Riemannian Geometry of Matrix Manifolds for Lagrangian Uncertainty Quantification of Stochastic Fluid Flows

by

Florian Feppon

Ingénieur diplômé de l'École polytechnique (2015)

Submitted to the Center for Computational Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Computation for Design and Optimization

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2017

Author .....
Center for Computational Engineering
December 2016

Certified by .....
Pierre F.J. Lermusiaux
Associate Professor, Department of Mechanical Engineering
Thesis Supervisor

Accepted by .....
Youssef Marzouk
Co-Director, Computation for Design and Optimization

Riemannian Geometry of Matrix Manifolds for Lagrangian Uncertainty Quantification of Stochastic Fluid Flows

by

Florian Feppon

Submitted to the Center for Computational Engineering
on December 2016, in partial fulfillment of the
requirements for the degree of
Master of Science in Computation for Design and Optimization

Abstract

This work focuses on developing theory and methodologies for the analysis of material transport in stochastic fluid flows. In a first part, two dominant classes of techniques for extracting Lagrangian Coherent Structures are reviewed and compared and some improvements are suggested for their pragmatic applications on realistic high-dimensional deterministic ocean velocity fields. In the stochastic case, estimating the uncertain Lagrangian motion can require to evaluate an ensemble of realizations of the flow map associated with a random velocity flow field, or equivalently realizations of the solution of a related transport partial differential equation. The Dynamically Orthogonal (DO) approximation is applied as an efficient model order reduction technique to solve this stochastic advection equation. With the goal of developing new rigorous reduced-order advection schemes, the second part of this work investigates the mathematical foundations of the method. Riemannian geometry providing an appropriate setting, a framework free of tensor notations is used to analyze the embedded geometry of three popular matrix manifolds, namely the fixed rank manifold, the Stiefel manifold and the isospectral manifold. Their extrinsic curvatures are characterized and computed through the study of the Weingarten map. As a spectacular by-product, explicit formulas are found for the differential of the truncated Singular Value Decomposition, of the Polar Decomposition, and of the eigenspaces of a time dependent symmetric matrix. Convergent gradient flows that achieve related algebraic operations are provided. A generalization of this framework to the non-Euclidean case is provided, allowing to derive analogous formulas and dynamical systems for tracking the eigenspaces of non-symmetric matrices. In the geometric setting, the DO approximation is a particular case of projected dynamical systems, that applies instantaneously the SVD truncation to optimally constrain the rank of the reduced solution. It is obtained that the error committed by the DO approximation is controlled under the minimal geometric condition that the original solution stays close to the low-rank manifold. The last part of the work focuses on the practical implementation of the DO methodology for the stochastic advection equation. Fully linear, explicit central schemes are selected to ensure stability, accuracy and efficiency of the method. Riemannian matrix optimization is applied for the dynamic evaluation of the dominant SVD of a given matrix and is integrated to the DO time-stepping. Finally the technique is illustrated numerically on the uncertainty quantification of the Lagrangian motion of two bi-dimensional benchmark flows.

Thesis Supervisor: Pierre F.J. Lermusiaux
Title: Associate Professor, Department of Mechanical Engineering

Acknowledgments

> Hommage à Toi, mon propre Soi [...] sous ta forme concrète d'univers dont l'essence transcende le déploiement des phénomènes.
> En qui Tu es Toi, et moi je suis moi, en qui Toi seul es et moi je ne suis pas, en qui il n'y a ni Toi ni moi, à Celui-là je rends hommage!
>
> Abhinavagupta, Twenty Verses on the Great Teaching, *Trad. Lilian Silburn*

This work marks the end of both my 15 month master at MIT and four year curriculum of engineering studies at École polytechnique. In a first place, it seems appropriate for me to thank providence for the experience I have been allowed to live during my stay here in the US and at MIT. As my fellow labmate Jing once said, "one needs the time of a PhD to understand the whole research the MSEAS group is doing", hence the decision of joining a particular research group instead of another, or more generally any decision regarding one's own future orientation, can only *look* like a choice : it is only after experience that one may decide whether the "choice" was appropriate or not. These fifteen months have been a lot of excitement and they have largely broaden my perspectives: for that, I am bound to thank luck for having guided me to this rich environment and to the many persons who each in their own way contributed to change a bit of myself.

Coming to physical people, I am most grateful towards Pierre. I thank him for having patiently beared all my questions and doubts and for having helped me to address them by providing crucial ideas and hints at the right moments, and for his careful revisions and corrections when writing this work. Beyond his qualities as a research adviser, Pierre always shows a great concern towards the well-being of his students. Maybe most importantly, I felt he constantly tries hard to adapt to each of them so as to allow them to work according to their personal sensibilities and to develop their best potential. Thank you for having guided me and enabled me to achieve this work !

I thank then my lab-mates for the every day life at the office and for welcoming my "frenchness", especially Jon, Arkopal, Jing, Chinmay, Sydney, Deepak, Johnathan a.k.a. *Jvo*, and Corbin towards I feel especially indebted for a few piano lessons and rewarding moments playing Schubert's Military March. I thank as well the other people of the MSEAS group, for all what I learned by working in a team consituted by people with such different backgrounds: Matt, Wael, Yukino, Sudip, Chris, Pat, Marcia, and the legendary Tapovan Lolla. I address then my thoughts to all the people that have contributed to enrich my multiple lives at MIT. Two doors from my office, thank you Saviz for all the passionate research discussions that have heavily contributed to this work, for the time on bike, at the pool, in the mountains, and for trying to convince me to stay in the U.S. by the multiple excursions at *Dim Sum* restaurant. Among our common very closed circle of friends from the Geneva area, thank you Salman for all the laughs, the meringue, and the discussions about the Grassmanian. Many thanks to Parnika and the hours spent on practicing Hindi and French and discussing whether the english "t" is retroflex or dental. A thought for Shubhayu, whom I have been really pleased to meet and be impressed by his musical and drawing skills. Many thanks to Viral and Dragosh for the awesome week-end trips in the mountains, playing board game or cooking international *recipes*. I want also to mention Sophie, Virgile, Thomas, Pablo, Victor and Malivai for the nice moments together in the company of people nostalgic of good bread, and Urmi, Alex, Rishon, and Boyan, for the ties created from class works. I also have a thought for all friends who supported me from "distance", or while reunioning during the summer: Maxence, Benjamin, Nicole, Mathieu,

Camille, Christophe. Among the other persons towards whom I feel grateful for dedicating their energy to me, I thank Prof. R. Delacy from Harvard University for welcoming me in his Hindi class and helping me in my travel plans in India, and Annie for taking care of my spine and being a great Yoga teacher.

I feel grateful towards MIT for providing an amazing international work environment filled with people united by their passion for scientific research. I feel as well deeply indebted towards École polytechnique for having offered me four incredible years where every moment seemed to be the most important period of the curriculum, and for providing me the appropriate background for my stay at MIT. In particular I thank Grégoire Allaire, with whom discussions and prior works provided me a few important ideas and relevant references for the work presented hereafter.

With the MSEAS group, I am grateful to the Office of Naval Research for support under grants N00014-14-1-0476 (Science of Autonomy LEARNS) and N00014-14-1-0725 (Bays-DA), and to the National Science Foundation for support under grants EAR-1520825 (NSF-ALPHA), to the Massachusetts Institute of Technology.

Un grand merci à ma mère, mon père et sa femme, Nadine, mes frères Kévin et Johann ainsi que tous les autres membres de ma famille, qui me ramènent à mes racines et avec lesquels les contacts réguliers ont été un soutien grandement apprécié loin de mon pays natal.

Contents

Notations

General notations relative to Riemannian geometry

E	Finite dimensional space
E^*	Dual space of E
$< \cdot, \cdot >$	Scalar product or duality bracket on E
$\|\| \cdot \|\| =< \cdot, \cdot >^{\frac{1}{2}}$	Euclidean norm on E
$\mathrm{Span}(A)$	Image space of a linear operator A
$\mathrm{Ker}(A)$	Kernel of a linear operator A
$\mathscr{M}$	Smooth manifold $\mathscr{M} \subset E$ embedded in E
$\mathcal{T}(R)$	Tangent space at $R \in \mathscr{M}$
$\mathcal{N}(R)$	Normal space at $R \in \mathscr{M}$
$\Pi_{\mathcal{T}(R)}$	Orthogonal projection onto the tangent space $\mathcal{T}(R)$
$\Pi_{\mathscr{M}}$	Orthogonal projection onto $\mathscr{M}$
$\overline{\Omega}$	Closure of a set $\Omega \subset E$
$\partial\mathscr{M}$	Boundary $\partial\mathscr{M} = \overline{\mathscr{M}} \backslash \mathscr{M}$ of $\mathscr{M}$
$\mathrm{Sk}(\mathscr{M})$	Skeleton of $\mathscr{M}$
$\mathfrak{R}$	Point $\mathfrak{R} \in E$ of the ambient space
$\mathfrak{X}$	Vector $\mathfrak{X} \in E$ of the ambient space attached to some point $\mathfrak{R} \in E$
R	Point $R \in \mathscr{M}$ of the manifold
X	Vector $X \in \mathcal{T}(R)$ tangent to the manifold at some point $R \in \mathscr{M}$
$R(t)$	Smooth curve $R(t) \in \mathscr{M}$ drawn on $\mathscr{M}$ and defined on an open interval around the initial time $t = 0$
$\dot{R} = \mathrm{d}R/\mathrm{d}t$	Time derivative of a trajectory $R(t)$
$\exp_R(tX)$	Geodesic curve on $\mathscr{M}$ starting from $R \in \mathscr{M}$ in the tangent direction $X \in \mathcal{T}(R)$
N	Vector $N \in \mathcal{N}(R)$ in the normal space at $R \in \mathscr{M}$
$\kappa_i(N)$	Principal curvatures in the direction N
$\kappa_\infty(R)$	Maximal curvature at the point $R \in \mathscr{M}$
I	Identity mapping
$\mathrm{D}_X f(R)$	Differential of a function f defined on a manifold $\mathscr{M} \subset E$
$\mathrm{D}\Pi_{T(R)}(X) \cdot Y$	Differential of the projection operator $\Pi_{T(R)}$ applied to Y

Some attention must be given to the notation used for the differentials. The differential of a smooth function f at the point R belonging to some manifold $\mathscr{M}$ (this includes possibly $\mathscr{M} = E$) in the direction $X \in T(R)$ is denoted $\mathrm{D}_X f(R)$:

$$\mathrm{D}_X f(R) = \frac{\mathrm{d}}{\mathrm{d}t} f(R(t))\Big|_{t=0} = \lim_{\Delta t \to 0} \frac{f(R(t+\Delta t)) - f(R(t))}{\Delta t},$$

where $R(t)$ is a curve of $\mathscr{M}$ such that $R(0) = R$ and $\dot{R}(0) = X$. The differential of the orthogonal projection operator $R \mapsto \Pi_{T(R)}$ at $R \in \mathscr{M}$, in the direction $X \in T(R)$ and applied to $Y \in \mathcal{M}_{l,m}$ is denoted $\mathrm{D}\Pi_{T(R)}(X) \cdot Y$:

$$\mathrm{D}\Pi_{T(R)}(X) \cdot Y = \left[\frac{\mathrm{d}}{\mathrm{d}t} \Pi_{\mathcal{T}(R(t))} \Big|_{t=0} \right] (Y) = \left[\lim_{\Delta t \to 0} \frac{\Pi_{\mathcal{T}(R(t+\Delta t))} - \Pi_{\mathcal{T}(R(t))}}{\Delta t} \right] (Y).$$

Notations used in the context of matrix spaces

$\mathcal{M}_{l,m}$	Space of l-by-m real matrices
Sym_n	Space of n-by-n symmetric matrices
$\mathcal{M}^*_{m,r}$	Space of m-by-r matrices that have full rank
$\mathrm{rank}(R)$	Rank of a matrix $R \in \mathcal{M}_{l,m}$
$\mathscr{M} = \{R \in \mathcal{M}_{l,m} \mid \mathrm{rank}(R) = r\}$	Fixed rank matrix manifold
$\mathrm{St}_{n,p} = \{U \in \mathcal{M}_{n,p} \mid U^T U = I\}$	Stiefel Manifold of n-by-p matrices
$\mathcal{O}_n = \{P \in \mathcal{M}_{n,n} \mid P^T P = I\}$	Group of r-by-r orthogonal matrices
$\mathscr{I}$	Isospectral manifold
$\mathcal{G}$	Grassman manifold
A^T	Transpose of a matrix A
$< A, B >= \mathrm{Tr}(A^T B)$	Frobenius matrix scalar product
$\|\|A\|\| = \mathrm{Tr}(A^T A)^{1/2}$	Frobenius norm
$\sigma_1(A) \geq \ldots \geq \sigma_{\mathrm{rank}(A)}(A)$	Non zeros singular values of a matrix A
$\lambda_i(A)$	Eigenvalues of a real matrix A
$\Re(\lambda_i(A))$	Real parts of the eigenvalues of a matrix A
$\mathrm{Span}(U)$	Vector space spanned by the columns of the matrix U
$\mathrm{sym}(\mathfrak{X}) = (\mathfrak{X} + \mathfrak{X}^T)/2$	Symmetric part of a square matrix $\mathfrak{X}$
$\mathrm{skew}(\mathfrak{X}) = (\mathfrak{X} - \mathfrak{X}^T)/2$	Skew-symmetric part of a square matrix $\mathfrak{X}$
δ_{ij}	Kronecker symbol. $\delta_{ii} = 1$ and $\delta_{ij} = 0$ for $i \neq j$

Introduction

The motivation that constitutes the guideline of this work is the need for novel fundamental theory and computational schemes that rigorously integrate uncertainty in Lagrangian predictions, that is in the analysis of material transport associated with dynamic time and space dependent uncertain velocity fields. Such concern is typically encountered in ocean and weather forecasting for which predictions often include some amount of uncertainty [94, 92]. The first results that integrate stochastic data-assimilative ocean modeling with Lagrangian predictions for stochastic Lagrangian Coherent Structures schemes were obtained for the Monterey Bay region [95]. Accurately and efficiently estimating how material transport is organized in such uncertain environments however still corresponds to fundamental and computational challenges that must for example be addressed when developing hazard response capabilities [105] or path planning of autonomous naval systems [96, 140].

Advection and Lagrangian Coherent Structures We start in chapter 1 with a review of the state of the art techniques available for analyzing material transport in non-autonomous dynamical systems. This involves notably the theory of Lagrangian Coherent Structures (LCS) [66, 51], that seeks to extract and provide tools for visualizing the relevant features of the flow map of the particle ODE (Ordinary Differential Equation)

$$\begin{cases} \dot{\boldsymbol{x}} = \boldsymbol{v}(t, \boldsymbol{x}) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0. \end{cases} \tag{1}$$

We examine and compare two classes of methods for defining and extracting relevant surfaces or sub-regions of the flow that exhibit material coherence. These two classes split between those which study the flow map of the ODE (1) directly (*diffeomorphism based methods*) versus those which consider its global action on scalar functions (*operator based methods*). This action is given by the resolvent operator of the advection equation

$$\begin{cases} \partial_t \boldsymbol{\psi} + \boldsymbol{v}(t, \boldsymbol{x}) \cdot \nabla \boldsymbol{\psi} = 0 \\ \boldsymbol{\psi}(0, \boldsymbol{x}) = \boldsymbol{\psi}_0(\boldsymbol{x}), \end{cases} \tag{2}$$

that maps an initial data $\boldsymbol{\psi}_0(\boldsymbol{x})$ to the advected data $\boldsymbol{\psi}(t, \boldsymbol{x})$ at time t. We discuss the efficient and pragmatic implementation of these methods for realistic ocean velocity data using both the Lagrangian (1) and the Eulerian (2) points of view. We also propose some new criteria for defining LCSs, as well as some reformulations of the existing theories. We compare these various techniques on three benchmark numerical examples including a highly resolved set of realistic ocean velocities.

Model order reduction for stochastic advection We then focus on devising methodologies that would adapt and extend these tools to uncertain velocity fields. For such cases, one needs to estimate the statistics of an ensemble of realizations of the flow maps of the ODE (1) where $\boldsymbol{v}(t, \boldsymbol{x}; \omega)$ becomes a stochastic variable depending on a random event ω, or equivalently, of the solution $\boldsymbol{\psi}(t, \boldsymbol{x}; \omega)$ of the stochastic advection equation

$$\begin{cases} \partial_t \boldsymbol{\psi} + \boldsymbol{v}(t, \boldsymbol{x}; \omega) \cdot \nabla \boldsymbol{\psi} = 0 \\ \quad \boldsymbol{\psi}(0, \boldsymbol{x}; \omega) = \boldsymbol{\psi}_0(\boldsymbol{x}; \omega). \end{cases} \tag{3}$$

This PDE formulation is not necessary more advantageous for evaluating material transport in the case of a single deterministic simulation [27], but one of its exceptional feature is that it transforms the nonlinear dynamics of the ODE (1) into a linear process. We will see later on that it makes (3) especially suited for applying dynamical model order reduction methods that avoid brute-force Monte-Carlo simulations by taking advantage of the low-rank structure of the stochastic solutions.

Equation (3) belongs to the more general class of stochastic PDEs of the form

$$\partial_t \boldsymbol{u} = \mathscr{L}(t, \boldsymbol{u}; \omega)\,, \tag{4}$$

where t is time, $\boldsymbol{u}$ the uncertain dynamical fields and $\mathscr{L}$ a differential operator. Finding efficient reduced dynamical systems for such stochastic PDEs is an issue commonly encountered in a wide variety of domains involving intensive computations and expensive high-fidelity simulations [126, 116, 84, 26]. For deterministic but parametric dynamical systems, ω may also represent a large set of possible parameter values that need to be accounted for by the model-order reduction. In general, reduced order methods seek for a dynamic approximation $\boldsymbol{u}_{\mathrm{DO}}$ of the solution $\boldsymbol{u}$ that can decompose onto a finite number of r spatial modes, $\boldsymbol{u}_i(t, x)$, and stochastic coefficients, $\zeta_i(t, \omega)$,

$$\boldsymbol{u}(t, \boldsymbol{x}; \omega) \simeq \boldsymbol{u}_{\mathrm{DO}} = \sum_{k=0}^{r} \zeta_i(t, \omega)\, \boldsymbol{u}_i(t, \boldsymbol{x}). \tag{5}$$

The existence of such approximation is motivated by the Karuhnen-Loève (KL) decomposition [99, 73], for which selecting the first r modes yields an optimal orthonormal basis $(\boldsymbol{u}_i)$. Many methods have been proposed to evolve either these modes or coefficients, but not both, the most popular being polynomial chaos expansions [154], Fourier decomposition [151], or Proper Orthogonal Decomposition [73]. In order to solve (3), we consider the Dynamically Orthogonal (DO) approximation introduced in 2009 for stochastic PDEs (4) by Sapsis and Lermusiaux [124]. In contrast with the previous methods, the DO method does not assume anything more than the dependence with respect to time of the modes $\boldsymbol{u}_i(t, \boldsymbol{x})$ and coefficients $\zeta_i(t; \omega)$. The reduced model is a coupled system of PDE that seeks to most accurately update the approximation (5) :

$$\begin{cases} \partial_t \zeta_i = < \mathscr{L}(t, \boldsymbol{u}_{\mathrm{DO}}; \omega), \boldsymbol{u}_i > \\ \displaystyle\sum_{j=1}^{r} \mathbb{E}[\zeta_i \zeta_j] \partial_t \boldsymbol{u}_j = \mathbb{E}\left[\zeta_i \left(\mathscr{L}(t, \boldsymbol{u}_{\mathrm{DO}}; \omega) - \sum_{j=1}^{r} < \mathscr{L}(t, \boldsymbol{u}_{\mathrm{DO}}; \omega), \boldsymbol{u}_j > \boldsymbol{u}_j \right) \right] , \end{cases} \tag{6}$$

$\mathbb{E}$ being the expectation operator and $<, >$ the standard scalar product over L^2 functions (an integral over the spatial domain). Numerical schemes were developed by Ueckermann et. al.

[147] for these PDEs that allowed to obtain efficient simulations of stochastic Navier-Stokes equations. Nevertheless, a gap remains today regarding, (i) the rigorous understanding of the error committed by this approximation in a general framework, (ii) the selection of appropriate numerical schemes in these methods for problems dominated by advection. In general, the discretization of the advection term $-\boldsymbol{v} \cdot \nabla \boldsymbol{\psi}$ in (2) requires particular care to obtain stability and accuracy [108], and it is not a priori clear how up-winding rules and other nonlinear schemes can be efficiently adapted to the stochastic case.

Riemannian geometry of matrix manifolds and low-rank methods In order to develop rigorous, stable and accurate schemes for (3), we investigated the foundations of low-rank model order reduction methods for stochastic PDEs. After discretization of (4) with respectively l and m spatial and stochastic degrees of freedoms, one is interested in the numerical solution of a large system of ODEs of the form

$$\dot{\mathfrak{R}} = \mathcal{L}(t, \mathfrak{R}), \tag{7}$$

where $\mathcal{L}$ is an operator acting on the space $\mathcal{M}_{l,m}$ of l-by-m matrices $\mathfrak{R}$. As noticed by the recent work of Musharbash in [103], it turns out that in this discrete setting, the DO method coincides with the "dynamical low-rank" approximation introduced in 2007 by Koch and Lubich [83, 103]. The decomposition (5) is written under the form of rank r approximation $\mathfrak{R} \simeq UZ^T$ where U is a l-by-r matrix containing the discretization of the basis functions $(\boldsymbol{u}_i)$, and Z is a m-by-r matrix containing the realizations of the stochastic coefficients (ζ_i).

Any model order reduction method attempts to evolve a point $R = UZ^T$ in the subset

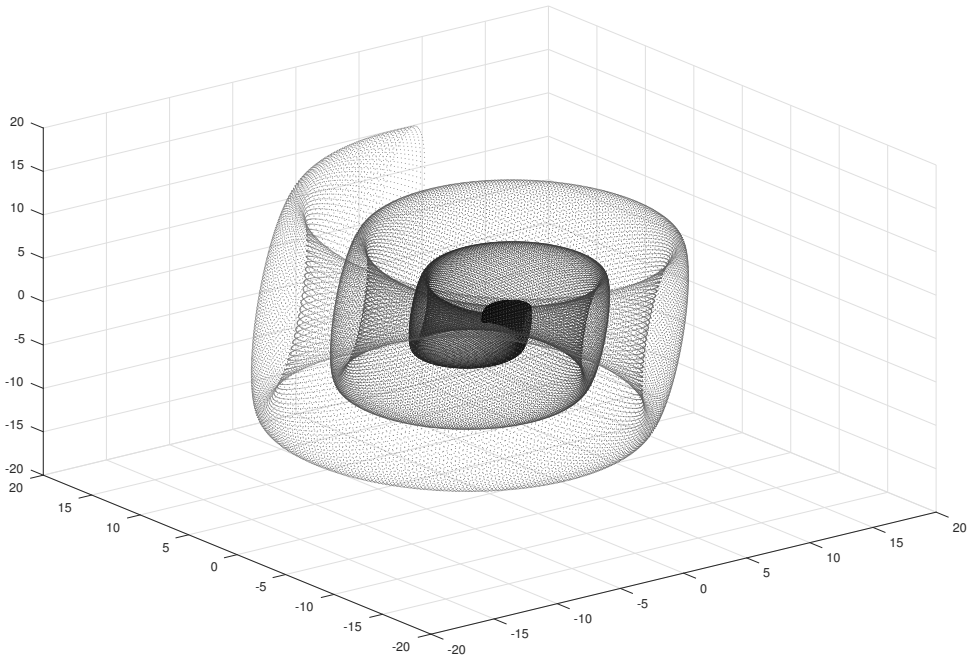
$$\mathscr{M} = \{R \in \mathcal{M}_{l,m} | \text{rank}(R) = r\}$$

constituted by rank r matrices. As already exploited in the optimization community [4], a crucial feature is that $\mathscr{M}$ has a *smooth* shape in the space $\mathcal{M}_{l,m}$ that gives to this set a lot of structure: in mathematical jargon, $\mathscr{M}$ *is a smooth manifold*. To provide a geometric intuition, a 3D projection of two 2-dimensional subsurfaces of the set of rank one 2-by-2 matrices (a three dimensional manifold embedded in a four dimensional space) has been plotted on Figure 1. This figure has been obtained by using the parameterization

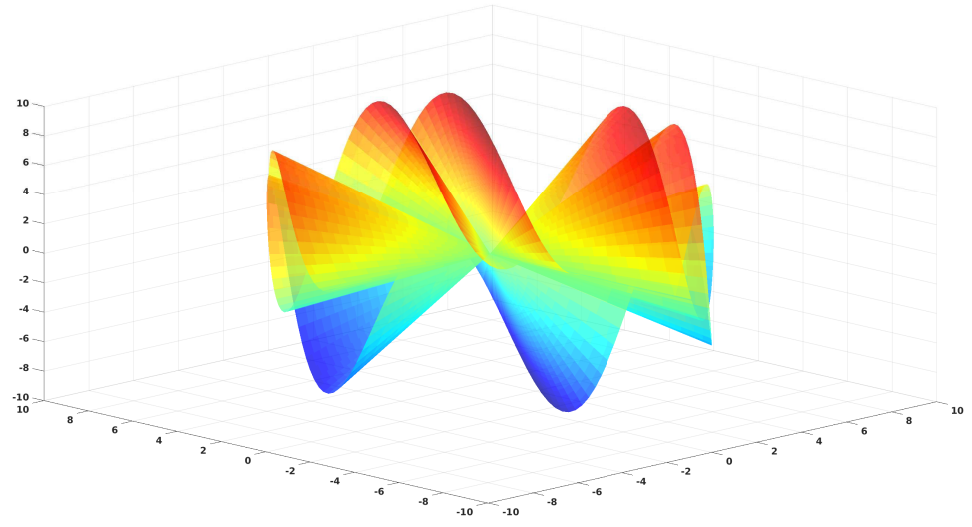
$$R(\rho, \theta, \boldsymbol{\phi}) = \rho \begin{pmatrix} \sin(\theta)\sin(\boldsymbol{\phi}) & \sin(\theta)\cos(\boldsymbol{\phi}) \\ \cos(\theta)\sin(\boldsymbol{\phi}) & \cos(\theta)\cos(\boldsymbol{\phi}) \end{pmatrix}, \ \rho > 0, \theta \in [0, 2\pi], \ \boldsymbol{\phi} \in [0, 2\pi],$$

and projecting orthogonally two subsurfaces by plotting the first three elements R_{11}, R_{12} and R_{21}. We will prove in chapter 2 that the maximal curvature of $\mathscr{M}$ is proportional to the inverse of the lowest singular value, which is consistent with the spiraling shape visible on Figure 1a near the origin (note that in this example, the smallest singular value is $\sigma_2(R) = \rho$). Simultaneously, $\mathscr{M}$ is the union of r-dimensional affine subspaces of $\mathcal{M}_{l,m}$ supported by the manifold of strictly lower rank matrices, which is to some extent illustrated on Figure 1b.

Geometrically, a dynamical system (7) can be seen as a time dependent vector field $\mathcal{L}(t, \cdot)$ that assigns the velocity $\mathcal{L}(t, \mathfrak{R})$ at time t at each point $\mathfrak{R}$ of the ambient space $\mathcal{M}_{l,m}$ of l-by-m matrices (this is illustrated on Figure 2). Similarly, any rank r model order reduction can be viewed as a vector field $L(t, \cdot)$ that must be everywhere tangent to the



(a) $R(\rho, \pi\rho, \boldsymbol{\phi})$



(b) $R(\rho, \boldsymbol{\phi}, \rho - \boldsymbol{\phi})$

Figure 1: Two subsurfaces of the rank-1 manifold $\mathscr{M}$ of 2 by 2 matrices. R is defined in the text.

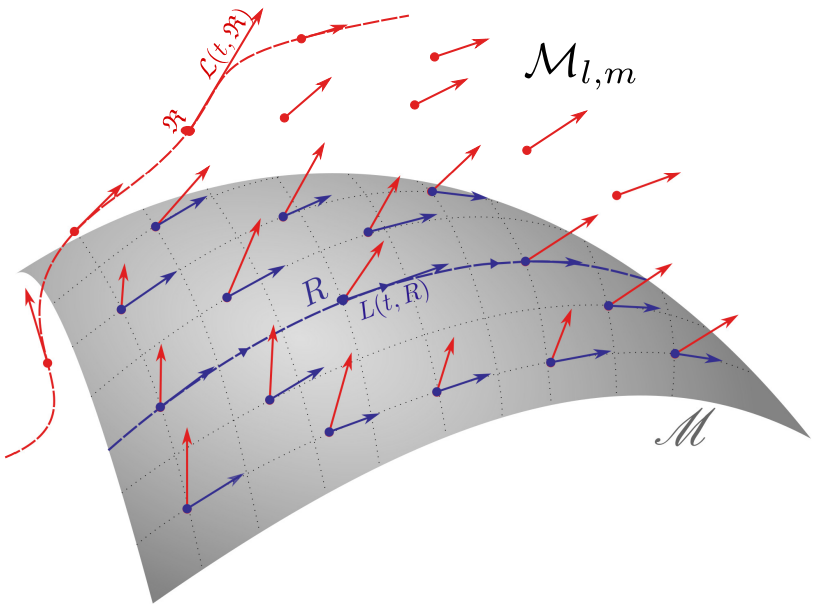


Figure 2: Dynamical systems as vector fields $\mathcal{L}$ in the ambient space $\mathcal{M}_{l,m}$ (in *red*), or as vector fields L tangent to the manifold $\mathscr{M}$ (in *blue*). The DO approximation corresponds to the dynamical system $L(t,\cdot)$ obtained by the orthogonal projection of the vector field $\mathcal{L}(t,\cdot)$ onto the local tangents to the manifold $\mathscr{M}$, i.e. the "tangent projection" in short.

manifold $\mathscr{M}$ of rank r matrices. The corresponding dynamical system is

$$\dot{R} = L(t,R) \in \mathcal{T}(R), \tag{8}$$

where $\mathcal{T}(R)$ denotes the tangent space of $\mathscr{M}$ at R. From this point of view, the DO or dynamical low rank approximation is obtained by "combing the hair" formed by the original vector field $\mathcal{L}$ on the manifold $\mathscr{M}$, that is by setting $L(t,R) = \Pi_{\mathcal{T}(R)}(\mathcal{L}(t,\mathfrak{R}))$ to be the time-dependent orthogonal projection of each vector $\mathfrak{X} = \mathcal{L}(t,R)$ onto the tangent space $\mathcal{T}(R)$. Making the projection $\Pi_{\mathcal{T}(R)}$ explicit, the DO solution $R(t) = U(t)Z(t)^T$ can be obtained by solving

$$\left\{ \begin{array}{ll} \dot{Z} = & \mathcal{L}(t,UZ^T)^T U \\ \dot{U} = & (I - UU^T)\mathcal{L}(t,UZ^T)Z(Z^TZ)^{-1}, \end{array} \right. \tag{9}$$

which is equivalent to the system given by Koch and Lubich in [83] and a discrete version of the DO equations (6) from Sapsis and Lermusiaux in [124]. Such relationships between matrix dynamical systems that learn subspaces and stochastic subspace predictions and data assimilation where also investigated by Lermusiaux [89] using ideas and methods from Brockett et al [20, 21].

Curvature of matrix manifolds, differentiability of the truncated SVD and Polar decomposition, Projected dynamical systems Riemannian geometry of the fixed rank manifold is therefore an adapted framework for understanding low-rank methods. In chapter 2 we review and reformulate some background material about the Riemannian geometry of embedded smooth manifolds in a formalism free of tensor notations. This

allows to conveniently analyze the embedded geometry of popularly encountered matrix manifolds, namely the Stiefel manifold (that includes the group of orthogonal matrices), the isospectral manifold (that includes the Grassman manifold), and the fixed rank manifold, the latter being the object of study in model order reduction. We obtain explicit formulas for the Weingarten map and the extrinsic or principal curvatures on these manifolds with respect to a given normal direction [139, 36, 11]. Combined with a result relating these curvatures to the differential of the orthogonal projection onto a given manifold $\mathscr{M}$ (the application mapping a point of the ambient space to the closest point on $\mathscr{M}$), we derive expressions for the differential of algebraic operations such as truncated SVD (Singular Value Decomposition), Polar Decomposition and of the projectors over the eigenspaces of a symmetric matrix. We extend the methodology to the case of a non Euclidean ambient space, which allows to apply a similar methodology to find time derivative of the eigenspaces of non symmetric matrices and to derive dynamical systems that compute them.

In this geometric framework, the DO approximation becomes a projected dynamical system on a manifold $\mathscr{M}$ embedded in an euclidean space E in the particular case where $\mathscr{M}$ is the fixed rank manifold and E the space $\mathcal{M}_{l,m}$ of l-by-m matrices. We provide an analysis of such projected dynamical system in a general setting and we show that the error committed remains controlled under the geometric condition that the non-reduced solution remains close to the low-rank manifold $\mathscr{M}$, or more precisely under the condition that it does not cross the skeleton of $\mathscr{M}$, which is also equivalent to the condition that the orthogonal projection of this solution onto $\mathscr{M}$ remains differentiable. Applied to the DO methodology later on in chapter 3, we find that the error of the DO approximation (9) is explicitly related to the gap $\sigma_r(\mathfrak{R}) - \sigma_{r+1}(\mathfrak{R})$ between the singular values of order r and $r+1$ of the original solution $\mathfrak{R}$, which is an improvement over the initial error analysis of Koch and Lubich in [83] and Musharbash in [103].

DO numerical schemes for stochastic advection Having at our disposal the geometric analysis, we investigate in chapter 3 the implementation in practice of the DO method for the stochastic advection equation with random velocity field (3). To our knowledge, the first results that coupled stochastic data-assimilative ocean modeling and Lagrangian Coherent Structures schemes were obtained by Lermusiaux and Lekien [95]. The Error Subspace Statistical Estimation scheme was used for uncertainty forecasting and data assimilation, and its ensemble of velocity fields was employed to compute the corresponding LCS uncertainties. LCS statistics were then studied for three different dynamical regimes and periods in the Monterey Bay region [94, 92]. Presently, we develop new theory and numerical schemes for such stochastic advection and Lagrangian transport studies, using the DO equations which allow much larger ensemble sizes and the novel geometrical methods obtained above which allow more robust and accurate integrations.

Deterministic-stochastic consistent numerical schemes are obtained by discretizing *first* in space before applying the DO method in the matrix setting, rather than looking for the discretization of the continuous DO equations as in [147]. For the method to be efficient, accurate, and directly consistent with deterministic realizations, we find that fully linear high order schemes that include some amount of artificial diffusion are appropriate. Various strategies are presented for selecting a time-stepping that accounts for the curvature of the fixed-rank manifold and the error related to closely singular coefficient matrices. Exploiting the relation between the DO method and the Singular Value Decomposition, we show that Riemannian optimization onto the fixed rank manifold can be integrated in the time

marching scheme to (i) improve the accuracy of the DO time-stepping by accounting for the curvature of the manifold and (ii) update at will the rank of the reduced solution. Finally, we demonstrate the applicability of the method on two numerical examples and provide comparisons with Monte-Carlo simulations.

We become able to efficiently estimate a truncated KL expansion of a stochastic flow map and hence evaluate its statistics such as mean position or standard deviation from the mean position of a particle.

This work led to the preparation of two preprint articles: most of the content of chapter 2 and chapter 3 are intended to appear in [44] and [45] respectively.

Chapter 1

Numerical methods for Lagrangian Coherent Structures and material transport analysis

1.1 Introduction

1.1.1 Advection and material transport

Advection plays a major role in a wide variety of physical processes and engineering applications of fluid mechanics [73, 14], neutronic transport, chemical transports, atmospheric sciences [121] and ocean sciences [60, 106]. At its most fundamental level, the pure advection process is understood through the transport partial differential equation (PDE),

$$\begin{cases} (\partial_t + \boldsymbol{v}(t,\boldsymbol{x}) \cdot \nabla)\boldsymbol{\psi} = 0 \\ \qquad\qquad \boldsymbol{\psi}(0,\boldsymbol{x}) = \boldsymbol{\psi}_0(\boldsymbol{x}), \end{cases} \tag{1.1}$$

that models the material transport of a passive (scalar or vectorial) tracer field $\boldsymbol{\psi}$ under a velocity field $\boldsymbol{v}$, having initially its values distributed as $\boldsymbol{\psi}_0$ over a physical domain of positions $\boldsymbol{x}$. This property is found by relating (1.1) to the solutions of the ordinary differential equation (ODE)

$$\begin{cases} \dot{\boldsymbol{x}} = \boldsymbol{v}(t,\boldsymbol{x}) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0, \end{cases} \tag{1.2}$$

where $\boldsymbol{x}(t)$ is physically understood as the position at time t of a particle initially located at $\boldsymbol{x}_0$ and whose instantaneous velocity is $\boldsymbol{v}(t,\boldsymbol{x}(t))$. If $\boldsymbol{\phi}_0^t$ is the flow map of this ODE, mapping initial positions $\boldsymbol{x}_0$ to those $\boldsymbol{\phi}_0^t(\boldsymbol{x}_0) = \boldsymbol{x}(t)$ at time t, then under sufficient regularity conditions on the velocity field $\boldsymbol{v}$ [34, 12], the solution $\boldsymbol{\psi}$ of the advection eqn. (1.1) is obtained by "carrying $\boldsymbol{\psi}_0$ values along particles' paths":

$$\boldsymbol{\psi}(t,\boldsymbol{x}) = \boldsymbol{\psi}_0(\boldsymbol{\phi}_0^{-t}(\boldsymbol{x})), \tag{1.3}$$

where $\boldsymbol{\phi}_0^{-t} = (\boldsymbol{\phi}_0^t)^{-1}$ is the backward or inverse flow map (Figure 1-1). This is related to the *renormalization* property of the "physical" solutions of this PDE, namely if ρ is a solution, then $b(\rho)$ is a solution for any function b with some regularity assumptions [17] (another existence and uniqueness mathematical framework to obtain solutions being the

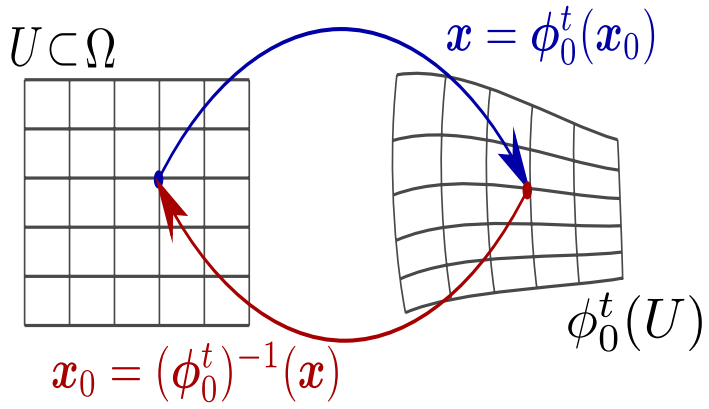


Figure 1-1: Illustration of the action of the forward and backward flow map on a subdomain $U \subset \Omega$ of the physical domain Ω. $\boldsymbol{\phi}_0^t$ maps initial particle positions $\boldsymbol{x}_0$ to their position at time t, and $(\boldsymbol{\phi}_0^t)^{-1}$ is the reciprocal map.

theory of viscosity solutions, see [29]). In fact, (1.1) and (1.2) are equivalent mathematical descriptions of material transport, as setting $\boldsymbol{\psi}_0(\boldsymbol{x}) = \boldsymbol{x}$ in (1.3) yields $\boldsymbol{\psi}(t, \boldsymbol{x}) = (\boldsymbol{\phi}_0^t)^{-1}(\boldsymbol{x})$. Similarly, solving backward in time the transport equation with a terminal condition,

$$\begin{cases} (\partial_s + \boldsymbol{v}(s, \boldsymbol{x}) \cdot \nabla)\boldsymbol{\rho} = 0 \\ \qquad\qquad \boldsymbol{\rho}(t, \boldsymbol{x}) = \boldsymbol{\rho}^t(\boldsymbol{x}), \end{cases} \tag{1.4}$$

allows to retrieve the forward flow map from the relation $\boldsymbol{\rho}(s, \boldsymbol{x}) = \boldsymbol{\rho}^t(\boldsymbol{\phi}_s^t(\boldsymbol{x}))$ by setting $\boldsymbol{\rho}^t(\boldsymbol{x}) = \boldsymbol{x}$. This shows that the flow map $\boldsymbol{\phi}_0^t$ can be obtained from a solution of the transport PDE (1.1) and vice versa. This property has been thoroughly investigated on the theoretical side to give a mathematical meaning to the solutions of the ODE (1.2) for velocity fields $\boldsymbol{v}$ with weak regularity [34, 12, 17], and more recently in numerical computations, as it offers an alternative method to direct particle advection for the evaluation of the flow map $\boldsymbol{\phi}_0^t$ [97, 98].

1.1.2 Lagrangian Coherent Structures

Material transport cannot be understood directly from the direct observation of the velocity field [66, 129], which must be integrated over a time-window. The examples to follow will provide a clear illustration that particle trajectories advected under an even rather "simple" time-dependent velocity field are not trivial. The concept of Lagrangian Coherent Structures (LCS) has emerged recently [71] to improve understanding of material transport in time-dependent fluid flows. Coherent structures is a notion initially derived from the observation of turbulent fluid flows and used to refer to persistent *eulerian* large scale patterns visible in the velocity fields. These patterns preserve their kinetic energy and allow for energy dissipation towards smaller scales in the Kolmogorov energy cascade [42]. *Lagrangian* Coherent Structures deal with the observation of the material flow itself, and therefore refers to the persistence of material sub-domains in the flow [39, 68, 66]. LCS are expected to allow for improved Lagrangian hazard predictions and prevention, typical applications being pollution tracking in oceans [88, 28] and other environmental flow hazards [1].

To date, several definitions of LCS, that do not fully coincide, have been proposed by different authors [39, 53, 66, 81, 107, 112, 129, 142] and, there are as many computational

methodologies to extract them from time-dependent (*non-autonomous*) velocity fields. In this chapter, we review the concept of LCS and the two dominant methods that have emerged to compute them, and their pragmatic applicability to realistic, highly-resolved and multi-scale ocean velocity data (note that a recent comparison of these methods is available on [9]). A time dependent velocity field $\boldsymbol{v}(t, \boldsymbol{x})$ is assumed to be given, which is not necessary a solution of the Navier-Stokes equations. Incompressibility, *i.e.* $\mathrm{div}(\boldsymbol{v}) = 0$, is often assumed but it is also not a requirement.

With the flow map $\boldsymbol{\phi}_0^t$ or equivalently the PDE (1.1) giving all the information about material transport, one could say everything needed is there and stop: for example, level set methods [108] or the direct use of the flow map function can be used to track the evolution of a domain spanned by a pollutant. Nevertheless, the information contained in the flow map or the operator (1.6) is not convenient to display: this is because the flow map is a function (a diffeomorphism) mapping a domain $\Omega \subset \mathbb{R}^n$ into itself, with the spatial dimension n being in general $n = 2$ or 3. Defining and computing LCS is therefore all about finding adequate visualizations of the flow map, that allow to extract the ominant structures of material transport through an "understandable" picture, for example by plotting a scalar function $LCS : \mathbb{R}^n \to \mathbb{R}$ that will contain an information that best represent the one contained in the flow map. This scalar function can be a field whose particular level-sets which indicate coherent sub-domains. It can also be an indicator function for the location of barriers sub-manifolds. Since the information contained in the vector mapping

$$\boldsymbol{\phi}_0^t : \mathbb{R}^n \to \mathbb{R}^n \tag{1.5}$$

can hardly be reproduced without loss in a simpler representation $LCS : \mathbb{R}^n \to \mathbb{R}$, several complementary techniques may be used to obtain a picture of how material transport is organized.

One sees therefore that computing LCS in practice is requires two main ingredients : (i) an estimation of the flow map, obtained directly by advecting particles, or indirectly by computing solutions of the PDE (1.1) and (ii) an adequate visualization of the flow map to extract coherent structures. LCS approaches can mainly be classified into two types: those which attempt to plot relevant features of the flow map diffeomorphism $\boldsymbol{\phi}_0^t$, *versus* those which focus on relevant features of the functional operator (1.1)

$$f \mapsto f \circ \boldsymbol{\phi}_0^{-t}, \tag{1.6}$$

that maps initial densities $f_0 \in L^2(\Omega)$ to the corresponding solution of the advection equation (1.1) at time t. The first class of approaches considers rather the action induced by the flow map on individual particle trajectories, while the second studies its global action on density distributions.

In all these methods, some issues are recurrent when it comes to applying these these computational methodologies to realistic data: dependency of coherence with a particular space-scale or a finite time window $[0, t]$, presence of inlets and outlets [142], noise or approximations in data measurement [64].

1.1.3 Three benchmark numerical examples

In the following we review the available techniques and bring some suggestions to improve LCS estimations and computations. Keeping in mind pragmatic ocean applications for which velocity data are gridded and often highly resolved, we will apply and compare these

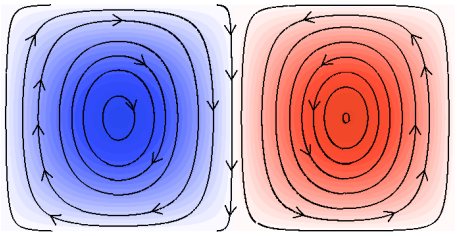


Figure 1-2: Streamlines and vorticity of the Double Gyre Flow at $t = 10$

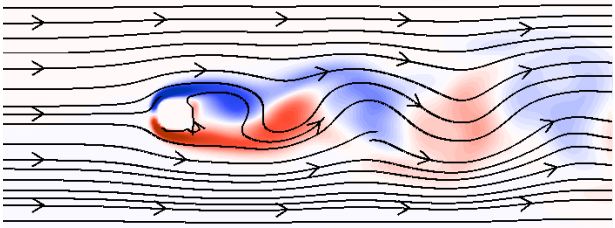


Figure 1-3: Streamlines and vorticity of the Flow Past Cylinder Flow at $t = 0$

on three benchmark examples: the analytic double-gyre flow that is a popular example of most of LCS works [129, 66, 97], the flow past a cylinder that is a typical example of Navier-Stokes fluid flow, and a realistic set of ocean velocity data over the region of the island of Palau.

Analytic double-gyre

The double gyre is a 2D benchmark example for studying Lagrangian coherence of particle motions [129, 97, 66]. This flow is constituted by two vortices oscillating horizontally (Figure 1-2). We use the analytic expression of this flow proposed by Shadden et al. [129] :

$$\boldsymbol{v}(t, \boldsymbol{x}) = (-\partial_y \phi, \partial_x \phi) \text{ with } \phi(\boldsymbol{x}, t; \omega) = A \sin[\pi f(x,t)] \sin(\pi y), \tag{1.7}$$

where $f(x,t) = \epsilon \sin(\omega t)x^2 + (1 - 2\epsilon \sin(\omega t))x$ and $\boldsymbol{x} = (x, y)$. The 2D domain is $\Omega = [0, 2] \times [0, 1]$ and the values considered for the parameters are $A = 0.1$, $\epsilon = 0.1$ and $\omega = 2\pi/10$. We consider a 512x256 grid and the flow is integrated between $t = 0$ and $t = 15$.

Flow past a cylinder

The second example is a numerical simulation of a flow past a cylinder (Figure 1-3). The flow is set on a domain $\Omega = [0, 16] \times [0, 6]$ discretized with a 240×90 grid. The Reynolds number is Re=100. The cylinder is a disc of center $(x_c, y_c) = (4.5, 3)$ and of radius $R = 0.5$. The flow enters at the left side on the domain with a velocity $\boldsymbol{v} = (1, 0)$. Neumann boundary conditions are considered at the top and bottom walls, while the second normal derivative is set to $\partial^2 \boldsymbol{v}/\partial n^2 = 0$ at the right bottom wall. We consider a time window $t \in [0, 10]$ over which the periodic regime is established.

Realistic velocity data over the Fleat Palau region

The third example is a realistic velocity field (Figure 1-4) over the Fleat Palau region, obtained from an oceanic prediction of the MSEAS data assimilative PE model [63] for this

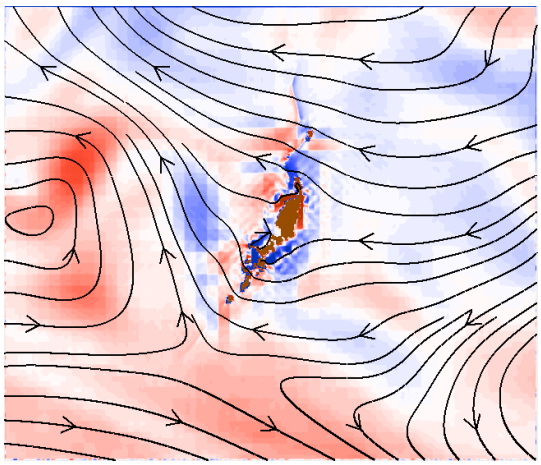


Figure 1-4: Streamlines and vorticity around the Palau region at $t = 0$

region. The resolution of the domain is 853x728 for a duration of 144 hours. This example is more challenging since (i) the resolution of this example is relatively high (ii) the domain includes inlets, outlets, and an inner obstacle with complex geometry, and (iii) the flow is only approximately divergence-free.

1.2 Diffeomorphism based LCS methods

Although most works belonging to this class of methods have been using terms from the theory of autonomous dynamical systems (such as "invariant", "attractive", "repelling" *manifolds*), most of these techniques find relevant features of the flow map in its "discontinuities" (sharp gradients) or irregularities: intuitively particles that are located from either side of a discontinuity surface of $\boldsymbol{\phi}_0^t$ have large diverging trajectories while particles located where the flow map is locally constant tend to reach nearby positions. Hence LCS are sought as material co-dimension 1 lines (in 2D) or surfaces (in 3D) that exhibit extremal properties of repulsion or attraction, ideally globally, or at least in the neighborhood of these surfaces. Several attempts have been made to identify such lines, mainly by Georges Häller [65, 69, 104, 128], but also [81, 129]. Another technique is the use of braids [10].

1.2.1 Extracting LCS from the SVD of the differential of the flow map

Several kinds of barriers have been defined, depending on how one identifies what is a codimension 1 "discontinuity" surface for the diffeomorphism $\boldsymbol{\phi}_0^t$. Irregularity can be quantified by the magnitude of the smallest Lipschitz constant L such that the inequality

$$||\boldsymbol{\phi}_0^t(\boldsymbol{x}_1) - \boldsymbol{\phi}_0^t(\boldsymbol{x}_2)|| \leq L||\boldsymbol{x}_1 - \boldsymbol{x}_2|| \tag{1.8}$$

is satisfied for $\boldsymbol{x}_2$ in the neighborhood of $\boldsymbol{x}_1$. Bounding this constant is one of the main interests of some works [12] that have aimed at defining the flow map from the transport PDE (1.1). In fact, for a regular flow map, inequality (1.8) becomes

$$||\mathrm{D}\boldsymbol{\phi}_0^t(\boldsymbol{x})\delta\boldsymbol{x}|| \leq L||\delta\boldsymbol{x}||,$$

where $\mathrm{D}\boldsymbol{\phi}_0^t$ is the differential of the flow map. In the following we will denote

$$\mathrm{D}\boldsymbol{\phi}_0^t(\boldsymbol{x}) = \sum_{i=1}^{n} \sigma_i^t(\boldsymbol{x})\boldsymbol{\eta}_i^t(\boldsymbol{x})\boldsymbol{\xi}_i^{tT}(\boldsymbol{x}), \quad \mathrm{D}\boldsymbol{\phi}_0^{-t}(\boldsymbol{y}) = \sum_{i=1}^{n} \sigma_i^{-t}(\boldsymbol{y})\boldsymbol{\eta}_i^{-t}(\boldsymbol{y})\boldsymbol{\xi}_i^{-tT}(\boldsymbol{y}), \tag{1.9}$$

the Singular Value Decomposition ([75]) of the differential $\mathrm{D}\boldsymbol{\phi}_0^t$ (resp. $\mathrm{D}\boldsymbol{\phi}_0^{-t}$) of the (*resp.*, backward) flow map at the position $\boldsymbol{x} \in \Omega \subset \mathbb{R}^n$ (*resp.* $\boldsymbol{y} \in \boldsymbol{\phi}_0^t(\Omega) \subset \mathbb{R}^n$), where singular values are assumed to be given in a decreasing order, (*i.e.* $\sigma_1^t(\boldsymbol{x}) \geq \sigma_2^t(\boldsymbol{x}) \geq \ldots \geq \sigma_n^t(\boldsymbol{x}) > 0$), and singular vectors are normalized ($||\boldsymbol{\xi}_i^s(\boldsymbol{x})|| = ||\boldsymbol{\eta}_i^s(\boldsymbol{x})|| = 1$). With these notations, the best constant L in equation (1.8) is the maximum singular value $\sigma_1^t(\boldsymbol{x})$, or as it is usually referred to in the LCS literature [66], the square root of the maximum eigenvalue of the Cauchy Green tensor $\mathrm{D}\boldsymbol{\phi}_0^t(\boldsymbol{x})^T\mathrm{D}\boldsymbol{\phi}_0^t(\boldsymbol{x})$. The corresponding maximum stretching direction in the initial domain is given by the right singular vector $\boldsymbol{\xi}_i^t(\boldsymbol{x})$, and is aligned with its matching left singular vector $\boldsymbol{\eta}_i^t(\boldsymbol{x})$ in the advected configuration.

The *forward* and *backward FTLE field* (Finite Time Lyapunov Exponent) on the time window $[0, t]$, respectively FTLE_0^t and FTLE_0^{-t} are defined by a logarithmic rescaling of the maximal singular value:

$$\mathrm{FTLE}_0^t(\boldsymbol{x}) = \frac{\log(\sigma_1^t(\boldsymbol{x}))}{t}, \qquad \mathrm{FTLE}_0^{-t}(\boldsymbol{y}) = \frac{\log(\sigma_1^{-t}(\boldsymbol{y}))}{t}. \tag{1.10}$$

Ridges of the forward FTLE have been used to find repelling LCS, while those of the backward FTLE were used to extract backward LCS [129]. There is a relationship between the two as it will be illustrated in corollary 1.1. One should keep in mind that the forward FTLE is a quantity defined in the initial configuration Ω while the backward FTLE is defined on the advected domain $\boldsymbol{\phi}_0^t(\Omega)$.

The logarithm is used as trajectories tend to diverge exponentially in time, hence $\mathrm{FTLE}_0^t(\boldsymbol{x})$ is a measure of the maximum rate of strain. This exponential growth of trajectories can be justified by the following observation:

Lemma 1.1. *Assume that the singular value $\sigma_1^t(\boldsymbol{x})$ of order i of $\mathrm{D}\boldsymbol{\phi}_0^t(\boldsymbol{x})$ remains simple on $]0, t]$. Then*

$$\sigma_i^t(\boldsymbol{x}) = \exp\left[\int_0^t \boldsymbol{\eta}_i^{sT}(\boldsymbol{x}) \left(\frac{\nabla\boldsymbol{v}(s, \boldsymbol{\phi}_0^t(\boldsymbol{x})) + \nabla\boldsymbol{v}(s, \boldsymbol{\phi}_0^t(\boldsymbol{x}))^T}{2}\right) \boldsymbol{\eta}_i^s(\boldsymbol{x})\mathrm{d}s\right] \leq e^{\rho(\boldsymbol{x})t}, \tag{1.11}$$

with

$$\rho(\boldsymbol{x}) = \sup_{s\in[0,t]} \left(\frac{\nabla\boldsymbol{v}(s, \boldsymbol{\phi}_0^t(\boldsymbol{x})) + \nabla\boldsymbol{v}(s, \boldsymbol{\phi}_0^t(\boldsymbol{x}))^T}{2}\right).$$

Proof. It is well known that $\frac{\mathrm{d}}{\mathrm{d}t}\sigma_i^t(\boldsymbol{x}) = \boldsymbol{\eta}_i^{tT}(\boldsymbol{x})\frac{\mathrm{d}}{\mathrm{d}t}\left(\nabla\boldsymbol{\phi}_0^t(\boldsymbol{x})\right)\boldsymbol{\xi}_i^t(\boldsymbol{x})$ (see e.g. [86]). Hence

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\sigma^t(\boldsymbol{x}) &= \boldsymbol{\eta}_i^{tT}(\boldsymbol{x})(\nabla\boldsymbol{v}(t, \boldsymbol{\phi}_0^t(\boldsymbol{x}))\mathrm{D}\boldsymbol{\phi}_0^t(\boldsymbol{x}))\boldsymbol{\xi}_i(\boldsymbol{x}) \\
&= \sigma_i^t(\boldsymbol{x})\boldsymbol{\eta}_i^{tT}(\boldsymbol{x})\nabla\boldsymbol{v}(t, \boldsymbol{\phi}_0^t(\boldsymbol{x}))\boldsymbol{\eta}_i^t(\boldsymbol{x}) \\
&= \left(\boldsymbol{\eta}_i^{tT}(\boldsymbol{x}) \left(\frac{\nabla\boldsymbol{v}(t, \boldsymbol{\phi}_0^t(\boldsymbol{x})) + \nabla\boldsymbol{v}(t, \boldsymbol{\phi}_0^t(\boldsymbol{x}))^T}{2}\right) \boldsymbol{\eta}_i^t(\boldsymbol{x})\right) \sigma_i^t(\boldsymbol{x}),
\end{aligned}$$

which yields (1.11) by integration. □

Remark 1.1. Equation (1.11) shows that the evolution of $\sigma_1^t(\boldsymbol{x})$ results from a competition between its history over the interval $[0,t]$ and the current eigendirection of maximal stretching of the instantaneous deformation tensor $(\nabla \boldsymbol{v} + \nabla \boldsymbol{v}^T)/2)$. This further emphasizes the dependency of *coherence* to a fixed time window $[0,t]$.

