UCLouvain

epl

École polytechnique de Louvain

Bounds on the geodesic distances on the Stiefel manifold

Author: **Simon MATAIGNE**
Supervisors: **Pierre-Antoine ABSIL, Nina MIOLANE**
Reader: **Estelle MASSART**
Academic year 2022–2023
Master [120] in Mathematical Engineering

Abstract

The past few years have seen the emergence of many applications for statistics on manifolds. The manifold-based statistical tools often rely on the notion of mean of two points, i.e., the midpoint of a minimizing geodesic between these two points. An important manifold for which we cannot guarantee the computation of a minimizing geodesic is the Stiefel manifold, which is the set of orthogonal p-frames in $\mathbb{R}^n$. The goal of this master's thesis is to provide new insights into the Stiefel manifold to enhance the performances of minimizing geodesic computation algorithms. To this aim, we provide new results on the bilipschitz equivalence of a previously proposed one-parameter family of Riemannian metrics. We also show that the geodesic distances induced by this family of Riemannian metrics are equivalent to the easy-to-compute Frobenius distance. Subsequently, we obtain tight bounds between the geodesic and the Frobenius distances. Many operations on the Stiefel manifold, and in particular the computation of a minimizing geodesic, require the computation of the matrix exponential of skew-symmetric matrices. To conclude this work, we present the SkewLinearAlgebra.jl package, a Julia library specialized in the efficient computation of the eigenvalue decomposition and the exponential of skew-symmetric matrices.

Acknowledgments

This master's thesis was supervised by Pierre-Antoine Absil, Professor at the Department of Mathematical Engineering at UCLouvain and Nina Miolane, Assistant Professor of Electrical and Computer Engineering at UC Santa Barbara. Their interest for the geodesic endpoint problem initiated this work and I would like to thank them for their support.

In particular, I would like to thank Pierre-Antoine Absil for the frequent discussions we had about this work. His observations and remarks were decisive and were at the basis of many statements. The complete freedom that was given to me was invaluable. Thank you also for the many opportunities and encounters that were granted to me.

I would also like to thank Nina Miolane, co-supervisor of this master's thesis, for her important feedbacks and for her participation in the communication of this work.

Thank you also to the group of students of the local a.025 at the Euler building. This studious but warm assembly provided a productive framework.

Finally, I wanted to pay a special tribute to Bernadette Petit who passed on the passion for mathematics to many students through her charism and exigence.

Contents

Notations

$\mathbb{R}, \mathbb{C}$	The real and complex numbers
$\mathbb{R}_+$	The positive reals
$\mathbb{R}^{n\times p}$	The real matrices of size $n \times p$
$\mathbf{A}^T, \mathbf{A}^*$	Transpose and conjugate transpose of $\mathbf{A}$
$\mathrm{Tr}(\mathbf{A}), \det(\mathbf{A})$	Trace and determinant of $\mathbf{A}$
$\lVert\mathbf{A}\rVert_{\mathrm{F}}$	$:= \sqrt{\mathrm{Tr}(\mathbf{A}^T\mathbf{A})},$ Frobenius norm
$\lVert\mathbf{A}\rVert_2$	$:= \max\{\sigma \mid \sigma \text{ is a singular value of } \mathbf{A}\},$ spectral norm
$\mathrm{Sym}(n)$	$:= \{\mathbf{A} \in \mathbb{R}^{n\times n} \mid \mathbf{A} = \mathbf{A}^T\}$
$\mathrm{sym}(\mathbf{A})$	$:= \dfrac{\mathbf{A}+\mathbf{A}^T}{2}$
$\mathrm{Skew}(n)$	$:= \{\mathbf{A} \in \mathbb{R}^{n\times n} \mid \mathbf{A} = -\mathbf{A}^T\}$
$\mathrm{skew}(\mathbf{A})$	$:= \dfrac{\mathbf{A}-\mathbf{A}^T}{2}$
$\mathbf{I}_n$	$:= n \times n$ identity matrix
$\mathbf{I}_{n\times p}$	$:= \begin{bmatrix} \mathbf{I}_p \\ \mathbf{0}_{(n-p)\times p} \end{bmatrix}$
$\mathrm{O}(n)$	$:= \{\mathbf{Q} \in \mathbb{R}^{n\times n} \mid \mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}_n\}$
$\mathrm{SO}(n)$	$:= \{\mathbf{Q} \in \mathbb{R}^{n\times n} \mid \mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}_n \text{ and } \det(\mathbf{Q}) = 1\}$
$\mathrm{St}(n,p)$	$:= \{\mathbf{U} \in \mathbb{R}^{n\times p} \mid \mathbf{U}^T\mathbf{U} = \mathbf{I}_p\}$
$T_{\mathbf{U}}\mathrm{St}(n,p)$	Tangent space at $\mathbf{U} \in \mathrm{St}(n,p)$
$N_{\mathbf{U}}\mathrm{St}(n,p)$	Normal space at $\mathbf{U} \in \mathrm{St}(n,p)$
$\mathbf{U}_\perp$	Orthogonal completion of $\mathbf{U} \in \mathrm{St}(n,p)$
$\dfrac{\mathrm{d}f}{\mathrm{d}t}, \dot{f}$	Time derivative of a function $f(t)$
$\langle\cdot,\cdot\rangle,\ g(\cdot,\cdot)$	Riemannian metric
$\lVert\cdot\rVert$	Norm operator
$\mathrm{Exp}_{\mathbf{U}}(\boldsymbol{\Delta})$	Exponential map at $\mathbf{U}$ with velocity $\boldsymbol{\Delta}$
$\mathrm{Log}_{\mathbf{U}}(\widetilde{\mathbf{U}})$	Logarithmic map at $\mathbf{U}$ evaluated in $\widetilde{\mathbf{U}}$
$d(\mathbf{U},\widetilde{\mathbf{U}})$	Distance function between $\mathbf{U}$ and $\widetilde{\mathbf{U}}$
γ	Curve
$l(\gamma)$	Length of a curve γ
$\exp(\mathbf{A}),\ e^{\mathbf{A}}$	Matrix exponential of $\mathbf{A} \in \mathbb{R}^{n\times n}$

Introduction

Non-Euclidean geometry has proved itself one of the most prolific branch of mathematics since it revolutionized physics in the 20th century. For its part, matrix theory has nothing to be ashamed of in terms of impact on science and society. From numerical simulations to machine learning, computers have revealed the striking efficiency of linear algebra. The Stiefel manifold, which is the focus of this master's thesis, lies at the intersection of these two foreground branches of mathematics. The Stiefel manifold is the set of orthogonal p-frames in $\mathbb{R}^n$. As a manifold, it inherits from the geometric tools and theoretical notions of non-Euclidean geometry. As a set of matrices, it benefits from the efficient tools of linear algebra and matrix theory.

At this stage, we find it interesting to recall the origin of the word "manifold". It comes from the German word "mannigfaltigkeit" introduced by Bernhard Riemann in 1854 to denote sets that are locally homeomorphic to the Euclidean space. "Mannig" means "many" while "faltig" means "folded". Hence, the term "manifold" denotes a structure that is "very folded".

The Stiefel manifold arises naturally in the definition of many optimization problems as a search space. We present a few of them to motivate our work. In Principal Component Analysis, the goal is to find a set of orthogonal vectors, hence a point on Stiefel, that conserves the maximum of variance of the data set. [16] proposes an orthogonal Procrustes problem on the Stiefel manifold. The objective is to find the best orthogonal mapping to transform a matrix into another. [39] introduces a blind source separation problem. It consists of separating, for example, several voices mixed in recordings. In this context, orthogonality is expected to correspond to different voices. More recently, [24] proposed an application for weight normalization in deep neural networks. The weights are constrained to belong to the Stiefel manifold. This supplementary constraint leads to have normalized and de-corellated activations, which ultimately provides a faster training among several other advantages. The implementation of the Stiefel manifold is available in software such as `Geomstats` [30] and `Manopt` [8]. It is also involved in several applications in algorithm development [15] and statistics [11].

Given two points on a geodesically complete manifold M[1] endowed with a Riemannian metric g, it is commonly needed to compute their geodesic distance, i.e. the length of the shortest curve remaining on the manifold and binding the two points. This is essential for instance to compute the Karcher mean (generalization of a mean on a manifold) of a set of points. The Stiefel manifold is an example where we are currently unable to compute the correct distance between two points. More precisely, existing algorithms to compute a path can not guarantee to find the *shortest* path [48]. It can thus seem paradoxical that we are unable to perform an operation as simple as a mean on the Stiefel manifold.

The goal of this master's thesis is to improve the characterization of the geodesic distances on the Stiefel manifold. In particular, we show the bilipschitz equivalence of the geodesic distances induced by a previously proposed one-parameter family of Riemannian metrics. The one-parameter family contains the well-known *canonical* and *Euclidean* metrics. We investigate the link between these geodesic distances with the easy-to-compute Frobenius distance by providing tight upper and lower bounds. We also propose families of matrices reaching these

[1]For any point $p \in M$, all the geodesics $\gamma(t)$ emanating from p are defined for all $t \in \mathbb{R}$ [35, Chapter III].

bounds. Moreover, we give geometric properties of the minimizing geodesics under the *canonical* metric.

Many algorithms to compute geodesics on the Stiefel manifold rely on the efficient computation of the matrix exponential of skew-symmetric matrices. At the exception of the `FORTRAN` library [42] from 1976, this corner of numerical linear algebra has been left quite unattended. Therefore, we present the `SkewLinearAlgebra.jl` package. A `Julia` package which offers a specialized and fast framework to compute the eigenvalue decomposition of a skew-symmetric matrix.

The structure of this master's thesis is as follows. It is divided in four chapters. The first chapter introduces the necessary background in differential geometry and linear algebra to understand this work. The second chapter is short and reviews the questions related to the geodesic distance on the Special Orthogonal group. The third chapter addresses the study of Riemannian metrics on the Stiefel manifold. It contains the main theoretical results of this report. The last chapter presents the `SkewLinearAlgebra.jl` package, a `Julia` package specialized in the resolution of eigenproblems for skew-symmetric matrices.

Certainly with a bit of malice, Henri Poincaré declared in *Science and Hypothesis*: "*Verification differs from proof precisely because it is analytical, and because it leads to nothing. It leads to nothing because the conclusion is nothing but the premises translated into another language. A real proof, on the other hand, is fruitful, because the conclusion is in a sense more general than the premises.*". The author's hope is that the present work contains more than mere verifications, although it may include a few.

Chapter 1

State of the art

This chapter gives an overview of the mathematical background needed to go through this master's thesis. We define the Stiefel manifold and the available theoretical tools that were developed in previous researches. In particular, we explain the notions of metric, tangent space, geodesic, geodesic distance, exponential and logarithmic maps. We concentrate our explanations on the Stiefel manifold because the general theory of Riemannian manifolds is highly abstract so that it applies to a broad range of problems. However, maintaining this level of abstraction in our current context would be unnecessarily complex and would require introducing concepts that are not relevant to our discussion. Afterwards, we review the properties of the skew-symmetric matrices. Finally, we briefly introduce the algorithms available to try to compute logarithmic maps on the Stiefel manifold.

1.1 The Stiefel manifold

The Stiefel manifold is the set of orthogonal p-frames in $\mathbb{R}^n$. It is defined, e.g. in [1, Section 3.3.2], as

$$\mathrm{St}(n,p) := \left\{ \mathbf{U} \in \mathbb{R}^{n\times p} \mid \mathbf{U}^T\mathbf{U} = \mathbf{I}_p,\ n \geq p \right\}. \tag{1.1}$$

As a manifold, $\mathrm{St}(n,p)$ inherits from a bunch of associated features that we review in what follows.

1.1.1 Tangent space

We obtain the definition of the tangent space to the Stiefel manifold following the approach proposed in [9, Section 7.3]. Let us consider $h : \mathbb{R}^{n\times p} \to \mathrm{Sym}(p) : \mathbf{U} \to \mathbf{U}^T\mathbf{U} - \mathbf{I}_p$. h is a defining function for $\mathrm{St}(n,p)$ (see e.g. [1, Section 3.3.2]). The differential of h at $\mathbf{U} \in \mathrm{St}(n,p)$ in a direction $\mathbf{\Delta} \in \mathbb{R}^{n\times p}$ is given by

$$\begin{aligned} Dh(\mathbf{U})[\mathbf{\Delta}] &:= \lim_{t\to 0} \frac{h(\mathbf{U}+t\mathbf{\Delta}) - h(\mathbf{U})}{t} \\ &= \mathbf{U}^T\mathbf{\Delta} + \mathbf{\Delta}^T\mathbf{U}. \end{aligned}$$

Using this expression for the differential of h at $\mathbf{U}$, we can obtain an expression for $T_{\mathbf{U}}\mathrm{St}(n,p)$, the tangent space at $\mathbf{U}$:

$$T_{\mathbf{U}}\mathrm{St}(n,p) = \ker Dh(\mathbf{U}) = \left\{ \mathbf{\Delta} \in \mathbb{R}^{n\times p} \mid \mathbf{U}^T\mathbf{\Delta} + \mathbf{\Delta}^T\mathbf{U} = \mathbf{0} \right\}. \tag{1.2}$$

Let $\begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix} \in \mathrm{SO}(n)$ where $\mathbf{U}_\perp \in \mathrm{St}(n, n-p)$ is called the orthogonal completion of $\mathbf{U}$. We explain in section 1.2 how to build such a completion. It follows from (1.2) that $\mathbf{\Delta}$ can be expressed as

$$\mathbf{\Delta} = \mathbf{U}\mathbf{A} + \mathbf{U}_{\perp}\mathbf{B}, \tag{1.3}$$

where $\mathbf{A} \in \mathrm{Skew}(p)$ and $\mathbf{B} \in \mathbb{R}^{(n-p)\times p}$. Introducing (1.3) in (1.4), we obtain the following definition for the tangent space at $\mathbf{U}$:

$$T_{\mathbf{U}}\mathrm{St}(n,p) = \left\{ \mathbf{U}\mathbf{A} + \mathbf{U}_{\perp}\mathbf{B} \mid \mathbf{A} \in \mathrm{Skew}(p),\ \mathbf{B} \in \mathbb{R}^{(n-p)\times p} \right\}. \tag{1.4}$$

1.1.2 Stiefel as a quotient space

Following [15], we highlight that the Stiefel manifold can be defined as a quotient space. The equivalence class of $\mathbf{Q} \in \mathrm{O}(n)$ is $[\mathbf{Q}] = \left\{ \hat{\mathbf{Q}} \mid \hat{\mathbf{Q}} \sim \mathbf{Q}, \hat{\mathbf{Q}} \in \mathrm{O}(n) \right\}$ where $\sim$ is an equivalence relation and $\mathrm{O}(n) := \{\mathbf{Q} \in \mathbb{R}^{n\times n} \mid \mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}_n\}$. Given $\mathbf{Q}, \hat{\mathbf{Q}} \in \mathrm{O}(n)$, we define the equivalence relation

$$\begin{aligned}
\mathbf{Q} \sim \hat{\mathbf{Q}} &\iff \mathbf{Q}\mathbf{I}_{n\times p} = \hat{\mathbf{Q}}\mathbf{I}_{n\times p} \\
&\iff \mathbf{Q} = \hat{\mathbf{Q}} \begin{bmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{n-p} \end{bmatrix} \text{ for some } \mathbf{R}_{n-p} \in \mathrm{O}(n-p).
\end{aligned}$$

Hence, we obtain the definition of $[\mathbf{Q}]$:

$$[\mathbf{Q}] = \left\{ \mathbf{Q} \begin{bmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{n-p} \end{bmatrix} \mid \mathbf{R}_{n-p} \in \mathrm{O}(n-p) \right\}. \tag{1.5}$$

This definition of equivalence class allows to define Stiefel as a quotient space:

$$\mathrm{St}(n,p) = \mathrm{O}(n)/\mathrm{O}(n-p) \tag{1.6}$$

which is the set of all equivalence classes induced by the similarity $\sim$ on $\mathrm{O}(n)$.

Remark: It is also usual in the literature to see $\mathrm{St}(n,p) = \mathrm{SO}(n)/\mathrm{SO}(n-p)$ where $\mathrm{SO}(n)$, the Special Orthogonal group, is the restriction of $\mathrm{O}(n)$ to matrices with positive unit determinant. The two definitions are equivalent if $n \neq p$. The advantage of this second definition is that $\mathrm{St}(n,p)$ remains connected when $n = p$ since $\mathrm{SO}(n)$ is connected while $\mathrm{O}(n)$ is not.

1.1.3 Metrics

Still in [15], A. Edelman et al. presented two Riemannian metrics (*metrics* in short) for the Stiefel manifold. The *Euclidean* metric and the *canonical* metric. In [27], Hüper et al. proposed a one-parameter family of metrics generalizing the Euclidean and the canonical metrics. We recall in this section the definitions of these Riemannian metrics. There are two equivalent standard notations for the metric of a Riemannian manifold: $\langle\cdot,\cdot\rangle$ and $g(\cdot,\cdot)$. In this work, we usually prefer $\langle\cdot,\cdot\rangle$. However, it may be more convenient to use g to precise the endowed metric associated to the manifold M: (M, g).

The Euclidean metric For all $\mathbf{\Delta}, \widetilde{\mathbf{\Delta}} \in T_{\mathbf{U}}\mathrm{St}(n,p)$, the Euclidean metric is defined as

$$\begin{aligned}
\langle \mathbf{\Delta}, \widetilde{\mathbf{\Delta}} \rangle_{\mathrm{E}} := g_{\mathrm{E}}(\mathbf{\Delta}, \widetilde{\mathbf{\Delta}}) &:= \mathrm{Tr}\, \mathbf{\Delta}^T \widetilde{\mathbf{\Delta}} && (1.7) \\
&= \mathrm{Tr}\, \mathbf{A}^T \widetilde{\mathbf{A}} + \mathrm{Tr}\, \mathbf{B}^T \widetilde{\mathbf{B}}. && (1.8)
\end{aligned}$$

The Euclidean metric is the metric inherited from the Euclidean space $\mathbb{R}^{n\times p}$. $\mathrm{St}(n,p)$ is considered as a subset of $\mathbb{R}^{n\times p}$ and the length of a curve joining two points is the length of the curve computed in the ambient space $\mathbb{R}^{n\times p}$ with the classical Euclidean distance.

The canonical metric For all $\mathbf{\Delta}, \widetilde{\mathbf{\Delta}} \in T_{\mathbf{U}}\mathrm{St}(n,p)$, the canonical metric is defined as

$$\langle \mathbf{\Delta}, \widetilde{\mathbf{\Delta}} \rangle_{\mathrm{c}} := g_{\mathrm{c}}(\mathbf{\Delta}, \widetilde{\mathbf{\Delta}}) := \operatorname{Tr} \mathbf{\Delta}^T (\mathbf{I} - \frac{1}{2}\mathbf{U}\mathbf{U}^T)\widetilde{\mathbf{\Delta}} \tag{1.9}$$

$$= \frac{1}{2} \operatorname{Tr} \mathbf{A}^T \widetilde{\mathbf{A}} + \operatorname{Tr} \mathbf{B}^T \widetilde{\mathbf{B}}. \tag{1.10}$$

This metric is the one inherited from the quotient structure of Stiefel: $\mathrm{St}(n,p) = \mathrm{O}(n)/\mathrm{O}(n-p)$.

The β-metrics The Euclidean and the canonical metric were united in [27] by a one-parameter family of metrics. This family is also investigated in the algorithms of [48]. It is parameterized in [27] by $\alpha \in \mathbb{R} \setminus \{-1\}$. We find more convenient to define $\beta = \frac{1}{2(\alpha+1)} \in \mathbb{R} \setminus \{0\}$. For all $\mathbf{\Delta}, \widetilde{\mathbf{\Delta}} \in T_{\mathbf{U}}\mathrm{St}(n,p)$, the β-metric is defined as

$$\langle \mathbf{\Delta}, \widetilde{\mathbf{\Delta}} \rangle_{\beta} := g_{\beta}(\mathbf{\Delta}, \widetilde{\mathbf{\Delta}}) := \operatorname{Tr} \mathbf{\Delta}^T (\mathbf{I} - (1-\beta)\mathbf{U}\mathbf{U}^T)\widetilde{\mathbf{\Delta}} \tag{1.11}$$

$$= \beta \operatorname{Tr} \mathbf{A}^T \widetilde{\mathbf{A}} + \operatorname{Tr} \mathbf{B}^T \widetilde{\mathbf{B}}. \tag{1.12}$$

Notice that g_β is a Riemannian metric if and only if $\beta > 0$. This family does not derive from a geometric interpretation. However, these β-metrics provide a very convenient framework to establish general proofs that are true at once for both the Euclidean and the canonical metrics. For all these Riemannian metrics, we can define the β-norm induced by the β-metric:

$$\|\mathbf{\Delta}\|_\beta = \sqrt{\langle \mathbf{\Delta}, \mathbf{\Delta} \rangle_\beta} = \sqrt{\beta \|\mathbf{A}\|_{\mathrm{F}}^2 + \|\mathbf{B}\|_{\mathrm{F}}^2}. \tag{1.13}$$

1.1.4 Normal space

The normal space to $\mathbf{U} \in \mathrm{St}(n,p)$ is the set of matrices of $\mathbb{R}^{n\times p}$ that have a zero inner product with all $\mathbf{\Delta} \in T_{\mathbf{U}}\mathrm{St}(n,p)$. [15] and [9] both provide an expression for the normal space at $\mathbf{U}$ using the Euclidean metric. Let us derive an equivalent expression using the parameterized β-metric:

$$\begin{aligned}
N_{\mathbf{U}}\mathrm{St}(n,p) &= \{\mathbf{N} \in \mathbb{R}^{n\times p} \mid \forall \mathbf{\Delta} \in T_{\mathbf{U}}\mathrm{St}(n,p) : \langle \mathbf{N}, \mathbf{\Delta} \rangle_\beta = 0\} \\
&= \{\mathbf{N} \in \mathbb{R}^{n\times p} \mid \forall \mathbf{A} \in \mathrm{Skew}(p),\ \forall \mathbf{B} \in \mathbb{R}^{(n-p)\times p}, \langle \mathbf{U}\mathbf{S} + \mathbf{U}_\perp \mathbf{C}, \mathbf{U}\mathbf{A} + \mathbf{U}_\perp \mathbf{B} \rangle_\beta = 0\} \\
&= \{\mathbf{N} \in \mathbb{R}^{n\times p} \mid \forall \mathbf{A} \in \mathrm{Skew}(p),\ \forall \mathbf{B} \in \mathbb{R}^{(n-p)\times p}, \operatorname{Tr} \mathbf{S}^T \mathbf{A} = 0 \text{ and } \operatorname{Tr} \mathbf{C}^T \mathbf{B} = 0\} \\
&= \{\mathbf{U}\mathbf{S} \mid \mathbf{S} \in \mathrm{Sym}(p)\},
\end{aligned}$$

where $\mathrm{Sym}(p)$ is the set of $p \times p$ symmetric matrices. It turns out that the normal space at $\mathbf{U}$ is the same whatever the β-metric.

1.1.5 Orthogonal projection

The orthogonal projection of $\widetilde{\mathbf{U}} \in \mathbb{R}^{n\times p}$ on the tangent space at $\mathbf{U}$, denoted by $\mathrm{Proj}_{\mathbf{U}}(\widetilde{\mathbf{U}})$, must satisfy several conditions. The first condition is that $\widetilde{\mathbf{U}} - \mathrm{Proj}_{\mathbf{U}}(\widetilde{\mathbf{U}})$ belong to $N_{\mathbf{U}}\mathrm{St}(n,p)$, the normal space at $\mathbf{U}$. This leads to

$$\widetilde{\mathbf{U}} - \mathrm{Proj}_{\mathbf{U}}(\widetilde{\mathbf{U}}) = \mathbf{U}\mathbf{S} \in N_{\mathbf{U}}\mathrm{St}(n,p), \tag{1.14}$$

for some $\mathbf{S} \in \mathrm{Sym}(p)$. Moreover, the projection must belong to $T_{\mathbf{U}}\mathrm{St}(n,p)$. This yields

$$\mathrm{Proj}_{\mathbf{U}}(\widetilde{\mathbf{U}})^T \mathbf{U} + \mathbf{U}^T \mathrm{Proj}_{\mathbf{U}}(\widetilde{\mathbf{U}}) = \mathbf{0}. \tag{1.15}$$

Gathering Equations (1.14) and (1.15), we obtain:

$$\mathrm{Proj}_{\mathbf{U}}(\widetilde{\mathbf{U}}) = \mathbf{U}\,\mathrm{skew}(\mathbf{U}^T\widetilde{\mathbf{U}}) + (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\widetilde{\mathbf{U}}, \tag{1.16}$$

where $\mathrm{skew}(\mathbf{U}^T\widetilde{\mathbf{U}}) = \frac{\mathbf{U}^T\widetilde{\mathbf{U}} - \widetilde{\mathbf{U}}^T\mathbf{U}}{2}$ is the skew-symmetric part of $\mathbf{U}^T\widetilde{\mathbf{U}}$.

1.1.6 Lengths and distances

Let Stiefel be endowed with any β-metric, $(\mathrm{St}(n,p), g_\beta)$. Given a continuously differentiable curve $\gamma : [0,1] \to \mathrm{St}(n,p)$, we define $l_\beta(\gamma)$, the length of γ by

$$l_\beta(\gamma) = \int_0^1 \|\dot{\gamma}(t)\|_\beta \, \mathrm{d}t. \tag{1.17}$$

This definition easily extends to piecewise smooth curves. Once the length of a curve is defined, we can define the distance function $d_\beta(\mathbf{U}, \widetilde{\mathbf{U}})$ for all $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$ by

$$d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) := \inf \left\{ l_\beta(\gamma) \mid \gamma(0) = \mathbf{U},\ \gamma(1) = \widetilde{\mathbf{U}} \right\}. \tag{1.18}$$

[35, Chapter II, Proposition 1.1] states that Definition 1.18 verifies the three axioms of a distance function. These axioms are (i) positive definiteness, (ii) symmetry and (iii) the triangular inequality. Geodesics (see next section) of $(\mathrm{St}(n,p), g_\beta)$ are well-defined and exist for all times [27]. Hence, the Hopf-Rinow Theorem (see e.g. [35, Chapter III, Theorem 1.1]) applies: $(\mathrm{St}(n,p), g_\beta)$ is *geodesically complete*. We recall [35, Chapter III, Corollary 1.3]:

Corollary 1.1 ([35]). *For any two points p, q of a complete Riemannian manifold M, there exists a minimizing geodesic γ joining p to q (i.e., a geodesic with $l(\gamma) = d(p,q)$).*

In consequence, if we define Γ_β as the set of all the β-geodesics (i.e., the geodesics of $(\mathrm{St}(n,p), g_\beta)$), we can write

$$d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) = \min \left\{ l_\beta(\gamma_\beta) \mid \gamma_\beta \in \Gamma_\beta,\ \gamma_\beta(0) = \mathbf{U},\ \gamma_\beta(1) = \widetilde{\mathbf{U}} \right\}. \tag{1.19}$$

1.1.7 Geodesics

Geodesics are the generalization of straight lines on a manifold. These are defined as curves such that the acceleration remains in the normal space to the manifold when travelled at constant speed. We provide expressions from the literature for the geodesics on the Stiefel manifold endowed with the previously proposed Riemannian metrics. Exceptionally in this section, we use the short notation γ for $\gamma(t)$. We avoid this practice in general since γ denotes a curve while $\gamma(t)$ denotes a point on $\mathrm{St}(n,p)$.

The Euclidean metric Let $\gamma : \mathbb{R} \to \mathrm{St}(n,p)$ be a differentiable curve. The conditions for a curve to be a Euclidean geodesic γ_{E} are the following:

1. $\gamma_{\mathrm{E}}^T \gamma_{\mathrm{E}} = \mathbf{I}_p$ so that γ_{E} belongs to $\mathrm{St}(n,p)$.
2. $\ddot{\gamma}_{\mathrm{E}} = -\gamma_{\mathrm{E}} S$ where $\mathbf{S} \in \mathrm{Sym}(p)$ so that the "acceleration" always remains in the normal space to γ.

The condition $\gamma_{\mathrm{E}}^T \ddot{\gamma}_{\mathrm{E}} + 2\dot{\gamma}_{\mathrm{E}}^T \dot{\gamma}_{\mathrm{E}} + \ddot{\gamma}_{\mathrm{E}}^T \gamma_{\mathrm{E}} = 0$ can be obtained by differentiating twice the first condition. Combining the second and this last condition, we obtain that Euclidean geodesics must verify the condition

$$\ddot{\gamma}_{\mathrm{E}} + \gamma_{\mathrm{E}}(\dot{\gamma}_{\mathrm{E}}^T \dot{\gamma}_{\mathrm{E}}) = \mathbf{0}. \tag{1.20}$$

Among all the geodesics, the goal is in general to have a *minimizing* geodesic. In [15], Edelman et al. provide the optimization problem yielding the minimizing geodesic:

$$\min_{\gamma_{\mathrm{E}}} \int_0^1 \sqrt{\mathrm{Tr}(\dot{\gamma}_{\mathrm{E}}^T \dot{\gamma}_{\mathrm{E}})} \mathrm{d}t \quad \text{such that} \quad \gamma_{\mathrm{E}}(0) = \mathbf{U} \text{ and } \gamma_{\mathrm{E}}(1) = \widetilde{\mathbf{U}}. \tag{1.21}$$

They derived that solutions to the Equation (1.20) are only those of the form

$$\gamma_{\mathrm{E}}(t) = \begin{bmatrix}\mathbf{U} & \boldsymbol{\Delta}\end{bmatrix} \exp\left(t \begin{bmatrix}\mathbf{A} & -\boldsymbol{\Delta}^T\boldsymbol{\Delta} \\ \mathbf{I}_p & \mathbf{A}\end{bmatrix}\right) \begin{bmatrix}\mathbf{I}_p \\ \mathbf{0}\end{bmatrix} \exp\left(-\mathbf{A}t\right), \tag{1.22}$$

where $\boldsymbol{\Delta} = \dot{\gamma}_{\mathrm{E}}(0)$ and $\mathbf{A} = \mathbf{U}^T\boldsymbol{\Delta} \in \mathrm{Skew}(p)$. [25] obtained an alternative formula involving only exponentials of skew-symmetric matrices at the cost of higher dimensional matrices:

$$\gamma_{\mathrm{E}}(t) = \exp\left(t\left(\boldsymbol{\Delta}\mathbf{U}^T - \mathbf{U}\boldsymbol{\Delta}^T\right)\right)\mathbf{U}\exp\left(-\mathbf{A}t\right). \tag{1.23}$$

The canonical metric [15] also provides a necessary and sufficient condition for the canonical geodesics γ_{c}:

$$\ddot{\gamma}_{\mathrm{c}} + \dot{\gamma}_{\mathrm{c}}\dot{\gamma}_{\mathrm{c}}^T\gamma_{\mathrm{c}} + \gamma_{\mathrm{c}}((\gamma_{\mathrm{c}}^T\dot{\gamma}_{\mathrm{c}})^2 + \dot{\gamma}_{\mathrm{c}}^T\dot{\gamma}_{\mathrm{c}}) = \mathbf{0}. \tag{1.24}$$

This condition is obtained with the calculus of variations on the problem

$$\min_{\gamma_{\mathrm{c}}} \int_0^1 \sqrt{\langle\dot{\gamma}_{\mathrm{c}}, \dot{\gamma}_{\mathrm{c}}\rangle_{\mathrm{c}}}\mathrm{d}t \quad \text{such that} \quad \gamma_{\mathrm{c}}(0) = \mathbf{U} \text{ and } \gamma(1)_{\mathrm{c}} = \widetilde{\mathbf{U}}.$$

We can verify that paths of the form

$$\gamma_{\mathrm{c}}(t) = \mathbf{Q}\exp(\mathbf{S}t)\mathbf{I}_{n\times p}, \;\; \mathbf{Q} \in \mathrm{SO}(n), \;\; \mathbf{S} \in \mathrm{Skew}(n) \tag{1.25}$$

verify Equation (1.24) if and only if

$$(\mathbf{I} - \mathbf{I}_{n\times p}\mathbf{I}_{n\times p}^T)\mathbf{S}(\mathbf{I} - \mathbf{I}_{n\times p}\mathbf{I}_{n\times p}^T) = \mathbf{0},$$

or equivalently,

$$\mathbf{S} = \begin{bmatrix}\mathbf{A} & -\mathbf{B}^T \\ \mathbf{B} & \mathbf{0}\end{bmatrix},$$

for some $\mathbf{A} \in \mathrm{Skew}(p)$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$.

