A*, 90(4):043858.

[Golub and Van Loan, 2012] Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU press.

[Guillemin and Pollack, 2010] Guillemin, V. and Pollack, A. (2010). *Differential topology*, volume 370. American Mathematical Soc.

[Hamilton, 1866] Hamilton, W. R. (1866). *Elements of quaternions*. Longmans, Green, & Company.

[He, 2003] He, J.-H. (2003). Homotopy perturbation method: a new nonlinear analytical technique. *Applied Mathematics and computation*, 135(1):73–79.

[Helmke and Moore, 2012] Helmke, U. and Moore, J. B. (2012). *Optimization and dynamical systems*. Springer Science & Business Media.

[Kanasewich, 1981] Kanasewich, E. R. (1981). *Time sequence analysis in geophysics*. University of Alberta.

[Karimi et al., 2020] Karimi, H., Nutini, J., and Schmidt, M. (2020). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition.

[Karstensen et al., 2016] Karstensen, J., Atamanchuk, D., Bendinger, A., Fried, N., Fupsova, K., Handmann, P., Klein, A., Körner, M., Oltmanns, M., Schmidt, C., et al. (2016). Subpolar gyre variability cruise no. msm54 may 12–june 04, 2016 st. john's (canada)–reykjavik (iceland).

[Kohan, 2014] Kohan, M. (2014). Upper bound on the number of charts needed to cover a topological manifold. Mathematics Stack Exchange. (version: 2014-05-21).

[Lang, 2012] Lang, S. (2012). *Fundamentals of differential geometry*, volume 191. Springer Science & Business Media.

[Le Bihan et al., 2016] Le Bihan, N., Chatelain, F., Manton, J. H., et al. (2016). Isotropic multiple scattering processes on hyperspheres. *IEEE Trans. Information Theory*, 62(10):5740–5752.

[Lee, 2000] Lee, J. M. (2000). *Introduction to smooth manifolds, version 3.0*. University of Washington, Department of Mathematics.

[Lee, 2013] Lee, J. M. (2013). Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 1–31. Springer.

[Li and Rhee, 1989] Li, T. and Rhee, N. (1989). Homotopy algorithm for symmetric eigenvalue problems. *Numerische Mathematik*, 55(3):265–280.

[Li, 1997] Li, T.-Y. (1997). Numerical solution of multivariate polynomial systems by homotopy continuation methods. *Acta numerica*, 6:399–436.

[Liao, 2009] Liao, M. (2009). Markov processes invariant under a lie group action. *Stochastic Processes and their Applications*, 119(4):1357–1367.

[Lilly, 2011] Lilly, J. M. (2011). Modulated oscillations in three dimensions. *IEEE Transactions on Signal Processing*, 59(12):5930–5943.

[Lilly and Olhede, 2009] Lilly, J. M. and Olhede, S. C. (2009). Bivariate instantaneous frequency and bandwidth. *IEEE Transactions on Signal Processing*, 58(2):591–603.

[Lilly and Olhede, 2012] Lilly, J. M. and Olhede, S. C. (2012). Analysis of modulated multivariate oscillations. *IEEE Transactions on Signal Processing*, 60(2):600–612.

[M. Criton, 2016] M. Criton, B. H. (2016). La querelle des maths modernes.

[Manton, 2012] Manton, J. H. (2012). Optimisation geometry. *arXiv preprint arXiv:1212.1775*.

[Manton, 2015] Manton, J. H. (2015). A framework for generalising the Newton method and other iterative methods from Euclidean space to manifolds. *Numerische Mathematik*, 129(1):91–125.

[Matthew R. and Kosowsky, 2006] Matthew R., F. and Kosowsky, A. (2006). The construction of spinors in geometric algebra. *arXiv preprint arXiv:0403040*.

[Olhede, 2013] Olhede, S. (2013). Modulated oscillations in many dimensions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110551.

[Pervin and Webb, 1982] Pervin, E. and Webb, J. A. (1982). Quaternions in computer vision and robotics. Technical report, Carnegie-Mellon Univ Pittsburgh, Dept. of Computer Science.

[Peyré, 2004] Peyré, G. (2004). *L'algèbre discrète de la transformée de Fourier*. Ellipses.