There is a duality between the SVD of the forward and backward flow maps and FTLE as stated in [70, 40, 81] :

Proposition 1.1. *The differential* $\mathrm{D}\boldsymbol{\phi}_0^{-t}$ *of the inverse flow map and of the flow map* $\mathrm{D}\boldsymbol{\phi}_0^t$ *are related by the formula:*

$$\mathrm{D}\boldsymbol{\phi}_0^{-t} = (\mathrm{D}\boldsymbol{\phi}_0^t)^{-1} \circ (\boldsymbol{\phi}_0^t)^{-1}.$$

Therefore the singular value decomposition of $\mathrm{D}\boldsymbol{\phi}_0^{-t}$ *is given by*

$$(\mathrm{D}\boldsymbol{\phi}_0^{-t})(\boldsymbol{y}) = \sum_{i=1}^{n} \sigma_i^t(\boldsymbol{\phi}_0^{-t}(\boldsymbol{y}))^{-1} \boldsymbol{\xi}^t(\boldsymbol{\phi}_0^{-t}(\boldsymbol{y})) \boldsymbol{\eta}^{tT}(\boldsymbol{\phi}_0^{-t}(\boldsymbol{y})).$$

In other words:

- *The* singular values *of* $\mathrm{D}\boldsymbol{\phi}_0^{-t}$ *are inverse of those of* $\mathrm{D}\boldsymbol{\phi}_0^t$ *advected backward in time:*

$$\sigma_i^{-t}(\boldsymbol{y}) = \sigma_{n-i+1}^t(\boldsymbol{\phi}_0^{-t}(\boldsymbol{y}))^{-1}.$$

- *The* right (*resp.* left) singular vectors *of* $\mathrm{D}\boldsymbol{\phi}_0^{-t}$ *are the corresponding left (resp. right) singular vectors of* $\mathrm{D}\boldsymbol{\phi}_0^t$ *advected backward in time:*

$$\boldsymbol{\xi}_i^{-t}(\boldsymbol{y}) = \boldsymbol{\eta}_{n-i+1}^t(\boldsymbol{\phi}_0^{-t}(\boldsymbol{y}))$$

$$\boldsymbol{\eta}_i^{-t}(\boldsymbol{y}) = \boldsymbol{\xi}_{n-i+1}^{-t}(\boldsymbol{\phi}_0^{-t}(\boldsymbol{y})).$$

Corollary 1.1. *In 2D* ($n = 2$) *and for a divergent free velocity field* $\boldsymbol{v}$*, the backward FTLE coincides with the forward FTLE advected backward in time:*

$$\mathrm{FTLE}_0^{-t}(\boldsymbol{y}) = \mathrm{FTLE}_0^t(\boldsymbol{\phi}_0^{-t}(\boldsymbol{y})).$$

Proof. This is an immediate consequence of the fact that under these assumptions,

$$\sigma_1^{-t}(\boldsymbol{y}) = \frac{1}{\sigma_2^t(\boldsymbol{\phi}_0^{-t}(\boldsymbol{y}))} = \sigma_1^t(\boldsymbol{\phi}_0^{-t}(\boldsymbol{y})),$$

since the free divergence condition implies $\det(\mathrm{D}\boldsymbol{\phi}_0^t) = \sigma_1^t(\boldsymbol{x})\sigma_2^t(\boldsymbol{x}) = 1$. □

The leading right singular vector at time t, $\boldsymbol{\xi}^t(\boldsymbol{x})$, is the direction maximizing the stretch (*i.e.* the discontinuity) $||\mathrm{D}\boldsymbol{\phi}_0^t(\boldsymbol{x})\delta\boldsymbol{x}||$ locally among all infinitesimal displacements $\delta\boldsymbol{x}$. Therefore, an natural idea is that LCS should be surfaces constantly normal to the vector field $\boldsymbol{\xi}_1^t(\boldsymbol{x})$. In [65], Haller proposed to define LCS as such surfaces that would satisfy the additional requirement that the magnitude of the discontinuity (i.e. $\sigma_1^t(\boldsymbol{x})$) is locally maximal in the normal direction to the surface. An issue with this approach is that requiring both conditions yields to an overdetermined definition. As a result, Haller proposed later a less restrictive theory in [68] where he defined attracting, repelling, and hyperbolic LCS, which can be interpreted as three possible ways of seeking "discontinuity fronts" of the flow map:

- *Repelling LCS* are surfaces everywhere normal to the leading vector field $\boldsymbol{\xi}_1^t(\boldsymbol{x})$, these surfaces are shown [104, 66] to be the most locally repelling surfaces $\mathcal{S}$ when allowing local deformations of the surface by rotations of the normal $\boldsymbol{n}$ around $\boldsymbol{x} \in \mathcal{S}$ (but without changing the position of $\boldsymbol{x}$). They are extracted in 2D as integral curves of the ODE $\mathrm{d}\boldsymbol{x}/\mathrm{d}s = \boldsymbol{\xi}_2^t(\boldsymbol{x}(s))$, with some care to handle possible sign discontinuities of the eigenvector field $\boldsymbol{\xi}_2$ (see [39]).

- *Elliptic LCS* are surfaces maximizing the tangential shear, i.e. associated to the maximum value of the constant L over all $\delta\boldsymbol{x}$ satisfying $\delta\boldsymbol{x}^T \mathrm{d}\boldsymbol{\phi}_0^t \delta\boldsymbol{x} = 0$. For a surface normal to such extremal direction $\delta\boldsymbol{x}$, the transport of the infinitesimal material perturbation $\delta\boldsymbol{x}$ is by definition tangential to the surface. They are extracted as integral curves of a field obtained from $\boldsymbol{\xi}_1^t$ and $\boldsymbol{\xi}_2^t$ (see [66]).

- *Attracting LCS* are surfaces that are everywhere normal to the eigenvector field $\boldsymbol{\xi}_n^t(\boldsymbol{x})$ associated with the smallest eigenvalue $\sigma_n^t(\boldsymbol{x})$. They are obtained as integral curves of the ODE $\mathrm{d}\boldsymbol{x}/\mathrm{d}s = \boldsymbol{\xi}_1^t(\boldsymbol{x}(s))$.

Nevertheless this locally optimal approach does not yield globally coherent structures: a LCS can be drawn from every point of the domain, and it is unclear how to select the most influential ones. In [66], Haller proposes to select the curves that go along global maxima of the FTLE field, but it it is not a priori guaranteed that a globally maximizing property is maintained all the way along the curve.

Other variants of this approach have been considered in [69] as well as some instantaneous techniques [128], valid for small integration times or for autonomous systems. To date, these methods have been tested on realistic flows in [112, 61].

1.2.2 Polar distance and rigid sets

In our review of the available LCS literature and our conclusions about the relations between LCS and the regularity of the flow map, we were surprised not too find any reference to the following well known theorem in continuum mechanics [120]:

Theorem 1.1. *Suppose that $\boldsymbol{\phi}$ is a transformation whose jacobian is a rotation at any point of the domain Ω, namely*

$$\forall \boldsymbol{x} \in \Omega, \mathrm{D}\boldsymbol{\phi}(\boldsymbol{x})^T \mathrm{D}\boldsymbol{\phi}(\boldsymbol{x}) = I.$$

Then $\boldsymbol{\phi}$ is a rigid transformation (i.e. a translation plus a rotation): there exists $P \in \mathcal{O}_{nn}$ a rotation (i.e. $P^T P = I$) independent of $\boldsymbol{x}$ such that

$$\forall \boldsymbol{x} \in \Omega, \boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x}_0) + P(\boldsymbol{x} - \boldsymbol{x}_0).$$

Proof. We give a proof inspired from [137, 118]. Two proofs are popular, the first consists in differentiating the equality $\mathrm{D}\boldsymbol{\phi}(\boldsymbol{x})^T \mathrm{D}\boldsymbol{\phi}(\boldsymbol{x}) = I$ and using the symmetry of the Hessian (namely Schwarz theorem) to obtain that $\mathrm{D}\boldsymbol{\phi}(\boldsymbol{x}) = 0$. The existence of the Hessian is actually not required as explained more transparently in the following. Consider the arc $\boldsymbol{\phi}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))$ image of the segment joining $\boldsymbol{x}$ to $\boldsymbol{y}$ by $\boldsymbol{\phi}$. Its length is

$$\int_0^1 \left|\left| \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{\phi}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \right|\right| \mathrm{d}t = ||\boldsymbol{y} - \boldsymbol{x}||.$$

Also this arc must have a length greater than the length of the segment joining $\boldsymbol{\phi}(\boldsymbol{x})$ to $\boldsymbol{\phi}(\boldsymbol{y})$, therefore $||\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{y})|| \leq ||\boldsymbol{y} - \boldsymbol{x}||$. Since the differential of the inverse map of $\boldsymbol{\phi}$ is $\mathrm{D}\boldsymbol{\phi}(\boldsymbol{\phi}^{-1}(\boldsymbol{x}))^{-1}$ is also an orthogonal matrix, one deduces that $||\boldsymbol{\phi}^{-1}(\boldsymbol{\phi}(\boldsymbol{y})) - \boldsymbol{\phi}^{-1}(\boldsymbol{\phi}(\boldsymbol{x}))|| = ||\boldsymbol{y} - \boldsymbol{x}|| \leq ||\boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{\phi}(\boldsymbol{x})||$, hence $||\boldsymbol{y} - \boldsymbol{x}|| = ||\boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{\phi}(\boldsymbol{x})||$: $\boldsymbol{\phi}$ preserves distances. This implies that $\boldsymbol{\phi}$ is a rigid motion. Indeed,

$$||\boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{\phi}(\boldsymbol{x}_0) + \boldsymbol{\phi}(\boldsymbol{x}_0) - \boldsymbol{\phi}(\boldsymbol{x})||^2 = ||\boldsymbol{y} - \boldsymbol{x}_0||^2 - 2 < \boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{\phi}(\boldsymbol{x}_0), \boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{x}_0) > + ||\boldsymbol{x} - \boldsymbol{x}_0||^2,$$

showing that

$$\forall \boldsymbol{x}, \boldsymbol{x}_0, \boldsymbol{y} \in \Omega, \ < \boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{\phi}(\boldsymbol{x}_0), \boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{x}_0) > = < \boldsymbol{y} - \boldsymbol{x}_0, \boldsymbol{x} - \boldsymbol{x}_0 > .$$

Denote P the matrix whose columns are $(\boldsymbol{\phi}(\boldsymbol{x}_0 + t\boldsymbol{e}_i) - \boldsymbol{\phi}(\boldsymbol{x}_0))_j$, for a t sufficiently small and $\boldsymbol{e}_i$ an orthonormal basis of $\boldsymbol{P}^n$, so that $\boldsymbol{x}_0 + t\boldsymbol{e}_i \in \Omega$. Then for any $\boldsymbol{x} = \boldsymbol{x}_0 + t\sum_{i=1}^n x^i \boldsymbol{e}_i$, by linearity of P with respect to the scalar product

$$\begin{aligned}
&< \boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{\phi}(\boldsymbol{x}_0), \boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{x}_0) > \\
&\qquad = < \boldsymbol{y} - \boldsymbol{x}_0, t\sum_{i=1}^n x_i \boldsymbol{e}_i > = \sum_{i=1}^n x_i < \boldsymbol{y} - \boldsymbol{x}_0, t\boldsymbol{e}_i > \\
&\qquad = \sum_{i=1}^n x_i < \boldsymbol{\phi}(y) - \boldsymbol{\phi}(\boldsymbol{x}_0), \boldsymbol{\phi}(\boldsymbol{x}_0 + t\boldsymbol{e}_i) - \boldsymbol{\phi}(\boldsymbol{x}_0) > \\
&\qquad = \sum_{i=1}^n x_i < \boldsymbol{\phi}(y) - \boldsymbol{\phi}(\boldsymbol{x}_0), P\boldsymbol{e}_i > = < \boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{\phi}(\boldsymbol{x}_0), P(\boldsymbol{x} - \boldsymbol{x}_0) > .
\end{aligned}$$

Hence one finally concludes that

$$\begin{aligned}
&||\boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{\phi}(\boldsymbol{x}_0) - P(\boldsymbol{y} - \boldsymbol{x}_0)||^2 \\
&\qquad = ||\boldsymbol{y} - \boldsymbol{x}_0||^2 - 2 < \boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{\phi}(\boldsymbol{x}_0), P(\boldsymbol{y} - \boldsymbol{x}_0) > + ||\boldsymbol{y} - \boldsymbol{x}_0||^2 \\
&\qquad = 2||\boldsymbol{y} - \boldsymbol{x}_0||^2 - 2||\boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{\phi}(\boldsymbol{x}_0)||^2 = 0.
\end{aligned}$$

□

Therefore, a way to quantify how far the flow map $\boldsymbol{\phi}_0^t$ is from being a rigid transformation, can intuitively be done by measuring how far the jacobian $\mathrm{D}\boldsymbol{\phi}_0^t(\boldsymbol{x})$ is from being a rotation at every point. It turns out that John (1961) has shown that theorem 1.1 is "stable under perturbations" in the following sense:

Theorem 1.2 (John [78], see also chapter 5, Theorem. 2.2 in [118])**.** *Let $B(\boldsymbol{x}_0, \rho) \subset \Omega$ be the ball centered at $\boldsymbol{x}_0$ and of radius ρ. Assume that there exists $\epsilon > 0$ such that*

$$\forall \boldsymbol{x} \in B(\boldsymbol{x}_0, \rho), \forall 1 \leq i \leq n, \ |\sigma_i(\boldsymbol{x}) - 1| \leq \epsilon,$$

where $\sigma_i(\boldsymbol{x}) = \sigma_i(\mathrm{D}\boldsymbol{\phi}(\boldsymbol{x}))$ is the i-th singular value of the Jacobian $\mathrm{D}\boldsymbol{\phi}(\boldsymbol{x})$. Then there exists a constant C dependent only of the dimension n of $\mathbb{R}^n$ and a rotation P independent of $\boldsymbol{x}$ such that $\boldsymbol{\phi}$ is close to be a rigid transformation on $B(\boldsymbol{x}_0, \rho)$:

$$\forall \boldsymbol{x} \in B(\boldsymbol{x}_0, \rho), ||\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{x}_0) - P(\boldsymbol{x} - \boldsymbol{x}_0)|| \leq C\rho\epsilon.$$

In the following, we say a set $\mathcal{A}_{rigid}$ is *rigid* between the instants 0 and t if the restriction

of flow map $\boldsymbol{\phi}_0^t$ to $\mathcal{A}_{rigid}$ is close to be a rigid motion. We also designate in the following by "*polar distance*" of a matrix the quantity

$$\mathcal{P}(F) = \left(\sum_{i=1}^{n} (1 - \sigma_i(F))^2 \right)^{\frac{1}{2}}, \tag{1.12}$$

which is the euclidean distance of the matrix F to the orthogonal group $\mathcal{O}_n$, (see proposition 2.22 in chapter 2). Hence theorem 1.2 states that *rigid* sets $\mathcal{A}_{rigid}$ may be obtained by a simply thresholding the polar distance:

$$\mathcal{A}^{\epsilon}_{rigid} = \{\boldsymbol{x} \in \Omega | \mathcal{P}(\mathrm{D}\boldsymbol{\phi}_0^t(\boldsymbol{x})) \leq \epsilon\}, \tag{1.13}$$

the parameter ϵ allowing some tolerance over the scale at which one seeks such rigidity. Connected components of $\mathcal{A}^{\mathrm{FTLE}}_{rigid}$ are transformed by $\boldsymbol{\phi}_0^t$ in a approximate rigid manner, with a possible stretching proportional to $\epsilon T \rho$ where ρ is the size of the each component set (a natural scaling factor for ϵ being $||\nabla v||_{L^\infty([0,T],L^2)}$, coming from the remark of equation (1.11)). In 2D, *i.e.* $n = 2$, and for a divergence free field, this criterion is more or less equivalent to thresholding the FTLE field: indeed, the relation $\sigma_1^t(\boldsymbol{x})\sigma_2^t(\boldsymbol{x}) = 1$ holds for all times t, therefore

$$\{\boldsymbol{x} \in \Omega | \mathrm{FTLE}_0^t(\boldsymbol{x}) \leq \log(1 + \epsilon/\sqrt{2})/t\} \subset \mathcal{A}^{\epsilon}_{rigid} \subset \{\boldsymbol{x} \in \Omega | \mathrm{FTLE}_0^t(\boldsymbol{x}) \leq \log(1 + \epsilon)/t\}.$$

Hence ridges of the FTLE that delimit regions where the FTLE is small may be considered as true boundaries of rigid sets. This addresses the critic made on the ability of FTLE to detect rigid structures: examples given in [65] for which FTLE ridges are disqualified to be LCS, are particular in the sense that the value of the FTLE is high everywhere in the domain. In these cases, the flow map $\boldsymbol{\phi}_0^t$ exhibits high stretching everywhere in the domain and the proposed criterion (1.13) can not be satisfied, even if the FTLE admits ridges.

Remark 1.2. Exploiting the duality of proposition 1.1, a possibly better definition for the polar distance can be

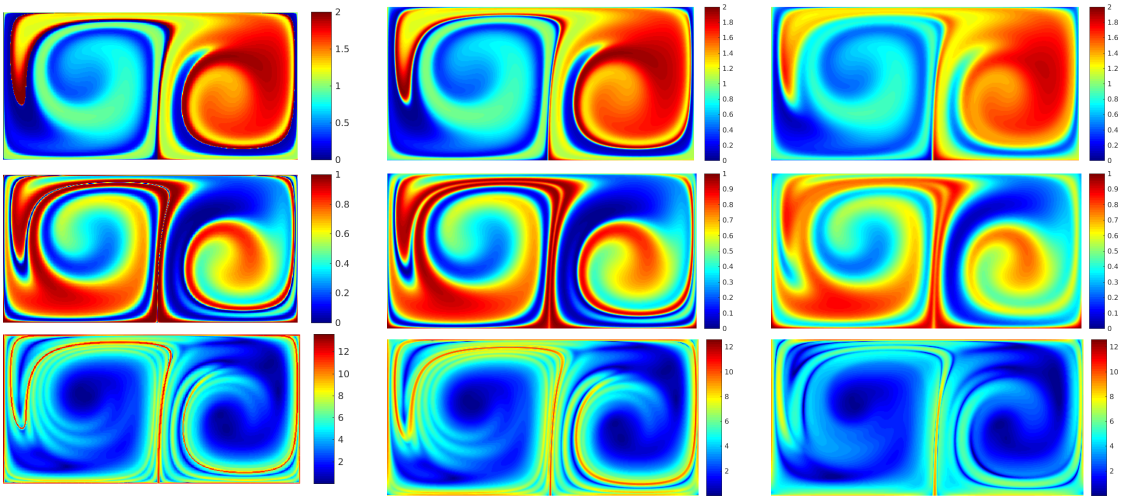
$$\mathcal{P}(F) = \left(\sum_{i=1}^{n} \frac{(1 - \sigma_i(F))^2}{\sigma_i(F)} \right)^{\frac{1}{2}}.$$

Indeed, this allows to obtain rigid sets in a way that is independent of the choice of the initial or final time: $\mathcal{P}(\mathrm{D}\boldsymbol{\phi}_0^{-t}) = \mathcal{P}(\mathrm{D}\boldsymbol{\phi}_0^t) \circ \boldsymbol{\phi}_0^{-t}$. With this definition, rigid sets obtained from the thresholding of equation (1.13) at time 0 coincide with those that would be obtained at time t from a thresholding of the backward flow map ($\mathcal{A}^{\epsilon}_{rigid,t} = \{\boldsymbol{x} \in \Omega | \mathcal{P}(\mathrm{D}\boldsymbol{\phi}_0^{-t}(\boldsymbol{x})) \leq \epsilon\}$), advected backward in time.

1.2.3 Numerical experiments and comparisons

Estimating the flow map and the FTLE field: particles vs. eulerian method

In most of the works mentioned above, the flow map $\boldsymbol{\phi}_0^t$ is estimated through direct particle advection (with their refinements [22]). In [97], Leung suggests to obtain the flow map by solving the transport PDE (1.1). This Eulerian method has the advantage of not requiring velocity interpolation outside grid points and to yield an accuracy that his uniform over the grid. Nevertheless, these schemes are subject to a CFL condition that constrains the time step $\Delta t \leq C \Delta x$ to be kept proportional to the grid spacing. Conversely, particle



(a) Particles (b) WENO (c) 1st order Upwind scheme

Figure 1-5: Numerical schemes comparisons for the Double Gyre example (From top to bottom: coordinates x and y of the flow map $\boldsymbol{\phi}_0^t$ and forward FTLE)

methods are easy to implement, easily parallelizable and are found to be very efficient when solving advection dominated problems [27]. In addition these methods may use a time step independent of the grid resolution, and are less subject to numerical diffusion inherent to advection schemes over large integration times. Diffusion tends to smear out sharp features of the computed flow map, such as FTLE ridges which are especially desirable in LCS computation. Nevertheless, this comes at the cost of furnishing an accuracy that is spatially dependent (proportional to the local lyapunov exponent $\sigma_1^t(\boldsymbol{x})$), which is visible by the presence of some numerical noise around sharp gradients of the flow map.

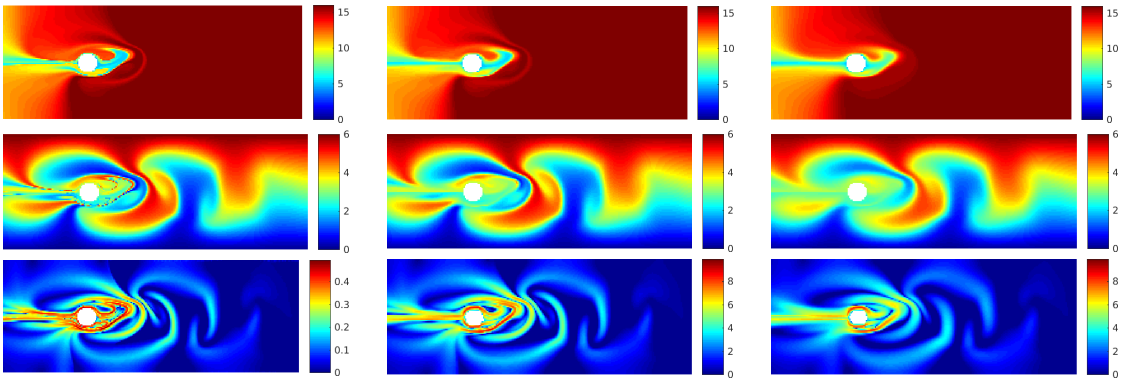
These effects are illustrated on Figs. 1-5 to 1-7 for the use of direct particle advection, the TVDRK3/WENO scheme [108] that has high order accuracy and low diffusivity, and the first order upwind scheme with euler time integration. For each example, we plot in vertical order the x and y coordinates of the computed flow map and the corresponding FTLE field. We used a CFL constant $CFL = 10$ for the particle simulation and $CFL = 0.9$ for the eulerian methods.

Repelling LCS

For each of the three examples, we extract a few LCS curves by integrating the vector field of the right singular vector $\boldsymbol{\xi}_2(\boldsymbol{x})$ of $\mathrm{D}\boldsymbol{\phi}_0^t$ associated with the lowest eigenvalue, according to the definition of LCS proposed by Haller [66]. These are plotted on Figure 1-8. It is interesting to notice that these curves follow qualitatively the ridges of the FTLE field, implying that the condition $< \nabla\sigma_1^t(\boldsymbol{x}), \boldsymbol{\xi}_1(\boldsymbol{x}) > \simeq 0$ is satisfied for the three examples (see [39, 15]), although there is no available justification in the literature about why it seems to be often the case.

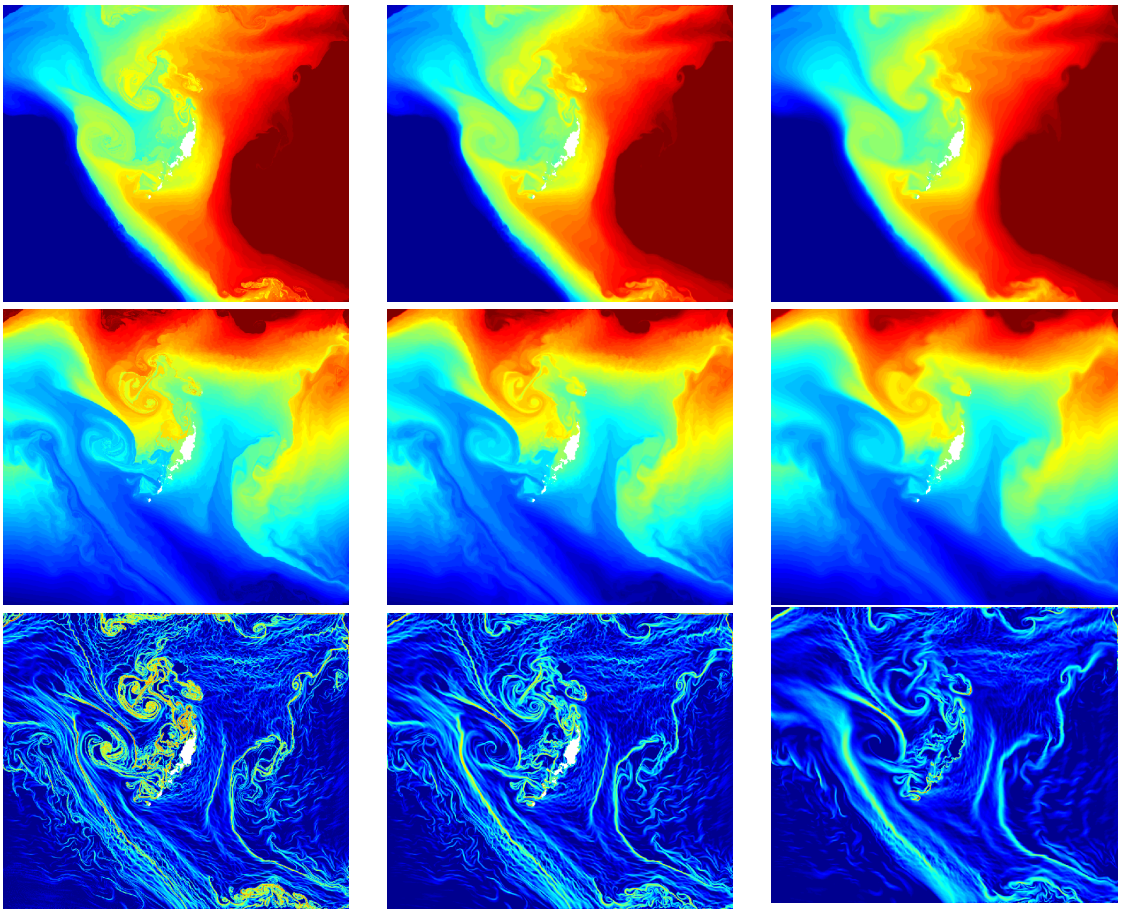
Polar distance criterion

For each of the three examples, we compute and plot the polar distance defined in eqn. (1.12) on Figure 1-9. "Rigid sets" on which the flow map acts approximately as a rigid



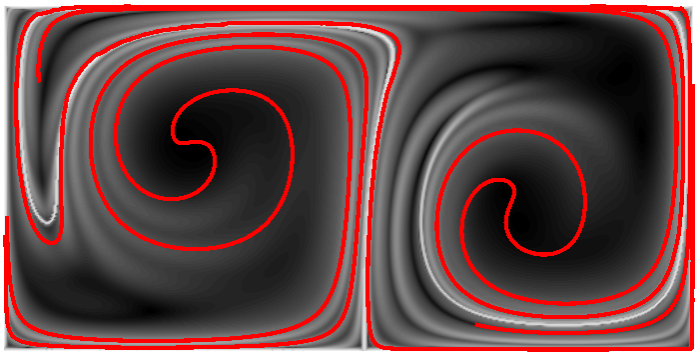
(a) Particles (b) WENO (c) 1st order Upwind scheme

Figure 1-6: Numerical schemes comparisons for the Flow Past a Cylinder example (From top to bottom: coordinates x and y of the flow map $\boldsymbol{\phi}_0^t$ and forward FTLE)

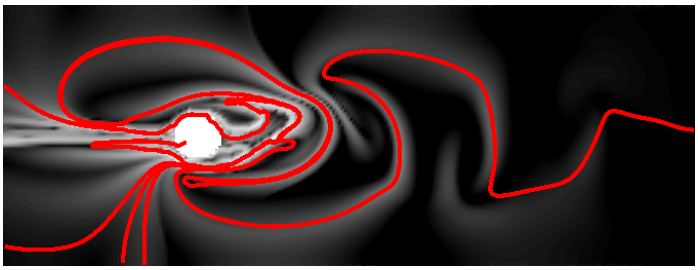


(a) Particles (b) WENO (c) 1st order Upwind scheme

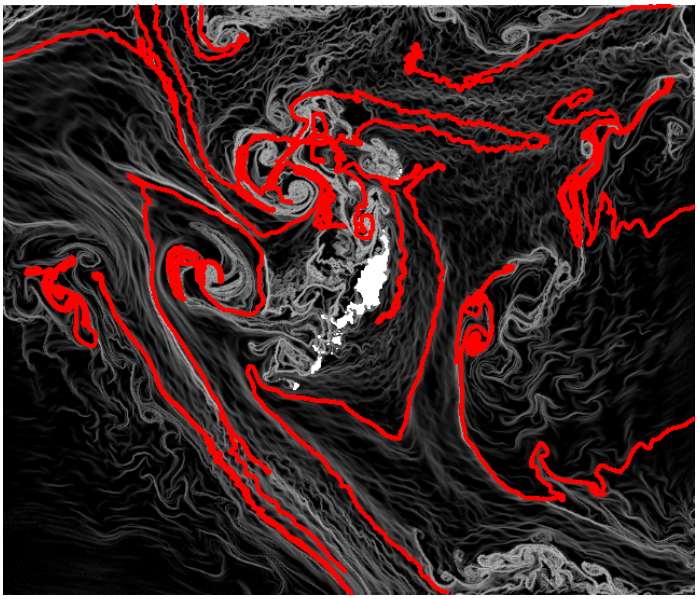
Figure 1-7: Numerical schemes comparisons for the Fleat region example (From top to bottom: coordinates x and y of the inverse flow map $\boldsymbol{\phi}_0^{-t}$ and backward FTLE)



(a) Repelling LCS for the Double Gyre example

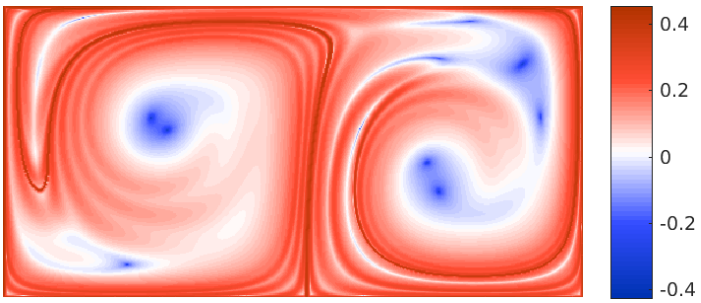


(b) Repelling LCS for the Flow Past a Cylinder

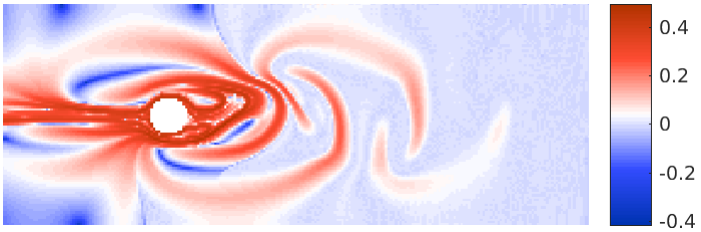


(c) Attracting LCS for the Palau region

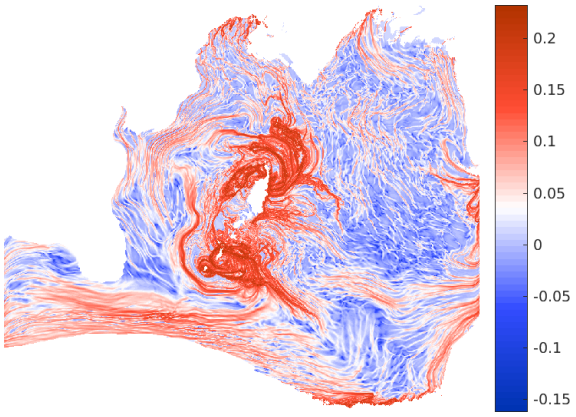
Figure 1-8: A few LCS (*in red*) obtained as tensor lines of the Cauchy Green Tensor $D\boldsymbol{\phi}_0^t$ (methodology of [66]). The FTLE field is plotted in greyscale in background



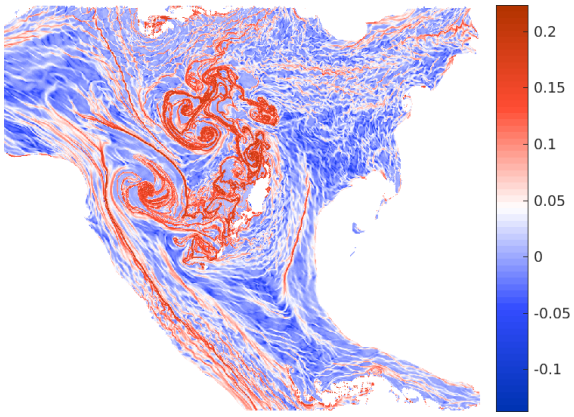
(a) Double Gyre example



(b) Flow Past a Cylinder



(c) Palau region (Initial time)



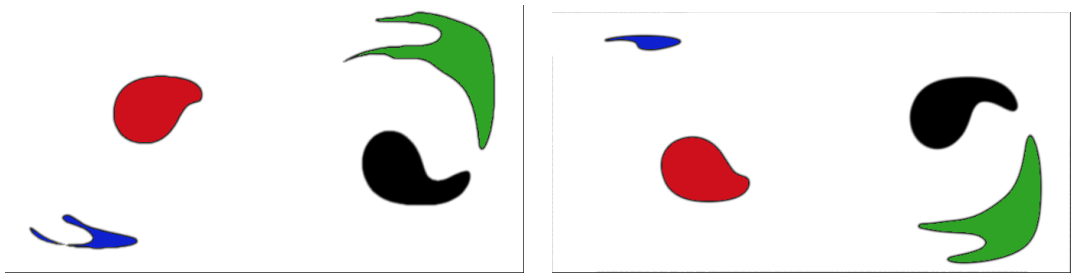
(d) Palau region (final time)

Figure 1-9: The polar distance $\log(\mathcal{P}(\mathrm{D}\boldsymbol{\phi}_0^t(\boldsymbol{x})))/t$ in logarithmic scale plotted for the three examples

transformation are obtained as the blue regions. Since the flow is divergence free, one sees a clear analogy with the FTLE field. These plots emphasize the fact that FTLE ridges have a "thickness". These examples show that the thresholding criterion (1.13) may be used to identify key subregions that are advected in a rigid manner. It is interesting to observe that the flow map acts as a rigid rotation and translation on each of the connected components obtained, but these rotations and translations may be different for each rigid region. Surrounding regions are characterized by an increased stretch.

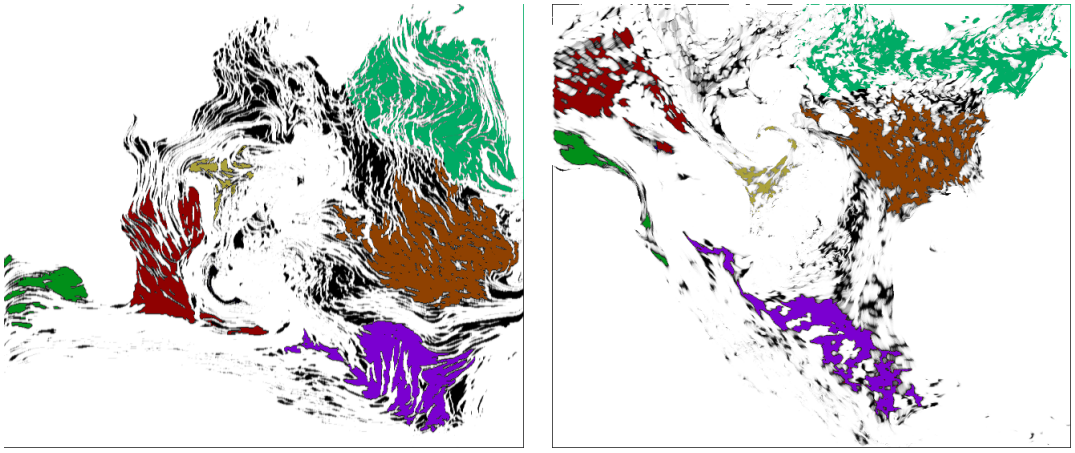
1.3 Operator based LCS methods

In this section, we review and reformulate the transfer operator methods developed by Froyland [46, 52, 47], and suggest a few improvements in order to apply them on highly resolved, realistic velocity fields. We demonstrate their applicability on the three benchmark



(a) Double Gyre

(b) Flow Past a Cylinder



(c) Palau region

Figure 1-10: Initial and final configurations of rigid sets. Colors have been added to help the reader identify corresponding pairs of rigid sets.

examples considered previously to allow for comparison.

1.3.1 Summary of the theory

Instead of analyzing the flow map $\boldsymbol{x} \mapsto \boldsymbol{\phi}_0^t(\boldsymbol{x})$ that describes the motion of particles individually, a second class of methods relies on functional operators of the form (1.6). In these methods, "relevant" information that allows to best understand the dynamics is found by analyzing the spectral or Singular Value Decompositions of these operators, in the same way the SVD of the linearized dynamics induced by $\mathrm{D}\boldsymbol{\phi}_0^t$ was used to extract the most relevant Lagrangian directions in section 1.2. In fact, one can also interpret the method as an attempt to select geometrically simple sets that remain simple when advected by the flow, which is achieved by finding the functions that are the most regular in both the initial and advected configuration. As it will be illustrated in the figures of the following numerical examples, this method yields an information that is somewhat different and complementary to the one obtained with FTLEs. Transfer operator methods have been used on realistic data in [48]. Note and another technique of this class is proposed in [23].

Transfer operators A key tool in the study of dynamical systems [87] is the introduction of the *Perron-Frobenius* operator

$$\begin{array}{rccc} \boldsymbol{P} & : & L^2(\Omega) & \to & L^2(\Omega) \\ & & f & \mapsto & f \circ (\boldsymbol{\phi}_0^t)^{-1} |\det \nabla (\boldsymbol{\phi}_0^t)^{-1}|, \end{array} \tag{1.14}$$

and of its adjoint, the *Koopman* operator

$$\begin{array}{rccc} \boldsymbol{P}^T & : & L^2(\Omega) & \to & L^2(\Omega) \\ & & f & \mapsto & f \circ \boldsymbol{\phi}_0^t. \end{array} \tag{1.15}$$

For autonomous flows, the first eigenvector of these operators allow to analyze long-term characteristics of the system dynamics, such as ergodicity or the existence of an invariant measure [87]. To extract coherent sets in a non-autonomous or time-dependent dynamics, it is natural to consider the *push forward*

$$\begin{array}{rccc} \boldsymbol{L} & : & L^2(\Omega, \mathrm{d}\boldsymbol{x}) & \to & L^2(\Omega, \boldsymbol{P}(\mathrm{d}\boldsymbol{x})) \\ & & f & \mapsto & f \circ (\boldsymbol{\phi}_0^t)^{-1}, \end{array} \tag{1.16}$$

and the *push backward* operator

$$\begin{array}{rccc} \boldsymbol{L}^* & : & L^2(\Omega, \boldsymbol{P}(\mathrm{d}\boldsymbol{x})) & \to & L^2(\Omega, \mathrm{d}\boldsymbol{x}) \\ & & f & \mapsto & f \circ \boldsymbol{\phi}_0^t. \end{array} \tag{1.17}$$

These operators are dual one another once an appropriate change of scalar product on the L^2 spaces has been introduced: the notation $L^2(\Omega, \mathrm{d}\boldsymbol{x})$ and $L^2(\Omega, \boldsymbol{P}(\mathrm{d}\boldsymbol{x}))$ refers to the L^2 spaces endowed respectively with the scalar products induced by the Lebesgue measure $\mathrm{d}\boldsymbol{x}$ and its image $\boldsymbol{P}(\mathrm{d}\boldsymbol{x}) = \boldsymbol{P}(1_\Omega)\mathrm{d}\boldsymbol{x}$. With these scalar products, the duality between the operators $\boldsymbol{L}$ and $\boldsymbol{L}^*$ is written

$$\begin{aligned} < \boldsymbol{L}f, g >_{\boldsymbol{P}(\mathrm{d}\boldsymbol{x})} &= \int_\Omega f(\boldsymbol{\phi}_0^{-t}(\boldsymbol{x})) g(\boldsymbol{x}) \boldsymbol{P}(\mathrm{d}\boldsymbol{x}) = \int_\Omega f(\boldsymbol{\phi}_0^{-t}(\boldsymbol{x})) g(\boldsymbol{x}) |\det \nabla (\boldsymbol{\phi}_0^t)^{-1}| \mathrm{d}\boldsymbol{x} \\ &= \int_\Omega f(\boldsymbol{x}) g(\boldsymbol{\phi}_0^t(\boldsymbol{y})) \mathrm{d}\boldsymbol{y} =< f, \boldsymbol{L}^* g >_{\mathrm{d}\boldsymbol{x}} . \end{aligned}$$

Froyland has shown that level-sets of the second dominant singular vectors of a diffusive approximation of $\mathcal{L}(t)$ may allow to extract *coherent sets*, that are expected to be subdomains characterized by a slow mixing with their complementary when advected by the flow [46, 51, 15, 53, 54, 52]. Under pure advection, the mass of a function f is conserved in the sense that $||f||_{\mathrm{d}\boldsymbol{x}} = ||\boldsymbol{L}f||_{\boldsymbol{P}(\mathrm{d}\boldsymbol{x})}$ (the change of scalar product allowing to remove the effects related to compressibility). If a small amount ϵ of diffusion ϵ is added (e.g. by the introduction of some amount of gaussian noise in the particle motions in (1.2) or a Laplacian operator in (1.1), see [49]), $\boldsymbol{L}$ is approximated as $\boldsymbol{L} \simeq \boldsymbol{L}_\epsilon$ and one has instead $||\boldsymbol{L}_\epsilon(t)f||_{\boldsymbol{P}(\mathrm{d}\boldsymbol{x})} \leq ||f||_{\mathrm{d}\boldsymbol{x}}$. In that setting, $\boldsymbol{L}_\epsilon$ becomes a compact operator and admits a non-trivial SVD (without diffusion $\boldsymbol{L}$ is a unitary operator since $\boldsymbol{L}\boldsymbol{L}^* = \boldsymbol{L}^*\boldsymbol{L} = I$, all singular values are equal to one) [52]. Since singular singular vectors (f_i, g_i) (with respect to the appropriate scalar products) are the solution of the maximization problem

$$\sigma_i = u_i^T \boldsymbol{L}_\epsilon(t) v_i = \max_{\substack{||f||_{\boldsymbol{P}(\mathrm{d}\boldsymbol{x})}=||g||_{\mathrm{d}\boldsymbol{x}}=1 \\ g\in \mathrm{Span}(g_j)_{j<i}^\perp \\ f\in \mathrm{Span}(f_j)_{j<i}^\perp}} < g, \boldsymbol{L}_\epsilon(t) f >_{\boldsymbol{P}(\mathrm{d}\boldsymbol{x})}, \tag{1.18}$$

the subspace $\mathrm{span}(f_i)_{i\leq r}$ can naturally naturally be understood as the r dimensional subspace of initial data that is *the most resistant to diffusivity*, since functions belonging to this subspace see their norm being reduced by a factor of at most σ_r (with $\sigma_1 = 1$).

SVD of the push-forward operator seen as a compact operator More generally, one can justify the method and the use of further dominant eigenvectors of $\boldsymbol{L}_\epsilon(t)$ to extract coherent sets by the intuition that these are level sets of functions f such that f and $\boldsymbol{L}f$ are simultaneously the smoothest as possible. This can be justified as follows: suppose the velocity field is divergence free ($\boldsymbol{P}(\mathrm{d}\boldsymbol{x}) = \mathrm{d}\boldsymbol{x}$, the extension to the general case being possible, see [50]) and that the diffusive approximation $\boldsymbol{L}_\epsilon(t)$ of $levelsetsofL$ is given by $\boldsymbol{L}_\epsilon(t) = (I - \epsilon\Delta)^{-\frac{1}{2}} \boldsymbol{L} (I - \epsilon\Delta)^{-\frac{1}{2}}$ where Δ is the laplacian equipped with suitable boundary conditions. $\boldsymbol{L}_\epsilon$ is a compact operator from the Sobolev space $H^1(\Omega)$ to itself and hence admits a singular value decomposition [30]. Equipping $H^1(\Omega)$ with the equivalent scalar product $< f, g >_{H^1,\epsilon} = \int_\Omega fg\mathrm{d}\boldsymbol{x} + \epsilon \int_\Omega \nabla f \nabla g \mathrm{d}\boldsymbol{x}$ for $\epsilon > 0$, one can rewrite (1.18) as

$$\sigma_i =< \boldsymbol{L}f_i, g_i >_{L^2} = \max_{\substack{f,g\in H_1(\Omega) \\ ||f||_{H^1,\epsilon}=1, ||g||_{H^1,\epsilon}=1, \\ f\in \mathrm{Span}(f_j)_{j<i}, g\in \mathrm{Span}(g_j)_{j<i}}} < \boldsymbol{L}f, g >_{L^2}, \tag{1.19}$$

where $f_i, g_i \in H^1(\Omega)$, $||f||_{H^1,\epsilon} = ||f||_{L^2}^2 + \epsilon||\nabla f||_{L^2}^2$ and $||\cdot||_{L^2}$ and $<\cdot,\cdot>_{L^2}$ are the standard L^2 norm and scalar product. In that setting, we can write a "true" singular value decomposition for the pure advective operator

$$\boldsymbol{L} : H^1(\Omega) \subset L^2(\Omega) \longrightarrow L^2(\Omega) \subset H^{-1}(\Omega)$$

seen as a compact map from $H^1(\Omega)$ to its dual $H^{-1}(\Omega)$:

$$\forall f \in H^1(\Omega),\ \boldsymbol{L}f = \sum_{i=1}^{\infty} \sigma_i < f, f_i >_{H^1,\epsilon} g_i,$$

where the equality must be understood in $H^{-1}(\Omega)$. In other words, $\boldsymbol{L}f_i = g_i$ means that the induced linear forms on $H^1(\Omega)$ are equal:

$$\forall \boldsymbol{\phi} \in H^1(\Omega), < \boldsymbol{L}f_i, \boldsymbol{\phi} >_{L^2} = < g_i, \boldsymbol{\phi} >_{H^1,\epsilon} .$$

Note that in that setting, $< \boldsymbol{L}f, g >_{L^2}$ in (1.19) ($\boldsymbol{L}f$ and g seen as L^2 functions) is equal to $< \boldsymbol{L}f, g >_{H^{-1}}$ ($\boldsymbol{L}$ and f seen as elements of $H^{-1}(\Omega)$) where $< \cdot, \cdot >_{H^{-1}}$ is the natural scalar product on $H^{-1}(\Omega)$ that is inferred from the identification $H^1(\Omega) \simeq H^{-1}(\Omega)$: for any $f^*, g^* \in H^{-1}(\Omega)$, the scalar product on $H^{-1}(\Omega)$ is defined by $< f^*, g^* >_{H^{-1}} = < f, g >_{H^1,\epsilon}$ where f and g are the functions of $H^1(\Omega)$ satisfying $\forall \boldsymbol{\phi} \in H^1$, $f^*(\boldsymbol{\phi}) = < f, \boldsymbol{\phi} >_{H^1,\epsilon}$ and $g^*(\boldsymbol{\phi}) = < g, \boldsymbol{\phi} >_{H^1,\epsilon}$.

Zero diffusion limit If the flow map is smooth enough, then $\boldsymbol{L}$ is also a map from $H^1(\Omega)$ to $H^1(\Omega)$ because $||\nabla(\boldsymbol{L}f)||_{L^2} \leq C||\nabla \boldsymbol{\phi}_0^t||_{L^2}||\nabla f||_{L^2}$ for a given constant depending only on the domain Ω. Then one can obtain (recall also that $\boldsymbol{L}^T\boldsymbol{L} = I$)

$$\forall f, g \in H^1(\Omega),$$
$$\left\langle \frac{I - (I - \epsilon\Delta)^{-\frac{1}{2}}\boldsymbol{L}^T(I - \epsilon\Delta)^{-1}\boldsymbol{L}(I - \epsilon\Delta)^{-\frac{1}{2}}}{\epsilon} f, g \right\rangle \xrightarrow[\epsilon \to 0]{} \int_\Omega \nabla f \nabla g + \nabla(\boldsymbol{L}f)\nabla(\boldsymbol{L}g) \mathrm{d}\boldsymbol{x}, \tag{1.20}$$

the limit being the dynamic laplacian operator obtained by Froyland in [47] (equation (3)). Froyland has in fact shown that the convergence still holds if the smoothing operator $(I - \Delta)^{\frac{1}{2}}$ is replaced with any isotropic regularizing kernel.

Therefore right singular vectors f_i are expected to converge in some sense (this has not been yet proven) to the eigenvectors of the dynamic Laplace operator $-\Delta - \boldsymbol{L}^*\Delta\boldsymbol{L}$. These eigenvectors satisfy the initial requirement of being functions f such that both f and $\boldsymbol{L}f$ are the smoothest as possible since they minimize $|\nabla f|^2 + |\nabla(\boldsymbol{L}f)|^2$ while satisfying $||f||_{L^2}^2 = 1$. Note that the dynamic laplacian seems not to exist if the inclusion $\boldsymbol{L}(H^1(\Omega)) \subset H^1(\Omega)$ does not hold. We mention that the application of the Courant Nodal domain theorem (see Theorem 13p111 in [30], vol. 3) would possibly allow to show that the number of connected coherent sets obtained for the k-th pair of singular vectors is less or equal to k, which is observed in the numerical examples to come.

Numerical methods The general procedure to extract coherent sets consists therefore in building a finite-dimensional approximation of the operator $\boldsymbol{L}$, before estimating the singular value decomposition of either $\boldsymbol{L}_\epsilon$ for a small diffusivity $\epsilon > 0$ (as in [52, 48]) or the eigenvectors of $\Delta + \boldsymbol{L}^*\Delta\boldsymbol{L}$. Coherent sets are obtained by thresholding functions in the dominant subspaces spanned by the singular vectors. Note that it seems a priori that evaluating the SVD of $\boldsymbol{L}_\epsilon$ is more stable numerically and less expensive that targeting at extracting eigenvectors of the limit $-\Delta - \boldsymbol{L}^*\Delta\boldsymbol{L}$. Indeed, iteratives methods for computing the eigenvectors associated with smallest eigenvalues of a positive definite operator require matrix inversion. In contrast, dominant singular vectors of $\boldsymbol{L}_\epsilon$ do not require such inversion (but slow convergence can be achieved if singular values are not sufficiently separated).

Numerically, the implementation of the method in the general case where the velocity field is not assumed divergence-free requires an estimation of the image measure $\boldsymbol{P}(\mathrm{d}\boldsymbol{x})$. This raises difficulties since the image measure should be found such that without diffusion,

$\boldsymbol{L}_\epsilon$ is self-adjoint with respect to the change of scalar product. The inherent diffusivity of numerical computations may result in poor estimates. Next, the SVD of the matrix $\boldsymbol{L}$, must be evaluated with respect to the change of scalar product. The procedure used in [52] is summarized in algorithm 1. Last,

Algorithm 1 Computing coherent sets in the non-autonomous setting (from [52])

1: Compute a finite dimensional approximation of the Koopman operator $\boldsymbol{P}^T(t) = \boldsymbol{L}^*$ (denoted $\overline{P}$ in [52]). If for example l spatial cells of same dimension are used for the discretization of both the starting and image spaces $L^2(\Omega, \mathrm{d}\boldsymbol{x})$ and $L^2(\Omega, \boldsymbol{P}(\mathrm{d}\boldsymbol{x}))$, $\boldsymbol{L}^* \in \mathcal{M}_{l,l}$ is a l-by-l matrix.

2: Assuming $\boldsymbol{L}^* \mathbf{1}_\Omega = \mathbf{1}_\Omega$, compute

$$q = \boldsymbol{L}^{*T} \mathbf{1}_\Omega, \tag{1.21}$$

to estimate $|\det \nabla(\boldsymbol{\phi}_0^t)^{-1}|$ and set $Q = \mathrm{diag}(q)$ the Gram matrix of the scalar product induced by $\boldsymbol{P}(\mathrm{d}\boldsymbol{x})$, where $\Delta\boldsymbol{x}$ is the volume of one grid cell. In other words,

$$< u, v >_{\boldsymbol{P}(\mathrm{d}\boldsymbol{x})} = \int_\Omega u(\boldsymbol{x}) v(\boldsymbol{x}) |\det \nabla(\boldsymbol{\phi}_0^t)^{-1}| \mathrm{d}\boldsymbol{x} \simeq \sum_k u^k v^k q^k |\Delta\boldsymbol{x}| = u^T Q v \Delta\boldsymbol{x}.$$

3: Define $\boldsymbol{L} = Q^{-1}\boldsymbol{L}^{*T}$ so that the properties $\boldsymbol{L}\mathbf{1}_\Omega = \mathbf{1}_\Omega$ and $\boldsymbol{x}^T Q \boldsymbol{L} \boldsymbol{y} = \boldsymbol{x}^T \boldsymbol{L}^{*T}(t) \boldsymbol{y}$ for any $\boldsymbol{x}, \boldsymbol{y}$ are satisfied.

4: Compute the SVD of $Q^{\frac{1}{2}}\boldsymbol{L} = A\Sigma V^T$, with $A^T A = I$ and $V^T V = I$. Set $U = Q^{-\frac{1}{2}} A$. The SVD of $\boldsymbol{L}$ with respect to the starting and image spaces is given by

$$\boldsymbol{L} = U\Sigma V^T \text{ with } U^T Q U = I \text{ and } V^T V = I.$$

5: Obtain initial and final positions of the coherent sets by thresholding the level-sets of respectively the right and left eigenvectors V and U.

In the following, we therefore seek to evaluate the SVD of $\boldsymbol{L}_\epsilon$ as an approximate, regularized method to evaluate the eigenvectors of the dynamic laplacian.

1.3.2 An efficient matrix-free method for computing coherent sets for highly resolved velocity data

Estimating numerically the operators $\boldsymbol{P}(t)$ or $\boldsymbol{L} \in \mathcal{M}_{l,l}$ is more expensive than a simple flow map computation as in section 1.2. In the works of Froyland mentioned above, the space $L^2(\Omega)$ is discretized with the Ulam Galerkin method, that is popular in the dynamical system community (see e.g. chapter 4 in [15]). Ulam's method estimates the image of characteristic functions $\mathbf{1}_{B_i}$ associated with a finite partition $\Omega = \cup_i B_i$ of the domain. This is done through trajectory integration of particles initially located in B_i. This method is relatively fast for moderate spatial resolution as it is easily parallelizable and yields moderately sparse operator matrices [51, 52], but has the drawback of furnishing a numerical approximation of $\boldsymbol{P}(t)$ that has the resolution of the decomposition $\cup_i B_i$ and not of the number of particles employed. A sufficiently large number of particles must be integrated per subdomain B_i, therefore the method can be expensive for highly-resolved velocity fields. For instance, the numerical examples considered in [52] were advecting 400 particles for each cell of a 256x128 grid. Hence for works involving realistic ocean velocities, the computation of the transfer operators $\boldsymbol{P}$ or $\boldsymbol{L}$ was restricted to a subdomain of the area considered [48].

We show in algorithm 2 that it is in fact possible to reduce the computational cost by using a method that uses a number of particles that is identical to the grid resolution. We exploit the fact that computing the matrix-vector product $\boldsymbol{L}f$ in an argument $f \in \boldsymbol{L}^2(\Omega)$ can be achieved very efficiently. We apply this method on the three benchmark examples

Algorithm 2 Matrix free method for coherent sets extraction

1: One computes a numerical approximation of the inverse flow map $\boldsymbol{\phi}_0^{-t}$, as in section 1.2.3, using e.g. particle methods at the grid-resolution.
2: For any finite-dimensional approximation of a function f, one can compute $\boldsymbol{L}f = f \circ \boldsymbol{\phi}_0^{-t}$ and $\boldsymbol{L}^* f = f \circ \boldsymbol{\phi}_0^{t}$ without solving the advection equation (1.1) by using linear interpolation (e.g. mathrm in Matlab)
3: The "diffusive" approximation $\boldsymbol{L}_\epsilon(t)$ of $\boldsymbol{L}$ can be obtained by composition with a smoothing operator j_ϵ (see [52]):

$$\boldsymbol{L}_\epsilon = j_\epsilon \boldsymbol{L} j_\epsilon,\ \boldsymbol{L}_\epsilon^* = j_\epsilon \boldsymbol{L}^* j_\epsilon.$$

The smoothing operator j_ϵ can for instance the inverse laplacian $(I - \epsilon\Delta)^{-\frac{1}{2}}$ used above, or a power of a shapiro filter as an explicit approximation of this inverse ($\Delta_\epsilon = (\mathcal{F}^{(k)})^p$ for $p \in \mathbb{N}$ and $\mathcal{F}^{(k)}$ defined at eqn. (3.32) in chapter 3).
4: An iterative method (e.g. Lanczos or Arnoldi iterations, see [143]) that requires to evaluate only matrix-vector multiplications, can be used to estimate dominant eigenvectors of $\boldsymbol{L}_\epsilon \boldsymbol{L}_\epsilon^*$.

introduced in section 1.2.3. For convenience we assume free divergence of the velocity fields (that is we use $Q = I$ with the notations of algorithm 1) and we used the diffusive operator $\Delta_\epsilon = (\mathcal{F}^{(1)})^3$. To allow for comparison with previous available works, we also plotted on Figure 1-12 singular vectors corresponding to the double gyre with the parameters used in [52] ($A = 0.25, \epsilon = 0.25, \omega = 2\pi$ with the notations[1] of (1.7), integrated from $t = 0$ to $t = 2$ on a 256x128 grid). Results are plotted on Figs. 1-11 to 1-14. *Right* singular vectors, that correspond to the initial configuration, have been plotted on the *left* and vice-versa. We note the ability of the method to deal (i) with flows having outlets, by putting all the "mass" of the eigenvectors in the region of the flow that is not exiting the domain, (ii) highly resolved velocity fields (see the Palau example of Figure 1-14). Coherent partitions were extracted from the zero level sets of these eigenvectors, the color scale being set such that red and blue correspond to positive and negative values respectively. It is interesting to compare these figures to those of section 1.2.3 and observe the influence of high forward or backward FTLE values on the shape of the coherent sets in respectively the initial or final configuration. Note that the coherent sets identified with this criterion are somewhat different from the rigid sets found from the thresholding of the polar distance in eqn. (1.13): coherent sets partition the domain into regions that "mix" slowly but may allow large stretching within their boundaries. This operator method yields therefore an information that is complementary to the one obtained from the polar distance, as observed by comparing the previous figures with Figure 1-10.