The β-metrics The β-geodesics were derived in [27] and [48] summarized the work using notations we follow in this master's thesis. The expression of the β-geodesics γ_β is

$$\gamma_\beta(t) = \exp\left(t((\beta-2)\mathbf{U}\mathbf{A}\mathbf{U}^T + \boldsymbol{\Delta}\mathbf{U}^T - \mathbf{U}\boldsymbol{\Delta}^T)\right)\mathbf{U}\exp\left((1-2\beta)\mathbf{A}t\right). \tag{1.26}$$

This formula is interesting but poorly manipulable. A better formula for analytic developments and interpretation is given by [48, Proposition 1]:

Proposition 1.1 ([48]). *Let $p \leq \frac{n}{2}$. For $\mathbf{U} \in \mathrm{St}(n,p)$, $\boldsymbol{\Delta} \in T_{\mathbf{U}}\mathrm{St}(n,p)$,*

$$\gamma_\beta(1) = \begin{bmatrix}\mathbf{U} & \mathbf{Q}\end{bmatrix}\exp\left(\begin{bmatrix}2\beta\mathbf{A} & -\mathbf{B}^T \\ \mathbf{B} & \mathbf{0}\end{bmatrix}\right)\mathbf{I}_{n\times p}\exp\left((1-2\beta)\mathbf{A}\right), \tag{1.27}$$

where $\mathbf{A} = \mathbf{U}^T\boldsymbol{\Delta} \in \mathrm{Skew}(p)$ and $\mathbf{Q}\mathbf{B} = (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\boldsymbol{\Delta} \in \mathbb{R}^{n\times p}$ is any matrix decomposition where $\mathbf{Q} \in \mathrm{St}(n,p)$ and $\mathbf{B} \in \mathbb{R}^{p\times p}$.

Besides, the formula (1.27) remains true if $p > \frac{n}{2}$. Of course, the $2p \times 2p$ matrix exponential should be then replaced by a $n \times n$ matrix exponential instead.

1.1.8 Exponential and the logarithmic map

Given $\mathbf{U} \in \mathrm{St}(n,p)$ endowed with a β-metric, the exponential map at $\mathbf{U}$ with an initial velocity $\boldsymbol{\Delta} \in T_{\mathbf{U}}\mathrm{St}(n,p)$ is $\mathrm{Exp}_{\mathbf{U},\beta}(\boldsymbol{\Delta}) := \gamma_\beta(1)$. For $\widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$, the logarithmic map, or geodesic endpoint problem, consists of finding $\boldsymbol{\Delta} \in T_{\mathbf{U}}\mathrm{St}(n,p)$ such that $\mathrm{Exp}_{\beta,\mathbf{U}}(\boldsymbol{\Delta}) = \widetilde{\mathbf{U}}$ and $\gamma_\beta(t) := \mathrm{Exp}_{\beta,\mathbf{U}}(t\boldsymbol{\Delta})$ is a minimizing geodesic for all $t \in [0,1]$. Given such a $\boldsymbol{\Delta}$, we say $\boldsymbol{\Delta} \in \mathrm{Log}_{\beta,\mathbf{U}}(\widetilde{\mathbf{U}})$. The problem of computing $\boldsymbol{\Delta} \in \mathrm{Log}_{\beta,\mathbf{U}}(\widetilde{\mathbf{U}})$ is an open problem on $\mathrm{St}(n,p)$. The best numerical methods available to compute candidates for $\boldsymbol{\Delta}$ is reviewed in [48]. We briefly explain and review the algorithms we have implemented in section 1.6. In order to get some insight about the logarithm problem on $\mathrm{St}(n,p)$, we recall Theorem 1.1. This theorem uses the notion of injectivity radius, $r_{\mathrm{St}}(\mathbf{U})$ (see e.g. [35, Chapter III, Definition 4.12]), defined by

$$r_{\mathrm{St}}(\mathbf{U}) := \sup\left\{r > 0 \mid \mathrm{Exp}_{\beta,\mathbf{U}}(\boldsymbol{\Delta}) \text{ is a diffeomorphism for all } \boldsymbol{\Delta} \text{ with } \|\boldsymbol{\Delta}\|_\beta \le r\right\}.$$

Theorem 1.1 ([48]). *Let* $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$ *with* $d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) < r_{\mathrm{St}}(\mathbf{U})$ *and let*

$$\mathbf{M} := \mathbf{U}^T\widetilde{\mathbf{U}},\ \mathbf{QN} := (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\widetilde{\mathbf{U}} = \begin{bmatrix}\mathbf{Q}_r & \mathbf{Q}_{p-r}\end{bmatrix}\begin{bmatrix}\mathbf{N}_r \\ \mathbf{0}\end{bmatrix},\ r \le p, \tag{1.28}$$

with $\mathbf{Q} \in \mathrm{St}(n,p)$ *and* $r = \mathrm{rank}(\mathbf{N}_r) = \mathrm{rank}(\mathbf{N})$. *Then* $\boldsymbol{\Delta} = \mathrm{Log}_{\mathbf{U},\beta}(\widetilde{\mathbf{U}})$ *features a representation*

$$\boldsymbol{\Delta} = \mathbf{UA} + \mathbf{QB} = \mathbf{UA} + \begin{bmatrix}\mathbf{Q}_r & \mathbf{Q}_{p-r}\end{bmatrix}\begin{bmatrix}\mathbf{B}_r \\ \mathbf{0}\end{bmatrix}, \tag{1.29}$$

where $\mathbf{A} \in \mathrm{Skew}(p)$ *and* $\mathbf{B}_r \in \mathbb{R}^{r\times p}$ *and* $\mathbf{A}, \mathbf{B}$ *solve Problem 1.1.*

The key aspect to solve the logarithm problem on $\mathrm{St}(n,p)$ was highlighted in [48] as the problem of solving Problem 1.1 that we describe hereunder.

Problem 1.1 ([48]). *For all* $\mathbf{U} \in \mathrm{St}(n,p)$, *Let* $\mathbf{M}, \mathbf{N}$ *from Theorem 1.1 and* $\begin{bmatrix}\mathbf{X}_0 \\ \mathbf{Y}_0\end{bmatrix} \in \mathrm{St}(2p,p)$ *any completion of* $\begin{bmatrix}\mathbf{M} \\ \mathbf{N}\end{bmatrix} \in \mathrm{St}(2p,p)$. *Find* $\mathbf{A}, \mathbf{C} \in \mathrm{Skew}(p)$, $\mathbf{B} \in \mathbb{R}^{n\times p}$ *such that*

$$F\left(\begin{bmatrix}\mathbf{A} & -\mathbf{B}^T \\ \mathbf{B} & \mathbf{C}\end{bmatrix}\right) = \begin{bmatrix}\mathbf{M} & \mathbf{X}_0 \\ \mathbf{N} & \mathbf{Y}_0\end{bmatrix},\ \textit{where } F : \mathrm{Skew}(2p) \to \mathrm{SO}(2p),$$
$$\textit{and } F\left(\begin{bmatrix}\mathbf{A} & -\mathbf{B}^T \\ \mathbf{B} & \mathbf{C}\end{bmatrix}\right) = \exp\left(\begin{bmatrix}2\beta\mathbf{A} & -\mathbf{B}^T \\ \mathbf{B} & \mathbf{0}\end{bmatrix}\right)\begin{bmatrix}\exp((1-2\beta)\mathbf{A}) & \mathbf{0} \\ \mathbf{0} & \exp(-\mathbf{C})\end{bmatrix}$$

The formulation of Theorem 1.1 is not convenient due to the very restrictive hypothesis $d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) < r_{\mathrm{St}}(\mathbf{U})$. We propose a more general formulation of the same theorem :

Theorem 1.2 (adapted from [48]). *Let* $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$ *with* $\widetilde{\mathbf{U}} \notin C_{\mathbf{U}}$ *(cut locus defined in section 1.4) and let*

$$\mathbf{M} := \mathbf{U}^T\widetilde{\mathbf{U}},\ \mathbf{QN} := (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\widetilde{\mathbf{U}} = \begin{bmatrix}\mathbf{Q}_r & \mathbf{Q}_{p-r}\end{bmatrix}\begin{bmatrix}\mathbf{N}_r \\ \mathbf{0}\end{bmatrix},\ r \le p, \tag{1.30}$$

with $\mathbf{Q} \in \mathrm{St}(n,p)$ *and* $r = \mathrm{rank}(\mathbf{N}_r) = \mathrm{rank}(\mathbf{N})$. *Then* $\boldsymbol{\Delta} \in \mathrm{Log}_{\mathbf{U},\beta}(\widetilde{\mathbf{U}})$ *features a representation*

$$\boldsymbol{\Delta} = \mathbf{UA} + \mathbf{QB} = \mathbf{UA} + \begin{bmatrix}\mathbf{Q}_r & \mathbf{Q}_{p-r}\end{bmatrix}\begin{bmatrix}\mathbf{B}_r \\ \mathbf{0}\end{bmatrix}, \tag{1.31}$$

where $\mathbf{A} \in \mathrm{Skew}(p)$ *and* $\mathbf{B}_r \in \mathbb{R}^{r\times p}$ *where* $\mathbf{A}, \mathbf{B}$ *is a solution of Problem 1.1 such that* $\|\boldsymbol{\Delta}\|_\beta$ *is minimal.*

Problem 1.1 offers a formulation for the logarithm problem, which forms the basis of computationally efficient algorithms, as e.g. [48, Algorithm 4]. It is important to notice that solving this problem only yields a *geodesic*, which may not necessarily be a *minimizing geodesic*. A minimizing geodesic is a solution $(\mathbf{A}, \mathbf{B})$ to Problem 1.1 that satisfies the condition $\sqrt{\beta\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2}$ being minimal.

1.2 Orthogonal completion

When given a matrix $\mathbf{U} \in \mathrm{St}(n, p)$, it is often needed to obtain an orthonormal completion $\mathbf{U}_\perp \in \mathrm{St}(n, n-p)$ (i.e., such that $\mathbf{U}_\perp^T \mathbf{U} = 0$). To obtain $\mathbf{U}_\perp$ in practice, we can initialize a normally-distributed random matrix $\mathbf{M} \in \mathbb{R}^{n \times n-p}$. Since we want the column space of $\mathbf{U}_\perp$ to be orthogonal to the column space of $\mathbf{U}$, we project $\mathbf{M}$ on this orthogonal space by taking $\widetilde{\mathbf{M}} = (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{M}$ where $\mathbf{I} - \mathbf{U}\mathbf{U}^T$ is the orthogonal projector (see e.g.[21, Section 2.6]). It is now expected that the column space of $\widetilde{\mathbf{M}}$ spans the column space of $\mathbf{U}_\perp$, i.e. $\widetilde{\mathbf{M}}$ is full column rank. This happens with probability one since $\mathbf{M}$ was a normally-distributed random matrix. To obtain $\mathbf{U}_\perp$, it is sufficient to apply a Gram-Schmidt process on $\widetilde{\mathbf{M}}$. We present in Algorithm 1 its numerically stable version: the Modified Gram-Schmidt algorithm [41, Algorithm 8.1]. In particular, we present a "in-place" version that only returns the orthogonalized matrix and not the complete QR factorization. $\mathbf{q}_i$ stands for the ith column of $\mathbf{Q}$.

Algorithm 1 [41, Algorithm 8.1] Modified Gram-Schimdt orthogonalization

Input $\mathbf{Q} \in \mathbb{R}^{n \times n-p}$
for $i = 1$ to $n-p$ **do**
  $\mathbf{q}_i = \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|}$
  for $j = i+1$ to $n-p$ **do**
    $\mathbf{q}_j = \mathbf{q}_j - \langle \mathbf{q}_j, \mathbf{q}_i \rangle \mathbf{q}_i$
  end for
end for
return $\mathbf{Q} \in \mathrm{St}(n, n-p)$

Algorithm 1 only features `BLAS-1` (vector-vector) operations [7]. As detailed in [38], state-of-the-art linear algebra routines rely on blocked algorithms that introduce `BLAS-3` (matrix-matrix) operations. The operation count is preserved but the spatial locality of `BLAS-3` operations increases the cache performance and decreases the number of accesses to the main memory. These algorithms are implemented in `LAPACK` [2].

1.3 The Special Orthogonal group

When $p = n$, the Stiefel manifold becomes the Orthogonal group of matrices $\mathrm{O}(n)$. Matrices of $\mathrm{O}(n)$ are divided into two connected components by the sign of their determinant. The Special Orthogonal group $\mathrm{SO}(n)$ is the subset of matrices with unit determinant.

$$\mathrm{SO}(n) := \left\{ \mathbf{Q} \in \mathbb{R}^{n \times n} \mid \mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}_n,\ \det(\mathbf{Q}) = 1 \right\}. \tag{1.32}$$

$\mathrm{SO}(n)$ is also often denoted as the set of rotations of size $n \times n$. The expression of the tangent space at all $\mathbf{Q} \in \mathrm{SO}(n)$ simplifies to

$$T_{\mathbf{Q}}\mathrm{SO}(n) = \{\mathbf{Q}\mathbf{A} \mid \mathbf{A} \in \mathrm{Skew}(n)\} = \{\mathbf{A}\mathbf{Q} \mid \mathbf{A} \in \mathrm{Skew}(n)\}. \tag{1.33}$$

The second equality of (1.33) stands because for all $\mathbf{A} \in \mathrm{Skew}(n)$ and $\mathbf{Q} \in \mathrm{SO}(n)$, we have $\mathbf{Q}\mathbf{A} = \mathbf{Q}\mathbf{A}\mathbf{Q}^T\mathbf{Q} = \widetilde{\mathbf{A}}\mathbf{Q}$ with $\widetilde{\mathbf{A}} = \mathbf{Q}\mathbf{A}\mathbf{Q}^T \in \mathrm{Skew}(n)$. The Special Orthogonal group inherits from the β-metrics of the Stiefel manifold. However, all the β-metrics become identical up to a scaling factor. Fixing $\beta = 1$ is a natural choice to avoid constants. This yields for all $\mathbf{\Delta}, \widetilde{\mathbf{\Delta}} \in T_{\mathbf{Q}}\mathrm{SO}(n)$

$$\langle \mathbf{\Delta}, \widetilde{\mathbf{\Delta}} \rangle := g(\mathbf{\Delta}, \widetilde{\mathbf{\Delta}}) := \mathrm{Tr}\, \mathbf{A}^T \widetilde{\mathbf{A}}. \tag{1.34}$$

Taking the special case $\beta = 1$ and $p = n$ of Stiefel's geodesics, we get that geodesics are curves of the form

$$\gamma(t) = \mathbf{Q} \exp(\mathbf{A}t) \text{ if } \mathbf{\Delta} = \mathbf{QA}, \tag{1.35}$$

$$\text{or equivalently } \gamma(t) = \exp(\mathbf{A}t)\mathbf{Q} \text{ if } \mathbf{\Delta} = \mathbf{AQ}. \tag{1.36}$$

1.4 Cut Locus

In this section, we introduce the notion of cut locus of a manifold M at a point $\mathbf{U}$. The first time a notion close to the cut locus was given was in 1905 by Henri Poincaré [33]. It was then called "ligne de partage". In 1935, J.H.C Whitehead [44] extended Poincare's notion to give the current definition of the cut locus. The notions presented hereunder are taken from [17] and [35, Chapter III, Section 4].

Let (M, g) be a compact geodesically complete Riemannian manifold of dimension n. Recall that $(\mathrm{St}(n,p), g_\beta)$ is geodesically complete for $n > p$ as a consequence of the Hopf-Rinow theorem (see e.g. [35], Theorem 1.1).

Definition 1.1. *The cut time* $\mathcal{T}_{\mathbf{U}}(\mathbf{\Delta})$ *for* $\mathbf{\Delta} \in T_{\mathbf{U}}M$ *is defined as*

$$\mathcal{T}_{\mathbf{U}}(\mathbf{\Delta}) := \sup\{t > 0 \mid \mathrm{Exp}_{\mathbf{U}}(s\mathbf{\Delta}) \textit{is a minimizing geodesic for all } s \in [0,t]\} \tag{1.37}$$

By definition of $\mathcal{T}_{\mathbf{U}}(\mathbf{\Delta})$, the geodesic defined by $\mathrm{Exp}_{\mathbf{U}}(t\mathbf{\Delta})$ is minimizing if and only if $t \in [0, \mathcal{T}_{\mathbf{U}}(\mathbf{\Delta})]$. This minimizing geodesic is unique for all $t \in [0, \mathcal{T}_{\mathbf{U}}(\mathbf{\Delta}))$ but not at $t = \mathcal{T}_{\mathbf{U}}(\mathbf{\Delta})$.

Definition 1.2. *The injectivity domain is* $\widehat{I}_{\mathbf{U}} := \{t\mathbf{\Delta} \mid \mathbf{\Delta} \in T_{\mathbf{U}}M,\ 0 < t < \mathcal{T}_{\mathbf{U}}(\mathbf{\Delta})\}$. $I_{\mathbf{U}} = \mathrm{Exp}_{\mathbf{U}}(\widehat{I}_{\mathbf{U}})$ *is called the interior set at* $\mathbf{U}$.

Definition 1.3. *The tangent cut locus is* $\widehat{C}_{\mathbf{U}} := \{\mathcal{T}_{\mathbf{U}}(\mathbf{\Delta})\mathbf{\Delta} \mid \mathbf{\Delta} \in T_{\mathbf{U}}M\}$ *and the cut locus* $C_{\mathbf{U}}$ *is defined as* $C_{\mathbf{U}} := \mathrm{Exp}_{\mathbf{U}}(\widehat{C}_{\mathbf{U}})$

The interior set and the cut locus are functions of the cut time. Therefore, the study of these sets essentially reduces to the study of the cut time. The characterization of the cut time is an efficient way of solving the logarithm problem. If a geodesic curve going from $\mathbf{U}$ to $\widetilde{\mathbf{U}}$ is known, it is minimizing if and only if its initial velocity belongs to $\hat{I}_{\mathbf{U}} \cup \hat{C}_{\mathbf{U}}$. One of our goals is to provide the tightest possible upper bounds for the cut time.

1.5 The Skew-symmetric matrices

Definition 1.4. *A matrix* $\mathbf{S} \in \mathbb{C}^{n \times n}$ *is skew-Hermitian if* $\mathbf{S} = -\bar{\mathbf{S}}^T =: -\mathbf{S}^*$. *If* $\mathbf{S} \in \mathbb{R}^{n\times n}$, *then* $\mathbf{S} = -\mathbf{S}^T$ *and we say that* $\mathbf{S}$ *is skew-symmetric or anti-symmetric.*

We will refer to skew-symmetric matrices using the letter $\mathbf{S}$ or $\mathbf{A}$. $\mathbf{S}$ stands for skew-symmetric and $\mathbf{A}$ for anti-symmetric.

Skew-symmetric matrices play a critical role in the study of orthogonal matrices. They are closely related to matrices in the Special Orthogonal group $\mathrm{SO}(n)$, which consists of square orthogonal matrices with unit determinant (i.e., rotation matrices). Specifically, every rotation matrix can be expressed as the matrix exponential of a skew-symmetric matrix, meaning that the logarithm of a rotation matrix always features a skew-symmetric representation. In this work, our objective is to expand our understanding of the logarithm on Stiefel. As a result, we will frequently encounter skew-symmetric matrices. Therefore it is important to review and clarify some properties and notations related to these matrices. It is surprisingly challenging to find a suitable and comprehensive reference on skew-symmetric matrices, especially when compared to their symmetric counterparts. For the properties mentioned below, we recommend consulting [23].

Property 1.1. *Every* $\mathbf{S} \in \mathbb{R}^{n\times n}$ *skew-symmetric can be diagonalized under unitary transformation* $\mathbf{V} \in \mathbb{C}^{n\times n}$. *The eigenvalues of* $\mathbf{S}$ *are pure imaginaries. The eigenvectors of* $\mathbf{S}$ *are orthonormal, i.e form a (complex) unitary matrix:* $\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^*$ *with* $\mathbf{V}^*\mathbf{V} = \mathbf{V}\mathbf{V}^* = \mathbf{I}_n$ *and* $\mathbf{R}e(\mathbf{\Lambda}_{ii}) = 0$ *for* $i \in \{1, ..., n\}$

Proof. $\mathbf{S}$ can be diagonalized under unitary transformations because skew-symmetric matrices are normal (i.e. $\mathbf{S}\mathbf{S}^T = (-\mathbf{S}^T)(-\mathbf{S}) = \mathbf{S}^T\mathbf{S}$) and all the normal matrices are diagonalizable under unitary transformations (see e.g. [21, Corollary 7.1.4]). The eigenvalues are pure imaginaries because $i\mathbf{S}$ is Hermitian ($\mathbf{A} = \bar{\mathbf{A}}^T,\ \mathbf{A} \in \mathbb{C}^{n\times n}$) and Hermitian matrices have real eigenvalues (see e.g. [23, Theorem 4.1.3]). $\square$

In the framework of rotation matrices, the eigenvalues of skew-symmetric matrices can be interpreted as angles, historically denoted by the Greek letter θ. Therefore, it is convenient to write the eigendecomposition $\mathbf{S} = \mathbf{V}i\mathbf{\Theta}\mathbf{V}^*$ where $i\mathbf{\Theta} = \mathbf{\Lambda}$. Lemma 1.1 establishes a useful property of the eigenvectors of skew-symmetric matrices.

Lemma 1.1. *The eigenvalues of* $\mathbf{S} \in \mathrm{Skew}(n)$ *are conjugate pure imaginary numbers and the eigenvectors of* $\mathbf{S} \in \mathrm{Skew}(n)$ *associated to conjugate eigenvalues are conjugate:* $\mathbf{S}\mathbf{v} = \lambda\mathbf{v}$ *if and only if* $\mathbf{S}\bar{\mathbf{v}} = \bar{\lambda}\bar{\mathbf{v}}$.

Proof. We show that $\mathbf{S}\mathbf{v} = \lambda\mathbf{v}$ and $\mathbf{S}\bar{\mathbf{v}} = \bar{\lambda}\bar{\mathbf{v}}$ yield the same necessary and sufficient conditions. We have $\mathbf{v} \in \mathbb{C}^n$ and $\lambda \in \mathbb{C}$. Pose $\mathbf{v} = \mathbf{v}_R + i\mathbf{v}_I$ with $\mathbf{v}_R, \mathbf{v}_I \in \mathbb{R}^n$ and $\lambda = i\theta$ with $\theta \in \mathbb{R}$. This yields

$$\begin{cases} \mathbf{S}\mathbf{v} = \mathbf{S}(\mathbf{v}_R + i\mathbf{v}_I) = \mathbf{S}\mathbf{v}_R + i\mathbf{S}\mathbf{v}_I \\ \lambda\mathbf{v} = i\theta(\mathbf{v}_R + i\mathbf{v}_I) = -\theta\mathbf{v}_I + i\theta\mathbf{v}_R \end{cases} \iff \begin{cases} \mathbf{S}\mathbf{v}_R = -\theta\mathbf{v}_I \\ \mathbf{S}\mathbf{v}_I = \theta\mathbf{v}_R \end{cases} \tag{1.38}$$

Similarly, we have

$$\begin{cases} \mathbf{S}\bar{\mathbf{v}} = \mathbf{S}(\mathbf{v}_R - i\mathbf{v}_I) = \mathbf{S}\mathbf{v}_R - i\mathbf{S}\mathbf{v}_I \\ \bar{\lambda}\bar{\mathbf{v}} = -i\theta(\mathbf{v}_R - i\mathbf{v}_I) = -\theta\mathbf{v}_I - i\theta\mathbf{v}_R \end{cases} \iff \begin{cases} \mathbf{S}\mathbf{v}_R = -\theta\mathbf{v}_I \\ \mathbf{S}\mathbf{v}_I = \theta\mathbf{v}_R \end{cases} \tag{1.39}$$

It turns out that Systems 1.38 and 1.39 are the same. Therefore, $\mathbf{S}\mathbf{v} = \lambda\mathbf{v}$ if and only if $\mathbf{S}\bar{\mathbf{v}} = \bar{\lambda}\bar{\mathbf{v}}$. $\square$

Property 1.2. *The matrix exponential* $e^{\mathbf{S}}$ *of* $\mathbf{S} \in \mathrm{Skew}(n)$ *belongs to* $\mathrm{SO}(n)$*:* $e^{\mathbf{S}} \in \mathrm{SO}(n)$.

Proof. Given $\mathbf{S} = \mathbf{V}i\mathbf{\Theta}\mathbf{V}^*$, we have $e^{\mathbf{S}} = \mathbf{V}e^{i\mathbf{\Theta}}\mathbf{V}^*$ by definition of the matrix exponential. Hence, all the eigenvalues of $e^{\mathbf{S}}$ lie on the complex unit circle and have a modulus 1. Since $e^{\mathbf{S}}$ is a real matrix by definition, we have $e^{\mathbf{S}} \in \mathrm{O}(n)$. Moreover, we have $\det(e^{\mathbf{S}}) = \det(e^{i\mathbf{\Theta}}) = 1$ since the eigenvalues of $\mathbf{S}$ come in conjugate pairs. This finally leads to $e^{\mathbf{S}} \in \mathrm{SO}(n)$. $\square$

Although we have shown in Lemma 1.1 that the existence of an eigenvector $\mathbf{v}$ implies that $\bar{\mathbf{v}}$ is also an eigenvector, it is clear that the complex eigenvector matrix $\mathbf{V}$ is not uniquely defined since every $e^{i\alpha}\mathbf{V}$ is also a valid matrix of eigenvectors for all $\alpha \in \mathbb{R}$. There is thus no straightforward reason for which $\mathbf{v}$ and $\bar{\mathbf{v}}$ should be eigenvectors of the same representation $\mathbf{V}$. We provide a constructive proof in the next Theorem.

Theorem 1.3. *Every* $\mathbf{S} \in \mathrm{Skew}(n)$ *admits an eigendecomposition such that conjugate eigenvectors are associated to conjugate eigenvalues.*

Proof. We have a so-called *constructive proof*. We build an eigendecomposition meeting the statement of the theorem. Any skew-symmetric matrix can be put in the real Schur form:

$$\mathbf{S} = \mathbf{Q}\mathbf{A}\mathbf{Q}^T, \tag{1.40}$$

where $\mathbf{Q} \in \mathrm{O}(n)$ and $\mathbf{A}$ is a skew-symmetric block-diagonal matrix where diagonal blocks are $\begin{bmatrix} 0 \end{bmatrix}$ or $\begin{bmatrix} 0 & \theta_j \\ -\theta_j & 0 \end{bmatrix}$. To go from the real Schur form to an eigenvalue decomposition, it is sufficient diagonalize $\mathbf{A}$. The eigenvalue decomposition of a 2×2 block is

$$\begin{bmatrix} 0 & \theta_j \\ -\theta_j & 0 \end{bmatrix} = \frac{\sqrt{2}}{2} \begin{bmatrix} -i & i \\ 1 & 1 \end{bmatrix} \begin{bmatrix} i\theta_j & 0 \\ 0 & -i\theta_j \end{bmatrix} \begin{bmatrix} i & 1 \\ -i & 1 \end{bmatrix} \frac{\sqrt{2}}{2} \tag{1.41}$$

We observe that the eigenvectors of the 2×2 blocks are conjugate vectors. Applying this decomposition on every block finishes the proof. □

1.6 Algorithms to compute logarithms on the Stiefel manifold

This section is intended for the reader interested on how numerical results are produced. This master's thesis is however mainly theoretical and understanding the algorithms briefly explained below is not necessary to go through this work. However, this thesis ultimately aims at improving these algorithms so that they are essential to motivate and stimulate questions.

Several algorithms already exist to (try to) solve the geodesic endpoint problem given the Euclidean metric, the canonical metric or any β-metric. The main references on the subject are [10] and the reviewing paper [48]. We briefly explain the ones we implemented to perform numerical experiments in this master's thesis. In 2017, Bryner [10] introduced two numerical schemes to address the geodesic endpoint problem with the Euclidean metric. The first scheme is a shooting method and the second is a path-straightening method. On the other hand, [48] proposes a very efficient method, based on a simple idea, designed for the canonical metric. This *vacuum algorithm* was originally proposed in [34]. It was then enhanced in [47] and again in [48].

1.6.1 The shooting method

The shooting method was developed in [10] and benchmarked in [48]. Given $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$, we wish to find *a* vector $\mathbf{\Delta} \in T_{\mathbf{U}}(\mathrm{St}(n,p))$ such that $\widetilde{\mathbf{U}} = \mathrm{Exp}_{\mathbf{U}}(\mathbf{\Delta})$[1]. The idea is the following: set $\mathbf{\Delta}$ to some initial guess in $T_{\mathbf{U}}(\mathrm{St}(n,p))$ and update $\mathbf{\Delta}$ until the exponential map is sufficiently close to $\widetilde{\mathbf{U}}$. To do this, the exponential map is discretized along $m+1$ points $\left\{\widetilde{\mathbf{U}}^s_j\right\}$ for $\{t_j\}$ and $j = 0, 1, ..., m$. First, the points $\{\widetilde{\mathbf{U}}^s_j\}$ along the geodesic are *exactly* computed. Then, the velocity error $\mathbf{\Delta}^s \propto \mathrm{Proj}_{\widetilde{\mathbf{U}}^s_m}(\widetilde{\mathbf{U}}^s_m - \widetilde{\mathbf{U}})$ is parallel transported back to $T_{\mathbf{U}}\mathrm{St}(n,p)$ in order to update $\mathbf{\Delta} \leftarrow \mathbf{\Delta} - \mathbf{\Delta}^s$. Since we can not perform this operation exactly, the parallel transport is *approximated* by successive projection and re-scaling on $T_{\widetilde{\mathbf{U}}^s_j}\mathrm{St}(n,p)$ for $j = m, m-1, ..., 0$. The procedure is detailed in Algorithm 2.

[1]Notice already the impossibility of looking for *the* vector $\mathbf{\Delta}$

Algorithm 2 Shooting method from [48], adapted from [10].

Input: $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$, $\epsilon > 0$ and $\{t_j\}$, $j = 0, 1, ..., m$ with $t_0 = 0$ and $t_m = 1$.
$\nu \leftarrow \|\widetilde{\mathbf{U}} - \mathbf{U}\|_\mathrm{F}$
$\mathbf{\Delta} \leftarrow \nu \frac{\mathrm{Proj}_\mathbf{U}(\widetilde{\mathbf{U}})}{\|\mathrm{Proj}_\mathbf{U}(\widetilde{\mathbf{U}})\|_\mathrm{F}}$
while $\nu > \epsilon$ **do**
  for $j = 1, ..., m$ **do**
    $\widetilde{\mathbf{U}}_j^s \leftarrow \mathrm{Exp}_{\beta,\mathbf{U}}(t_j \mathbf{\Delta})$
  end for
  $\mathbf{\Delta}^s \leftarrow \widetilde{\mathbf{U}}_m^s - \widetilde{\mathbf{U}}$
  $\nu \leftarrow \|\mathbf{\Delta}^s\|_\mathrm{F}$
  for $j = m, ..., 0$ **do**
    $\mathbf{\Delta}^s \leftarrow \nu \frac{\mathrm{Proj}_{\widetilde{\mathbf{U}}_j^s}(\mathbf{\Delta}^s)}{\|\mathrm{Proj}_{\widetilde{\mathbf{U}}_j^s}(\mathbf{\Delta}^s)\|_\mathrm{F}}$
  end for
  $\mathbf{\Delta} \leftarrow \mathbf{\Delta} - \mathbf{\Delta}^s$
end while
return $\mathbf{\Delta}$

1.6.2 The path-straightening method

The idea behind the path-straightening method is the following: take an initial path going from $\mathbf{U}$ to $\widetilde{\mathbf{U}}$. Using a gradient descent on its "energy", update the path until its curvature is minimal. The path-straightening method involves advanced theory and would ask too much details to be properly explained. Therefore, we invite the interested reader to consult [10]. Understanding how this method works is however not required to read this master's thesis and we don't use it much.