[Picinbono, 1997] Picinbono, B. (1997). On instantaneous amplitude and phase of signals. *IEEE Transactions on signal processing*, 45(3):552–560.

[Rihaczek, 1968] Rihaczek, A. (1968). Signal energy distribution in time and frequency. *IEEE Transactions on information Theory*, 14(3):369–374.

[Roux, 2011] Roux, S. (2011). Pour une étude des formes de la mathématisation.

[Sadeghi et al., 2012] Sadeghi, P., Kennedy, R. A., and Khalid, Z. (2012). Commutative anisotropic convolution on the 2-sphere. *IEEE Transactions on Signal Processing*, 60(12):6697–6703.

[Schreier and Scharf, 2010] Schreier, P. J. and Scharf, L. L. (2010). *Statistical signal processing of complex-valued data: the theory of improper and noncircular signals*. Cambridge university press.

[Simon, 1996] Simon, B. (1996). *Representations of finite and compact groups*. Number 10. American Mathematical Soc.

[Stillwell, 2008] Stillwell, J. (2008). *Naive lie theory*. Springer Science & Business Media.

[Stokes, 1852] Stokes, G. G. (1852). *On the Composition and Resolution of Streams of Polarized Light from different Sources*, volume 3 of *Cambridge Library Collection - Mathematics*, page 233–258. Cambridge University Press.

[Taras I., 2021] Taras I., L. (2021). Simple euler methods and its modifications. http://www.cems.uvm.edu/~tlakoba/math337/notes_1.pdf. retrieved on 31-03-2021.

[Thomson and Emery, 2014] Thomson, R. E. and Emery, W. J. (2014). *Data analysis methods in physical oceanography*. Newnes.

[Van der Pol, 1946] Van der Pol, B. (1946). The fundamental principles of frequency modulation. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(23):153–158.

[Vicci, 2001] Vicci, L. (2001). Quaternions and rotations in 3-space: The algebra and its geometric interpretation. *TR01–014*, pages 1–11.

[Ville, 1948] Ville, J. (1948). Theorie et application de la notion de signal analytic, cables et transmissions, 2a (1), 61-74, paris, france, 1948. *Translation by I. Selin, Theory and applications of the notion of complex signal, Report T-92, RAND Corporation, Santa Monica, CA*.

[Vlachos et al., 2002] Vlachos, M., Domeniconi, C., Gunopulos, D., Kollios, G., and Koudas, N. (2002). Non-linear dimensionality reduction techniques for classification and visualization. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 645–651.

[Wandelt and Gorski, 2001] Wandelt, B. D. and Gorski, K. M. (2001). Fast convolution on the sphere. *Physical Review D*, 63(12):123002.

[Wie and Barba, 1985] Wie, B. and Barba, P. M. (1985). Quaternion feedback for spacecraft large angle maneuvers. *Journal of Guidance, Control, and Dynamics*, 8(3):360–365.

[Wikipedia contributors, 2018] Wikipedia contributors (2018). Coordinate-induced basis — Wikipedia, the free encyclopedia. [Online; accessed 16-September-2021].

[Wikipedia contributors, 2021a] Wikipedia contributors (2021a). Root-finding algorithms — Wikipedia, the free encyclopedia. [Online; accessed 14-September-2021].

[Wikipedia contributors, 2021b] Wikipedia contributors (2021b). Spinor — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Spinor&oldid=1028404981. [Online; accessed 26-August-2021].

[Wikipedia contributors, 2021c] Wikipedia contributors (2021c). Three-body problem — Wikipedia, the free encyclopedia. [Online; accessed 14-September-2021].

[Wikipédia, 2019] Wikipédia (2019). Algorithme de faddeev-leverrier — wikipédia, l'encyclopédie libre. [En ligne; Page disponible le 3-mars-2019].

[Xing-hua and Dan-fu, 1990] Xing-hua, W. and Dan-fu, H. (1990). On dominating sequence method in the point estimate and smale theorem. *Science in China*, page 02.

[Yeo et al., 2008] Yeo, B. T., Ou, W., and Golland, P. (2008). On the construction of invertible filter banks on the 2-sphere. *IEEE Transactions on Image Processing*, 17(3):283–300.

[Zhou and So, 2017] Zhou, Z. and So, A. M.-C. (2017). A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165(2):689–728.