1.3.3 A DO "infinitesimal operator" approach

In this part, we discuss briefly another approach to estimate coherent sets, that could be used, for example, in the case where computing many Lanczos iterations would be too costly.

[1] $\epsilon = 0.25$ being not the diffusivity parameter but the one of the Double Gyre example.

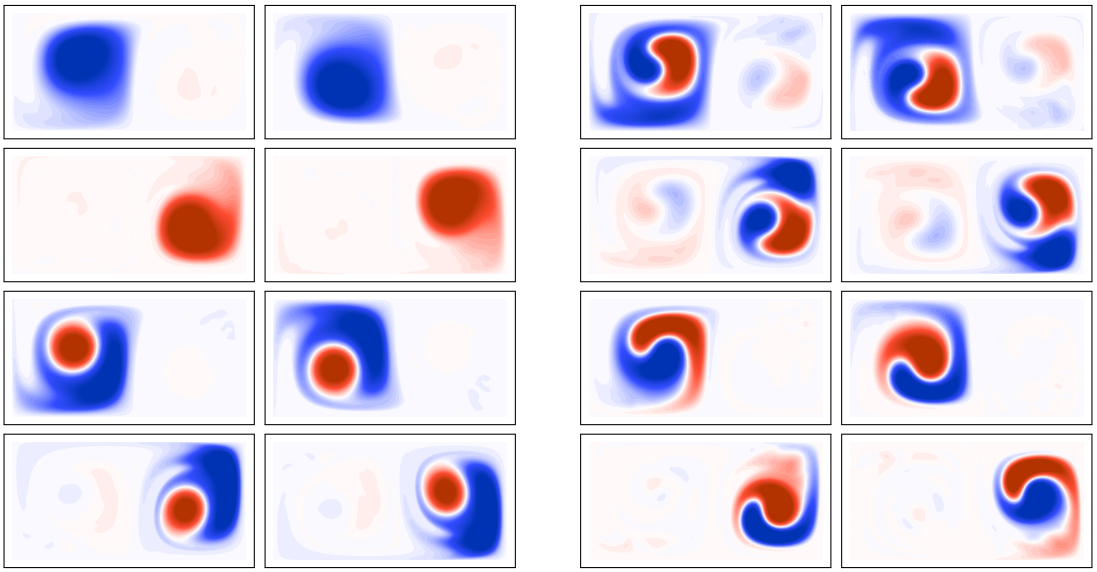


Figure 1-11: Pairs of corresponding right and left singular vectors number 1 to 8 of the diffusive operator $\boldsymbol{L}_\epsilon(t)$ for the Double Gyre Example (as defined in section 3.4.1).

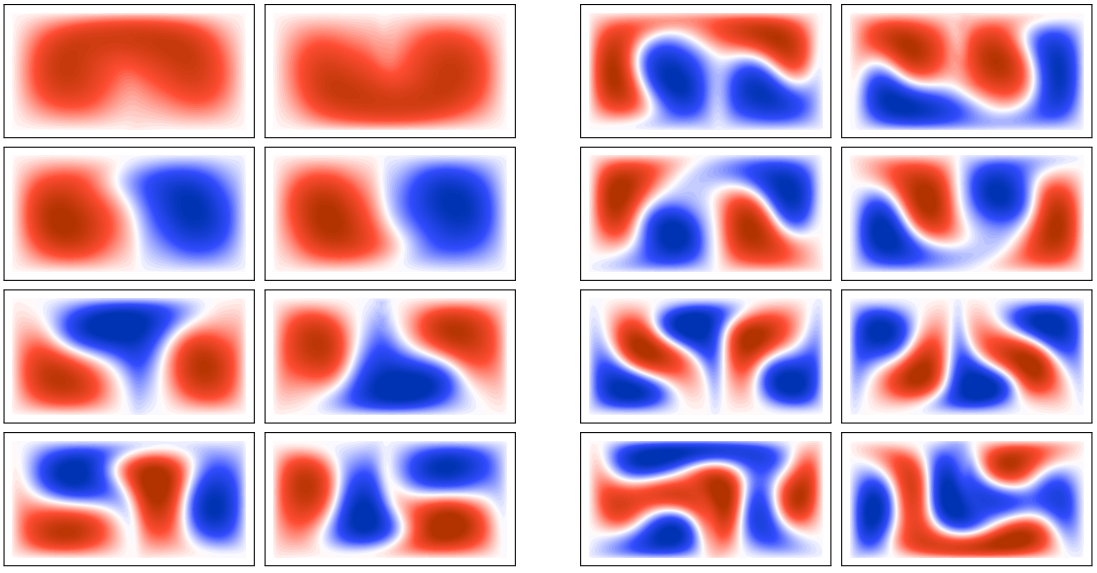


Figure 1-12: Pairs of corresponding right and left singular vectors number 1 to 8 of the diffusive operator $\boldsymbol{L}_\epsilon(t)$ for the Double Gyre Example (as defined in [52], for comparison).

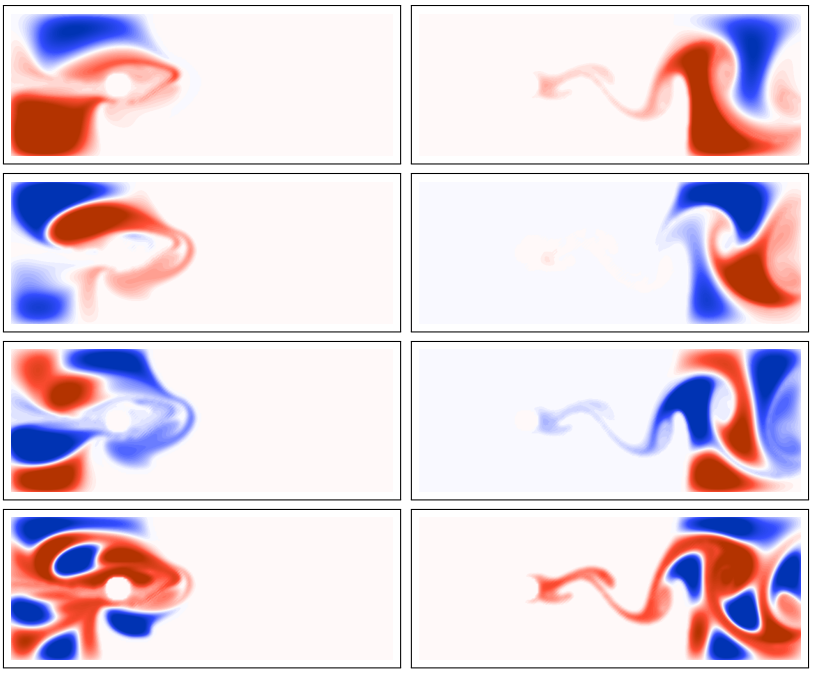


Figure 1-13: Pairs of corresponding right and left singular vectors number 2,3,5 and 12 of the diffusive operator $\boldsymbol{L}_\epsilon(t)$ for the Flow Past A Cylinder Example.

In fact the method reduces somehow to estimate singular vectors with a single Lanczos iteration and from a "good" initial guess value. In general the main computational cost of the operator method comes from the fact that one needs to estimate the singular vectors of the matrix $\boldsymbol{L}$ that has large dimensions and is dense. In view of the model order reduction methods that will be developed in the next chapters, the DO methodology is natural: one can evolve a low-rank approximation L of $\boldsymbol{L}$ and obtain at a low cost approximate singular vectors.

To achieve this goal, consider a fully linear advection schemes for the transport PDE (1.1) (such will be developed in section 3.3.1 of chapter 3) : computing $\boldsymbol{L}f_0$ is achieved by solving the transport PDE (1.1) forward in time with the initial data f_0. At the discrete level, this is done by solving an ODE of the form

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} f = A(t) f \\ f(0) = f_0, \end{cases} \tag{1.22}$$

where $A(t) \in \mathcal{M}_{l,l}$ is the matrix discretization of the operator $f \mapsto -\boldsymbol{v}(t,\cdot)\cdot\nabla f$ (that includes diffusivity, e.g. with artificial diffusion). Since $A(t)$ is a linear operator, the operator $\boldsymbol{L}$ is the resolvent of (1.22) and obtained as the solution of the matrix ODE

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{L} = A(t) \boldsymbol{L} \\ \boldsymbol{L}(0) = I. \end{cases} \tag{1.23}$$

This is naturally the discrete transposition of the fact that the unbounded operator $f \mapsto \boldsymbol{v}(t,\cdot)\cdot\nabla f$ is the infinitesimal generator of the semi-group of transformations $\boldsymbol{L}$ [87, 19,

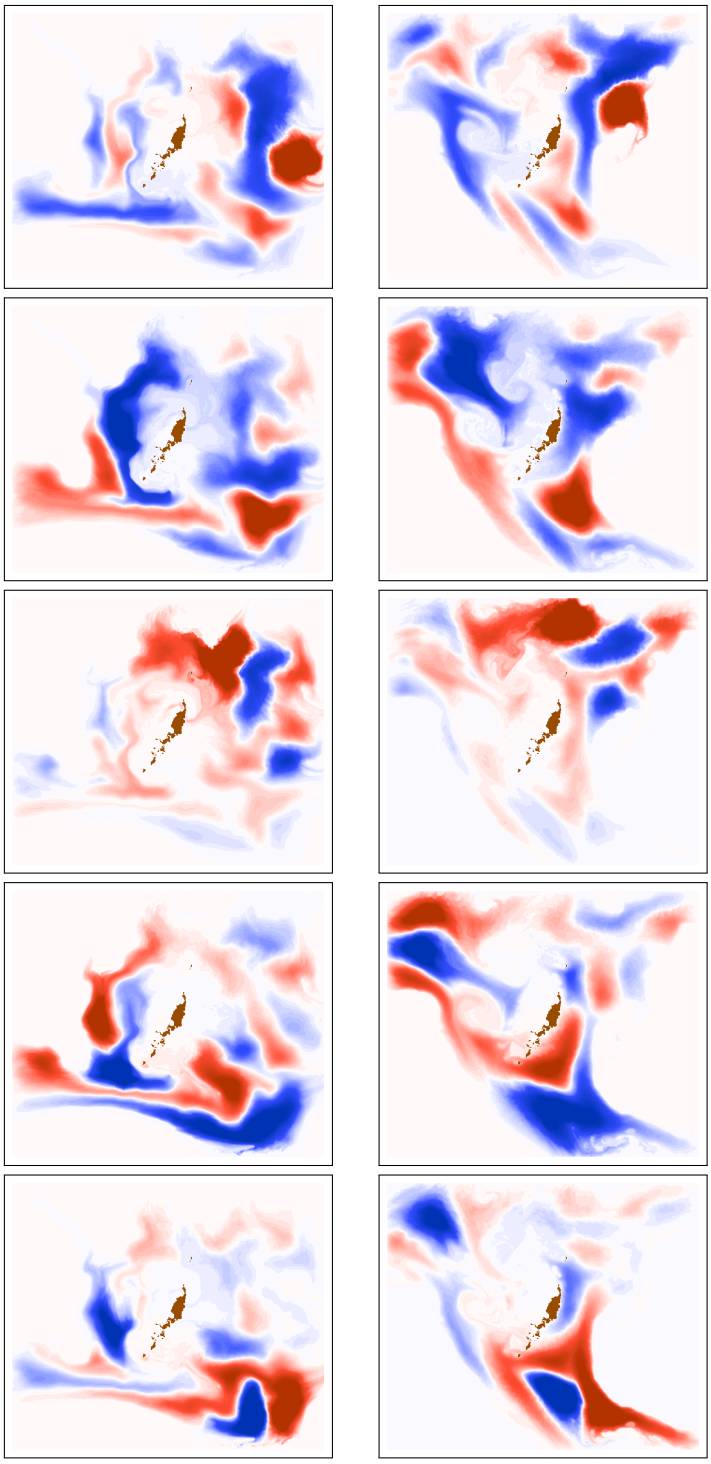


Figure 1-14: Pairs of corresponding right and left singular vectors number 3,7, 12, 17 and 20 of the diffusive koopman operator $\boldsymbol{L}_\epsilon(t)$ for the Palau Region. (1.24).

49]. Note that non-linear methods such as WENO or limiter-based finite-volumes schemes involve an operator $A(t)$ that is also non-linear.

In this setting, for convenience, we still assume that the velocity field $\boldsymbol{v}$ is divergent free, hence we consider $Q = I$ and algorithm 1 reduces to estimate directly the SVD of $\boldsymbol{L}$. A particular reason for that is that the formula (1.21) may produce a poor estimation of q due to the numerical diffusion induced by the numerical scheme itself, or because of the way boundary conditions are handled.

The DO methodology (see [124] or later on in section 2.2.1) applied to (1.23) seeks to evolve a low-rank approximation L of $\boldsymbol{L}$. Since initially $\boldsymbol{L} = I$, it is natural to initialize L as a projector on a low dimensional space of smooth functions: if r is the rank of the approximation, one sets

$$L(0) = U_0 U_0^T,$$

where $U_0 \in \mathcal{M}_{l,r}$ contains typically the r first eigenvectors of a smoothing operator (for example $-\Delta$ or $\mathcal{F} - I$ where $\mathcal{F}$ is a shapiro filter). With the notations of chapter 2, the DO approximation is set initially as $Z(0) = U(0)^T$. The operator $\mathcal{L}$ is $\mathcal{L}(t, R) = A(t)R$ and DO equations for modes and coefficients are written

$$\begin{cases} \dot{Z} = \mathcal{L}(t, UZ^T)^T U = Z(U^T A(t)^T U) \\ \dot{U} = (I - UU^T)\mathcal{L}(t, UZ^T)Z(Z^T Z)^{-1} = A(t)U - U(U^T A(t)U). \end{cases} \tag{1.24}$$

Alternatively, another way to build a low-rank approximation of $\mathcal{L}(t)$ is to estimate the image of the subspace spanned by U_0, *i.e.* to solve the transport PDE for each of the columns of U_0 and to obtain the low-rank approximation by approximating $\mathcal{L}(t)$ by its restriction on the subspace U_0:

$$L \simeq (\boldsymbol{L}U_0)U_0^T. \tag{1.25}$$

It turns out that eqn. (1.24) and (1.25) yield the same approximation $L(t)$. This fact can be seen can be seen by verifying that the DO solution $U(t)Z(t)^T$ of (1.24) and the matrix $(\boldsymbol{L}U_0)U_0^T$ are solution of a same ODE. More directly with the framework of chapter 2, this is an immediate consequence of the geometric fact that the vector field $\mathcal{L}$ in the matrix space $\mathcal{M}_{l,l}$ is tangent to the manifold of r rank matrices: $\mathcal{L}(t, UZ^T) = A(t)UZ^T$ is of the form $\delta U Z^T + U\delta Z^T$ with $\delta U = A(t)U$ and $\delta Z = 0$. Hence the DO approximation (2.57) for an initial rank r data coincides with the exact time integration.

Remark 1.3. One can notice that the equation for the mode matrix U coincides with the equation for *OTD* modes introduced by Babaee and Sapsis in [13] with $A(t)$ being the linearized operator of the dynamics. One therefore sees that the subspace spanned by OTD modes coincides with the image of the initial subspace by the resolvant associated with the system (1.23).

Numerically, the formulation (1.25) can be interesting because it preserves the evolution of the subspace spanned by $\mathcal{L}U_0$, while columns of $\mathcal{L}U_0$ tends to align exponentially fast along dominant left singular vectors (see also the discussion in [13, 41]). Nevertheless computing directly $\mathcal{L}U_0$ by solving (1.23) directly does not require constant reorthonormalization and can be achieved by using the particle method described in algorithm 2.

We evaluated the method on the three benchmark example. Once the low-rank approximation is obtained, the SVD can be straightforwardly estimated by using algorithm 3 of chapter 2 (page 105). We plot the result of this method on Figs. 1-15 to 1-17 for the double-gyre with the settings of [52], the Flow Past a Cylinder example and the Palau

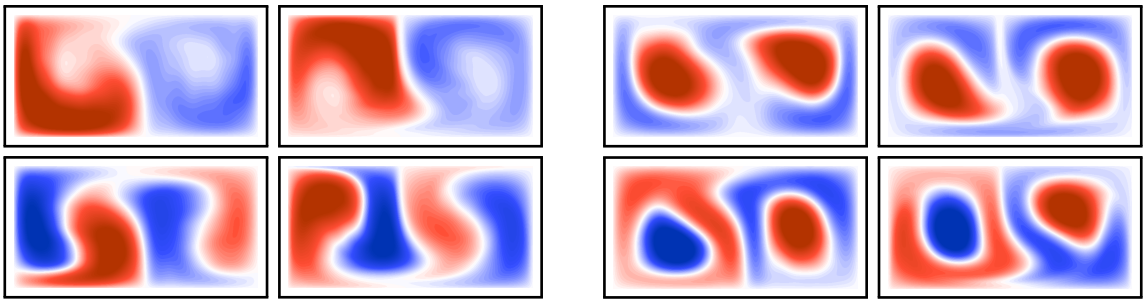


Figure 1-15: Coherent sets obtained from the 0 level-set thresholding of the approximated right and left singular vectors 2,4,7 and 8 of the push forward operator $\boldsymbol{L}$ computed by using the DO approximation (1.25) with $r = 150$ modes.

region. Note that a relatively large number r of modes is needed to obtain moderately accurate coherent sets. The method yields moderately accurate left singular vectors but more poorly estimates of the right singular vectors because these belong to the initial space $\mathrm{Span}(U_0)$. An additional Lanczos iteration (applying this method backward in time) would allow to increase the accuracy.

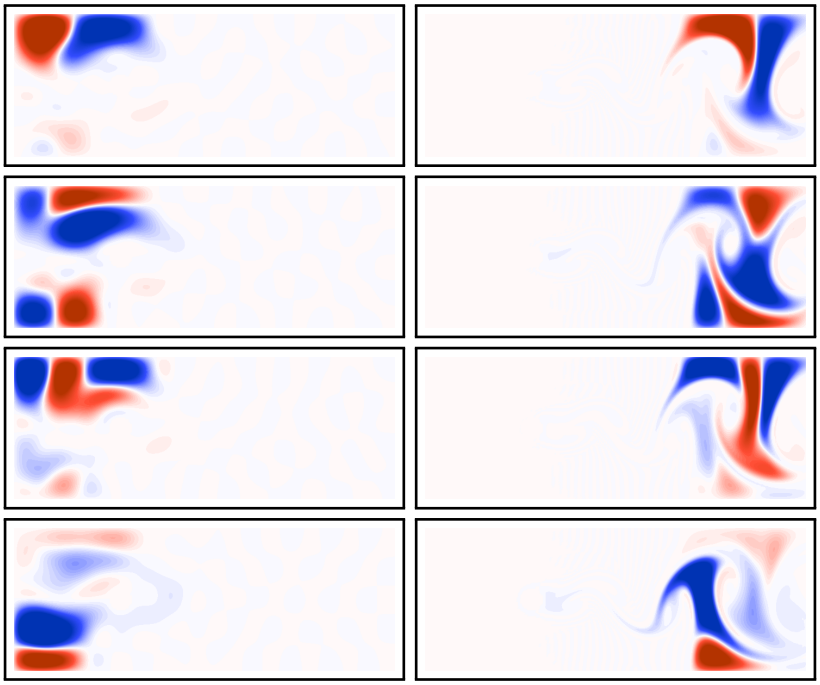


Figure 1-16: Coherent sets for the Flow Past a Cylinder example obtained from the 0 level-set thresholding of the approximated right and left singular vectors 3,4,6 and 7 of the push forward operator $\boldsymbol{L}$ computed by using the DO approximation (1.25) with $r = 150$ modes.

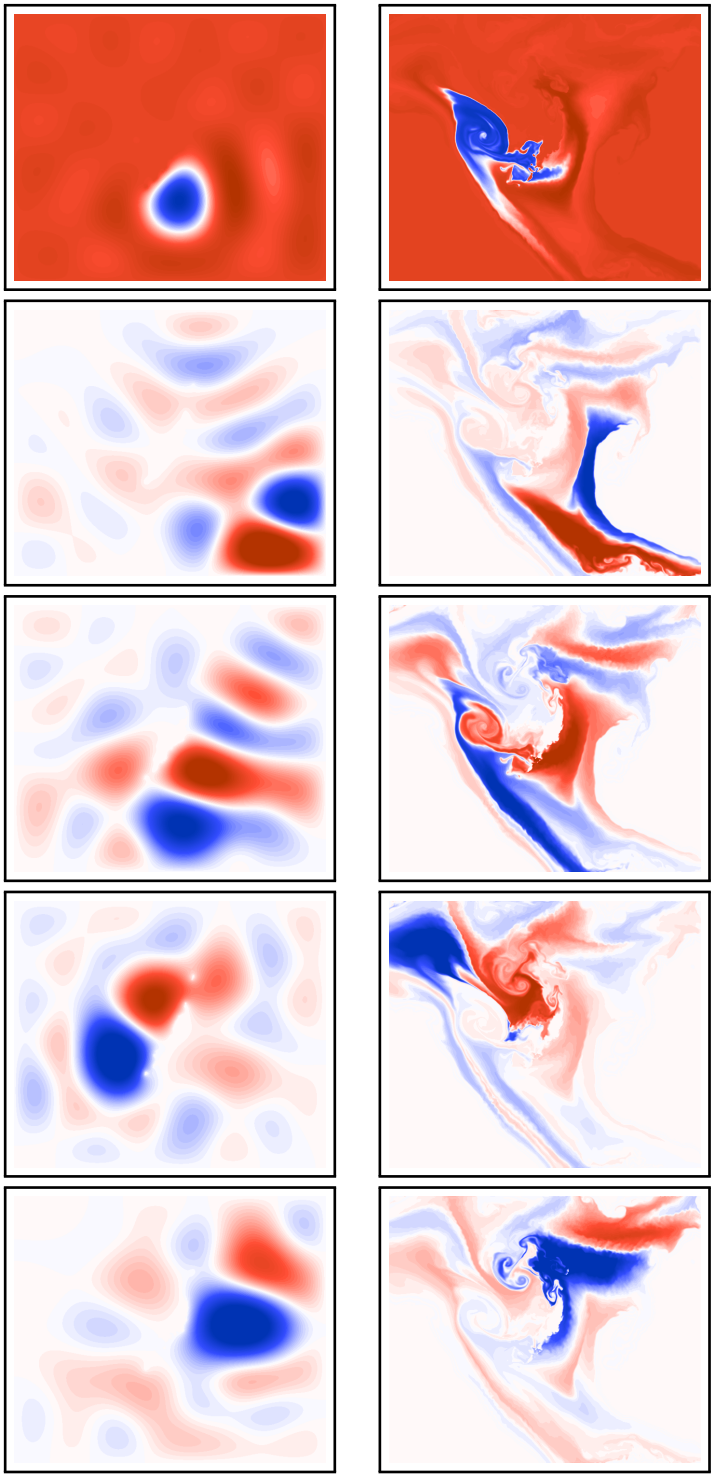


Figure 1-17: Coherent sets for the Palau Region obtained from the 0 level-set thresholding of the approximated right and left singular vectors 1,4,5,7 and 17 of the operator $\boldsymbol{L}$ computed by using a DO approximation with $r = 60$ modes (1.24).

Chapter 2

Embedded geometry of matrix manifolds and dynamic approximation

Efficient and rigorous numerical schemes for the stochastic advection equation that will be presented in chapter 3 require a rigorous understanding of the method (Dynamically Orthogonal approximation) developed for stochastic PDEs by Sapsis and Lermusiaux in 2009 [124]. Following the prior analysis of Koch and Lubich [83] and Musharbash [103], Riemannian geometry over matrix manifolds is an appropriate mathematical framework to understand and analyze the method. In this chapter, we investigate the foundations of the mathematical framework.

We start by reviewing the necessary background material relative to Riemannian geometry on embedded manifolds in section 2.1. Most of the results of this section can be found in classical literature [139, 11] but we have attached a particular importance on using tensor free notations, which will turn to be especially convenient when computing geometric quantities on matrix manifolds. This is achieved by expressing these almost exclusively in terms of the projection operator onto tangent spaces and its differential. With these notations, we provide definitions of the Weingarten map and of the extrinsic or principal curvatures of the manifold with respect to a normal direction. The most important result is theorem 2.1, a restatement of a result available in [11], that expresses the differential of the orthogonal projection onto an embedded manifold in term of these curvatures. We investigate then in section 2.1.3 a generalization of the framework to manifolds embedded in finite dimensional spaces that are not necessary euclidean.

We apply this setting to matrix manifolds in section 2.2. We evaluate explicitely the tangent and normal spaces, geodesics equations and principal curvatures for the fixed-rank, the Stiefel and the Isospectral manifolds. For each of these, the orthogonal projection onto the manifold is related to an algebraic operation, e.g. polar decomposition for the Orthogonal group, truncated Singular Value Decomposition for the fixed-rank manifold and projectors over the eigenspaces of symmetric matrices for the Isospectral manifold. Having computed principal curvatures, we are able to provide explicit formulas for the differential of these operations. We also apply the non-euclidean generalization to study the differentiability of the eigenspaces of non-symmetric matrices, those map being not orthogonal projections but sharing similar characteristics.

Finally, we study in section 2.3 the approximation of a dynamical system by mean of a

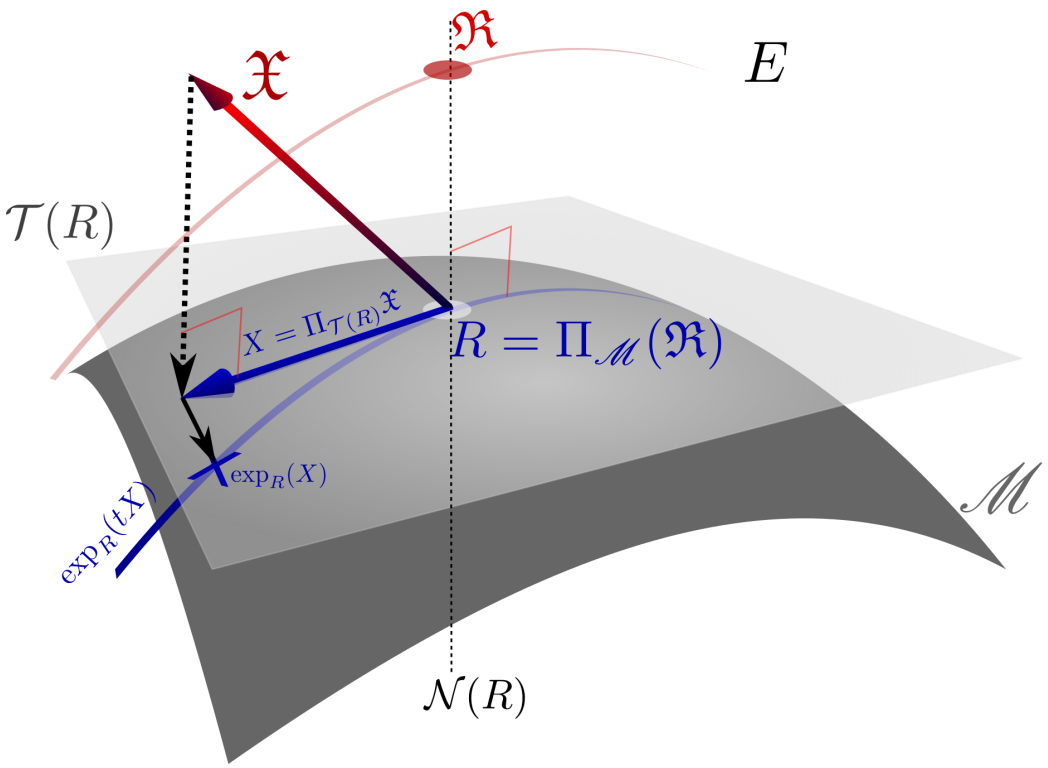


Figure 2-1: Illustration of an embedded manifold $\mathscr{M} \subset E$ and geometric concept of interests.

reduced dynamical system set on a manifold (of small dimension). This approximation, that consists in replacing the vector field of the original dynamical system with its tangent projection onto the manifold turns to be exactly the *Dynamically Orthogonal* or the *Dynamical low-rank* methods introduced by Sapsis and Lermusiaux [124] and Koch and Lubich [83] respectively, when the manifold considered is the set of fixed-rank matrices. We perform an error analysis that extends the results of [83] and provide geometric interpretation of the conditions under which the approximation error remains controlled.

Important geometric concepts that are the matter of this chapter are illustrated on Figure 2-1.

2.1 Background material: Extrinsic geometry on Riemannian manifolds

2.1.1 Tangent space, normal space, metric and geodesics

Let E be a finite dimensional euclidean space.

Definition 2.1. An embedded smooth manifold $\mathscr{M}$ of dimension d is a subset of E that can be parameterized locally with d coordinates: for every point $R \in \mathscr{M}$ there exists an open set $U \subset \mathbb{R}^d$, an open neighborhood $V \subset E$ of R and a $\mathcal{C}^\infty$ diffeomorphism $\boldsymbol{\phi} : U \to V \cap \mathscr{M}$ that is called a local coordinate chart.

Definition 2.2. The tangent space $\mathcal{T}(R) \subset E$ at R is the set of all velocity vectors $\dot{R}(0)$

at time 0 for a $\mathcal{C}^1$ curve $R(t)$ drawn on $\mathscr{M}$ satisfying $R(0) = R$. This is a d dimensional vector space. The normal space $\mathcal{N}(R)$ is defined to be the orthogonal complement of $\mathcal{T}(R)$.

The main object of interest in embedded differential geometry is the orthogonal projection $\Pi_{\mathcal{T}(R)}$ onto the tangent space $\mathcal{T}(R)$ at a point R on $\mathscr{M}$. This map projects displacements $\mathfrak{X} = \dot{\mathfrak{R}} \in \mathbb{E}$ of a matrix $\mathfrak{R}$ of the ambient space E to the tangent directions $X = \Pi_{\mathcal{T}(R)}\mathfrak{X} \in \mathcal{T}(R)$. Most of geometric quantities defined later on are obtained from this map.

Definition 2.3. We denote $\Pi_{\mathcal{T}(R)} : E \to \mathcal{T}(R)$ the orthogonal projection onto the tangent space, that is for any vector $\mathfrak{X} \in E$, $\Pi_{\mathcal{T}(R)}$ satisfies

$$||\mathfrak{X} - \Pi_{\mathcal{T}(R)}\mathfrak{X}|| = \min_{X \in \mathcal{T}(R)} ||\mathfrak{X} - X||.$$

The map $R \mapsto \Pi_{\mathcal{T}(R)}$ is of class $\mathcal{C}^\infty$ for smooth manifold $\mathscr{M}$.

Proof. Consider $x \in \mathscr{M}$ and a local coordinate map $\boldsymbol{\phi} : \mathbb{R}^p : E$ such that $\mathscr{M} = \boldsymbol{\phi}(\mathbb{R}^p)$. Then the tangent space at x is $\mathcal{T}(x) = \text{span}(\mathrm{D}\boldsymbol{\phi}(x))$ and the projector on $\mathcal{T}(x)$ is $\Pi_{\mathcal{T}(x)} = \mathrm{D}\boldsymbol{\phi}(x)(\mathrm{D}\boldsymbol{\phi}(x)^T\mathrm{D}\boldsymbol{\phi}(x))^{-1}\mathrm{D}\boldsymbol{\phi}(x)$ which is a map of class $\mathcal{C}^{k-1}$ if $\boldsymbol{\phi}$ is of class $\mathcal{C}^k$. □

A metric on $\mathscr{M}$ defines how distances are measured on the manifold, by prescribing a smoothly varying scalar product on each tangent space. Since E is an Euclidean space, any embedded manifold $\mathscr{M}$ is a Riemannian manifold with the metric induced by the scalar product $< \cdot, \cdot >$ of E. In the language of Riemannian geometry, the value g_R of the metric g at the point R is the bilinear form over $\mathcal{T}(R)$ defined by $g_R(X, Y) =< X, Y >$ for any tangent vectors $X, Y \in \mathcal{T}(R)$.

In differential geometry, one distinguishes the geometric properties that are *intrinsic*, *i.e.* that depend only on the metric g defined on the manifold, from the ones that are *extrinsic*, *i.e.* that depend on the ambient space in which the manifold $\mathscr{M}$ is defined. The following proposition recalls the link between the extrinsic projection $\Pi_{\mathcal{T}(R)}$ and the intrinsic notion of derivation onto a manifold. For embedded manifolds, *i.e.* defined as subsets of an ambient space, the covariant derivative at $R \in \mathscr{M}$ is obtained by projecting the usual derivative onto the tangent space $\mathcal{T}(R)$, and the Christoffel symbol corresponds to the normal component that has been removed [36].

Proposition 2.1. *Let X and Y be two tangent vector fields defined on a neighborhood of $R \in \mathscr{M}$. The covariant derivative $\nabla_X Y$ with respect to the metric inherited from the ambient space is the projection of $\mathrm{D}_X Y$ onto the tangent space $\mathcal{T}(R)$:*

$$\nabla_X Y = \Pi_{\mathcal{T}(R)}(\mathrm{D}_X Y).$$

The Christoffel symbol $\Gamma(X, Y)$ is defined by the relationship $\nabla_X Y = \mathrm{D}_X Y + \Gamma(X, Y)$ and is characterized by the formula

$$\Gamma(X, Y) = -(I - \Pi_{\mathcal{T}(R)})\mathrm{D}_X Y = -\mathrm{D}\Pi_{\mathcal{T}(R)}(X) \cdot Y. \tag{2.1}$$

The Christoffel symbol is symmetric: $\Gamma(X, Y) = \Gamma(Y, X)$.

Proof. (see also [139], Vol.3, Ch.1.) The first fact comes from

$$\begin{aligned} \mathrm{D}_Z < X, Y > &=< \mathrm{D}_Z X, Y > + < X, \mathrm{D}_Z Y > \\ &=< \Pi_{\mathcal{T}(R)} \mathrm{D}_Z X, Y > + < X, \Pi_{\mathcal{T}(R)} \mathrm{D}_Z Y > \\ &=< \nabla_Z X, Y > + < X, \nabla_Z Y >, \end{aligned}$$

which is exactly the requirement that the connection is compatible with the metric. The second fact comes from the definition of $\Gamma(X,Y)$ as $\Gamma(X,Y) = \nabla_X Y - D_X Y$, and by differentiation of the equality $\Pi_{\mathcal{T}(R)} Y = Y$ along the direction X. The third fact reflects that the Lie bracket $[X,Y] = D_X Y - D_Y X$ lies in the tangent space: $(I - \Pi_{\mathcal{T}(R)})(\mathrm{D}_X Y - \mathrm{D}_Y X) = 0$. A proof is obtained as follows: consider a local coordinate chart $\boldsymbol{\phi} : U \subset \mathbb{R}^d \to \mathscr{M}$ and $\mathfrak{R} = \boldsymbol{\phi}(u)$. Any tangent vector X can be written as $X = \mathrm{D}\boldsymbol{\phi}(u)(x)$ for some $x \in \mathbb{R}^d$. Denote $Y = \mathrm{D}\boldsymbol{\phi}(u)(y)$. Then one can check that $\mathrm{D}_Y X - \mathrm{D}_X Y = \left[\mathrm{D}^2\boldsymbol{\phi}(u)(x,y) - \mathrm{D}^2\boldsymbol{\phi}(u)(y,x)\right] + \mathrm{D}\boldsymbol{\phi}(u)(\mathrm{D}_y x - \mathrm{D}_x y)$ where the bracket term vanishes because of Schwartz theorem. Hence $[X,Y] \in \mathrm{Span}(\mathrm{D}\boldsymbol{\phi}) = \mathcal{T}(R)$. □

Remark 2.1. An important feature of equation (2.1) is that the Christoffel symbol $\Gamma(X,Y)$ depends only on the projection map Π_T at the point R and not on neighboring values of the tangent vectors X, Y, which is *a priori* not clear from the equality $\Gamma(X,Y) = -(I - \Pi_{\mathcal{T}(R)})\mathrm{D}_X Y$.

The covariant derivative allows to obtain equations for the geodesics of the manifold $\mathscr{M}$. These geodesics (Figure 2-1) are the shortest paths among all possible smooth curves drawn on $\mathscr{M}$ joining two points sufficiently close.

Definition 2.4. A geodesic $R(t)$ on $\mathscr{M}$ is a curve satisfying either of the following properties :

- $R(t)$ is a stationary curve joining two given points for the perimeter functional $R(t) \mapsto S(R(t)) = \int_0^1 ||\dot{R}(t)|| \mathrm{d}t$.
- the acceleration $\ddot{R}(t) \in \mathcal{N}(R)$ lies in the normal space to $\mathscr{M}$ at all instants ([36, 139])
- the velocity $\dot{R}$ is stationary under the covariant derivative, *i.e.*

$$\nabla_{\dot{R}} \dot{R} = \ddot{R} - \mathrm{D}\Pi_{\mathcal{T}(R)}(\dot{R}) \cdot \dot{R} = 0. \tag{2.2}$$

Proof. (see also [139]) Consider a tangent variation $\delta R(t) \in \mathcal{T}(R(t))$ of $R(t)$. Then

$$\frac{\partial S}{\partial R} \cdot \delta R = \int_0^1 < \frac{\dot{R}}{||\dot{R}||}, \delta \dot{R} > \mathrm{d}t = -\int_0^1 < \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\dot{R}}{||\dot{R}||}\right), \delta R > \mathrm{d}t$$

must vanish for all tangent variations $\delta R \in \mathcal{T}(R(t))$. Consider $\gamma(t)$ such that $||R(\gamma(t))||$ is constant (note that S is invariant under such change of parameterization). The stationary condition writes then $\ddot{R}(\gamma(t)) \in \mathcal{N}(R(\gamma(t)))$ for all $t \in [0,1]$. Differentiating the relation $\dot{R} = \Pi_{\mathcal{T}(R)}(\dot{R})$ yields

$$\ddot{R} = \mathrm{D}\Pi_{\mathcal{T}(R(t))}(\dot{R}(t))\dot{R}(t) + \Pi_{\mathcal{T}(R(t))}\ddot{R}(t).$$

Thus $R(t)$ is stationary curves for S if and only if $R(\gamma^{-1}(t))$ is a solution of (2.2). □

Geodesics allow to define the exponential map and parallel transport [139], which indicate respectively how to walk on the manifold from a point $R \in \mathscr{M}$ along a straight direction $X \in \mathcal{T}(R)$ and how to transfer tangent vectors from one point to another.

Definition 2.5. The exponential map $\exp_R$ at $R \in \mathscr{M}$ is the application

$$\begin{array}{rccc} \exp_R : & \mathcal{T}(R) & \to & \mathscr{M} \\ & X & \mapsto & R(1), \end{array} \tag{2.3}$$

where $R(1)$ is the value at time 1 of the solution of the geodesic equation (2.32) with initial conditions $R(0) = R$ and $\dot{R}(0) = X$. The value of $\tau_{RR(1)} = \dot{R}(1)$, is called is called the parallel transport of X from R to $R(1)$.

2.1.2 Curvature and differentiability of the orthogonal projection

Differentiability results for the orthogonal projection onto smooth embedded manifolds, as presented with tensor notations in [11], are now centralized and restated with notations using only the orthogonal projection $\Pi_{\mathcal{T}(R)}$. The main motivation is that algebraic operations on matrices such as SVD truncation (section 2.2.1) or the polar decomposition (section 2.2.2) can be directly related to orthogonal projections onto matrix manifolds. Hence general geometric differentiability results for the projections transpose directly into formulas for the differential of these algebraic matrix operations.

Definition 2.6. The orthogonal projection of a point $\mathfrak{R}$ onto $\mathscr{M}$ is defined whenever there is a unique point of $\mathscr{M}$, denoted $\Pi_{\mathscr{M}}(\mathfrak{R})$, minimizing the Euclidean distance $R \mapsto ||\mathfrak{R} - R||$ to $\mathfrak{R}$. When this occurs, the residual $\mathfrak{R} - \Pi_{\mathscr{M}}(\mathfrak{R})$ must be normal to $\mathscr{M}$ at $\Pi_{\mathscr{M}}(\mathfrak{R})$, namely

$$\mathfrak{R} - \Pi_{\mathscr{M}}(\mathfrak{R}) \in \mathcal{N}(\Pi_{\mathscr{M}}(\mathfrak{R})) \Leftrightarrow \Pi_{\mathcal{T}(\Pi_{\mathscr{M}}(\mathfrak{R}))}(\mathfrak{R} - \Pi_{\mathscr{M}}(\mathfrak{R})) = 0. \tag{2.4}$$

Proof. For any tangent vector $X \in \mathcal{T}(R)$, consider a curve $R(t)$ drawn on $\mathscr{M}$ such that $R(0) = R$ and $\dot{R}(0) = X$ where R is minimizing $J(R) = \frac{1}{2}||\mathfrak{R} - R||^2$. Then the stationary condition $\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} J(R(t)) = - < \mathfrak{R} - R, X >= 0$ states precisely (2.4). □

Remark 2.2. The normality of $\mathfrak{R} - R$ for the point $R = \Pi_{\mathscr{M}}(\mathfrak{R})$ being the orthogonal projection of $\mathfrak{R}$ onto $\mathscr{M}$ is geometrically illustrated on Figure 2-1.

We motivate the introduction of the Weingarten map and extrinsic curvatures by the following observation (also present in the proofs of [83]):

Proposition 2.2. *Suppose the projection $\Pi_{\mathscr{M}}$ is defined and differentiable at $\mathfrak{R}$. Then the differential $\mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R})$ of $\Pi_{\mathscr{M}}$ at the point $\mathfrak{R}$ in the direction $\mathfrak{X} \in E$ satisfies :*

$$\mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R}) = \Pi_{\mathcal{T}(\Pi_{\mathscr{M}}(\mathfrak{R}))}(\mathfrak{X}) + \mathrm{D}\Pi_{\mathcal{T}(\Pi_{\mathscr{M}}(\mathfrak{R}))}(\mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R})) \cdot (\mathfrak{R} - \Pi_{\mathscr{M}}(\mathfrak{R})). \tag{2.5}$$

Proof. Differentiating equation (2.4) along the direction $\mathfrak{X}$ yields

$$\mathrm{D}\Pi_{\mathcal{T}(\Pi_{\mathscr{M}}(\mathfrak{R}))}(\mathrm{D}\Pi_{\mathscr{M}}(\mathfrak{R})(\mathfrak{X})) \cdot (\mathfrak{R} - \Pi_{\mathscr{M}}(\mathfrak{R})) + \Pi_{\mathcal{T}(\Pi_{\mathscr{M}}(\mathfrak{R}))}(\mathfrak{X} - \mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R})) = 0.$$

Since $\Pi_{\mathscr{M}}(\mathfrak{R}) \in \mathscr{M}$ for any $\mathfrak{R}$, the differential $\mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R})$ is a tangent vector, and the results follows from the relation $\Pi_{\mathcal{T}(\Pi_{\mathscr{M}}(\mathfrak{R}))}(\mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R})) = \mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R})$. □

Let $R = \Pi_{\mathscr{M}}(\mathfrak{R})$ be the projection of the point $\mathfrak{R}$ on $\mathscr{M}$ and $N = \mathfrak{R} - R$ the corresponding normal residual vector. Solving (2.5) for the differential $X = \mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R})$ requires to invert the linear operator $I - L_R(N)$ where $L_R(N)$ is the map $X \mapsto \mathrm{D}\Pi_{\mathcal{T}(R)}(X) \cdot N$. $L_R(N)$ would be zero if $\mathscr{M}$ were to be a "flat" vector subspace and can be interpreted as a curvature correction. $L_R(N)$ is called the Weingarten map and turns to be a symmetric endomorphism on $\mathcal{T}(R)$ whose eigenvalues are by definition the principal curvatures.

Definition 2.7 (Weingarten map)**.** For any point $R \in \mathscr{M}$, tangent and normal vector fields $X, Y \in \mathcal{T}(R)$ and $N \in \mathcal{N}(R)$ defined on a neighborhood of R, the following relation, called *Weingarten identity* holds:

$$< \Pi_{\mathcal{T}(R)}(\mathrm{D}_X N), Y >=< N, \Gamma(X, Y) > . \tag{2.6}$$

Also, the tangent variation $\Pi_{\mathcal{T}(R)}(\mathrm{D}_X N)$ depend only on the value of the normal vector field N at R as it can be seen from the identity

$$\mathrm{D}\Pi_{\mathcal{T}(R)}(X) \cdot N = -\Pi_{\mathcal{T}(R)}(\mathrm{D}_X N). \tag{2.7}$$

The application

$$\begin{array}{rccc} L_R(N) & : & \mathcal{T}(R) & \to & \mathcal{T}(R) \\ & & X & \mapsto & \mathrm{D}\Pi_{\mathcal{T}(R)}(X) \cdot N, \end{array}$$

is therefore a symmetric map of the tangent space into itself and is called the Weingarten map in the normal direction N. The corresponding eigenvectors and eigenvalues are respectively called the *principal directions* and *principal curvatures* of $\mathscr{M}$ in the normal direction N. The induced symmetric bilinear form on the tangent space,

$$\mathrm{II}(N) : \ (X, Y) \mapsto < L_R(N)X, Y >= - < N, \Gamma(X, Y) >, \tag{2.8}$$

is called the second fundamental form in the direction N.

Proof. (See also [135] or the proof Theorem 5 of [139], vol.3, ch.1.) Differentiating the relation $\Pi_{\mathcal{T}(R)}(N) = 0$ along a direction X yields equation (2.7). Also, since the vector field N is normal to $\mathscr{M}$ on a neighborhood of R, differentiating the relation $< N, X >= 0$ in the Y directions allows to obtain $< \mathrm{D}_Y N, X >= - < N, D_X Y >$. Now, the identity follows from the series of equalities

$$\begin{array}{ll} < \Pi_{\mathcal{T}(R)}(\mathrm{D}_X N), Y > & =< \mathrm{D}_X N, Y >= - < N, D_Y X > \\ & = - < N, (I - \Pi_{\mathcal{T}(R)}(X))\mathrm{D}_Y X >=< N, \Gamma(X, Y) > . \end{array}$$

□

Remark 2.3. The notion of Weingarten map and principal curvatures is maybe more commonly encountered in the literature on differential geometry for hypersurfaces, where one can define $L_R(N)X = \mathrm{D}_X N$ for a given normal vector (whose direction is unique since $\dim(\mathcal{N}(R)) = 1$. Nevertheless the more general definition definition 2.7 is present in [135, 11, 5, 3, 139].

The Weingarten map is related to the covariant Hessian (see also [4, 5]), which will be useful in the proof of the next theorem, and also for developing optimization onto the fixed rank manifold later on in section 3.3.3.

Definition 2.8. Let J a smooth function defined on $\mathscr{M}$ and $R \in \mathscr{M}$. The covariant gradient of J at R is the unique vector $\nabla J \in \mathcal{T}(R)$ such that

$$\forall X \in T_R,\ J(\exp_R(tX)) = J(R) + t < \nabla J, X > +o(t).$$

The covariant Hessian $\mathcal{H}J$ of J at R is the linear map on $\mathcal{T}(R)$ defined by

$$\mathcal{H}J(X) = \nabla_X \nabla J,$$

and the following second order Taylor approximation of J holds:

$$J(\exp_R(tX)) = J(R) + t < \nabla J, X > + \frac{t^2}{2} < X, \mathcal{H}J(X) > +o(t^2).$$

The following proposition (see [5]) explains how these quantities are related to the usual gradient and Hessian, so that they become accessible for computations.

Proposition 2.3. *Let J be a smooth function defined in the ambient space E and denote $\mathrm{D}J$ and D^2J its respective euclidean gradient and Hessian. Then the covariant gradient and Hessian are given by*

$$\nabla J = \Pi_{\mathcal{T}(R)}(\mathrm{D}J), \tag{2.9}$$

$$\mathcal{H}J(X) = \Pi_{\mathcal{T}(R)}(\mathrm{D}^2J(X)) + \mathrm{D}\Pi_{\mathcal{T}(R)}(X) \cdot \left[(I - \Pi_{\mathcal{T}(R)})(\mathrm{D}J)\right]. \tag{2.10}$$

Proof. Using the compatibility of the connection ∇_X with the metric:

$$\begin{aligned} < \mathcal{H}JX, Y > &=< \nabla_X \nabla J, Y >= \mathrm{D}_X < \nabla J, Y > - < \nabla_X Y, \nabla J > \\ &= \mathrm{D}_X(\mathrm{D}_Y J) - < \mathrm{D}J, \mathrm{D}_X Y > - < \Gamma(X,Y), \mathrm{D}J > \end{aligned}$$

the second lines using the decomposition $\nabla_X Y = \mathrm{D}_X Y + \Gamma(X,Y)$. Checking (using some system of coordinates) that $\mathrm{D}_X(\mathrm{D}_Y J) =< X, \mathrm{D}^2J(Y) > + < \mathrm{D}J, \mathrm{D}_X Y >$, one obtains finally the identity

$$\forall X, Y \in \mathcal{T}(R), < \mathcal{H}J(X), Y >=< \mathrm{D}^2J(X), Y > - < \Gamma(X,Y), \mathrm{D}J > . \tag{2.11}$$

Now, equation (2.10) follows by using the Weingarten identity (2.6). □

The differentiability of the projection map for arbitrary sets has been studied in [153, 2] and more recently in the context of smooth manifolds in [11, 55, 24] with recent applications in shape optimization [8]. The following theorem reformulates these results in tensor-free notations. The proof given in the following is essentially a justification that one can indeed invert the operator $I - L_R(N)$ in (2.5) by using its eigendecomposition.

Recall that the closure $\overline{\mathscr{M}}$ is the set of limit points of $\mathscr{M}$ and we refer to the set $\partial\mathscr{M} = \overline{\mathscr{M}} \backslash \mathscr{M}$ as the *boundary* of the manifold $\mathscr{M}$ (we have therefore $\partial\mathscr{M} \cap \mathscr{M} = \emptyset$: points on the boundary are excluded of $\mathscr{M}$). The skeleton of $\mathscr{M}$ (see e.g. [32]) is the set $\mathrm{Sk}(\mathscr{M})$ of all points that admit at least two possible minimizer of $||\mathfrak{R} - R||$ on $\overline{\mathscr{M}}$ (note that $\Pi_{\mathscr{M}}$ is defined when there is a unique minimizer $R \in \overline{\mathscr{M}}$ and that in addition $R \in \mathscr{M}$).

Theorem 2.1. *Let $\Omega \subset E$ be an open set of E over which $\Pi_{\mathscr{M}}$ is defined and such that $\Omega \cap \overline{Sk(\mathscr{M})} = \emptyset$. For $\mathfrak{R} \in \Omega$, denote $\kappa_i(N)$ and Φ_i the respective eigenvalues and eigenvectors of the Weingarten map $L_R(N)$ at $R = \Pi_{\mathscr{M}}(\mathfrak{R})$ with the normal direction $N = \mathfrak{R} - \Pi_{\mathscr{M}}(\mathfrak{R})$.*

Then all the principal curvatures satisfy $\kappa_i(N) < 1$ and the projection $\Pi_{\mathscr{M}}$ is differentiable at $\mathfrak{R}$. The differential $\mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R})$ at $\mathfrak{R}$ in the direction $\mathfrak{X}$ satisfies

$$\begin{aligned}\mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R}) &= \sum_{\kappa_i(N)} \frac{1}{1-\kappa_i(N)} < \Phi_i, \mathfrak{X} > \Phi_i \\ &= \Pi_{T(\Pi_{\mathscr{M}}(R))}(\mathfrak{X}) + \sum_{\kappa_i(N)\neq 0} \frac{\kappa_i(N)}{1-\kappa_i(N)} < \Phi_i, \mathfrak{X} > \Phi_i.\end{aligned} \tag{2.12}$$

Proof. The proof of this theorem (see also [11]) is done in three steps:

Step 1: Under the conditions of theorem 2.1, the projection $\Pi_{\mathscr{M}}$ is continuous on Ω. Consider a sequence $\mathfrak{R}_n \in \Omega$ converging in E to $\mathfrak{R}$ and denote $\Pi_{\mathscr{M}}(\mathfrak{R}_n)$ the corresponding projections. Let $\epsilon > 0$ be a real such that $\forall n \geq 0, ||\mathfrak{R}_n - \mathfrak{R}|| < \epsilon$. Since

$$\begin{aligned}||\Pi_{\mathscr{M}}(\mathfrak{R}_n) - \mathfrak{R}|| &\leq ||\Pi_{\mathscr{M}}(\mathfrak{R}_n) - \mathfrak{R}_n|| + ||\mathfrak{R}_n - \mathfrak{R}|| \\ &\leq ||\mathfrak{R}_n - \Pi_{\mathscr{M}}(\mathfrak{R})|| + ||\mathfrak{R}_n - \mathfrak{R}|| \\ &\leq 2\epsilon + ||\mathfrak{R} - \Pi_{\mathscr{M}}(\mathfrak{R})||,\end{aligned}$$

the sequence $\Pi_{\mathscr{M}}(\mathfrak{R}_n)$ is bounded. Denote $R \in \overline{\mathscr{M}}$ a limit point of this sequence. Passing to the limit the inequality $||\mathfrak{R}_n - \Pi_{\mathscr{M}}(\mathfrak{R}_n)|| \leq ||\mathfrak{R}_n - \Pi_{\mathscr{M}}(\mathfrak{R})||$, one obtains $||\mathfrak{R} - R|| \leq ||\mathfrak{R} - \Pi_{\mathscr{M}}(\mathfrak{R})||$. The unicity of the projection, and the fact that there is no $R \in \overline{\mathscr{M}}\backslash\mathscr{M}$ satisfying this inequality, shows that $R = \Pi_{\mathscr{M}}(\mathfrak{R})$. Since the bounded sequence $(\Pi_{\mathscr{M}}(\mathfrak{R}_n))$ has a unique limit point, one deduces the convergence $\Pi_{\mathscr{M}}(\mathfrak{R}_n) \to \Pi_{\mathscr{M}}(\mathfrak{R})$ and hence the continuity of the projection map at $\mathfrak{R}$.

Step 2: At any point $\mathfrak{R} \in \Omega$, any principal curvature $\kappa_i(N)$ in the direction N at $\Pi_{\mathscr{M}}(\mathfrak{R})$ must satisfy $\kappa_i(N) < 1$.

A consequence of the formula (2.10) is that the covariant Hessian of the distance function $J(R) = \frac{1}{2}||\mathfrak{R} - J||^2$ at $R = \Pi_{\mathscr{M}}(\mathfrak{R})$ is given by

$$\begin{array}{rccc}\mathcal{H}J & : & \mathcal{T}(R) & \to & \mathcal{T}(R) \\ & & X & \mapsto & X - L_R(N)(X),\end{array} \tag{2.13}$$

where N is the normal direction $N = \mathfrak{R} - \Pi_{\mathscr{M}}(\mathfrak{R})$. Since $R = \Pi_{\mathscr{M}}(\mathfrak{R})$ must be a local minimum of J, this Hessian must be positive, namely any eigenvalue $\kappa_i(N)$ of the Weingarten map $L_R(N)$ must satisfy $1 - \kappa_i(N) \geq 0$. Now, consider $s > 1$ such that $R + sN \in \Omega$ and notice that $||R + sN - \Pi_{\mathscr{M}}(\mathfrak{R})|| = s||N||$. Since

$$||R + sN - \Pi_{\mathscr{M}}(R + sN)|| \leq ||R + sN - \Pi_{\mathscr{M}}(\mathfrak{R})|| = s||N||,$$

the uniqueness of the projection in Ω implies that $\Pi_{\mathscr{M}}(R + sN) = R$ (*i.e.* the projection is invariant along orthogonal rays). The linearity of the Weingarten map in N implies $\kappa_i(sN) = s\kappa_i(N)$, hence $\kappa_i(N) \leq \frac{1}{s} < 1$, which concludes the proof.

Step 3: Application of the implicit function theorem.

Consider the function $f(\mathfrak{R}, R) = \Pi_{\mathcal{T}(R)}(R - \mathfrak{R})$ defined on $\mathscr{M} \times E$. The differential of f with respect to the variable R in a direction $X \in \mathcal{T}(R)$ at $R = \Pi_{\mathscr{M}}(\mathfrak{R})$ is the application

$$X \mapsto \Pi_{\mathcal{T}(R)}X - \mathrm{D}_X\Pi_{\mathcal{T}(R)}(\mathfrak{R} - R) = (I - L_R(N))(X).$$

Step 2 implies that the Jacobian $\partial_{R,X}f$ has no zero eigenvalue and hence is invertible.

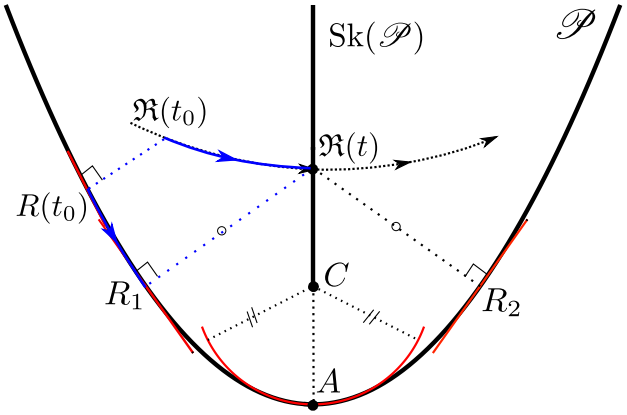


Figure 2-2: A parabola $\mathscr{M} = \mathscr{P}$ and its skeleton set $\mathrm{Sk}(\mathscr{P})$. The orthogonal projection $\Pi_{\mathscr{M}}$ is not differentiable on the adherence $\overline{\mathrm{Sk}(\mathscr{P})}$. Projected values $R(t) = \Pi_{\mathscr{M}}(\mathfrak{R}(t))$ jump from R_1 to R_2 when $\mathfrak{R}(t)$ crosses the skeleton. A center of curvature C, for which $\kappa_i(C - \Pi_{\mathscr{M}}(C)) = 1$, may admit a unique projection, A, but is a limit point of the skeleton $\mathrm{Sk}(\mathscr{P})$.