1.6.3 The vacuum algorithm

The idea behind the vacuum algorithm is to solve Problem 1.1 iteratively. This is in general hard except for $(\mathrm{St}(n,p), g_\mathrm{c})$. Given an initial path $\gamma(t) := \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix} \exp\left(\begin{bmatrix}\mathbf{A}_0 & -\mathbf{B}_0^T \\ \mathbf{B}_0 & \mathbf{C}_0\end{bmatrix} t\right) \mathbf{I}_{2p\times p}$ with $\gamma(1) = \widetilde{\mathbf{U}}$, we want to find $\mathbf{A}, \mathbf{C} \in \mathrm{Skew}(p), \mathbf{B} \in \mathbb{R}^{p\times p}$ such that

$$\exp\left(\begin{bmatrix}\mathbf{A}_0 & -\mathbf{B}_0^T \\ \mathbf{B}_0 & \mathbf{C}_0\end{bmatrix}\right) = \exp\left(\begin{bmatrix}\mathbf{A} & -\mathbf{B}^T \\ \mathbf{B} & \mathbf{0}\end{bmatrix}\right) \begin{bmatrix}\mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \exp(-\mathbf{C})\end{bmatrix}. \tag{1.42}$$

The idea is to iteratively obtain $\mathbf{A}_k, \mathbf{B}_k, \mathbf{C}_k$ so that $\|\mathbf{C}_k\|_2$ converges to zero. This is done by choosing $\mathbf{\Gamma}_k$ and computing

$$\begin{bmatrix}\mathbf{A}_{k+1} & -\mathbf{B}_{k+1}^T \\ \mathbf{B}_{k+1} & \mathbf{C}_{k+1}\end{bmatrix} := \log\left(\exp\left(\begin{bmatrix}\mathbf{A}_k & -\mathbf{B}_k^T \\ \mathbf{B}_k & \mathbf{C}_k\end{bmatrix}\right)\begin{bmatrix}\mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \exp(\mathbf{\Gamma}_k)\end{bmatrix}\right). \tag{1.43}$$

It was shown in [47] that $\mathbf{\Gamma}_k = -\mathbf{C}_k$ is a convergent choice. [48, Equation 19] introduced an update rule for $\mathbf{\Gamma}_k$ requiring to solve a *Sylverster equation* (see Algorithm 3). This update brings $\mathbf{C}_{k+1}$ faster to zero by cancelling more terms of the BCH series of the log. [48, Proposition 7] shows that this update rule yields at least a linear asymptotic convergence of $\|\mathbf{C}_{k+1}\|_2$ to zero. The procedure is detailed in Algorithm 3.

Algorithm 3 Vacuum algorithm, canonical metric ($\beta = \frac{1}{2}$) from [48]

Input: $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$, $\epsilon > 0$
$\mathbf{M} \leftarrow \mathbf{U}^T\widetilde{\mathbf{U}}$
$\mathbf{QN} \leftarrow \widetilde{\mathbf{U}} - \mathbf{UM}$
$\mathbf{V}_0 \leftarrow \begin{bmatrix} \mathbf{M} & \mathbf{X}_0 \\ \mathbf{N} & \mathbf{Y}_0 \end{bmatrix} \in \mathrm{SO}(2p)$
for $k = 0, 1, 2...$ **do**
  $\begin{bmatrix} \mathbf{A}_k & -\mathbf{B}_k^T \\ \mathbf{B}_k & \mathbf{C}_k \end{bmatrix} \leftarrow \log_m(\mathbf{V}_k)$
  if $\|\mathbf{C}_k\|_2 \leq \epsilon$ **then**
    break
  end if
  $\mathbf{S}_k \leftarrow \frac{1}{12}\mathbf{B}_k\mathbf{B}_k^T - \frac{1}{2}\mathbf{I}_p$
  Solve $\mathbf{C}_k = \mathbf{S}_k\mathbf{\Gamma} + \mathbf{\Gamma}\mathbf{S}_k$ for $\mathbf{\Gamma}$
  $\mathbf{\Phi}_k \leftarrow \exp_m(\mathbf{\Gamma})$
  $\mathbf{V}_{k+1} \leftarrow \mathbf{V}_k \begin{bmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}_k \end{bmatrix}$
end for
return $\mathbf{\Delta} := \mathbf{U}\mathbf{A}_k + \mathbf{Q}\mathbf{B}_k$

Chapter 2

Geodesics and distances on the Special Orthogonal group

This chapter provides a review of the characteristics of the Special Orthogonal group of matrices $\mathrm{SO}(n)$. We define $\mathrm{SO}(n)$ as the set of matrices $\mathbf{Q} \in \mathbb{R}^{n\times n}$ such that $\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}$ and $\det(\mathbf{Q}) = 1$. It is commonly referred to as the set of n-dimensional *rotations*, in contrast to the set of n-dimensional *reflections*, which consists of orthonormal matrices with *negative* unit determinant. Studying $\mathrm{SO}(n)$ is a relevant starting point for analyzing the properties of the Stiefel manifold $\mathrm{St}(n, p)$, which is a more complex case. Indeed, many of the useful geometric interpretations that we use in our analysis of $\mathrm{St}(n, p)$ are derived from the properties of $\mathrm{SO}(n)$. For instance, since $\mathrm{St}(n, p) = \mathrm{SO}(n)/\mathrm{SO}(n-p)$, the canonical geodesics of $\mathrm{St}(n, p)$ are geodesics of $\mathrm{SO}(n)$ from which only the first p columns are kept. This highlights the importance of first reviewing the properties of the Special Orthogonal group. While this chapter is relatively brief, it is kept separate from the following chapter to clearly distinguish new results from previously known properties.

2.1 Minimizing geodesics and distance on $\mathrm{SO}(n)$

The goal of this section is to provide a closed form expression for the distance function on $\mathrm{SO}(n)$. A formula is provided by Edelman et al. [15]. This first section is dedicated to the re-derivation of this distance function on $\mathrm{SO}(n)$.

Lemma 2.1. *Let $\gamma : [0, 1] \to \mathrm{SO}(n) : t \to \mathbf{Q}_0 e^{\mathbf{A}t}$ where $\mathbf{Q}_0 \in \mathrm{SO}(n)$ and $\mathbf{A} \in \mathrm{Skew}(n)$. If $\mathbf{A} = \mathbf{V} i \boldsymbol{\Theta} \mathbf{V}^*$ is the eigenvalue decomposition of $\mathbf{A}$, then $l(\gamma) = \sqrt{\mathrm{Tr}(\boldsymbol{\Theta}^T\boldsymbol{\Theta})} = \|\mathbf{A}\|_{\mathrm{F}}$*

Proof. By the definition (1.17) of the length of a curve γ, we have

$$l(\gamma) = \int_0^1 \left\| \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{Q}_0 e^{\mathbf{A}t}) \right\| \,\mathrm{d}t = \int_0^1 \|\mathbf{Q}_0 e^{\mathbf{A}t}\mathbf{A}\| \,\mathrm{d}t.$$

The definition of the Riemannian metric (1.34) yields

$$\begin{aligned}
\|\mathbf{Q}_0 e^{\mathbf{A}t}\mathbf{A}\|^2 &= \mathrm{Tr}\left((\mathbf{Q}_0 e^{\mathbf{A}t}\mathbf{A})^T \mathbf{Q}_0 e^{\mathbf{A}t}\mathbf{A}\right) \\
&= \mathrm{Tr}\left(\mathbf{A}^T (e^{\mathbf{A}t})^T \mathbf{Q}_0^T \mathbf{Q}_0 e^{\mathbf{A}t}\mathbf{A}\right) \\
&= \mathrm{Tr}(\mathbf{A}^T\mathbf{A}) && \text{Exponential of skew-symmetric is unitary} \\
&= \mathrm{Tr}(\boldsymbol{\Theta}^T\boldsymbol{\Theta}). && \text{Unitary invariance, see e.g., [21, Section 2.5]}
\end{aligned}$$

Finally, we have $l(\gamma) = \sqrt{\mathrm{Tr}(\mathbf{\Theta}^T\mathbf{\Theta})} = \|\mathbf{A}\|_{\mathrm{F}}$. □

Lemma 2.1 emphasizes the relation between the length of a geodesic and the eigenvalues of the matrix $\mathbf{A}$ characterizing the geodesic. Although slightly modified, the essence of Lemma 2.1 is conserved on $(\mathrm{St}(n,p), g_c)$. Now, we can prove a necessary and sufficient condition to have a minimizing geodesic on $\mathrm{SO}(n)$. The Theorem 2.1 and its strategy of proof are essentially equivalent to [13, Theorem 4.11].

Theorem 2.1 ([13]). *For all $\mathbf{Q}_0 \in \mathrm{SO}(n)$ and $\mathbf{A} \in \mathrm{Skew}(n)$, let $\gamma_{\mathbf{A}} : [0,1] \to \mathrm{SO}(n) : t \to \mathbf{Q}_0 e^{\mathbf{A}t}$ be a geodesic curve. Then $\gamma_{\mathbf{A}}$ is a minimizing geodesic from $\mathbf{Q}_0$ to $\gamma_{\mathbf{A}}(1)$ if and only if $\|\mathbf{A}\|_2 \leq \pi$.*

Proof. ($\Longrightarrow$): We show the contrapositive. If $\|\mathbf{A}\|_2 > \pi$, then there exists a shorter curve connecting $\mathbf{Q}_0$ to $\gamma_{\mathbf{A}}(1)$. The definition of $\gamma_{\mathbf{A}}(t)$ yields:

$$\gamma_{\mathbf{A}}(t) = \mathbf{Q}_0 e^{\mathbf{A}t} \iff \mathbf{Q}_0^T \gamma_{\mathbf{A}}(t) = e^{\mathbf{A}t} = \mathbf{V} e^{i\mathbf{\Theta}t}\mathbf{V}^*$$

where $\mathbf{A} = \mathbf{V} i \mathbf{\Theta} \mathbf{V}^*$. Suppose there exists $(i,j) \in \{1,...,n\}^2$, with $i \neq j$ and $s \in \mathbb{N} \setminus \{0\}$ such that $\mathbf{\Theta}_{ii} = -\mathbf{\Theta}_{jj} > 0$ and $\pi s < |\mathbf{\Theta}_{ii}| = |\mathbf{\Theta}_{jj}| \leq \pi(s+1)$. Then build $\widetilde{\mathbf{\Theta}}$ such that:

$$\begin{cases} \widetilde{\mathbf{\Theta}}_{kk} = \mathbf{\Theta}_{kk} \text{ for all } k \in \{1,...,n\} \setminus \{i,j\} \\ \widetilde{\mathbf{\Theta}}_{ii} = \mathbf{\Theta}_{ii} - 2\pi\lceil \frac{s}{2} \rceil \ \in [-\pi, \pi] \\ \widetilde{\mathbf{\Theta}}_{jj} = \mathbf{\Theta}_{jj} + 2\pi\lceil \frac{s}{2} \rceil \ \in [-\pi, \pi] \end{cases}.$$

Let $\mathbf{B} := \mathbf{V} i \widetilde{\mathbf{\Theta}} \mathbf{V}^*$ so that $e^{\mathbf{A}} = \mathbf{V} e^{i\mathbf{\Theta}} \mathbf{V}^* = \mathbf{V} e^{i\widetilde{\mathbf{\Theta}}} \mathbf{V}^* = e^{\mathbf{B}}$, i.e., $\mathbf{B}$ defines a geodesic $\gamma_{\mathbf{B}}$ connecting $\mathbf{Q}_0$ and $\gamma_{\mathbf{A}}(1)$. Comparing the lengths of $\gamma_{\mathbf{A}}$ and $\gamma_{\mathbf{B}}$, we obtain

$$\mathrm{Tr}(\mathbf{\Theta}^T\mathbf{\Theta}) > \mathrm{Tr}(\widetilde{\mathbf{\Theta}}^T\widetilde{\mathbf{\Theta}}) \Longrightarrow l(\gamma_{\mathbf{A}}) > l(\gamma_{\mathbf{B}}).$$

The geodesic defined by $\mathbf{B}$ is strictly shorter than the one defined by $\mathbf{A}$.

($\Longleftarrow$): Suppose $\|\mathbf{A}\|_2 \leq \pi$. Again, we have $\mathbf{Q}_0^T\gamma_{\mathbf{A}}(1) = e^{\mathbf{A}}$. Notice that $e^{\mathbf{A}}$ is uniquely defined because so is $\mathbf{Q}_0^T\gamma_{\mathbf{A}}(1)$. Then for all $\mathbf{B} \in \mathrm{Skew}(n)$ such that $e^{\mathbf{B}} = e^{\mathbf{A}}$, we have that $\mathbf{A}$ and $\mathbf{B}$ share the same eigenvectors (up to a unit modulus complex factor) and can only differ by their eigenvalues. We write

$$\mathbf{A} = \mathbf{V} i \mathbf{\Theta} \mathbf{V}^* \text{ and } \mathbf{B} = \mathbf{V} i \widetilde{\mathbf{\Theta}} \mathbf{V}^*.$$

The necessary condition $e^{\mathbf{B}} = e^{\mathbf{A}}$ yields $e^{i\widetilde{\mathbf{\Theta}}} = e^{i\mathbf{\Theta}}$ or equivalently $\widetilde{\mathbf{\Theta}}_{jj} = \mathbf{\Theta}_{jj} + 2k_j\pi$ for some $k_j \in \mathbb{N}$ for all $j \in \{1,...,n\}$. Hence $\mathrm{Tr}(\mathbf{\Theta}^T\mathbf{\Theta}) \leq \mathrm{Tr}(\widetilde{\mathbf{\Theta}}^T\widetilde{\mathbf{\Theta}}) \Longrightarrow l(\gamma_{\mathbf{A}}) \leq l(\gamma_{\mathbf{B}})$. Thus $\gamma_{\mathbf{A}}$ is a geodesic of minimal length connecting $\mathbf{Q}_0$ and $\gamma_{\mathbf{A}}(1)$. □

Corollary 2.1. *For all $\mathbf{A} \in \mathrm{Skew}(n)$ and $\mathbf{Q}_0 \in \mathrm{SO}(n)$, $\gamma : \mathbb{R}_+ \to \mathrm{SO}(n) : t \to \mathbf{Q}_0 e^{\mathbf{A}t}$ is a minimizing geodesic from $\mathbf{Q}_0$ to $\gamma_{\mathbf{A}}(t)$ if and only if $t \leq \frac{\pi}{\|\mathbf{A}\|_2}$.*

Proof. This is a direct consequence of Theorem 2.1. $\|\mathbf{A}t\|_2 \leq \pi$ if and only if $t \leq \frac{\pi}{\|\mathbf{A}\|_2}$. □

Theorem 2.1 recalls the geometric interpretation inherited from the hypersphere where the cut locus of any point is the antipodal point. The analogous of an antipodal point for any matrix $\mathbf{Q} \in \mathrm{SO}(n)$ is thus the set of matrices obtained by applying a rotation which has an eigenvalue of modulus π exactly. Theorem 2.2 finally yields the "straightforward" proposition of Edelman et al. [15, Equation 2.15].

Theorem 2.2 ([15]). *For all* $\mathbf{Q}, \widetilde{\mathbf{Q}} \in \mathrm{SO}(n)$*, let* $\widetilde{\mathbf{Q}} = \mathbf{Q}e^{\mathbf{A}}$ *where* $\mathbf{A} \in \mathrm{Skew}(n)$*. If* $\mathbf{A} = \mathbf{V}i\mathbf{\Theta}\mathbf{V}^*$ *is an eigenvalue decomposition and* $\|\mathbf{A}\|_2 \leq \pi$*. Then* $d(\mathbf{Q}, \widetilde{\mathbf{Q}}) = \sqrt{\mathbf{\Theta}^T\mathbf{\Theta}} = \|\mathbf{A}\|_{\mathrm{F}}$.

Proof. Using Theorem 2.1, $\gamma : [0,1] \to \mathrm{SO}(n) : t \to \mathbf{Q}e^{\mathbf{A}t}$ is a minimizing geodesic from $\mathbf{Q}$ to $\widetilde{\mathbf{Q}}$. If Γ is the set of all the geodesics on $\mathrm{SO}(n)$, we know from the Hopf-Rinow theorem that

$$d(\mathbf{Q}, \widetilde{\mathbf{Q}}) = \min\{l(\gamma) \mid \gamma \in \Gamma,\ \gamma(0) = \mathbf{Q},\ \gamma(1) = \widetilde{\mathbf{Q}}\}. \tag{2.1}$$

The latter expression yields $d(\mathbf{Q}, \widetilde{\mathbf{Q}}) = \sqrt{\mathbf{\Theta}^T\mathbf{\Theta}} = \|\mathbf{A}\|_{\mathrm{F}}$ using Lemma 2.1. □

Corollary 2.2 ([13]). *The diameter of* $\mathrm{SO}(n)$ *is* $\pi\sqrt{2\left\lfloor \frac{n}{2} \right\rfloor}$.

Proof. For all $\widetilde{\mathbf{Q}} := \mathbf{Q}e^{\mathbf{A}} \in \mathrm{SO}(n)$, in view of Theorem 2.2, the maximal achievable distance is given when all the eigenvalues of $\mathbf{A}$ have a modulus π except the mandatory zero eigenvalue of $\mathbf{A}$ if n is odd, which is responsible for $\lfloor \cdot \rfloor$. □

2.2 The cut locus on SO(n)

The results of section 2.3 characterize the cut locus on $\mathrm{SO}(n)$.

Property 2.1. *Let* $\mathbf{A} \in \mathrm{Skew}(n)$*. The cut time at* $\mathbf{Q}$ *in the direction* $\mathbf{\Delta} = \mathbf{Q}\mathbf{A} \in T_{\mathbf{Q}}\mathrm{SO}(n)$ *is* $\mathcal{T}_{\mathbf{Q}}(\mathbf{\Delta}) := \frac{\pi}{\|\mathbf{A}\|_2}$.

Proof. This is a direct consequence of Corollary 2.1 □

From this property, we immediately get the injectivity domain, the interior set, the tangent cut locus and the cut locus of $\mathrm{SO}(n)$

Property 2.2. *The tangent cut locus at* $\mathbf{Q} \in \mathrm{SO}(n)$ *is* $\hat{C}_{\mathbf{Q}} = \{\mathbf{Q}\mathbf{A} \in \mathrm{Skew}(n) \mid \|\mathbf{A}\|_2 = \pi\}$.

Property 2.3. *The cut locus at* $\mathbf{Q} \in \mathrm{SO}(n)$ *is*

$$\begin{aligned} C_{\mathbf{Q}} &= \{\mathbf{Q}e^{\mathbf{A}} \mid \mathbf{A} \in \mathrm{Skew}(n) \text{ and } \|\mathbf{A}\|_2 = \pi\} \\ &= \{\mathbf{Q}\hat{\mathbf{Q}} \mid \hat{\mathbf{Q}} \in \mathrm{SO}(n) \text{ and } -1 \text{ belongs to the spectrum of } \hat{\mathbf{Q}}\}. \end{aligned}$$

The latter expression holds since $e^{i\pi} = -1$, the praised Euler formula.

2.3 Properties of rotational invariance

We show that distances on $\mathrm{SO}(n)$ are invariant under right and left rotations.

Lemma 2.2. *For all* $\mathbf{Q}, \widetilde{\mathbf{Q}}, \mathbf{R} \in \mathrm{SO}(n)$*, we have* $d(\mathbf{Q}, \widetilde{\mathbf{Q}}) = d(\mathbf{Q}\mathbf{R}, \widetilde{\mathbf{Q}}\mathbf{R}) = d(\mathbf{R}\mathbf{Q}, \mathbf{R}\widetilde{\mathbf{Q}})$

Proof. Let $\mathbf{\Delta} \in \mathrm{Log}_{\mathbf{Q}}(\widetilde{\mathbf{Q}})$ and $\mathbf{\Delta} = \mathbf{Q}\mathbf{A}_1 = \mathbf{A}_2\mathbf{Q}$ where $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Skew}(n)$ and $\mathbf{A}_2 = \mathbf{Q}\mathbf{A}_1\mathbf{Q}^T$. By definition, we have $\widetilde{\mathbf{Q}} = \mathbf{Q}e^{\mathbf{A}_1} = e^{\mathbf{A}_2}\mathbf{Q}$. This yields,

$$\begin{aligned} \widetilde{\mathbf{Q}} &= \mathbf{Q}e^{\mathbf{A}_1} \qquad \text{and} \qquad \widetilde{\mathbf{Q}} = e^{\mathbf{A}_2}\mathbf{Q} \\ \iff \mathbf{R}\widetilde{\mathbf{Q}} &= \mathbf{R}\mathbf{Q}e^{\mathbf{A}_1} \qquad \text{and} \qquad \widetilde{\mathbf{Q}}\mathbf{R} = e^{\mathbf{A}_2}\mathbf{Q}\mathbf{R}. \end{aligned}$$

Using Theorem 2.1, we have $\mathbf{R}\mathbf{\Delta} \in \mathrm{Log}_{\mathbf{R}\mathbf{Q}}(\mathbf{R}\widetilde{\mathbf{Q}})$ and $\mathbf{\Delta}\mathbf{R} \in \mathrm{Log}_{\mathbf{Q}\mathbf{R}}(\widetilde{\mathbf{Q}}\mathbf{R})$. Hence $d(\mathbf{Q}, \widetilde{\mathbf{Q}}) = d(\mathbf{Q}\mathbf{R}, \widetilde{\mathbf{Q}}\mathbf{R}) = d(\mathbf{R}\mathbf{Q}, \mathbf{R}\widetilde{\mathbf{Q}})$. □

Leveraging Lemma 2.2 is a first opportunity to make a link between $\mathrm{St}(n,p)$ and $\mathrm{SO}(n)$. We show in Corollary 2.3 that if one is looking for the closest pair of matrices in $\mathrm{SO}(n)$ such that the p first columns of both matrices are fixed, then one can arbitrarily complete the first matrix entirely and only look for the optimal completion of the second matrix.

Corollary 2.3. *For all* $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$ *and* $\mathbf{U}_\perp, \widetilde{\mathbf{U}}_\perp \in \mathrm{St}(n, n-p)$ *respective orthogonal completions , we have* $\min_{\mathbf{Q},\widetilde{\mathbf{Q}} \in \mathrm{SO}(n-p)} d\left(\begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{I}_p & \mathbf{0}\\ \mathbf{0} & \mathbf{Q}\end{bmatrix}, \begin{bmatrix}\widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{I}_p & \mathbf{0}\\ \mathbf{0} & \widetilde{\mathbf{Q}}\end{bmatrix}\right)$
$= \min_{\mathbf{Q} \in \mathrm{SO}(n-p)} d\left(\begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}, \begin{bmatrix}\widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{I}_p & \mathbf{0}\\ \mathbf{0} & \mathbf{Q}\end{bmatrix}\right)$

Proof. Apply Lemma 2.2 with $\mathbf{R} = \begin{bmatrix}\mathbf{I}_p & \mathbf{0}\\ \mathbf{0} & \mathbf{Q}^{-1}\end{bmatrix}$. □

2.4 Conclusion

The goal of this chapter was to understand and re-derive by ourselves the definition and the properties of the geodesic distance on $\mathrm{SO}(n)$. To do so, we have used theoretical tools and abstract concepts that can be leveraged to extend the current knowledge about the Stiefel manifold, especially when it is endowed with the Euclidean or canonical metrics.

Chapter 3

Geodesics and distances on the Stiefel manifold

In this chapter, we present the new results we have obtained in the framework of this master's thesis with the aim to improve the performances of minimizing geodesic computation algorithms. In the first part, we show the bilipschitz equivalence of the β-metrics between each other and w.r.t the Frobenius distance. Afterwards, we leverage geometric insights from the hypersphere in $\mathbb{R}^{n\times p}$ to obtain (i) tight bounds on the β-distances in terms of the easy-to-compute Frobenius distance and (ii) the families of matrices of the Stiefel manifold reaching these bounds. Formally, we try to obtain closed form formulas for $\underline{b}_\beta(\delta) := \min\{d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) \mid \mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p), \|\mathbf{U} - \widetilde{\mathbf{U}}\|_\mathrm{F} = \delta\}$ and $\overline{b}_\beta(\delta) := \max\{d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) \mid \mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p), \|\mathbf{U} - \widetilde{\mathbf{U}}\|_\mathrm{F} = \delta\}$. For some triplet (β, n, p) we succeed to obtain these bounds. For others, we are left with data-driven conjectures. In particular, we obtain the diameter of $(\mathrm{St}(n,p), g_\mathrm{E})$. To conclude, we prove several necessary conditions for *canonical* geodesics to be minimizing.

3.1 Equivalence of the geodesic distances

In this section, we show that for any pair $\beta_1, \beta_2 > 0$, the distances d_{β_1} and d_{β_2} are bilipschitz equivalent, i.e., there is $\alpha, \delta > 0$ such that for all $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$, we have $\alpha d_{\beta_1}(\mathbf{U}, \widetilde{\mathbf{U}}) \leq d_{\beta_2}(\mathbf{U}, \widetilde{\mathbf{U}}) \leq \delta d_{\beta_1}(\mathbf{U}, \widetilde{\mathbf{U}})$.

Lemma 3.1. *If* $\|\mathbf{\Delta}\|_\mathrm{a} \leq \mu \|\mathbf{\Delta}\|_\mathrm{b}$ *for all* $\mathbf{U} \in \mathrm{St}(n,p)$ *and* $\mathbf{\Delta} \in T_\mathbf{U}\mathrm{St}(n,p)$*, then* $d_\mathrm{a}(\mathbf{U}, \widetilde{\mathbf{U}}) \leq \mu d_\mathrm{b}(\mathbf{U}, \widetilde{\mathbf{U}})$ *for all* $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$.

Proof. Let $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$. For every differentiable curve $\gamma : [0,1] \to \mathrm{St}(n,p)$ with $\gamma(0) = \mathbf{U}$ and $\gamma(1) = \widetilde{\mathbf{U}}$, it holds that

$$d_\mathrm{a}(\mathbf{U}, \widetilde{\mathbf{U}}) \leq l_\mathrm{a}(\gamma) = \int_0^1 \|\dot{\gamma}(t)\|_\mathrm{a} \,\mathrm{d}t \leq \mu \int_0^1 \|\dot{\gamma}(t)\|_\mathrm{b} \,\mathrm{d}t = \mu l_\mathrm{b}(\gamma).$$

Hence, we obtain

$$d_\mathrm{a}(\mathbf{U}, \widetilde{\mathbf{U}}) \leq \mu \inf\{l_b(\gamma) \mid \gamma(0) = \mathbf{U},\ \gamma(1) = \widetilde{\mathbf{U}}\} = \mu d_\mathrm{b}(\mathbf{U}, \widetilde{\mathbf{U}}).$$

$\square$

Theorem 3.1.

(a) If $0 < \beta_2 \le \beta_1$*, then* $d_{\beta_1}(\mathbf{U}, \tilde{\mathbf{U}}) \le \sqrt{\frac{\beta_1}{\beta_2}} d_{\beta_2}(\mathbf{U}, \tilde{\mathbf{U}})$ *for all* $\mathbf{U}, \tilde{\mathbf{U}} \in \mathrm{St}(n,p)$*.*

(b) If $0 < \beta_1 \le \beta_2$*, then* $d_{\beta_1}(\mathbf{U}, \tilde{\mathbf{U}}) \le d_{\beta_2}(\mathbf{U}, \tilde{\mathbf{U}})$ *for all* $\mathbf{U}, \tilde{\mathbf{U}} \in \mathrm{St}(n,p)$*.*

Proof. (a) Let $0 < \beta_2 \le \beta_1$. For all $\hat{\mathbf{U}} \in \mathrm{St}(n,p)$ and all $\mathbf{\Delta} \in T_{\hat{\mathbf{U}}}\mathrm{St}(n,p)$, we have

$$\|\mathbf{\Delta}\|_{\beta_1}^2 = \beta_1\|\mathbf{A}\|_{\mathrm{F}}^2 + \|\mathbf{B}\|_{\mathrm{F}}^2 \le \beta_1\|\mathbf{A}\|_{\mathrm{F}}^2 + \frac{\beta_1}{\beta_2}\|\mathbf{B}\|_{\mathrm{F}}^2 = \frac{\beta_1}{\beta_2}\|\mathbf{\Delta}\|_{\beta_2}^2, \tag{3.1}$$

hence $\|\mathbf{\Delta}\|_{\beta_1}^2 \le \frac{\beta_1}{\beta_2}\|\mathbf{\Delta}\|_{\beta_2}^2$. By Lemma 3.1, the claim follows.

(b) Let $0 < \beta_1 \le \beta_2$. For all $\hat{\mathbf{U}} \in \mathrm{St}(n,p)$ and all $\mathbf{\Delta} \in T_{\hat{\mathbf{U}}}\mathrm{St}(n,p)$, we have

$$\|\mathbf{\Delta}\|_{\beta_1}^2 = \beta_1\|\mathbf{A}\|_{\mathrm{F}}^2 + \|\mathbf{B}\|_{\mathrm{F}}^2 \le \beta_2\|\mathbf{A}\|_{\mathrm{F}}^2 + \|\mathbf{B}\|_{\mathrm{F}}^2 = \|\mathbf{\Delta}\|_{\beta_2}^2, \tag{3.2}$$

hence $\|\mathbf{\Delta}\|_{\beta_1}^2 \le \|\mathbf{\Delta}\|_{\beta_2}^2$. By Lemma 3.1, the claim follows.

$\square$

Corollary 3.1 summarizes the bilipchitz equivalence between all the geodesic distances under the β-metrics.

Corollary 3.1. $\min\left\{1, \sqrt{\frac{\beta_1}{\beta_2}}\right\} d_{\beta_2}(\mathbf{U}, \tilde{\mathbf{U}}) \le d_{\beta_1}(\mathbf{U}, \tilde{\mathbf{U}}) \le \max\left\{1, \sqrt{\frac{\beta_1}{\beta_2}}\right\} d_{\beta_2}(\mathbf{U}, \tilde{\mathbf{U}})$ *for all* $\mathbf{U}, \tilde{\mathbf{U}} \in \mathrm{St}(n,p)$*.*

Proof. Combine Theorems 3.1.(a) and 3.1.(b). $\square$

Corollary 3.2. $\frac{\sqrt{2}}{2} d_{\mathrm{E}}(\mathbf{U}, \tilde{\mathbf{U}}) \le d_{\mathrm{c}}(\mathbf{U}, \tilde{\mathbf{U}}) \le d_{\mathrm{E}}(\mathbf{U}, \tilde{\mathbf{U}}) \le \sqrt{2} d_{\mathrm{c}}(\mathbf{U}, \tilde{\mathbf{U}})$ *for all* $\mathbf{U}, \tilde{\mathbf{U}} \in \mathrm{St}(n,p)$*.*

Proof. Recall that $d_{\mathrm{E}} = d_{\beta=1}$ and $d_{\mathrm{c}} = d_{\beta=\frac{1}{2}}$. $\square$

3.2 Tightness of the bounds on the β-metrics

In this section, we show that all the bounds of section 3.1 are tight. By tight, we mean here that it is not possible to find better lipschitz constants.

Theorem 3.2.

(a) The bound of Theorem 3.1.(a) is tight.

(b) The bound of Theorem 3.1.(b) is tight.