The implicit function theorem ensures the existence of a diffeomorphism $\boldsymbol{\phi}$ mapping an open neighborhood $\Omega_E \subset E$ of $\mathfrak{R}$ to an open neighborhood $\Omega_{\mathscr{M}} \subset \mathscr{M}$ of R, such that for any $\mathfrak{R}' \in \Omega_E$, $\boldsymbol{\phi}(\mathfrak{R}')$ is the unique element of $\Omega_{\mathscr{M}}$ satisfying $f(\mathfrak{R}', \boldsymbol{\phi}(\mathfrak{R}')) = 0$. By continuity of the projection obtained in step 1, one can assume, by replacing Ω_E with the open subset $\Omega_E \cap \Pi_{\mathscr{M}}^{-1}(\Omega_{\mathscr{M}})$, that $\Pi_{\mathscr{M}}(\Omega_E) \subset \Omega_{\mathscr{M}}$. Then, the equality $f(\mathfrak{R}', \Pi_{\mathscr{M}}(\mathfrak{R}')) = 0$ implies by uniqueness: $\boldsymbol{\phi}(\mathfrak{R}') = \Pi_{\mathscr{M}}(\mathfrak{R}')$. Hence $\Pi_{\mathscr{M}} = \boldsymbol{\phi}$ on Ω_E, and, in particular, $\Pi_{\mathscr{M}}$ is differentiable. Finally, for a given $X \in E$, one can now solve (2.5) by projection onto the eigenvectors of $L_R(N)$ and obtain (2.12). □

Remark 2.4. One cannot expect the projection map to be differentiable at points that are in the adherence $\overline{\mathrm{Sk}(\mathscr{M})}$, as there is a "jump" of the projected values across $\mathrm{Sk}(\mathscr{M})$, as illustrated on Figure 2-2.

A useful remark coming from the step 2 of the proof is the following:

Corollary 2.1. *A necessary condition for $R \in \mathscr{M}$ to be a local minimum of the distance functional $J(R) = \frac{1}{2}||\mathfrak{R} - R||^2$ is that $N = \mathfrak{R} - R \in \mathcal{N}(R)$ is a normal vector at R and that all the eigenvalues of the Weingarten map $L_R(N)$ satisfy $\kappa_i(N) \leq 1$. Therefore if the following condition holds*

$$\forall R \neq \Pi_{\mathscr{M}}(\mathfrak{R}) \in \mathscr{M},\ N = \mathfrak{R} - R \in \mathcal{N}(R) \Rightarrow \max_i \kappa_i(N) > 1, \tag{2.14}$$

then a solution $R(t)$ of the gradient flow

$$\dot{R} = -\nabla J(R) = \Pi_{\mathcal{T}(R)}(\mathfrak{R} - R). \tag{2.15}$$

converges almost surely to $\Pi_{\mathscr{M}}(\mathfrak{R})$. In particular, if condition (2.14) is satisfied for any $\mathfrak{R} \in \mathscr{M}$ and if $\partial\mathscr{M} = \emptyset$, then $\mathscr{M}$ is connected.

Proof. We refer the reader to chapter 7 of [79] (Morse theory) for the proofs that the solution of a sufficiently smooth gradient flow converges almost surely to a local minimum. □

Remark 2.5. The geometry of a connected manifold $\mathscr{M}$ is therefore relatively "simple" when the condition (2.14) holds. We will see that it is the case for each of the matrix manifolds studied in section 2.2. The gradient flow (2.15) is extremely powerful as it allows to compute either the projection $\Pi_{\mathscr{M}}(\mathfrak{R})$ or a path between two given points R_1, R_2 on $\mathscr{M}$ (by considering $\mathfrak{R} = R_2$ and $R(0) = R_1$), which can be useful when no analytic expression is available for the geodesic connecting R_1 to R_2. As an application, we obtain in section 2.2 dynamical systems (2.15) that achieve algebraic operations, in the continuity of [20, 31].

The following two results establish bounds for the differential of the projection operators $\Pi_{\mathscr{M}}$ and $\mathrm{D}\Pi_{\mathcal{T}(R)}$ that will be useful later for the error analysis of section 2.3.

Lemma 2.1. *For any $\mathfrak{R} \in E$ for which $\Pi_{\mathscr{M}}(\mathfrak{R})$ is defined and $\mathfrak{X} \in E$, denoting $N = \mathfrak{R} - \Pi_{\mathscr{M}}(\mathfrak{R}) \in \mathcal{N}(R)$:*

$$||\mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R}) - \Pi_{\mathcal{T}(\Pi_{\mathscr{M}}(\mathfrak{R}))}\mathfrak{X}|| \leq \max_i \frac{\kappa_i(N)}{1-\kappa_i(N)}||\mathfrak{X}||.$$

Proof. This is immediate from the decomposition (2.12). □

Lemma 2.2. *For any points $R \in \mathscr{M}$, $\mathfrak{X} \in E$ and tangent vector $X \in \mathcal{T}(R)$,*

$$||\mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot\mathfrak{X}|| \leq \left(\max_{\substack{N\in\mathcal{N}(R)\\||N||=1}} \kappa_i(N)\right) ||X||\,||\mathfrak{X}||.$$

The constant $\kappa_\infty(R) = \max\limits_{\substack{N\in\mathcal{N}(R)\\||N||=1}} \kappa_i(N) < \infty$ is optimal and is called the maximal curvature of $\mathscr{M}$ at $\mathfrak{R}$.

Proof. $N \mapsto L_R(N)$ is a continuous map from $\mathcal{N}(R)$ to the space of endomorphisms over $\mathcal{T}(R)$ (as a linear functional over finite dimensional spaces). As a consequence the eigenvalue maps $N \mapsto \kappa_i(N)$ are continuous functions, implying $\kappa_\infty(R) < \infty$. If $N \in \mathcal{N}(R)$ is a normal vector, the maximum eigenvalue of the map $L_R(N) : X \mapsto \mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot N$ is $\max \kappa_i(N)$. As a result,

$$\forall X \in \mathcal{T}(R),\ ||\mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot N|| \leq \max_{N\in\mathcal{N}(R)} \kappa_i(N)||X|| \leq \kappa_\infty(R)||X||\,||N||. \tag{2.16}$$

If Y is a tangent vector, the vector $\Gamma(X,Y) = \mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot Y$ is normal (proposition 2.1) and the Weingarten identity (2.6) yields

$$\begin{aligned}
||\mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot Y||^2 &= < \mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot Y, \Gamma(X,Y) > \\
&= - < \mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot \Gamma(X,Y), Y > \\
&\leq ||\mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot\Gamma(X,Y)||\,||Y|| \\
&\leq \kappa_\infty(R)||X||\,||\mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot Y||\,||Y||\,.
\end{aligned}$$

Therefore, one finds that (2.16) holds also for tangent vectors Y:

$$||\mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot Y|| \leq \kappa_\infty(R)||X||\,||Y||.$$

Hence, for any $\mathfrak{X} = N + Y$ where N is normal and Y tangent,

$$||\mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot\mathfrak{X}||^2 = ||\mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot N||^2 + ||\mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot Y||^2$$
$$\leq \kappa_\infty(R)^2||X||^2\,(||N||^2 + ||Y||^2) = \kappa_\infty(R)^2||X||^2\,||\mathfrak{X}||^2,$$

which proves the inequality claimed. □

We mention a simple observation that results from this lemma and that allows to understand what are the conditions for a geodesic to not be defined for all times on an embedded manifold:

Corollary 2.2. *A geodesic $R(t)$ on a finite dimensional manifold blows up in finite time if and only if it reaches the boundary $\partial\mathscr{M} = \overline{\mathscr{M}}\backslash\mathscr{M}$ in finite time. There are two kinds of scenario:*

- *the maximal curvature of $\mathscr{M}$ remains bounded and $\mathscr{M}$ could be extended on a neighborhood of the exiting point on $\partial\mathscr{M}$.*
- *the maximal curvature $\kappa_\infty(R(t))$ blows up in finite time, that is $R(t)$ reaches a singularity around which $\mathscr{M}$ has a spiraling shape. $\mathscr{M}$ cannot be further extended at the exiting point.*

In particular, on an embedded closed finite dimensional manifold, geodesic are always defined for all times.

Proof. We recall that the norm of the velocity $||\dot{R}||$ is constant on a geodesic $R(t)$. Assume that there is a finite time blowing up of $R(t)$ at time T. The boundedness of $\dot{R}$ on the interval $[0, T[$ implies that there exists a limit point $R(T) \in \overline{\mathscr{M}}$ such that $R(t) \to R$ when $t \to T$. Therefore the blowing up occurs at time T if and only if $||\ddot{R}|| \to \infty$ when $t \to T$. Applying now the bound of lemma 2.2 to the geodesic equation (2.2) yields $||\ddot{R}|| \leq \kappa_\infty(R)||\dot{R}||^2$. Assume $R(T) \in \mathscr{M}$. Then the blowing up implies $\kappa_\infty(R(T)) = \infty$ which is a contradiction. Therefore $R(T) \in \partial\mathscr{M}$. □

2.1.3 Oblique projections and generalization to embedded manifolds in non-euclidean spaces

As the applications of section 2.2 will motivate, we investigate in this part a generalization of the previous results to a wider class of projection maps $\Pi_{\mathscr{M}}(\mathfrak{R})$ that are not defined by a minimization principle as in definition 2.6 but share similar properties. The generalization consists in giving up the assumption that the ambient space is euclidean. Instead, we assume to be given a *bundle* of "normal" subspaces $\mathcal{N}(R)$ satisfying $E = \mathcal{T}(R) \oplus \mathcal{N}(R)$ for each point of the manifold, where these normal subspaces are not necessary orthogonal anymore. In this setting, one can consider the *oblique* projection $\Pi_{\mathscr{M}}$ that maps a point $\mathfrak{R} \in E$ to a point $R \in \mathscr{M}$ such that $\mathfrak{R} - R \in \mathcal{N}(R)$. If E is euclidean, then the orthogonal projection $\Pi_{\mathscr{M}}$ is a special case of oblique projection where the normal spaces $\mathcal{N}(R)$ are orthogonal to the tangent spaces $\mathcal{T}(R)$. The applications mapping a matrix to its truncated SVD, polar part or projector over the eigenspaces of a symmetric matrix are examples of orthogonal projections $\Pi_{\mathscr{M}}$. Examples of oblique projections that are not orthogonal include the applications that map a real matrix to the orthogonal linear projector over its stable dominant subspaces (associated with the complex eigenvalues of maximal real parts), or

to the non orthogonal linear projector whose image is the dominant subspace and whose kernel is the complement stable subspace associated with the remaining eigenvalues. We find generalizations of formula (2.12) to obtain the differential of such maps, and of equation (2.15) to derive dynamical systems for which $\Pi_{\mathscr{M}}(\mathfrak{R})$ is a stable equilibrium point.

In all this part, we consider $\mathscr{M} \subset E$ a smooth manifold embedded in a finite-dimensional vector space E, but E is not assumed to be euclidean anymore: one does not have a natural scalar product inducing a metric on $\mathscr{M}$.

Definition 2.9. We say that an application $\Pi_{\mathscr{M}}$ is an *oblique* projection onto $\mathscr{M}$ if the following conditions are satisfied:

1. There exists an open neighborhood $\mathcal{V} \subset E$ containing $\mathscr{M}$ and such that $\Pi_{\mathscr{M}} : \mathcal{V} \to \mathscr{M}$ is an application from $\mathcal{V}$ onto $\mathscr{M}$.

2. $\forall R \in \mathscr{M}$, $\Pi_{\mathscr{M}}(R) = R$.

3. $\forall R \in \mathscr{M}$, there exists a vector space $\mathcal{N}(R)$ such that

$$\mathcal{U}(R) = \{N \in E \,|\, R + N \in \mathcal{V},\, \Pi_{\mathscr{M}}(R+N) = R\}.$$

   is an open neighborhood of $\mathcal{N}(R)$ containing 0. We call $\mathcal{N}(R)$ the *normal space at* R.

The concept of oblique projection is illustrated on Figure 2-3. Geometrically, the third condition means that $\Pi_{\mathscr{M}}$ maps all points of the affine subspace $R+\mathcal{N}(R)$ sufficiently close to R onto the same point. More informally, the bundle of normal spaces $\mathcal{N}(R)$ can be understood as a set of "hairs" on the manifolds, and $\Pi_{\mathscr{M}}$ is a point mapping a point on a hair to its root on the manifold. When two oblique normal spaces intersect, there is possibly an ambiguity in the definition of $\Pi_{\mathscr{M}}(\mathfrak{R})$, which explains why one must restrict the domain of $\Pi_{\mathscr{M}}$ to a neighborhood $\mathcal{V}$.

We now observe that if an oblique projection is sufficiently smooth, then one obtains a bundle of normal spaces $\mathcal{N}(R)$ as well as a smooth map of linear projectors $R \mapsto \Pi_{\mathcal{T}(R)}$ whose image is $\mathcal{T}(R)$ and kernel $\mathcal{N}(R)$.

Proposition 2.4. *If $\Pi_{\mathscr{M}}$ is a differentiable oblique projection, then for any $R \in \mathscr{M}$,*

$$\Pi_{\mathcal{T}(R)} : \mathfrak{X} \mapsto \mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(R)$$

is the linear projector whose image is $\mathcal{T}(R)$ and whose kernel is $\mathcal{N}(R)$. In particular the direct sum decomposition $E = \mathcal{T}(R) \oplus \mathcal{N}(R)$ holds and $\Pi_{\mathscr{M}}$ satisfies

$$\forall \mathfrak{R} \in \mathcal{V},\, \Pi_{\mathcal{T}(\Pi_{\mathscr{M}}(\mathfrak{R}))}(\mathfrak{R} - \Pi_{\mathscr{M}}(\mathfrak{R})) = 0.$$

Proof. 1. $\Pi_{\mathcal{T}(R)}$ is a projector. One obtains by differentiating $\Pi_{\mathscr{M}}(\Pi_{\mathscr{M}}(\mathfrak{R})) = \Pi_{\mathscr{M}}(\mathfrak{R})$ with respect to $\mathfrak{R}$ in the direction $\mathfrak{X}$ the relation

$$\forall \mathfrak{X} \in E,\, \mathrm{D}_{\mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R})}\Pi_{\mathscr{M}}(\Pi_{\mathscr{M}}(\mathfrak{R})) = \mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R}).$$

Setting $\mathfrak{R} = R \in \mathscr{M}$ yields $\mathrm{D}_{\mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(R)}\Pi_{\mathscr{M}}(R) = \mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(R)$, and the result follows from the identity $\Pi_{\mathcal{T}(R)} \circ \Pi_{\mathcal{T}(R)} = \Pi_{\mathcal{T}(R)}$.

2. $\mathrm{Span}(\Pi_{\mathcal{T}(R)}) = \mathcal{T}(R)$. Differentiating $\Pi_{\mathscr{M}}(R) = R$ with respect to R in a direction $X \in \mathcal{T}(R)$, yields $\Pi_{\mathcal{T}(R)}X = X$ hence $\mathcal{T}(R) \subset \mathrm{Span}(\Pi_{\mathcal{T}(R)})$. Because $\Pi_{\mathscr{M}}$ is a map onto

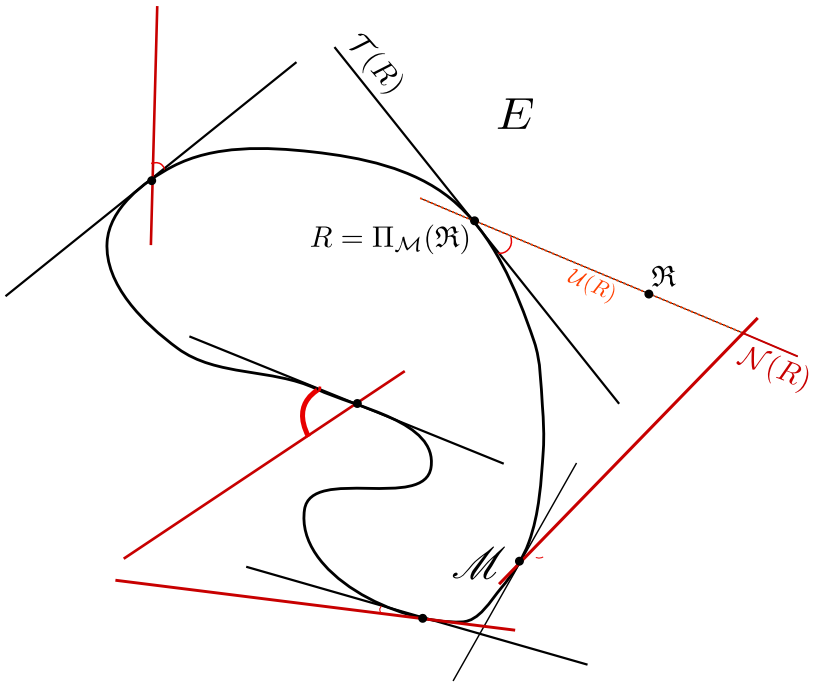


Figure 2-3: Oblique projection and bundle of tangent spaces $\mathcal{T}(R)$ and oblique normal spaces $\mathcal{N}(R)$ onto a manifold $\mathscr{M}$.

the manifold $\mathscr{M}$, the rank of $\mathfrak{X} \mapsto \mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R})$ must be lower the dimension of $\mathscr{M}$ and this implies $\mathrm{Span}(\Pi_{\mathcal{T}(R)}) = \mathcal{T}(R)$.

3. $\mathrm{Ker}(\Pi_{\mathcal{T}(R)}) = \mathcal{N}(R)$. For a given $N \in \mathcal{N}(R)$, differentiating at $t = 0$ the relation $\Pi_{\mathscr{M}}(R + tN) = R$ yields $\mathrm{D}_N\Pi_{\mathscr{M}}(R) = 0$ hence $\mathcal{N}(R) \subset \mathrm{Ker}(\Pi_{\mathcal{T}(R)})$. Reciprocally, note that the differential of the map $N \mapsto \Pi_{\mathscr{M}}(R + N) - R$ has constant rank $r = \dim(\mathscr{M})$. Therefore there exists a neighborhood $\mathcal{W} \subset E$ such that the set $\{N \in \mathcal{W} \,|\, R + N \in \mathcal{V},\, \Pi_{\mathscr{M}}(R+N) = R\}$ is a smooth manifold of dimension $\dim E - \dim \mathscr{M}$. But by definition of an oblique projection, this set is included in $\mathcal{U}(R)$, an open neighborhood of $\mathcal{N}(R)$, therefore $\dim \mathcal{N}(R) \geq \dim E - \dim \mathcal{T}(R) = \dim \mathrm{Ker}(\Pi_{\mathcal{T}(R)})$. Hence $\mathcal{N}(R) = \mathrm{Ker}(\mathcal{T}(R))$. □

We aim now at showing that *conversely*, if a manifold $\mathscr{M}$ is equipped with a differentiable map of linear projectors $R \mapsto \Pi_{\mathcal{T}(R)}$, then one can define a differentiable projection map $\Pi_{\mathscr{M}}(\mathfrak{R})$ associated with the normal spaces $\mathcal{N}(R) = \mathrm{Ker}(\Pi_{\mathcal{T}(R)})$. We first notice that most of the properties obtained in section 2.1 remain valid for projections $\Pi_{\mathcal{T}(R)}$ not necessary orthogonal.

Proposition 2.5. *Let $\mathscr{M} \subset E$ be an embedded smooth manifold equipped with a differentiable map $R \mapsto \Pi_{\mathcal{T}(R)}$ of linear projections over the tangent spaces at $\mathscr{M}$. Consider X and Y two differentiable tangent vector fields in a neighborhood of $R \in \mathscr{M}$. Then $\Pi_{\mathcal{T}(R)}$ defines a torsion-free connection on $\mathscr{M}$ by the formula*

$$\forall X, Y \in \mathcal{T}(R),\ \nabla_X Y = \Pi_{\mathcal{T}(R)}(\mathrm{D}_X Y). \tag{2.17}$$

In other words, one has Gauss formula

$$\forall X, Y \in \mathcal{T}(R),\ \nabla_X Y = \mathrm{D}_X Y + \Gamma(X, Y),$$

where the Christoffel symbol $\Gamma(X,Y)=\Gamma(Y,X)=-(I-\Pi_{\mathcal{T}(R)})(\mathrm{D}_X Y)\in\mathcal{N}(R)$ *is symmetric and has values in the normal space* $\mathcal{N}(R)$*. Furthermore, its value depends only on the values of the vector fields* X *and* Y *at* R *as it is visible from*

$$\Gamma(X,Y)=-\mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot Y.$$

Additionally, the Weingarten map

$$\begin{array}{rccc} L_R(N) & : & \mathcal{T}(R) & \rightarrow & \mathcal{T}(R) \\ & & X & \mapsto & \mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot N=-\Pi_{\mathcal{T}(R)}(\mathrm{D}_X N), \end{array} \tag{2.18}$$

defines a linear application of the tangent space $\mathcal{T}(R)$ *into itself.*

Proof. The proof is strictly identical to those given in proposition 2.1 and definition 2.7: it suffices to differentiate the relations $\Pi_{\mathcal{T}(R)}(Y)=Y$ and $\Pi_{\mathcal{T}(R)}(N)=0$ with respect to Y for given tangent vector fields X and N, and to use the fact that the Lie Bracket is a tangent vector. □

It is not clear that one can find a riemannian metric associated with the torsion-free connection ∇ defined from $\Pi_{\mathcal{T}(R)}$[1]. One can also wonder whether a *Weingarten* identity analogous to (2.6) still holds. The answer is positive provided the scalar product is replaced with the duality bracket. In the following, we denote by E^* the dual space of a finite dimensional vector space E and $<,>$ the duality bracket, i.e. $<v,x>=v(x)$ for any linear form $v\in E^*$ and vector $x\in E$. Recall that if A is a linear endomorphism of E, one can define the transpose of A to be the linear endomorphism A^* of E^* defined by the relation $<A^*v,x>=<v,Ax>$ for any $x\in E$ and $v\in E^*$.

Proposition 2.6. *For any* $R\in\mathscr{M}$*, the direct sum* $E^*=\mathcal{T}(R)^*\oplus\mathcal{N}(R)^*$ *holds where* $\mathcal{T}(R)^*=\Pi^*_{\mathcal{T}(R)}E^*$ *and* $\mathcal{N}(R)^*=(I-\Pi^*_{\mathcal{T}(R)})E^*$*. In particular,* $\Pi^*_{\mathcal{T}(R)}$ *is the linear projector whose image is* $\mathcal{T}(R)^*$ *and whose kernel is* $\mathcal{N}(R)^*$*. The map of projections* $R\mapsto\Pi^*_{\mathcal{T}(R)}$ *induces a connection over the dual bundle* $\mathcal{T}(R)^*$ *by the formula:*

$$\forall V\in\mathcal{T}(R)^*,\forall X\in\mathcal{T}(R),\ \nabla_X V=\Pi^*_{\mathcal{T}(R)}(\mathrm{D}_X V). \tag{2.19}$$

The connection ∇ *defined by* (2.17) *and* (2.19) *is compatible with the duality bracket :*

$$\forall X,Y\in\mathcal{T}(R),V\in\mathcal{T}(R)^*,\ \mathrm{D}_X<V,Y>=<\nabla_X V,Y>+<V,\nabla_X Y>.$$

One has Gauss formula

$$\forall V\in\mathcal{T}(R)^*,X\in\mathcal{T}(R),\ \nabla_X V=\mathrm{D}_X V+\Gamma(X,V),$$

where the Christoffel symbol $\Gamma(X,V)=-(I-\Pi^*_{\mathcal{T}(R)})(\mathrm{D}_X V)\in\mathcal{N}(R)^*$ *has values in the normal dual space* $\mathcal{N}(R)*$*. Furthermore,* $\Gamma(X,V)$ *depends only on the value of the tangent vector and dual fields* X *and* V *at* R *as it is visible from the formula*

$$\Gamma(X,V)=-\mathrm{D}\Pi^*_{\mathcal{T}(R)}(X)\cdot V.$$

[1]The question of under which condition a torsion-free connection is the Levi-Civita connection of a Riemannian metric has been investigated in [127].

Finally, for any $N \in \mathcal{N}(R)^*$, the dual Weingarten map

$$\begin{array}{rccc} L_R^*(N) & : & \mathcal{T}(R) & \to & \mathcal{T}(R)^* \\ & & X & \mapsto & \mathrm{D}\Pi^*_{\mathcal{T}(R)}(X)\cdot N = -\Pi^*_{\mathcal{T}(R)}(\mathrm{D}_X N), \end{array}$$

defines a linear application of the tangent space $\mathcal{T}(R)$ into its dual $\mathcal{T}(R)^*$ and the following Weingarten identities *holds:*

$$\forall X, Y \in \mathcal{T}(R), N \in \mathcal{N}(R)^*, < N, \mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot Y >=< \mathrm{D}\Pi^*_{\mathcal{T}(R)}(X)\cdot N, Y >,$$

$$\forall V \in \mathcal{T}(R)^*, X \in \mathcal{T}(R), N \in \mathcal{N}(R), < \mathrm{D}\Pi^*_{\mathcal{T}(R)}(X)\cdot V, N >=< V, \mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot N > .$$

Proof. The proof is almost identical to the ones of propositions 2.1 and 2.5 and definition 2.7 and is left to the reader. □

We now show that given a manifold $\mathscr{M}$ equipped with a map of linear projectors $R \mapsto \Pi_{\mathcal{T}(R)}$, one can find an open neighborhood $\mathcal{V}$ of $\mathscr{M}$ and a unique differentiable projection map $\Pi_{\mathscr{M}}$ satisfying the conditions of definition 2.9.

Proposition 2.7. *The set $\mathcal{Q} = \{(R, N) \in \mathscr{M} \times E | N \in \mathcal{N}(R)\}$ is a submanifold of $\mathscr{M} \times E$ of dimension* $\dim(E)$. *If $\mathscr{M}$ is compact, there exists an open neighborhood $V \subset \mathcal{Q}$ containing $\mathscr{M} \times \{0\}$ such that the map*

$$\begin{array}{rccc} \Phi & : & \mathcal{Q} & \to & E \\ & & (R, N) & \mapsto & R + N \end{array}$$

is a diffeomorphism from V onto its image $\mathcal{V} = \Phi(V)$. The reciprocal map $\Pi_{\mathscr{M}} : \mathcal{V} \to \mathscr{M}$ defined by $\Pi_{\mathscr{M}}(R+N) = R$ satisfies the properties of definition 2.9 and is called the oblique projection onto $\mathscr{M}$ relative to the normal subspaces $\mathcal{N}(R)$.

Proof. The proof is identical to the one of the "tubular neighborhood theorem" (see Theorem IV.5.4 in [117] and Theorem II.2.4 in [18]). Consider the map

$$\begin{array}{rccc} \Psi & : & \mathcal{M} \times E & \to \mathcal{Q} & \\ & & (R, \mathfrak{X}) & \mapsto & (R, (I - \Pi_{\mathcal{T}(R)})\mathfrak{X}). \end{array}$$

whose differential at $(R, \mathfrak{X})$ is $(X, \mathfrak{A}) \mapsto (X, -\mathrm{D}\Pi_{\mathcal{T}(R)}(X)\cdot\mathfrak{X}) + (0, (I - \Pi_{\mathcal{T}(R)})\mathfrak{A})$ for $(X, \mathfrak{A}) \in \mathcal{T}(R) \times E$. This differential has constant rank that is equal to $\dim(\mathcal{T}(R)) + \dim(\mathcal{N}(R)) = \dim(E)$. Hence $\mathcal{Q} = \Psi(\mathcal{M} \times E)$ is a smooth manifold of dimension $\dim E$. The tangent space at $(R, N) \in \mathcal{Q}$ is the set $\{(X - L_R(N)X, A) | (X, A) \in \mathcal{T}(R) \times \mathcal{N}(R)\}$. Hence the differential of Φ at $(X - L_R(N)X, A) \mapsto (X - L_R(N)X) + A$ which is invertible for $||N||$ sufficiently small, and in particular on the subset $\mathscr{M} \times \{0\}$. Then the local inversion theorem ensures that for every $R_0 \in \mathscr{M}$, there exists a ball

$$B(R_0, \delta) = \{(R, N) \in \mathcal{Q} |\ ||R - R_0|| \leq \delta \text{ and } ||N|| \leq \delta\}$$

of radius δ such that Φ is a local diffeomorphism from $B(R, \delta)$ to $\Phi(B(R, \delta))$.

By compacity of $\mathscr{M}$, one can extract a finite family $(B(R_i, \delta_i))$ of these balls such that $\mathscr{M} \subset \bigcup_i \Phi(B(R_i, \delta_i))$. Denote δ the minimal radius of these. We show that there exists $0 < \epsilon < \delta$ such that Φ is injective on the set

$$V = \{(R, N) \in \mathcal{Q} |\ ||N|| \leq \epsilon\}.$$

If the contrary is false, there exists two sequences (R_1^n, N_1^n) and (R_2^n, N_2^n) such that $R_1^n + N_1^n = R_2^n + N_2^n$ with $R_1^n \neq R_2^n$, $||N_1^n|| \to 0$ and $||N_2^n|| \to 0$ when $n \to \infty$. By compacity, one can assume up to extract a subsequence that $R_1^n \to R$. But $||R_1^n - R_2^n|| \leq ||N_1^n|| + ||N_2^n|| \to 0$ implies $R_2^n \to R$. Hence for n large enough (R_1^n, N_1^n) and (R_2^n, N_2^n) belong to one of the balls $B(R_i, \delta_i)$, which is a contradiction. □

Before stating the result regarding the differentiability of $\Pi_{\mathscr{M}}$, we first recall some simple facts about eigenvectors of non symmetric endomorphisms. Consider A a linear endomorphism over a finite-dimensional complex vector space E that can be diagonalized. Denote $\lambda_i \in \mathbb{C}$ and $(u_i)_{1 \leq i \leq n} \in E$ a corresponding basis of eigenvectors. Consider now E^* the dual of E (the space of linear forms over $\mathbb{C}$) and $(v_i)_{1 \leq i \leq n} \in E^*$ the dual basis of $(u_i)_{1 \leq i \leq n}$, *i.e.* for all $x \in E$, $< v_j, x >= v_j(x)$ is the coordinate of x along the vector u_j in the basis $(u_i)_{1 \leq i \leq n}$. Then by definition of a dual basis, (u_i) and (v_i) are *bi-orthogonal*, in the sense

$$\forall 1 \leq i, j \leq n, < v_i, u_j >= \delta_{ij},$$

and we can rewrite the action of A along the eigendecomposition as

$$Ax = \sum_{i=1}^{n} \lambda_i u_i < v_i, x > . \tag{2.20}$$

The dual family (v_i) forms a basis of eigenvectors for A^* with eigenvalues λ_i: $A^* v_i = \lambda_i v_i$. If E is a subspace of $\mathbb{C}^n$ equipped with the hermitian product $\forall x, y \in \mathbb{C}^n$, $< x, y >= \overline{x}^T y$, then dual eigenvectors can be identified to vectors $v_i \in E$ by the relation $\forall x \in E$, $< v_i, x >= \overline{v}_i^T x$. One can rewrite (2.20) as

$$A = \sum_{i=1}^{n} \lambda_i u_i \overline{v}_i^T. \tag{2.21}$$

The vectors v_i are eigenvectors of $\overline{A}^T$ are also the columns of $\overline{P}^{-T}$ where P is the invertible matrix such that $P^{-1}AP = \mathrm{diag}(\lambda_i)_{1 \leq i \leq n}$. Since $Au_i = \lambda_i u_i$ and $\overline{v}_i^T A = \lambda_i \overline{v}_i^T$, one refers to (see [75]) u_i and v_i as being respectively the *right* and *left* eigenvectors of A. In section 2.2.4, it will be convenient to use the representation (2.21) in the case of A being a complex or a real matrix, but we will keep in mind the representation (2.20) for more general linear maps. We denote $\mathrm{sp}(A)$ the set of complex eigenvalues of a linear operator A.

Theorem 2.2. *The neighborhood $\mathcal{V} \subset E$ in proposition 2.7 can be chosen such that for any $\mathfrak{R} = R + N \in \mathcal{V}$ with $R \in \mathscr{M}$, $N \in \mathcal{N}(R)$, 1 is not an eigenvalue of the Weingarten map $L_R(N)$: $1 \notin \mathrm{sp}(L_R(N))$. The oblique projection $\Pi_{\mathscr{M}}$ is differentiable on $\mathcal{V}$ and the differential is the map*

$$\mathfrak{X} \mapsto \mathrm{D}_{\mathfrak{X}} \Pi_{\mathscr{M}}(\mathfrak{R}) = (I - L_R N)^{-1} \Pi_{\mathcal{T}(R)}(\mathfrak{X}), \tag{2.22}$$

In particular, if $L_R(N)$ is diagonalizable in $\mathbb{C}$, and if we denote $\kappa_i(N)$ the (complex) eigenvalues of $L_R(N)$ associated with a basis of eigenvectors $(\Phi_i)_i$ and its dual basis (Φ_i^*), then

$$\forall \mathfrak{X} \in E, \ \mathrm{D}_{\mathfrak{X}} \Pi_{\mathscr{M}}(\mathfrak{R}) = \sum_i \frac{1}{1 - \kappa_i(N)} < \Phi_i^*, \Pi_{\mathcal{T}(R)} \mathfrak{X} > \Phi_i. \tag{2.23}$$

Proof. The proof is a consequence of the implicit function theorem applied to the function $(\mathfrak{R}, R) \mapsto f(\mathfrak{R}, R) = \Pi_{\mathcal{T}(R)}(\mathfrak{R} - R)$, whose partial differential with respect to R is $L_R(N) - I$. Equations (2.22) and (2.23) follow similarly as in theorem 2.1. □

Remark 2.6. It can be useful to notice that formula (2.22) holds globally in E for any $\mathfrak{R} = R + N$ with $N \in \mathcal{N}(R)$ and $1 \notin \mathrm{sp}(L_R(N))$, because the implicit function theorem ensures the existence of a local inverse map ϕ (here $\Pi_{\mathscr{M}}$) in a neighborhood of $\mathfrak{R}$ onto a neighborhood of R such that $\mathfrak{R} - \phi(\mathfrak{R}) \in \mathcal{N}(\phi(\mathfrak{R}))$ remains normal when varying $\mathfrak{R}$.

Formula (2.23) is the generalization of (2.12) to the case of oblique projections. We also obtain that $\Pi_{\mathscr{M}}(\mathfrak{R})$ can also be computed by using a dynamical system, that coincides with the gradient flow (2.15) when the projection $\Pi_{\mathscr{M}}$ is orthogonal.

Proposition 2.8. *If the neighborhood $\mathcal{V}$ of proposition 2.7 is chosen such that*

$$\mathcal{V} \subset \mathcal{W} = \{R + N \in E | R \in \mathscr{M}, N \in \mathcal{N}(R) \text{ and } sp(L_R(N)) \subset \{\lambda \in \mathbb{C} | \Re(\lambda) < 1\}\}.$$

then for $\mathfrak{R} \in \mathcal{V}$, $\Pi_{\mathscr{M}}(\mathfrak{R})$ is an asymptotically stable equilibrium point of

$$\dot{R} = \Pi_{\mathcal{T}(R)}(\mathfrak{R} - R). \tag{2.24}$$

Proof. Note that $\mathcal{W}$ defined as above remain an open neighborhood of $\mathscr{M}$ in E. The result is a trivial consequence of the fact that $\Pi_{\mathscr{M}}(\mathfrak{R})$ is an equilibrium of (2.24) and that the linear operator of the linearized dynamic is $L_R(N) - I$, with $N = \mathfrak{R} - \Pi_{\mathscr{M}}(\mathfrak{R})$ and $R = \Pi_{\mathscr{M}}(\mathfrak{R})$. By assumption, $\mathfrak{R} \in \mathcal{V}$, a region where the eigenvalues of $L_R(N) - I$ are all strictly negative. □

As observed previously the local stability may become global if $R = \Pi_{\mathscr{M}}(\mathfrak{R})$ is defined and is the only point such that the spectrum of $L_R(\mathfrak{R} - R)$ lies in $\{\lambda \in \mathbb{C} | \Re(\lambda) < 1\}$. The local stability of the dynamical system (2.24) is a particularly powerful result, as it yields systematically a continuous time algorithm to find the value of $\Pi_{\mathscr{M}}(\mathfrak{R} + \delta\mathfrak{R})$ given the knowledge of $\Pi_{\mathscr{M}}(\mathfrak{R})$ and a small perturbation $\delta\mathfrak{R}$.

We have therefore at our disposal a framework to find in a systematic way the differential of oblique projections and derive dynamical systems that compute them. The procedure can be summed up in the following steps. Given a "candidate" continuous projection $\Pi_{\mathscr{M}}$,

1. Identify the image manifold $\mathscr{M}$, on which $\Pi_{\mathscr{M}}$ is the identity, as well as the tangent spaces $\mathcal{T}(R)$ of $\mathscr{M}$.

2. Identify the normal space $\mathcal{N}(R)$ characterized by $\Pi_{\mathscr{M}}(R+tN) = R$ for all t sufficiently small.

3. Compute the linear projector $\Pi_{\mathcal{T}(R)}$ whose image is $\mathcal{T}(R)$ and kernel is $\mathcal{N}(R)$.

4. If the map $R \mapsto \Pi_{\mathcal{T}(R)}$ is differentiable, and if $\Pi_{\mathscr{M}}$ is continuous, then one can apply theorem 2.2 or its remark to obtain that $\Pi_{\mathscr{M}}(\mathfrak{R})$ is differentiable. Compute the Weingarten map $L_R(N)$ given a normal vector. If possible, diagonalize $L_R(N)$ and obtain the differential of $\Pi_{\mathscr{M}}$ with the formula (2.23).

5. Derive the dynamical system (2.24) to obtain a continuous algorithm for computing $\Pi_{\mathscr{M}}(\mathfrak{R})$.

2.2 Embedded geometry and curvature of matrix manifolds

In the following we apply the tools developed in the previous section to the exhaustive study of three embedded matrix manifolds: the fixed-rank manifold, the Stiefel manifold and the isospectral manifold. For each, we compute tangent space, normal space, geodesics, Weingarten map and principal curvatures. We relate the orthogonal projection onto the manifold to an algebraic operation, namely the truncated SVD, polar decomposition, and the operation of replacing the eigenvalues of a symmetric matrix. Applying theorem theorem 2.4, we obtain their differential, and we provide gradient flows that achieve these algebraic operations. Such may be useful in practical computations when one is for example interested in updating an algebraic decomposition of a time dependent matrix as in [20, 89, 31]. We apply also the framework of *oblique-projections* to derive dynamical systems tracking the eigenspaces of non-symmetric matrices.

2.2.1 The fixed rank manifold and the differentiability of the SVD truncation

This section establishes the geometric framework of low-rank approximation, by unifying results sparsely available in [83, 124, 103], and by providing expressions for classical geometric characteristics such as geodesics and covariant derivative.

Definition 2.10. The set of l-by-m matrices of rank r is denoted by $\mathscr{M}$:

$$\mathscr{M} = \{R \in \mathcal{M}_{l,m} | \text{rank}(R) = r\}. \tag{2.25}$$

The following lemma [114] fixes the parametrization of $\mathscr{M}$ by conveniently representing its elements R in terms of mode and coefficient matrices, U and Z, respectively.

Lemma 2.3. *Any matrix $R \in \mathscr{M}$ can be decomposed as*

- $R = UZ^T$ *where* $U \in \mathcal{M}^*_{l,r}$ *and* $Z \in \mathcal{M}^*_{m,r}$ *(*$\text{rank}(U) = \text{rank}(Z) = r$*). This decomposition is unique modulo an invertible matrix* $A \in \text{GL}_r$*, namely if* $U_1, U_2 \in \mathcal{M}^*_{l,r}$*,* $Z_1, Z_2 \in \mathcal{M}^*_{m,r}$*,*

$$U_1 Z_1^T = U_2 Z_2^T \Leftrightarrow \exists A \in \text{GL}_r,\ U_2 = U_1 A \text{ and } Z_2 = Z_1 A^{-T}.$$

- $R = UZ^T$*, where* $U \in \text{St}_{l,r}$ *and* $Z \in \mathcal{M}^*_{m,r}$*,* i.e. $U^T U = I$ *and* $\text{rank}(Z) = r$*, respectively. This decomposition is unique modulo a rotation matrix* $P \in O_r$*, namely if* $U_1, U_2 \in \mathcal{M}_{l,r}$*,* $Z_1, Z_2 \in \mathcal{M}_{m,r}$*, and* $U_1^T U_1 = U_2^T U_2 = I$*, then*

$$U_1 Z_1^T = U_2 Z_2^T \Leftrightarrow \exists P \in \text{O}_r, U_1 = U_2 P \text{ and } Z_1 = Z_2 P.$$

Proof. The existence is immediate by writing the SVD $R = U\Sigma V^T$. For the uniqueness, we proceed as follows. Since U_1, Z_1 have full rank, one can define $A = Z_2^T Z_1 (Z_1^T Z_1)^{-1} \in \mathcal{M}_{r,r}$ and notice that

$$\begin{aligned}&(U_1^T U_1)^{-1} U_1^T U_2 A \\ &\quad = ((U_1^T U_1)^{-1} U_1^T)(U_2 Z_2^T)(Z_1 (Z_1^T Z_1)^{-1}) = ((U_1^T U_1)^{-1} U_1^T)(U_1 Z_1^T)(Z_1 (Z_1^T Z_1)^{-1}) = I.\end{aligned}$$

Therefore A is invertible and $A^{-1} = (U_1^T U_1)^{-1} U_1^T U_2$. This proves the first point since $U_1 = U_2 Z_2^T Z_1 (Z_1^T Z_1)^{-1} = U_2 A$ and $Z_1^T = (U_1^T U_1)^{-1} U_1^T U_2 Z_2^T = A^{-1} Z_2^T$. When in addition $U_1^T U_1 = U_2^T U_2 = I$, one can write

$$A = (U_2^T U_2) Z_2^T Z_1 (Z_1^T Z_1)^{-1} = U_2^T (U_1 Z_1^T) Z_1 (Z_1^T Z_1)^{-1} = U_2^T U_1,$$

therefore $A^{-1} = U_1^T U_2 = A^T$, proving $A \in \mathcal{O}_r$. □

We will use in the following exclusively and extensively the second factorization, and the statement "let $UZ^T \in \mathscr{M}$" will always implicitly assumes $U \in \mathcal{M}_{l,r}$, $Z \in \mathcal{M}_{m,r}$, $U^T U = I$, and $\mathrm{rank}(Z) = r$. Other parameterizations of $\mathscr{M}$ are possible and give equivalent results [101].

Proposition 2.9. *$\mathscr{M}$ is a smooth manifold of dimension $(l+m)r - r^2$.*

Proof. (see also [139]). Consider the map

$$\begin{array}{rccc} \phi & : & \mathrm{St}_{l,r} \times \mathcal{M}_{m,r} & \longrightarrow & \mathcal{M}_{l,m} \\ & & (U, Z) & \longmapsto & UZ^T. \end{array}$$

This map is a smooth submersion. Furthermore, the following proposition will show that its differential $\mathrm{D}\phi$ has constant rank $(l+m)r - r^2$. The property is then an immediate consequence of the constant rank theorem. □

Remark 2.7. The same argument will be used for the two other manifolds studied: one considers a smooth parameterization of $\mathscr{M}$, then one derives a candidate "tangent space". If that candidate tangent space has a constant dimension, then $\mathscr{M}$ is a smooth manifold by application of the constant rank theorem.

The tangent space $\mathcal{T}(UZ^T)$ at a point $R = UZ^T$ is the set of all possible vectors tangent to smooth curves $R(t) = U(t)Z(t)^T$ drawn on the manifold $\mathscr{M}$. Therefore, any such tangent vector at $R(0) = UZ^T$ is of the form $\dot{R} = \dot{U}Z^T + U\dot{Z}^T$, where $\dot{U}$ and $\dot{Z}$ are the time derivatives of the matrices $U(t)$ and $Z(t)$ at time $t = 0$. In the following, the notations X_U, X_Z, and $X = X_U Z^T + U X_Z^T$ will be used to denote the tangent directions $\dot{U}$, $\dot{Z}$, and $\dot{R}$ for the respective matrices U, Z and R. The orthogonality condition that $U^T U = I$ must hold for all times implies that X_U must satisfy $\dot{U}^T U + U^T \dot{U} = X_U^T U + U^T X_U = 0$. Nevertheless, this is not sufficient to parameterize uniquely tangent vectors X from the displacements X_U and X_Z for U and Z, since $X = X_U Z^T + U X_Z^T$ is invariant under the transformation $X_U \leftarrow X_U \Omega$ and $X_Z \leftarrow X_Z \Omega$ for any skew-symmetric matrix $\Omega = -\Omega^T$. Intuitively, this is related to the fact that rotations of the columns of the mode matrix, U, do not change the subspace $\mathrm{span}(\boldsymbol{u}_i)$ supporting the modal decomposition (5) (see [124]). A unique parameterization of the tangent space can be obtained by adding the condition that this subspace must evolve orthogonally to itself, i.e. by requiring $U^T X_U = 0$.

Proposition 2.10. *The tangent space of $\mathscr{M}$ at $R = UZ^T \in \mathscr{M}$ is the set*

$$\mathcal{T}(UZ^T) = \{X_U Z^T + U X_Z^T \mid X_U \in \mathcal{M}_{l,r},\, X_Z \in \mathcal{M}_{m,r},\, U^T X_U + X_U^T U = 0\}. \tag{2.26}$$

$\mathcal{T}(UZ^T)$ is uniquely parameterized by the horizontal space

$$\mathcal{H}_{(U,Z)} = \{(X_U, X_Z) \in \mathcal{M}_{l,r} \times \mathcal{M}_{m,r} \mid U^T X_U = 0\}, \tag{2.27}$$

that is for any tangent vector $X \in \mathcal{T}(UZ^T)$, there exists a unique $(X_U, X_Z) \in \mathcal{H}_{(U,Z)}$ such that $X = X_U Z^T + U X_Z^T$.

Proof. (see also [83]) One can always write a tangent vector $X = UX_Z^T + X_U Z^T \in \mathcal{T}(UZ^T)$ as $X = U(X_Z^T + U^T X_U Z^T) + ((I - UU^T)X_U)Z^T$, implying that $\mathcal{T}(UZ^T) = \{X_U Z^T + UX_Z^T | (X_U, X_Z) \in \mathcal{H}_{(U,Z)}\}$. Furthermore, if $X = UX_Z^T + X_U Z^T$ with $U^T X_U = 0$, then $X_Z = X^T U$ and $X_U = (I - UU^T)XZ(Z^T Z)^{-1}$ showing that $(X_U, X_Z) \in \mathcal{H}_{(U,Z)}$ is defined uniquely from X. □

Remark 2.8. The denomination "*horizontal space*" for the set $\mathcal{H}_{(U,Z)}$ (2.27) refers to the definition of a non-ambiguous representation of the tangent space $\mathcal{T}(UZ^T)$ (2.26). This notion is developed rigorously in the theory of quotient manifolds e.g. [101, 36].

In the following, the notation $X = (X_U, X_Z)$ is used equivalently to denote a tangent vector $X = X_U Z^T + U X_Z^T \in \mathcal{T}(UZ^T)$, where $U^T X_U = 0$ is implicitly assumed.

Definition 2.11. At each point $UZ^T \in \mathscr{M}$, the metric g on $\mathscr{M}$ is the scalar product acting on the tangent space $\mathcal{T}(UZ^T)$ that is inherited from the scalar product of $\mathcal{M}_{l,m}$:

$$\begin{aligned} g((X_U, X_Z), (Y_U, Y_Z)) &= \mathrm{Tr}((X_U Z^T + U X_Z^T)^T (Y_U Z^T + U Y_Z^T)) \\ &= \mathrm{Tr}(Z^T Z X_U^T Y_U + X_Z^T Y_Z). \end{aligned} \tag{2.28}$$

Remark 2.9. In [101] and other works of matrix optimization e.g. [6, 148, 130], one uses the metric induced by the parametrization of the manifold $\mathscr{M}$: the norm of a tangent vector $(X_U, X_Z) \in \mathcal{H}_{(U,Z)}$ is defined to be $||(X_U, X_Z)||^2 = ||X_U||^2_{\mathrm{St}_{l,r}} + ||X_Z||^2_{\mathcal{M}_{m,r}}$ where $||\ ||_{\mathrm{St}_{l,r}}$ is a canonical norm on the Stiefel Manifold (see [36]) and $||\ ||_{\mathcal{M}_{m,r}}$ is the Frobenius norm on $\mathcal{M}_{m,r}$. In this work, one is rather interested in the metric inherited from the ambient full space $\mathcal{M}_{l,m}$, since it is the metric used to estimate the distance from a matrix $\mathfrak{R} \in \mathcal{M}_{l,m}$ to its best r-rank approximation, namely the error committed by the truncated SVD.

Proposition 2.11. *At every point $UZ^T \in \mathscr{M}$, the orthogonal projection $\Pi_T(UZ^T)$ onto the tangent space $\mathcal{T}(UZ^T)$ is the application*

$$\begin{array}{rccc} \Pi_{\mathcal{T}(UZ^T)} & : & \mathcal{M}_{l,m} & \to & \mathcal{H}_{(U,Z)} \\ & & \mathfrak{X} & \mapsto & ((I - UU^T)\mathfrak{X}Z(Z^T Z)^{-1}, \mathfrak{X}^T U). \end{array} \tag{2.29}$$

Proof. (see also [83]) $\Pi_{\mathcal{T}(R)}\mathfrak{X}$ is obtained as the unique minimizer of the convex functional $J(X_U, X_Z) = \frac{1}{2}||\mathfrak{X} - X_U Z^T - UX_Z^T||^2$ on the space $\mathcal{H}_{(U,Z)}$. The minimizer (X_U, X_Z) is characterized by the vanishing of the gradient of J:

$$\forall \Delta \in \mathcal{M}_{l,r},\ \Delta^T U = 0 \Rightarrow \frac{\partial J}{\partial X_U} \cdot \Delta = - < \mathfrak{X} - X_U Z^T - UX_Z^T, \Delta Z^T >= 0,$$

$$\forall \Delta \in \mathcal{M}_{m,r},\ \frac{\partial J}{\partial X_Z} \cdot \Delta = - < \mathfrak{X} - X_U Z^T - UX_Z^T, U\Delta^T >= 0,$$

yielding respectively $X_U = (I - UU^T)\mathfrak{X}Z(Z^T Z)^{-1}$ and $X_Z = \mathfrak{X}^T U$. □

The orthogonal complement of the tangent space $\mathcal{T}(R)$ is obtained from the identity $(I - \Pi_{\mathcal{T}(UZ^T)}) \cdot \mathfrak{X} = (I - UU^T)\mathfrak{X}(I - Z(Z^T Z)^{-1}Z^T)$:

Proposition 2.12. *The normal space at $R = UZ^T$ is :*

$$\begin{aligned}\mathcal{N}(R) &= \{N \in \mathcal{M}_{l,m} | (I - UU^T)N(I - Z(Z^TZ)^{-1}Z^T) = N\} \\ &= \{N \in \mathcal{M}_{l,m} \mid U^TN = 0 \text{ and } NZ = 0\}.\end{aligned} \tag{2.30}$$

In model order reduction, a matrix $R = UZ^T \in \mathscr{M}$ is usually a low rank-r approximation of a full rank matrix $\mathfrak{R} \in \mathcal{M}_{l,m}$. The following proposition shows that the normal space at R, $\mathcal{N}(R)$, can be understood as the set of all possible completions of the approximation (5) :

Proposition 2.13. *Let N be a given normal vector $N \in \mathcal{N}(R)$ at $R = UZ^T \in \mathscr{M}$. Then there exists an orthonormal basis of vectors $(u_i)_{1\leq i\leq l}$ in $\mathbb{R}^l$, an orthonormal basis $(v_i)_{1\leq i\leq m}$ of $\mathbb{R}^M$, and $r + k$ non zero singular values $(\sigma_i)_{1\leq i\leq r+k}$ such that*

$$UZ^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T \text{ and } N = \sum_{i=1}^{k} \sigma_{r+i} u_{r+i} v_{r+i}^T.$$

Proof. Consider $N = U_N \Theta V_N^T$ the SVD decomposition of N [74]. Since $U^TN = 0$, r columns of U_N are spanned by U and associated with zero singular values of N, therefore u_i is obtained from the columns of U for $1 \leq i \leq r$ and from the left singular vectors of N associated with non zero singular values for $r + 1 \leq i \leq r + k$, $k \geq 0$. The vectors v_i and v_{r+j} are obtained similarly. The singular values σ_i are obtained by reunion of the respective r and k non-zeros singular values of Z and N. □

Proposition 2.14. *Consider $X = (X_U, X_Z) \in \mathcal{H}_{(U,Z)}$ and $Y = (Y_U, Y_Z) \in \mathcal{H}_{(U,Z)}$ two tangent vector fields. The covariant derivative $\nabla_X Y$ on $\mathscr{M}$ is given by*

$$\nabla_X Y = (D_X Y_U + UX_U^T Y_U + (X_U Y_Z^T + Y_U X_Z^T)Z(Z^TZ)^{-1},\ D_X Y_Z - ZY_U^T X_U) \in \mathcal{H}_{(U,Z)}. \tag{2.31}$$

Therefore, geodesic equations on $\mathscr{M}$ are given by

$$\begin{cases} \ddot{U} + U\dot{U}^T\dot{U} + 2\dot{U}\dot{Z}^TZ(Z^TZ)^{-1} = & 0 \\ \ddot{Z} - Z\dot{U}^T\dot{U} = & 0. \end{cases} \tag{2.32}$$

Proof. Writing $X = X_U Z^T + UX_Z^T$ and $Y = Y_U Z^T + UY_Z^T$, one obtains:

$$\begin{aligned} D_X Y &= D_X Y_U Z^T + Y_U X_Z^T + X_U Y_Z^T + U D_X Y_Z^T \\ &= D_X Y_U Z^T + U D_X Y_Z^T + X_U Y_Z^T + Y_U X_Z^T. \end{aligned}$$

Applying the projection $\Pi_T(UZ^T)$ using eqn. (2.29), *i.e.*

$$\nabla_X Y = \Pi_{(U,Z)}(D_X Y) = ((I - UU^T)D_X Y Z(Z^TZ)^{-1}, D_X Y^T U),$$

yields in the coordinates of the horizontal space:

$$\nabla_X Y = ((I - UU^T)D_X(Y_U) + (X_U Y_Z^T + Y_U X_Z^T)Z(Z^TZ)^{-1}, D_X(Y_Z) + ZD_X(Y_U^T)U).$$

(2.31) is obtained by differentiating the constraint $U^TY_U = 0$ along the direction X, *i.e.* $X_U^TY_U + U^TD_XY_U = 0$, and replacing accordingly $U^TD_XY_U$ into the above expression. Since $\mathrm{D}_{(\dot{U},\dot{Z})}(\dot{U}) = \ddot{U}$ and $\mathrm{D}_{(\dot{U},\dot{Z})}(\dot{Z}) = \ddot{Z}$, $\nabla_{(\dot{U},\dot{Z})}(\dot{U}, \dot{Z}) = 0$ yields eqs. (2.32). □

It is well known [59, 75] that the truncated SVD, *i.e.* the map that set all singular values of a matrix $\mathfrak{R}$ to zero except the r highest, yields the best rank r approximation.

Definition 2.12. Let $\mathfrak{R} \in \mathcal{M}_{l,m}$ a matrix of rank at least r, i.e. $r+k, k \geq 0$, and denote $\mathfrak{R} = \sum_{i=1}^{r+k} \sigma_i(\mathfrak{R}) u_i v_i^T$ its singular value decomposition. If $\sigma_r(\mathfrak{R}) > \sigma_{r+1}(\mathfrak{R})$, then the rank r truncated SVD

$$\Pi_{\mathscr{M}}(\mathfrak{R}) = \sum_{i=1}^{r} \sigma_i(\mathfrak{R}) u_i v_i^T \in \mathscr{M}, \tag{2.33}$$

is the unique matrix $R \in \mathscr{M}$ minimizing the Euclidean distance $R \mapsto ||\mathfrak{R} - R||$.

Proof. Definition 2.6 and proposition 2.13 show that $N = \mathfrak{R} - R$ belongs to the normal space $\mathcal{N}(R)$ and hence must be of the form $N = \sum_{i \in S} \sigma_i u_i v_i^T$ where S is a subset of k integers between 1 and $r+k$. Since $||N||^2 = \sum_{i \in S} \sigma_i^2$, it is clear that $J(R)$ is minimized by setting $R = \sum_{i=1}^{r} \sigma_i(\mathfrak{R}) u_i v_i^T$. The condition $\sigma_r(\mathfrak{R}) > \sigma_{r+1}(\mathfrak{R})$ ensures the uniqueness of the minimizer R. □

Remark 2.10. The skeleton of $\mathscr{M}$ (Fig. 2-2) is therefore the set

$$\mathrm{Sk}(\mathscr{M}) = \{\mathfrak{R} \in \mathcal{M}_{l,m} | \sigma_r(\mathfrak{R}) = \sigma_{r+1}(\mathfrak{R})\}.$$

characterized by the crossing of the singular values of order r and $r+1$.

Proposition 2.15. *The Weingarten map $L_R(N)$ of the fixed rank manifold $\mathscr{M}$ in the normal direction $N \in \mathcal{N}(R)$ is the application:*

$$\begin{array}{rccc} L_R(N) \ : & \mathcal{H}_{(U,Z)} & \longrightarrow & \mathcal{H}_{(U,Z)} \\ & (X_U, X_Z) & \longmapsto & (N X_Z (Z^T Z)^{-1}, N^T X_U). \end{array} \tag{2.34}$$

The second fundamental form is given by:

$$\mathrm{II} \ : \ (X, Y) \mapsto < X, L_R(N)(Y) > = \mathrm{Tr}((X_U Y_Z^T + Y_U X_Z^T)^T N). \tag{2.35}$$

Proof. Differentiating (2.29) along the tangent direction $(X_U, X_Z) \in \mathcal{H}_{(U,Z)}$, and using the relations $U^T N = 0$ and $NZ = 0$, yields

$$L_R(N) = U X_U^T N + N X_Z (Z^T Z)^{-1} Z^T. \tag{2.36}$$

The normality of N implies that $(N X_Z (Z^T Z)^{-1}, N^T X_U)$ is a vector of the horizontal space and therefore equation (2.34) follows. One obtains (2.35) by evaluating the scalar product $< X, L_R(N)(Y) >$ with the metric g (equation (2.28)). □

Remark 2.11. The Christoffel symbol is deduced from equations (2.35) and (2.8):

$$\Gamma(X, Y) = -(I - \Pi_{\mathcal{T}(R)})(X_U Y_Z^T + Y_U X_Z^T). \tag{2.37}$$

Theorem 2.3. *Consider a point $R = U Z^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T \in \mathscr{M}$ and a normal vector $N = \sum_{j=1}^{k} \sigma_{r+j} u_{r+j} v_{r+j}^T \in \mathcal{N}(R)$ (no ordering of the singular values is assumed). At R and in the direction N, there are $2kr$ non-zero principal curvatures*

$$\kappa_{i,j}^{\pm}(N) = \pm \frac{\sigma_{r+j}}{\sigma_i},$$

for all possible combinations of non-zero singular values σ_{r+j}, σ_i *for* $1 \leq i \leq r$ *and* $1 \leq j \leq k$*. The normalized corresponding principal directions are the tangent vectors*

$$\Phi^{\pm}_{i,r+j} = \frac{1}{\sqrt{2}}(u_{r+j}v_i^T \pm u_i v_{r+j}^T). \tag{2.38}$$

The other principal curvatures are null and associated with the principal subspace

$$\text{span}\{(u_i v^T)_{1\leq i\leq r} | Nv = 0\} \oplus \text{span}\{(uv_i^T)_{1\leq i\leq r} | u^T N = u^T U = 0\}.$$

Proof. A principal curvature $\kappa^{\pm}_{i,j}(N)$ at $R = UZ^T$ associated with a principal direction $X = (X_U, X_Z)$ must satisfy the eigenvalue problem $L_R(N)(X) = \kappa^{\pm}_{i,j}(N)X$. $\Phi^{\pm}_{i,r+j}$ is indeed a tangent vector as one can write $\Phi^{\pm}_{i,r+j} = X_U Z^T \pm U X_Z^T$ with:

$$(X_U, X_Z) = \frac{1}{\sqrt{2}\sigma_{r+j}\sigma_i}(N v_{r+j} u_i^T U, N^T u_{r+j} v_i^T Z).$$

One finds that $\Phi^{\pm}_{i,r+j}$ is an eigenvector by using $\sqrt{2}X_U Z^T = u_{r+j}v_i^T$ and $\sqrt{2}UX_Z^T = u_i v_{r+j}^T$ and replacing them into (2.36). This also leads to $\kappa^{\pm}_{i,j}(N) = \pm\frac{\sigma_{r+j}}{\sigma_i}$. One then checks that $lr+mr-r^2$ eigenvalues (including multiplicities) have been obtained: there are $2kr$ non-zeros principal curvatures. The dimension of $\text{span}\{(u_i v^T)_{1\leq i\leq r} | Nv = 0\}$ is $(m-k)r$ and the one of $\text{span}\{(uv_i^T)_{1\leq i\leq r} | u^T N = u^T U = 0\}$ is $(l-k-r)r$. The total is $(m-k)r+(l-k-r)r+2kr = mr + lr - r^2$ as expected. □

This theorem shows that the maximal curvature of $\mathscr{M}$ (for normalized normal directions $||N|| = 1$) is $\kappa_\infty(R) = \sigma_r(R)^{-1}$ and hence diverges as the smallest singular value goes to 0. This fact (as well as proposition 2.16 and lemma 2.2) confirms what is visible on Figure 1: the manifold $\mathscr{M}$ can be seen as a collection of cones or as a multidimensional spiral, whose axes are the lower dimensional manifolds of matrices of rank less than $r-1$. Applying directly the formula (2.12) of theorem 2.1, one obtains an explicit expression for the differential of the truncated SVD:

Theorem 2.4. *Consider* $\mathfrak{R} \in \mathcal{M}_{l,m}$ *with rank greater than* r *and denote* $\mathfrak{R} = \sum_{i=1}^{r+k} \sigma_i u_i v_i^T$ *its SVD decomposition, where the singular values are ordered decreasingly:* $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{r+k}$*. Suppose that the orthogonal projection* $\Pi_{\mathscr{M}}(\mathfrak{R}) = UZ^T$ *of* $\mathfrak{R}$ *onto* $\mathscr{M}$ *is uniquely defined, that is* $\sigma_r > \sigma_{r+1}$*. Then* $\Pi_{\mathscr{M}}$*, the truncated SVD of order* r*, is differentiable at* $\mathfrak{R}$ *and the differential* $\mathrm{D}_{\mathfrak{X}}\Pi(\mathfrak{R})$ *in a direction* $\mathfrak{X} \in \mathcal{M}_{l,m}$ *is given by the formula*

$$\mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R}) = \Pi_{T\ (\Pi_{\mathscr{M}}(R))}(\mathfrak{X}) + \sum_{\substack{1\leq i\leq r\\ 1\leq j\leq k}} \left[\frac{\sigma_{r+j}}{\sigma_i - \sigma_{r+j}} < \mathfrak{X}, \Phi^+_{i,r+j} > \Phi^+_{i,r+j} - \frac{\sigma_{r+j}}{\sigma_i + \sigma_{r+j}} < \mathfrak{X}, \Phi^-_{i,r+j} > \Phi^-_{i,r+j}\right], \tag{2.39}$$

where $\Phi^{\pm}_{i,r+j}$ *are the principal directions of equation (2.38). More explicitly,*

$$\begin{aligned}\mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R}) = (I-UU^T)\mathfrak{X}Z(Z^TZ)^{-1}Z^T + UU^T\mathfrak{X} \\ + \sum_{\substack{1\leq i\leq r\\ 1\leq j\leq k}} \frac{\sigma_{r+j}}{\sigma_i^2-\sigma_{r+j}^2}[(\sigma_i u_{r+j}^T\mathfrak{X}v_i + \sigma_{r+j}u_i^T\mathfrak{X}v_{r+j})u_{r+j}v_i^T \\ + (\sigma_{r+j}u_{r+j}^T\mathfrak{X}v_i + \sigma_i u_i^T\mathfrak{X}v_{r+j})u_i v_{r+j}^T]. \quad (2.40)\end{aligned}$$

Proof. The set $\Omega = \{\mathfrak{R}\in\mathcal{M}_{l,m} | \sigma_{r+1}(\mathfrak{R}) > \sigma_r(\mathfrak{R})\}$ is open by continuity of the singular values and the boundary $\overline{\mathscr{M}}\backslash\mathscr{M}$ is the set of matrices of rank strictly lower than r, hence condition of theorem 2.1 are fulfilled. Equation (2.39) follows by replacing $\kappa_i(N)$ and Φ_i in (2.12) by the corresponding curvature eigenvalues $\pm\frac{\sigma_{r+j}}{\sigma_i}$ and eigenvectors $\Phi^{\pm}_{i,r+j}$ of theorem 2.3. □

Corollary 2.3. *Let* $\mathfrak{R}(t) = \sum_{i=1}^{r+k}\sigma_i(t)u_i(t)v_i(t)^T \in \mathcal{M}_{l,m}$ *the SVD of a time dependent matrix derivable with respect to* t *such that* $\sigma_r(t) > \sigma_{r+1}(t)$ *for all time. Then a dynamical system tracking the truncated SVD of* $\mathfrak{R}$ *is*

$$\begin{cases}\dot{U} = (I-UU^T)\dot{\mathfrak{R}}Z(Z^TZ)^{-1} \\ \qquad + \left[\sum_{\substack{1\leq i\leq r\\ 1\leq j\leq k}} \frac{\sigma_{r+j}}{\sigma_i^2-\sigma_{r+j}^2}(\sigma_i u_{r+j}^T\dot{\mathfrak{R}}v_i + \sigma_{r+j}u_i^T\dot{\mathfrak{R}}v_{r+j})u_{r+j}v_i^T\right]Z(Z^TZ)^{-1} \\ \dot{Z} = \dot{\mathfrak{R}}^TU + \left[\sum_{\substack{1\leq i\leq r\\ 1\leq j\leq k}} \frac{\sigma_{r+j}}{\sigma_i^2-\sigma_{r+j}^2}(\sigma_{r+j}u_{r+j}^T\dot{\mathfrak{R}}v_i + \sigma_i u_i^T\dot{\mathfrak{R}}v_{r+j})v_{r+j}u_i^T\right]U.\end{cases} \quad (2.41)$$

Proof. One just needs to apply the projection $\Pi_{\mathcal{T}(\Pi_{\mathscr{M}}(R))}$ to $\mathrm{D}_{\dot{\mathfrak{R}}}\Pi_{\mathscr{M}}(\mathfrak{R})$ with formula (2.29) to retrieve the coordinates of the horizontal space. □

In practice (2.41) is not very convenient to evolve the truncated SVD because it requires to keep track of all singular values and vectors. The following dynamical system allows to compute the truncated SVD from an approximate initial guess:

Proposition 2.16. *Consider* $\mathfrak{R}\in\mathcal{M}_{l,m}$ *such that its projection* $\Pi_{\mathscr{M}}(\mathfrak{R})$ *onto* $\mathscr{M}$ *is well defined, that is* $\sigma_r(\mathfrak{R}) > \sigma_{r+1}(\mathfrak{R})$*. Then the distance function* $J(R) = ||\mathfrak{R}-R||^2$ *admits no other local minima on* $\mathscr{M}$ *than* $\Pi_{\mathscr{M}}(\mathfrak{R})$*. In other words, for almost any initial rank* r *matrix* $U(0)Z(0)^T$*, the solution* $U(t)Z(t)^T$ *of the gradient flow*

$$\begin{cases}\dot{U} = (I-UU^T)\mathfrak{R}Z(Z^TZ)^{-1} \\ \dot{Z} = \mathfrak{R}^TU - Z\end{cases} \quad (2.42)$$

converges to $\Pi_{\mathscr{M}}(\mathfrak{R})$*, the truncated SVD of* $\mathfrak{R}$*.*

Proof. Let $\mathfrak{R} = \sum_{i=1}^{r+k}\sigma_i(\mathfrak{R})u_iv_i^T$ be the SVD of $\mathfrak{R}$ and $R\in\mathscr{M}$ such that $\nabla J(R) = 0$. It is known from definition 2.12 that such a point is of the form $R = \sum_{i\in A}\sigma_i u_i v_i^T$ where A is a subset of r indices $1\leq i\leq r+k$. corollary 2.1 states that R can be a local minimum of J only if all curvatures in the normal direction $N = \mathfrak{R}-R$ satisfy $\kappa_{ij}(N)\leq 1$. The

maximum of such curvatures being $\frac{\max_{j\notin A}\sigma_j(\mathfrak{R})}{\min_{i\in A}\sigma_i(\mathfrak{R})}$, this condition holds only if $A=\{1,\dots,r\}$, that is if $R=\Pi_{\mathscr{M}}(\mathfrak{R})$. The dynamical system (2.42) is obtained by replacing $\nabla J(R)$ with the expression (2.15). □

Remark 2.12. We cannot use directly corollary 2.1 to conclude that $\mathscr{M}$ is connected because $\mathscr{M}$ is not closed. Nevertheless, the existence of the SVD and the connectedness of the Stiefel manifold for $r<n$ implies that $\mathscr{M}$ is connected when $r<n$. Note that when $r=n$, $\mathscr{M}$ has two connected component constituted of the matrices of determinant respectively strictly positive and negative.

2.2.2 Stiefel Manifold, Orthogonal group and differentiability of the Polar decomposition

Definition 2.13. Let n be an integer greater than 2 and $p\leq n$. The Stiefel manifold is the set $\mathrm{St}_{n,p}$ embedded in the ambient space $\mathcal{M}_{n,p}$

$$\mathrm{St}_{n,p}=\{U\in\mathcal{M}_{n,p}|U^TU=I\}.$$

The orthogonal group $\mathcal{O}_n$ is the Stiefel manifold for the particular case $p=n$:

$$\mathcal{O}_n=\{P\in\mathcal{M}_{n,n}|P^TP=PP^T=I\}.$$

Proposition 2.17. *The tangent space $\mathcal{T}(UU^T)$ at $U\in\mathrm{St}_{n,p}$ is the set*

$$\begin{aligned}\mathcal{T}(UU^T)&=\{X\in\mathcal{M}_{n,p}|X^TU+XU^T=0\}\\&=\{\Delta+U\Omega|\Delta\in\mathcal{M}_{n,p},\,\Omega\in\mathcal{M}_{p,p}\text{ and }\Delta^TU=0,\,\Omega^T=-\Omega.\}\end{aligned}$$

Therefore $\mathrm{St}_{n,p}$ is a smooth manifold of dimension $np-p^2+p(p-1)/2=np-p(p+1)/2$.

Proof. (see also [36]) The first equality is obvious by differentiating $U^TU=I$ and the second equality is obtained by writing $X=UU^TX+(I-UU^T)X$ and setting $\Delta=(I-UU^T)X$ and $\Omega=U^TX$. □

We denote $\mathcal{H}_U=\{(\Delta,\Omega)\in\mathcal{M}_{n,p}\times\mathcal{M}_{p,p}|\Delta^TU=0,\,\Omega^T=-\Omega\}$ which we refer to as the horizontal space at U and which is just a convenient parameterization of the tangent space $\mathcal{T}(UU^T)$. In the following we identify therefore an element $(\Delta_X,\Omega_X)\in\mathcal{H}_U$ with a tangent vector $X=\Delta_X+U\Omega_X\in\mathcal{T}(UU^T)$. The embedded metric over $\mathrm{St}_{n,p}$ is given by

$$g(X,Y)=\mathrm{Tr}(\Delta_X^T\Delta_Y+\Omega_X^T\Omega_Y).$$

Proposition 2.18. *The projection $\Pi_{\mathcal{T}(UU^T)}$ on the tangent space $\mathcal{T}(UU^T)$ is the map*

$$\begin{array}{rccl}\Pi_{\mathcal{T}(UU^T)}&:&\mathcal{M}_{n,p}&\longrightarrow&\mathcal{T}(UU^T)\\&&\mathfrak{X}&\longmapsto&(I-UU^T)\mathfrak{X}+U\mathrm{skew}(U^T\mathfrak{X}),\end{array}$$

where $\mathrm{skew}(\mathfrak{X})=(\mathfrak{X}-\mathfrak{X}^T)/2$.

Proof. (see also [36]) $\Pi_{\mathcal{T}(UU^T)}\mathfrak{X}$ is the minimizer of the distance functional $J(\Delta,\Omega)=||\Delta+U\Omega-\mathfrak{X}||^2$ on the set $\mathcal{H}_U$. The vanishing of the gradient on that space is written:

$$\forall\delta\in\mathcal{M}_{n,p},\,U^T\delta=0\Rightarrow<\Delta+U\Omega-\mathfrak{X},\delta>=0,$$

$$\forall \delta \in \mathcal{M}_{p,p},\ \delta^T = -\delta \Rightarrow < \Delta + U\Omega - \mathfrak{X}, U\delta >=< \Omega - U^T\mathfrak{X}, \delta >= 0,$$

which yield respectively $\Delta = (I - UU^T)\mathfrak{X}$ and $\Omega = \mathrm{skew}(U^T\mathfrak{X})$. □

Remark 2.13. This proposition is also a reformulation of the "minimization principle" proposed by Babaee and Sapsis in Theorem 2.1. of [13]. Note that the authors also found $X_U = (I - UU^T)\mathfrak{X}$ but forgot to minimize J with respect to the antisymmetric matrix Ω.

Proposition 2.19. *The normal space at U is*

$$\mathcal{N}(U) = \{UT | T \in \mathrm{Sym}_p\}.$$

Proof. This is an immediate consequence of

$$\mathcal{N}(U) = (I - \Pi_{\mathcal{T}(UU^T)})\mathcal{M}_{n,p} = \{UU^T\mathfrak{X} - U\mathrm{skew}(U^T\mathfrak{X}) | \mathfrak{X} \in \mathcal{M}_{n,p}\}.$$

□

Proposition 2.20. *The covariant derivative on $\mathrm{St}_{n,p}$ is given by*

$$\nabla_X Y = (I - UU^T)\mathrm{D}_X(\Delta_Y) + \Delta_X\Omega_Y + U\mathrm{skew}(U^T\mathrm{D}_X\Delta_Y + \Omega_X\Omega_Y + \mathrm{D}_X\Omega_Y).$$

In particular geodesic equations on $\mathrm{St}_{n,p}$ are given by

$$\begin{cases} \dot{\Delta} = -U\Delta^T\Delta - \Delta\Omega \\ \dot{\Omega} = 0, \end{cases}$$

or more explicitly

$$\ddot{U} + U\dot{U}^T\dot{U} = 0,$$

with $\Omega = U^T\dot{U}$ being a constant.

Proof. We write that $\mathrm{D}_X Y = \mathrm{D}_X\Delta_Y + (\Delta_X + U\Omega_X)\Omega_Y + U\mathrm{D}_X\Omega_Y$ and then we obtain $\Delta_X Y$ by applying the projection $\Pi_{\mathcal{T}(UU^T)}$. To obtain geodesic equations, the condition $\nabla_{\dot{U}}\dot{U}$ become $(I - UU^T)\dot{\Delta} + \Delta\Omega = 0$ and $\dot{\Omega} + \mathrm{skew}(U^T\dot{\Delta} + \Omega^2) = 0$. Differentiating the condition $\Delta^T U = 0$, we obtain $U^T\dot{\Delta} = -\Delta^T\Delta$ and hence $\dot{\Delta} = UU^T\dot{\Delta} + (I - UU^T)\dot{\Delta} = -U\Delta^T\Delta - \Delta\Omega$ and $\dot{\Omega} = 0$. To obtain the equation in terms of U, we differentiate $\dot{U} = \Delta + U\Omega$ to write

$$\ddot{U} = \dot{\Delta} + (\Delta + U\Omega)\Omega = U(\Omega^2 - \Delta^T\Delta) = -U\dot{U}^T\dot{U}.$$

□

Remark 2.14. The reader will find an explicit formula for $U(t)$ in [36]. When $n = p$, the geodesic equation reduces to $\dot{\Omega} = 0$ and hence $\dot{U} = U\Omega$ implies $U(t) = U(0)e^{t\Omega}$.

Proposition 2.21. *Let $N = UT \in \mathcal{N}(U)$ a normal vector and $T = \sum_{i=1}^p \lambda_i(T)u_i u_i^T$ the eigenvalue decomposition of the symmetric matrix T. The Weingarten map of $\mathrm{St}_{n,p}$ with respect to the normal direction N is the application*

$$\begin{array}{rccc} L_U(N) & : & \mathcal{T}(UU^T) & \longrightarrow & \mathcal{T}(UU^T) \\ & & \Delta + U\Omega & \longmapsto & -\Delta T - U(\Omega T + T\Omega)/2. \end{array} \tag{2.43}$$

The principal curvatures in the direction N are constituted by the p real numbers

$$\kappa_i(T) = -\lambda_i(T),$$

associated with the $n-p$ dimensional eigenspaces

$$\{vu_i^T | v \in \mathrm{Span}(U)^\perp\},$$

and the $p(p-1)/2$ real numbers

$$\kappa_{ij}(T) = -\frac{\lambda_i(T) + \lambda_j(T)}{2},$$

associated with the normalized eigenvectors

$$\Phi_{ij} = \frac{U}{\sqrt{2}}(u_i u_j^T - u_j u_i^T).$$

Proof. We differentiate $\Pi_{\mathcal{T}(UU^T)}\mathfrak{X}$ with respect to U in the direction $X = \Delta + U\Omega$ before setting $\mathfrak{X} = N$. We obtain as such

$$\begin{aligned}
&\mathrm{D}\Pi_{\mathcal{T}(UU^T)}(X)\cdot N \\
&\qquad = -2\mathrm{sym}((\Delta + U\Omega)U^T)N + (\Delta + U\Omega)\mathrm{skew}(U^T N) + U\mathrm{skew}((\Delta + U\Omega)^T N) \\
&\qquad = -(\Delta U^T + U\Delta^T)N - U\mathrm{skew}(\Omega T),
\end{aligned}$$

which yields (2.43) by setting $N = UT$. Therefore an eigenvector $(\Delta, \Omega) \in \mathcal{H}_U$ of $L_U(N)$ with an eigenvalue λ satisfies

$$\begin{cases} -\Delta T = \lambda\Delta \\ -\dfrac{1}{2}(\Omega T + T\Omega) = \lambda\Omega. \end{cases}$$

One checks then that the vectors $(vu_i^T, 0)$ with v a vector in $\mathrm{Span}(U)^\perp$ and $(0, U(u_i u_j^T - u_j u_i^T)/\sqrt{2})$ are solution to this problem with the eigenvalues claimed. Because the total dimension formed by these eigenspaces coincides with the dimension of the tangent space, there are no other eigenvalues. □

Remark 2.15. The maximal curvature of $\mathrm{St}_{n,p}$ is therefore equal to $\kappa_\infty(R) = 1$ when $p < n$ and $\kappa_\infty(R) = \sqrt{2}/2$ when $p = n$. We also obtain that Christoffel symbol is given by $\Gamma(X, Y) = U\mathrm{sym}(\Delta_X^T \Delta_Y - \Omega_X \Omega_Y)$.

Proposition 2.22. *If $\mathfrak{R} \in \mathcal{M}_{n,p}^*$ is a full-rank n-by-p matrix, there exists a unique orthogonal projection $\Pi_{\mathrm{St}_{n,p}}(\mathfrak{R}) \in \mathrm{St}_{n,p}$ minimizing the Euclidean distance $U \mapsto ||\mathfrak{R} - U||$ from $\mathfrak{R}$ to a point U in the Stiefel manifold. This unique matrix is the polar part $P \in \mathrm{St}_{n,p}$ in the polar decomposition $\mathfrak{R} = PS$ with $S \in \mathcal{M}_{p,p}$ symmetric definite positive. The distance to $\mathrm{St}_{n,p}$ is*

$$||\mathfrak{R} - \Pi_{\mathrm{St}_{n,p}}(\mathfrak{R})||^2 = \sum_{i=1}^{n}(1 - \sigma_i(\mathfrak{R}))^2.$$

Note that if $\mathfrak{R} = U\Sigma V^T$ is the SVD of $\mathfrak{R}$ with $U \in \mathrm{St}_{n,p}$, $\Sigma \in \mathcal{M}_{p,p}$ diagonal and $V \in \mathcal{O}_p$, then $P = UV^T$ and $S = V\Sigma V^T$.

Proof. (see also [144]) A necessary condition for P to be the minimizer $P = \Pi_{\mathrm{St}_{n,p}}(\mathfrak{R})$ is that the residual $\mathfrak{R} - P$ must be a normal vector at P, namely $\mathfrak{R} = P(I + T)$ for some symmetric matrix $T \in \mathcal{M}_{p,p}$. Then the eigenvalues of T are of the form $\lambda_i(T) = \sigma_i(\mathfrak{R}) - 1$ or $\lambda_i(T) = -(\sigma_i(\mathfrak{R}) + 1)$, and the distance to $\mathrm{St}_{n,p}$ is given by $||\mathfrak{R} - P||^2 = \sum_{i=1}^{n}(\pm\sigma_i(\mathfrak{R}) - 1)^2$. This summation is minimized only when $\lambda_i(T) = \sigma_i(\mathfrak{R}) - 1$, i.e. for $S = I + T$ symmetric and positive. The condition that in addition, the singular values $\sigma_i(\mathfrak{R})$ are non zero, that is S definite, ensures that that the polar part is uniquely defined [58], i.e. the uniqueness of the orthogonal projection. □

Proposition 2.23. *Consider the polar decomposition $\mathfrak{R} = PS$ of a full rank matrix $\mathfrak{R} \in \mathcal{M}^*_{n,p}$ with $S \in \mathcal{M}_{p,p}$ symmetric positive definite and $P \in \mathrm{St}_{n,p}$. Denote $S = \sum_{i=1}^{r} \sigma_i(\mathfrak{R}) u_i u_i^T$ the eigendecomposition of S. Then the orthogonal projection $\Pi_{\mathrm{St}_{n,p}}$, namely the application $\mathfrak{R} \mapsto P$ is differentiable at $\mathfrak{R}$ and the differential in the direction $\mathfrak{X}$ is given by the formula*

$$\begin{aligned} \mathrm{D}_{\mathfrak{X}}\Pi_{\mathrm{St}_{n,p}}(\mathfrak{R}) = &\sum_{\{i,j\}} \frac{2}{\sigma_i(\mathfrak{R}) + \sigma_j(\mathfrak{R})} \left(u_i^T \mathrm{skew}(P^T\mathfrak{X}) u_j\right) P(u_i u_j^T - u_j u_i^T) \\ &+ \sum_{i=1}^{p} \frac{1}{\sigma_i(\mathfrak{R})}(I - PP^T)\mathfrak{X} u_i u_i^T. \end{aligned} \tag{2.44}$$

Proof. The equation is immediately obtained by applying formula (2.12) of theorem 2.4 with the normal vector $N = \mathfrak{R} - P = P(S - I)$, for which one finds

$$1 - \kappa_i(N) = 1 - (1 - \sigma_i(\mathfrak{R})) = \sigma_i(\mathfrak{R}),$$

$$\sum_{\mathrm{Span}(e_i) = \mathrm{Span}(P)^\perp} < \mathfrak{X}, e_i u_i^T > e_i u_i^T = \sum_{\mathrm{Span}(e_i) = \mathrm{Span}(P)^\perp} ((e_i^T \mathfrak{X} u_i) e_i) u_i^T = (I - PP^T)\mathfrak{X} u_i u_i^T,$$

$$1 - \kappa_{ij}(N) = 1 - \big(1 - (\sigma_i(\mathfrak{R}) + \sigma_j(\mathfrak{R}))/2\big) = (\sigma_i(\mathfrak{R}) + \sigma_j(\mathfrak{R}))/2,$$

$$\begin{aligned} < \Phi_{ij}, \mathfrak{X} > \Phi_{ij} &=< P\mathrm{skew}(u_i u_j^T), \mathfrak{X} > P(u_i u_j^T - u_j u_i^T) \\ &=< u_i u_j^T, \mathrm{skew}(P^T\mathfrak{X}) > P(u_i u_j^T - u_j u_i^T). \end{aligned}$$

□

Remark 2.16. Formula (2.44) has been obtained by Chen for the orthogonal group (that is in the case $n = p$ for which the second summation is 0) by using purely algebraic (see p. 181 in [25]). In continuum mechanics, one often consider $\mathfrak{R}(t) = F(t) = \nabla \boldsymbol{\phi}_{\mathbf{0}}^{\boldsymbol{t}}(\boldsymbol{x})$ as being the gradient of a transformation $\boldsymbol{\phi}_{\mathbf{0}}^{\boldsymbol{t}}(\boldsymbol{x})$, and $\mathfrak{X} = \nabla \boldsymbol{v}$ as the gradient of the velocity field evolving the transformation (namely $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\phi}_0^t(\boldsymbol{x}) = \boldsymbol{v}(t, \boldsymbol{\phi}_0^t(\boldsymbol{x}))$). The orthogonal matrix P and the symmetric matrix S of the polar decomposition $F(t) = P(t)S(t)$ are respectively interpreted as the rotation and the stretching components of the transformation [144, 67]. At $t = 0$, F is the identity matrix and equation (2.44) becomes $\dot{P}\Big|_{t=0} = (\nabla \boldsymbol{v} - \nabla \boldsymbol{v}^T)/2$. Formula (2.44) generalizes the well known result that the instantaneous rotation rate of a transformation is half the vorticity (see e.g. [67]).

The following proposition gives a dynamical systems that achieves the polar decomposition analogously as proposition 2.16:

Proposition 2.24. *Consider a full rank matrix $\mathfrak{R} \in \mathcal{M}^*_{n,p}$ and $\mathfrak{R} = \sum_{i=1}^{p} \sigma_i(\mathfrak{R}) u_i v_i^T$ its Singular Value Decomposition.*

- *If $p < n$, then $\Pi_{\mathrm{St}_{n,p}}(\mathfrak{R})$ is the unique local minimum of the distance function $J : U \mapsto \frac{1}{2}||\mathfrak{R} - U||^2$, and therefore, for almost any initial data $U(0) \in \mathrm{St}_{n,p}$, the solution $U(t)$ of the gradient flow*

$$\dot{U} = \mathfrak{R} - \frac{1}{2}(UU^T\mathfrak{R} + U\mathfrak{R}^T U) \tag{2.45}$$

converges to the polar part $\Pi_{\mathrm{St}_{n,p}}(\mathfrak{R}) = \sum_{i=1}^{p} u_i v_i^T$ of $\mathfrak{R}$. In particular, $\mathrm{St}_{n,p}$ is connected for $p < n$.

- *If $n = p$, then J admits other local minima that are the matrices $U \in \mathrm{St}_{n,p}$ of the form*

$$U = \sum_{i=1}^{n-1} u_i v_i^T - u_n v_n^T, \tag{2.46}$$

where u_n is a singular vector corresponding to the smallest singular value $\sigma_n(\mathfrak{R})$. Therefore any solution $U(t)$ of the gradient flow (2.45) converges almost surely to the polar part $\Pi_{\mathrm{St}_{n,p}}(\mathfrak{R})$ provided the initial data $U(0)$ lies in the same connected component of $\mathcal{O}_n$. Otherwise, $U(t)$ converges almost surely towards an element $U \in \mathcal{O}_n$ of the form (2.46).

Proof. Let $U \in \mathrm{St}_{n,p}$ such that $\nabla J(U) = 0$. Then (1.12) shows that $\mathfrak{R} = U(I + T)$ with $\lambda_i(T) = \sigma_i(\mathfrak{R}) - 1$ or $\lambda_i(T) = -(\sigma_i(\mathfrak{R}) + 1)$. Denote $N = \mathfrak{R} - U = UT$ the residual normal vector. If $p < n$ then the condition $\forall 1 \leq i \leq p$, $\kappa_i(N) = -\lambda_i(T) \leq 1$ required for U to be a local cannot be satisfied if there exists i such that $\lambda_i(T) = -(\sigma_i(\mathfrak{R}) + 1)$. This proves that the only local minimum is achieve by $\Pi_{\mathscr{M}}(\mathfrak{R})$.

If $p = n$, then the condition for U to be a minimum is that $\kappa_{ij}(N) = -\frac{1}{2}(\lambda_i(T) + \lambda_j(T)) \leq 1$ for all pair $\{i, j\}$. This condition cannot be satisfied if there exists at least two indices i and j such that $\lambda_i(T) = -(\sigma_i(\mathfrak{R}) + 1)$ and $\lambda_j(T) = -(\sigma_j(\mathfrak{R}) + 1)$. If i is an index such that $\lambda_i(T) = -(\sigma_i(\mathfrak{R}) + 1)$, then $\kappa_{ij}(N) \leq 1$ implies $\forall j \neq i$, $\sigma_i(\mathfrak{R}) \leq \sigma_j(\mathfrak{R})$ therefore $i = n$ and U is of the form (2.46). Finally the gradient flow is obtained by making $\dot{U} = -\nabla J(U)$ explicit with $\nabla J(U) = (I - UU^T)(U - \mathfrak{R}) + U\mathrm{skew}(U^T(U - \mathfrak{R}))$. □

2.2.3 The isospectral manifold, the Grassmannian, and the geometry of mutually orthogonal subspaces

In this part we denote Sym_n the set of n-by-n symmetric matrices. The focus is on the isospectral manifold (also studied by [20, 31]), for which each point can be considered as a collection of m subspaces orthogonal one another of a n dimensional Euclidean space, whose given dimensions are $n_1, \ldots, n_m$ and such that $n = n_1 + \cdots + n_m$. A symmetric matrix S having m eigenvalues with multiplicities $n_1, \ldots, n_m$ can be used to represent such collection of subspaces: one identifies S to the collection of its eigenspaces. In the particular case where $m = 2$, the isospectral manifold is more commonly known as the Grassman manifold or Grassmanian, whose points are p dimensional subspaces. This leads to the following definition:

Definition 2.14. The isospectral manifold is the set $\mathscr{I}$ of all symmetric n-by-n matrices $S \in \mathrm{Sym}_n$ whose spectrum is constituted of m distinct eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_m$ with respective multiplicities $n_1, \ldots, n_m$. Denoting Λ such a matrix, $\mathscr{I}$ admits two convenient

parameterizations :

$$\begin{aligned}\mathscr{I} &= \{P\Lambda P^T | P \in \mathcal{O}_n\} \\ &= \{\sum_{i=1}^{m} \lambda_i U_i U_i^T | U_i \in \mathcal{M}_{n,m_i},\, U_i^T U_j = \delta_{ij}\}.\end{aligned}$$

The Grassman manifold $\mathcal{G}$ is defined to be the set of all projectors over a d dimensional subspace :

$$\mathcal{G} = \{UU^T \in \mathcal{M}_{n,n} | U^T U = I \text{ and } U \in \mathcal{M}_{n,p}\}.$$

A set of matrices $U_i \in \mathcal{M}_{n,m_i}$ may be used to describe points on the manifold $\mathscr{I}$, where each U_i represent each eigenspace $\mathrm{Span}(U_i)$. When $m = 2$, a single matrix U can be used more conveniently to represent points on the Grassmannian $\mathcal{G}$, since the knowledge of a vector space or of its orthogonal complement are equivalent.

The time derivative of a trajectory $U_i(t)$ can be decomposed along the basis given by the union of the U_k as $\dot{U}_i = \sum_{j=1}^{m} U_j \Delta_i^j$ with $\Delta_i^j \in \mathcal{M}_{m_j,m_i}$. The matrix Δ_i^j is related to the magnitude of the rotation of the subspace $\mathrm{Span}(U_i)$ around the axis given by the subspace $\mathrm{Span}(U_j)$. So as to remain orthogonal one another, a condition must be satisfied by the Δ_i^j :

Proposition 2.25. *The tangent space $\mathcal{T}(S)$ at $S \in \mathscr{I}$ is the set*

$$\begin{aligned}\mathcal{T}(S) &= \{[\Omega, S] = \Omega S - S\Omega | \Omega \in \mathcal{M}_{n,n},\, \Omega^T = -\Omega\} \\ &= \left\{ \sum_{\{i,j\}\subset\{1,\dots,m\}} (\lambda_i - \lambda_j)(U_j \Delta_i^j U_i^T - U_i \Delta_j^i U_j^T) \Big| \Delta_i^j \in \mathcal{M}_{n_j,n_i},\ \Delta_j^i = -(\Delta_i^j)^T \right\} \\ &= \left\{ \sum_{i\neq j} (\lambda_i - \lambda_j) U_j \Delta_i^j U_i^T \Big| \Delta_i^j \in \mathcal{M}_{n_j,n_i},\, \Delta_j^i = -(\Delta_i^j)^T. \right\}\end{aligned} \tag{2.47}$$

The Δ_i^j defined in the above expressions for each pair $\{i, j\} \subset \{1, \dots, m\}$ parameterizes uniquely the tangent space $\mathcal{T}(S)$. Therefore the $\mathscr{I}$ is a smooth manifold of dimension $\left(n^2 - \sum_{i=1}^{m} n_i^2\right)/2$.

Proof. (see also [31]). Consider first the parameterization of $\mathscr{I}$ by $\mathcal{O}_n$, i.e. write $S = P\Lambda P^T$. Denote Ω a skew-symmetric matrix. Differentiating with respect to P in the direction $P\Omega$ tangent to $\mathcal{O}_n$ at P, one obtains that a tangent vector X is of the form $X = P\Omega\Lambda - \Lambda\Omega P^T = P\Omega P^T P\Lambda P^T - P\Lambda P^T P\Omega P^T = [P\Omega P^T, S]$. This shows the first equality. Write now $S = \sum_{i=1}^{m} \lambda_i U_i U_i^T$ with $U_i \in \mathcal{M}_{n,m_i}$ and $U_i^T U_j = \delta_{ij}$. Differentiating the constraint $U_i^T U_j = 0$ yields $(\Delta_i^j)^T = -\Delta_j^i$. This implies that a tangent vector is of the form $X = \sum_{i,j=1}^{m} \lambda_i (U_j \Delta_i^j U_i^T - U_i \Delta_j^i U_j^T)$ which gives the expression claimed after reordering. If X is a tangent vector, the formula $\Delta_i^j = U_j^T X U_i/(\lambda_i - \lambda_j)$ determines uniquely Δ_i^j. Finally, the dimension of $\mathcal{T}(S)$ is

$$\sum_{\{i,j\}} n_j n_i = \frac{1}{2} \sum_{i\neq j} n_i n_j = \frac{1}{2} \left(\sum_{i,j} n_i n_j - \sum_{i=1}^{m} n_i^2 \right) = \left(n^2 - \sum_{i=1}^{m} n_i^2 \right)/2.$$

□

Remark 2.17. For the Grassman manifold, the tangent space $\mathcal{T}(UU^T)$ at UU^T is more conveniently (and uniquely) described as

$$\mathcal{T}(UU^T) = \{\Delta U^T + U\Delta | \Delta^T U = 0, \Delta \in \mathcal{M}_{n,p}\}. \tag{2.48}$$

where the matrix Δ can be understood as the time derivative $\Delta = \dot{U}$ of the column matrix $U \in \mathcal{M}_{n,p}$ used to represent the point $UU^T \in \mathcal{G}$. The dimension of the Grassman manifold is $p(n-p)$.

We denote $\mathcal{H}$ the set of all $\Delta_i^j \in \mathcal{M}_{n_j,n_i}$ which we refer to as the horizontal space. In the following, for a given $X \in \mathcal{T}(S)$, we will denote $(\Delta_X)_i^j$ the associated coordinates in the horizontal space. Note that the metric on $\mathscr{I}$ is given by

$$< X, Y >= \sum_{i,j} (\lambda_i - \lambda_j)^2 \mathrm{Tr}[(\Delta_X)_i^{jT} (\Delta_Y)_i^j].$$

Proposition 2.26. *The projection $\Pi_{\mathcal{T}(S)}$ on the tangent space $\mathcal{T}(S)$ is the map*

$$\begin{array}{rccc} \Pi_{\mathcal{T}(S)} & : & \mathrm{Sym}_n & \longrightarrow & \mathcal{T}(S) \\ & & \mathfrak{X} & \longmapsto & \sum\limits_{\{i,j\} \subset \{1,\dots,m\}} (U_j U_j^T \mathfrak{X} U_i U_i^T + U_i U_i^T \mathfrak{X} U_j U_j^T), \end{array}$$

that is with the coordinates of the horizontal space, $\Delta_i^j = U_j^T \mathfrak{X} U_i / (\lambda_i - \lambda_j)$.

Proof. This is obtained by differentiating $||\mathfrak{X} - X||^2$ with respect to Δ_i^j, for a tangent vector $X \in \mathcal{H}$ written with the coordinates Δ_i^j of the horizontal space. □

Remark 2.18. For the Grassman manifold, with the coordinates of remark 2.17, this projection is more conveniently written as $\Delta_{\Pi_{\mathcal{T}(UU^T)}\mathfrak{X}} = (I - UU^T)\mathfrak{X}U$.

Proposition 2.27. *The normal space $\mathcal{N}(S)$ at S is the set of all symmetric matrices N that let stable each eigenspace $\mathrm{Span}(U_i)$ of S:*

$$\mathcal{N}(S) = \left\{ \sum_{i=1}^m U_i U_i^T \mathfrak{X} U_i U_i^T | \mathfrak{X} \in \mathrm{Sym}_n \right\}.$$

In other words, it is the set of all matrices $N \in \mathrm{Sym}_n$ of the form

$$N = \sum_{i=1}^m \sum_{a=1}^{n_i} \lambda_{i,a}(N) u_{i,a} u_{i,a}^T,$$

where for each $1 \leq i \leq m$, $\lambda_{i,a}(N)_{1 \leq a \leq n_i}$ is a set of n_i real eigen-values associated with n_i eigenvectors $(u_{i,a})_{1 \leq a \leq n_i}$ forming a basis of the eigenspace $\mathrm{Span}(U_i)$.

Proof. This is an immediate consequence of $\mathcal{N}(S) = \{(I - \Pi_{\mathcal{T}(S)})\mathfrak{X} | \mathfrak{X} \in \mathrm{Sym}_n\}$. □

Proposition 2.28. *The covariant derivative on $\mathscr{I}$ is given by, in the coordinates of the horizontal space:*

$$(\Delta_{\nabla_X Y})_i^j = D_X(\Delta_Y)_i^j + \sum_{k=1}^m \left[\frac{\lambda_i - \lambda_k}{\lambda_i - \lambda_j} (\Delta_X)_k^j (\Delta_Y)_i^k + \frac{\lambda_j - \lambda_k}{\lambda_i - \lambda_j} (\Delta_Y)_k^j (\Delta_X)_i^k \right].$$

Therefore geodesic equations on $\mathscr{I}$ are given by

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}U_i = U_j\Delta_i^j \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\Delta_i^j = -\displaystyle\sum_{k=1}^{m} \frac{\lambda_i - 2\lambda_k + \lambda_j}{\lambda_i - \lambda_j}\Delta_k^j\Delta_i^k. \end{cases}$$

Proof. Let Y a tangent vector parameterized by $(\Delta_Y)_i^j$ as in (2.47). We have

$$\mathrm{D}_X Y = \sum_{i\neq j}(\lambda_i - \lambda_j)\Big[U_k(\Delta_X)_j^k(\Delta_Y)_i^j U_i^T + U_j\mathrm{D}_X(\Delta_Y)_i^j U_i^T - U_j(\Delta_Y)_i^j(\Delta_X)_k^i U_k^T\Big],$$

where summation over the repeated index k is assumed. Writing $(\Delta_{\nabla_X^Y})_i^j = \frac{1}{\lambda_i-\lambda_j}U_j^T\mathrm{D}_X Y U_i$ yields the result. □

Remark 2.19. For the Grassman manifold $\mathcal{G}$, denoting V a matrix spanning the orthogonal complement of $\mathrm{Span}(U)$, we find that $\Delta_1^2 = V^TU$ is a constant and geodesic equations are written $\ddot{U} = U\Delta_1^2\Delta_2^1 = -U\dot{U}^T\dot{U}$. This is the same equation than the one for the embedded Stiefel manifold given previously. Noticing that $\dot{U}^T\dot{U}$ is a constant, one can find an explicit formula for $U(t)$ (see [36]).

Proposition 2.29. *Let $N = \sum_{i=1}^m\sum_{a=1}^{n_i}\lambda_{i,a}(N)u_{i,a}u_{i,a}^T$ the eigenvalue decomposition of a normal vector N at S. The Weingarten map in the direction N on $\mathscr{I}$ is given by*

$$\begin{array}{rccc} L_S(N) & : & \mathcal{H} & \longrightarrow & \mathcal{H} \\ & & (\Delta_i^j) & \longmapsto & \Big(\frac{1}{\lambda_i-\lambda_j}(U_j^TNU_j\Delta_i^j - \Delta_i^jU_i^TNU_i)\Big)_i^j. \end{array} \tag{2.49}$$

The principal curvatures are the real

$$\kappa_{i,a}^{j,b} = \frac{\lambda_{j,b}(N) - \lambda_{i,a}(N)}{\lambda_i - \lambda_j},$$

for all pairs $\{i,j\} \subset \{1,\dots,m\}$ and couples (a,b) with $1 \leq a \leq n_i$ and $1 \leq b \leq n_j$. The corresponding normalized eigendirections are the tangent vectors

$$\Phi_{i,a}^{j,b} = \frac{1}{\sqrt{2}}(u_{i,a}u_{j,b}^T + u_{j,b}u_{i,a}^T).$$

Proof. Differentiating $\Pi_{\mathcal{T}(S)}N$ with respect to the tangent direction $X = (\Delta_i^j)$ yields

$$\mathrm{D}\Pi_{\mathcal{T}(S)}(X)\cdot N = \sum_{i\neq j}\Big[(U_k\Delta_j^kU_j^T - U_j\Delta_k^jU_k^T)NU_iU_i^T + U_jU_j^TN(U_k\Delta_i^kU_i^T - U_i\Delta_k^iU_k^T)\Big],$$

with summation over repeated indices k. Using the fact that N is a normal vector, we obtain:

$$\mathrm{D}\Pi_{\mathcal{T}(S)}(X)\cdot N = \sum_{i\neq j}\Big[-U_j\Delta_i^jU_i^TNU_iU_i^T + U_jU_j^TNU_j\Delta_i^jU_i^T\Big].$$

Expression (2.49) follows from $(\Delta_{\mathrm{D}\Pi_{\mathcal{T}(S)}(X)\cdot N})_i^j = U_j^T(\mathrm{D}\Pi_{\mathcal{T}(S)}(X)\cdot N)U_i/(\lambda_i-\lambda_j)$. It is easy then to check that $\Delta_{i,a}^{j,b} = U_j^Tu_{j,b}u_{i,a}^TU_i$ provide a basis of eigenvectors with eigenvalues

$\kappa_{i,a}^{j,b}$, whence the result after normalization. □

Remark 2.20. The maximal curvature of the isospectral manifold is therefore given by $\kappa_\infty(S) = \max_{\{i,j\}} \frac{\sqrt{2}}{|\lambda_i - \lambda_j|}$.

Proposition 2.30. *Let $\mathbf{S} \in \mathrm{Sym}_n$ be a symmetric matrix and denote*

$$\mathbf{S} = \sum_{i=1}^{m} \sum_{a=1}^{n_i} \lambda_{i,a}(\mathbf{S}) u_{i,a} u_{i,a}^T$$

its eigenvalue decomposition, where the eigenvalues have been ordered decreasingly, i.e

$$\forall 1 \leq a_i \leq n_i,\ \lambda_{1,a_1} \geq \lambda_{2,a_2} \geq \ldots \geq \lambda_{m,a_m},$$
$$\forall 1 \leq i \leq m,\ \lambda_{i,1} \geq \lambda_{i,2} \geq \ldots \geq \lambda_{i,n_i}.$$

If for any $1 \leq i \leq m-1$, $\lambda_{i+1,1}(\mathbf{S}) > \lambda_{i,n_i}(\mathbf{S})$, that is the eigenspaces of $\mathbf{S}$ are well separated relatively to the ordering given by Λ, then $\mathbf{S}$ admits a unique projection onto $\mathscr{I}$. This projection is obtained by replacing the eigenvalues of $\mathbf{S}$ by those of Λ :

$$\Pi_{\mathscr{I}}(\mathbf{S}) = \sum_{i=1}^{m} \sum_{a=1}^{n_i} \lambda_i u_{i,a} u_{i,a}^T.$$

The distance of $\mathbf{S}$ to the manifold $\mathscr{I}$ is given by $||\mathbf{S} - \Pi_{\mathscr{I}}(\mathbf{S})||^2 = \sum_{i=1}^{m} \sum_{a=1}^{n_i} (\lambda_{i,a} - \lambda_i)^2$.

Proof. (see also [31]) For a given $S \in \mathscr{I}$, $N = \mathbf{S} - S$ is a normal vector at S if $\mathbf{S} = S + N$ with S and N that can be diagonalized in the same basis. To make notations easy, denote $\mathbf{S} = \sum_{l=1}^{n} \lambda_l(\mathbf{S}) u_l u_l^T$ the eigen decomposition of $\mathbf{S}$ (no ordering assumed). We have $N = \sum_{l=1}^{n} (\lambda_l(\mathbf{S}) - \Lambda_{\sigma(l)}) u_l u_l^T$ where the Λ_l are the eigenvalues λ_i of Λ ordered decreasingly (including multiplicities), and σ a permutation. Noticing that for any given numbers satisfying $a < b$ and $c < d$, we have $(a-c)^2 + (b-d)^2 < (a-d)^2 + (b-c)^2$, we see that the norm of N is minimized by selecting the permutation σ to be the identity. □

Remark 2.21. A possible interpretation of this result in the case of the Grassman manifold, is that each symmetric matrix $\mathbf{S}$ of the ambient space can be interpreted as a collection of n orthogonal directions $(u_i)_{1 \leq i \leq n}$ that have been assigned a "score" $\lambda_i(\mathbf{S})$. A point on the Grassman manifold corresponds to selecting p directions u_i with the score 1 and rejecting other directions. One projects a point $\mathbf{S}$ of the ambient space by selecting the p directions that have the highest score.

Proposition 2.31. *Let $\mathbf{S} \in \mathrm{Sym}_n$ is a symmetric matrix defined as in proposition 2.30. The projection onto $\mathscr{I}$ is differentiable at $\mathbf{S}$ and the derivative in a direction $\mathfrak{X} \in \mathrm{Sym}_n$ is given by*

$$\mathrm{D}_{\mathfrak{X}} \Pi_{\mathscr{I}}(\mathbf{S}) = \sum_{\substack{\{i,j\} \subset \{1,\ldots,m\} \\ 1 \leq a \leq n_i \\ 1 \leq b \leq n_j}} \frac{\lambda_i - \lambda_j}{\lambda_{i,a}(\mathbf{S}) - \lambda_{j,b}(\mathbf{S})} (u_{i,a}^T \mathfrak{X} u_{j,b})(u_{i,a} u_{j,b}^T + u_{j,b} u_{i,a}^T). \tag{2.50}$$

Proof. We apply the formula of theorem 2.1 with $N = \sum_{i=1}^{m} \sum_{a=1}^{n_i} (\lambda_{i,a}(\mathbf{S}) - \lambda_i) u_{i,a} u_{i,a}^T$. We find

$$1 - \kappa_{i,a}^{j,b}(N) = \frac{\lambda_{j,b}(\mathbf{S}) - \lambda_{i,a}(\mathbf{S})}{\lambda_i - \lambda_j},$$

$$< \mathfrak{X}, \Phi^{j,b}_{i,a} > \Phi^{j,b}_{i,a} = \frac{1}{2} < \mathfrak{X}, 2\mathrm{sym}(u_{i,a}u_{j,b}^T) > (u_{i,a}u_{j,b}^T + u_{j,b}u_{i,a}^T),$$

which yields the expression claimed. □

Corollary 2.4. *Consider* $\mathbf{S}(t) = \sum_{i=1}^n \lambda_i(t)u_i(t)u_i(t)^T$ *the eigendecomposition of a time dependent symmetric matrix. Consider* $I \subset \{1, \dots, n\}$ *a subset of* p *indices such that the subspace* $\mathcal{U}(t) = \mathrm{Span}(u_i(t))_{i\in I}$ *spanned by the corresponding* p *eigenvectors is well defined for all time, i.e. there is no couple of indices* $i \in I$ *and* $j \notin I$ *such that* $\lambda_i(t) = \lambda_j(t)$*. Then the subspace* $\mathcal{U}(t)$ *is differentiable with respect to* t[2] *and an ODE for the evolution of a corresponding orthonormal basis of vectors* $U \in \mathcal{M}_{n,p}$ *satisfying* $\mathrm{Span}(U(t)) = \mathcal{U}(t)$ *is*

$$\dot{U} = \sum_{\substack{i\in I \\ j\notin I}} \frac{1}{\lambda_i(t) - \lambda_j(t)} (u_i^T \dot{\mathbf{S}} u_j) u_j u_i^T U. \tag{2.51}$$

Proof. Consider the isospectral manifold $\mathscr{I}$ with the multiplicity of the spectrum Λ being adapted to I. The case where I is a set of p indices corresponding to a subspace $\mathrm{Span}(U_k)$ spanned by p consecutive eigenvectors is obtained by writing $\dot{U}_k = U_j \Delta^j_k$ with $\Delta^j_k = U_j^T \mathrm{D}_{\dot{\mathbf{S}}} \Pi_{\mathscr{I}}(\mathbf{S}) U_k / (\lambda_k - \lambda_j)$. For the general case, one writes the projector onto $\mathcal{U}$ as $UU^T = \sum_{k\in\tilde{I}} U_k U_k^T$, $\tilde{I}$ being a set of indices such that $\mathrm{Span}(U_k)_{k\in\tilde{I}} = \mathrm{Span}(u_i)_{i\in I}$ and each matrix U_k spanning sets of consecutive eigenvectors. Then one obtains a time derivative for U by writing $\dot{U} = (I - UU^T)\left(\sum_{k\in\tilde{I}} \frac{\mathrm{d}}{\mathrm{d}t}(U_k U_k^T)\right) U$, which gives the result. □

As an application we provide a dynamical system that finds the dominant subspaces of a symmetric matrix, which is the analogous of proposition 2.16 and proposition 2.24.

Proposition 2.32. *Consider a symmetric matrix* $\mathbf{S} = \sum_{i=1}^m \sum_{a=1}^{n_i} \lambda_{i,a}(\mathbf{S}) u_{i,a} u_{i,a}^T \in \mathrm{Sym}_n$ *such that its projection* $\Pi_{\mathscr{I}}(\mathbf{S})$ *is uniquely defined. Then the distance functional* $S \mapsto ||\mathbf{S} - S||^2$ *admits no other local minimum on* $\mathscr{I}$ *than* $\Pi_{\mathscr{I}}(\mathbf{S})$*. Therefore, for almost any initial data* $S(0) \in \mathrm{Sym}_n$*, the solution* $S(t) = \sum_{i=1}^m \lambda_i U_i U_i^T$ *of the gradient flow*

$$\dot{U}_i = \sum_{j\neq i} \frac{1}{\lambda_i - \lambda_j} U_j U_j^T \mathbf{S} U_i \tag{2.52}$$

converges to $\Pi_{\mathscr{I}}(\mathbf{S})$*, or in other words, each of the matrices* U_i *converge to a matrix spanning the same subspace as* $\mathrm{Span}(u_{i,a})_{1\leq a\leq n_i}$*. In particular, the isospectral manifold is connected.*

Proof. Denote $N = \mathbf{S} - S$ the residual normal vector of a critical point S of the distance functional. The condition for S to be a local minimum is that all curvatures in the direction N satisfy $\kappa^{j,b}_{i,a}(N) \leq 1$, which is equivalent to the condition $\frac{\lambda_{i,a} - \lambda_{j,b}}{\lambda_i - \lambda_j} \geq 0$. This condition can be satisfied only for $S = \Pi_{\mathscr{I}}(\mathbf{S})$. □

Remark 2.22. In [20], Brockett considered the double-bracket flow $\dot{H} = [H, [H, \mathbf{S}]]$ on $\mathscr{I}$ to achieve the diagonalization of a symmetric matrix starting from $H(0) = \Lambda$ and proved analogous convergence results (also in [31]). This flows coincides with the gradient descent for the distance functional $J(P) = ||P\Lambda P^T - \mathbf{S}||^2$ with respect to $P \in \mathcal{O}_n$. The reader can check that the corresponding expression in the horizontal coordinates is

$$\dot{U}_i = \sum_{j\neq i} (\lambda_i - \lambda_j) U_j U_j^T \mathbf{S} U_i,$$

[2] in the sense that the projector over this subspace is differentiable.

instead of (2.52). In other words, it reduces to use a gradient descent where each of the components Δ_i^j of the covariant gradient $\nabla J = \Pi_{\mathcal{T}(S)}(\mathbf{S} - S)$ have been rescaled by the positive numbers $(\lambda_i - \lambda_j)^2$.

Remark 2.23. Applying this result to the particular case of the Grassman manifold, we obtain that for almost any initial data $U(0) \in \mathrm{St}_{n,p}$, the solution $U(t)$ of the gradient flow

$$\dot{U} = (I - UU^T)\mathbf{S}U \tag{2.53}$$

converges to a matrix $U \in \mathrm{St}_{n,p}$ whose columns span the dominant subspace of $\mathbf{S}$. A generalization of this result for the case where $\mathbf{S}$ is not necessary symmetric has also been found recently by Babaee and Sapsis (Theorem 2.3 in [13]).

2.2.4 Non euclidean grassmanian, biorthogonal manifold, and derivative of eigenspaces of nonsymmetric matrices

In [13], it has been found that *even if* $\boldsymbol{S}$ *is not symmetric*, the solution $U(t)$ of the dynamical system (2.52) converges almost surely , and the limit U generates the same subspace than the one spanned by the eigenvectors of $\boldsymbol{S}$ associated with the eigenvalues of maximal real parts. In the following, we recast this result in the framework of oblique projections developped in section 2.1.3. We consider the matrix space $\mathcal{M}_{n,n}$ and an integer $p \leq n$. Given a matrix $\mathfrak{R} \in \mathcal{M}_{n,n}$ we denote $\lambda_i(\mathfrak{R})$ its complex eigenvalues, where these eigenvalues have been ordered according to the real parts: $\Re(\lambda_1(\mathfrak{R})) \geq \Re(\lambda_2(\mathfrak{R})) \geq \ldots \geq \Re(\lambda_n(\mathfrak{R}))$. When $\Re(\lambda_p(\mathfrak{R})) > \Re(\lambda_{p+1}(\mathfrak{R}))$, there exist a unique p-dimensional stable subspace on which the complex eigenvalues of $\mathfrak{R}$ are $(\lambda_i(\mathfrak{R}))_{1\leq i\leq p}$ (see [75]), which we refer to as the *dominant subspace of* $\mathfrak{R}$. If in addition $\mathfrak{R} = \sum_{i=1}^n \lambda_i(\mathfrak{R}) u_i \overline{v}_i^T$ is diagonalizable in $\mathbb{C}$, this dominant subspace is given by $\mathrm{Span}(\Re(u_i), \Im(u_i))_{1\leq i\leq p}$ where u_i and v_i are the respective right and left eigenvectors of $\mathfrak{R}$ satisfying $\overline{v}_i^T u_j = \delta_{ij}$ (eqn. (2.21)).

In the following we investigate the differential of two applications, and provide dynamical system to compute them:

1. $\mathfrak{R} \mapsto UU^T$ where $UU^T \in \mathcal{G}$ is the orthogonal projector over the dominant subspace of $\mathfrak{R}$.

2. $\mathfrak{R} \mapsto \sum_{i=1}^p u_i \overline{v}_i^T$ (when $\mathfrak{R}$ is diagonalizable), or in other words the application mapping $\mathfrak{R}$ to the (non-orthogonal) linear projector whose image is the dominant subspace of $\mathfrak{R}$ and kernel is the stable subspace spanned by the remaining eigenvectors.