Proof. (a) Let $0 < \beta_2 \le \beta_1$, $\mathbf{U} \in \mathrm{St}(n,p)$ and $\mathbf{\Delta} = \mathbf{U}\mathbf{A} + \mathbf{U}_\perp\mathbf{B} \in T_{\mathbf{U}}\mathrm{St}(n,p)$. Take $\mathbf{B} = 0$ and observe that (1.26) reduces to

$$\begin{aligned}\gamma_\beta(t) &= \exp\left(2\beta t\mathbf{U}\mathbf{A}\mathbf{U}^T\right)\mathbf{U}\exp\left[(1-2\beta)\mathbf{A}t\right] \\ &= \mathbf{U}\exp\left(2\beta t\mathbf{A}\right)\exp\left[(1-2\beta)\mathbf{A}t\right] \\ &= \mathbf{U}e^{\mathbf{A}t},\end{aligned}$$

where we have used the definition of the matrix exponential and the fact that $\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$. Hence $\gamma_\beta(t)$ does not depend on β. Further take $\mathbf{A}$ small enough for γ_β to be minimizing between $\mathbf{U} = \gamma_\beta(0)$ and $\widetilde{\mathbf{U}} := \gamma_\beta(1)$. Then

$$d_{\beta_1}(\mathbf{U}, \widetilde{\mathbf{U}}) = l_{\beta_1}(\gamma_\beta) = \|\mathbf{\Delta}\|_{\beta_1} = \sqrt{\frac{\beta_1}{\beta_2}}\|\mathbf{\Delta}\|_{\beta_2} = \sqrt{\frac{\beta_1}{\beta_2}} l_{\beta_2}(\gamma_\beta) = \sqrt{\frac{\beta_1}{\beta_2}} d_{\beta_2}(\mathbf{U}, \widetilde{\mathbf{U}}).$$

(b) Let $0 < \beta_1 \leq \beta_2$, $\mathbf{U} \in \mathrm{St}(n,p)$ and $\mathbf{\Delta} = \mathbf{U}\mathbf{A} + \mathbf{U}_\perp\mathbf{B} \in T_{\mathbf{U}}\mathrm{St}(n,p)$. Take $\mathbf{A} = 0$ and observe that (1.26) reduces to

$$\begin{aligned}\gamma_\beta(t) &= \exp\left(t(\mathbf{U}_\perp\mathbf{B}\mathbf{U}^T - \mathbf{U}\mathbf{B}^T\mathbf{U}_\perp^T)\right)\mathbf{U} \\ &= \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}\exp\left(t\begin{bmatrix}0 & -\mathbf{B}^T \\ \mathbf{B} & 0\end{bmatrix}\right)\mathbf{I}_{n\times p}.\end{aligned}$$

Again, $\gamma_\beta(t)$ does not depend on β. Further take B small enough for γ_β to be minimizing between $\mathbf{U} = \gamma_\beta(0)$ and $\widetilde{\mathbf{U}} := \gamma_\beta(1)$. Then

$$d_{\beta_1}(\mathbf{U}, \widetilde{\mathbf{U}}) = l_{\beta_1}(\gamma_\beta) = \|\mathbf{\Delta}\|_{\beta_1} = \|\mathbf{\Delta}\|_{\beta_2} = l_{\beta_2}(\gamma_\beta) = d_{\beta_2}(\mathbf{U}, \widetilde{\mathbf{U}}).$$

□

3.3 Length of the β-geodesics

In the following theorem, we establish a formula to compute the β_2-length of a β_1-geodesic given its initial velocity.

Theorem 3.3. *For all $\beta_1, \beta_2 > 0$, given γ_{β_1} a β_1-geodesic, $\mathbf{U} := \gamma_{\beta_1}(0)$, and $\mathbf{\Delta} := [\dot{\gamma}_{\beta_1}(t)]_{t=0} = \mathbf{U}\mathbf{A} + \mathbf{U}_\perp\mathbf{B}$. Then $l_{\beta_2}(\gamma_{\beta_1}) = \sqrt{\beta_2\|\mathbf{A}\|_\mathrm{F}^2 + \|\mathbf{B}\|_\mathrm{F}^2}$.*

Proof. Let $\mathbf{S} := 2(\beta_1 - 1)\mathbf{U}\mathbf{A}\mathbf{U}^T + \mathbf{\Delta}\mathbf{U}^T - \mathbf{U}\mathbf{\Delta}^T$. Starting from (1.26) and taking the derivative of γ_{β_1} yields

$$\dot{\gamma}_{\beta_1}(t) = e^{\mathbf{S}t}\mathbf{S}\mathbf{U}e^{(1-2\beta_1)\mathbf{A}t} + (1-2\beta_1)e^{\mathbf{S}t}\mathbf{U}\mathbf{A}e^{(1-2\beta_1)\mathbf{A}t}.$$

Then, for all t,

$$\begin{aligned}\|\dot{\gamma}_{\beta_1}(t)\|_{\beta_2}^2 &= \mathrm{Tr}\left(\dot{\gamma}_{\beta_1}(t)^T(\mathbf{I} - (1-\beta_2)\gamma_{\beta_1}(t)\gamma_{\beta_1}(t)^T)\dot{\gamma}_{\beta_1}(t)\right) \\ &= \mathrm{Tr}\,(\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} - (2\beta_1 - \beta_2)\mathbf{U}^T\mathbf{S}^T\mathbf{U}\mathbf{A} - (2\beta_1 - 1)\mathbf{A}^T\mathbf{U}^T\mathbf{S}\mathbf{U} \\ &\qquad + (2\beta_1 - 1)(2\beta_1 - \beta_2)\mathbf{A}^T\mathbf{A}) \\ &= [4\beta_1^2 - 2\beta_1(2\beta_1 - \beta_2) - 2\beta_1(2\beta_1 - 1) \\ &\qquad + (2\beta_1 - 1)(2\beta_1 - \beta_2)]\,\|\mathbf{A}\|_\mathrm{F}^2 + \|\mathbf{B}\|_\mathrm{F}^2 \\ &= \beta_2\|\mathbf{A}\|_\mathrm{F}^2 + \|\mathbf{B}\|_\mathrm{F}^2.\end{aligned}$$

The statement of the theorem follows directly from (1.17) since $\|\dot{\gamma}_{\beta_1}(t)\|_{\beta_2}$ is constant over time. □

3.4 A lower bound on the β-distance by the Frobenius distance

3.4.1 Stiefel and the hypersphere

Let $\mathbf{U} \in \mathrm{St}(n,p)$. Define $\mathrm{vec}(\mathbf{U}) \in \mathbb{R}^{np}$ as the vector obtained by stacking the columns of $\mathbf{U}$ one under the other such that $\mathbf{U}_{i,j} = \mathrm{vec}(\mathbf{U})_{(j-1)n+i}$. Let $\mathrm{vec}(\mathrm{St}(n,p))$ denote the set of vectors obtained by applying this operation on all the matrices of $\mathrm{St}(n,p)$. We also introduce the notation $\mathbb{S}^k_r$ for the hypersphere of dimension k, of radius r, centered at the origin. For all $\mathbf{U} \in \mathrm{St}(n,p)$, since $\|\mathbf{U}\|_\mathrm{F} = \|\mathrm{vec}(\mathbf{U})\|_2 = \sqrt{p}$, it is straightforward that $\mathrm{vec}(\mathrm{St}(n,p)) \subseteq \mathbb{S}^{np-1}_{\sqrt{p}}$.

We recall a useful property from trigonometry to compute the length of an arc on the hypersphere given the chord length. The property is found by working in the blue triangle of Figure 3.1.

Property 3.1. *For all $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{S}^k_r$, the Euclidean geodesic distance (arc-length) is given by* $d_\mathrm{E}(\mathbf{v}_1,\mathbf{v}_2) = 2r \arcsin\left(\frac{\|\mathbf{v}_1-\mathbf{v}_2\|_2}{2r}\right)$.

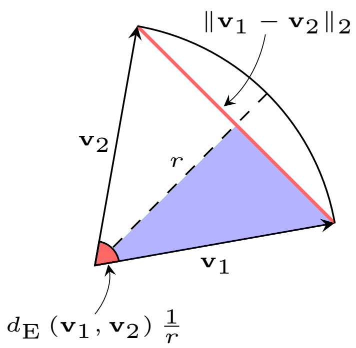

Figure 3.1: Illustration of Property 3.1

The next lemma establishes the equivalence of the Euclidean lengths in $\mathrm{St}(n,p)$ and $\mathrm{vec}(\mathrm{St}(n,p))$.

Lemma 3.2. *The length of a continuously differentiable curve $\gamma : [0,1] \to \mathrm{St}(n,p)$ under the Euclidean metric, written $l_\mathrm{E}(\gamma)$, is equal to the Euclidean length of $\mathrm{vec}(\gamma)$ in $\mathbb{S}^{np-1}_{\sqrt{p}}$, written $l_\mathrm{E}(\mathrm{vec}(\gamma))$.*

Proof. $l_\mathrm{E}(\gamma) = \int_0^1 \|\dot\gamma(t)\|_\mathrm{E}\,\mathrm{d}t = \int_0^1 \|\mathrm{vec}(\dot\gamma(t))\|_2\,\mathrm{d}t = l_\mathrm{E}(\mathrm{vec}(\gamma))$. □

Theorem 3.4 is a direct consequence of the fact that $\mathrm{St}(n,p)$ can be seen as a subset of $\mathbb{S}^{np-1}_{\sqrt{p}}$. The Euclidean shortest path between two points in $\mathrm{St}(n,p)$ can thus not be shorter than the shortest path in $\mathbb{S}^{np-1}_{\sqrt{p}}$, the last one being the well known arc-length between the two points on the sphere.

Theorem 3.4. $d_\mathrm{E}(\mathbf{U},\widetilde{\mathbf{U}}) \geq 2\sqrt{p}\arcsin\left(\frac{\|\mathbf{U}-\widetilde{\mathbf{U}}\|_\mathrm{F}}{2\sqrt{p}}\right)$ *for all* $\mathbf{U},\widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$.

Proof. Using Property 3.1, we have $d_{\mathrm{E}}(\mathrm{vec}(\mathbf{U}), \mathrm{vec}(\widetilde{\mathbf{U}})) = 2\sqrt{p}\arcsin\left(\frac{\|\mathbf{U}-\widetilde{\mathbf{U}}\|_{\mathrm{F}}}{2\sqrt{p}}\right)$ since $\|\mathrm{vec}(\mathbf{U})-\mathrm{vec}(\widetilde{\mathbf{U}})\|_2 = \|\mathbf{U}-\widetilde{\mathbf{U}}\|_{\mathrm{F}}$. Moreover, we have

$$\begin{aligned}
d_{\mathrm{E}}\left(\mathrm{vec}(\mathbf{U}), \mathrm{vec}(\widetilde{\mathbf{U}})\right) &= \min\left\{l_{\mathrm{E}}(\mathrm{vec}(\gamma)) \mid \mathrm{vec}(\gamma(t)) \in \mathbb{S}^{np}_{\sqrt{p}},\ \gamma(0) = \mathbf{U},\ \gamma(1) = \widetilde{\mathbf{U}}\right\} \\
&\leq \min\left\{l_{\mathrm{E}}(\mathrm{vec}(\gamma)) \mid \mathrm{vec}(\gamma(t)) \in \mathrm{vec}(\mathrm{St}(n,p)), \gamma(0) = \mathbf{U}, \gamma(1) = \widetilde{\mathbf{U}}\right\} \\
&= \min\left\{l_{\mathrm{E}}(\gamma) \mid \gamma(t) \in \mathrm{St}(n,p), \gamma(0) = \mathbf{U}, \gamma(1) = \widetilde{\mathbf{U}}\right\} \quad \text{Using Lemma 3.2} \\
&= d_{\mathrm{E}}(\mathbf{U}, \widetilde{\mathbf{U}}).
\end{aligned}$$

□

Corollary 3.3. $d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) \geq \min\{1, \sqrt{\beta}\}2\sqrt{p}\arcsin\left(\frac{\|\mathbf{U}-\widetilde{\mathbf{U}}\|_{\mathrm{F}}}{2\sqrt{p}}\right)$ *for all* $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$.

Proof. Combine Corollary 3.1 and Theorem 3.4. □

To conclude this section, we prove that the bound of Corollary 3.3 is tight when p is even and $\beta \in (0, 1]$. To this end, consider the 2-by-2 rotation $\mathbf{B}(\theta) := \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$ and the block diagonal matrix $\mathbf{G}(\theta)$ where

$$\mathbf{G}(\theta) := \begin{bmatrix} \mathbf{B}(\theta) & 0 & 0 & \dots & 0 & 0 \\ 0 & \mathbf{B}(\theta) & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & \mathbf{B}(\theta) \end{bmatrix} \in \mathrm{SO}(p)$$

for any $\theta \in [-\pi, \pi]$.

Theorem 3.5. *Let p be even and $\beta \in (0, 1]$. Then for all $\mathbf{U} \in \mathrm{St}(n,p)$ and all $\theta \in [-\pi, \pi]$, we have* $d_\beta(\mathbf{U}, \mathbf{U}\mathbf{G}(\theta)) = 2\sqrt{\beta p}\arcsin\left(\frac{\|\mathbf{U}-\mathbf{U}\mathbf{G}(\theta)\|_{\mathrm{F}}}{2\sqrt{p}}\right)$.

Proof. Let $\theta \in [-\pi, \pi]$ and $\mathbf{\Delta} = \mathbf{U}\mathbf{A}(\theta)$ where

$$\mathbf{A}(\theta) := \begin{bmatrix} \mathbf{\Omega}(\theta) & 0 & 0 & \dots & 0 & 0 \\ 0 & \mathbf{\Omega}(\theta) & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & \mathbf{\Omega}(\theta) \end{bmatrix} \in \mathrm{Skew}(p)$$

and $\mathbf{\Omega}(\theta) = \begin{bmatrix} 0 & \theta \\ -\theta & 0 \end{bmatrix}$. As shown in the proof of Theorem 3.2, (1.26) reduces to $\gamma_\beta(t) = \mathbf{U}\exp\left[\mathbf{A}(\theta)t\right] = \mathbf{U}\mathbf{G}(\theta t)$. This yields $\dot{\gamma}_\beta(t) = \mathbf{U}\exp\left[\mathbf{A}(\theta)t\right]\mathbf{A}(\theta)$ and $\|\dot{\gamma}_\beta(t)\|^2_\beta = \beta\|\mathbf{A}(\theta)\|^2_{\mathrm{F}} = \beta p\theta^2$. It follows from (1.17) that $l_\beta(\gamma_\beta) = \sqrt{\beta p}|\theta|$.

Morover, letting $c := \cos(\theta)$ and $s := \sin(\theta)$, we get

$$\begin{aligned}
\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}^2 &= \|\mathbf{U}(\mathbf{G}(\theta) - \mathbf{I})\|_{\mathrm{F}}^2 \\
&= \sum_{i=1,i+=2}^{p-1} \sum_{j=1}^{n} ((c-1)\mathbf{U}_{j,i} - s\mathbf{U}_{j,i+1})^2 + (s\mathbf{U}_{j,i} + (c-1)\mathbf{U}_{j,i+1})^2 \\
&= \sum_{i=1,i+=2}^{p-1} \sum_{j=1}^{n} (c-1)^2\mathbf{U}_{j,i}^2 + s^2\mathbf{U}_{j,i+1}^2 + s^2\mathbf{U}_{j,i}^2 + (c-1)^2\mathbf{U}_{j,i+1}^2 \\
& \qquad\qquad\qquad\qquad \text{Since the columns are orthogonal} \\
&= p[(c-1)^2 + s^2] \qquad\qquad \text{Since the columns are normed} \\
&= 2p(1 - \cos(\theta)) \\
&= 4p\sin^2(\frac{\theta}{2}).
\end{aligned}$$

This yields $|\theta| = 2\arcsin\frac{\|\mathbf{U}-\widetilde{\mathbf{U}}\|_{\mathrm{F}}}{2\sqrt{p}}$ and $l_\beta(\gamma_\beta) = 2\sqrt{\beta p}\arcsin\left(\frac{\|\mathbf{U}-\widetilde{\mathbf{U}}\|_{\mathrm{F}}}{2\sqrt{p}}\right)$. In view of Corollary 3.3, γ_β is thus a *minimizing* geodesic. Hence, letting $\widetilde{\mathbf{U}} := \gamma_\beta(1)$, we have

$$d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) = l_\beta(\gamma_\beta) = 2\sqrt{\beta p}\arcsin\left(\frac{\|\mathbf{U} - \widetilde{\mathbf{U}}\|_{\mathrm{F}}}{2\sqrt{p}}\right).$$

$\square$

Remark on the matrices of Theorem 3.5 The eigenvalues of $\mathbf{G}(\theta)$ are invariant under similarity permutations ($\mathbf{P}\mathbf{G}(\theta)\mathbf{P}^T$ where $\mathbf{P}$ is a permutation matrix). Hence, if we define $\widetilde{\mathbf{U}} = \mathbf{U}\mathbf{P}\mathbf{G}(\theta)\mathbf{P}^T$, the development of Theorem 3.5 computing the Frobenius distance in terms of θ still holds. Indeed, if $\mathbf{G}(\theta)$ is permuted, the combinations are made with different couples of columns but the subsequent steps remain true.

Remark on antipodal points We still consider p even and $\beta \in (0, 1]$. For all $\mathbf{U} \in \mathrm{St}(n, p)$, we define the antipodal point $\widetilde{\mathbf{U}} := \mathbf{U}\mathbf{G}(\pi) = -\mathbf{U}$. Using Theorem 3.5, antipodal points have a β-distance equal to $\pi\sqrt{\beta p}$. In particular, we have $d_{\mathrm{c}}(\mathbf{U}, -\mathbf{U}) = \pi\frac{\sqrt{2p}}{2}$ and $d_{\mathrm{E}}(\mathbf{U}, -\mathbf{U}) = \pi\sqrt{p}$. This distance is exactly the diameter of $\mathbb{S}^{np-1}_{\sqrt{p}}$ in the ambient space $\mathbb{R}^{np}$. In consequence, the β-geodesics between antipodal points with p even must correspond to great circles in $\mathbb{S}^{np-1}_{\sqrt{p}}$. Further in this work, we will see that $\pi\sqrt{p}$ is in fact the diameter of Stiefel endowed with its Euclidean metric, $(\mathrm{St}(n, p), g_{\mathrm{E}})$.

3.4.2 Numerical corroboration of the bounds

We have shown theoretically the tightness of the bound of Theorem 3.4 and Corollary 3.3 for p even and $\beta \in (0, 1]$. Let us perform a numerical experiment using Algorithm 3 ([48, Algorithm 4]) and $\beta = \frac{1}{2}$ (canonical metric). This will allow to (i) corroborate the bounds obtained and (ii) verify the ability of Algorithm 3 to compute admissible distances between the points. In order to build Figure 3.2, we sampled pairs of matrices $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n, p)$ following the procedure of Appendix A. Appendix A emphasizes the difficulty encountered to *randomly* sample pairs of matrices providing all possible Frobenius distances and geodesic distances due to the so-called "*curse of dimensionality*".

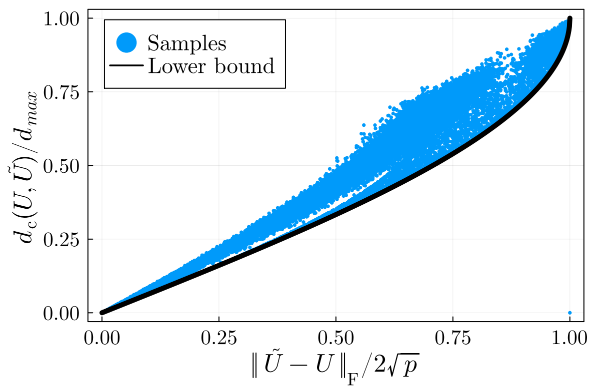

Figure 3.2: Distances computed on 25000 random samples $(n,p) = (8,4)$, $d_{max} := \frac{\pi\sqrt{2p}}{2}$.

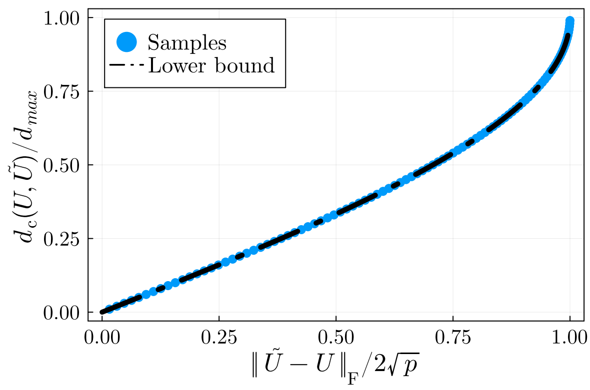

Figure 3.3: Distances computed on matrices $\widetilde{\mathbf{U}} = \mathbf{UG}(\theta)$, $(n,p) = (8,4)$, $d_{max} := \frac{\pi\sqrt{2p}}{2}$.

Figure 3.2 allows to verify by experience that the lower bound is always respected. For $\widetilde{\mathbf{U}} = \mathbf{UG}(\theta)$, Figure 3.3 shows that Algorithm 3 computes the correct distances since the lower bound is retrieved numerically.

3.4.3 The case p odd

After handling the scenario where p is even, we examine the case where p is odd. This case is more difficult and we build our intuition from the numerical experiments of Figure 3.4.

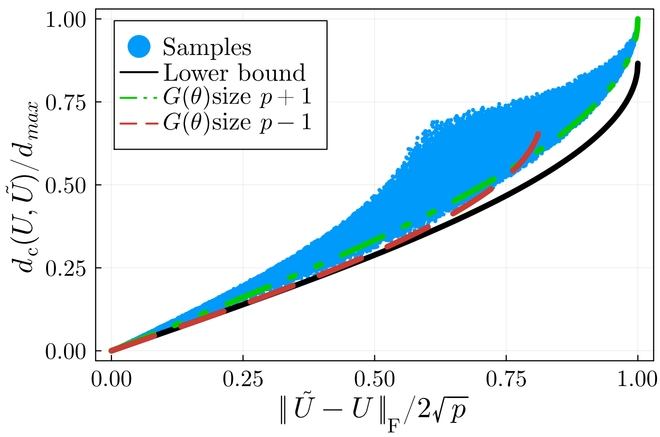

Figure 3.4: Distances computed on 100000 random samples. $(n,p) = (6,3)$, $d_{max} := \frac{\pi\sqrt{2(p+1)}}{2}$.

From the experiences, we observe that the lower bound, which was reached for p even, is not for p odd. From the observation of Figure 3.4, we conjecture the lower bound

$$\min\left[\sqrt{2(p-1)}\arcsin\left(\frac{\|\mathbf{U}-\widetilde{\mathbf{U}}\|_F}{2\sqrt{p-1}}\right), \sqrt{2(p+1)}\arcsin\left(\frac{\|\mathbf{U}-\widetilde{\mathbf{U}}\|_F}{2\sqrt{p}}\right)\right]. \tag{3.3}$$

This conjecture is based on an educated guess: the first term is the bound obtained after a rotation using a matrix $\mathbf{G}(\theta)$ of size $p-1$, thus the exact lower bound for $p-1$ (even). The second term is obtained using a matrix $\mathbf{G}(\theta)$ of size $p+1$ (even). We observe on Figure 3.4 that the transition between the two terms is not perfect so that the true bound should create a transition between the 2 curves. However, for $p \gg 1$, notice that the conjecture converges to the lower bound of Corollary 3.3. The gap becomes relatively thinner as p grows. Said otherwise, experiences indicate that there is few incentive to prove the tight lower bound for p odd given that we know it for p even.

3.5 Geodesics preserving the column space and their cut locus

There is a case where it is very simple to solve the geodesic endpoint problem on Stiefel. It is when the matrices are the same up to a right *rotation*. In this case, the problem is similar to the problem on $\mathrm{SO}(p)$ and we can solve it. The result is suggested in [47] but the arguments are not developed (in particular the argument of minimality of the geodesic). Moreover, [47] allows reflections (negative determinant), which we think should not be allowed since the matrix logarithm can not be guaranteed real. The error actually spans across the entire paper [47]. However, the author acknowledged and rectified the mistake in a subsequent publication [48].

Lemma 3.3. *For all* $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$ *such that* $\mathrm{Col}(\mathbf{U}) = \mathrm{Col}(\widetilde{\mathbf{U}})$, *the matrix* $\mathbf{Q} \in \mathrm{O}(p)$ *such that* $\widetilde{\mathbf{U}} = \mathbf{UQ}$ *exist and is unique.*

Proof. Existence: the existence of $\mathbf{Q} \in \mathbb{R}^{p\times p}$ is straightforward since $\mathrm{Col}(\mathbf{U}) = \mathrm{Col}(\widetilde{\mathbf{U}})$. It remains to prove that $\mathbf{Q} \in \mathrm{O}(p)$. It is given by

$$\widetilde{\mathbf{U}}^T\widetilde{\mathbf{U}} = \mathbf{I} \Longrightarrow \mathbf{Q}^T\mathbf{U}^T\mathbf{U}\mathbf{Q} = \mathbf{Q}^T\mathbf{Q} = \mathbf{I} \Longrightarrow \mathbf{Q} \in \mathrm{O}(p). \tag{3.4}$$

Uniqueness: $\widetilde{\mathbf{U}} = \mathbf{UQ}$ yields $\mathbf{Q} = \mathbf{U}^T\widetilde{\mathbf{U}}$. $\mathbf{Q}$ is thus uniquely defined. □

Lemma 3.4. *For all* $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$, *such that* $\mathrm{Col}(\mathbf{U}) = \mathrm{Col}(\widetilde{\mathbf{U}})$ *and* $\widetilde{\mathbf{U}}^T\mathbf{U} \in \mathrm{SO}(p)$, *at least a minimal geodesic has the form* $\gamma_\beta(t) = \mathbf{U}e^{\mathbf{A}t}$ *with* $\gamma(1) = \widetilde{\mathbf{U}}$.

Proof. Suppose $\widetilde{\mathbf{U}} \notin C_{\mathbf{U}}$. By Theorem 1.2, $\mathrm{Col}(\mathbf{U}) = \mathrm{Col}(\widetilde{\mathbf{U}})$ implies that $(\mathbf{I} - \mathbf{U}\mathbf{U}^T)\widetilde{\mathbf{U}} = 0$. Hence $\mathbf{\Delta} \in \mathrm{Log}_{\mathbf{U}}(\widetilde{\mathbf{U}})$ features a representation $\mathbf{\Delta} = \mathbf{UA}$ for some $\mathbf{A} \in \mathrm{Skew}(p)$. This implies that the minimizing geodesic can be written $\gamma_\beta(t) = \mathbf{U}e^{\mathbf{A}t}$.

Suppose now $\widetilde{\mathbf{U}} \in C_{\mathbf{U}}$. Since $\widetilde{\mathbf{U}}$ can be approached by a geodesic $\gamma_\beta(t) = \mathbf{U}e^{\mathbf{A}t}$ by Lemma 3.3. Lemma 3.4 is true for any sequence of points on γ_β converging to $\widetilde{\mathbf{U}}$, hence the lemma is also true at $\widetilde{\mathbf{U}}$ by continuity. □

Lemmas 3.3 and 3.4 allow to characterize the cut locus for geodesics preserving the column space.

Property 3.2. *Given* $\mathbf{A} \in \mathrm{Skew}(p)$, $\mathbf{U} \in \mathrm{St}(n,p)$ *and a initial velocity* $\mathbf{\Delta} = \mathbf{UA} \in T_{\mathbf{U}}\mathrm{St}(n,p)$,

1. *the cut time is* $\mathcal{T}_{\mathbf{U}}(\mathbf{\Delta}) = \frac{\pi}{\|\mathbf{A}\|_2}$.
2. *the tangent cut locus is* $\hat{C}_{\mathbf{U}}(\mathbf{\Delta}) = \mathbf{\Delta}\mathcal{T}_{\mathbf{U}}(\mathbf{\Delta}) = \mathbf{UA}\mathcal{T}_{\mathbf{U}}(\mathbf{\Delta})$.
3. *the cut locus is* $C_{\mathbf{U}}(\mathbf{\Delta}) = \mathrm{Exp}_u(\mathbf{\Delta}) = \mathbf{U}e^{\mathbf{A}\mathcal{T}_{\mathbf{U}}(\mathbf{\Delta})}$.

Proof. For all $\mathbf{A} \in \mathrm{Skew}(p)$ and $\mathbf{U} \in \mathrm{St}(n,p)$, we know that $\widetilde{\mathbf{U}} := \mathbf{U}e^{\mathbf{A}}$ verifies Lemma 3.4. Moreover, we know by Lemma 3.3 that if $\widetilde{\mathbf{U}} = \mathbf{U}e^{\mathbf{B}}$ then $e^{\mathbf{B}} = e^{\mathbf{A}}$. Hence, the geodesic endpoint problem between matrices spanning the same column space reduces to the exact same problem we encountered for $\mathrm{SO}(p)$. Since $e^{\mathbf{A}}$ is unique, the matrix $\mathbf{A} \in \mathrm{Skew}(p)$ appearing in $\mathbf{UA} \in \mathrm{Log}_{\mathbf{U}}(\widetilde{\mathbf{U}})$ is a matrix logarithm of $e^{\mathbf{A}} = \mathbf{U}^T\widetilde{\mathbf{U}}$ such that $\|\mathbf{A}\|_2 \leq \pi$ following the same argument as Theorem 2.1. The cut time, the tangent cut locus and the cut locus are direct consequences. □

3.6 An upper bound on the β-distance by the Frobenius distance

We introduce geometric sets that will provide several theorems further in this work. Given two points $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$, we define a latitude hyperplane $\Pi_{\mathbf{U},\widetilde{\mathbf{U}}}$ as the intersection of the hypersphere

$$\mathbb{S}^{n\times p}_{\sqrt{p}} := \left\{\mathbf{X} \in \mathbb{R}^{n\times p} \mid \mathrm{Tr}(\mathbf{X}^T\mathbf{X}) = p\right\}, \tag{3.5}$$

with a second hypersphere $\Sigma_{\mathbf{U},\widetilde{\mathbf{U}}}$ for which $\mathbf{U}$ and $\widetilde{\mathbf{U}}$ are antipodal points. Let $\mathbf{C} = \frac{\mathbf{U}+\widetilde{\mathbf{U}}}{2} \neq 0$ and $r = \frac{\|\widetilde{\mathbf{U}}-\mathbf{U}\|_{\mathrm{F}}}{2}$. The hypersphere $\Sigma_{\mathbf{U},\widetilde{\mathbf{U}}}$ is defined as

$$\Sigma_{\mathbf{U},\widetilde{\mathbf{U}}} := \left\{\mathbf{X} \in \mathbb{R}^{n\times p} \mid \mathrm{Tr}((\mathbf{X}-\mathbf{C})^T(\mathbf{X}-\mathbf{C})) = r^2\right\}. \tag{3.6}$$

If $\widetilde{\mathbf{U}} \neq -\mathbf{U}$, combining the Equations (3.5) and (3.6) leads to the intersection hyperplane $\Pi_{\mathbf{U},\widetilde{\mathbf{U}}}$ defined as the set

$$\Pi_{\mathbf{U},\widetilde{\mathbf{U}}} := \left\{\mathbf{X} \in \mathbb{R}^{n\times p} \mid \mathrm{Tr}(\mathbf{X}^T(\mathbf{U}+\widetilde{\mathbf{U}})) = p + \mathrm{Tr}(\mathbf{U}^T\widetilde{\mathbf{U}})\right\}. \tag{3.7}$$

The hyperplane $\Pi_{\mathbf{U},\widetilde{\mathbf{U}}}$ in $\mathbb{R}^{n\times p}$ cuts a spherical cap of $\mathbb{S}^{n\times p}_{\sqrt{p}}$ (the cap is defined as the part that does not include the origin) and divides $\Sigma_{\mathbf{U},\widetilde{\mathbf{U}}}$ into two symmetrical parts. The sets $\mathbb{S}^{n\times p}_{\sqrt{p}}, \Sigma_{\mathbf{U},\widetilde{\mathbf{U}}}$ and $\Pi_{\mathbf{U},\widetilde{\mathbf{U}}}$ are very important to obtain geometric insights on $\mathrm{St}(n,p)$. These sets are at the core of Theorems 3.4, 3.6, 3.8, 3.11 and 3.12.

3.6.1 An upper bound on the Euclidean distance

For all $\mathbf{U} \in \mathrm{St}(n,p)$, we define the augmentation of $\mathbf{U}$, $\mathbf{U}_{\mathrm{aug}} \in \mathrm{St}(n+1,p)$, as the matrix obtained by adding a last row of zeros to $\mathbf{U}$.