Oblique projection on the Grassman manifold

In this part we consider again the Grassman manifold, embedded in $\mathcal{M}_{n,n}$ instead of Sym_n :

$$\mathcal{G} = \{UU^T \in \mathcal{M}_{n,n} | U \in \mathcal{M}_{n,p} \text{ and } U^T U = I\}.$$

We have seen (remark 2.17) that its tangent space is

$$\mathcal{T}(R) = \{U\Delta^T + \Delta U^T | \Delta \in \mathcal{M}_{n,p},\, U^T\Delta = 0\}.$$

Applying the methodology explained at the end of section 2.1.3, here is our candidate of oblique projection:

Proposition 2.33. *Consider $\Pi_{\mathscr{M}}$ the application mapping a matrix $\mathfrak{R} = \sum_{i=1}^{n} \lambda_i(\mathfrak{R}) u_i \overline{v}_i^T$ to UU^T the orthogonal projector over the dominant p-dimensional subspace $\mathrm{Span}(U) = \mathrm{Span}(u_i)_{1 \leq i \leq p}$ of $\mathfrak{R}$. Then $\Pi_{\mathscr{M}}$ is an oblique projection on $\mathcal{G}$, and the respective normal space at $R = UU^T \in \mathcal{G}$ is the set of matrices $\mathfrak{R} \in \mathcal{M}_{n,n}$ for which $\mathrm{Span}(U)$ is stable by $\mathfrak{R}$:*

$$\begin{aligned} \mathcal{N}(UU^T) &= \{N \in E | \mathrm{Span}(NU) \subset \mathrm{Span}(U)\} \\ &= \{N \in E | (I - UU^T) N U U^T = 0\}. \end{aligned}$$

Proof. We check the conditions of definition 2.9. The continuity of the eigenvalues of a matrix imply that $\Pi_{\mathscr{M}}$ is well posed on an open neighborhood $\mathcal{V} \subset \mathcal{M}_{n,n}$ containing $\mathcal{G}$. It is clear that $\Pi_{\mathscr{M}}(UU^T) = UU^T$ and that if $N \in \mathcal{V}$ satisfies $\Pi_{\mathscr{M}}(UU^T + N) = UU^T$, the subspace spanned by U must be stable by $N = (N + UU^T) - UU^T$. Reciprocally consider $N \in \mathcal{N}(UU^T)$ and denote $(\lambda_i(N))_{1 \leq i \leq n}$ the eigenvalues of N where the first p are associated with the stable subspace $\mathrm{Span}(U)$. $\mathrm{Span}(U)$ is stable by $N + UU^T$, with eigenvalues $\lambda_i(N) + 1$ for $1 \leq i \leq p$, while $\mathrm{Span}(U)^{\perp}$ is stable by $N^T + UU^T$, with associated eigenvalues $\lambda_j(N)$ for $p+1 \leq j \leq n$. Since $N^T + UU^T$ and $N + UU^T$ share the same eigenvalues, we deduce by continuity that for $N \in \mathcal{N}(R)$ in a neighborhood of 0, $1 + \lambda_i(N) > \lambda_{p+j}(N)$ for $1 \leq i \leq p$ and $1 \leq j \leq n - p$. Hence $\Pi_{\mathscr{M}}(UU^T + N) = UU^T$. □

Proposition 2.34. *The linear projector $\Pi_{\mathcal{T}(UU^T)}$ whose image is the tangent space $\mathcal{T}(UU^T)$ and whose kernel $\mathcal{N}(UU^T)$ is given by:*

$$\begin{array}{rccc} \Pi_{\mathcal{T}(UU^T)} & : & \mathcal{M}_{n,n} & \to & \mathcal{T}(UU^T) \\ & & \mathfrak{X} & \mapsto & (I - UU^T)\mathfrak{X}UU^T + UU^T\mathfrak{X}^T(I - UU^T). \end{array}$$

Or with the coordinate of the horizontal space, $\Delta_{\Pi_{\mathcal{T}(UU^T)}\mathfrak{X}} = (I - UU^T)\mathfrak{X}U$.

Proof. It is clear that $\Pi_{\mathcal{T}(UU^T)} = \Pi_{\mathcal{T}(UU^T)} \circ \Pi_{\mathcal{T}(UU^T)}$ which shows that $\Pi_{\mathcal{T}(UU^T)}$ is a linear projector. One can check then that $\mathrm{Ker}(\Pi_{\mathcal{T}(UU^T)}) = \mathcal{N}(UU^T)$ and $\mathrm{Span}(\Pi_{\mathcal{T}(UU^T)}) \subset \mathcal{T}(UU^T)$. Noticing that $\mathcal{T}(R) \cap \mathcal{N}(R) = \{0\}$, one deduces $\mathrm{Span}(\Pi_{\mathcal{T}(UU^T)}) = \mathcal{T}(R)$. □

Remark 2.24. Notice the difference with the orthogonal projection that assigned $\Delta = (I - UU^T)\mathrm{sym}(\mathfrak{X})U$.

Proposition 2.35. *The Weingarten map in a direction $N \in \mathcal{N}(UU^T)$ for the manifold $\mathcal{G}$ equipped with the map of projectors $UU^T \mapsto \Pi_{\mathcal{T}(UU^T)}$ is given by*

$$\begin{array}{rccc} L_{UU^T}(N) & : & \mathcal{T}(R) & \to & \mathcal{T}(UU^T) \\ & & X & \mapsto & 2 \times \mathrm{sym}((I - UU^T)NXUU^T - XUU^TNUU^T) \end{array}$$

i.e. with the coordinate of the horizontal space, the map $\Delta \mapsto (I - UU^T)N\Delta - \Delta U^T N U$. If $N = \sum_{i=1}^{n} \lambda_i u_i \overline{v}_i^T$ is diagonalizable and $\mathrm{Span}(U) = \mathrm{Span}(u_i)_{1 \leq i \leq p}$, then $L_{UU^T}(N)$ is also diagonalizable and the $p(n-p)$ eigenvalues are given by

$$\kappa_{i(p+j)}(N) = \lambda_{p+j}(N) - \lambda_i(N),\ 1 \leq i \leq p \text{ and } 1 \leq j \leq n - p.$$

A corresponding basis of eigenvectors $\Phi_{ij} \in \mathcal{T}(UU^T)$ is given by

$$\Phi_{i(p+j)} = UU^T \overline{v}_i u_{p+j}^T (I - UU^T) + (I - UU^T) u_{p+j} \overline{v}_i^T UU^T,$$

associated with the dual basis of left eigenvectors defined by:

$$\forall X \in \mathcal{T}(R), < \Phi^*_{ip+j}, X >= \overline{v}^T_{p+j} X u_i.$$

Proof. The derivation of the expression of the Weingarten map is analogous to proposition 2.29 and is omitted. The reader is invited to check that the proposed expression for Φ_{ij} yields indeed a basis of eigenvectors of $L_{UU^T}(N)$. To find the dual basis, one considers $X = \sum_{ij} \alpha_{ij} \Phi_{ij} \in \mathcal{T}(R)$ and checks that $\alpha_{ij} = \overline{v}^T_j X u_i$ as claimed. □

Corollary 2.5. *Let $\mathfrak{R}(t) \in \mathcal{M}_{n,n}$ a time dependent matrix and denote $\lambda_i(t)$ its eigenvalues. Assume that $\Pi_{\mathscr{M}}(\mathfrak{R}(t)) = U(t)U(t)^T$ is defined for all times, that is $\Re(\lambda_p(t)) > \Re(\lambda_{p+1}(t))$. Then a dynamical system for $U(t)$ such that $\mathrm{Span}(U(t))$ is the dominant subspace of $\mathfrak{R}(t)$ at all times is*

$$\dot{U} = \sum_{\substack{1 \le i \le p \\ 1 \le j \le n-p}} \frac{1}{\lambda_i - \lambda_{p+j}} \left[\overline{v}^T_{p+j} \dot{\mathfrak{R}} u_i \right] (I - UU^T) u_{p+j} \overline{v}^T_i U, \tag{2.54}$$

where $(u_i)_{1 \le i \le n}$ and $(v_i)_{1 \le i \le n}$ are the right and left eigenvectors of $\mathfrak{R}(t) - U(t)U(t)^T$, associated with the eigenvalues $\lambda_i(t) - 1$ for $1 \le i \le p$ and $\lambda_{p+j}(t)$ for $1 \le j \le n-p$.

Proof. theorem 2.2 ensures the existence of a differentiable trajectory $u(t)u^T(t)$ such that $u(t)$ is stable by $\mathfrak{R}(t)$ and $u(0) = U(0)$. The continuity of eigenvalues imply $u(t)u(t)^T = U(t)U(t)^T = \Pi_{\mathcal{G}}(\mathfrak{R})$. Formula (2.54) follows identically as in corollary 2.4. □

Remark 2.25. Note that $(u_i)_{1 \le i \le p}$ and $(v_{p+j})_{1 \le j \le n-p}$ in (2.54) are also right and left eigenvectors of $\mathfrak{R}$, but not $(u_{p+j})_{1 \le j \le n-p}$ and $(v_i)_{1 \le i \le p}$.

We know already from the result of proposition 2.8 that $\mathrm{d}(UU^T)/\mathrm{d}t = \Pi_{\mathcal{T}(UU^T)}(\mathfrak{R} - UU^T)$ (eqn. (2.24)) is a dynamical system for which $\Pi_{\mathcal{G}}(\mathfrak{R})$ is a stable equilibrium point provided $\mathfrak{R}$ is sufficiently close to $\mathcal{G}$. Examining the eigenvalues of the Weingarten map, we recover Theorem 2.3 of [13].

Corollary 2.6. *Let $\mathfrak{R} \in \mathcal{M}_{n,n}$ such that $\Pi_{\mathcal{G}}(\mathfrak{R})$ is defined. Then $\Pi_{\mathcal{G}}(\mathfrak{R})$ is the unique stable equilibrium point of the dynamical system defined on $\mathcal{G}$ by*

$$\dot{U} = (I - UU^T)\mathfrak{R}U.$$

Proof. This dynamical system coincides with (2.24) since $(I - UU^T)(\mathfrak{R} - UU^T)U = (I - UU^T)\mathfrak{R}U$. Equilibrium points UU^T are those for which $N = \mathfrak{R} - UU^T \in \mathcal{N}(R)$, *i.e.* those such that U spans a subspace of $\mathfrak{R}$ formed by p eigenvectors. Denote $(\lambda_i)_{1 \le i \le p}$ the corresponding eigenvalues and $(\lambda_{p+j})_{1 \le j \le n-p}$ the remaining one. Then the eigenvalues of the Weingarten map are $\kappa_{i(p+j)}(N) = \lambda_{p+j} - (\lambda_i - 1)$. The stability condition $\Re(\kappa_{i(p+j)}(N)) < 1$ can be satisfied for all eigenvalues only if $UU^T = \Pi_{\mathcal{G}}(\mathfrak{R})$. □

Oblique projection on the "bi-Grassman" manifold

In this part we get interested in the derivative of the application mapping a matrix (possibly non symmetric) to the projector whose image is the dominant subspace and kernel the complement stable subspace.

Definition 2.15. We denote $\mathscr{M}$ the set of *rank-p* linear projectors of $\mathcal{M}_{n,n}$:

$$\begin{aligned}\mathscr{M} &= \{R \in \mathcal{M}_{n,n} | R^2 = R \text{ and } \text{rank}(R) = p\} \\ &= \{UV^T | U \in \mathcal{M}_{n,p}, V \in \mathcal{M}_{n,p},\, V^T U = I\}.\end{aligned}$$

Proof. We prove the ensemble equality. It is clear that with these notations, UV^T is a projector of rank p, because $UV^T UV^T = UV^T$ and $\text{Tr}(UV^T) = \text{Tr}(V^T U) = I$. Now, if R is a projector of rank p, denote P an invertible matrix diagonalizing R. Then U is obtained from the first p columns of P and V is obtained from the first p columns of P^{-T}. □

Remark 2.26. UV^T is the unique projector whose image is $\text{Span}(UV^T) = \text{Span}(U)$ and whose kernel is $\text{Ker}(UV^T) = \text{Span}(V)^\perp$.

A tangent vector $X \in \mathcal{T}(UV^T)$ has the form $X = X_U V^T + U X_V^T$ where X_U and X_V can be understood as the time derivatives of the matrices U and V. Similarly as with the grassman manifold, a "DO-condition" appears to uniquely parameterize the tangent spaces of $\mathscr{M}$:

Proposition 2.36. *The tangent space of $\mathscr{M}$ is*

$$\begin{aligned}\mathcal{T}(UV^T) &= \{X_U V^T + U X_V^T | X_U \in \mathcal{M}_{n,p},\, X_V \in \mathcal{M}_{n,p}\,, V^T X_U + U^T X_V = 0\} \\ &= \{X_U V^T + U X_V^T | X_U \in \mathcal{M}_{n,p},\, X_V \in \mathcal{M}_{n,p}\,, V^T X_U = U^T X_V = 0\}\end{aligned}$$

We refer to $\mathcal{H}_{UV^T} = \{(X_U, X_V) \in \mathcal{M}_{n,p} \times \mathcal{M}_{n,p} | X_U^T V = X_V^T U = 0\}$ as the horizontal space *at the point $R = UV^T$ and the map $(X_U, X_V) \mapsto X_U V^T + U X_V^T$ from $\mathcal{H}_{UV^T}$ to $\mathcal{T}(UV^T)$ is an isomorphism. Hence $\mathscr{M}$ is a smooth manifold of dimension $2p(n-p)$.*

Proof. We first prove the inclusion $\subset$, the inclusion $\supset$ being obvious. Consider two given X_U and X_V and write

$$\begin{cases} X_U = (I - UV^T)X_U + UV^T X_U \\ X_V = (I - VU^T)X_V + VU^T X_V \end{cases}$$

Denote $\Omega = V^T X_U$, the condition $V^T X_U + U^T X_V = 0$ is equivalent to write

$$\begin{cases} X_U = (I - UV^T)X_U + U\Omega \\ X_V = (I - VU^T)X_V - V\Omega^T \end{cases}$$

Denote now $X_U' = (I - UV^T)X_U$ and $X_V' = (I - VU^T)X_V$. Then one has $V^T X_U' = U^T X_V' = 0$ and $X_U V^T + U X_V^T = X_U' V^T + U(X_V')^T$ showing the ensemble equality. In addition, if $X = X_U V^T + U X_V^T$ with $U^T X_V = V^T X_U = 0$ then one can obtain $X_U = XU$ and $X_V = X^T V$ showing the uniqueness of the tangent space parameterization. □

Proposition 2.37. *Consider $\Pi_{\mathscr{M}}$ the application mapping a matrix $\mathfrak{R} \in \mathcal{M}_{n,n}$ to the linear projector whose image is the dominant subspace of $\mathfrak{R}$ and kernel the complementary stable subspace spanned by the remaining eigenvectors:*

$$\begin{array}{rccc} \Pi_{\mathscr{M}} & : & E & \to & \mathscr{M} \\ & & \mathfrak{R} = \sum_{i=1}^n \lambda_i(\mathfrak{R}) u_i \overline{v}_i^T & \mapsto & \Pi_{\mathscr{M}}(\mathfrak{R}) = \sum_{i=1}^n u_i \overline{v}_i^T. \end{array}$$

Then $\Pi_{\mathscr{M}}$ is an oblique projection on $\mathscr{M}$, and the respective normal space $\mathcal{N}(UV^T)$ at $R = UV^T \in \mathscr{M}$ is the set of matrices $\mathfrak{R} \in \mathcal{M}_{n,n}$ for which both $\text{Span}(U)$ *and* $\text{Span}(V)^\perp$

are stable by $\mathfrak{R}$*:*

$$\begin{aligned}\mathcal{N}(UV^T) &= \{N \in \mathcal{M}_{n,n} | \text{Span}(NU) \subset \text{Span}(U) \text{ and } N[\text{Span}(V)^\perp] \subset \text{Span}(V)^\perp\} \\ &= \{N \in \mathcal{M}_{n,n} | N = (I - UV^T)N(I - UV^T) + UV^T NUV^T\}.\end{aligned}$$

Proof. The proof is identical to the one of proposition 2.33. □

Proposition 2.38. *The linear projector* $\Pi_{\mathcal{T}(UV^T)}$ *whose image is the tangent space* $\mathcal{T}(UV^T)$ *and whose kernel is* $\mathcal{N}(UV^T)$ *is given by:*

$$\begin{array}{rccc}\Pi_{\mathcal{T}(UV^T)} & : & \mathcal{M}_{n,n} & \to & \mathcal{T}(UV^T) \\ & & \mathfrak{X} & \mapsto & (I - UV^T)\mathfrak{X}UV^T + UV^T\mathfrak{X}(I - UV^T),\end{array}$$

or with the coordinates of the horizontal space, $X_U = (I - UV^T)\mathfrak{X}U$ *and* $X_V = (I - VU^T)\mathfrak{X}^T V$*.*

Proof. The proof is analogous to the one of proposition 2.34 and is left to the reader. □

Proposition 2.39. *The Weingarten map* $L_{UV^T}(N)$ *with respect to a normal vector* $N \in \mathcal{N}(UV^T)$ *is given by*

$$\mathrm{D}\Pi_{\mathcal{T}(UV^T)}(X) \cdot N = NXUV^T + UV^T XN - XUV^T NUV^T - UV^T NUV^T X,$$

or with the coordinates of the horizontal space ;

$$\begin{array}{rccc}L_R(N) & : & \mathcal{H}_{UV^T} & \to & \mathcal{H}_{UV^T} \\ & & (X_U, X_V) & \mapsto & (NX_U - X_U V^T NU, N^T X_V - X_V U^T N^T V).\end{array}$$

Denote $N = \sum_{i=1}^n \lambda_i(N) u_i \overline{v}_i^T$ *the eigendecomposition of* N *in* $\mathbb{C}$ *with* $UV^T = \sum_{i=1}^p u_i \overline{v}_i^T$*. The eigenvalues of* $L_R(N)$ *are the* $n(n-p)$ *reals*

$$\kappa_{ij} = \lambda_j(N) - \lambda_i(N) \, \forall 1 \leq i \leq p, \, p+1 \leq j \leq n.$$

For each eigenvalues one can find two independent eigenvectors:

$$\Phi^*_{ij,U} = u_i \overline{v}_j^T, \, \Phi_{ij,V} = u_j \overline{v}_i^T,$$

with respective dual forms

$$< \Phi^*_{ij,U}, X >= \overline{v}_i^T X u_j, < \Phi^*_{ij,V}, X >= \overline{v}_j^T X u_i.$$

Proof. The derivation of the Weingarten map is identical to the one of proposition 2.29 and is omitted. The dual basis is obtained by writing $X = \sum_{ij} \alpha_{ij,U}\Phi_{ij,U} + \alpha_{ij,V}\Phi_{ij,V}$ for a given $X \in \mathcal{T}(R)$ and extracting $\alpha_{ij,U}$ and $\alpha_{ij,V}$ by right and left multiplications by u_i and $\overline{v}_i^T$. □

Corollary 2.7. *The oblique projection* $\Pi_{\mathscr{M}}$ *is differentiable and its differential is given by*

$$\mathrm{D}_{\mathfrak{X}}\Pi_{\mathscr{M}}(\mathfrak{R}) = \sum_{\substack{1 \leq i \leq p \\ 1 \leq j \leq n-p}} \frac{1}{\lambda_i(\mathfrak{R}) - \lambda_{p+j}(\mathfrak{R})} \left[(\overline{v}_i^T \mathfrak{X} u_{p+j}) u_i \overline{v}_{p+j}^T + (\overline{v}_{p+j}^T \mathfrak{X} u_i) u_{p+j} \overline{v}_i^T \right].$$

In other words, if $\mathfrak{R}(t) \in \mathcal{M}_{n,n}$ *is a time dependent matrix whose eigenvalue decomposition is given by* $\mathfrak{R}(t) = \sum_{i=1}^{n} \lambda_i(t) u_i \overline{v}_i^T$ *(where the eigenvalues have been ordered according to their real parts) and if* $\Pi_{\mathscr{M}}(\mathfrak{R}(t))$ *is defined for all times, or in other words,* $\Re(\lambda_p(t)) > \Re(\lambda_{p+1}(t))$*, then a dynamical system tracking the dominant stable subspace* $\mathrm{Span}(U) = \mathrm{Span}(u_i)_{1\leq i\leq p}$ *and its stable complementary* $\mathrm{Span}(V)^{\perp} = \mathrm{Span}(u_{p+j})_{1\leq j\leq n-p}$ *is*

$$\begin{cases} \dot{U} = \displaystyle\sum_{\substack{1\leq i\leq p \\ 1\leq j\leq n-p}} \frac{1}{\lambda_i(t) - \lambda_{p+j}(t)} (\overline{v}_{p+j}^T \dot{\mathfrak{R}} u_i) u_{p+j} \overline{v}_i^T U \\ \dot{V} = \displaystyle\sum_{\substack{1\leq i\leq p \\ 1\leq j\leq n-p}} \frac{1}{\lambda_i(t) - \lambda_{p+j}(t)} (\overline{v}_i^T \dot{\mathfrak{R}} u_{p+j}) \overline{v}_{p+j} u_i^T V. \end{cases}$$

Finally, we explicit as well the dynamical system that allows to compute the projection map $\Pi_{\mathscr{M}}$.

Corollary 2.8. *If* $\mathfrak{R} = \sum_{i=1}^{n} \lambda_i(\mathfrak{R}) u_i \overline{v}_i^T$ *is a real matrix diagonalizable in* $\mathbb{C}$ *and such that* $\Re(\lambda_p(\mathfrak{R})) > \Re(\lambda_{p+1}(\mathfrak{R}))$*, then* $UV^T = \sum_{i=1}^{p} u_i \overline{v}_i^T = \Pi_{\mathscr{M}}(\mathfrak{R})$ *is the unique asymptotically stable equilibrium point of the dynamical system*

$$\begin{cases} \dot{U} = (I - UV^T)\mathfrak{R}U \\ \dot{V} = (I - VU^T)\mathfrak{R}^T V. \end{cases}$$

Proof. The proof follows again immediately by noticing that $N = \mathfrak{R} - R$ with $R = \Pi_{\mathscr{M}}(\mathfrak{R})$ is the unique normal vector N and point R such that $\mathfrak{R} - R$ is a normal vector and the real parts of the eigenvalues of $L_R(\mathfrak{R})$ are strictly lower than one. □

2.3 Projected dynamical systems and dynamic approximation

In this part we apply the results of section 2.1.2 to analyze projected dynamical systems onto smooth manifolds. The results presented here after will be readily applicable to the fixed rank manifold, for which the corresponding projected dynamical system will be directly related to the Dynamically Orthogonal approximation, which will be the focus of the next chapter.

Consider a dynamical system associated with a time dependent vector field $\mathcal{L}(t, \cdot) \in E$ in the full Euclidean space E:

$$\dot{\mathfrak{R}} = \mathcal{L}(t, \mathfrak{R}), \tag{2.55}$$

If we know that the trajectory $\mathfrak{R}(t)$ lies approximately on some manifold $\mathscr{M}$ one can seek to find an approximation of $\mathfrak{R}(t)$ obtained by evolving an approximation $R(t)$ directly on $\mathscr{M}$, *i.e.* one look for a dynamical system of the form

$$\dot{R} = L(t, R) \ \in \mathcal{T}(R), \tag{2.56}$$

where $L(t, \cdot)$ is a tangent vector field on $\mathscr{M}$. Such approximation can naturally be obtained by replacing the vector field $\mathcal{L}(t, \cdot)$ of the ambient space with its tangent projection everywhere on $\mathscr{M}$: this is the idea of "combing" the hair formed by the vector field $\mathcal{L}(t, \cdot)$ in E onto the manifold $\mathscr{M}$ as illustrated in the introduction on Figure 2-1.

Definition 2.16. The *projected* dynamical system on $\mathscr{M}$,

$$\begin{cases} \dot{R} &= \Pi_{\mathcal{T}(R)}(\mathcal{L}(t,R)) \\ R(0) &= \Pi_{\mathscr{M}}(\mathfrak{R}(0)), \end{cases} \tag{2.57}$$

is called the *dynamic* approximation of (2.55), and $R(t)$ is called the *reduced* solution.

Remark 2.27. For model order reduction applications as motivated in the introduction, $E = \mathcal{M}_{l,m}$ is the space of all discretized solutions of (4), and $\mathscr{M}$ is the fixed rank manifold, over which a rank r approximation by evolving modes and coefficient separately. This will be the focus of chapter 3.

For the approximation (2.57) to be perfect, the reduced solution $R(t)$ should coincide at all times with the projection $\Pi_{\mathscr{M}}(\mathfrak{R}(t))$ of the original solution. Nevertheless, $\Pi_{\mathscr{M}}(\mathfrak{R}(t))$ is not the solution of a reduced system of the form (2.56) since its time derivative depends directly on the value of $\mathfrak{R}(t)$ in the full space E: replacing $\mathfrak{X}$ by $\dot{\mathfrak{R}} = \mathcal{L}(t,\mathfrak{R})$ in (2.5) yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\Pi_{\mathscr{M}}(\mathfrak{R}(t)) = \Pi_{\mathcal{T}(\Pi_{\mathscr{M}}(\mathfrak{R}))}(\mathcal{L}(t,\mathfrak{R})) + \mathrm{D}\Pi_{\mathcal{T}(\Pi_{\mathscr{M}}(\mathfrak{R}))}\left(\frac{\mathrm{d}}{\mathrm{d}t}\Pi_{\mathscr{M}}(\mathfrak{R}(t))\right)\cdot(\mathfrak{R} - \Pi_{\mathscr{M}}(\mathfrak{R})). \tag{2.58}$$

One therefore sees from this equation that the projected dynamical system (2.57) is obtained by replacing $\mathfrak{R}(t)$ in (2.58) with its approximation $R(t) \in \mathscr{M}$, for which $R = \Pi_{\mathscr{M}}(R)$ makes the curvature term $\mathrm{D}\Pi_{\mathcal{T}(\Pi_{\mathscr{M}}(R))}\left(\frac{\mathrm{d}}{\mathrm{d}t}\Pi_{\mathscr{M}}(R)\right)\cdot(R - \Pi_{\mathscr{M}}R))$ vanish. Accounting for this term in all generality would require external information to estimate the neglected normal component $\mathfrak{R} - \Pi_{\mathscr{M}}(\mathfrak{R})$.

The above is related to a "computational" interpretation of the dynamic approximation (2.57). Consider a time integration of the dynamical system (2.55) over (t^n, t^{n+1}),

$$\mathfrak{R}^{n+1} = \mathfrak{R}^n + \Delta t\overline{\mathcal{L}}(t^n, \mathfrak{R}^n, \Delta t), \tag{2.59}$$

where $\overline{\mathcal{L}}(t,\mathfrak{R},\Delta t)$ denotes the full-space integral $\overline{\mathcal{L}}(t,\mathfrak{R},\Delta t) = \frac{1}{\Delta t}\int_t^{t+\Delta t}\mathcal{L}(s,\mathfrak{R}(s))\mathrm{d}s$ for the exact integration or the increment function [62] for a numerical integration. Examples of the latter include $\overline{\mathcal{L}}(t,\mathfrak{R},\Delta t) = \mathcal{L}(t,\mathfrak{R})$ for forward Euler and $\overline{\mathcal{L}}(t,\mathfrak{R},\Delta t) = \mathcal{L}(t+\Delta t/2, \mathfrak{R} + \Delta t/2\,\mathcal{L}(t,\mathfrak{R}))$ for a second-order Runge-Kutta scheme. Assume that the solution $\mathfrak{R}^n$ at time t^n is well approximated by an estimation $R^n \in \mathscr{M}$. A natural way to obtain an estimation of $\Pi_{\mathscr{M}}(\mathfrak{R}^{n+1})$ at the next time step is to set

$$\begin{cases} R^{n+1} = \Pi_{\mathscr{M}}(R^n + \Delta t\overline{\mathcal{L}}(t, R^n, \Delta t)) \\ R^0 = \Pi_{\mathscr{M}}(\mathfrak{R}(0)). \end{cases} \tag{2.60}$$

It turns out that (2.60) can be seen as a discretization of the dynamic approximation (2.57). One finds that, for any point $R \in \mathscr{M}$ on the manifold,

$$\frac{\Pi_{\mathscr{M}}(R + \Delta t\overline{\mathcal{L}}(t,R,\Delta t)) - R}{\Delta t} \underset{\Delta t\to 0}{\longrightarrow} \Pi_{\mathcal{T}(R)}(\mathcal{L}(t,R)) \tag{2.61}$$

holds true since the curvature term vanishes in (2.58), and $\overline{\mathcal{L}}(t,R,0) = \mathcal{L}(t,R)$ by consistency of the time marching with the exact integration (2.59) [62]. This implies, under sufficient regularity condition on $\mathcal{L}$, that the continuous limit of the scheme (2.60) is the projected dynamical system (2.57).

Theorem 2.5. *Assume that the reduced solution (2.57) is defined on a time interval $[0,T]$. Consider N_T time steps $\Delta t = T/N_T$, denote $t^n = n\Delta t$ and consider R^n the sequence obtained from the scheme (2.60). Assume that $\mathcal{L}$ is Lipschitz continuous, that is there exists a constant K such that*

$$\forall t \in [0,T],\ \forall \mathfrak{R}_1, \mathfrak{R}_2 \in E,\ ||\mathcal{L}(t,\mathfrak{R}_1) - \mathcal{L}(t,\mathfrak{R}_2)|| \leq K||\mathfrak{R}_1 - \mathfrak{R}_2||. \tag{2.62}$$

Then the sequence R^n converges uniformly to the reduced solution $R(t)$ in the following sense:

$$\sup_{0\leq n\leq N_T} ||R^n - R(t^n)|| \underset{\Delta t\to 0}{\longrightarrow} 0.$$

Proof. It is sufficient to check that the scheme (2.60) is both consistent and stable (see [62]). Denote Φ the increment function of the scheme (2.60):

$$\Phi(t,R,\Delta t) = \frac{\Pi_{\mathscr{M}}(R + \Delta t\overline{\mathcal{L}}(t,R,\Delta t)) - R}{\Delta t} = \frac{1}{\Delta t}\int_0^1 \frac{\mathrm{d}}{\mathrm{d}\tau}\Pi_{\mathscr{M}}(g(R,t,\tau,\Delta t))\mathrm{d}\tau \tag{2.63}$$

with $g(R,t,\tau,\Delta t) = R + \tau\Delta t\overline{\mathcal{L}}(t,R,\Delta t)$. Consider a compact neighborhood $\mathcal{U}$ of E containing the trajectory $R(t)$ on the interval $[0,T]$ and sufficiently thin such that $\mathcal{U}$ does not intersect the set (open, see [32]) where $\Pi_{\mathscr{M}}$ is not differentiable. On that set, $\Pi_{\mathscr{M}}$ is Lipschitz continuous. The consistency of (2.60) and continuity of Φ on $[0,T]\times\mathcal{U}\times\mathbb{R}$ follows from (2.61). For usual time marching schemes (e.g. Runge Kutta), the Lipschitz condition (2.62) also holds for the map $R \mapsto \overline{\mathcal{L}}(t,R,\Delta t)$. Therefore Φ is also Lipschitz continuous with respect to R on $\mathcal{U}$ by composition. This is a sufficient stability condition. □

Remark 2.28. The same results hold true if one uses

$$R^{n+1} = \Pi_{\mathscr{M}}(R^n + \Delta t\Pi_{\mathcal{T}(R^n)}\overline{\mathcal{L}}(t,R,\Delta t))$$

instead of (2.60), since (2.61) still holds[3]. This can be of interest in practice if the projection operator $\Pi_{\mathscr{M}}$ is more easily computed in that fashion (this is the case for the fixed rank manifold, as explained later on in algorithm 3).

Using theorem 2.4, a bound for the growth of the error committed by the reduced solution is obtained. In order to state the result, we need to introduce a notation to quantify the Lipschitz behavior of the projection onto tangent spaces $R \mapsto \Pi_{\mathcal{T}(R)}$:

Definition 2.17. For any point $R \in \mathscr{M}$ there exists a constant ρ such that

$$\forall R' \in \mathscr{M},\ ||\Pi_{\mathcal{T}(R)} - \Pi_{\mathcal{T}(R')}|| \leq \rho||R - R'||, \tag{2.64}$$

where the norm of the left hand-side is the operator norm relative to the euclidean norm $||\,||$ on E. We denote $\rho_\infty(R)$ the minimum constant ρ such that the inequality above is satisfied which we refer to as the "local" Lipschitz constant of $\Pi_{\mathcal{T}}$ at R.

Proof. Consider a neighborhood $\mathcal{V}$ of 0 in $\mathcal{T}(R)$ and $\mathcal{U}$ a neighborhood of R in $\mathscr{M}$ for which the exponential map $\exp_R$ is a local diffeomorphism from $\mathcal{V}$ onto $\mathcal{U}$ (see [139]). On that neighborhood, the map $R \mapsto \Pi_{\mathcal{T}(R)}$ is smooth, hence there exists a constant A such that

$$\forall X \in \mathcal{V},\ ||\Pi_{\mathcal{T}(R)} - \Pi_{\mathcal{T}(\exp_R(X))}|| \leq A||X||$$

[3] We thank P.A. Absil for this remark.

Since $\exp_R$ is a local diffeomorphism and has hence a smooth inverse, there exists also a constant B such that $||X|| = ||X - 0|| \leq B||\exp_R(X) - R||$. Hence we have shown that there exists $\boldsymbol{\eta}$ such that

$$\forall R' \in \mathscr{M},\, ||R - R'|| \leq \boldsymbol{\eta} \Rightarrow ||\Pi_{\mathcal{T}(R)} - \Pi_{\mathcal{T}(R')}|| \leq AB||R - R'||.$$

Now it is a well known fact that for two projectors $\Pi_{\mathcal{T}(R)}$ and $\Pi_{\mathcal{T}(R')}$, $||\Pi_{\mathcal{T}(R)} - \Pi_{\mathcal{T}(R')}|| \leq 1$ (see e.g. Theorem 2.6.1 in [58]). Therefore (2.64) holds with $\rho = \min(\frac{1}{\boldsymbol{\eta}}, AB)$. □

Remark 2.29. The local Lipschitz constant $\rho_\infty(R)$ on the fixed rank manifold is bounded by $2/\sigma_r(R)$, see lemma 3.1 in chapter 3.

Theorem 2.6. *Consider $\mathfrak{R}(t) \in E$ the solution of the original dynamical system (2.55). Assume that the following conditions hold on a time interval $[0, T]$:*

1. *$\mathcal{L}$ is Lipschitz continuous,* i.e. *equation (2.62) holds.*

2. *The original solution $\mathfrak{R}(t)$ stays close to the manifold $\mathscr{M}$, in the sense that $\mathfrak{R}(t)$ remains in a domain where $\Pi_{\mathscr{M}}$ is differentiable. In particular, $\mathfrak{R}(t)$ does not cross the skeleton of $\mathscr{M}$ on $[0, T]$ and*

$$\forall t \in [0, T],\ \max_i \kappa_i(\mathfrak{R}(t) - \Pi_{\mathscr{M}}(\mathfrak{R}(t))) < 1,$$

   where the κ_i are the curvatures introduced in definition 2.7.

Then, the error of the reduced approximation $R(t)$ (eqn. (2.57)) remains controlled by the best approximation error $||\mathfrak{R} - \Pi_{\mathscr{M}}(\mathfrak{R}(t))||$ on $[0, T]$:

$$\forall t \in [0, T],\ ||R(t) - \Pi_{\mathscr{M}}(\mathfrak{R}(t))|| \leq \int_0^t ||\mathfrak{R}(s) - \Pi_{\mathscr{M}}(\mathfrak{R}(s))|| \left(K + \frac{\kappa_\infty(\Pi_{\mathscr{M}}(\mathfrak{R}(t)))||\mathcal{L}(s, \mathfrak{R}(s))||}{1 - \max_i \kappa_i(\mathfrak{R}(s) - \Pi_{\mathscr{M}}(\mathfrak{R}(s)))} \right) e^{\boldsymbol{\eta}(t-s)} \mathrm{d}s, \quad (2.65)$$

where κ_∞ is the maximal curvature defined at lemma 2.2, $\boldsymbol{\eta}$ is the constant

$$\boldsymbol{\eta} = K + \sup_{t \in [0,T]} \left(||\mathcal{L}(t, \mathfrak{R}(t))||\, \rho_\infty(\Pi_{\mathscr{M}}(\mathfrak{R}(t))) \right) \quad (2.66)$$

and $\rho_\infty(\Pi_{\mathscr{M}}(\mathfrak{R}(t)))$ the local Lipschitz constant of $\Pi_{\mathcal{T}}$ at $\Pi_{\mathscr{M}}(\mathfrak{R}(t))$.

Proof. The proof compares the derivative of the best approximation $\Pi_{\mathscr{M}}(\mathfrak{R}(t))$, obtained explicitly with the formula (2.12), to the derivative $\dot{R}$ of the dynamic approximation (2.57), before applying Gronwall's lemma. Denote $R^*(t) = \Pi_{\mathscr{M}}(\mathfrak{R}(t))$ and $N(t) = \mathfrak{R}(t) - \Pi_{\mathscr{M}}(\mathfrak{R}(t))$. Since $\dot{R}^*(t) = \mathrm{D}_{\dot{\mathfrak{R}}}\Pi_{\mathscr{M}}(\mathfrak{R}(t))$, bounding (2.5) and using lemma 2.1 yields:

$$\begin{aligned} ||\dot{R} - \dot{R}^*|| &\leq ||\Pi_{\mathcal{T}(R^*)}(\mathcal{L}(t, \mathfrak{R})) - \Pi_{\mathcal{T}(R)}(\mathcal{L}(t, R))|| + \max_i \frac{\kappa_i(N(t))}{1 - \kappa_i(N(t))} ||\mathcal{L}(t, \mathfrak{R})|| \\ &\leq ||\Pi_{\mathcal{T}(R^*)}(\mathcal{L}(t, \mathfrak{R})) - \Pi_{\mathcal{T}(R)}(\mathcal{L}(t, R))|| + \frac{\kappa_\infty(R^*)||N(t)||}{1 - \max_i \kappa_i(N(t))} ||\mathcal{L}(t, \mathfrak{R})||. \end{aligned}$$

Furthermore,

$$\begin{aligned}||\Pi_{\mathcal{T}(R^*)}(\mathcal{L}(t,\mathfrak{R})) - \Pi_{\mathcal{T}(R)}(\mathcal{L}(t,R))|| \leq\ & ||\Pi_{\mathcal{T}(R^*)}(\mathcal{L}(t,\mathfrak{R})) - \Pi_{\mathcal{T}(R)}(\mathcal{L}(t,\mathfrak{R}))|| \\ &+ ||\Pi_{\mathcal{T}(R)}(\mathcal{L}(t,\mathfrak{R})) - \Pi_{\mathcal{T}(R)}(\mathcal{L}(t,R^*))|| \\ &+ ||\Pi_{\mathcal{T}(R)}(\mathcal{L}(t,R^*)) - \Pi_{\mathcal{T}(R)}(\mathcal{L}(t,R))||.\end{aligned}$$

The Lipschitz continuity of $\mathcal{L}$ and definition 2.17 together imply

$$\begin{aligned}||\Pi_{\mathcal{T}(R^*)}(\mathcal{L}(t,\mathfrak{R})) - \Pi_{\mathcal{T}(R)}(\mathcal{L}(t,\mathfrak{R}))|| &\leq \rho_\infty(R^*)||R-R^*||\,||\mathcal{L}(t,\mathfrak{R})||, \\ ||\Pi_{\mathcal{T}(R)}(\mathcal{L}(t,\mathfrak{R})) - \Pi_{\mathcal{T}(R)}(\mathcal{L}(t,R^*))|| &\leq K||\mathfrak{R}-R^*||, \\ ||\Pi_{\mathcal{T}(R)}(\mathcal{L}(t,R^*)) - \Pi_{\mathcal{T}(R)}(\mathcal{L}(t,R))|| &\leq K||R-R^*||.\end{aligned}$$

Finally, the following inequality is derived, combining all above equations together:

$$||\dot{R}-\dot{R}^*|| \leq (K+\rho_\infty(R^*)||\mathcal{L}(t,\mathfrak{R})||)\,||R-R^*|| + \left(K + \frac{\kappa_\infty(R^*)||\mathcal{L}(t,\mathfrak{R})||}{1-\max_i \kappa_i(N(t))}\right)||\mathfrak{R}-R^*||.$$

An application of Gronwall's Lemma (see corollary 4.3. in [72]) yields (2.65). □

Remark 2.30. This statement improves the result expressed in Theorem 5.1 of [83] (in the case of $\mathscr{M}$ being the fixed rand manifold), since no assumption is made on the smallness of the best approximation error $||\mathfrak{R}-\Pi_{\mathscr{M}}(\mathfrak{R})||$, nor on the boundedness of $||R-\Pi_{\mathscr{M}}(\mathfrak{R})||$. Note also that in the case one has access to the original derivative $\mathcal{L}(t,\mathfrak{R}(t))$ and consider $\dot{R} = \Pi_{\mathcal{T}(R)}(\mathcal{L}(t,\mathfrak{R}(t)))$ instead of (2.57) (this is actually the framework considered in [83]) then the bounds above hold with $K=0$.

Theorem 3.1 highlights two sufficient conditions for the error committed by the DO approximation to remain small :

Condition 1 The discrete operator $\mathcal{L}$ must not be too sensitive to the error $\mathfrak{R}(t)-R(t)$, namely the Lipschitz constant K must be small. This error is commonly encountered by any approximation made for evaluating the operator of a dynamical system (as a consequence of Gronwall's lemma [72]). If the Lipschitz constant is too big then one can expect (2.57) not to be a really good approximation.

Condition 2 Independently of the choice of the reduced order model, the solution of the initial system (2.55), $\mathfrak{R}(t)$, must remain close to the manifold $\mathscr{M}$, or in other words, must remain far from the skeleton $\mathrm{Sk}(\mathscr{M})$ of $\mathscr{M}$. As visible on Figure 2-2, the best rank r approximation $\Pi_{\mathscr{M}}(\mathfrak{R})$ of $\mathfrak{R}$ exhibits a jump when $\mathfrak{R}$ crosses the skeleton. (*i.e.* when $\sigma_r(\mathfrak{R}) = \sigma_{r+1}(\mathfrak{R})$ occurs for the fixed rank manifold). At that point, the discontinuity of $\Pi_{\mathscr{M}}(\mathfrak{R}(t))$ cannot be tracked or any smooth dynamic approximation of the form (2.56).

Last, it should be noted that the growth rate $\boldsymbol{\eta}$ (equation (2.66)) of the error is related to the local lipschitz constant $\rho_\infty(\Pi_{\mathscr{M}}(\mathfrak{R}))$ of $\Pi_{\mathcal{T}}$. This is related to the fact the tangent projection Π_T in (2.57) is applied at the location of the DO solution $R(t)$ instead of the one of the best approximation $\Pi_{\mathscr{M}}(\mathfrak{R})$. If $R(t)$ and $\Pi_{\mathscr{M}}(\mathfrak{R}(t))$ are too far from one another, these tangent spaces may be oriented very differently because of the curvature of $\mathscr{M}$.

Chapter 3

Efficient simulation of stochastic advection and Lagrangian transport

3.1 Introduction

A typical challenge encountered in environmental Lagrangian flow predictions is the need for dealing with velocity data that include a certain level of uncertainty, resulting from sparse acquisitions, noise in direct measurements, or errors in the inferred numerical predictions [94]. Uncertainty is modeled by "adding" randomness to the velocity field, each realization $\boldsymbol{v}(t, \boldsymbol{x}; \omega)$ corresponding to a particular possible scenario ω. An issue of great interest in hazard predictions [88], is to quantify how this uncertainty reverberates in the Lagrangian motion. A basic Monte-Carlo (MC) approach would then solve either the stochastic ODE

$$\begin{cases} \dot{\boldsymbol{x}} = \boldsymbol{v}(t, \boldsymbol{x}; \omega) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0, \end{cases} \tag{3.1}$$

or the stochastic partial differential equation (SPDE)

$$\begin{cases} \partial_t \boldsymbol{\psi} + \boldsymbol{v}(t, \boldsymbol{x}; \omega) \cdot \nabla \boldsymbol{\psi} = 0 \\ \boldsymbol{\psi}(0, \boldsymbol{x}) = \boldsymbol{x}, \end{cases} \tag{3.2}$$

for a large number of realizations, ω. While performance of as well Monte-Carlo and particle methods [27] can be optimized through parallelism, such methodologies are computationally demanding for cases requiring high resolution in both the spatial and stochastic domains, *i.e.* a large number of particles and realizations. Hence, while they have been useful in a variety of applicationsm particle and MC methods are very expensive for uncertain advection.

A substantial benefit of the PDE formulation (3.2) is its compatibility with dynamical model order reduction that take direct advantage of the spatial structures in the solution. Classic reduced order methods aim to evolve low-rank decompositions such as $\boldsymbol{\psi}(t, \boldsymbol{x}; \omega) \simeq \sum_{i=1}^{r_\Psi} \zeta_i(t; \omega) \boldsymbol{u}_i(\boldsymbol{x})$ or $\boldsymbol{\psi}(t, \boldsymbol{x}; \omega) \simeq \sum_{i=1}^{r_\Psi} \zeta_i(\omega) \boldsymbol{u}_i(t, \boldsymbol{x})$ at a cost much smaller than the direct realization methods [141, 57] by evolving independently a small number r_Ψ of spatial modes, $\boldsymbol{u}_i$, or stochastic coefficients, ζ_i. Challenges remain however because advection tends to create fine features in the solution, with sharp gradients or shocks that evolve in time and

space, and these shocks require to be handled with extreme care by numerical schemes [109, 108, 134, 100]. For stochastic advection, the classic methods ranging from Polynomial Chaos to stochastic Galerkin schemes [110, 77, 37] that either assume *a priori* choices of time-independent modes $\boldsymbol{u}_i(\boldsymbol{x})$, or rely on strong hypotheses on the probability distribution of the coefficients ζ_i, may not be well suited for capturing these fine spatial patterns (because requiring a high number of Fourier modes), or non Gaussian behaviors of the coefficients. In addition, upwinding, total variation diminishing (TVD), or Essentially Non Oscillatory (ENO) rules, must be adapted for reduced-order numerical advection schemes which is challenging [145, 147, 125]. This is in part why most classic stochastic advection attempts have essentially restricted themselves to one dimensional applications [57, 77, 37, 110] or simplified 2D cases that do not exhibit strong shocks [149].

In contrast with classic reduced order methods, the Dynamically Orthogonal (DO) methodology [124, 122] solves equations to simultaneously evolve a time-dependent basis of modes, $\boldsymbol{u}_i(t,\boldsymbol{x})$, and coefficients, $\zeta_i(t;\omega)$,

$$\boldsymbol{\psi}(t,\boldsymbol{x};\omega) \simeq \sum_{i=1}^{r_\Psi} \zeta_i(t;\omega)\boldsymbol{u}_i(t,\boldsymbol{x})\,. \tag{3.3}$$

Such dynamic approaches [91] can efficiently capture the evolving spatial flow features and their variability. Numerical schemes for DO equations have been derived for a variety of dynamics, from stochastic Navier-Stokes [147] to Hamilton-Jacobi [140] equations. The question adapting advection schemes to the DO framework while maintaining consistency of between determistic and stochastic integration has been investigated by Ueckermann [147, 146] and has remained since then an active area of research in the MSEAS group.

The results developped in chapter 2 and provide a rigorous framework to further develop the implementation and the numerical integration of the DO equations, understood as a dynamic approximation over the fixed rank manifold (section 2.2.1). This chapter addresses the implementation of the DO methodology [124] in view of its geometric interpretation, and derives new numerical schemes for the stochastic advection equation (3.2) and Lagrangian transports. We develop a deterministic/stochastic consistent methodology that attempts at minimizing the between the MC and DO approximation "realization by realization", in contrast with an approach that would target at preserving only the statistical properties. As an immediate benefit, an efficient computational methodology for evaluating an ensemble of flow maps $\boldsymbol{\psi}(t,\boldsymbol{x};\omega) = \boldsymbol{\phi}_0^t(\boldsymbol{x};\omega)$ of the ODE (3.1) with random velocity is obtained. The issue of capturing shocks is addressed in this work by considering fully linear but stabilized advection schemes. This allows to apply in a compatible fashion reduced order methods that rely on tensor decompositions of either the solution, $\boldsymbol{\psi}$, or of its time derivative $-\boldsymbol{v}\cdot\nabla\boldsymbol{\psi}$. The schemes presented herein are not restricted to pure transport, they are also applicable to stochastic PDEs that include advection terms of the form $\boldsymbol{v}\cdot\nabla$, such as the Navier Stokes equations.

A synopsis of the coupled DO PDEs as initially introduced by [124] and of the methodology that allows to derive a set of coupled PDEs "in the continuous domain" for the evolution of the tensor decomposition (3.3) is given in section 3.2. Numerical schemes for this set of PDEs are obtained by applying the DO methodology directly onto the spatial discretization of the stochastic transport PDE rather than its continuous version (3.2). In that discrete setting, it is found that the formally derived DO equations coincide with the projected dynamical system defined in definition 2.12 of the previous chapter, with $\mathscr{M}$ being the fixed rank-r manifold. Applying directly the results developped previously, we obtain that (i)

the Dynamically Orthogonal approximation (DO) coincides with the reduced order method that that applies the SVD truncation after every time step Δt with $\Delta t \to 0$ (ii) the approximation error is controlled over large integration times provided the original solution remains close to the low rank manifold $\mathscr{M}$, in the sense that it remains far from the skeleton of $\mathscr{M}$. This geometric condition can be expressed as an explicit dependence of the error on the gaps between singular values of order r and $r+1$. Section 3.3 focuses on the implementation in practice of the DO machinery to solve the stochastic transport PDE (3.2). Factorization properties of the advection operator must be preserved at the discrete level to avoid additional level of approximations. This is ensured through the selection of a fully linear advection scheme, whose accuracy and stability is obtained by the use of high order spatial and temporal discretization combined with linear filtering, a technique popular in ocean engineering [131]. It is explained how stochastic boundary conditions can be accounted for by the model order reduced method in an optimal and convenient manner. Different possible time discretization strategies for the DO approximation are discussed, as well as the issue of modifying dynamically the stochastic dimensionality r_Ψ of the tensor approximation (3.3). It is explained how Riemannian gradient descents and geodesic equations can be integrated to the time stepping to adaptively track the truncated SVD of the stochastic solution, and to account for the curvature of the low-rank manifold. Finally, as a requirement of both the DO method and multi-steps time marching schemes, an efficient method is proposed for preserving the orthonormality of the modal basis $(\boldsymbol{u}_i)$ during the time integration, as well as the smooth evolution of this basis and the coefficients ζ_i. Numerical results of the overall methodology are presented in section 3.4 using the bi-dimensional stochastic analytic double-gyre flow and stochastic flow past a cylinder, both of which include sharp gradients. The DO results are finally contrasted with those of direct Monte-Carlo.

3.2 Model order reduction of the stochastic transport equation with the Dynamically Orthogonal approximation

3.2.1 Mathematical setting for the transport PDE

The stochastic transport PDE (3.2) is set on a smooth bounded domain Ω of $\mathbb{R}^d$ where d denotes the spatial dimension. The flow map $\boldsymbol{\phi}_0^t$ of the ODE (3.1) is defined for all time if particle trajectories don't leave the domain Ω, which is ensured if the normal flux $\boldsymbol{v}\cdot\boldsymbol{n}$ vanishes on the boundary $\partial\Omega$, $\boldsymbol{n}$ denoting the outward normal of Ω. In the following, one deals with the more general case where $\boldsymbol{v}\cdot\boldsymbol{n}$ may have an arbitrary sign on $\partial\Omega$. Inlet and outlet boundaries are denoted respectively

$$\partial\Omega_-(t;\omega) = \{x \in \partial\Omega | \boldsymbol{v}(t,x;\omega)\cdot\boldsymbol{n} < 0\}$$
$$\partial\Omega_+(t;\omega) = \{x \in \partial\Omega | \boldsymbol{v}(t,x;\omega)\cdot\boldsymbol{n} \geq 0\},$$

and several works [30, 16, 17] have shown that the transport eqn. (3.2) is well posed (under suitable regularity assumptions on $\boldsymbol{v}$), provided a Dirichlet boundary condition is prescribed at the inlet $\partial\Omega_-(t;\omega)$. Following Leung [97], we consider the dirichlet boundary condition

$$\boldsymbol{\psi}(t,\boldsymbol{x};\omega) = \boldsymbol{x} \text{ on } \partial\Omega_-(t;\omega), \tag{3.4}$$

which ensures that the solution $\boldsymbol{\psi}(t,\boldsymbol{x};\omega)$ carries the value of the initial entering location of the particle that arrived in $\boldsymbol{x}$ at time t. Theoretically no boundary conditions is required

on the outlet boundary $\partial\Omega_+(t;\omega)$, but those may be needed for convenience in numerical schemes. In numerical applications of section 3.4, similarly as in [97], the Neumann boundary condition was considered:

$$\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{n}}(t,\boldsymbol{x};\omega) = 0 \text{ on } \partial\Omega_+(t;\omega). \tag{3.5}$$

This boundary condition naturally arises when considering $\boldsymbol{\psi}$ as a viscous limits of eqn. (3.2) (see Theorem 4.1 in [17]). For simplicity, it is assumed that a modal decomposition of the stochastic velocity field $\boldsymbol{v}$ is available:

$$\boldsymbol{v}(t,\boldsymbol{x};\omega) = \sum_{k=1}^{r_{\boldsymbol{v}}} \beta_k(t;\omega)\boldsymbol{v}_k(t,\boldsymbol{x}). \tag{3.6}$$

3.2.2 Derivation of DO field equations in the continuous setting

The DO methodology uses equations to evolve adaptively modes $\boldsymbol{u}_i(t,\boldsymbol{x})$ and stochastic coefficients $\zeta_i(t;\omega)$ considered both as time-dependent quantities, so as to most accurately update the modal approximation (3.3). Such equations can formally be found [124] by replacing the solution $\boldsymbol{\psi}$ with its tensor approximation (3.3) in the transport equation (3.2) :

$$(\partial_t\zeta_j)\boldsymbol{u}_j + \zeta_j\partial_t\boldsymbol{u}_j + \zeta_j\beta_k\boldsymbol{v}_k\cdot\nabla\boldsymbol{u}_j = 0, \tag{3.7}$$

where the Einstein summation convention over repeated indexes is used. The family of modes is assumed orthonormal, namely

$$\forall 1\leq i,j\leq r \;\; <\boldsymbol{u}_i,\boldsymbol{u}_j> = \int_\Omega u_i(t,\boldsymbol{x})u_j(t,\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \delta_{ij}, \tag{3.8}$$

where $<,>$ denotes the scalar product on $L^2(\Omega)$. Furthermore, the "dynamically orthogonal condition"

$$\forall 1\leq i,j\leq r, <\partial_t\boldsymbol{u}_i,\boldsymbol{u}_j> = 0 \tag{3.9}$$

is imposed to remove the redundancy in (3.3), coming from the fact that the modal decomposition is invariant under rotations of modes $\boldsymbol{u}_i$ and coefficients ζ_i [124]. Assuming these conditions, equations for the coefficients, ζ_i, are obtained by L^2 projection of eqn. (3.7) onto the modes, $\boldsymbol{u}_i$:

$$\forall 1\leq i\leq r, \partial_t\zeta_i + \zeta_j\beta_k < \boldsymbol{v}_k\cdot\nabla\boldsymbol{u}_j,\boldsymbol{u}_i> = 0. \tag{3.10}$$

Governing equations for the modes, $\boldsymbol{u}_i$, are obtained by L^2 projection in the probability space onto the coefficients: multiplying eqn. (3.7) by ζ_i, replacing $\partial_t\zeta_i$ by using eqn. (3.10), applying the expectation and multiplying by the inverse $(\mathbb{E}[\zeta_i\zeta_j])^{-1}$ of the moment matrix $(\mathbb{E}[\zeta_i\zeta_j])_{1\leq i,j\leq k}$ yields :

$$\partial_t\boldsymbol{u}_i + (\mathbb{E}[\zeta_i\zeta_j])^{-1}\mathbb{E}[\zeta_j\beta_k]\boldsymbol{v}_k\cdot\nabla\boldsymbol{u}_j = (\mathbb{E}[\zeta_i\zeta_j])^{-1}\mathbb{E}[\zeta_j\beta_k] < \boldsymbol{v}_k\cdot\nabla\boldsymbol{u}_j,\boldsymbol{u}_l>\boldsymbol{u}_l. \tag{3.11}$$

Deriving boundary conditions is slightly more delicate as (3.4) and (3.5) involve a stochastic partition $\partial\Omega = \partial\Omega_-(t;\omega)\cup\partial\Omega_+(t;\omega)$ of the boundary. They are obtained by rewriting

equations (3.4) and (3.5) as

$$\sum_{j=1}^{r}\left[\zeta_j \boldsymbol{u}_j \mathbf{1}_{\boldsymbol{v}\cdot\boldsymbol{n}<0} + \zeta_j \frac{\partial \boldsymbol{u}_j}{\partial \boldsymbol{n}} \mathbf{1}_{\boldsymbol{v}\cdot\boldsymbol{n}\geq 0}\right] = \boldsymbol{x}\mathbf{1}_{\boldsymbol{v}\cdot\boldsymbol{n}<0} \text{ on } \partial\Omega,$$

where $\mathbf{1}_{\boldsymbol{v}\cdot\boldsymbol{n}<0}(t,\boldsymbol{x};\omega)$ is the random indicator variable equal to 1 when $\boldsymbol{v}\cdot\boldsymbol{n} < 0$ and 0 otherwise, and $\mathbf{1}_{\boldsymbol{v}\cdot\boldsymbol{n}\geq 0} = 1 - \mathbf{1}_{\boldsymbol{v}\cdot\boldsymbol{n}<0}$. Projecting again onto the coefficients, ζ_i, yields mixed boundary conditions for the modes, $\boldsymbol{u}_i$:

$$\mathbb{E}[\zeta_i\zeta_j \mathbf{1}_{\beta_k \boldsymbol{v}_k\cdot\boldsymbol{n}<0}]\boldsymbol{u}_j + \mathbb{E}[\zeta_i\zeta_j \mathbf{1}_{\beta_k \boldsymbol{v}_k\cdot\boldsymbol{n}\geq 0}]\frac{\partial \boldsymbol{u}_j}{\partial \boldsymbol{n}} = \mathbb{E}[\zeta_i \mathbf{1}_{\beta_k \boldsymbol{v}_k\cdot\boldsymbol{n}<0}]\boldsymbol{x} \text{ on } \partial\Omega. \tag{3.12}$$

So far, the methodology followed by works that have used DO equations to solve stochastic PDEs [124, 140, 103] has been deriving a set of coupled PDEs for modes and coefficients such as (3.10) to (3.12), before attempting to design appropriate time and spatial numerical schemes for these equations. However this approach makes unclear how to deal with numerical difficulties that are encountered in the discretization of the original PDE (1.1) and will arise in a similar manner in the model order reduced system. In the same way unadapted discretizations of the convective terms $\boldsymbol{v}\cdot\nabla\boldsymbol{\psi}$ in eqn. (1.1) lead to the blowing up of the numerical solution, a great deal of attention must be given to the discretization of the fluxes $\boldsymbol{v}_k\cdot\nabla\boldsymbol{u}_j$. Popular advection schemes [108] are using up-winding, in the sense that spatial derivatives are discretized according to the orientation of the velocity, $\boldsymbol{v}$, but it is unclear how this rule translates when the velocity $\boldsymbol{v}$ becomes stochastic. These difficulties were actually acknowledged in previous works dealing with stochastic Navier-Stokes equations. Ueckermann [147] proposed as an empirical remedy to average numerical fluxes according to the probability distribution of the velocity direction.