Lemma 3.5. *For all $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$, $\beta > 0$, we have $d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) = d_\beta(\mathbf{U}_{\mathrm{aug}}, \widetilde{\mathbf{U}}_{\mathrm{aug}})$.*

Proof. Suppose $\widetilde{\mathbf{U}} \notin C_{\mathbf{U}}$ and $\widetilde{\mathbf{U}}_{\mathrm{aug}} \notin C_{\mathbf{U}_{\mathrm{aug}}}$. The proof of this lemma is guided by Theorem 1.2. We show that if $\mathbf{\Delta} \in \mathrm{Log}_{\mathbf{U},\beta}(\widetilde{\mathbf{U}})$, then $\mathbf{\Delta}_{\mathrm{aug}} := \begin{bmatrix}\mathbf{\Delta}\\ 0\end{bmatrix} \in \mathrm{Log}_{\mathbf{U}_{\mathrm{aug}},\beta}(\widetilde{\mathbf{U}}_{\mathrm{aug}})$. Let

$$\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p), \quad \mathbf{M} = \mathbf{U}^T\widetilde{\mathbf{U}}, \quad \mathbf{QN} = (\mathbf{I} - \mathbf{UU}^T)\widetilde{\mathbf{U}}, \tag{3.8}$$

where $\mathbf{Q} \in \mathrm{St}(n,p)$. Then Theorem 1.2 ensures that $\mathbf{\Delta} \in \mathrm{Log}_{\mathbf{U},\beta}(\widetilde{\mathbf{U}})$ features a representation $\mathbf{\Delta} = \mathbf{UA} + \mathbf{QB}$ where $\mathbf{A} \in \mathrm{Skew}(p)$, $\mathbf{B} \in \mathbb{R}^{p\times p}$ and $\mathbf{A}, \mathbf{B}$ solve Problem 1.1. As mentioned in [48], the key here is that $\mathbf{Q}$ can be an arbitrary orthonormal base of the column space of $(\mathbf{I} - \mathbf{UU}^T)\widetilde{\mathbf{U}}$ without affecting $\mathbf{\Delta}$. Thus if we want to solve the logarithm problem for $\mathbf{U}_{\mathrm{aug}}, \widetilde{\mathbf{U}}_{\mathrm{aug}} \in \mathrm{St}(n+1,p)$, we can arbitrarily take $\mathbf{Q}_{\mathrm{aug}} = \begin{bmatrix}\mathbf{Q}\\ 0\end{bmatrix} \in \mathrm{St}(n+1,p)$ such that $\mathbf{Q}_{\mathrm{aug}}\mathbf{N} = (\mathbf{I} - \mathbf{U}_{\mathrm{aug}}\mathbf{U}_{\mathrm{aug}}^T)\widetilde{\mathbf{U}}_{\mathrm{aug}}$. Moreover, $\mathbf{M} = \mathbf{U}_{\mathrm{aug}}^T\widetilde{\mathbf{U}}_{\mathrm{aug}} = \mathbf{U}^T\widetilde{\mathbf{U}}$. Notice that $\mathbf{M}$ and $\mathbf{N}$ are left unchanged compared to their definition of (3.8). In consequence, the Problem 1.1 that needs to be solved to obtain $\mathbf{\Delta}_{\text{aug}} \in \text{Log}_{\mathbf{U}_{\text{aug}},\beta}(\widetilde{\mathbf{U}}_{\text{aug}})$ is same as the one to get $\mathbf{\Delta} \in \text{Log}_{\mathbf{U},\beta}(\widetilde{\mathbf{U}})$. This leads to $\mathbf{\Delta}_{\text{aug}} = \mathbf{U}_{\text{aug}}\mathbf{A} + \mathbf{Q}_{\text{aug}}\mathbf{B}$ where $\mathbf{A}, \mathbf{B}$ are the same as for $\mathbf{\Delta}$. Finally, we have

$$d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) = \|\mathbf{\Delta}\|_\beta = \sqrt{\beta\|\mathbf{A}\|_{\text{F}}^2 + \|\mathbf{B}\|_{\text{F}}^2} = \|\mathbf{\Delta}_{\text{aug}}\|_\beta = d_\beta(\mathbf{U}_{\text{aug}}, \widetilde{\mathbf{U}}_{\text{aug}}). \tag{3.9}$$

Suppose $\widetilde{\mathbf{U}} \in C_{\mathbf{U}}$ or $\widetilde{\mathbf{U}}_{\text{aug}} \in C_{\mathbf{U}_{\text{aug}}}$. From [35, Chapter III, Lemma 4.4], we know that $C_{\mathbf{U}}$ and $C_{\mathbf{U}_{\text{aug}}}$ are null sets of $\text{St}(n, p)$ and $\text{St}(n+1, p)$ respectively. Hence, we can always build two sequences $\left\{\widehat{\mathbf{U}}_i\right\}_{i\in\mathbb{N}}$ and $\left\{\widehat{\mathbf{U}}_{i,\text{aug}}\right\}_{i\in\mathbb{N}}$ such that $\lim_{i\to\infty} \widehat{\mathbf{U}}_i = \widetilde{\mathbf{U}}$, $\lim_{i\to\infty} \widehat{\mathbf{U}}_{i,\text{aug}} = \widetilde{\mathbf{U}}_{\text{aug}}$, $\widehat{\mathbf{U}}_i \notin C_{\mathbf{U}}$ and $\widehat{\mathbf{U}}_{i,\text{aug}} \notin C_{\mathbf{U}_{\text{aug}}}$ for all $i \in \mathbb{N}$. For all such sequences, we have

$$\begin{aligned} d_\beta(\mathbf{U}, \widehat{\mathbf{U}}_i) &= d_\beta(\mathbf{U}_{\text{aug}}, \widehat{\mathbf{U}}_{i,\text{aug}}) \quad \text{for all } i \in \mathbb{N} \\ \implies \lim_{i\to\infty} d_\beta(\mathbf{U}, \widehat{\mathbf{U}}_i) &= \lim_{i\to\infty} d_\beta(\mathbf{U}_{\text{aug}}, \widehat{\mathbf{U}}_{i,\text{aug}}) \\ \implies d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) &= d_\beta(\mathbf{U}_{\text{aug}}, \widetilde{\mathbf{U}}_{\text{aug}}), \end{aligned}$$

where the last step stands by continuity of the β-distance function. □

Corollary 3.4. *Let N be a positive integer. Every pair $\mathbf{U}, \widetilde{\mathbf{U}} \in \text{St}(n, p)$ can be augmented by N rows of zeros without changing the β-distances.*

Proof. Straightforward induction on Lemma 3.5 □

Now we prove a lemma on the existence of a solution to a particular nonlinear system of equations. These equations come from essential conditions to prove Theorem 3.6.

Lemma 3.6. *For all $\mathbf{U}, \widetilde{\mathbf{U}} \in \text{St}(n, p)$ and given $\mathbf{R} = \frac{\widetilde{\mathbf{U}}-\mathbf{U}}{2}$, the system $\begin{cases} \mathbf{U}^T\mathbf{X} + \mathbf{X}^T\mathbf{U} = \mathbf{0} \\ \widetilde{\mathbf{U}}^T\mathbf{X} + \mathbf{X}^T\widetilde{\mathbf{U}} = \mathbf{0} \\ \mathbf{X}^T\mathbf{X} = \mathbf{R}^T\mathbf{R} \end{cases}$ has at least a solution $\mathbf{X} \in \mathbb{R}^{n\times p}$ if $n \geq 3p$.*

Proof. Since $n \geq 3p$, we can always take $\hat{\mathbf{U}} \in \text{St}(n, p)$ such that $\hat{\mathbf{U}}^T\mathbf{U} = \hat{\mathbf{U}}^T\widetilde{\mathbf{U}} = \mathbf{0}$. For all $\mathbf{Y} \in \mathbb{R}^{p\times p}$, $\mathbf{X} := \hat{\mathbf{U}}\mathbf{Y}$ satisfies the two first equations. The third equation yields $\mathbf{Y}^T\mathbf{Y} = \mathbf{R}^T\mathbf{R}$ that has always a solution since $\mathbf{Y}$ is free. For instance, $\mathbf{Y} = (\mathbf{R}^T\mathbf{R})^{\frac{1}{2}}$ is always a real solution since $\mathbf{R}^T\mathbf{R}$ is symmetric semi-positive definite. □

Theorem 3.6. *For all $\mathbf{U}, \widetilde{\mathbf{U}} \in \text{St}(n, p)$, we have $d_{\text{E}}(\mathbf{U}, \widetilde{\mathbf{U}}) \leq \frac{\pi}{2}\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\text{F}}$*

Proof. First of all, we augment $\mathbf{U}$ and $\widetilde{\mathbf{U}}$ with rows of zeros such that $n \geq 3p$. Due to Corollary 3.4, the β-distances are left unchanged and Lemma 3.6 holds. Let us define a curve γ for $t \in [0, \pi]$ in the space of matrices $\mathbb{R}^{n\times p}$ defined as

$$\gamma(t) := \mathbf{C} + \cos(t)\mathbf{R} + \sin(t)\mathbf{S},$$

where $\mathbf{C} = \frac{\mathbf{U}+\widetilde{\mathbf{U}}}{2}$ and $\mathbf{R} = \frac{\mathbf{U}-\widetilde{\mathbf{U}}}{2}$. We impose conditions on $\mathbf{S} \in \mathbb{R}^{n\times p}$ to get a curve γ that belongs to $\text{St}(n, p)$ and draws a half-circle:

$$\begin{cases} \mathbf{C}^T\mathbf{S} + \mathbf{S}^T\mathbf{C} = \mathbf{0} \\ \mathbf{R}^T\mathbf{S} + \mathbf{S}^T\mathbf{R} = \mathbf{0} \\ \mathbf{S}^T\mathbf{S} = \mathbf{R}^T\mathbf{R} \end{cases} \iff \begin{cases} \mathbf{U}^T\mathbf{S} + \mathbf{S}^T\mathbf{U} = \mathbf{0} \\ \widetilde{\mathbf{U}}^T\mathbf{S} + \mathbf{S}^T\widetilde{\mathbf{U}} = \mathbf{0} \\ \mathbf{S}^T\mathbf{S} = \mathbf{R}^T\mathbf{R} \end{cases}. \tag{3.10}$$

The System (3.10) has always a solution since Lemma 3.6 holds. If these conditions are met, then $\gamma(t)$ belongs to the hypersphere $\Sigma_{\mathbf{U},\widetilde{\mathbf{U}}}$ for all $t \in \mathbb{R}$. We recall the definition of $\Sigma_{\mathbf{U},\widetilde{\mathbf{U}}}$:

$$\Sigma_{\mathbf{U},\widetilde{\mathbf{U}}} := \left\{\mathbf{X} \in \mathbb{R}^{n\times p} \mid \operatorname{Tr}\left((\mathbf{X}-\mathbf{C})^T(\mathbf{X}-\mathbf{C})\right) = r^2\right\},$$

where $r = \frac{\|\widetilde{\mathbf{U}}-\mathbf{U}\|_{\mathrm{F}}}{2}$. For all $t \in \mathbb{R}$, we have

$$\begin{aligned}
\operatorname{Tr}\left((\gamma(t)-\mathbf{C})^T(\gamma(t)-\mathbf{C})\right) &= \operatorname{Tr}\left((\cos(t)\mathbf{R}+\sin(t)\mathbf{S})^T(\cos(t)\mathbf{R}+\sin(t)\mathbf{S})\right)\\
&= \operatorname{Tr}(\mathbf{R}^T\mathbf{R}) \qquad \text{Since } \mathbf{R}^T\mathbf{S}+\mathbf{S}^T\mathbf{R} = \mathbf{0} \text{ and } \mathbf{S}^T\mathbf{S} = \mathbf{R}^T\mathbf{R}.\\
&= r^2.
\end{aligned}$$

Since γ is an affine combination of two vectors, it lies in a two-dimensional plane. Because γ also lies on the border of the hypersphere $\Sigma_{\mathbf{U},\widetilde{\mathbf{U}}}$, γ must draw a great-circle of Euclidean length $\pi r = \frac{\pi}{2}\|\widetilde{\mathbf{U}}-\mathbf{U}\|_{\mathrm{F}}$. More rigorously, we have

$$\begin{aligned}
l_{\mathrm{E}}(\gamma) &= \int_0^\pi \sqrt{\operatorname{Tr}(\dot{\gamma}(t)^T\dot{\gamma}(t))}\,\mathrm{d}t\\
&= \int_0^\pi \sqrt{\operatorname{Tr}\left((\cos(t)\mathbf{S}-\sin(t)\mathbf{R})^T(\cos(t)\mathbf{S}-\sin(t)\mathbf{R})\right)}\,\mathrm{d}t\\
&= \int_0^\pi \sqrt{\operatorname{Tr}(\mathbf{R}^T\mathbf{R})}\,\mathrm{d}t \qquad \text{Since } \mathbf{S}^T\mathbf{S} = \mathbf{R}^T\mathbf{R} \text{ and } \mathbf{R}^T\mathbf{S}+\mathbf{S}^T\mathbf{R} = \mathbf{0}.\\
&= \frac{\pi}{2}\|\widetilde{\mathbf{U}}-\mathbf{U}\|_{\mathrm{F}}.
\end{aligned}$$

We claim that $\gamma(t)$ also lies in $\mathrm{St}(n,p)$ for all $t \in \mathbb{R}$. Using the definition of γ, we get

$$\begin{aligned}
\gamma(t)^T\gamma(t) &= \mathbf{C}^T\mathbf{C} + \cos(t)(\mathbf{C}^T\mathbf{R}+\mathbf{R}^T\mathbf{C}) + \sin(t)(\mathbf{C}^T\mathbf{S}+\mathbf{S}^T\mathbf{C}) + \cos^2(t)(\mathbf{R}^T\mathbf{R}) + \sin^2(t)(\mathbf{S}^T\mathbf{S})\\
&= \mathbf{C}^T\mathbf{C} + \mathbf{R}^T\mathbf{R} \qquad \text{Since } \mathbf{C}^T\mathbf{R}+\mathbf{R}^T\mathbf{C} = \mathbf{0} \text{ and } \mathbf{C}^T\mathbf{S}+\mathbf{S}^T\mathbf{C} = \mathbf{0}.\\
&= \mathbf{I}.
\end{aligned}$$

By definition (1.17) of the distance function, we have

$$d_{\mathrm{E}}(\mathbf{U},\widetilde{\mathbf{U}}) \le l_{\mathrm{E}}(\widetilde{\gamma}) \text{ for all curves } \widetilde{\gamma} : [0,1] \to \mathrm{St}(n,p) \text{ with } \widetilde{\gamma}(0) = \mathbf{U} \text{ and } \widetilde{\gamma}(1) = \widetilde{\mathbf{U}}.$$

Hence, we have in particular $d_{\mathrm{E}}(\mathbf{U},\widetilde{\mathbf{U}}) \le \frac{\pi}{2}\|\widetilde{\mathbf{U}}-\mathbf{U}\|_{\mathrm{F}}$. □

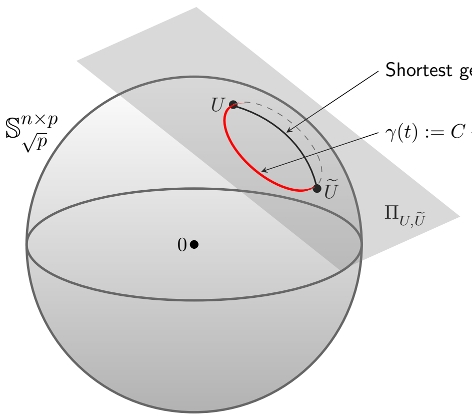

Figure 3.5: Illustration of Theorem 3.6

3.6.2 The Euclidean diameter of Stiefel

Theorem 3.7. *The diameter of* $(\mathrm{St}(n,p), g_{\mathrm{E}})$ *is* $\pi\sqrt{p}$. *Moreover, for all* $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$, $d_{\mathrm{E}}(\mathbf{U}, \widetilde{\mathbf{U}}) = \pi\sqrt{p}$ *if and only if* $\widetilde{\mathbf{U}} = -\mathbf{U}$.

Proof. The combination of Theorem 3.4 and Theorem 3.6 provides the Euclidean diameter of Stiefel. Indeed, knowing that the Frobenius distance is bounded from above by $2\sqrt{p}$ on $\mathrm{St}(n,p)$, Theorem 3.6 leads to the upper bound $d_{\mathrm{E}}(\mathbf{U}, -\mathbf{U}) \leq \pi\sqrt{p}$. Moreover, Theorem 3.4 gives the lower bound

$$d_{\mathrm{E}}(\mathbf{U}, -\mathbf{U}) \geq 2\sqrt{p}\arcsin\left(\frac{\|2\mathbf{U}\|_{\mathrm{F}}}{2\sqrt{p}}\right) = \pi\sqrt{p}.$$

Therefore, we know that (i) the Euclidean geodesic distance between $\mathbf{U}$ and $-\mathbf{U}$ is $\pi\sqrt{p}$ for all $p > 0$ and (ii) for all $\mathbf{U} \in \mathrm{St}(n,p)$, the pair $(\mathbf{U}, -\mathbf{U})$ reaches the upper bound on the Euclidean geodesic distance. The diameter of $(\mathrm{St}(n,p), g_{\mathrm{E}})$ is thus $\pi\sqrt{p}$ for all p and, for all $\mathbf{U} \in \mathrm{St}(n,p)$, $-\mathbf{U}$ is the only matrix such that the distance reaches the diameter. $\square$

More intuitively, we see on Figure 3.6 that the lower and upper bounds cross at $(\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}},\ d_{\mathrm{E}}(\mathbf{U}, \widetilde{\mathbf{U}})) = (2\sqrt{p}, \pi\sqrt{p})$, which corresponds to $\widetilde{\mathbf{U}} = -\mathbf{U}$.

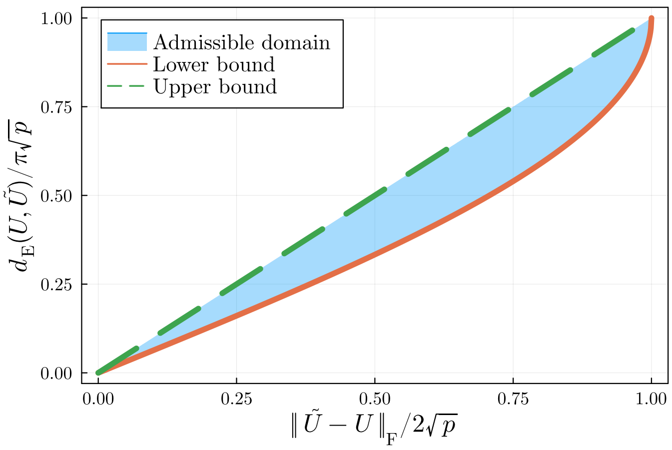

Figure 3.6: Admissible domain for any pair (n,p).

3.7 Corroboration of the Euclidean bounds

Let us corroborate the bounds obtained for the Euclidean geodesic distance. We evaluate the performances of the two algorithms proposed in [10]: the shooting method and the path straightening method. We recall that the shooting method is Algorithm 2 ([48, Algorithm 1]) and the path straightening method is [10, Algorithms 2 to 8].

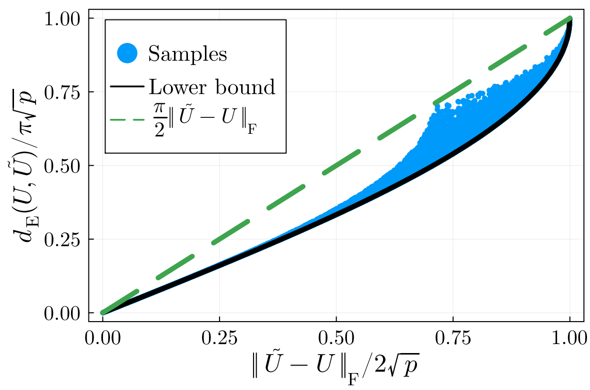

Figure 3.7: Distances computed with the shooting method. 10000 samples. $(n, p) = (4, 2)$

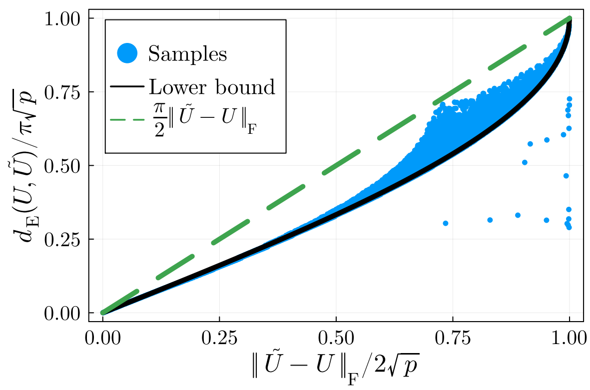

Figure 3.8: Distances computed with the path straightening method. 10000 samples. $(n, p) = (4, 2)$

Figures 3.7 and 3.8 show the results obtained using the shooting method and the path-straightening algorithm with a discretization of $N = 5\left\lceil\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}\right\rceil$ and $N = 30\left\lceil\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}\right\rceil$ points respectively for the discrete sampling of the geodesic. On Figure 3.7, we observe that all the experiments verified the bound and the diameter of $(\mathrm{St}(n,p), g_{\mathrm{E}})$. For the path straightening method of Figure 3.8, we observe that a few computed points do not respect the bounds obtained. Therefore we can assess that the method failed for these few points (18 out of 10000). Knowing that the shooting method is cheaper to apply, we experienced that the shooting method should be preferred to the path-straightening method.

3.8 Equivalence of the β-distance and the Frobenius distance

Theorem 3.6, combined with Theorem 3.4 and Corollary 3.1, allows to conclude to the bilipschitz equivalence of the geodesic distance and the Frobenius distance.

Corollary 3.5. *For all* $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$*, we have* $\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}} \le d_{\mathrm{E}}(\widetilde{\mathbf{U}}, \mathbf{U}) \le \frac{\pi}{2}\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}$

Proof. By Theorem 3.4, we have

$$d_{\mathrm{E}}(\widetilde{\mathbf{U}}, \mathbf{U}) \ge 2\sqrt{p}\arcsin\left(\frac{\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}}{2\sqrt{p}}\right) \ge \|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}. \tag{3.11}$$

The last inequality holds since $\theta \to 2\sqrt{p}\arcsin\left(\frac{\theta}{2\sqrt{p}}\right)$ is a convex function for $\theta \in [0, 2\sqrt{p}]$ and $\theta \to \theta$ is the tangent at $\theta = 0$. Combined with Theorem 3.6 to obtain the upper bound $\frac{\pi}{2}\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}$, the proof is completed. □

The result of Corollary 3.5 can be generalized to all the β-distances using the equivalence of the β-distances (Corollary 3.1).

Corollary 3.6. *We have* $\min\{1, \sqrt{\beta}\}\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}} \le d_{\beta}(\widetilde{\mathbf{U}}, \mathbf{U}) \le \max\{1, \sqrt{\beta}\}\frac{\pi}{2}\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}$ *for all* $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$,

Proof. Combine Corollary 3.5 and Corollary 3.1. □

Notice that the value of β such that the two lipschitz constants of Corollary 3.6 are the closest from each other is $\beta = 1$, i.e under the Euclidean metric. Hence, the Euclidean geodesic distance is the most similar to the Frobenius distance.

3.9 Reaching the linear upper bound

Are there pairs $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$ such that $d_{\mathrm{E}}(\mathbf{U}, \widetilde{\mathbf{U}}) = \frac{\pi}{2}\|\widetilde{\mathbf{U}} - \mathbf{U}\|_F$? To answer that question, we need to recall the definition of the hyperplane $\Pi_{\mathbf{U},\widetilde{\mathbf{U}}}$ cutting a spherical cap from $\mathbb{S}^{n\times p}_{\sqrt{p}}$. Defining $f_{\mathbf{U},\widetilde{\mathbf{U}}}(\mathbf{X}) = \mathrm{Tr}\left(\mathbf{X}^T(\mathbf{U} + \widetilde{\mathbf{U}})\right) - p - \mathrm{Tr}(\mathbf{U}^T\widetilde{\mathbf{U}})$, we have

$$\Pi_{\mathbf{U},\widetilde{\mathbf{U}}} := \left\{\mathbf{X} \in \mathbb{R}^{n\times p} \mid f_{\mathbf{U},\widetilde{\mathbf{U}}}(\mathbf{X}) = 0\right\}. \tag{3.12}$$

$\Pi_{\mathbf{U},\widetilde{\mathbf{U}}}$ divides $\mathbb{S}^{n\times p}_{\sqrt{p}}$ in two parts. The part such that $f_{\mathbf{U},\widetilde{\mathbf{U}}}(\mathbf{X}) > 0$ is called the spherical cap. Notice that $\Pi_{\mathbf{U},\widetilde{\mathbf{U}}}$ is only well-defined for $\widetilde{\mathbf{U}} \neq -\mathbf{U}$.

Lemma 3.7. *For all $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$ with $\widetilde{\mathbf{U}} \neq -\mathbf{U}$, let $\gamma : [0,1] \to \mathrm{St}(n,p)$ be a curve such that $\gamma(0) = \mathbf{U}$ and $\gamma(1) = \widetilde{\mathbf{U}}$. If $f_{\mathbf{U},\widetilde{\mathbf{U}}}(\gamma(t)) \leq 0$ for all $t \in [0,1]$, then $l_{\mathrm{E}}(\gamma) \geq \frac{\pi}{2}\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}$.*

Proof. We recall the notation $\mathbf{C} = \frac{\mathbf{U}+\widetilde{\mathbf{U}}}{2}$ and $r = \frac{\|\mathbf{U}-\widetilde{\mathbf{U}}\|_{\mathrm{F}}}{2}$. If $f_{\mathbf{U},\widetilde{\mathbf{U}}}(\gamma(t)) \leq 0$ for all $t \in [0,1]$, since we have $\mathrm{Tr}(\gamma(t)^T\gamma(t)) = p$, we also have

$$\mathrm{Tr}\left((\gamma(t) - \mathbf{C})^T(\gamma(t) - \mathbf{C})\right) \geq r^2. \tag{3.13}$$

Indeed,

$$\begin{aligned}
& 0 \geq f_{\mathbf{U},\widetilde{\mathbf{U}}}(\gamma(t)) \\
\iff\ & 0 \geq \mathrm{Tr}\left(\gamma(t)^T(\mathbf{U} + \widetilde{\mathbf{U}})\right) - p - \mathrm{Tr}(\mathbf{U}^T\widetilde{\mathbf{U}}) \\
\iff\ & 0 \geq \mathrm{Tr}\left(\gamma(t)^T(\mathbf{U} + \widetilde{\mathbf{U}})\right) - \mathrm{Tr}(\gamma(t)^T\gamma(t)) - \left[\mathrm{Tr}(\mathbf{C}^T\mathbf{C}) - \frac{1}{4}\mathrm{Tr}((\mathbf{U} - \widetilde{\mathbf{U}})^T(\mathbf{U} - \widetilde{\mathbf{U}}))\right] \\
\iff\ & r^2 \leq \mathrm{Tr}\left((\gamma(t) - \mathbf{C})^T(\gamma(t) - \mathbf{C})\right).
\end{aligned}$$

Hence, γ never enters the interior of the ball included in the sphere $\Sigma_{\widetilde{\mathbf{U}},\mathbf{U}}$ in $\mathbb{R}^{n\times p}$ centered at $\mathbf{C}$ of radius r. Since γ starts and ends on the boundary of this ball, it can not be shorter (in the Euclidean sense) than the shortest path on $\Sigma_{\widetilde{\mathbf{U}},\mathbf{U}}$ which has a length $\pi r = \frac{\pi}{2}\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}$. Finally, we have $l_{\mathrm{E}}(\gamma) \geq \frac{\pi}{2}\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}$. □

Lemma 3.7 was motivated by the fact that we also know $d_{\mathrm{E}}(\mathbf{U}, \widetilde{\mathbf{U}}) \leq \frac{\pi}{2}\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}$ from Theorem 3.6. Therefore if we find pairs of points $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$ such that

$$\left\{\mathbf{X} \in \mathbb{R}^{n\times p} \mid f_{\mathbf{U},\widetilde{\mathbf{U}}}(\mathbf{X}) > 0\right\} \cap \mathrm{St}(n,p) = \emptyset,$$

we ensure $f_{\mathbf{U},\widetilde{\mathbf{U}}}(\gamma(t)) \leq 0$ for all $t \in [0,1]$, for all $\gamma : [0,1] \to \mathrm{St}(n,p)$ going from $\mathbf{U}$ to $\widetilde{\mathbf{U}}$. In consequence, we will have $d_{\mathrm{E}}(\mathbf{U}, \widetilde{\mathbf{U}}) = \frac{\pi}{2}\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}$ leveraging Lemma 3.7 and Theorem 3.6.

Now, we build such pairs $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$. Take $k \in \mathbb{N} \cap [1,p]$. For all $\mathbf{U} \in \mathrm{St}(n,p)$, we define $h_k : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ such that

$$h_k(\mathbf{U})_{i,j} := \begin{cases} -\mathbf{U}_{i,j} \text{ if } j \leq k \\ \mathbf{U}_{i,j} \text{ otherwise} \end{cases}, \tag{3.14}$$

i.e., $h_k(\mathbf{U})$ flips the first k columns of $\mathbf{U}$.

Theorem 3.8. *For all $\mathbf{U} \in \mathrm{St}(n,p)$ and $k \in \mathbb{N} \cap [1,p]$, we have $d_{\mathrm{E}}(\mathbf{U}, h_k(\mathbf{U})) = \frac{\pi}{2}\|h_k(\mathbf{U}) - \mathbf{U}\|_{\mathrm{F}}$.*

Proof. Theorem 3.8 is already proven for $k = p$ in Theorem 3.7. Hence, we suppose $k \leq p-1$ so that $h_k(\mathbf{U}) \neq -\mathbf{U}$. By definition of $h_k(\mathbf{U})$, we have $\mathrm{Tr}(\mathbf{U}^T h_k(\mathbf{U})) = p - 2k$. Moreover,

$$\mathbf{U}_{i,j} + h_k(\mathbf{U})_{i,j} := \begin{cases} 0 \text{ if } j \leq k \\ 2\mathbf{U}_{i,j} \text{ otherwise} \end{cases}. \tag{3.15}$$

Hence, for all $\mathbf{X} \in \mathrm{St}(n,p)$, we have $\mathrm{Tr}\left(\mathbf{X}^T(\mathbf{U} + h_k(\mathbf{U}))\right) \leq 2(p-k)$. Finally, for all $\mathbf{X} \in \mathrm{St}(n,p)$, we have

$$\begin{aligned} f_{\mathbf{U},h_k(\mathbf{U})}(\mathbf{X}) &= \mathrm{Tr}\left(\mathbf{X}^T(\mathbf{U} + h_k(\mathbf{U}))\right) - p - \mathrm{Tr}(\mathbf{U}^T h_k(\mathbf{U})) \\ &\leq 2(p-k) - p - (p-2k) \\ &= 0. \end{aligned}$$

Therefore, we obtain $d_{\mathrm{E}}(\mathbf{U}, h_k(\mathbf{U})) = \frac{\pi}{2}\|h_k(\mathbf{U}) - \mathbf{U}\|_{\mathrm{F}}$ leveraging Lemma 3.7 and Theorem 3.6. □

In the previous proof, h_k was arbitrarily defined as a function that flips the columns of $\mathbf{U}$ in ascending order. It is however clear that we can take any subset of k columns among p and Theorem 3.8 still holds. For each value of k, this makes $\begin{pmatrix} p \\ k \end{pmatrix} := \frac{p!}{k!(p-k)!}$ possible choices. Hence we highlighted a total of $2^p = \sum_{k=0}^{p} \begin{pmatrix} p \\ k \end{pmatrix}$ matrices reaching the bound.

3.10 A better upper bound ?

Section 3.9 showed the tightness of the upper bound $\frac{\pi}{2}\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}$ for the Euclidean geodesic distance. It is clear however that the bound is only tight for some Frobenius distances and that a better bound exists, i.e. tight for every Frobenius distance. We show that there is few incentive to look for this better bound. For $s \in \{0, 1, ..., p-1\}$ and $2\sqrt{s} \leq \|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}} \leq 2\sqrt{s+1}$, we could conjecture from Figure 3.9 (see below) an expression of the form

$$\begin{aligned} d_{\mathrm{E}}(\mathbf{U}, \widetilde{\mathbf{U}}) &\leq \pi\sqrt{s} + 2(\sqrt{s+1} - \sqrt{s}) \arcsin\left(\frac{\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}} - 2\sqrt{s}}{2(\sqrt{s+1} - \sqrt{s})}\right) \\ &=: b_s\left(\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}\right). \end{aligned}$$

For convenience, we can define $b_*(x) := b_s(x)$ if $2\sqrt{s} \leq x \leq 2\sqrt{s+1}$ for $s = 0, 1, ..., p$. The true tight bound may differ in negligible terms from b_*. If we can get an intuition for this bound b_*, it seems there is no straightforward way to prove it. However, in the context of applications from machine learning and statistics on manifolds, it is natural to consider p large. For p large, the maximum gap between b^* and $\frac{\pi}{2}\|\widetilde{\mathbf{U}}-\mathbf{U}\|_{\mathrm{F}}$ becomes negligible in comparison to the diameter $\pi\sqrt{p}$. Indeed, we have

$$\begin{aligned}
\max_{x\in[0,2\sqrt{p}]}\left|\frac{\pi}{2}x-b_*(x)\right| &= \max_{s\in\{0,1,\dots,p\}}\max_{x\in[2\sqrt{s},2\sqrt{s+1}]}\left|\frac{\pi}{2}x-b_s(x)\right| \\
&= \max_{s\in\{0,1,\dots,p\}}\max_{x\in[0,2]}\left|\frac{\pi}{2}x-2\arcsin\left(\frac{x}{2}\right)\right|\left(\sqrt{s+1}-\sqrt{s}\right) \\
&\leq \max_{s\in\{0,1,\dots,p\}}\max_{x\in[0,2]}\left|\frac{\pi}{2}x-x\right|\left(\sqrt{s+1}-\sqrt{s}\right) \\
&= \pi-2.
\end{aligned}$$

Hence, the maximum relative error with respect to the diameter becomes negligible as p grows:

$$\lim_{p\to\infty}\frac{\max_{x\in[0,2\sqrt{p}]}\left|\frac{\pi}{2}x-b_*(x)\right|}{\pi\sqrt{p}}=0. \tag{3.16}$$

The conclusion of this section is that even if there exists a proof for the best upper bound, the improvement compared to the linear bound we have proven becomes negligible as p grows.