3.2.3 DO approximation in the discrete matrix setting: projected dynamical system on the fixed rank manifold

Instead of seeking numerical schemes for the continuous DO partial differential equations (3.10) to (3.12), a more rigorous approach is found by applying the DO methodology directly on the spatial discretization chosen for the original SPDE (3.2). This idea may indicate in general what should be the proper discretization of DO equations, assuming these are well-posed, given available numerical schemes to simulate each deterministic realization.

In the following, we denote by $\boldsymbol{\Psi}_{i,j}(t) = \boldsymbol{\psi}(t,\boldsymbol{x}_i;\omega_j) \in \mathcal{M}_{l,m}$ the entries of a l-by-m matrix of realizations stored in computer memory. Here, l denotes the spatial dimension (typically l/d nodes $\boldsymbol{x}_i$ are used for a d-dimensional domain) and m is the number of realizations ω_j are considered. The numerical solution $\boldsymbol{\Psi}(t)$ of the SPDE (3.2) is obtained by solving the matrix ODE in the ambiant space of l-by-m matrices $\mathcal{M}_{l,m}$:

$$\dot{\boldsymbol{\Psi}} = \mathcal{L}(t,\boldsymbol{\Psi}) \in \mathcal{M}_{l,m}, \tag{3.13}$$

where $\mathcal{L}$ is a matrix operator that includes spatial discretization of the realizations of the fluxes $-\boldsymbol{v}\cdot\nabla\boldsymbol{\psi}$, and of the boundary conditions (3.4). As in section 2.2.1, we denote

$$\Psi(t) = U(t)Z(t)^T \simeq \boldsymbol{\Psi}(t) \tag{3.14}$$

a rank r approximation $\Psi(t) \in \mathscr{M}$, where $U(t)$ and $Z(t)$ are respectively lower dimen-

sional l-by-r_Ψ and m-by-r_Ψ matrices containing the discretizations $U_{ik}(t) = \boldsymbol{u}_k(t, \boldsymbol{x}_i)$ and $Z_{jk}(t) = \zeta_k(t; \omega_j)$ of the modes and coefficients appearing in the decomposition (3.3). The orthogonality of the modes (3.8) and the DO condition (3.9) are written at the discrete level as

$$U^T U = I \text{ and } U^T \dot{U} = 0. \tag{3.15}$$

In this setting, the low-rank manifold is $\mathscr{M} = \{\Psi \in \mathcal{M}_{l,m} | \text{rank}(\Psi) = r_\Psi\}$ and the DO approximation $\Psi(t)$ is defined to be the solution of the projected dynamical system on $\mathscr{M}$ studied in the general case in section 2.3:

$$\left\{ \begin{array}{ll} \dot{\Psi} & = \Pi_{\mathcal{T}(\Psi)}(\mathcal{L}(t, \Psi)) \\ \Psi(0) & = \Pi_{\mathscr{M}}(\boldsymbol{\Psi}(0)). \end{array} \right. \tag{3.16}$$

Using the expression (2.29) for the projection $\Pi_{\mathcal{T}(\Psi)}$, this ODE system can be written as the set of coupled evolution equations for the mode and coefficient matrices U and Z, that turns to be exactly a discrete version of the continuous DO equations (3.10) and (3.11):

$$\left\{ \begin{array}{ll} \dot{Z} = & \mathcal{L}(t, UZ^T)^T U \\ \dot{U} = & (I - UU^T)\mathcal{L}(t, UZ^T) Z (Z^T Z)^{-1}. \end{array} \right. \tag{3.17}$$

These equations are exactly those presented as DO equations in [124]. With the notation of (4) and (5), using $<\cdot, \cdot>$ to denote the continuous dot product operator (an integral over the spatial domain) and $\mathbb{E}$ the expectation, they were written as the following set of coupled stochastic PDEs:

$$\left\{ \begin{array}{c} \partial_t \zeta_i = < \mathscr{L}(t, \boldsymbol{u}_{\text{DO}}; \omega), \boldsymbol{u}_i > \\ \displaystyle\sum_{j=1}^{r} \mathbb{E}[\zeta_i \zeta_j] \partial_t \boldsymbol{u}_j = \mathbb{E}\left[\zeta_i \left(\mathscr{L}(t, \boldsymbol{u}_{\text{DO}}; \omega) - \sum_{j=1}^{r} < \mathscr{L}(t, \boldsymbol{u}_{\text{DO}}; \omega), \boldsymbol{u}_j > \boldsymbol{u}_j \right) \right] . \end{array} \right. \tag{3.18}$$

However, when dealing with infinite dimensional Hilbert spaces, the vector space of solutions of (4) depends on the PDEs, which complicates the derivation of a general theory for (3.18). Considering the DO approximation as a computational method for evolving low rank matrices relaxes these issues through the finite-dimensional setting.

Applying directly theorem 2.5, one finds that the DO approximation (3.17) corresponds to the limit when the time step goes to 0 of a time integrating scheme for the original ODE (3.13) that would apply the truncated SVD after each time step to remove the optimal amount of information required to constrain the rank of the solution. Therefore, other reduced order models of the form (2.56) are characterized by larger errors on short integration times for solutions whose initial value are low-rank.

Remark 3.1. Other dynamical systems that perform instantaneous matrix operations have been studied in [20, 138, 31] (e.g. see Lemma 3.4 and Corollary 3.5) for tracking the full SVD or QR decomposition. Continuous SVD has been combined with adaptive Kalman filtering in uncertainty quantification to continuously adapt the dominant subspace supporting the stochastic solution [89, 90, 91]. The dominant singular vectors of state transition matrices and other operators have also found varied applications in atmospheric and ocean sciences [43, 111, 76, 93, 102, 80, 33].

To obtain a precise bound for the approximation error, we need to state a result for the

local lipschitz constant of the projection operator of the fixed rank manifold:

Lemma 3.1. *For any points $R^1, R^2 \in \mathscr{M}$ such that $||R^1 - R^2|| < \frac{1}{2}(\sigma_r(R^1) + \sigma_r(R^2))$,*

$$||\Pi_{\mathcal{T}(R^1)} - \Pi_{\mathcal{T}(R^2)}|| \leq \frac{||R^1 - R^2||}{\frac{1}{2}(\sigma_r(R^1) + \sigma_r(R^2)) - ||R^1 - R^2||} \leq \frac{||R^1 - R^2||}{\sigma_r(R^1) - \frac{3}{2}||R^1 - R^2||}, \quad (3.19)$$

where the norm of the left-handside is the operator norm. Also the following more global estimate holds (see [150] and Theorem 2.6.1 in [59]):

$$||\Pi_{\mathcal{T}(R^1)} - \Pi_{\mathcal{T}(R^2)}|| \leq \max\left(1, \frac{2}{\sigma_r(R^1)}||R^1 - R^2||\right). \quad (3.20)$$

Therefore the local lipschitz constant $\rho_\infty(R)$ of the projection operator $\Pi_{\mathcal{T}}$ on the fixed rank manifold satisfies $\rho_\infty(R) \leq 2/\sigma_r(R)$.

Proof. We give a purely geometric proof for the estimate (3.19), locally stronger than (3.20) (although the better global bound (3.20) is obtained algebraically in [150] and sufficient for our application). The idea is to consider a smooth curve $R(t) \in \mathscr{M}$ joining $R^1 = R(0)$ to $R^2 = R(1)$ and to use lemma 2.2 to bound

$$||\Pi_{\mathcal{T}(R^1)} - \Pi_{\mathcal{T}(R^2)}|| = \left|\left|\int_0^1 \mathrm{D}\Pi_{\mathcal{T}(R(t))}(\dot{R}(t))\mathrm{d}t\right|\right| \leq \sup_{t\in[0,1]} ||\mathrm{D}\Pi_{\mathcal{T}(R(t))}(\dot{R}(t))||. \quad (3.21)$$

Following [83], one defines $R(t) = \Pi_{\mathscr{M}}(\mathfrak{R}(t))$ as being the orthogonal projection of the straight line $\mathfrak{R}(t) = (1-t)R^1 + tR^2$ joining R^1 to R^2 onto the manifold. Since $\sigma_{r+1}(R^1) = \sigma_{r+1}(R^2) = 0$, the following bounds hold (see p.448 in [75]):

$$\sigma_{r+1}(\mathfrak{R}(t)) \leq \min(t, 1-t)||R^1 - R^2|| \leq \frac{1}{2}||R^1 - R^2||,$$
$$\begin{cases} \sigma_r(\mathfrak{R}(t)) \geq \sigma_r(R^1) - t||R^1 - R^2|| \\ \sigma_r(\mathfrak{R}(t)) \geq \sigma_r(R^2) - (1-t)||R^1 - R^2||. \end{cases}$$

Therefore

$$\sigma_r(\mathfrak{R}(t)) - \sigma_{r+1}(\mathfrak{R}(t)) \geq \frac{\sigma_r(R^1) + \sigma_r(R^2)}{2} - ||R^1 - R^2||. \quad (3.22)$$

A consequence of lemma 2.1 is that:

$$||\dot{R}(t)|| = ||\mathrm{D}_{\dot{\mathfrak{R}}}\Pi_{\mathscr{M}}(\mathfrak{R}(t))|| \leq \frac{\sigma_r(\mathfrak{R}(t))}{\sigma_r(\mathfrak{R}(t)) - \sigma_{r+1}(\mathfrak{R}(t))}||R^2 - R^1||. \quad (3.23)$$

Therefore, combining the bound of lemma 2.2 with (3.22) and (3.23):

$$||\mathrm{D}\Pi_{\mathcal{T}(R(t))}(\dot{R}(t))|| \leq \frac{1}{\sigma_r(R(t))}||\dot{R}(t)|| \leq \frac{||R^2 - R^1||}{\frac{1}{2}(\sigma_r(R^1) + \sigma_r(R^2)) - ||R^2 - R^1||},$$

which together with (3.21) and $\sigma_r(R_2) \geq \sigma_r(R_1) - ||R^1 - R^2||$ proves the result. Notice that (3.19) combined with $||\Pi_{\mathcal{T}(R^1)} - \Pi_{\mathcal{T}(R^2)}|| \leq 1$ allows to obtain $||\Pi_{\mathcal{T}(R^1)} - \Pi_{\mathcal{T}(R^2)}|| \leq \frac{5}{2\sigma_r(R^1)}||R^1 - R^2||$, but the constant 2 is an improvement of [150]. □

Using this lemma and the fact that the maximal curvature of the fixed rank manifold at a point $R \in \mathscr{M}$ is $\kappa_\infty = 1/\sigma_r(R)$ (see section 2.2.1), we can restate theorem 2.6 as follows :

Theorem 3.1. *Consider $\mathbf{\Psi}(t) \in \mathcal{M}_{l,m}$ the solution of the original dynamical system (3.13). Assume that the following conditions hold on a time interval $[0,T]$:*

1. *$\mathcal{L}$ is Lipschitz continuous,* i.e. *equation (2.62) holds.*
2. *The original solution $\mathbf{\Psi}(t)$ remains close to the low rank manifold $\mathscr{M}$, in the sense that $\mathbf{\Psi}(t)$ does not cross the skeleton of $\mathscr{M}$ on $[0,T]$,* i.e. *that is there is no crossing of the singular value of order r:*

$$\forall t \in [0,T],\ \sigma_r(\mathbf{\Psi}(t)) > \sigma_{r+1}(\mathbf{\Psi}(t)).$$

Then, the error of the DO approximation $\Psi(t)$ (equation (3.17)) remains controlled by the best approximation error $||\mathbf{\Psi} - \Pi_{\mathscr{M}}(\mathbf{\Psi}(t))||$ on $[0,T]$:

$$\forall t \in [0,T],\ ||\Psi(t) - \Pi_{\mathscr{M}}(\mathbf{\Psi}(t))|| \leq \int_0^t ||\mathbf{\Psi}(s) - \Pi_{\mathscr{M}}(\mathbf{\Psi}(s))|| \left(K + \frac{||\mathcal{L}(s, \mathbf{\Psi}(s))||}{\sigma_r(\mathbf{\Psi}(s)) - \sigma_{r+1}(\mathbf{\Psi}(s))} \right) e^{\boldsymbol{\eta}(t-s)} \mathrm{d}s, \quad (3.24)$$

where $\boldsymbol{\eta}$ is the constant

$$\boldsymbol{\eta} = K + \sup_{t \in [0,T]} \frac{2}{\sigma_r(\mathbf{\Psi}(t))} ||\mathcal{L}(t, \mathbf{\Psi}(t))||. \quad (3.25)$$

As observed numerically in [103], the DO solution may diverge sharply from the SVD truncation after a crossing of the singular values $\sigma_r(\mathbf{\Psi})$ and $\sigma_{r+1}(\mathbf{\Psi})$. From the point of view of model order reduction, the resulting error can be related to the evolution of the residual $\mathbf{\Psi} - \Pi_{\mathscr{M}}(\mathbf{\Psi})$ that is not accounted for by the reduced order model. When the crossing of singular values occurs, neglected modes in the approximation (5) become "dominant", but cannot be captured by a reduced order model that has evolved only the first modes initially dominant. Note that if one would track exactly the best L^2 approximation of $\mathbf{\Psi}(t)$, one would need to use the dynamical system (2.41) of corollary 2.3, that require the knowledge of the non reduced solution $\mathbf{\Psi}(t)$, or a closure to model the unknown modes and singular values.

3.3 Implementation of the DO approximation for stochastic advection

In the following, it is explained how the DO approximation (3.17) can be implemented in practice to solve the stochastic transport equation (3.2).

3.3.1 Motivations for linear advection schemes

The DO approximation is computationally attractive because it allows to replace the dynamical system (3.13) that requires to evolve lm matrix coefficients with (3.17), that evolves a solution set on the manifold of the – much smaller – dimension $(l+m)r - r^2$, by evolving the $lr+mr$ coefficients of the matrices U and Z. Nevertheless, the DO matrix system (3.17) offers a true gain of computational efficiency *only* if the evaluation of l-by-m matrices can be avoided. This is not a priori achievable in this formulation as the operator $\mathcal{L}$ needs to be evaluated on the l-by-m matrix $\Psi = UZ^T$. In the case where all lm coefficients of Ψ need

to be computed from U and Z, the method provides no computational benefit other than a reduction of memory storage in comparison with solving the initial non-reduced system (3.13).

The gain of efficiency can be achieved if the operator $\mathcal{L}(t,\cdot)$ maps a rank r_Ψ decomposition $\Psi = UZ^T$ onto a factorization

$$\mathcal{L}(t, UZ^T) = L_U L_Z^T \tag{3.26}$$

of rank at most r', where L_U is a l-by-k matrix, L_Z a m-by-k matrix, and r' an integer typically largely inferior to l and m. In that case, the system (3.17) can be computed efficiently as

$$\begin{cases} \dot{U} = & [(I - UU^T)L_U][L_Z^T Z(Z^T Z)^{-1}] \\ \dot{Z} = & L_Z[L_U^T U]. \end{cases} \tag{3.27}$$

where brackets have been used to highlight products that allow to compute the derivatives $\dot{U}$ and $\dot{Z}$ without having to deal with l-by-m matrices. Such factorization occurs for instance when $\mathcal{L}(t,\cdot)$ is polynomial of order d, for which rank r_Ψ matrices are mapped onto rank $r' \leq r^d$ matrices.

In the spatially continuous view point, the differential operator $\boldsymbol{\psi} \mapsto \boldsymbol{v}\cdot\nabla\boldsymbol{\psi}$ satisfies this condition, as the rank r_Ψ decomposition (3.3) is mapped to a one of rank $r_L = r_\Psi \times r_{\boldsymbol{v}}$:

$$\boldsymbol{v}\cdot\nabla\boldsymbol{\psi} = \sum_{\substack{1\leq j\leq r_\Psi \\ 1\leq k\leq r_{\boldsymbol{v}}}} \zeta_j \beta_k \boldsymbol{v}_k \cdot \nabla \boldsymbol{u}_j. \tag{3.28}$$

This equation further highlights why adapting advection schemes to model order reduction is challenging, as popular discretizations of $\boldsymbol{v}\cdot\nabla\boldsymbol{\psi}$ involve non-polynomial nonlinearities in the matrix operator $\mathcal{L}$. These schemes rely indeed on the use of min-max functions required by upwinding or high order discretizations such as ENO or TVD schemes that are selecting a smooth approximation of the spatial derivative $\nabla\boldsymbol{\psi}$. In theses cases, the nonlinearity of the operator $\mathcal{L}$ prevents the decomposition (3.28) to hold at the discrete level without introducing further approximations, which may alter drastically the stability of time integration and the accuracy of the numerical solution. A very natural approach followed by [124, 147] is to assume that the decomposition (3.28) holds before applying non linear schemes to discretize the fluxes $\boldsymbol{v}_k\cdot\nabla\boldsymbol{u}_j$ in (3.10) and (3.11). A key issue includes maintaining the consistency between the deterministic MC and DO solutions. In the example considered in section 3.4 for which high gradients especially occur, such approaches were observed to lead to either numerical explosion or very large errors on long integration time.

Consequently, this work investigated the use of linear central advection schemes that do not require upwinding and that have the property to preserve the decomposition (3.28). Therefore, the flux $-\boldsymbol{v}\cdot\nabla\boldsymbol{\psi}$ is discretized as

$$\mathcal{L}(t,\boldsymbol{\Psi})_{i,\alpha} = -\boldsymbol{v}(t,\boldsymbol{x}_i;\omega_\alpha)\cdot \mathrm{D}\boldsymbol{\Psi}_{i,\alpha} \tag{3.29}$$

where D is a linear finite-difference operator approximating the gradient ∇. With $\Psi = UZ^T$ as in (3.14), this allows to obtain $\mathcal{L}(t,\Psi) = L_U L_Z^T$ as required in (3.27), with L_U and L_Z the l-by-r_L and m-by-r_L matrices

$$(L_U)_{i,jk} = \boldsymbol{v}_k(t,\boldsymbol{x}_i)\cdot \mathrm{D}\boldsymbol{u}_j(t,\boldsymbol{x}_i),\ (L_Z)_{\alpha,jk} = \zeta_j(t;\omega_\alpha)\beta_k(t,\omega_\alpha).$$

In one dimension, the gradient can be approximated by the second order operator

$$\mathrm{D}\boldsymbol{\Psi}_{i,\alpha} = \frac{\boldsymbol{\Psi}_{i+1,\alpha} - \boldsymbol{\Psi}_{i-1,\alpha}}{2\Delta x}, \tag{3.30}$$

and we will also consider the sixth order finite difference operator

$$\mathrm{D}\boldsymbol{\Psi}_{i,\alpha} = \frac{3}{2}\frac{\boldsymbol{\Psi}_{i+1,\alpha} - \boldsymbol{\Psi}_{i-1,\alpha}}{2\Delta x} - \frac{3}{5}\frac{\boldsymbol{\Psi}_{i+2,\alpha} - \boldsymbol{\Psi}_{i-2,\alpha}}{4\Delta x} + \frac{1}{10}\frac{\boldsymbol{\Psi}_{i+3,\alpha} - \boldsymbol{\Psi}_{i-3,\alpha}}{6\Delta x}, \tag{3.31}$$

where Δx denotes the spatial resolution and a natural numbering is assumed for the index i. These formula are adapted in a straightforward manner to discretize partial derivatives in higher dimension [108]. This approach might seem unexpected, since central schemes are known to be numerically unstable under Euler time integration. In addition, Godunov theorem expresses that it is not possible to devise a linear scheme higher than first order accuracy that do not create false extrema in numerical solutions [56]. These extrema are produced by numerical dispersion and manifest under the form of spurious oscillations. In fact, it is possible to contain this phenomenon near shocks, and obtain high order accuracy where the solution is smooth. Stability and the removal of part of the oscillations can be achieved by the introduction of a right amount of numerical dissipation, either by using artificial viscosity [136] or filtering [131, 38, 85, 115, 35]. Shapiro filters are especially attractive because easy to implement, fully linear, and designed to remove optimally the shortest resolvable numerical frequency without affecting other wave components [131, 132, 133]. In one dimension, denoting δ^2 the operator $\delta^2\Psi_{i,\alpha} = \Psi_{i+1,\alpha} - 2\Psi_{i,\alpha} + \Psi_{i+1,\alpha}$, the Shapiro filters $\mathcal{F}^{(i)}$ of order $i = 2$, 4 and 8 are defined by the formulas (see [131])

$$\begin{array}{ll} \mathcal{F}^{(2)}\Psi_{i,\alpha} & = (1 + \delta^2/4)\Psi_{i,\alpha} \\ \mathcal{F}^{(4)}\Psi_{i,\alpha} & = (1 - \delta^2/4)(1 + \delta^2/4)\Psi_{i,\alpha} \\ \mathcal{F}^{(8)}\Psi_{i,\alpha} & = (1 + \delta^4/16)(1 - \delta^4/16)\Psi_{i,\alpha}. \end{array} \tag{3.32}$$

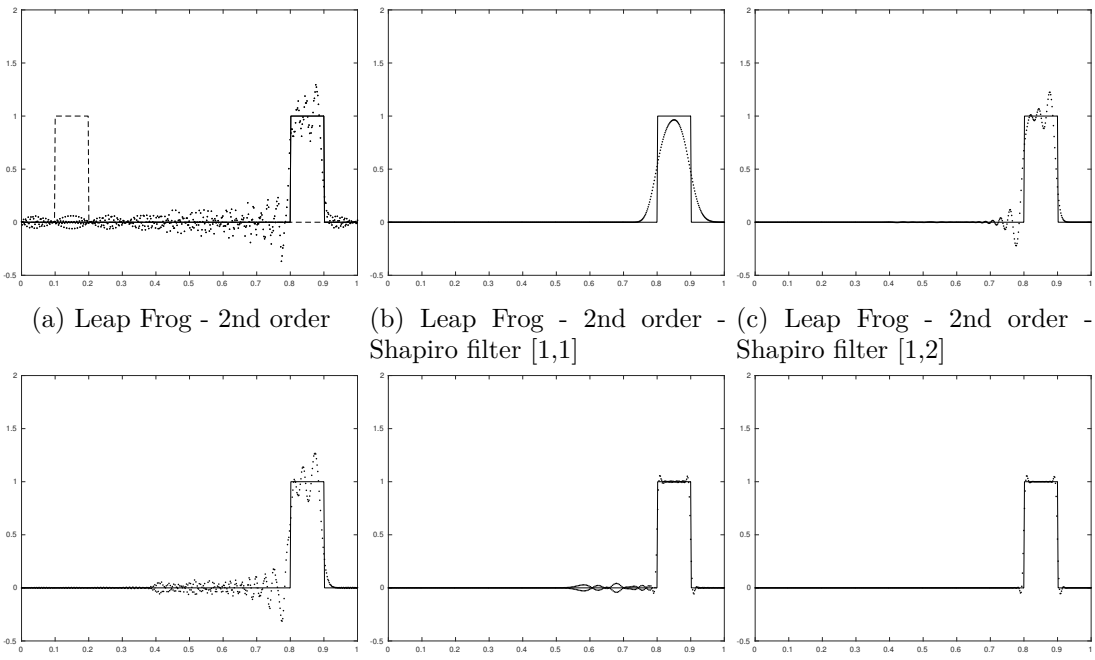
Their linearity allows to filter the decomposition $\boldsymbol{\psi} = \zeta_i \boldsymbol{u}_i$ efficiently by filtering the discretization of the modes $\boldsymbol{u}_i$, or in other words, $\mathcal{F}^{(i)}(UZ^T) = (\mathcal{F}^{(i)}U)Z^T$. The order and frequency of applications can be turned to the desired filter-spectrum [89], and linear limiters may be combined with Shapiro filters. To achieve further stability, higher order discretizations of the temporal derivative are generally used in complement to these filters. Popular linear multi-step methods range from Leap-Frog [152], Runge-Kutta and Adam Bashforth [35]. For instance, the second order Leap-Frog scheme evolves the value $\boldsymbol{\Psi}^n$ of the numerical solution $\boldsymbol{\Psi}$ at time t^n according to the rule

$$\frac{\boldsymbol{\Psi}^{n+1} - \boldsymbol{\Psi}^{n-1}}{2\Delta t} = \mathcal{L}(t^n, \boldsymbol{\Psi}^n), \tag{3.33}$$

while the third order Runge Kutta (RK3) method uses

$$\frac{\boldsymbol{\Psi}^{n+1} - \boldsymbol{\Psi}^n}{\Delta t} = \frac{k_1^n + 4k_2^n + k_3^n}{6} \text{ with } \left\{ \begin{array}{ll} k_1^n & = \mathcal{L}(t^n, \boldsymbol{\Psi}^n) \\ k_2^n & = \mathcal{L}\left(t^n + \Delta t/2, \boldsymbol{\Psi}^n + k_1^n \Delta t/2\right) \\ k_3^n & = \mathcal{L}(t^n + \Delta t, \boldsymbol{\Psi}^n + \Delta t(2k_2^n - k_1^n)). \end{array} \right. \tag{3.34}$$

A comparison of several combinations of these techniques is illustrated on Figure 3-1 for the one dimensional advection equation $\partial_t \boldsymbol{\psi} + v\partial_x \boldsymbol{\psi} = 0$, a benchmark case for selecting an appropriate linear scheme for the transport eqn. (3.2) in higher dimension. A boxcar



(a) Leap Frog - 2nd order

(b) Leap Frog - 2nd order - Shapiro filter [1,1]

(c) Leap Frog - 2nd order - Shapiro filter [1,2]

(d) RK3 - 2nd order

(e) RK3 - 6th order

(f) RK3 - 6th order - Shapiro filter [10,3]

Figure 3-1: Comparison of different linear centered schemes for the 1D advection equation. The mention Shapiro filter $[n_1, n_2]$ indicates that the Shapiro filter of order 2^{n_2} (see [131]) has been applied after every n_1 iterations. The initial box-car function is visible in dashed line on the first plot.

function is advected to the right with a velocity $v = 0.7$ in the domain $[0, 1]$ until the time $t = 10$. The spatial resolution is set to $\Delta x = 0.002$ and the CFL condition $\Delta t \leq 0.6 v \Delta x$ is used to define the time increment Δt. The figure illustrates how accuracy and stability can be achieved by (i) using multi-step time marching schemes, (ii) using high order spatial discretization and (iii) adding a proper amount of numerical dissipation to remove spurious oscillations.

3.3.2 Boundary conditions

Boundary conditions of the reduced solution have been formally obtained in section 3.2. They could be treated more rigorously by incorporating original boundary conditions (3.4) and (3.5) explicitly in the discretization operator $\mathcal{L}$. However, such view point is inconvenient in the implementation. In this work, boundary nodes are stored in a $l_{bc} \times m$ "ghost" matrix and it is assumed that the l-by-m matrix of realizations $\mathbf{\Psi}$ contains only the values at internal nodes. These ghost cells allow to evaluate conveniently the differential operator D in the definition (3.29) of $\mathcal{L}(t, \mathbf{\Psi})$. Their values are reinitialized at the beginning of each time step according to the boundary conditions (3.4) and (3.5). In the following, the operator which assigns the values of these boundary cells at time t is denoted $\mathcal{B}_C(t, \cdot)$. With this notation, the solution that includes both internal nodes and boundary values is the block matrix $\mathbf{\Psi}_{bc} = \begin{bmatrix} \mathcal{B}_C(t, \Psi) \\ \Psi \end{bmatrix}$. For example on the one-dimensional domain $\Omega = [0, 1]$, the value

of the boundary node $\boldsymbol{x}_1 = 0$ is determined by the relation

$$\mathcal{B}_C(t, \boldsymbol{\Psi})_{1,\alpha} = \begin{cases} 0 & \text{if } \boldsymbol{v}(t, 0; \alpha) \geq 0 \\ (18\boldsymbol{\Psi}_{2,\alpha} - 9\boldsymbol{\Psi}_{3,\alpha} + 2\boldsymbol{\Psi}_{4,\alpha})/11 & \text{if } \boldsymbol{v}(t, 0; \alpha) < 0, \end{cases}$$

if one uses a third order reconstruction for the Neumann boundary condition (3.5). The difficulty of determining how these boundary conditions should be accounted for by the reduced solution $\Psi = UZ^T$ comes from the fact that assigning boundary values does in general not preserve the rank: $\text{rank}(\boldsymbol{\Psi}_{bc}) > r$. Boundary conditions may be enforced on the reduced solution while ensuring minimal error by solving the minimization problem

$$\min_{\text{rank}(\Psi_{bc})=r} \left|\left| \Psi_{bc} - \begin{bmatrix} \mathcal{B}_C(t, \Psi) \\ \Psi \end{bmatrix} \right|\right|^2. \tag{3.35}$$

This yields the best rank r_Ψ approximation of the $(l + n_{BC})$-by-m matrix $\boldsymbol{\Psi}_{bc}$, whose decomposition $\Psi_{bc} = U_{bc}Z_{bc}^T$ allows to compute conveniently the discrete differential operator D in (3.29) requiring boundary values. The minimization can for example be achieved by using a gradient descent starting from the initial rank r_Ψ matrix Ψ, as explained in the next subsection and in [44, 101].

When boundary conditions are deterministic or homogeneous, they can be directly implemented as boundary conditions for the discretization of the modes, $\boldsymbol{u}_i$ [124]. For example, prescribing zero or Neumann boundary conditions for all the realizations of $\boldsymbol{\psi}$ is found by prescribing directly the same boundary conditions for the modes, $\boldsymbol{u}_i$. For more general cases, it can still be desirable to avoid solving (3.35) by seeking boundary values for the modes so as to optimally approximate the original boundary conditions. This is achieved by replacing the minimization problem (3.35) with a one for the l_{bc}-by-r_Ψ ghost matrix U_{BC} containing boundary values for the matrix U:

$$\min_{U_{BC} \in \mathcal{M}_{l_{bc}, r}} ||U_{BC}Z^T - \mathcal{B}_C(t, \Psi)||^2. \tag{3.36}$$

The solution of this linear regression problem is easily obtained by

$$U_{BC} = \mathcal{B}_C(t, \Psi)Z(Z^TZ)^{-1}, \tag{3.37}$$

and it turns out that this optimality condition is the discrete analogous of the original boundary conditions (3.12) obtained formally in section 3.2. The decomposition of the reduced solution including boundary values considered is therefore $\Psi_{bc} = \begin{bmatrix} U_{BC} \\ U \end{bmatrix} Z^T$.

3.3.3 Low-rank Time stepping and continuous SVD tracking

One issue commonly encountered in the time discretization of dynamical systems defined on differentiable manifolds is the fact that the discrete time stepping tends to make the numerical solution exit the manifold. If Ψ^n is a point on the manifold $\mathscr{M}$ at t^n, and $\dot{\Psi}^n \in \mathcal{T}(R)$ is the time derivative, any straight move $\Psi^n + \Delta t\dot{\Psi}^n$ leaves $\mathscr{M}$. An application, called *retraction* [6, 148, 4], must be used to convert the tangent direction $X = \Delta t\dot{\Psi}^n \in \mathcal{T}(\Psi^n)$ into a point $\rho_{\Psi^n}(X)$ back onto the manifold.

Definition 3.1. A retraction ρ_Ψ is an application from the tangent space at $\Psi \in \mathcal{T}(\Psi)$

onto the manifold,

$$\begin{array}{rccc} \rho_\Psi & : & \mathcal{T}(\Psi) & \to & \mathscr{M} \\ & & X & \mapsto & \rho_\Psi(X). \end{array}$$

that in addition satisfies the following consistency conditions:

1. $\rho_\Psi(0) = \Psi$
2. $\forall X \in \mathcal{T}(\Psi)$, $\mathrm{D}_X\rho_\Psi(0) = \frac{\mathrm{d}}{\mathrm{d}t}\rho_{\Psi^n}(tX)\Big|_{t=0} = X$.

The two conditions require a retraction to approximate at the first order the exponential map. They are restatement that (1.) if one walks on the manifold starting from Ψ with zero velocity, one stays at the initial position Ψ, and (2.), if $X \in \mathcal{T}(\Psi)$ is the argument of the retraction, then one indeed move on the manifold with a velocity X. One can show that if in addition, $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\rho_\Psi(tX)\Big|_{t=0} \in \mathcal{N}(\Psi)$ then the retraction is a second order approximation of the exponential map [6]. Depending on the choice of the retraction, several implementations can be considered for the explicit discretization of (3.17).

1. Direct time marching scheme for the DO system (3.17)

As in [147, 103], a very intuitive idea for moving a rank r_Ψ matrix $\Psi^n = U^n Z^{nT}$ onto a direction $\dot{\Psi}^n = \dot{U}^n Z^{nT} + U^n \dot{Z}^{nT}$ with a step Δt is to update independently the mode and coefficient matrices U^n and Z^n by using the following scheme, which is a direct time-discretization of the system (3.17):

$$\left\{ \begin{array}{rcl} Z^{n+1} = & Z^n + \Delta t \dot{Z}^n \\ U^{n+1} = & U^n + \Delta t \dot{U}^n, \end{array} \right. \tag{3.38}$$

where $\dot{Z}^n$ and $\dot{U}^n$ are the approximations of the time derivatives $\dot{U}$ and $\dot{Z}$ being used. This corresponds to use the retraction ρ_{UZ^T} defined by

$$\rho_{UZ^T}(\dot{U}Z^T + U\dot{Z}^T) = (U + \dot{U})(Z + \dot{Z})^T = UZ^T + (\dot{U}Z^T + U\dot{Z}^T) + \dot{U}\dot{Z}^T. \tag{3.39}$$

2. Geodesic equations in between time steps to deal with ill-conditioned matrices

The ideal retraction is the exponential map $\rho_{\Psi^n} = \exp_{\Psi^n}$ (see [6]) computed from geodesic paths $\gamma(s)$ on $\mathscr{M}$, which are the direct analogous of straight lines onto curved manifolds. These curves are parametrized as $\gamma(s) = \exp_{\Psi^n}(s\dot{\Psi}^n)$ (see Figure 2-1), indicating how to "walk" onto the manifold from Ψ^n into the straight direction $\dot{\Psi}^n = \dot{U}^n(Z^n)^T + U^n(\dot{Z}^n)^T$. It was shown in chapter 2 that the value of $\exp_{\Psi^n}(s\dot{\Psi}^n)$ is given by the solution $\gamma(s) = U(s)Z(s)^T$ at time s of the geodesic equations (eqn. (2.32)):

$$\left\{ \begin{array}{rcl} \ddot{U} + U\dot{U}^T\dot{U} + 2\dot{U}\dot{Z}^T Z(Z^T Z)^{-1} = & 0 \\ \ddot{Z} - Z\dot{U}^T\dot{U} = & 0. \\ U(0) = U^n,\ Z(0) & = Z^n \\ \dot{U}(0) = \dot{U}^n,\ \dot{Z}(0) & = \dot{Z}^n. \end{array} \right. \tag{3.40}$$

Without direct analytical solutions to (2.32), numerical schemes are used. Computing retractions that approximate well the exponential map is a challenge commonly encountered

in optimization on matrix manifolds with orthogonality constraints [101], as discussed in [6]. One can check that the retraction ρ_{UZ^T} of equation (3.39) is approximating the exponential map only to the first order (see [6]), which can lead to numerical errors at locations of high curvature on the manifold $\mathscr{M}$. The curvature of the rank r_Ψ manifold $\mathscr{M}$ at the point Ψ^n is inversely proportional to the lowest singular value $\sigma_{r_\Psi}(\Psi^n)$ [44]. As a consequence, errors can be incurred by the direct time stepping (3.38) when the matrix Z^n is ill conditioned. Equations (2.32) can be solved during the DO time integration in between time steps, to move more accurately on the manifold without the need for recomputing values of the operator $\mathcal{L}$. For instance, Euler steps (3.38) can be replaced with

$$U^{n+1}(Z^{n+1})^T = \exp_{\Psi^n}(\Delta t \dot{\Psi^n}). \tag{3.41}$$

This can be done using high order time marching schemes for the discretization of (2.32). The intermediate time step δt for these can be set adaptively: a rule of thumb is to use time steps having a length lower than the minimal curvature radius $\sigma_{r_\Psi}(Z)$ at the point UZ^T:

$$\delta t ||\dot{U} Z^T + U \dot{Z}^T|| < C \sigma_{r_\Psi}(Z),$$

where $C \simeq 1$ is a constant set by the user.

3. Algebraic computation of the truncated SVD after each time step

It has been highlighted in section 3.2 that DO eqs. (3.27) define a dynamical system that truncates the SVD at all instants so as to optimally constrain the rank of the reduced solution (eqn. theorem 2.5). Denoting $\Psi^n = U^n(Z^n)^T$ the DO solution at time t^n, integrating the non-reduced dynamical system (3.13) for a time step $[t^n, t^{n+1}]$ yields a rank $r_{\overline{L}} > r_\Psi$ prediction

$$\mathbf{\Psi}^{n+1} = \Psi^n + \Delta t \overline{\mathcal{L}(t^n, \Psi^n)}, \tag{3.42}$$

where $\overline{\mathcal{L}(t^n, \Psi^n)}$ represent the full-space integral for the exact integration or the increment function for a numerical integration. For the latter, it can be an approximation of the time derivative $\mathcal{L}(t^n, \Psi(t^n))$, e.g. $\overline{\mathcal{L}(t^n, \Psi^n)} = \mathcal{L}(t^n, \Psi^n)$ for explicit Euler.

One way to proceed for evolving the low rank approximation Ψ^n to Ψ^{n+1} is to compute directly the rank r_Ψ SVD truncation $\Pi_{\mathscr{M}}(\mathbf{\Psi}^{n+1})$ (eqn. (2.33))

$$\Psi^{n+1} = U^{n+1}(Z^{n+1})^T = \Pi_{\mathscr{M}}(\Psi^n + \Delta t \overline{\mathcal{L}(t^n, \Psi^n)}) \tag{3.43}$$

so as to obtain modes and coefficients U^{n+1} and Z^{n+1} at time $t^{n+1} = t^n + \Delta t$. Such scheme has been shown to be a consistent time-discretization of the DO equations (2.57) (see [44]). For an Euler step, it corresponds to using the retraction $\rho_\Psi(X) = \Pi_{\mathscr{M}}(\Psi + X)$, a second-order accurate approximation of the exponential map [6].

The scheme (3.43) can be computed efficiently an in a fully algebraic manner when the operator $\mathcal{L}$ factors as (3.26). Indeed, the linear approximation of the time derivative then admits a decomposition $\overline{\mathcal{L}(t^n, U^n(Z^n)^T)} = L_U^n (L_Z^n)^T$ of rank at most $r_{\overline{L}} = r_L \times p_t$, p_t being the order of the time integration scheme utilized. Therefore $\mathbf{\Psi}^{n+1}$ factors as

$$\begin{aligned} \mathbf{\Psi}^{n+1} &= U^n(Z^n)^T + \Delta t L_U^n (L_Z^n)^T \\ &= \mathbf{\Psi}_U^{n+1} (\mathbf{\Psi}_Z^{n+1})^T \text{ with } \mathbf{\Psi}_U^{n+1} = [U^n \;\; L_U^n] \text{ and } \mathbf{\Psi}_Z^{n+1} = [Z^n \;\; \Delta t L_Z^n], \end{aligned} \tag{3.44}$$

with $L_U^n \in \mathcal{M}_{l,r_{\overline{L}}}, L_Z^n \in \mathcal{M}_{m,r_{\overline{L}}}$. The rank of $\mathbf{\Psi}^{n+1}$ is therefore at most $\text{rank}(\mathbf{\Psi}^{n+1}) = r_{\mathbf{\Psi}} < r + r_{\overline{L}}$ which can be assumed to be largely inferior to l and m. This can be exploited to compute the truncated SVD through an algorithm that avoids computing large matrices of size l-by-m (see algorithm 3).

Algorithm 3 Rank r_Ψ truncated SVD of $\mathbf{\Psi} = \mathbf{\Psi}_U \mathbf{\Psi}_Z^T$ with $\mathbf{\Psi}_U \in \mathcal{M}_{l,r_{\mathbf{\Psi}}}$, $\mathbf{\Psi}_Z \in \mathcal{M}_{m,r_{\mathbf{\Psi}}}$ and $r < r_{\mathbf{\Psi}} << \min(l, m)$

1: Orthonormalize the columns of the matrix $\mathbf{\Psi}_U$ (see the discussion in section 3.3.5), *i.e* find a basis change matrix $A \in \mathcal{M}_{r_{\mathbf{\Psi}},r_{\mathbf{\Psi}}}$ such that $(\mathbf{\Psi}_U A)^T(\mathbf{\Psi}_U A) = I$ and set

$$\mathbf{\Psi}_U \leftarrow \mathbf{\Psi}_U A,\ \mathbf{\Psi}_Z \leftarrow \mathbf{\Psi}_Z A^{-T}$$

so as to preserve the product $\mathbf{\Psi} = \mathbf{\Psi}_U \mathbf{\Psi}_Z^T$.

2: Compute the "compact" SVD of the *smaller* m-by-$r_{\mathbf{\Psi}}$ matrix $\mathbf{\Psi}_Z$:

$$\mathbf{\Psi}_Z = V \Sigma P^T,$$

where Σ is a $r_{\mathbf{\Psi}}$-by-$r_{\mathbf{\Psi}}$ diagonal matrix of singular are values, and $V \in \mathcal{M}_{m,r_{\mathbf{\Psi}}}$ and $P \in \mathcal{M}_{r_{\mathbf{\Psi}},r_{\mathbf{\Psi}}}$ orthogonal matrices of singular vectors. This is achieved by computing the eigen decomposition of the "covariance" matrix $\mathbf{\Psi}_Z^T \mathbf{\Psi}_Z$.

3: The SVD of $\mathbf{\Psi} = \mathbf{\Psi}_U \mathbf{\Psi}_Z^T$ is given by $\mathbf{\Psi} = U \Sigma V^T$ with $U = \mathbf{\Psi}_U P$ an orthogonal l-by-$r_{\mathbf{\Psi}}$ matrix of left singular vectors. The truncated SVD of order r_Ψ is straightforwardly obtained from the first r_Ψ columns of U, V and Σ.

This first algorithm has some issues. First, reorthonormalizations and eigenvalue decompositions such as in steps 1 and 2 do not allow to keep track of the smooth evolution of the modes $U(t)$ and coefficients $Z(t)$ solutions of the system (3.17). Additional procedures are needed [147, 146]. Second, with the repeated use of such algebraic operations, additional round off errors may be introduced.

4. Riemannian gradient descent on the low-rank manifold for continuous updates of the truncated SVD

Alternatively, a gradient descent on the low-rank manifold $\mathscr{M}$ can be used to find the correction that needs to be added to modes U^n and coefficients Z^n, so as to evaluate the SVD truncation $\Psi^{n+1} = \Pi_{\mathscr{M}}(\Phi^n)$ (eqns. (3.43) and (3.44)). In a more general setting, consider $\mathbf{\Psi} \in \mathcal{M}_{l,m}$ a l-by-M matrix for which one wants to evaluate the truncated SVD $\Pi_{\mathscr{M}}(\mathbf{\Psi})$. By definition, $\Pi_{\mathscr{M}}(\mathbf{\Psi})$ is the solution of the minimization problem

$$\min_{UZ^T \in \mathscr{M}} J(UZ^T) = \frac{1}{2}||\mathbf{\Psi} - UZ^T||^2. \tag{3.45}$$

In the DO time stepping, $\mathbf{\Psi} = \Phi^n$ and $\Psi = U^n Z^{nT} \simeq \Pi_{\mathscr{M}}(\mathbf{\Psi})$ is a low-rank approximation of $\mathscr{M}$. Therefore a Riemannian gradient descent on the manifold $\mathscr{M}$ can be used (see also [4, 101, 148]) to improve the quality of this low-rank approximation Ψ at a cheap expense. As reviewed in [36], usual optimization algorithms such as gradient and Newton methods can be straightforwardly adapted to matrix manifolds. The differences with their euclidean counterparts is that: (i) usual gradient and Hessians must be replaced by their covariant equivalents (definition 2.8); (ii) one needs to follow geodesics instead of straight lines to

move on the manifold; and, (iii) directions followed at the previous time steps, needed for example in the conjugate gradient method, must be transported to the current location (definition 2.5).

In proposition 2.16, it has been proven that *as a geometric feature of the fixed-rank manifold* $\mathscr{M}$, the distance function J may admit several critical points, but a unique local, hence global, minimum on $\mathscr{M}$. As a consequence, saddle points of J are unstable equilibrium solutions of the gradient flow $\dot{\Psi} = -\nabla J(\Psi)$ and hence are expected to be avoided by gradient descent, which will converge in practice to the global minimum $\Pi_{\mathscr{M}}(\mathbf{\Psi})$. This does not apply to the Newton method which seeks a zero of the gradient ∇J rather than a true minimum. This result is a convergence guarantee for the gradient descent and may be compared to [113, 150].

Applying directly proposition 2.3, the gradient and the Hessian of J at $R = UZ^T \in \mathscr{M}$ are given by:

$$\nabla J = ((I - UU^T)(UZ^T - \mathbf{\Psi})Z(Z^TZ)^{-1}, (UZ^T - \mathbf{\Psi})^T U), \tag{3.46}$$

$$\begin{array}{rccc} \mathcal{H}J & : & \mathcal{H}_{(U,Z)} & \to & \mathcal{H}_{(U,Z)} \\ & & \begin{pmatrix} X_U \\ X_Z \end{pmatrix} & \mapsto & \begin{pmatrix} X_U - N_{UZ^T}(\mathbf{\Psi})X_Z(Z^TZ)^{-1} \\ X_Z - N_{UZ^T}(\mathbf{\Psi})^T X_U \end{pmatrix}, \end{array} \tag{3.47}$$

where $N_{UZ^T}(\mathbf{\Psi}) = (I - \Pi_{\mathcal{T}(UZ^T)})(\mathbf{\Psi} - UZ^T) = (I - UU^T)\mathbf{\Psi}(I - Z(Z^TZ)^{-1}Z^T)$ is the orthogonal projection of $\mathbf{\Psi} - R$ onto the normal space. The Newton direction X is found by solving the linear system $\mathcal{H}J(X) = -\nabla J(\Psi)$, that reduces to

$$\begin{cases} X_U A + BX_Z = E \\ B^T X_U + X_Z = F, \end{cases}$$

with $A = (Z^TZ)$, $B = -N_{UZ^T}(\mathbf{\Psi})$, $E = (I - UU^T)\mathbf{\Psi}Z$ and $F = -Z + \mathbf{\Psi}^T U$. This requires to solve the Sylvester equation $X_U A - BB^T X_U = E - BF$ for X_U, that can be done in theory by using standard techniques [82], before computing X_Z from $X_Z = F - B^T X_U$. As a "proof of concept", a matrix $\mathbf{\Psi} \in \mathcal{M}_{l,m}$ with $m = 100$ and $l = 150$ is considered, with singular values chosen to be equally spaced in the interval $[1, 10]$. Three optimization algorithms detailed in [36] (gradient descent with fixed step, conjugate gradient descent, and Newton method) are implemented to find the best rank $r = 5$ approximation of $\mathbf{\Psi}$, with a random initialization. Convergence curves are plotted on Figure 3-2: linear and quadratic rates characteristic of respectively gradient and Newton methods are obtained. As expected from proposition 2.16, gradient descents globally converge to the truncated SVD, while Newton iterations may be attracted to a saddle point. Turning back to the DO time stepping with $\mathbf{\Psi} = \Psi^{n+1} = U^{n+1}(Z^{n+1})^T$ one uses the procedure described in algorithm 4 as an efficient method to update the truncated SVD of $\mathbf{\Psi}^{n+1}$ given the initial guess $\Psi^n = U^n(Z^n)^T$. If Δt is small enough, the method is expected to converge after a small number of iterations, while preserving the continuous evolution of the mode and coefficient matrix U and Z. In comparison with the use of geodesics, this method offers the benefit to ensure the accuracy of the reduced solution, while being less sensitive to the singularity of the matrix Z. Also, this method is a direct improvement of the DO time stepping (3.38), as one can see that one step of (3.38) coincides with the first step of the gradient descent (3.48) starting from the current value $U^n(Z^n)^T$ and with $\mu = 1$.

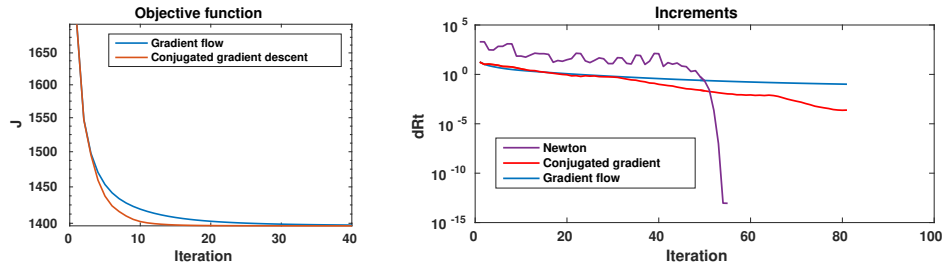


Figure 3-2: Convergence curves of optimization algorithms for minimizing the distance function J (equation (3.45)). Newton does not converge to the global minimum and hence is not represented on the left curve.

Algorithm 4 Gradient descent for updating a rank r_Ψ truncated SVD of $\mathbf{\Psi} = \mathbf{\Psi}_U \mathbf{\Psi}_Z^T$ with $\mathbf{\Psi}_U \in \mathcal{M}_{l,r_\Psi}$, $\mathbf{\Psi}_Z \in \mathcal{M}_{m,r_\Psi}$ and $r < r_\Psi << \min(l,m)$

1: Initialize a rank r guess $U_0 Z_0^T \simeq \mathbf{\Psi}$

2: So as to minimize $J(U,Z) = ||\mathbf{\Psi} - UZ^T||$ on $\mathscr{M}$, compute the gradient step

$$\begin{cases} Z_{k+1} = & Z_k - \mu \nabla J_U(U_k, Z_k) \\ U_{k+1} = & U_k - \mu \nabla J_Z(U_k, Z_k), \end{cases} \tag{3.48}$$

where μ is a constant set by the user and the gradient $(\nabla J_U, \nabla J_Z)$ of J on the manifold is given by [44]

$$\begin{cases} \nabla J_U(U,Z) = (-(I - UU^T)\mathbf{\Psi}_U[(\mathbf{\Psi}_Z)^T Z (Z^T Z)^{-1}] \\ \nabla J_Z(U,Z) = Z - \mathbf{\Psi}_Z[(\mathbf{\Psi}_U)^T U]), \end{cases} \tag{3.49}$$

brackets highlighting matrix products that render the computation efficient.

3: Orthonormalize the modes U_k (see section 3.3.5) after each iteration and repeat until convergence is achieved.

3.3.4 Increasing dynamically the rank of the approximation

In the SPDE (3.2), all realizations of the solution share the same initial value $\boldsymbol{\psi}(0, \boldsymbol{x}; \omega) = \boldsymbol{x}$. Hence the DO approximation coincides with the exact solution at time $t = 0$ and is given by the rank 1 decomposition $\Psi = UZ^T$ where U is a normalized column vector proportional to the discretization of the coordinate function $\boldsymbol{x}$, while Z is a column vector identically equal to the normalization factor. Obviously, $\boldsymbol{\psi}(t, \boldsymbol{x}; \omega)$ becomes random immediately after $t > 0$ and hence the rank of the DO solution must be modified dynamically to capture dominant stochastic subspaces that are forming throughout the time evolution of the solution. This is a common issue in model order reduction of stochastic PDEs.

Reducing the dimension r_Ψ of the DO stochastic subspace is straightforward: it is sufficient to truncate the SVD of the current DO solution $\Psi = UZ^T$, using for example algorithm 3, when the lowest singular value $\sigma_{r_\Psi}(\Psi) < \underline{\sigma}$ becomes lower than a threshold $\underline{\sigma}$ [123]. Increasing the stochastic dimension from r_Ψ to ${r_\Psi}' > r_\Psi$ is more involved, as ${r_\Psi}' - r_\Psi$ new dominant directions $\boldsymbol{u}_i$ supporting the decomposition (3.3) must be found. Sapsis and Lermusiaux [123] suggested to add modes aligned with the most sensitive directions of the operator $\mathcal{L}$ (i.e. with the gradient of $\mathcal{L}$ in the ambient space $\mathcal{M}_{l,m}$), but without the guarantee of tracking the best rank ${r_\Psi}'$ approximation at the next time step. An additional major difficulty lies in the issue of detecting when the dimension of the DO subspace must be increased. Sapsis and Lermusiaux [123] suggested to increase the rank r_Ψ when $\sigma_{r_\Psi}(\Psi) > \overline{\sigma}$ reaches another threshold $\overline{\sigma} > \underline{\sigma}$. Nevertheless, when the rank of the original solution remains equal to r_Ψ while $\sigma_{r_\Psi}(\Psi)$ increases, the dimensionality r_Ψ might be increased when it should not. Conversely, the rank of the original solution may truly increase while new singular values remain lower than the threshold $\overline{\sigma}$.

Theses issues can be solved by examining the component of the time derivative $\mathcal{L}(t, \Psi)$ that is normal to the manifold and neglected by the DO approximation (Figure 2-1). The value of this component is given by (see [44])

$$N(UZ^T) = (I - UU^T)\mathcal{L}(t, UZ^T)(I - Z(Z^TZ)^{-1}Z^T). \tag{3.50}$$

Since the singular value $\sigma_{r_\Psi+1}(\Psi^n + \Delta t\overline{L(t^n, \Psi^n)})$ after a step Δt is of typical magnitude of $\sigma_1(N(\Psi^n))\Delta t$ (see [75]), this first and other singular values of $N(UZ^T)$ are related to the speed at which the solution exits the rank r_Ψ matrix manifold $\mathscr{M}$. Therefore, a quantitative criterion that is expected to track accurately the rank of the true original solution is

$$\sigma_1(N(U^n(Z^n)^T))\Delta t > \overline{\sigma}. \tag{3.51}$$

A common value σ can be used for the threshold $\sigma = \overline{\sigma} = \underline{\sigma}$ to detect when the rank of the DO subspace must be decreased, hence the setting of this single σ provides a lower bound desired for the smallest singular value of the covariance matrix Z. Singular vectors of N contain the new dominant directions. They can be combined with a gradient descent similar to (3.48), so as to compute the rank ${r_\Psi}'$ (instead of r_Ψ) truncated SVD of $\boldsymbol{\Psi}^{n+1} = \Psi^n + \Delta t\overline{L(t^n, \Psi^n)}$, while preserving the smooth evolution of the first r_Ψ modes and coefficients (in contrast with the direct use of the algebraic algorithm 3). The procedure is summarized in algorithm 5.

Algorithm 5 Augmenting the rank of the DO solution

1: Compute $\Phi^n = U^n(Z^n)^T + \Delta L_U^n (L_Z^n)^T$ with $L_U^n \in \mathcal{M}_{l,r_{\overline{L}}}, L_Z^n \in \mathcal{M}_{m,r_{\overline{L}}}$ as in (3.44).
2: Compute the normal component (of rank at most $r_{\overline{L}}$).

$$N(U^n(Z^n)^T) = [(I - U^n(U^n)^T)L_U^n][(L_Z^n)^T(I - Z^n((Z^n)^T Z^n)^{-1}(Z^n)^T)].$$

3: Compute the rank $r_\Psi{}' - r_\Psi < r_{\overline{L}}$ truncated SVD $N_U^n(N_Z^n)^T$ of $N(U^n(Z^n)^T)$ by using algorithm 3.
4: Use the gradient descent (3.48) starting from the initialization values $U_0^n = [U^n N_U^n]$ and $[Z^n N_Z^n]$, so as to find the truncated SVD $U^{n+1}(Z^{n+1})^T$ of rank $r_\Psi{}' > r_\Psi$ of Φ^n.

3.3.5 Preserving the orthonormality of the mode matrix U

As highlighted in [147], an issue with time discretization, e.g. (3.38) or (3.48), is that in general, the matrix U^{n+1} obtained after a discrete time step does not exactly satisfy the orthogonality constraint ${U^{n+1}}^T U^{n+1} = I$. A numerical procedure must therefore be used to reduce the truncation errors committed by the discretization, even though the true trajectory $U(t)Z^T(t)$ on $\mathscr{M}$ and the DO equations (3.17) ensure and assume $U^T U = I$ at all instants. This procedure must be accurate as numerical orthonormalization may also introduce round off errors that can lead to significant error over large integration times. For example, standard and modified Gram Schmidt orthonormalization present numerical instabilities when UZ^T becomes close to being rank deficient (see [143]). For this reason, [146, 147] used the following procedure: compute the eigendecomposition of the Gram matrix $K = U^T U$,

$$PKP^T = \Sigma. \tag{3.52}$$

Then rotate and scale accordingly modes and coefficients by setting

$$\begin{cases} U \leftarrow UP\Sigma^{-1/2} \\ Z \leftarrow ZP\Sigma^{1/2}. \end{cases} \tag{3.53}$$

The eigenvalue problem (3.52) can be solved using Householder factorization which is known to be numerically stable in comparison with Gram Schmidt orthonormalization [143]. An issue is that this procedure may introduce permutations or sign changes, leading to artificial discontinuities in the time evolution of the mode and coefficient matrices U and Z. Figure 3-3 illustrates the problem by plotting the typical evolution of a coefficient of the matrix Z with this orthonormalization procedure. Even though sign checks alleviate the problem [147], they are a burden. Hence, to reinforce orthogonality between time steps and provide smooth evolutions for both U and Z (3.17), one can employ a gradient flow, as was done in the DO time-stepping (3.48). Reorthonormalization is then performed by finding an invertible matrix $A \in \mathcal{M}_{r,r}$ such that $(UA)^T(UA) = A^T K A = I$ and by setting $U \leftarrow UA$ and $Z \leftarrow ZA^{-T}$. Such matrix A actually solves the following optimization problem:

$$\min_{A \in \mathcal{M}_{r,r}} G(A) = \frac{1}{4}||A^T K A - I||^2.$$

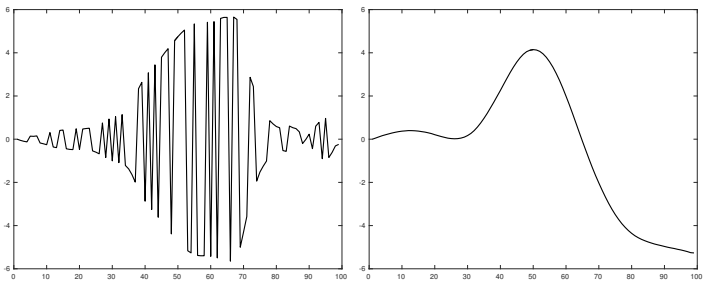


Figure 3-3: Time evolution of a coefficient of the matrix Z^n obtained by the time integration of (3.17). On the *left*, reorthonormalization of the matrix U^n is performed by solving the eigenvalue problem (3.53) while the gradient flow (3.54) has been used on the *right*. Eigenvalue decompositions introduce sign flips and permutations, that results in artificial discontinuities in the individual matrices U^n and Z^n.