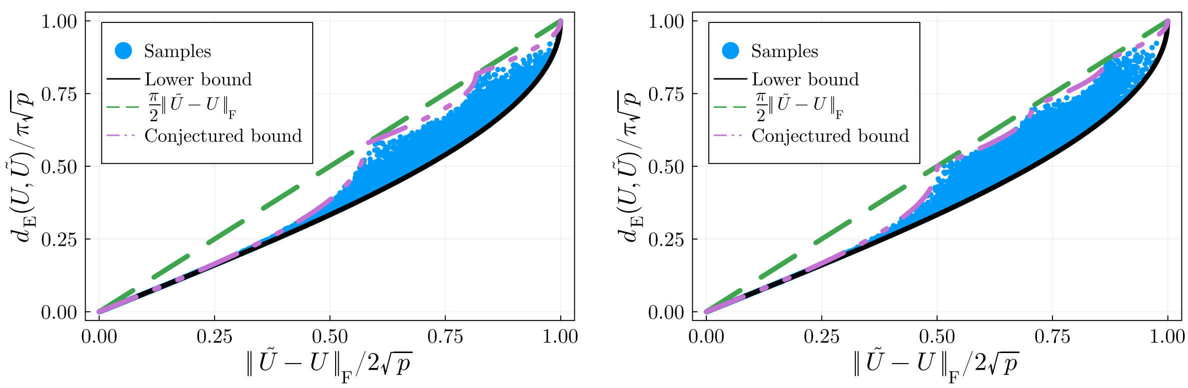

Figure 3.9: Distances computed with the shooting method. 10000 samples. $(n,p)=(5,3)$ and $(n,p)=(5,4)$.

3.11 The β-distance increases with the number of frames

We prove a lemma generalized for the β-metric with $\beta\geq\frac{1}{2}$ (thus including both the Euclidean an the canonical metrics). We introduce two convenient notations: (i) $\mathbf{U}_{1:j}$ denotes the matrix made of the j first columns of $\mathbf{U}$ and (ii) $\mathbf{U}_{1:i,1:j}$ denotes the submatrix of $\mathbf{U}$ made of the rows 1 to i and the columns 1 to j. Lemma 3.8 shows that adding columns to matrices of $\mathrm{St}(n,p)$ increases the geodesic distance. The added value of Lemma 3.8 is that it is often easy to deal with p even and difficult to solve the case p odd, as we experienced in section 3.4.3. Since an odd integer is always surrounded by even integers (groundbreaking news), Lemma 3.8 is convenient to extend a result when it is only proven for p even. The statement is relatively straightforward under the Euclidean metric but has proven itself a bit more touchy for the generalized β-metric.

Lemma 3.8. *Let $\beta\geq\frac{1}{2}$ and $p>1$, we have $d_\beta(\mathbf{U},\widetilde{\mathbf{U}})\geq d_\beta(\mathbf{U}_{1:p-1},\widetilde{\mathbf{U}}_{1:p-1})$ for all $\mathbf{U},\widetilde{\mathbf{U}}\in\mathrm{St}(n,p)$,* *Proof.* The author apologizes in advance for this very computational proof. Let $\mathbf{A}, \mathbf{B}$ such that:

$$\{\mathbf{A}, \mathbf{B}\} \subseteq \operatorname{argmin}_{\mathbf{A},\mathbf{B}} \Bigg\{ \sqrt{\beta\|\mathbf{A}\|_{\mathrm{F}}^2 + \|\mathbf{B}\|_{\mathrm{F}}^2} \mid \mathbf{A} \in \operatorname{Skew}(p),\ \mathbf{B} \in \mathbb{R}^{(n-p)\times p}, \widetilde{\mathbf{U}} = \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix} \exp\left(\begin{bmatrix} 2\beta\mathbf{A} & -\mathbf{B}^T \\ \mathbf{B} & \mathbf{0}\end{bmatrix}\right) \begin{bmatrix} e^{(1-2\beta)\mathbf{A}} \\ \mathbf{0}\end{bmatrix} \Bigg\}.$$

By definition, we have $d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) = \sqrt{\beta\|\mathbf{A}\|_{\mathrm{F}}^2 + \|\mathbf{B}\|_{\mathrm{F}}^2} = \frac{\sqrt{2}}{2}\|\mathbf{S}_\beta\|_{\mathrm{F}}$ where $\mathbf{S}_\beta := \begin{bmatrix} 2\beta\mathbf{A} & -\mathbf{B}^T \\ \mathbf{B} & \mathbf{0}\end{bmatrix}$.
Let us consider the curve $\widetilde{\gamma}$ given by $\widetilde{\gamma} := \gamma\mathbf{I}_{p\times(p-1)}$ that goes from $\mathbf{U}_{1:p-1}$ to $\widetilde{\mathbf{U}}_{1:p-1}$ in $\operatorname{St}(n, p-1)$. The length of $\widetilde{\gamma}$ in $\operatorname{St}(n, p-1)$ is $l_\beta(\widetilde{\gamma}) = \int_0^1 \|\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\gamma}(t)\|_\beta \,\mathrm{d}t$ where

$$\begin{aligned}
\left\|\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\gamma}(t)\right\|_\beta^2 &= \operatorname{Tr}\left(\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\gamma}(t)^T(\mathbf{I} - (1-\beta)\widetilde{\gamma}(t)\widetilde{\gamma}(t)^T)\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\gamma}(t)\right) \\
&= \operatorname{Tr}\left(\mathbf{I}_{(p-1)\times p}\frac{\mathrm{d}}{\mathrm{d}t}\gamma(t)^T(\mathbf{I} - (1-\beta)\gamma(t)\begin{bmatrix}\mathbf{I}_{p-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n-p+1}\end{bmatrix}\gamma(t)^T)\frac{\mathrm{d}}{\mathrm{d}t}\gamma(t)\mathbf{I}_{p\times(p-1)}\right).
\end{aligned}$$

We can write $\frac{\mathrm{d}}{\mathrm{d}t}\gamma(t) = \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix} e^{S_\beta t}\mathbf{S}_{\frac{1}{2}}\mathbf{R}\mathbf{I}_{n\times p}$ if we define $\mathbf{R} := \begin{bmatrix} e^{(1-2\beta)\mathbf{A}t} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-p}\end{bmatrix}$. This yields

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\gamma(t)^T(\mathbf{I} - (1-\beta)\gamma(t)\begin{bmatrix}\mathbf{I}_{p-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n-p+1}\end{bmatrix}\gamma(t)^T)\frac{\mathrm{d}}{\mathrm{d}t}\gamma(t) \\
&\qquad = \mathbf{I}_{p\times n}\mathbf{R}^T(\mathbf{S}_{\frac{1}{2}}^T\mathbf{S}_{\frac{1}{2}} - (1-\beta)\mathbf{S}_{\frac{1}{2}}^T\mathbf{R}\begin{bmatrix}\mathbf{I}_{p-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n-p+1}\end{bmatrix}\mathbf{R}^T\mathbf{S}_{\frac{1}{2}})\mathbf{R}\mathbf{I}_{n\times p}.
\end{aligned}$$

Therefore, we can write

$$\begin{aligned}
&\left\|\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\gamma}(t)\right\|_\beta^2 \\
&= \operatorname{Tr}\left(\mathbf{I}_{p-1\times n}\mathbf{R}^T(\mathbf{S}_{\frac{1}{2}}^T\mathbf{S}_{\frac{1}{2}} - (1-\beta)\mathbf{S}_{\frac{1}{2}}^T\mathbf{R}\begin{bmatrix}\mathbf{I}_{p-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n-p+1}\end{bmatrix}\mathbf{R}^T\mathbf{S}_{\frac{1}{2}})\mathbf{R}\mathbf{I}_{n\times p-1}\right) \\
&= \operatorname{Tr}\left(\mathbf{I}_{p-1\times n}\mathbf{R}^T(\mathbf{S}_{\frac{1}{2}}^T\mathbf{S}_{\frac{1}{2}} - (1-\beta)\mathbf{S}_{\frac{1}{2}}^T\mathbf{R}\left(\begin{bmatrix}\mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n-p}\end{bmatrix} - \begin{bmatrix}\mathbf{0}_{p-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}_{n-p+1}\end{bmatrix}\right)\mathbf{R}^T\mathbf{S}_{\frac{1}{2}})\mathbf{R}\mathbf{I}_{n\times p-1}\right) \\
&= \operatorname{Tr}\left(\mathbf{I}_{p-1\times p}(e^{(1-2\beta)\mathbf{A}t})^T(\beta\mathbf{A}^T\mathbf{A} + \mathbf{B}^T\mathbf{B})e^{(1-2\beta)\mathbf{A}t}\mathbf{I}_{p\times p-1}\right) \\
&+ \operatorname{Tr}\left(\mathbf{I}_{p-1\times p}(e^{(1-2\beta)\mathbf{A}t})^T((1-\beta)\mathbf{A}^Te^{(1-2\beta)\mathbf{A}t}\begin{bmatrix}\mathbf{0}_{p-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}_{n-p+1}\end{bmatrix}(e^{(1-2\beta)\mathbf{A}t})^T\mathbf{A})e^{(1-2\beta)\mathbf{A}t}\mathbf{I}_{p\times p-1}\right).
\end{aligned}$$

Here we can introduce three following properties: (i) $\operatorname{Tr}(\mathbf{I}_{p-1\times p}\mathbf{M}^T\mathbf{M}\mathbf{I}_{p\times p-1}) \leq \operatorname{Tr}(\mathbf{M}^T\mathbf{M})$, (ii) $\|\mathbf{Q}\mathbf{M}\|_{\mathrm{F}} = \|\mathbf{M}\|_{\mathrm{F}}$ if $\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = I$ and (iii) $\mathbf{A}$ and $e^{(1-2\beta)\mathbf{A}t}$ commute. This gives

$$l_\beta(\widetilde{\gamma}) \leq \int_0^1 \left(\beta\|\mathbf{A}_{1:p,1:p-1}\|_{\mathrm{F}}^2 + \|\mathbf{B}\|_{\mathrm{F}}^2 + (1-\beta)\|\mathbf{A}_{p,1:p-1}\|_{\mathrm{F}}^2\right)^{\frac{1}{2}} \mathrm{d}t.$$

As a last argument, for all $\beta \geq \frac{1}{2}$, we have $(1-\beta) \leq \beta$. This finally yields

$$l_\beta(\widetilde{\gamma}) \leq \sqrt{\beta(\|\mathbf{A}_{1:p,1:p-1}\|_{\mathrm{F}}^2 + \|\mathbf{A}_{p,1:p-1}\|_{\mathrm{F}}^2) + \|\mathbf{B}\|_{\mathrm{F}}^2} = \sqrt{\beta\|\mathbf{A}\|_{\mathrm{F}}^2 + \|\mathbf{B}\|_{\mathrm{F}}^2}.$$

By definition, we have $d_\beta(\mathbf{U}_{1:p-1}, \widetilde{\mathbf{U}}_{1:p-1}) \leq l_\beta(\widetilde{\gamma}) \leq \sqrt{\beta\|\mathbf{A}\|_{\mathrm{F}}^2 + \|\mathbf{B}\|_{\mathrm{F}}^2} = d_\beta(\mathbf{U}, \widetilde{\mathbf{U}})$. □

Conjecture 3.1. *Lemma 3.8 is true for all* $\beta > 0$.

3.12 The diameter conjecture under the β-metric

In this section, we try to find the diameter of the Stiefel manifold endowed with any β-metric, $(\mathrm{St}(n,p), g_\beta)$. The diameter of a compact manifold endowed with a metric is the maximum distance between points on this manifold. We already know it for $\beta = 1$: it is $\pi\sqrt{p}$. By the equivalence of the β-distances, we already know that $\pi\sqrt{p}$ is an upper bound on the diameter for every $\beta \in (0,1]$. Our numerical experiments until here clearly indicate that we should have $\pi\frac{\sqrt{2p}}{2} = \pi\sqrt{\beta p}$ for the canonical metric with p even and $\pi\frac{\sqrt{2(p+1)}}{2} = \pi\sqrt{\beta(p+1)}$ if p is odd. Can we do better than $\pi\sqrt{p}$? This section tries to answer that question. Our answer will however be subject to Conjecture 3.2 we present below.

3.12.1 The conjecture

Conjecture 3.2. *Let $0 \le r < 2\sqrt{p}$ and $\mathbf{U} \in \mathrm{St}(n,p)$. Define*

$$d_{max} = \max_{\widetilde{\mathbf{U}} \in \mathrm{St}(n,p)} \{d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) \mid \|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}} \le r\}.$$

Then, there exists $\hat{\mathbf{U}} \in \mathrm{St}(n,p)$ such that $\|\hat{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}} > r$ and $d_\beta(\mathbf{U}, \hat{\mathbf{U}}) \ge d_{max}$.

This Conjecture 3.2 is the only missing piece to obtain the proof of the diameter of the Stiefel manifold endowed with the β-metric.

Lemma 3.9. *Suppose Conjecture 3.2 holds. For all $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$, we have $d_\beta(\widetilde{\mathbf{U}}, \mathbf{U}) \le \max_{\hat{\mathbf{U}} \in \mathrm{St}(n,p)} \{d_\beta(\hat{\mathbf{U}}, \mathbf{U}) \mid \|\hat{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}} = 2\sqrt{p}\}$.*

Proof. Let $\{r_i\}_{\mathbb{N}}$ be any non-negative monotonically increasing sequence such that $\lim_{i\to\infty} r_i = 2\sqrt{p}$. We define two other sequences based on $\{r_i\}_{\mathbb{N}}$:

$$m_i = \max_{\widetilde{\mathbf{U}} \in \mathrm{St}(n,p)} \left\{d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) \mid \|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}} \le r_i\right\}, \tag{3.17}$$

$$M_i = \max_{\widetilde{\mathbf{U}} \in \mathrm{St}(n,p)} \left\{d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) \mid r_i \le \|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}} \le 2\sqrt{p}\right\}. \tag{3.18}$$

We can notice that the sequence $\{m_i\}_{\mathbb{N}}$ is non-decreasing and $\{M_i\}_{\mathbb{N}}$ is non-increasing. Using Conjecture 3.2, we have that

$$\begin{aligned}
& m_i \le M_i \quad \text{for all } i \in \mathbb{N} \\
\implies & \lim_{i\to\infty} m_i \le \lim_{i\to\infty} M_i \\
\implies & \max_{\widetilde{\mathbf{U}} \in \mathrm{St}(n,p)} \{d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) \mid \|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}} \le \lim_{i\to\infty} r_i\} \le \max_{\widetilde{\mathbf{U}} \in \mathrm{St}(n,p)} \{d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) \mid \lim_{i\to\infty} r_i \le \|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}} \le 2\sqrt{p}\} \\
\implies & \max_{\widetilde{\mathbf{U}} \in \mathrm{St}(n,p)} \{d_\beta(\mathbf{U}, \widetilde{\mathbf{U}}) \mid \|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}} \le 2\sqrt{p}\} \le \max_{\hat{\mathbf{U}} \in \mathrm{St}(n,p)} \{d_\beta(\hat{\mathbf{U}}, \mathbf{U}) \mid \|\hat{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}} = 2\sqrt{p}\} \\
\iff & \text{For all } \mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p) : d_\beta(\widetilde{\mathbf{U}}, \mathbf{U}) \le \max_{\hat{\mathbf{U}} \in \mathrm{St}(n,p)} \{d_\beta(\hat{\mathbf{U}}, \mathbf{U}) \mid \|\hat{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}} = 2\sqrt{p}\}.
\end{aligned}$$

$\square$

Theorem 3.9. *Suppose Conjecture 3.2 holds and* $\beta \in [\frac{1}{2}, 1]$. *For all* $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$, *we have* $d_\beta(\widetilde{\mathbf{U}}, \mathbf{U}) \leq \begin{cases} \pi\sqrt{\beta p} \text{ for } p \text{ even} \\ \pi\sqrt{\beta(p+1)} \text{ for } p \text{ odd} \end{cases}$.

Proof. First consider p even. Using Lemma 3.9, for all $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$, we have

$$d_\beta(\widetilde{\mathbf{U}}, \mathbf{U}) \leq \max_{\hat{\mathbf{U}} \in \mathrm{St}(n,p)} \left\{ d_\beta(\hat{\mathbf{U}}, \mathbf{U}) \mid \|\hat{\mathbf{U}} - \mathbf{U}\|_\mathrm{F} = 2\sqrt{p} \right\}.$$

But $\hat{\mathbf{U}} = -\mathbf{U}$ is the only matrix such that $\|\hat{\mathbf{U}} - \mathbf{U}\|_\mathrm{F} = 2\sqrt{p}$. This yields

$$\max_{\hat{\mathbf{U}} \in \mathrm{St}(n,p)} \left\{ d_\beta(\hat{\mathbf{U}}, \mathbf{U}) \mid \|\hat{\mathbf{U}} - \mathbf{U}\|_\mathrm{F} = 2\sqrt{p} \right\} = d_\beta(-\mathbf{U}, \mathbf{U}) = \pi\sqrt{\beta p}.$$

Now consider p odd. Using Lemma 3.8(hence the constraint $\beta \in [\frac{1}{2}, 1]$), the maximal distance for p is less or equal than the maximal distance for $p+1$. Since $p+1$ is even, we have

$$\max_{\hat{\mathbf{U}} \in \mathrm{St}(n,p)} \left\{ d_\beta(\hat{\mathbf{U}}, \mathbf{U}) \mid \|\hat{\mathbf{U}} - \mathbf{U}\|_\mathrm{F} = 2\sqrt{p} \right\} \leq \pi\sqrt{\beta(p+1)}.$$

□

3.12.2 Visualization of the bounds

Figure 3.10 and 3.11 show a numerical corroboration of the bounds and the conjectured diameter. The canonical distance is obtained using [48, Algorithm 4].

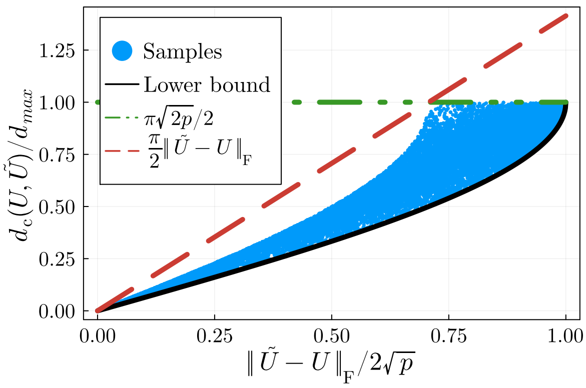

Figure 3.10: Distances computed on 15000 samples, $(n,p) = (4,2)$, $d_{max} := \pi\frac{\sqrt{2p}}{2}$.

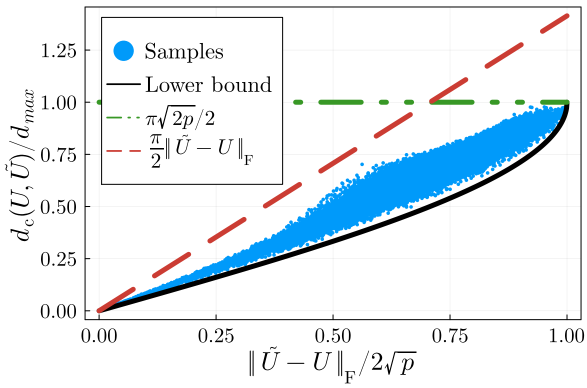

Figure 3.11: Distances computed on 40000 samples, $(n,p) = (8,4)$, $d_{max} := \pi\frac{\sqrt{2p}}{2}$.

3.13 Stiefel endowed with the canonical metric

In this section, we take a particular interest into Stiefel endowed with its canonical metric, $(\mathrm{St}(n,p), g_\mathrm{c})$. The canonical metric is particular since it is the one inherited from the quotient geometry of Stiefel, $\mathrm{St}(n,p) = \mathrm{SO}(n)/\mathrm{SO}(n-p)$. The canonical metric has been used in several applications, for instance in machine learning [28] or optimization [20, 46]. For this particular metric we can leverage our knowledge from $\mathrm{SO}(n)$ to prove more properties of the minimizing geodesics.

3.13.1 The canonical distance on $\mathrm{St}(n,p)$ as a distance on $\mathrm{SO}(n)$

We start by recalling that the problem of computing the canonical distance in $\mathrm{St}(n,p)$ is equivalent to a problem finding the shortest distance between equivalence classes in $\mathrm{SO}(n)$ (see (1.5)). Theorem 3.10 can also be seen as a particular case of Theorem 1.2 for the canonical metric ($\beta = \frac{1}{2}$).

Theorem 3.10. $d_c(\mathbf{U}, \widetilde{\mathbf{U}}) = \min\left\{ \frac{\sqrt{2}}{2} d\left(\begin{bmatrix} \mathbf{U} & \mathbf{U}_\perp \end{bmatrix}, \begin{bmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \mathbf{R} \end{bmatrix} \right) \mid \mathbf{R} \in \mathrm{SO}(n-p) \right\}$ *for all* $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$ *and* $\mathbf{U}_\perp, \widetilde{\mathbf{U}}_\perp$ *orthogonal completions in* $\mathrm{SO}(n)$.

Proof. The following development is more of a way to retrieve the link between $\mathrm{St}(n,p)$ and $\mathrm{SO}(n)$ rather than a proof, since a proper proof should use more fundamental arguments. Recall the definition of the canonical distance:

$$d_c(\mathbf{U}, \widetilde{\mathbf{U}}) = \min_{\mathbf{A},\mathbf{B}} \left\{ \sqrt{\frac{1}{2}\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2} \mid \mathbf{A} \in \mathrm{Skew}(p),\ \mathbf{B} \in \mathbb{R}^{(n-p)\times p}, \right.$$
$$\left. \widetilde{\mathbf{U}} = \begin{bmatrix} \mathbf{U} & \mathbf{U}_\perp \end{bmatrix} \exp\left(\begin{bmatrix} \mathbf{A} & -\mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \right) \mathbf{I}_{n\times p} \right\}.$$

Since $\exp\left(\begin{bmatrix} \mathbf{A} & -\mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \right) \in \mathrm{SO}(n)$ and because there is at least a geodesic that is a minimizing path, we can write

$$d_c(\mathbf{U}, \widetilde{\mathbf{U}}) = \min_{\mathbf{Q}} \left\{ \frac{\sqrt{2}}{2} \|\log(\mathbf{Q})\|_F \mid \mathbf{Q} \in \mathrm{SO}(n),\ \widetilde{\mathbf{U}} = \begin{bmatrix} \mathbf{U} & \mathbf{U}_\perp \end{bmatrix} \mathbf{Q} \mathbf{I}_{n\times p} \right\}. \tag{3.19}$$

Since $\{\widetilde{\mathbf{U}}_\perp \mathbf{R} \mid \mathbf{R} \in \mathrm{SO}(n-p)\} = \{\mathbf{X} \in \mathrm{St}(n, n-p) \mid \begin{bmatrix} \widetilde{\mathbf{U}} & \mathbf{X} \end{bmatrix} \in \mathrm{SO}(n)\}$, (3.19) can be written as

$$\begin{aligned} d_c(\mathbf{U}, \widetilde{\mathbf{U}}) &= \min_{\mathbf{Q},\mathbf{R}} \left\{ \frac{\sqrt{2}}{2} \|\log(\mathbf{Q})\|_F \mid \mathbf{Q} \in \mathrm{SO}(n), \mathbf{R} \in \mathrm{SO}(n-p),\ \begin{bmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \mathbf{R} \end{bmatrix} = \begin{bmatrix} \mathbf{U} & \mathbf{U}_\perp \end{bmatrix} \mathbf{Q} \right\} \\ &= \min_{\mathbf{R}} \left\{ \frac{\sqrt{2}}{2} d\left(\begin{bmatrix} \mathbf{U} & \mathbf{U}_\perp \end{bmatrix}, \begin{bmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \mathbf{R} \end{bmatrix} \right) \mid \mathbf{R} \in \mathrm{SO}(n-p) \right\} \qquad \text{Using Corollary 2.3.} \end{aligned}$$

Finding the canonical distance on $\mathrm{St}(n,p)$ is thus equivalent to extending the basis $\mathbf{U}$ arbitrarily and $\widetilde{\mathbf{U}}$ in $\mathrm{SO}(n)$ in an optimal manner to compute the distance in $\mathrm{SO}(n)$. □

Theorem 3.10 yields an interesting consequence for algorithms computing candidates for a minimizing geodesic between two points on $\mathrm{St}(n,p)$: the completion $\mathbf{U}_\perp$ can be chosen arbitrarily.

3.13.2 A spectral upper bound for the cut time of canonical geodesics

In the case of $\mathrm{SO}(n)$, the spectrum of the skew-symmetric matrix characterizing a geodesic (see (1.35)) provides a *necessary and sufficient* condition for the geodesic to be minimizing: for $\mathbf{Q} \in \mathrm{SO}(n)$ and $\mathbf{\Delta} = \mathbf{Q}\mathbf{A} \in T_{\mathbf{Q}}\mathrm{SO}(n)$, $\mathbf{\Delta}$ is a logarithm if and only if $\|\mathbf{A}\|_2 \leq \pi$ (see Corollary 2.1). This is due to the uniqueness of the rotation matrix mapping a point of $\mathrm{SO}(n)$ on another. With $(\mathrm{St}(n,p), g_c)$, we lose this property of uniqueness in general and thus, it would seem, the hope for a necessary and sufficient condition. However, it is still possible to obtain a *necessary* condition on the spectrum the skew-symmetric matrix characterizing the canonical geodesic. This condition will help to show several geometric properties of the *minimizing* canonical geodesics.

Lemma 3.10. *For all* $\mathbf{U} \in \mathrm{St}(n,p)$ *and all canonical geodesics* $\gamma_c : [0,1] \to \mathrm{St}(n,p) : t \to \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix} e^{\mathbf{S}t}\mathbf{I}_{n\times p}$ *where* $\mathbf{S} = \begin{bmatrix}\mathbf{A} & -\mathbf{B}^T \\ \mathbf{B} & \mathbf{0}\end{bmatrix} \in \mathrm{Skew}(n)$*: if* γ_c *is minimizing for all* $t \in [0,1]$*, then we have* $\|\mathbf{S}\|_2 \leq \pi$.

Proof. Suppose the statement does not hold. Then there exists $\widetilde{\mathbf{S}} := \begin{bmatrix}\widetilde{\mathbf{A}} & -\widetilde{\mathbf{B}}^T \\ \widetilde{\mathbf{B}} & \widetilde{\mathbf{C}}\end{bmatrix} \in \mathrm{Skew}(n)$ such that $e^{\mathbf{S}} = e^{\widetilde{\mathbf{S}}}$ and $\|\widetilde{\mathbf{S}}\|_2 \leq \pi$ (see construction of Theorem 2.1). We build $\widetilde{\gamma} : [0,1] \to \mathrm{St}(n,p) : t \to \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix} e^{\widetilde{\mathbf{S}}t}\mathbf{I}_{n\times p}$. The length of $\widetilde{\gamma}$ is

$$\begin{aligned} l_c(\widetilde{\gamma}) &= \sqrt{\frac{1}{2}\|\widetilde{\mathbf{A}}\|_F^2 + \|\widetilde{\mathbf{B}}\|_F^2} \\ &\leq \sqrt{\frac{1}{2}(\|\widetilde{\mathbf{A}}\|_F^2 + \|\widetilde{\mathbf{C}}\|_F^2) + \|\widetilde{\mathbf{B}}\|_F^2} \\ &= \frac{\sqrt{2}}{2}\|\widetilde{\mathbf{S}}\|_F \leq \frac{\sqrt{2}}{2}\|\mathbf{S}\|_F = l_c(\gamma_c). \end{aligned}$$

Hence, there would exists another curve $\widetilde{\gamma}$, not necessarily a canonical geodesic if $\widetilde{\mathbf{C}} \neq 0$, that has a shorter canonical length than γ_c . □

3.13.3 Length-minimal geodesics go onward

We have seen that $\mathrm{St}(n,p)$ can be seen as a subset of the hypersphere $\mathbb{S}^{n\times p}_{\sqrt{p}}$. On hyperspheres, minimizing geodesics constantly increase the Frobenius distance from the starting point. As a contradictory example, this property does not hold on an ellipsoid, minimizing geodesics do not constantly increase the Frobenius distance. We show in Theorem 3.11 that Stiefel under the canonical metric manifold is not alike an ellipsoid: minimizing geodesics constantly increase the Frobenius distance from the starting point.

Theorem 3.11. *For all* $\mathbf{U} \in \mathrm{St}(n,p)$ *and all canonical geodesics* $\gamma_c : [0,1] \to \mathrm{St}(n,p) : t \to \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix} e^{\mathbf{S}t}\mathbf{I}_{n\times p}$ *where* $\mathbf{S} = \begin{bmatrix}\mathbf{A} & -\mathbf{B}^T \\ \mathbf{B} & \mathbf{0}\end{bmatrix} \in \mathrm{Skew}(n)$*: if* γ_c *is minimizing for all* $t \in [0,1]$*, then* $\|\gamma_c(t) - \mathbf{U}\|_F$ *is a non-decreasing function of* t *for all* $t \in [0, \frac{\pi}{\|\mathbf{S}\|_2}]$.

Proof. Define $f : \mathbb{R} \to \mathbb{R}$, the squared Frobenius distance from $\gamma(t)$ to the initial point $\mathbf{U}$ and define $\mathbf{Q} := \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}$. We can compute

$$\begin{aligned} f(t) = \|\gamma_c(t) - \mathbf{U}\|_F^2 &= \mathrm{Tr}(\gamma_c(t) - \mathbf{U})^T(\gamma_c(t) - \mathbf{U})) \\ &= \mathrm{Tr}(\mathbf{Q}e^{\mathbf{S}t}\mathbf{I}_{n\times p} - \mathbf{U})^T(\mathbf{Q}e^{\mathbf{S}t}\mathbf{I}_{n\times p} - \mathbf{U})) \\ &= 2p - \mathrm{Tr}(\mathbf{I}_{p\times n}(e^{\mathbf{S}t} + (e^{\mathbf{S}t})^T)\mathbf{I}_{n\times p}) \\ &= 2p - \mathrm{Tr}(\mathbf{I}_{p\times n}(e^{\mathbf{S}t} + e^{-\mathbf{S}t})\mathbf{I}_{n\times p}) \\ &= 2p - 2\,\mathrm{Tr}(\mathbf{I}_{p\times n}\cosh(\mathbf{S}t)\mathbf{I}_{n\times p}). \end{aligned}$$

This yields the time derivative of f:

$$\frac{\mathrm{d}}{\mathrm{d}t}f(t) = -2\operatorname{Tr}(\mathbf{I}_{p\times n}\sinh(\mathbf{S}t)\mathbf{S}\mathbf{I}_{n\times p}).$$

We reformulate the quantity $\sinh(\mathbf{S}t)\mathbf{S}$ introducing the eigenvalue decomposition of $\mathbf{S} = \mathbf{V}i\mathbf{\Theta}\mathbf{V}^*$:

$$\begin{aligned}\sinh(\mathbf{S}t)\mathbf{S} &= \frac{\mathbf{V}e^{i\mathbf{\Theta}t}\mathbf{V}^* - \mathbf{V}e^{-i\mathbf{\Theta}t}\mathbf{V}^*}{2}\mathbf{V}i\mathbf{\Theta}\mathbf{V}^*\\ &= \mathbf{V}\frac{e^{i\mathbf{\Theta}t} - e^{-i\mathbf{\Theta}t}}{2}i\mathbf{\Theta}\mathbf{V}^*\\ &= -\mathbf{V}\sin(\mathbf{\Theta}t)\mathbf{\Theta}\mathbf{V}^*.\end{aligned}$$

Replacing the new expression for $\sinh(\mathbf{S}t)\mathbf{S}$ into $\frac{\mathrm{d}}{\mathrm{d}t}f(t)$ yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}f(t) &= -2\operatorname{Tr}(\mathbf{I}_{p\times n}\sinh(\mathbf{S}t)\mathbf{S}\mathbf{I}_{n\times p})\\ &= 2\operatorname{Tr}(\mathbf{I}_{p\times n}\mathbf{V}\sin(\mathbf{\Theta}t)\mathbf{\Theta}\mathbf{V}^*\mathbf{I}_{n\times p})\\ &= 2\sum_{i=1}^{p}\sum_{j=1}^{n}\theta_j\sin(\theta_j t)\bar{\mathbf{V}}_{i,j}\mathbf{V}_{i,j}\\ &\geq 2\sum_{i=1}^{p}\min_{j=1,..,n}(\theta_j\sin(\theta_j t))\sum_{j=1}^{n}\bar{\mathbf{V}}_{i,j}\mathbf{V}_{i,j}\\ &= 2p\min_{j=1,..,n}(\theta_j\sin(\theta_j t))\\ &\geq 0.\end{aligned}$$

At the last step, we used the fact that for all $c>0$, $\theta\in[-c,c]$, $t\in[0,\frac{\pi}{c}]$ we have $\sin(\theta t)\theta\geq 0$. The latter inequality is true by Lemma 3.10. Since the square root is a monotonically increasing bijection from $\mathbb{R}^+$ to itself, if $f(t)$ is a non-decreasing function, then so is $\sqrt{f(t)}$. □

Corollary 3.7. *For all $\mathbf{U}\in\mathrm{St}(n,p)$, let $\gamma_{\mathrm{c}}:\mathbb{R}^+\to\mathrm{St}(n,p)$ be a canonical geodesic with $\gamma_{\mathrm{c}}(0)=\mathbf{U}$. If $\frac{\mathrm{d}}{\mathrm{d}t}\Big[\|\gamma_{\mathrm{c}}(t)-\mathbf{U}\|_{\mathrm{F}}\Big]_{t=t^*}<0$ for some $t_*>0$, then $\gamma_{\mathrm{c}}:[0,t^*]\to\mathrm{St}(n,p)$ is not a minimizing geodesic, i.e, the canonical cut time $\mathcal{T}_{c,\mathbf{U}}(\mathbf{\Delta})$ satisfies $\frac{\mathrm{d}}{\mathrm{d}t}\Big[\|\gamma_{\mathrm{c}}(t)-\mathbf{U}\|_{\mathrm{F}}\Big]_{0\leq t\leq\mathcal{T}_{c,\mathbf{U}}(\mathbf{\Delta})}\geq 0$.*

Proof. Combine Theorem 3.11 and Lemma 3.10. □

3.13.4 Canonical geodesics lie in a spherical cap

For all $\mathbf{U},\widetilde{\mathbf{U}}\in\mathrm{St}(n,p)$, $\mathbf{U}\neq-\widetilde{\mathbf{U}}$, recall from Equations (3.5), (3.6) and (3.7) the definitions of $\mathbb{S}^{n\times p}_{\sqrt{p}}$, $\Sigma_{\mathbf{U},\widetilde{\mathbf{U}}}$ and $\Pi_{\mathbf{U},\widetilde{\mathbf{U}}}$. The hyperplane $\Pi_{\mathbf{U},\widetilde{\mathbf{U}}}$ cuts a spherical cap of $\mathbb{S}^{n\times p}_{\sqrt{p}}$ and divides the hypersphere $\Sigma_{\mathbf{U},\widetilde{\mathbf{U}}}$ into two symmetrical parts. Our claim is that a *minimizing* geodesic of Stiefel endowed with the canonical metric, $(\mathrm{St}(n,p),g_{\mathrm{c}})$, always lies in the spherical cap, i.e., the part of $\mathbb{R}^{n\times p}$ which does not contain the origin.

Theorem 3.12. *For all $\mathbf{U},\widetilde{\mathbf{U}}\in\mathrm{St}(n,p)$, $\mathbf{U}\neq-\widetilde{\mathbf{U}}$, if $\gamma_{\mathrm{c}}:[0,1]\to\mathrm{St}(n,p)$ is a canonical minimizing geodesic with $\mathbf{U}=\gamma_{\mathrm{c}}(0)$ and $\widetilde{\mathbf{U}}=\gamma_{\mathrm{c}}(1)$, then $\operatorname{Tr}(\gamma_{\mathrm{c}}(t)^T(\mathbf{U}+\widetilde{\mathbf{U}}))\geq p+\operatorname{Tr}(\mathbf{U}^T\widetilde{\mathbf{U}})$ for all $t\in[0,1]$.*

Proof. First, recall the formula of a canonical geodesic: $\gamma_c : [0,1] \to \mathrm{St}(n,p) : t \to \begin{bmatrix} \mathbf{U} & \mathbf{U}_\perp \end{bmatrix} e^{\mathbf{S}t}\mathbf{I}_{n\times p}$ where $\mathbf{S} := \begin{bmatrix} \mathbf{A} & -\mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \in \mathrm{Skew}(n)$. We transform the equation of the hyperplane $\Pi_{\mathbf{U},\widetilde{\mathbf{U}}}$ in a convenient expression by the means of the eigendecomposition of $\mathbf{S}$. From the equation of the geodesic, we get

$$\widetilde{\mathbf{U}} = \begin{bmatrix} \mathbf{U} & \mathbf{U}_\perp \end{bmatrix} e^{\mathbf{S}}\mathbf{I}_{n\times p} \Longrightarrow e^{\mathbf{S}}\mathbf{I}_{n\times p} = \begin{bmatrix} \mathbf{U}^T \\ \mathbf{U}_\perp^T \end{bmatrix} \widetilde{\mathbf{U}} = \begin{bmatrix} \mathbf{U}^T\widetilde{\mathbf{U}} \\ \mathbf{U}_\perp^T\widetilde{\mathbf{U}} \end{bmatrix}. \tag{3.20}$$

This key step (3.20) allows us to write (3.7) in terms of the eigendecomposition of $\mathbf{S} = \mathbf{V}i\mathbf{\Theta}\mathbf{V}^*$. We can write

$$\begin{aligned}
\mathrm{Tr}\left(\gamma_c(t)^T(\mathbf{U}+\widetilde{\mathbf{U}})\right) &= \mathrm{Tr}\left(\mathbf{I}_{p\times n}(e^{\mathbf{S}t})^T \begin{bmatrix} \mathbf{U}^T \\ \mathbf{U}_\perp^T \end{bmatrix} (\mathbf{U}+\widetilde{\mathbf{U}})\right) && \text{By definition of } \gamma \\
&= \mathrm{Tr}\left(\mathbf{I}_{p\times n}(e^{\mathbf{S}t})^T\mathbf{I}_{n\times p} + \mathbf{I}_{p\times n}(e^{\mathbf{S}t})^T \begin{bmatrix} \mathbf{U}^T\widetilde{\mathbf{U}} \\ \mathbf{U}_\perp^T\widetilde{\mathbf{U}} \end{bmatrix}\right) \\
&= \mathrm{Tr}\left(\mathbf{I}_{p\times n}(e^{\mathbf{S}t})^T\mathbf{I}_{n\times p} + \mathbf{I}_{p\times n}(e^{\mathbf{S}t})^T e^{\mathbf{S}}\mathbf{I}_{n\times p}\right) && \text{Using Equation (3.20)} \\
&= \mathrm{Tr}\left(\mathbf{I}_{p\times n}\mathbf{V}(e^{-i\mathbf{\Theta}t} + e^{-i\mathbf{\Theta}(t-1)})\mathbf{V}^*\mathbf{I}_{n\times p}\right) \\
&= \mathrm{Tr}\left(\mathbf{I}_{p\times n}\mathbf{V}(e^{i\mathbf{\Theta}t} + e^{i\mathbf{\Theta}(t-1)})\mathbf{V}^*\mathbf{I}_{n\times p}\right).
\end{aligned}$$

The right hand side of (3.7) can also be rewritten:

$$\begin{aligned}
p + \mathrm{Tr}(\mathbf{U}^T\widetilde{\mathbf{U}}) &= \mathrm{Tr}\left(\mathbf{I}_{p\times n}(\mathbf{I}+e^{\mathbf{S}})\mathbf{I}_{n\times p}\right) \\
&= \mathrm{Tr}\left(\mathbf{I}_{p\times n}\mathbf{V}(\mathbf{I}+e^{i\mathbf{\Theta}})\mathbf{V}^*\mathbf{I}_{n\times p}\right).
\end{aligned}$$

Combined, these new expressions show that the geodesics lie in the spherical cap. We consider n even so that every eigenvalue of $\mathbf{S}$ has a conjugate pair, but the case n odd brings no contradiction to the final conclusion. The reason is detailed further.

$$\begin{aligned}
\mathrm{Tr}\left(\gamma_c(t)^T(\mathbf{U}+\widetilde{\mathbf{U}})\right) - p - \mathrm{Tr}(\mathbf{U}^T\widetilde{\mathbf{U}}) &= \mathrm{Tr}\left(\mathbf{I}_{p\times n}\mathbf{V}(e^{i\mathbf{\Theta}t} + e^{i\mathbf{\Theta}(t-1)} - I - e^{i\mathbf{\Theta}})\mathbf{V}^*\mathbf{I}_{n\times p}\right) \\
&=: \mathrm{Tr}\left(\mathbf{I}_{p\times n}\mathbf{V}\mathbf{\Lambda}\mathbf{V}^*\mathbf{I}_{n\times p}\right) && \text{Short notation } \mathbf{\Lambda} \\
&= \sum_{i=1}^{p}\sum_{j=1}^{n} \lambda_j \mathbf{V}_{ij}\bar{\mathbf{V}}_{ij} \\
&= \sum_{i=1}^{p}\sum_{j=1}^{n} \frac{1}{2}(\lambda_j + \bar{\lambda}_j)\mathbf{V}_{ij}\bar{\mathbf{V}}_{ij} && \text{Using Lemma 1.1} \\
&\geq \sum_{i=1}^{p} \frac{1}{2} \min_{j=1,..,n} (\lambda_j + \bar{\lambda}_j) \sum_{j=1}^{n} \mathbf{V}_{ij}\bar{\mathbf{V}}_{ij} \\
&= \sum_{i=1}^{p} \frac{1}{2} \min_{j=1,..,n} (\lambda_j + \bar{\lambda}_j) \\
&\geq 0\ .
\end{aligned}$$

The last inequality can be shown by writing explicitly the sum of two conjugate eigenvalues. We omit the indexing j for clarity. This yields

$$\begin{aligned}
\lambda + \bar{\lambda} = 2\mathrm{Re}(\lambda) &= 2\left[\cos(\theta t) + \cos\left(\theta(t-1)\right) - (1+\cos(\theta))\right] \\
&= 4\cos\left(\frac{\theta}{2}\right)\left[\cos\left(\frac{\theta}{2}(1-2t)\right) - \cos\left(\frac{\theta}{2}\right)\right] && \text{Using Carnot and Simpson's rules} \\
&\geq 0 && \text{Using Lemma 3.10}
\end{aligned}$$

for all $t \in [0,1]$ and $\theta \in [-\pi, \pi]$. In this last step, we used Lemma 3.10 which provided the necessary condition $\theta \in [-\pi, \pi]$ for a canonical geodesic to be minimizing. In case of n odd, one eigenvalue does not come in conjugate pair and is real. This real eigenvalue is $\lambda = 0$ since it is associated to the eigenvalue $\theta = 0$ of $\mathbf{S}$. n odd brings thus no contradiction to the final result. Finally, we have $\mathrm{Tr}\left(\gamma_c(t)^T(\mathbf{U}+\widetilde{\mathbf{U}})\right) \geq p + \mathrm{Tr}(\mathbf{U}^T\widetilde{\mathbf{U}})$ for all $t \in [0,1]$. "The canonical minimizing geodesics lie in the spherical cap". □

The careful reader will notice that Theorems 3.11 and 3.12 are only consequences of Lemma 3.10. Hence, these are weaker statements. The justification for the existence of these theorems is that they provide cheaper conditions to verify. Indeed, the cost of obtaining the eigenvalues of a skew-symmetric matrix grows cubically with the dimension n while checking the theorems has a quadratic cost.

3.13.5 An educated guess for the tight upper bound

Suppose $n \geq 2p$. In order to obtain $\bar{b}(\delta) := \max\left\{d_c(\mathbf{U}, \widetilde{\mathbf{U}}) \mid \|\widetilde{\mathbf{U}} - \mathbf{U}\|_F = \delta\right\}$, one can be interested to answer the "dual" question: "At a fixed canonical distance, how can I minimize the Frobenius distance ?". Fix $d_c(\mathbf{U}, \widetilde{\mathbf{U}}) = |\theta| \in [0, \pi]$. We want to build $\gamma : [0,1] \to \mathrm{St}(n,p)$, defined by $\mathbf{Q} \in \mathrm{SO}(n)$ such that $l_c(\gamma) = \frac{\sqrt{2}}{2}\|\log(\mathbf{Q})\|_F = |\theta|$ and

$$\gamma(1) = \widetilde{\mathbf{U}} = \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}\mathbf{Q}\mathbf{I}_{n\times p}.$$

We want $\|\widetilde{\mathbf{U}} - \mathbf{U}\|_F$ as close as possible to zero. Our claim, guided by experiments from, e.g., Figure 3.10, is that we can not do better than a simple Givens rotation between a column of $\mathbf{U}$ and a column of $\mathbf{U}_\perp$. For this specific choice for $\mathbf{Q}$, we have:

$$\begin{aligned}
\|\widetilde{\mathbf{U}} - \mathbf{U}\|_F^2 &= \sum_{j=1}^{n}((\cos\theta - 1)\mathbf{U}_{j,i} - \sin\theta\mathbf{U}_{\perp,j,k})^2 \quad \text{For some } i \in \{1,...,p\} \text{ and } k \in \{p+1,...,n\}\\
&= \sum_{j=1}^{n}(\cos\theta - 1)^2\mathbf{U}_{j,i}^2 + \sin\theta^2\mathbf{U}_{\perp,j,k}^2 && \text{Columns are orthogonal}\\
&= (\cos\theta - 1)^2 + \sin\theta^2 && \text{Columns are normed}\\
&= 2(1-\cos\theta)\\
&= 4\sin^2(\frac{\theta}{2}).
\end{aligned}$$

Equivalently, we have

$$l_c(\gamma) = |\theta| = 2\arcsin\left(\frac{\|\widetilde{\mathbf{U}} - \mathbf{U}\|_F}{2}\right). \tag{3.21}$$

The domain of the result is limited: it is restricted to $\|\widetilde{\mathbf{U}} - \mathbf{U}\|_F \leq 2$. Indeed, the Frobenius distance can go up to $2\sqrt{p}$, thus (3.21) is only valid on a small range of the possible Frobenius distances. Hence we must generalize the approach in order to reach larger Frobenius distances. At $|\theta| = \pi$, the first column of $\mathbf{U}$ is flipped. Pursuing the same idea, we extend γ by rotating the second column similarly:

$$\widetilde{\mathbf{U}} = \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}
-1 & 0 & ... & 0 & 0 & ....\\
0 & \cos(\theta) & ... & 0 & -\sin(\theta) & ...\\
.. & ... & ... & ... & ... & ...\\
0 & 0 & ... & -1 & 0 & ...\\
0 & \sin\theta & ... & 0 & \cos(\theta) & ...\\
.. & ... & ... & ... & ... & I
\end{bmatrix}\mathbf{I}_p. \tag{3.22}$$

Dots ("...") stand for zeros. Because the eigenvalues of the matrix just above are $\pm i\pi$ and $\pm i\theta$, this rotation has a canonical length $l_c(\gamma) = \frac{\sqrt{2}}{2}\|\log(\mathbf{Q})\|_F = \sqrt{\pi^2+\theta^2} > \pi$. However, the same final transformation can be modelled at a cheaper canonical cost:

$$\widetilde{\mathbf{U}} = \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix} \begin{bmatrix} -1 & 0 & ... & 0 & 0 & .... \\ 0 & \cos(\theta) & ... & 0 & \sin(\theta) & ... \\ .. & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 1 & 0 & ... \\ 0 & \sin\theta & ... & 0 & -\cos(\theta) & ... \\ ... & ... & ... & ... & ... & I \end{bmatrix} \mathbf{I}_p. \tag{3.23}$$

The new transformation (3.23) has a canonical length π exactly, whatever θ. Indeed, the eigenvalues of a block $\begin{bmatrix}\cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta)\end{bmatrix}$ are -1 and 1. It is a reflector. The principal logarithms of these eigenvalues are $i\pi$ and 0 respectively. The two nested rotations of Equation (3.22) have been replaced by two cheaper reflections in Equation (3.23). Since the nesting of two reflections makes a rotation, the transformation belongs to $\mathrm{SO}(n)$. In consequence, once a column is fully rotated, another can thus be rotated as much as wanted without affecting the canonical distance. Moreover, we can notice that the skew-symmetric logarithm of the matrix of Equation (3.23) shows $\mathbf{0}_{(n-p)\times(n-p)}$ as a right lower block. Hence it is a geodesic.

What is the canonical distance if we rotate one more column? The second and third columns are rotated by angles θ_1 and θ_2 respectively. To be as general as possible, we consider k (with k odd) columns were already flipped previously, not only one. We know that the most expensive rotation can be done for "free" by encoding it as a reflection. The cost of the transformation is thus defined by $\min(|\theta_1|, |\theta_2|)$. Evaluating the cost of this transformation given the observation of Equation (3.23) allows to compute the canonical length of the curve γ:

$$l_c(\gamma) = \frac{\sqrt{2}}{2}\sqrt{(1+k)\pi^2 + 2\min(|\theta_1|, |\theta_2|)^2}. \tag{3.24}$$

On the other hand, we can compute the Frobenius distance induced by this transformation.

$$\begin{aligned}
\|\widetilde{\mathbf{U}} - \mathbf{U}\|_F^2 &= \sum_{j=1}^{p}\sum_{i=1}^{n}(\widetilde{\mathbf{U}}_{i,j} - \mathbf{U}_{i,j})^2 \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n} 4\mathbf{U}_{i,j}^2 + \sum_{i=1}^{n}((\cos(\theta_1)-1)\mathbf{U}_{i,k+1} - \sin(\theta_1)\mathbf{U}_{\perp,i,k+1})^2 \\
&\qquad + \sum_{i=1}^{n}(((\cos(\theta_2)-1)\mathbf{U}_{i,k+2} - \sin(\theta_2)\mathbf{U}_{\perp,i,k+2})^2 \\
&= 4k + 2(1-\cos(\theta_1)) + 2(1-\cos(\theta_2)).
\end{aligned}$$

Recall that we want to minimize the Frobenius distance at fixed canonical distance. We left θ_1 and θ_2 unconstrained to be general but we can now leverage the final objective to choose them. We want to solve

$$\begin{aligned}
\min_{\theta_1,\theta_2} \quad & -\cos(\theta_1) - \cos(\theta_2) \\
s.t \quad & \min(|\theta_1|, |\theta_2|) = \theta \text{ for some } 0 \le \theta \le \pi.
\end{aligned}$$

The constraint $\min(|\theta_1|, |\theta_2|) = \theta$ fixes the length $l_c(\gamma)$. The solution to this problem is $|\theta_1| = |\theta_2| = \theta$. To minimize the Frobenius distance, the two rotations must have the same amplitude.

The latter observation yields

$$
\begin{aligned}
&\|\widetilde{\mathbf{U}}-\mathbf{U}\|_{\mathrm{F}}^2 = 4(1+k-\cos(\theta)) \\
\iff\ &|\theta| = \arccos\left(1+k-\frac{\|\widetilde{\mathbf{U}}-\mathbf{U}\|_{\mathrm{F}}^2}{4}\right).
\end{aligned}
$$

Introducing Equation (3.24), we finally obtain

$$
\implies l_{\mathrm{c}}(\gamma) = \frac{\sqrt{2}}{2}\sqrt{(1+k)\pi^2 + 2\arccos^2\left(1+k-\frac{\|\widetilde{\mathbf{U}}-\mathbf{U}\|_{\mathrm{F}}^2}{4}\right)} \text{ for } k \text{ odd.} \tag{3.25}
$$

We conjecture that Equation (3.25) is the upper bound we are looking for and $d_{\mathrm{c}}(\mathbf{U},\widetilde{\mathbf{U}}) = l_{\mathrm{c}}(\gamma)$.

Conjecture 3.3. *For all $\mathbf{U},\widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$, $n \geq 2p$ and all positive integers $k \leq p-2$ with k odd, we have*

$$
d_{\mathrm{c}}(\mathbf{U},\widetilde{\mathbf{U}}) \leq \begin{cases} 2\arcsin\left(\frac{\|\widetilde{\mathbf{U}}-\mathbf{U}\|_{\mathrm{F}}}{2}\right) & \text{if } \|\widetilde{\mathbf{U}}-\mathbf{U}\|_{\mathrm{F}} \leq 2 \\ \frac{\sqrt{2}}{2}\sqrt{(1+k)\pi^2 + 2\arccos^2\left(1+k-\frac{\|\widetilde{\mathbf{U}}-\mathbf{U}\|_{\mathrm{F}}^2}{4}\right)} & \text{if } \|\widetilde{\mathbf{U}}-\mathbf{U}\|_{\mathrm{F}} \in 2\left[\sqrt{k},\sqrt{k+2}\right] \\ \frac{\sqrt{2p}}{2}\pi & \text{if } p \text{ even and } \|\widetilde{\mathbf{U}}-\mathbf{U}\|_{\mathrm{F}} \geq p-1 \end{cases}
$$

Conjecture 3.3 may look a bit odd but is nonetheless based on a detailed reasoning. As we will observe in next section, numerical experiments only confirm our bet.

3.13.6 Numerical corroboration of the conjecture

In this section, we evaluate the correctness of the Conjecture 3.3 by numerical experiments. Figure 3.13 shows that the matrices that we have defined to obtain Equation 3.25 achieve the expected canonical distance. more precisely, [48, Algorithm 4] recovers the same value. This observation corroborates that the curve described in section 3.13.5 is a minimizing geodesic. Figure 3.12 features random samples that have been generated using the method described in Appendix A. Indeed, special techniques need to be leveraged to defeat the *curse of dimensionality* to sample all possible Frobenius distances.

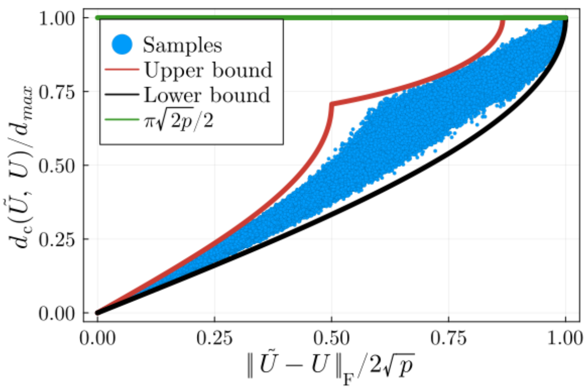

Figure 3.12: Distances computed on randomly generated matrices, $(n,p) = (8,4)$, $d_{max} := \pi\frac{\sqrt{2p}}{2}$.

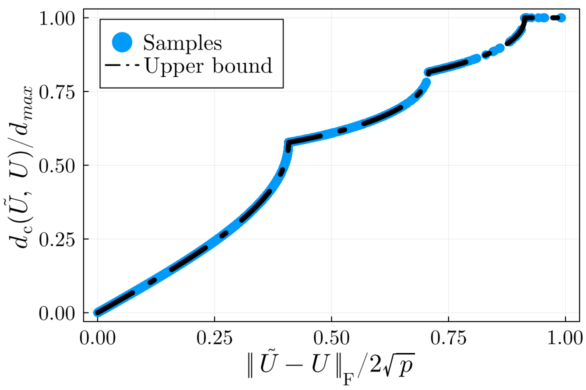

Figure 3.13: Distances computed on the matrices of section 3.13.5, $(n,p) = (12,6)$, $d_{max} := \pi\frac{\sqrt{2p}}{2}$.

3.13.7 A heuristic path on $\mathrm{St}(n,p)$ with $p \leq \frac{n}{2}$ using the SVD

In this section, we examine the performance of the SVD in providing an approximate solution to the geodesic endpoint problem. Theorem 3.10 illustrates that we can view the geodesic endpoint problem on $(\mathrm{St}(n,p), g_c)$ as the task of completing the bases $\mathbf{U}$ and $\widetilde{\mathbf{U}}$ in $\mathrm{SO}(n)$ in such a way that the resulting matrices are as close as possible to each other in $\mathrm{SO}(n)$. The theory of *canonical angles* (see [6, Theorem 1]) demonstrates that when given two bases $\mathbf{U}_\perp$ and $\widetilde{\mathbf{U}}_\perp$, the SVD provides two other bases $\mathbf{U}_\perp\mathbf{R}$ and $\widetilde{\mathbf{U}}_\perp\widetilde{\mathbf{R}}$ that span the same spaces, but with common subspaces being spanned by the same basis vectors in the representation $\mathbf{U}_\perp\mathbf{R}$ and $\widetilde{\mathbf{U}}_\perp\widetilde{\mathbf{R}}$. The solution we propose using the SVD is essentially the same as the initial iterate proposed by [47, Algorithm 3.1]. There, it is interpreted as the solution to an orthogonal Procrustes method based on [21, Theorem 2.6.3].

Lemma 3.11. *For all $\mathbf{U}_\perp, \widetilde{\mathbf{U}}_\perp \in \mathrm{St}(n, n-p)$ where $p \leq \frac{n}{2}$, let $\mathbf{U}_\perp^T\widetilde{\mathbf{U}}_\perp = \mathbf{R}\mathbf{\Sigma}\widetilde{\mathbf{R}}^T$ be the SVD. Then 1 is a singular value of multiplicity at least $n-2p$*

Proof. We rely on the theory of canonical angles [6, Theorem 1]. The multiplicity of the singular value 1 is equal to the dimension of the linear subspace defined by the intersection of the column spaces of $\mathbf{U}_\perp$ and $\widetilde{\mathbf{U}}_\perp$, written $\dim\left(\mathrm{Col}(\mathbf{U}_\perp) \cap \mathrm{Col}(\widetilde{\mathbf{U}}_\perp)\right)$. Since $\dim\left(\mathrm{Col}(\mathbf{U}_\perp)\right) = \dim\left(\mathrm{Col}(\widetilde{\mathbf{U}}_\perp)\right) = n-p$, we have

$$\begin{aligned}
&\dim\left(\mathrm{Col}(\mathbf{U}_\perp) \cup \mathrm{Col}(\widetilde{\mathbf{U}}_\perp)\right) = \dim\left(\mathrm{Col}(\mathbf{U}_\perp)\right) + \dim\left(\mathrm{Col}(\widetilde{\mathbf{U}}_\perp)\right) - \dim\left(\mathrm{Col}(\mathbf{U}_\perp) \cap \mathrm{Col}(\widetilde{\mathbf{U}}_\perp)\right) \\
&\implies 2(n-p) - \dim\left(\mathrm{Col}(\mathbf{U}_\perp) \cap \mathrm{Col}(\widetilde{\mathbf{U}}_\perp)\right) \leq n \\
&\implies \dim\left(\mathrm{Col}(\mathbf{U}_\perp) \cap \mathrm{Col}(\widetilde{\mathbf{U}}_\perp)\right) \geq n-2p.
\end{aligned}$$

The singular value 1 has thus a multiplicity at least $n-2p$. □

Theorem 3.13. *For all $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$ with $p \leq \frac{n}{2}$, we have $d_c(\mathbf{U}, \widetilde{\mathbf{U}}) \leq \pi\sqrt{p}$. Moreover, the SVD provides a curve of length less than $\pi\sqrt{p}$.*

Proof. Let $\mathbf{U}_\perp, \widetilde{\mathbf{U}}_\perp$ be orthogonal completions of $\mathbf{U}$ and $\widetilde{\mathbf{U}}$ respectively. Using Corollary 2.3 and Theorem 3.10, we get

$$d_c\left(\mathbf{U}, \widetilde{\mathbf{U}}\right) \leq d\left(\begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\mathbf{R}\end{bmatrix}, \begin{bmatrix}\widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp\widetilde{\mathbf{R}}\end{bmatrix}\right),$$

for all $\mathbf{R}, \widetilde{\mathbf{R}} \in \mathrm{SO}(n-p)$. We can build an orthogonal matrix $\mathbf{Q}$ using the SVD of $\mathbf{U}_\perp^T\widetilde{\mathbf{U}}_\perp = \mathbf{R}\mathbf{\Sigma}\widetilde{\mathbf{R}}^T$. Leveraging Lemma 3.11, we obtain

$$\mathbf{Q} := \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\mathbf{R}\end{bmatrix}^T \begin{bmatrix}\widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp\widetilde{\mathbf{R}}\end{bmatrix} = \begin{bmatrix} * & \mathbf{0} & * \\ \mathbf{0} & \mathbf{I}_{n-2p} & \mathbf{0} \\ * & \mathbf{0} & * \end{bmatrix}^{1},$$

where the $*$ stand for $p \times p$ non-zero matrices. The distance in $\mathrm{SO}(n)$ can be computed as $\frac{\sqrt{2}}{2}\|\log(\mathbf{Q})\|_{\mathrm{F}}$ by Theorem 2.2. Due to $\mathbf{I}_{n-2p}$ in the center of the matrix $\mathbf{Q}$, $\log(\mathbf{Q})$ has $n-2p$ zeros at the same location. Hence,

$$\log(\mathbf{Q}) = \begin{bmatrix} * & \mathbf{0} & * \\ \mathbf{0} & \mathbf{0}_{n-2p} & \mathbf{0} \\ * & \mathbf{0} & * \end{bmatrix} \implies d(\begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\mathbf{R}\end{bmatrix}, \begin{bmatrix}\widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp\widetilde{\mathbf{R}}\end{bmatrix}) = \frac{\sqrt{2}}{2}\|\log(\mathbf{Q})\|_{\mathrm{F}} \leq \pi\sqrt{p}.$$

[1] If $\mathbf{R}$ or $\widetilde{\mathbf{R}}$ is a reflection, we can transform them in rotations by multiplying the last column by -1.

The last inequality holds since $\|\log(\mathbf{Q})\|_\mathrm{F}^2 = \sum_{i=1}^{2p} |\lambda_i|^2$ where the λ_i are the eigenvalues of $\log(\mathbf{Q})$ and $|\lambda_i|$ is bounded from above by π. □

Theorem 3.13 is coherent with [15, Corollary 2.2]. The latter corollary states that a canonical geodesic can always be written using an orthogonal completion of $\mathbf{U}$ only made of p columns. We provide the pseudo-code to obtain the heuristic path provided by the SVD as built in Theorem 3.13.