Therefore, one can find a reorthonormalization matrix A close to the identity by solving the gradient flow

$$\frac{\mathrm{d}A}{\mathrm{d}s} = -\frac{\partial G}{\partial A} = -KA(A^T KA - I), \tag{3.54}$$

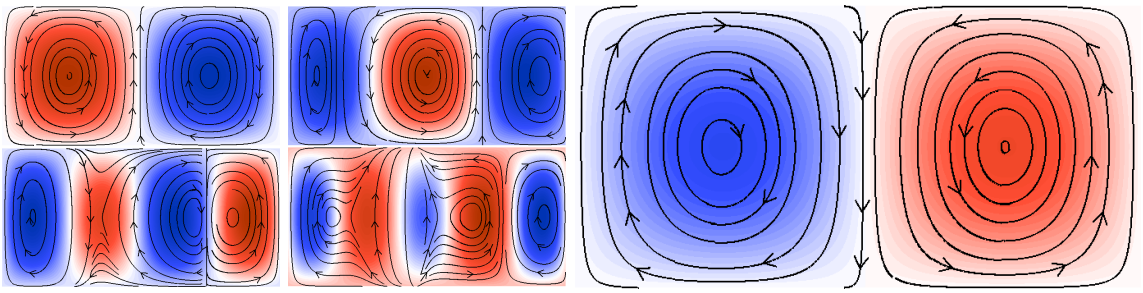
with the initial value $A(0) = I$. The inverse A^{-1} of A can be simultaneously tracked by solving the ODE

$$\frac{\mathrm{d}A^{-1}}{\mathrm{d}s} = -A^{-1}\frac{\mathrm{d}A}{\mathrm{d}s}A^{-1}.$$

The resulting numerical procedure is summarized in algorithm 6. Typically, one expects $A = I + O(||U^T U - I||)$ and hence both corrections $UA \simeq U$ and $ZA^{-T} \simeq Z$ will have an order of magnitude identical to the initial error, hence ensuring the smooth evolution of U and Z. Figure 3-3 shows the time evolution of a coefficient of the matrix Z using this method. Only a few number of Euler steps are necessary to obtain convergence, which makes the method efficient. The matrix $A \simeq I$ is well conditioned and algorithm 6 has small round off errors.

Algorithm 6 Reorthonormalization procedure

Define a tolerance parameter ϵ and a time step μ (typically $\mu \simeq 1$)
$K \leftarrow U^T U$
$A \leftarrow I$, $A^{-1} \leftarrow I$
while $||A_k^T K A_k - I||^2 > \epsilon$ **do**
    $\mathrm{d}A_k \leftarrow -KA_k(A_k^T KA_k - I)$
    $A_{k+1} \leftarrow A_k + \mu \mathrm{d}A_k$
    $A_{k+1}^{-1} \leftarrow A_k^{-1} - \mu A_k^{-1}(\mathrm{d}A_k)A_k^{-1}$
    $k \leftarrow k + 1$
end while
$U \leftarrow UA_k$
$Z \leftarrow ZA_k^{-T}$



(a) Spatial modes $\boldsymbol{v}_i(\boldsymbol{x})$ of the velocity

(b) Realization $\omega = 2\pi/10$ at $t = 10$

Figure 3-4: Streamlines of the double gyre flow with stochastic oscillation frequency. The intensity of the vorticity is displayed in background color.

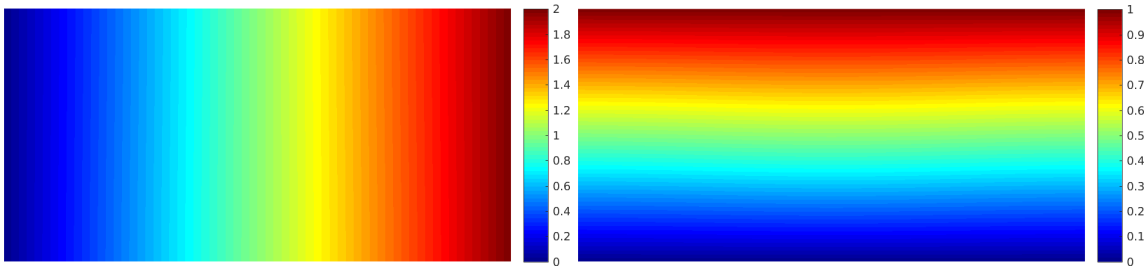
3.4 Numerical results

In this section we apply the numerical methodology described previously to two stochastic version of the double gyre flow and of the flow past a cylinder whose material transport was analyzed for individual deterministic cases in section 1.1.3. In both of the following examples, the threshold used for increasing the stochastic dimensionality (eqn. (3.51)) is set to $\sigma = 10^{-2}$ and the retraction used in the DO time-marching is the one described in section 3.3.3, computed with the gradient descent.

3.4.1 Stochastic double gyre flow

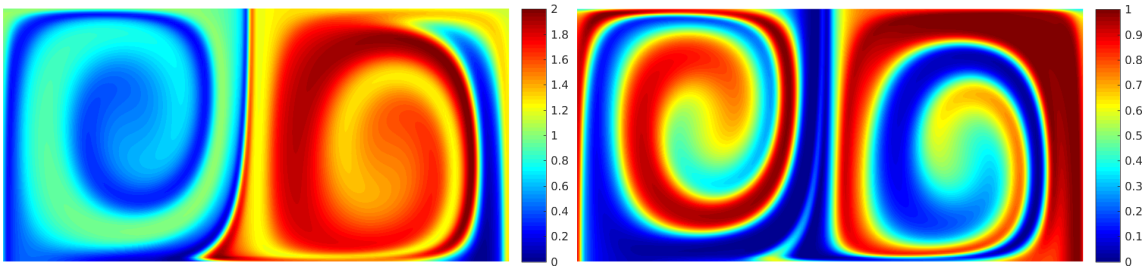
We consider again the benchmark double gyre flow of (1.7) in the setting of section 1.1.3. One can be interested how the Lagrangian motion of particles is impacted by the oscillation frequency ω, that is therefore the random parameter considered. In this numerical application, the stochastic PDE (3.2) is solved up to the time $t = 10$ and for ω uniformly distributed in $[\pi/10, 8\pi/10]$. The parameters are set according to $A = 0.1$, $\epsilon = 0.1$. For the DO computations, the spatial domain $[0, 2] \times [0, 1]$ is discretized using a 257×129 grid with $l_{bc} = 768$ boundary nodes, and the stochastic domain $[\pi/10, 8\pi/10]$ using $m = 10,000$ The velocity is decomposed onto 4 time-independent modes $\boldsymbol{v}_i(\boldsymbol{x})$ (Figure 3-4), and coefficients $\beta_i(t;\omega) =< \boldsymbol{v}_i(\boldsymbol{x}), \boldsymbol{v}(t,\boldsymbol{x};\omega) >$ are obtained by orthogonal projection. The initial value $\boldsymbol{\psi}(0,\boldsymbol{x};\omega) = \boldsymbol{x}$ of the solution is visible on Figure 3-5. The PDE (3.2) is solved directly with $\omega = 2\pi/10$ until $t = 10$ in order to validate the advection scheme selected in section 3.3.1. The result is confronted to the popular 5th order WENO scheme combined with the TVDRK3 time stepping [108] on Figure 3-6. Although the central scheme smears some small details on this example, both solutions obtained are fairly comparable, which demonstrates the broad applicability of this fully linear scheme for advection. This scheme is therefore used to solve the DO equations (3.17) as discussed in section 3.3.

The DO simulation is run with $r_\Psi = 20$ modes. For numerical stability, the 8th order Shapiro filter $\mathcal{F}^{(8)}$ (eqn. (3.32)) is applied at every step instead of 10. The first 4 modes obtained by the SVD truncation of the solution at $t = 10$ are displayed on Figure 3-7. This figure illustrates the ability of the DO solution to capture dominant modes that are far from being Fourier modes, and multi-modal distributions of the coefficients that are far from being Gaussian. Deterministic realizations, obtained by solving directly the transport PDE (3.2) for $\omega \in \{2\pi/10, 5\pi/10, 8\pi/10\}$, are compared against their DO solution on Figure 3-8. The figure shows an excellent agreement between results. The approximation of the solution by

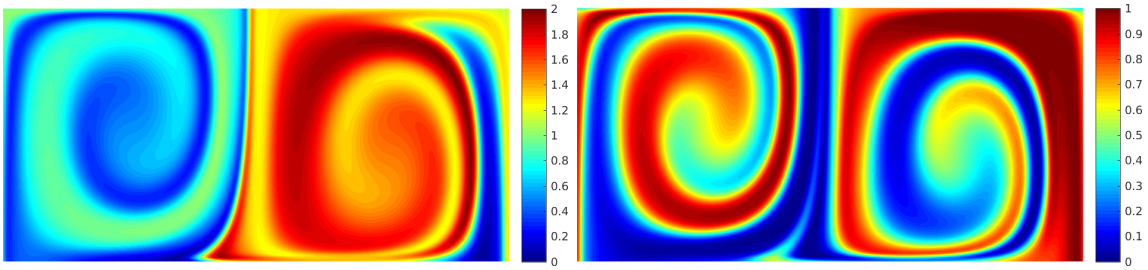


(a) x coordinate (b) y coordinate

Figure 3-5: Initial value $\boldsymbol{\psi}(0, \boldsymbol{x}; \omega) = \boldsymbol{x}$ of the advection eqn. (3.2)

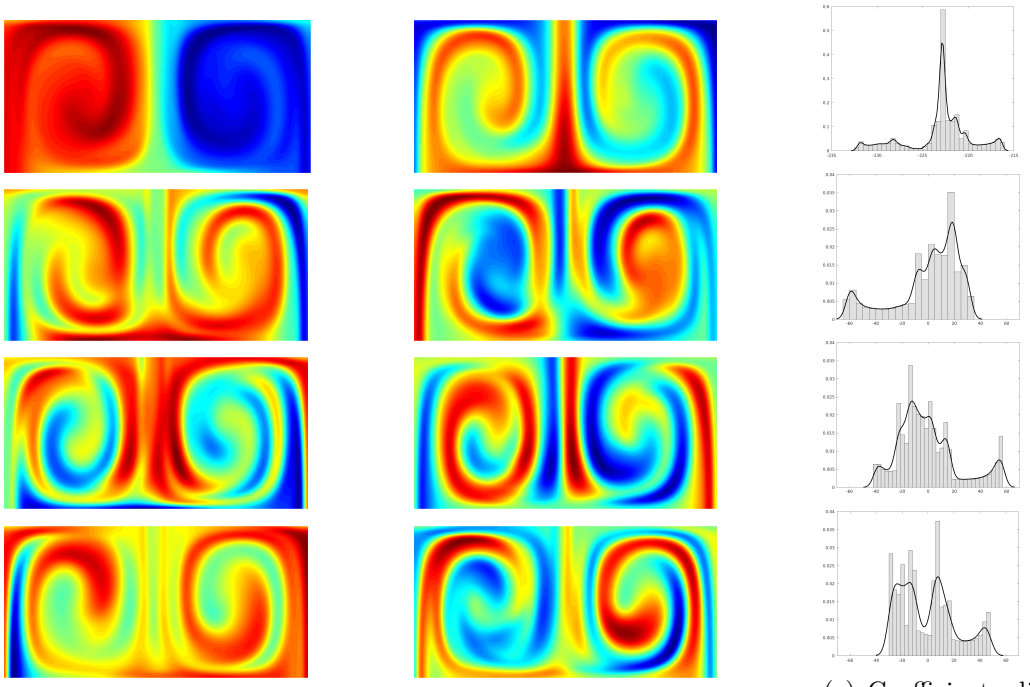


(a) 6th order central scheme - RK3 - Shapiro filter [10,3]



(b) 5th order WENO scheme - TVDRK3

Figure 3-6: Comparison between linear and non linear advection schemes for the direct resolution of (3.2) (without model order reduction) for the realization $\omega = 2\pi/10$.



(a) x coordinate (b) y coordinate (c) Coefficients distribution

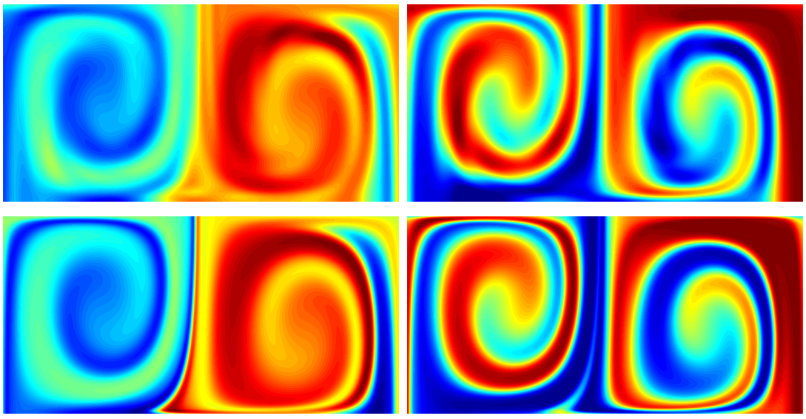
Figure 3-7: Dominant 4 first modes $\boldsymbol{\psi}_i$ and histogram of the corresponding distributions of the coefficients ζ_i of the solution $\boldsymbol{\psi}$ of the transport PDE (3.2) at $t = 10$.

20 modes incurs the loss of some sharp features, but the agreement between Monte-Carlo and DO realizations shows that the variability of the solution is well captured by the low dimensional time-dependent basis. The CPU time (estimated with Matlab) required by the DO simulation is $\text{CPU}_{\text{DO}} = 3530$ while each Monte-Carlo realization require $\text{CPU}_{\text{MC}} = 135$. The observed computational speed up is therefore of $\frac{\text{CPU}_{\text{MC}} \times m}{\text{CPU}_{\text{DO}}} \simeq 382$. This is coherent with the prediction given by the ratio $\frac{lm}{(l+m)r_\Psi - {r_\Psi}^2} \simeq 433$ between the dimension of the ambient space and the one of the manifold $\mathscr{M}$.

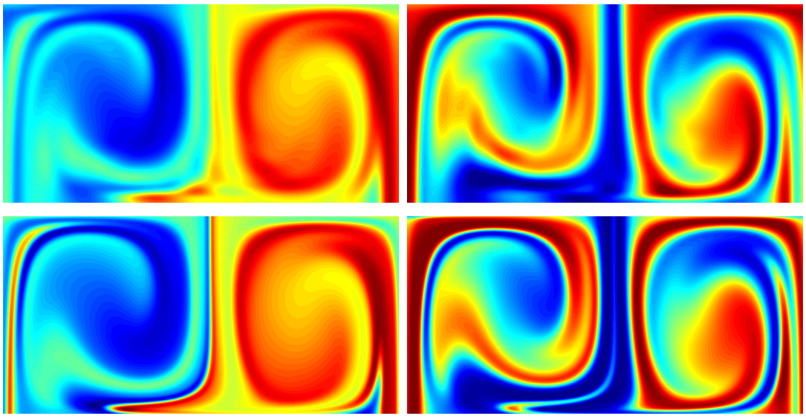
The mean and the variance of the solution are computed efficiently in a straightforward manner from the DO approximation and displayed on Figure 3-9. This figure highlights the mean behavior of the flow and the regions characterized by an increased level of uncertainty, which illustrates the applicability of the method for the study of Lagrangian motion under a stochastic velocity.

3.4.2 Stochastic flow past a cylinder

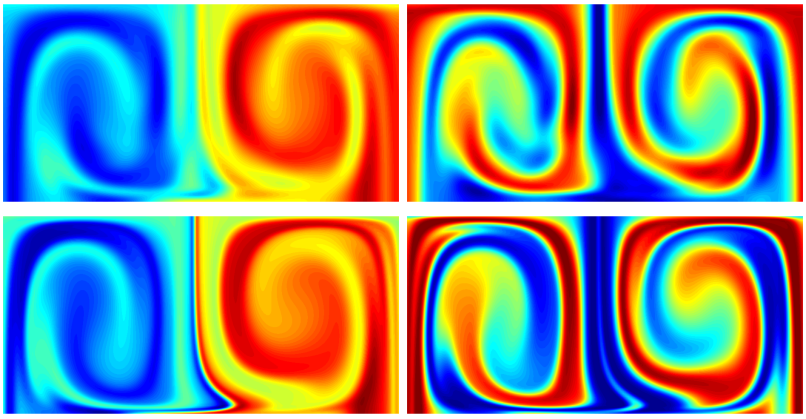
In this part, we consider again the Flow Past a Cylinder example of section 1.1.3. A random perturbation is used to initiate a stochastic flow $\boldsymbol{v}_0(t, \boldsymbol{x}; \omega)$ with periodic regime. $m = 10000$ realizations of this flow are computed by using a DO simulation with the numerical schemes described in [147]. The time window considered is [0,10], the time $t = 0$ being chosen once the periodic regime is established. The first four dominant modes of this flow along with one particular realization are displayed on Figure 3-11. The stochastic (forward) flow-map is computed analogously to the previous example with $r_\Psi = 20$ modes and the Shapiro filter $\mathcal{F}^{(8)}$ being applied at every time step. Figure 3-11 displays the values of the first 4



(a) $\omega = 2\pi/10$

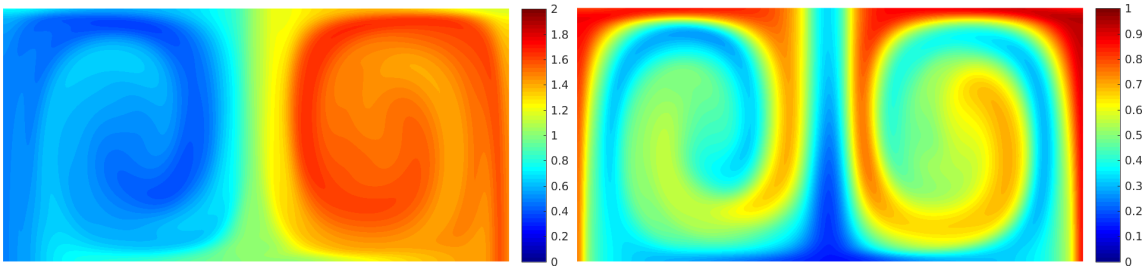


(b) $\omega = 5\pi/10$



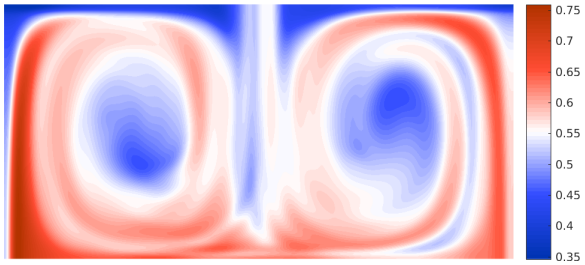
(c) $\omega = 8\pi/10$

Figure 3-8: Comparison between DO solution (*above*) versus direct Monte Carlo (*below*)



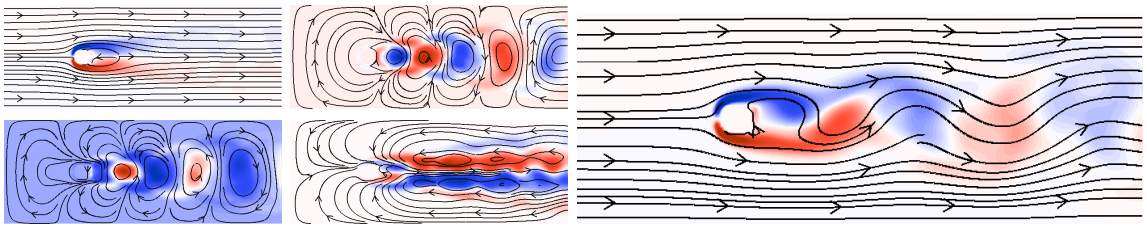
(a) x coordinate (b) y coordinate

(c) Mean $\mathbb{E}[\boldsymbol{\psi}]$ at $t = 10$



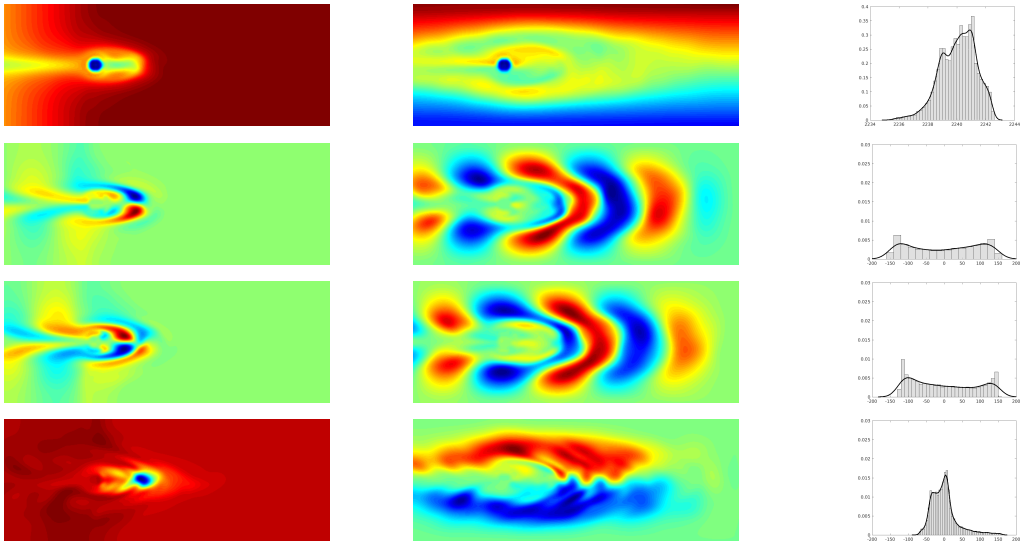
(d) Standard deviation distribution $\sigma_\Psi = \mathbb{E}[||\Psi - \mathbb{E}[\Psi]||^2]^{1/2}$ for the stochastic double gyre. Red highlights initial positions characterized with the most uncertainty.

Figure 3-9: Statistical quantities computed from the DO simulation



(a) Spatial modes $\boldsymbol{v}_i(0, \boldsymbol{x})$ of the velocity (b) A particular realization at $t = 0$

Figure 3-10: Streamlines of the stochastic flow past of a cylinder with stochastic initialization. The intensity of the vorticity is displayed in background color.



(a) x coordinate (b) y coordinate (c) Distribution

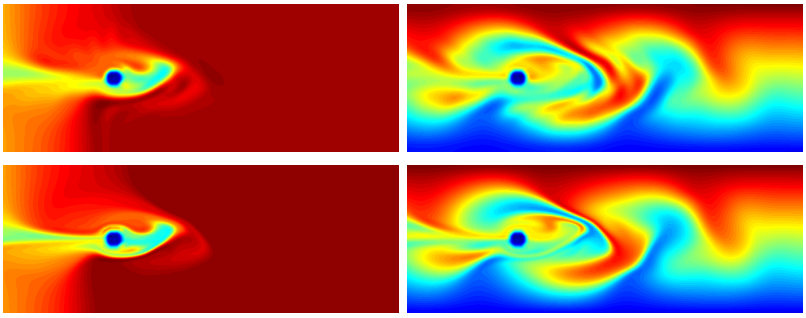
Figure 3-11: Dominant 4 first modes $\boldsymbol{\psi}_i$ and histogram of the corresponding distributions of the coefficients ζ_i of the solution $\boldsymbol{\psi}$ of the transport PDE (3.2) at $t = 10$.

dominant modes and the corresponding coefficient distributions of the resulting solution at time $t = 10$. Three particular realizations $\omega_1, \omega_2, \omega_3$ are evaluated directly and compared to the DO simulation on Figure 3-12. One still observe an excellent agreement between the Monte-Carlo realizations and the DO reconstructed solutions. Similarly as above, mean positions and variability of the resulting Lagrangian motion are plotted on Figure 3-13. Since particles may exit the domain, the value of $\boldsymbol{\psi}(10, \boldsymbol{x}; \omega)$ is the final position occupied by a particle initially located at $\boldsymbol{x}$ at time $t = 0$ if this particle does not leave the domain, or the position of where the particle left the domain otherwise. For this example, $l = 42848$ and $m = 10000$. The observed CPU times required for the DO simulation and one Monte-Carlo realization are respectively $\text{CPU}_{\text{DO}} = 940$ and $\text{CPU}_{\text{MC}} \simeq 32$ which yields an effective computational speed up of $\frac{\text{CPU}_{\text{MC}} \times m}{\text{CPU}_{\text{DO}}} \simeq 340$, still consistent with the prediction $\frac{lm}{(l+m)r_\Psi - {r_\Psi}^2} \simeq 405$.

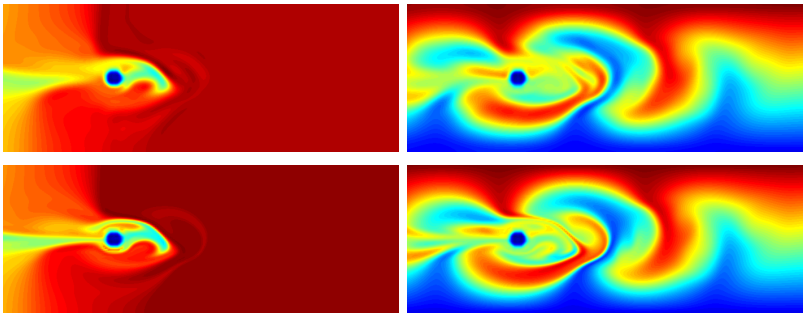
3.5 Conclusion and future works

The overall contribution of this work is twofold : we first provided a mathematical framework of *oblique projections* that allows to obtain the time derivative of implicit matrix maps and to derive convergent time matrix algorithms to compute them. Applying it to the fixed rank manifold, we obtained an error analysis of the DO method and new methodologies for its implementation. A future work could investigate whether this mathematical framework can be exported into the infinite dimensional setting. For example gradient descent and dynamical systems that achieve reinitializations have been used in level set methods when evolving signed distance functions[108].

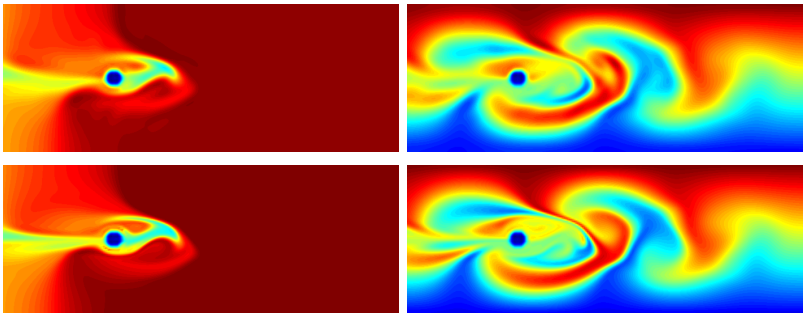
We then improved and reviewed the implementation of the Dynamically Orthogonal methodology by exploiting its relation to truncated Singular Value Decomposition. Its broad applicability to treat advection has been illustrated, offering a novel method for computing a large number of realizations of the flow map of an ODE with stochastic velocity. Fully



(a) ω_1

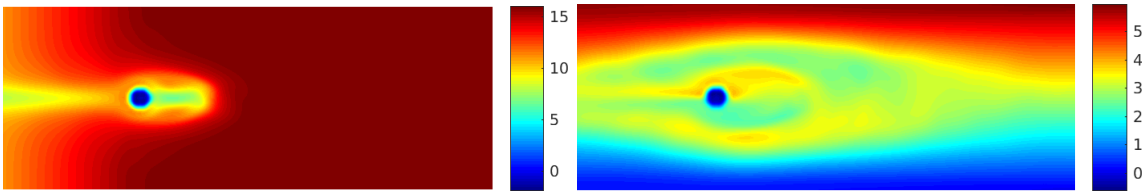


(b) ω_2



(c) ω_3

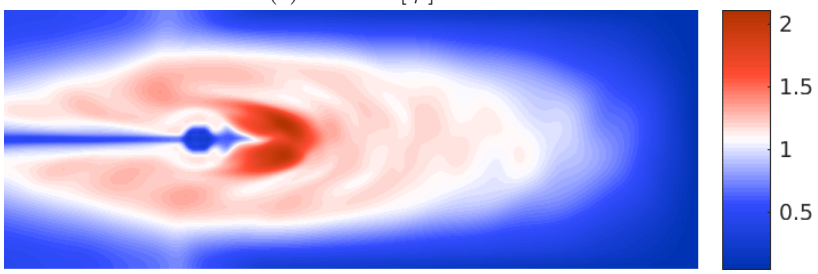
Figure 3-12: Comparison between DO solution (*above*) versus direct Monte Carlo realizations (*below*) for the flow past a cylinder



(a) x coordinate

(b) y coordinate

(c) Mean $\mathbb{E}[\psi]$ at $t = 10$



(d) Standard deviation distribution $\sigma_\Psi = \mathbb{E}[||\Psi - \mathbb{E}[\Psi]||^2]^{1/2}$ for the stochastic double gyre. Red highlights initial positions characterized with the most uncertainty.

Figure 3-13: Statistical quantities computed from the DO simulation for the flow past a cylinder example

linear advection scheme have been proven to be effective when integrated in linear model order reduction. One issue that must be acknowledged, still, is the fact that in general, advective processes are not well captured by low rank approximation [119]. Future works could investigate on finding what low-dimensional manifolds could better capture advective dynamics. One could think for example on wavelet expansions, and symmetry reductions, a challenge being that the inferred computational methodologies should remain at a moderate cost.

Regarding Lagrangian coherent structures, we believe this work opens directions towards the quantification of uncertainty in advection dominated systems. Our approximation tends to smear out sharp gradients (still in a minimized way thanks to the linear advection schemes) but preserve the overall behavior of the dynamics, and is expected to speed up the computation of averaged statistics such as mean and standard deviation. One will notice that this approach is not immediately connected to the LCSs methodologies presented in chapter 1, for example it yields flow maps and not FTLE realizations or coherent sets. Future research directions could focus on finding how relevant features of stochastic flow maps can be vizualized, and uncertainty of Lagrangian Structures be quantified. A challenge of great interest could be finding appropriate modelling of *shape statistics*, namely how to best "represent" the common features of a large number of realizations of shapes. The issue is mathematically not trivial due to the complexity of spaces of shapes [7], but one can expect it to open potential advances in this field.

Bibliography

[1] Advanced Lagrangian Predictions for Hazards Assessments (NSF-ALPHA). MSEAS Group Page. http://mseas.mit.edu/Research/NSF_ALPHA/index.html.

[2] Abatzoglou, T. The metric projection on C2 manifolds in Banach spaces. *Journal of Approximation Theory 26*, 3 (1979), 204–211.

[3] Absil, P., Trumpf, J., Mahony, R., and Andrews, B. All roads lead to Newton: Feasible second-order methods for equality-constrained optimization. Tech. rep., Citeseer, 2009.

[4] Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[5] Absil, P.-A., Mahony, R., and Trumpf, J. An extrinsic look at the Riemannian Hessian. In *Geometric Science of Information*. Springer, 2013, pp. 361–368.

[6] Absil, P.-A., and Malick, J. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization 22*, 1 (2012), 135–158.

[7] Allaire, G. *Shape optimization by the homogenization method*, vol. 146. Springer Science & Business Media, 2012.

[8] Allaire, G., Dapogny, C., Delgado, G., and Michailidis, G. Multi-phase structural optimization via a level set method. *ESAIM. Control, Optimisation and Calculus of Variations 20*, 2 (2014), 576.

[9] Allshouse, M. R., and Peacock, T. Lagrangian based methods for coherent structure detection. *Chaos: An Interdisciplinary Journal of Nonlinear Science 25*, 9 (2015), 097617.

[10] Allshouse, M. R., and Thiffeault, J.-L. Detecting coherent structures using braids. *Physica D: Nonlinear Phenomena 241*, 2 (2012), 95–105.

[11] Ambrosio, L. *Geometric evolution problems, distance function and viscosity solutions*. Springer, 2000.

[12] Ambrosio, L. Transport equation and Cauchy problem for non-smooth vector fields. In *Calculus of variations and nonlinear partial differential equations*. Springer, 2008, pp. 1–41.

[13] Babaee, H., and Sapsis, T. A minimization principle for the description of modes associated with finite-time instabilities. In *Proc. R. Soc. A* (2016), vol. 472, The Royal Society, p. 20150779.

[14] Bennett, A. *Lagrangian fluid dynamics*. Cambridge University Press, 2006.

[15] Bollt, E. M., and Santitissadeekorn, N. *Applied and Computational Measurable Dynamics*, vol. 18. SIAM, 2013.

[16] Boulanouar, M. L'opérateur d'advection. I. Existence d'un C0 semi-groupe. *Transport Theory and Statistical Physics 31*, 2 (2002), 169–176.

[17] Boyer, F. Trace theorems and spatial continuity properties for the solutions of the transport equation. *Differential and integral equations 18*, 8 (2005), 891–934.

[18] Bredon, G. E. *Topology and geometry*, vol. 139. Springer Science & Business Media, 2013.

[19] Brezis, H., Ciarlet, P. G., and Lions, J. L. *Analyse fonctionnelle: théorie et applications*, vol. 91. Dunod Paris, 1999.

[20] Brockett, R. W. Dynamical systems that sort lists, diagonalize matrices and solve linear programming problems. In *Decision and Control, 1988., Proceedings of the 27th IEEE Conference on* (1988), IEEE, IEEE, pp. 799–803.

[21] Brockett, R. W. Dynamical systems that learn subspaces. In *Mathematical System Theory*. Springer, 1991, pp. 579–592.

[22] Brunton, S. L., and Rowley, C. W. Fast computation of FTLE fields for unsteady flows: a comparison of methods. *Chaos 20*, 1 (2010), 017503.

[23] Budišić, M., and Mezić, I. Geometry of the ergodic quotient reveals coherent structures in flows. *Physica D: Nonlinear Phenomena 241*, 15 (2012), 1255–1269.

[24] Cannarsa, P., and Cardaliaguet, P. Representation of equilibrium solutions to the table problem of growing sandpiles. *Journal of the European Mathematical Society 6*, 4 (2004), 435–464.

[25] Chen, Y.-C., and Wheeler, L. Derivatives of the stretch and rotation tensors. *Journal of elasticity 32*, 3 (1993), 175–182.

[26] Constantine, P. G. *Active subspaces: Emerging ideas for dimension reduction in parameter studies*, vol. 2. SIAM, 2015.

[27] Cottet, G.-H., and Koumoutsakos, P. D. *Vortex methods: theory and practice*. Cambridge university press, 2000.

[28] Coulliette, C., Lekien, F., Paduan, J. D., Haller, G., and Marsden, J. E. Optimal pollution mitigation in Monterey Bay based on coastal radar data and nonlinear dynamics. *Environmental science & technology 41*, 18 (2007), 6562–6572.

[29] Crandall, M. G., Ishii, H., and Lions, P.-L. User's guide to viscosity solutions of second order partial differential equations. *Bulletin of the American Mathematical Society 27*, 1 (1992), 1–67.

[30] Dautray, R., and Lions, J.-L. *Mathematical Analysis and Numerical Methods for Science and Technology*. Springer Science Business Media, 2000.

[31] Dehaene, J. *Continuous-time matrix algorithms systolic algorithms and adaptive neural networks.* PhD thesis, Electrical Engineering Department (ESAT), K. U. Leuven, Belgium, 1995.

[32] Delfour, M. C., and Zolésio, J.-P. *Shapes and geometries: metrics, analysis, differential calculus, and optimization*, vol. 22. Siam, 2011.

[33] Diaconescu, E. P., and Laprise, R. Singular vectors in atmospheric sciences: A review. *Earth-Science Reviews 113*, 3-4 (2012), 161–175.

[34] DiPerna, R. J., and Lions, P.-L. Ordinary differential equations, transport theory and Sobolev spaces. *Inventiones mathematicae 98*, 3 (1989), 511–547.

[35] Durran, D. R. The third-order Adams-Bashforth method: An attractive alternative to leapfrog time differencing. *Monthly weather review 119*, 3 (1991), 702–720.

[36] Edelman, A., Arias, T. A., and Smith, S. T. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications 20*, 2 (1998), 303–353.

[37] El-Beltagy, M. A., Wafa, M. I., and Galal, O. H. Upwind Finite-Volume Solution of Stochastic Burgers' Equation. *AM 03*, 11 (2012), 1818–1825.

[38] Engquist, B., Lötstedt, P., and Sjögreen, B. Nonlinear filters for efficient shock computation. *Mathematics of Computation 52*, 186 (1989), 509–537.

[39] Farazmand, M., and Haller, G. Computing Lagrangian coherent structures from their variational theory. *Chaos 22*, 1 (2012), 1–12.

[40] Farazmand, M., and Haller, G. Attracting and repelling Lagrangian coherent structures from a single computation. *Chaos: An Interdisciplinary Journal of Nonlinear Science 23*, 2 (2013), 023101.

[41] Farazmand, M., and Sapsis, T. P. Dynamical indicators for the prediction of bursting phenomena in high-dimensional systems. *Phys. Rev. E 94* (Sep 2016), 032212.

[42] Farge, M. The continuous wavelet transform of two-dimensional turbulent flows. *Wavelets and their Applications* (1992), 275–302.

[43] Farrell, B. F., and Ioannou, P. J. Generalized stability theory. Part I: Autonomous operators. *Journal of the atmospheric sciences 53*, 14 (1996), 2025–2040.

[44] Feppon, F., and Lermusiaux, P. F. J. A geometric approach to dynamical model order reduction. *Submitted to SIAM Journal of Matrix Analysis* (2016).

[45] Feppon, F., and Lermusiaux, P. F. J. Dynamically Orthogonal Numerical Schemes for Efficient Stochastic Advection and Lagrangian Transport. *In preparation* (2016).

[46] Froyland, G. An analytic framework for identifying finite-time coherent sets in time-dependent dynamical systems. *Physica D: Nonlinear Phenomena 250* (2013), 1–19.

[47] Froyland, G. Dynamic isoperimetry and the geometry of Lagrangian coherent structures. *Nonlinearity 28*, 10 (2015), 3587.

[48] Froyland, G., Horenkamp, C., Rossi, V., and van Sebille, E. Studying an Agulhas ring's long-term pathway and decay with finite-time coherent sets. *Chaos: An Interdisciplinary Journal of Nonlinear Science 25*, 8 (2015), 083119.

[49] Froyland, G., Junge, O., and Koltai, P. Estimating long-term behavior of flows without trajectory integration: The infinitesimal generator approach. *SIAM Journal on Numerical Analysis 51*, 1 (2013), 223–247.

[50] Froyland, G., and Kwok, E. A dynamic Laplacian for identifying Lagrangian coherent structures on weighted Riemannian manifolds. *arXiv preprint arXiv:1610.01128* (2016).

[51] Froyland, G., Lloyd, S., and Santitissadeekorn, N. Coherent sets for nonautonomous dynamical systems. *Physica D: Nonlinear Phenomena 239*, 16 (2010), 1527–1541.

[52] Froyland, G., and Padberg, K. Almost-invariant and finite-time coherent sets: directionality, duration, and diffusion. In *Ergodic Theory, Open Dynamics, and Coherent Structures*. Springer, 2014, pp. 171–216.

[53] Froyland, G., Padberg, K., England, M. H., and Treguier, A. M. Detection of coherent oceanic structures via transfer operators. *Physical review letters 98*, 22 (2007), 224503.

[54] Froyland, G., Schwalb, M., Padberg, K., and Dellnitz, M. A transfer operator based numerical investigation of coherent structures in three-dimensional Southern Ocean circulation. *ratio 1* (2008), 1.

[55] Gilbarg, D., and Trudinger, N. S. *Elliptic partial differential equations of second order*. springer, 2015.

[56] Godunov, S. K. A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. *Matematicheskii Sbornik 89*, 3 (1959), 271–306.

[57] Goloviznin, V. M., Semenov, V. N., Korotkin, I. A., and Karabasov, S. A. A novel computational method for modelling stochastic advection in heterogeneous media. *Transport in porous media 66*, 3 (2007), 439–456.

[58] Golub, G. H., and Van Loan, C. F. Matrix computations. 1996. *Johns Hopkins University, Press, Baltimore, MD, USA* (1996), 374–426.

[59] Golub, G. H., and Van Loan, C. F. *Matrix computations*, vol. 3. JHU Press, 2012.

[60] Griffa, A., Kirwan Jr, A., Mariano, A. J., Özgökmen, T., and Rossby, H. T. *Lagrangian analysis and prediction of coastal and ocean dynamics*. Cambridge University Press, 2007.

[61] Hadjighasem, A., and Haller, G. Geodesic Transport Barriers in Jupiters Atmosphere: A Video-Based Analysis. *SIAM Rev. 58*, 1 (Jan. 2016), 69–89.

[62] Haier, E., Norsett, S., and Wanner, G. Solving Ordinary Differential Equations I, Nonstiff Problems. *Section III 8* (2000).

[63] Haley Jr, P. J., and Lermusiaux, P. F. Multiscale two-way embedding schemes for free-surface primitive equations in the "multidisciplinary simulation, estimation and assimilation system". *Ocean dynamics 60*, 6 (2010), 1497–1537.

[64] Haller, G. Lagrangian coherent structures from approximate velocity data. *Physics of Fluids (1994-present) 14*, 6 (2002), 1851–1861.

[65] Haller, G. A variational theory of hyperbolic Lagrangian coherent structures. *Physica D: Nonlinear Phenomena 240*, 7 (2011), 574–598.

[66] Haller, G. Lagrangian coherent structures. *Annual Review of Fluid Mechanics 47* (2015), 137–162.

[67] Haller, G. Dynamic rotation and stretch tensors from a dynamic polar decomposition. *Journal of the Mechanics and Physics of Solids 86* (2016), 70–93.

[68] Haller, G., and Beron-Vera, F. J. Geodesic theory of transport barriers in two-dimensional flows. *Physica D: Nonlinear Phenomena 241*, 20 (2012), 1680–1702.

[69] Haller, G., Hadjighasem, A., Farazmand, M., and Huhn, F. Defining coherent vortices objectively from the vorticity. *Journal of Fluid Mechanics 795* (2016), 136–173.

[70] Haller, G., and Sapsis, T. Lagrangian coherent structures and the smallest finite-time Lyapunov exponent. *Chaos: An Interdisciplinary Journal of Nonlinear Science 21*, 2 (2011), 023115.

[71] Haller, G., and Yuan, G. Lagrangian coherent structures and mixing in two-dimensional turbulence. *Physica D: Nonlinear Phenomena 147*, 3 (2000), 352–370.

[72] Hartman, P. *Ordinary Differential Equations*, second ed. Society for Industrial and Applied Mathematics, 2002.

[73] Holmes, P., Lumley, J. L., and Berkooz, G. *Turbulence, coherent structures, dynamical systems and symmetry*. Cambridge university press, 1998.

[74] Horn, R. A., and Johnson, C. R. *Topics in matrix analysis*. Cambridge Univ. Press Cambridge etc, 1991.

[75] Horn, R. A., and Johnson, C. R. *Matrix Analysis*. Cambridge University Press (CUP), 2009.

[76] Hoskins, B., Buizza, R., and Badger, J. The nature of singular vector growth and structure. *Quarterly Journal of the Royal Meteorological Society 126*, 566 (2000), 1565–1580.

[77] Jardak, M., Su, C.-H., and Karniadakis, G. E. Spectral polynomial chaos solutions of the stochastic advection equation. *Journal of Scientific Computing 17*, 1-4 (2002), 319–338.

[78] John, F. Rotation and strain. *Communications on Pure and Applied Mathematics 14*, 3 (1961), 391–413.

[79] Jost, J. *Riemannian geometry and geometric analysis*. Springer Science & Business Media, 2008.

[80] Kalnay, E. *Atmospheric modeling, data assimilation and predictability*. Cambridge university press, 2003.

[81] Karrasch, D. Attracting Lagrangian coherent structures on Riemannian manifolds. *Chaos: An Interdisciplinary Journal of Nonlinear Science 25*, 8 (2015), 087411.

[82] Kirrinnis, P. Fast algorithms for the Sylvester equation AX- XBT= C. *Theoretical Computer Science 259*, 1 (2001), 623–638.

[83] Koch, O., and Lubich, C. Dynamical low-rank approximation. *SIAM Journal on Matrix Analysis and Applications 29*, 2 (2007), 434–454.

[84] Kutz, J. N. *Data-driven modeling & scientific computation: methods for complex systems & big data*. OUP Oxford, 2013.

[85] Lafon, F., and Osher, S. High order filtering methods for approximating hyperbolic systems of conservation laws. *Journal of Computational Physics 96*, 1 (1991), 110–142.

[86] Lancaster, P. On eigenvalues of matrices dependent on a parameter. *Numerische Mathematik 6*, 1 (1964), 377–387.

[87] Lasota, A., and Mackey, M. C. *Chaos, fractals, and noise: stochastic aspects of dynamics*, vol. 97. Springer Science & Business Media, 2013.

[88] Lekien, F., Coulliette, C., Mariano, A. J., Ryan, E. H., Shay, L. K., Haller, G., and Marsden, J. Pollution release tied to invariant manifolds: A case study for the coast of Florida. *Physica D: Nonlinear Phenomena 210*, 1 (2005), 1–20.

[89] Lermusiaux, P. F. J. *Error subspace data assimilation methods for ocean field estimation: theory, validation and applications*. Harvard University, 1997.

[90] Lermusiaux, P. F. J. Estimation and Study of Mesoscale Variability in the Strait of Sicily. *Dynamics of Atmospheres and Oceans 29*, 2 (1999), 255–303.

[91] Lermusiaux, P. F. J. Evolving the subspace of the three-dimensional multiscale ocean variability: Massachusetts Bay. *Journal of Marine Systems 29*, 1 (2001), 385–422.

[92] Lermusiaux, P. F. J. Uncertainty estimation and prediction for interdisciplinary ocean dynamics. *Journal of Computational Physics 217*, 1 (2006), 176–199.

[93] Lermusiaux, P. F. J. Adaptive modeling, adaptive data assimilation and adaptive sampling. *Physica D: Nonlinear Phenomena 230*, 1 (2007), 172–196.

[94] Lermusiaux, P. F. J., Chiu, C.-S., Gawarkiewicz, G. G., Abbot, P., Robinson, A. R., Miller, R. N., Haley, Jr, P. J., Leslie, W. G., Majumdar, S. J., Pang, A., and Lekien, F. Quantifying Uncertainties in Ocean Predictions. *Oceanography 19*, 1 (2006), 92–105.

[95] Lermusiaux, P. F. J., and Lekien, F. Dynamics and Lagrangian coherent structures in the ocean and their uncertainty. In *Extended Abstract in report of the Dynamical System Methods in Fluid Dynamics Oberwolfach Workshop* (Germany, July 31st - August 6th 2005), J. E. Marsden and J. Scheurle, Eds., Mathematisches Forschungsinstitut Oberwolfach, p. 2.

[96] Lermusiaux, P. F. J., Lolla, T., Haley, Jr., P. J., Yigit, K., Ueckermann, M. P., Sondergaard, T., and Leslie, W. G. Science of Autonomy: Time-Optimal Path Planning and Adaptive Sampling for Swarms of Ocean Vehicles. In *Springer Handbook of Ocean Engineering: Autonomous Ocean Vehicles, Subsystems and Control*, T. Curtin, Ed. Springer, 2016, ch. 21.

[97] Leung, S. An Eulerian approach for computing the finite time Lyapunov exponent. *Journal of computational physics 230*, 9 (2011), 3500–3524.

[98] Leung, S. The backward phase flow method for the Eulerian finite time Lyapunov exponent computations. *Chaos: An Interdisciplinary Journal of Nonlinear Science 23*, 4 (2013), 043132.

[99] Loeve, M. Probability theory, vol. ii. *Graduate texts in mathematics 46* (1978), 0–387.

[100] Lolla, S. V. T. *Path planning and adaptive sampling in the coastal ocean.* PhD thesis, Massachusetts Institute of Technology, 2016.

[101] Mishra, B., Meyer, G., Bonnabel, S., and Sepulchre, R. Fixed-rank matrix factorizations and Riemannian low-rank optimization. *Computational Statistics 29*, 3-4 (Nov. 2014), 591–621.

[102] Moore, A. M., Arango, H. G., Di Lorenzo, E., Cornuelle, B. D., Miller, A. J., and Neilson, D. J. A comprehensive ocean prediction and analysis system based on the tangent linear and adjoint of a regional ocean model. *Ocean Modelling 7*, 1 (2004), 227–258.

[103] Musharbash, E., Nobile, F., and Zhou, T. Error Analysis of the Dynamically Orthogonal Approximation of Time Dependent Random PDEs. *SIAM Journal on Scientific Computing 37*, 2 (2015), A776–A810.

[104] Oettinger, D., Blazevski, D., and Haller, G. Global variational approach to elliptic transport barriers in three dimensions. *Chaos: An Interdisciplinary Journal of Nonlinear Science 26*, 3 (2016), 033114.

[105] Olascoaga, M., Beron-Vera, F., Brand, L., and Kocak, H. Tracing the early development of harmful algal blooms on the West Florida Shelf with the aid of

Lagrangian coherent structures. *Journal of Geophysical Research: Oceans 113*, C12 (2008).

[106] Onken, R., Robinson, A. R., Lermusiaux, P. F. J., Haley, P. J., and Anderson, L. A. Data-driven simulations of synoptic circulation and transports in the Tunisia-Sardinia-Sicily region. *Journal of Geophysical Research: Oceans 108*, C9 (2003).

[107] Onu, K., Huhn, F., and Haller, G. An Algorithmic Introduction to Lagrangian Coherent Structures, 2014.

[108] Osher, S., and Fedkiw, R. *Level set methods and dynamic implicit surfaces*, vol. 153. Springer Science & Business Media, 2006.

[109] Osher, S., and Shu, C.-W. High-order essentially nonoscillatory schemes for Hamilton-Jacobi equations. *SIAM Journal on numerical analysis 28*, 4 (1991), 907–922.

[110] Osnes, H., and Langtangen, H. P. A study of some finite difference schemes for a unidirectional stochastic transport equation. *SIAM Journal on Scientific Computing 19*, 3 (1998), 799–812.

[111] Palmer, T., Gelaro, R., Barkmeijer, J., and Buizza, R. Singular vectors, metrics, and adaptive observations. *Journal of the Atmospheric Sciences 55*, 4 (1998), 633–653.

[112] Peacock, T., and Haller, G. Lagrangian coherent structures: The hidden skeleton of fluid flows. *Physics Today 66*, 2 (2013), 41–47.

[113] Pitaval, R.-A., Dai, W., and Tirkkonen, O. Convergence of Gradient Descent for Low-Rank Matrix Approximation. *IEEE Transactions on Information Theory 61*, 8 (2015), 4451–4457.

[114] Piziak, R., and Odell, P. Full rank factorization of matrices. *Mathematics magazine 72*, 3 (1999), 193–201.

[115] Qiu, J., and Shu, C.-W. On the Construction, Comparison, and Local Characteristic Decomposition for High-Order Central WENO Schemes. *Journal of Computational Physics 183*, 1 (Nov. 2002), 187–209.

[116] Quarteroni, A., and Rozza, G. *Reduced Order Methods for Modeling and Computational Reduction*, vol. 9. Springer, 2014.

[117] Renard, D. Introduction à la géométrie différentielle. Cours de l' École Polytechnique, 2016.

[118] Reshetnyak, Y. G. *Stability theorems in geometry and analysis*, vol. 304. Springer Science & Business Media, 2013.

[119] Rowley, C. W., and Marsden, J. E. Reconstruction equations and the karhunen–loève expansion for systems with symmetry. *Physica D: Nonlinear Phenomena 142*, 1 (2000), 1–19.

[120] Salençon, J. *Mécanique des milieux continus: Concepts généraux*, vol. 1. Editions Ecole Polytechnique, 2005.

[121] Samelson, R. M., and Wiggins, S. *Lagrangian transport in geophysical jets and waves: The dynamical systems approach*, vol. 31. Springer Science & Business Media, 2006.

[122] Sapsis, T. P. *Dynamically Orthogonal Field Equations for Stochastic Fluid Flows and Particle Dynamics*. PhD thesis, Massachusetts Institute of Technology, Department of Mechanical Engineering, Cambridge, MA, February 2011.

[123] Sapsis, T. P., and Lermusiaux, P. F. Dynamical criteria for the evolution of the stochastic dimensionality in flows with uncertainty. *Physica D: Nonlinear Phenomena 241*, 1 (2012), 60–76.

[124] Sapsis, T. P., and Lermusiaux, P. F. J. Dynamically orthogonal field equations for continuous stochastic dynamical systems. *Physica D: Nonlinear Phenomena 238*, 23–24 (Dec. 2009), 2347–2360.

[125] Sapsis, T. P., Ueckermann, M. P., and Lermusiaux, P. F. J. Global analysis of Navier–Stokes and Boussinesq stochastic flows using dynamical orthogonality. *Journal of Fluid Mechanics 734* (2013), 83–113.

[126] Schilders, W. H., Van der Vorst, H. A., and Rommes, J. *Model order reduction: theory, research aspects and applications*, vol. 13. Springer, 2008.

[127] Schmidt, B. Conditions on a connection to be a metric connection. *Communications in Mathematical Physics 29*, 1 (1973), 55–59.

[128] Serra, M., and Haller, G. Objective Eulerian Coherent Structures. *arXiv preprint arXiv:1512.02112* (2015).

[129] Shadden, S. C., Lekien, F., and Marsden, J. E. Definition and properties of Lagrangian coherent structures from finite-time Lyapunov exponents in two-dimensional aperiodic flows. *Physica D: Nonlinear Phenomena 212*, 3 (2005), 271–304.

[130] Shalit, U., Weinshall, D., and Chechik, G. Online learning in the embedded manifold of low-rank matrices. *The Journal of Machine Learning Research 13*, 1 (2012), 429–458.

[131] Shapiro, R. Smoothing, filtering, and boundary effects. *Rev. Geophys. 8*, 2 (1970), 359.

[132] Shapiro, R. The use of linear filtering as a parameterization of atmospheric diffusion. *Journal of the Atmospheric Sciences 28*, 4 (1971), 523–531.

[133] Shapiro, R. Linear filtering. *Mathematics of Computation 29*, 132 (1975), 1094–1094.

[134] Shu, C.-W., and Osher, S. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *Journal of Computational Physics 77*, 2 (1988), 439–471.

[135] Simon, L., et al. Lectures on geometric measure theory. In *Proceedings of the Centre for Mathematical Analysis* (1983), vol. 3, The Australian National University, Mathematical Sciences Institute, Centre for Mathematics & its Applications.

[136] Sjögreen, B. High order centered difference methods for the compressible Navier-Stokes equations. *Journal of Computational Physics 117*, 1 (1995), 67–78.

[137] Small, C. G. *The statistical theory of shape*. Springer Science & Business Media, 2012.

[138] Smith, S. Dynamical systems that perform the singular value decomposition. *Systems & Control Letters 16*, 5 (1991), 319–327.

[139] Spivak, M. A comprehensive introduction to differential geometry. *Publish or Perish Inc 3* (1973).

[140] Subramani, D. N., and Lermusiaux, P. F. J. Energy-optimal path planning by stochastic dynamically orthogonal level-set optimization. *Ocean Modelling 100* (2016), 57–77.

[141] Tang, D., Schwartz, F. W., and Smith, L. Stochastic modeling of mass transport in a random velocity field. *Water Resources Research 18*, 2 (1982), 231–244.

[142] Tang, W., Chan, P. W., and Haller, G. Accurate extraction of Lagrangian coherent structures over finite domains with application to flight data analysis over Hong Kong International Airport. *Chaos: An Interdisciplinary Journal of Nonlinear Science 20*, 1 (2010), 017502.

[143] Trefethen, L. N., and Bau III, D. *Numerical linear algebra*, vol. 50. Siam, 1997.

[144] Truesdell, C., and Noll, W. The non-linear field theories of mechanics. *Encyclopedia of Physics 3* (1965), 3.

[145] Tryoen, J., Maître, O. L., Ndjinga, M., and Ern, A. Intrusive Galerkin methods with upwinding for uncertain nonlinear hyperbolic systems. *Journal of Computational Physics 229*, 18 (2010), 6485–6511.

[146] Ueckermann, M. P., Lermusiaux, P. F. J., and Sapsis, T. P. Numerical Schemes and Studies for Dynamically Orthogonal Equations of Stochastic Fluid and Ocean Flows. MSEAS Report 11, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, 2011.

[147] Ueckermann, M. P., Lermusiaux, P. F. J., and Sapsis, T. P. Numerical schemes for dynamically orthogonal equations of stochastic fluid and ocean flows. *Journal of Computational Physics 233* (Jan. 2013), 272–294.

[148] Vandereycken, B. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization 23*, 2 (2013), 1214–1236.

[149] Wan, X., Xiu, D., and Karniadakis, G. E. Stochastic solutions for the two-dimensional advection-diffusion equation. *SIAM Journal on Scientific computing 26*, 2 (2004), 578–590.

[150] Wei, K., Cai, J.-F., Chan, T. F., and Leung, S. Guarantees of Riemannian Optimization for Low Rank Matrix Completion. *arXiv preprint arXiv:1603.06610* (2016).

[151] Willcox, K., and Megretski, A. Fourier series for accurate, stable, reduced-order models in large-scale linear applications. *SIAM Journal on Scientific Computing 26*, 3 (2005), 944–962.

[152] Williams, P. D. Achieving seventh-order amplitude accuracy in leapfrog integrations. *Monthly Weather Review 141*, 9 (2013), 3037–3051.

[153] Wulbert, D. Continuity of metric projections. *Transactions of the American Mathematical Society 134*, 2 (1968), 335–341.

[154] Xiu, D., and Karniadakis, G. E. The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM journal on scientific computing 24*, 2 (2002), 619–644.