Algorithm 4 Heuristic path on Stiefel using the SVD

Step 1: Build $\mathbf{U}_\perp, \widetilde{\mathbf{U}}_\perp \in \mathrm{St}(n, n-p)$ such that $\mathbf{Q} = \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}$, $\widetilde{\mathbf{Q}} = \begin{bmatrix}\widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp\end{bmatrix}$ belong to $\mathrm{SO}(n)$.
Step 2: Compute the SVD of $\mathbf{U}_\perp^T \widetilde{\mathbf{U}}_\perp = \mathbf{R}\boldsymbol{\Sigma}\widetilde{\mathbf{R}}^T$.
Step 3: Assign $\mathbf{U}_\perp \leftarrow \mathbf{U}_\perp \mathbf{R}$ and $\widetilde{\mathbf{U}}_\perp \leftarrow \widetilde{\mathbf{U}}_\perp \widetilde{\mathbf{R}}$. Enforce $\det(\mathbf{Q}) = \det(\widetilde{\mathbf{Q}}) = 1$
Step 4: Compute $\hat{\mathbf{Q}} = \mathbf{Q}^T \widetilde{\mathbf{Q}}$.
Step 5: Compute $\mathbf{A} = \log(\hat{\mathbf{Q}})$. Since $\hat{\mathbf{Q}} = \mathbf{V}e^{i\boldsymbol{\Theta}}\mathbf{V}^* \implies \mathbf{A} = \mathbf{V}i\boldsymbol{\Theta}\mathbf{V}^*$. Take $\boldsymbol{\Theta}_{ii} \in [-\pi, \pi]$.
return $\gamma_\mathbf{A} : [0, 1] \to \mathrm{St}(n, p) : t \to \mathbf{U}e^{\mathbf{A}t}\mathbf{I}_{n\times p}$

The heuristic path is very similar to the initialization of [47, Algorithm 3.1]. Instead of the initialization $\begin{bmatrix}\mathbf{U}^T\widetilde{\mathbf{U}} & \mathbf{U}^T\widetilde{\mathbf{U}}_\perp\widetilde{\mathbf{R}} \\ \mathbf{R}^T\mathbf{U}_\perp^T\widetilde{\mathbf{U}} & \boldsymbol{\Sigma}\end{bmatrix}$, they use $\begin{bmatrix}\mathbf{U}^T\widetilde{\mathbf{U}} & \mathbf{U}^T\widetilde{\mathbf{U}}_\perp\widetilde{\mathbf{R}}\mathbf{R}^T \\ \mathbf{U}_\perp^T\widetilde{\mathbf{U}} & \mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^T\end{bmatrix}$, which is equivalent up to a similarity transformation. Theorem 3.13 shows that Algorithm 4 provides a heuristic initialization with theoretical guarantees. Since Theorem 3.13 provides a worst-case upper bound, we can even expect the SVD to provide better results in general. We benchmark this heuristic solution with the best solution available: [48, Algorithm 4] to compute canonical geodesics. An error criterion must be defined to compare the solutions. Let θ^Z be the canonical length of the solution provided by Zimmerman's algorithm and θ^SVD be the length of the heuristic. We perform N experiments indexed by i. The criterion is

$$\text{Mean Relative Error(MRE)} := \frac{1}{N}\sum_{i=1}^{N}\frac{|\theta_i^\mathrm{SVD} - \theta_i^\mathrm{Z}|}{\theta_i^\mathrm{Z}}. \tag{3.26}$$

The matrices are generated following the results of Appendix A.

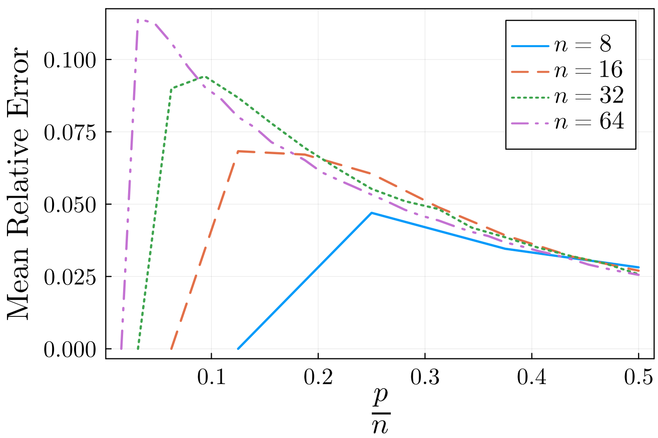

Figure 3.14: Mean Relative Error, $N = 500$.

The results of Figure 3.14 show that the heuristic solution is good, especially if p is close to

$\frac{n}{2}$. For $p = \frac{n}{2}$, the Mean Relative Error(MRE) reduces to less than 3%. Furthermore, we can notice that the worst case is when $n \gg p$: the larger is n compared to p, the larger is the MRE.

3.13.8 Computing an upper bound for the cut time

Gathering the results of previous sections, we expect that

$$d_c(\mathbf{U}, \widetilde{\mathbf{U}}) \leq h\left(\|\widetilde{\mathbf{U}} - \mathbf{U}\|_F\right), \tag{3.27}$$

for some continuous but non-smooth function $h : \mathbb{R} \to \mathbb{R}$. Given an initial velocity $\mathbf{\Delta} = \mathbf{U}\mathbf{A} + \mathbf{U}_\perp \mathbf{B} \in T_{\mathbf{U}}\mathrm{St}(n,p)$, the canonical distance after any sufficiently small time t is $d_c(\mathbf{U}, \widetilde{\mathbf{U}}(t)) = t\sqrt{\frac{1}{2}\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2} = t\|\mathbf{\Delta}\|_c$. An upper bound $\bar{t}$ on the cut time $\mathcal{T}_{\mathbf{U}}(\mathbf{\Delta})$ can thus be computed by solving the nonlinear equation

$$\bar{t}\|\mathbf{\Delta}\|_c = h\left(\|\mathrm{Exp}_{c,\mathbf{U}}\left(\bar{t}\mathbf{\Delta}\right) - \mathbf{U}\|_F\right). \tag{3.28}$$

The solution to Equation (3.28) can be computed using numerical methods. The formulation is here natural for the fixed point method (see e.g. [29, Definition 1.1]). If we restrict $\mathbf{\Delta}$ to 3 degrees of freedom, we can offer a visualization of the proposed method. Let us take for example $h = \min\left(\frac{\pi}{2}\|\mathrm{Exp}_{c,\,\mathbf{U}}\left(\bar{t}\mathbf{\Delta}\right) - \mathbf{U}\|_F, \pi\frac{\sqrt{2p}}{2}\right)$. Let $(n,p) = (4,2)$, $\mathbf{A} = \begin{bmatrix} 0 & -a \\ a & 0 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} b_1 & 0 \\ b_2 & 0 \end{bmatrix}$. Given any direction $(a, b_1, b_2) \in \mathbb{S}_1^3$ in the Euclidean space, Figure 3.15 represents the surface obtained when multiplying the triplet in each direction by $\bar{t}$, $\mathbb{S}_1^3 \to \mathbb{R}^3 : (a, b_1, b_2) \to \bar{t}(a, b_1, b_2)$. The result is a deformed sphere showing the directions where the bound on the cut time is reached at first.

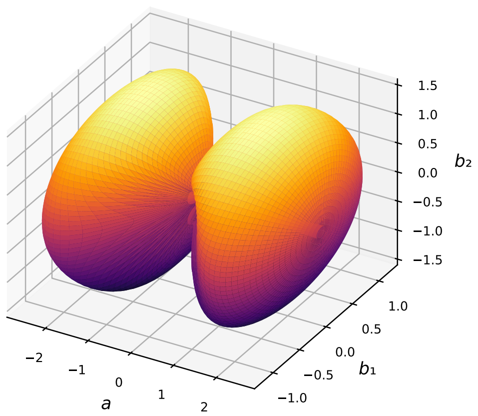

Figure 3.15: $\mathbb{S}_1^3 \to \mathbb{R}^3 : (a, b_1, b_2) \to \bar{t}(a, b_1, b_2)$

3.14 Conclusion

This chapter explored both theoretical and numerical aspects of the geodesic distances induced by the β-metrics. We first demonstrated that all these β-metrics were bilipschitz equivalent and obtained the best Lipschitz constants between them. In a second time, we have obtained lower and upper bounds on these β-distances by the easy-to-compute Frobenius distance. These bounds not only provided a geometric understanding of the Stiefel manifold but also offered simple conditions to check the output of algorithms that aim at computing logarithms. As a corollary, we discovered several families of matrices reaching these bounds. Moreover, we examined the geometric properties of the geodesics under the canonical metric, leveraging the eigenvalue decomposition of the skew-symmetric matrix characterizing the geodesic. Finally, we have highlighted the good performances of a heuristic approach to build a path between points on the Stiefel manifold.

Chapter 4

SkewLinearAlgebra.jl

Minimizing-geodesic computation algorithms on Stiefel make an intensive use of exponentials of skew-symmetric matrices. This is an example where we are interested in the fast computation of the matrix exponential. In the case of *symmetric* matrices, efficient routines are available in many languages. This is the case of `Python (Numpy)` [22], `MATLAB` [26], `Julia` [4], `C/C++` and `FORTRAN`. This is inherited from the `C/FORTRAN` libraries `BLAS` [7] and `LAPACK` [2].

Figure 4.1: Logo of `SkewLinearAlgebra.jl`

`BLAS` and `LAPACK` are imported in the previously cited languages/packages to perform efficient numerical linear algebra operations. However, `LAPACK` provides no specialized routines for *skew-symmetric* matrices. In consequence, the previous languages/packages provide no specialized methods for skew-symmetric matrices either. `SkewLinearAlgebra.jl` is a proposal to fill this unattended part of numerical linear algebra for eigenproblems. In 1976, [42] proposed a `FORTRAN` package with the same aims. This package was a bit forgotten due to the decline of `FORTRAN` and not updated since to the best of the author's knowledge. Besides, `SkewLinearAlgebra.jl` is not limited to real matrices, proposes a faster *blocked tridiagonalization* and many more options. The author developed the `SkewLinearAlgebra.jl` package at the Massachusetts Institute of Technology (MIT) under the supervision of Prof. Steven G. Johnson. It can be found at the address `https://github.com/JuliaLinearAlgebra/SkewLinearAlgebra.jl`.

4.1 Mimicking the symmetric problem

There are many ways to compute the matrix exponential. The paper [31] proposes at least 19 of them with various computational performances and numerical stability properties. In state-of-the-art linear algebra libraries such as `LinearAlgebra.jl` [4], the matrix exponential of *symmetric* matrices is implemented as follows. Let $\mathbf{A} \in \mathrm{Sym}(n)$. Compute the eigendecomposition $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, then return $\exp(\mathbf{A}) = \mathbf{V}\exp(\mathbf{\Lambda})\mathbf{V}^T$. We see that the symmetric exponential problem essentially reduces to the efficient computation of the eigendecomposition of $\mathbf{A}$. The `LAPACK` routine to obtain the eigendecomposition is `syevr` ("symmetric eigenvalue reduction"). The first step of `syevr` is to call `sytrd`, the routine performing a tridiagonal reduction $\mathbf{A} = \mathbf{Q}\mathbf{T}\mathbf{Q}^T$ where $\mathbf{Q} \in \mathrm{O}(n)$ and $\mathbf{T}$ is symmetric tridiagonal. Once the tridiagonalization is performed, various very efficient algorithms exist to obtain the eigendecomposition of $\mathbf{T}$. `syevr` implements a divide and conquer algorithm [12]. The praised QR algorithm [18] is another possible choice of comparable performance in single-threaded framework. It was not the selected algorithm in `LAPACK` due to its poor parallelizability. However, we will see that in the case of skew-symmetric matrices, QR iterations, or more precisely Francis's iterations [43], are particularly cheap in terms of computations.

4.2 The tridiagonalization: an Imitation Game©

As explained in section 4.1, the first step is to obtain the tridiagonalization of $\mathbf{A} \in \text{Skew}(n)$. The latter is a particular case of the so-called *Hessenberg reduction*. The Hessenberg reduction computes an almost triangular reduction of $\mathbf{A} = \mathbf{Q}\mathbf{H}\mathbf{Q}^T$ by the means of an orthogonal similarity transformations $\mathbf{Q} \in \text{O}(n)$. The first subdiagonal remains non-zero.

$$\mathbf{H} = \mathbf{Q}^T\mathbf{A}\mathbf{Q} = \begin{bmatrix} * & * & * & \dots & * & * & * \\ * & * & * & \dots & * & * & * \\ 0 & * & * & \dots & * & * & * \\ 0 & 0 & * & \dots & * & * & * \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & * & * \end{bmatrix}. \tag{4.1}$$

When $\mathbf{A}$ is skew-symmetric, it is clear that $\mathbf{H}$ is tridiagonal since the orthogonal similarity transformation $\mathbf{Q}$ preserves the skew-symmetry. Hence, we say $\mathbf{H} := \mathbf{T}$ and $\mathbf{A} = \mathbf{Q}\mathbf{T}\mathbf{Q}^T$. We will now see how this similarity transformation is computed in numerical linear algebra.

4.2.1 A naive skew-symmetric tridiagonalization

The earliest trace we could find of the specialized symmetric tridiagonalization is [45] (1962). The proposed way is to use Householder reflectors. A Householder reflector is an orthogonal matrix $\mathbf{Q} = \mathbf{I} - \tau\mathbf{v}\mathbf{v}^T \in \text{O}(n)$ such that for a chosen $\mathbf{x} \in \mathbb{R}^n$, we have

$$\mathbf{Q}\mathbf{x} = (\mathbf{I} - \tau\mathbf{v}\mathbf{v}^T)\mathbf{x} = \begin{bmatrix} \|\mathbf{x}\|_2 \\ 0 \\ \dots \\ 0 \end{bmatrix}. \tag{4.2}$$

The solution to (4.2) is to take $\mathbf{s} = \mathbf{x} + \|\mathbf{x}\|_2\text{sign}(\mathbf{x}_1)\mathbf{e}_1$, $\mathbf{v} = \frac{\mathbf{s}}{\mathbf{s}_1}$ and $\tau = \frac{\mathbf{s}_1}{\text{sign}(\mathbf{x}_1)\|\mathbf{x}\|_2}$. This transformation can be applied to tridiagonalize $\mathbf{A} \in \text{Skew}(n)$ by successive applications:

$$\hat{\mathbf{A}}_1 = (\mathbf{I} - \tau_1\mathbf{v}_1\mathbf{v}_1^T)\mathbf{A}(\mathbf{I} - \tau_1\mathbf{v}_1\mathbf{v}_1^T)^T = \begin{bmatrix} 0 & -\|\mathbf{a}_1\|_2 & 0 & \dots & 0 \\ \|\mathbf{a}_1\|_2 & & & & \\ 0 & & & & \\ \dots & & \tilde{\mathbf{A}}_1 & & \\ 0 & & & & \end{bmatrix},$$

$$\hat{\mathbf{A}}_2 = (\mathbf{I} - \tau_2\mathbf{v}_2\mathbf{v}_2^T)\hat{\mathbf{A}}_1(\mathbf{I} - \tau_2\mathbf{v}_2\mathbf{v}_2^T)^T = \begin{bmatrix} 0 & -\|\mathbf{a}_1\|_2 & 0 & 0 & \dots & 0 \\ \|\mathbf{a}_1\|_2 & 0 & -\|\tilde{\mathbf{a}}_2\|_2 & 0 & \dots & 0 \\ 0 & \|\tilde{\mathbf{a}}_2\|_2 & & & & \\ 0 & 0 & & & & \\ \dots & \dots & & \tilde{\mathbf{A}}_2 & & \\ 0 & 0 & & & & \end{bmatrix}, \quad \text{etc.}$$

For the symmetric case, it is shown in [45] that a step can be conveniently written as a symmetric update of the matrix $\hat{\mathbf{A}}_{i-1}$:

$$\begin{aligned} \hat{\mathbf{A}}_i &= \hat{\mathbf{A}}_{i-1} - \tau_i\mathbf{v}_i\mathbf{v}_i^T\hat{\mathbf{A}}_{i-1} - \tau_i\hat{\mathbf{A}}_{i-1}\mathbf{v}_i\mathbf{v}_i^T + \tau_i^2\mathbf{v}_i\mathbf{v}_i^T\hat{\mathbf{A}}_{i-1}\mathbf{v}_i\mathbf{v}_i^T \\ &= \hat{\mathbf{A}}_{i-1} - \mathbf{v}_i\mathbf{q}_i^T - \mathbf{q}_i\mathbf{v}_i^T, \end{aligned}$$

where $\mathbf{q}_i = \tau_i\hat{\mathbf{A}}_{i-1}\mathbf{v}_i - \frac{\tau_i^2}{2}\mathbf{v}_i^T\hat{\mathbf{A}}_{i-1}\mathbf{v}_i\mathbf{v}_i$. For the skew-symmetric case, we easily see that $\mathbf{v}_i^T\hat{\mathbf{A}}_{i-1}\mathbf{v}_i = 0$. Hence, we obtain skew-symmetric updates:

$$\hat{\mathbf{A}}_i = \hat{\mathbf{A}}_{i-1} + \mathbf{v}_i\mathbf{q}_i^T - \mathbf{q}_i\mathbf{v}_i^T, \tag{4.3}$$

where $\mathbf{q}_i = \tau_i\hat{\mathbf{A}}_{i-1}\mathbf{v}_i$. In terms of operation count, this procedure is the most efficient known and is very simple to implement. This was implemented under the name `TRIZD` in the previously proposed `FORTRAN` package [42]. According to [41, Lecture 26], the number of floating-point operations needed is $\frac{4n^3}{3}$. However, we can observe that it only features matrix-vector products that are called `BLAS-2` operations in numerical linear algebra. These operations are not efficient in terms of *flops* (floating-point operations per second). Only `BLAS-3` operations (matrix-matrix) reach the peak performance (highest achievable flops of the CPU). Therefore, the method we presented just above is not implemented as is in libraries like `LAPACK`, and `SkewLinearAlgebra.jl` by extension.

4.2.2 An efficient tridiagonalization

A major innovation for "reduction algorithms" such as tridiagonalization was the development of "blocked algorithms" that maximize the use of `BLAS-3` routines. `LAPACK`'s `sytrd` routine belongs to this family of algorithms. The scientific papers that allowed to go from `BLAS-2` to `BLAS-3` are [5] and [36]. These papers introduced a method to group successive Householder transformations called the "**WV** representation for products of Householder matrices". It was emphasized that it was possible to write

$$(\mathbf{I} - \tau_1\mathbf{v}_1\mathbf{v}_1^T)...(\mathbf{I} - \tau_l\mathbf{v}_l\mathbf{v}_l^T) = \mathbf{I} + \mathbf{W}\mathbf{V}^T, \tag{4.4}$$

for some well build $\mathbf{W}, \mathbf{V} \in \mathbb{R}^{n\times l}$. We incite the reader to consult [36] for more details on this construction. The consequence of this result is that if we knew in advance the Householder transformations we wanted to apply to tridiagonalize $\mathbf{A}$, we could do it using only `BLAS-3` operations thanks to Equation (4.4). The bottleneck is that we are only able to compute the reflector $\mathbf{v}_k$ to eliminate the kth column ($k < n$) once we have eliminated all the previous columns. The elimination of these columns can only be performed using `BLAS-2` operations. The trade-off between `BLAS-2` and `BLAS-3` operations gives birth to the "blocked algorithms". The method is the following. Take the $n_\mathrm{b} < n$ first columns of $\mathbf{A}$ (n_b is the *block size*). Reduce these n_b columns with `BLAS-2` operations but do not update any element of the $(n-n_\mathrm{b})\times(n-n_\mathrm{b})$ bottom right submatrix. Meanwhile build a **WV** representation with the Householder transformations computed. Once the n_b first columns are completely reduced, update the submatrix at once using the **WV** representation and `BLAS-3` routines. Repeat the operation on the submatrix $\mathbf{A}_{n_\mathrm{b}:n,n_\mathrm{b}:n}$ until the matrix is completely tridiagonalized.

The procedure described above is implemented in `sytrd` (or `hetrd` in the case of a complex Hermitian matrix). The code is memory-efficient in the sense that it maximizes cache reuse and minimizes the memory storage needed to tridiagonalize. The package `SkewLinearAlgebra.jl` implements `hessenberg!(A::SkewHermitian)`, an adapted version of `sytrd/hetrd` for skew-symmetric/skew-Hermitian matrices. However, an "unfair" advantage makes `sytrd` faster. As explained in [40], the blocked Hessenberg reduction features a remaining matrix-vector (m-v in short) product representing about 20% of the operations, but 70% of the running time (This highlights that `BLAS-3` performs the 80% remaining operations in 30% of the time only). `BLAS` proposes a routine called `symv` to perform a symmetric m-v product in half the time of the general m-v product `gemv`. `symv` has no equivalent skew-symmetric implementation at this day. Hence, `hessenberg!(A::SkewHermitian)` is compelled to use `gemv`. This is the only valuable difference creating a gap of performance between `LAPACK`'s symmetric tridiagonalization and the one implemented in `SkewLinearAlgebra.jl`. We stop the explanation of the tridiagonalization at this stage because it would take us too far from the purpose of this master's thesis. For the reader interested in high-performance scientific computing and detailed implementations, we encourage to visit the following web pages:

- `sytrd` implementation: https://netlib.org/lapack/explore-html/d3/db6/group__double_s_ycomputational_gaefcd0b153f8e0c36b510af4364a12cd2.html
- `hessenberg!(A::SkewHermitian)` implementation: https://github.com/JuliaLinearAlgebra/SkewLinearAlgebra.jl/blob/main/src/hessenberg.jl

4.3 The eigenproblem for tridiagonal matrices

We are now able to tridiagonalize $\mathbf{A} \in \mathrm{Skew}(n)$ as $\mathbf{A} = \mathbf{QTQ}^T$. We want to obtain the eigenvalue decomposition of $\mathbf{T} = \mathbf{V}i\mathbf{\Theta}\mathbf{V}^*$. Computing directly this decomposition would be very expensive since it would ask for complex arithmetic. Our aim is to produce an algorithm making only use of real arithmetic. The key here is to obtain the real Schur decomposition $\mathbf{T} = \widetilde{\mathbf{Q}}\mathbf{S}\widetilde{\mathbf{Q}}^T$. As explained in Theorem 1.3, it is quite straightforward to go from the real Schur decomposition to the eigendecomposition since the real Schur form $\mathbf{S}$ is a block diagonal matrix with blocks $\begin{bmatrix}0\end{bmatrix}$ or $\begin{bmatrix}0 & \theta\\ -\theta & 0\end{bmatrix}$ for some $\theta \in \mathbb{R}$. In `SkewLinearAlgebra.jl`, the algorithm chosen to compute the real Schur decomposition is the double shifted QR algorithm, also called *bulge chasing algorithm* or *Francis's algorithm* as a tribute to its discoverer. The historical papers introducing the QR algorithm and then its double-shifted version are [18] and [19]. The famous 2000 paper [14] ranked *Francis's algorithm* among the 10 most important algorithms of all times. Funny enough, according to the reviewing paper [43], Francis only learned in 2007 the importance of his contributions after Gene Golub's effort to retrieve his trace.

4.3.1 Francis's algorithm

In this section, we describe Francis's algorithm for general matrices. We explain further the significant advantage that skew-symmetry provides in its implementation. The first step is to present the QR algorithm with shifts, Algorithm 5, originally presented in [18].

Algorithm 5 QR algorithm with shifts

$\mathbf{H}_0 = \mathbf{Q}_0^*\mathbf{A}\mathbf{Q}_0$
for $k = 1, 2, ...$ **do**
  $\mathbf{Q}_k\mathbf{R}_k = \mathbf{H}_k - \mu_k\mathbf{I}$
  $\mathbf{H}_{k+1} = \mathbf{R}_k\mathbf{Q}_k + \mu_k\mathbf{I}$
end for

[18, Theorem 3] provides the original result about the convergence of $\mathbf{H}_k$ to the Schur form. It was then proved that proper shifts μ_k lead to the quadratic convergence of the diagonal elements of $\mathbf{H}_k$ to the eigenvalues of $\mathbf{A}$. The problem with Algorithm 5 is that it must deal with complex arithmetic to make appear the complex eigenvalues on the diagonal. The solution to this problem proposed in [19] is to group iterations two by two using conjugate shifts. Doing so, we converge to the real Schur form (real two by two diagonal blocks on the main diagonal), not the complex Schur form. This trick allows to keep real arithmetic. The new iteration takes the form

$$\mathbf{Q}_k\mathbf{R}_k = (\mathbf{H}_k - \mu_k\mathbf{I})(\mathbf{H}_k + \bar{\mu}_k\mathbf{I}) =: p(\mathbf{H}_k),$$
$$\mathbf{H}_{k+2} = \mathbf{Q}_k^T\mathbf{H}_k\mathbf{Q}_k.$$

The use of conjugate shifts has for consequence that $(\mathbf{H}_k - \mu_k\mathbf{I})(\mathbf{H}_k + \bar{\mu}_k\mathbf{I})$ is real. Indeed,

$$p(\mathbf{H}_k) := (\mathbf{H}_k - \mu_k\mathbf{I})(\mathbf{H}_k + \bar{\mu}_k\mathbf{I}) = \mathbf{H}_k^2 - 2\mathrm{Re}(\mu_k)\mathbf{H}_k + |\mu_k|^2\mathbf{I} \in \mathbb{R}^{n\times n}. \tag{4.5}$$

Used as is, the trick would be expensive as it would ask to compute $\mathbf{H}_k^2$. [19, Theorem 11], today known as the *Implicit Q theorem*, provides an astonishing result avoiding to compute $\mathbf{H}_k^2$. This theorem says that if $\mathbf{H}_k$ is taken such that its subdiagonal elements are positive, then $\mathbf{Q}_k$ is *uniquely* defined by the first column of $p(\mathbf{H}_k)$. Hence, it is only needed to compute the first column of $p(\mathbf{H}_k)$ and the Householder reflection eliminating this column, that we denote by $\mathbf{U}_1^T$. Applying the similarity transformation $\mathbf{U}_1$ to $\mathbf{H}_k$ leads to

$$\mathbf{U}_1^T\mathbf{H}_k\mathbf{U}_1 = \begin{bmatrix} * & * & * & * & * & \dots & * & * & * \\ * & * & * & * & * & \dots & * & * & * \\ * & * & * & * & * & \dots & * & * & * \\ * & * & * & * & * & \dots & * & * & * \\ 0 & 0 & 0 & * & * & \dots & * & * & * \\ 0 & 0 & 0 & 0 & * & \dots & * & * & * \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & * & * \end{bmatrix}. \tag{4.6}$$

Non-zero elements were introduced under the subdiagonal. Indeed, $\mathbf{U}_1^T$ was designed to eliminate the first column of $p(\mathbf{H}_k)$, not of $\mathbf{H}_k$. We call these subdiagonal non-zero elements the *bulge*. As $\mathbf{Q}_k$ is uniquely defined by $\mathbf{U}_1$ (Implicit Q theorem), all we have to do is to eliminate the bulge to retrieve a Hessenberg form. We call $\mathbf{U}_2^T$ the transformation that eliminates the first column of the bulge. Applying the similarity transformation yields

$$\mathbf{U}_2^T\mathbf{U}_1^T\mathbf{H}_k\mathbf{U}_1\mathbf{U}_2 = \begin{bmatrix} * & * & * & * & * & \dots & * & * & * \\ * & * & * & * & * & \dots & * & * & * \\ 0 & * & * & * & * & \dots & * & * & * \\ 0 & * & * & * & * & \dots & * & * & * \\ 0 & * & * & * & * & \dots & * & * & * \\ 0 & 0 & 0 & 0 & * & \dots & * & * & * \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & * & * \end{bmatrix}. \tag{4.7}$$

Observe that eliminating the first column of the bulge shifted this bulge to the right. Repeating this shifting operation $n-3$ more times leads to total elimination of the bulge, hence the name of *bulge chasing algorithm*. This process of chasing the bulge completely is known as *Francis's iteration*. $\mathbf{Q}_k$, however never computed, was implicitly applied on $\mathbf{H}_k$ by the iteration, hence the name of *Implicit Q theorem*. Finally, we obtain

$$\mathbf{Q}_k = \mathbf{U}_1\mathbf{U}_2 \dots \mathbf{U}_{n-1} \quad \text{and} \quad \mathbf{H}_{k+2} = \mathbf{Q}_k^T\mathbf{H}_k\mathbf{Q}_k. \tag{4.8}$$

4.3.2 The skew-symmetric case

In this section, we highlight the significant advantage that skew-symmetry provides to Francis's iteration. First, we recall that the Hessenberg form is here tridiagonal, we denote $\mathbf{H}_k := \mathbf{T}_k$. We take the "Wilkinson shifts". In the case of a tridiagonal skew-symmetric matrices, this shift is the last subdiagonal element of $\mathbf{T}_k$ times $\pm i$. The result on $p(\mathbf{T}_k)$ is

$$p(\mathbf{T}_k) = \mathbf{T}_k^2 + |\mu_k|^2\mathbf{I}. \tag{4.9}$$

The efficiency of Francis's iterations resides in three key observations that allow to create very cheap iterations. It turns out that these observations were already made in 1976 independently of me in the `FORTRAN` package [42].

Algorithm 6 Francis's algorithm

$\mathbf{H}_0 = \mathbf{Q}_0^T\mathbf{A}\mathbf{Q}_0$
for $k = 1, 2, ...$ **do**
  Compute $p(\mathbf{H}_k)\mathbf{e}_1$.
  Compute $\mathbf{U}_1^T$ to eliminate $p(\mathbf{H}_k)\mathbf{e}_1$.
  Initiate the bulge by performing $\mathbf{U}_1^T\mathbf{H}_k\mathbf{U}_1$.
  Compute and apply successively $\mathbf{U}_2, ..., \mathbf{U}_{n-1}$ to $\mathbf{H}_k$ to chase the bulge.
  The result $\mathbf{H}_{k+1} = \mathbf{Q}_k^T\mathbf{H}_k\mathbf{Q}_k$ is implicitly obtained.
end for

Observation 1: the initialized bulge is a scalar

Definition 4.1. *We define a* 3*-Givens rotation as a matrix* $\mathbf{G} = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix} \in \mathrm{SO}(3)$.

It is easy to verify that $p(\mathbf{T}_k)$ is a band-symmetric matrix with a band of two. Moreover, the first sub- and superdiagonal are zero. In consequence, the first column of $p(\mathbf{T}_k)$ features only two non-zero elements. The easiest orthogonal matrix to initialize the bulge is thus a 3-Givens rotations. Let us observe the result on $\mathbf{T}_0$:

$$\begin{bmatrix} c & . & s & . \\ . & 1 & . & . \\ -s & . & c & . \\ . & . & . & 1 \end{bmatrix} \begin{bmatrix} . & -t_1 & . & . \\ t_1 & . & -t_2 & . \\ . & t_2 & . & -t_3 \\ . & . & t_3 & . \end{bmatrix} \begin{bmatrix} c & . & -s & . \\ . & 1 & . & . \\ s & . & c & . \\ . & . & . & 1 \end{bmatrix} =$$
$$\begin{bmatrix} . & -ct_1 + st_2 & . & -st_3 \\ ct_1 - st_2 & . & -st_1 - ct_2 & . \\ . & st_1 + ct_2 & . & -ct_3 \\ st_3 & . & ct_3 & . \end{bmatrix}.$$

The bulge, that we denote by β, is reduced to a single number: $\beta = st_3$.

Observation 2: a chased scalar bulge remains scalar

Since the bulge β is a scalar, we only need a simple 3-Givens rotation to chase it. Let us observe the effect of this 3-Givens rotation on $\mathbf{T}_k$:

$$\begin{bmatrix} 1 & . & . & . & . \\ . & c & . & s & . \\ . & . & 1 & . & . \\ . & -s & . & c & . \\ . & . & . & . & 1 \end{bmatrix} \begin{bmatrix} . & -t_1 & . & -\beta & . \\ t_1 & . & -t_2 & . & . \\ . & t_2 & . & -t_3 & . \\ \beta & . & t_3 & . & -t_4 \\ . & . & . & t_4 & . \end{bmatrix} \begin{bmatrix} 1 & . & . & . & . \\ . & c & . & -s & . \\ . & . & 1 & . & . \\ . & s & . & c & . \\ . & . & . & . & 1 \end{bmatrix} =$$
$$\begin{bmatrix} . & -ct_1 - s\beta & . & . & . \\ ct_1 + s\beta & . & -ct_2 + st_3 & . & -st_4 \\ . & ct_2 - st_3 & . & -st_2 - ct_3 & . \\ . & . & st_2 + ct_3 & . & -ct_4 \\ . & st_4 & . & ct_4 & . \end{bmatrix}.$$

The scalar bulge remains scalar! The two previous observations are enough to obtain the eigenvalues of any skew-symmetric tridiagonal matrix at very low cost. $\mathbf{T}_k$ can be stored as a vector and β as a scalar. Performing Francis's iteration reduces to chase the bulge using the formulas obtained above.

Observation 3: $\mathbf{H}_k$ is a "chessboard matrix"

This last observation is only interesting if we want to retrieve the eigenvectors as well. We will observe that the eigenvectors can be assembled at "low" cost (we will precise what we mean by "low").

Definition 4.2. *A chessboard matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ is a matrix such that $\mathbf{M}_{i,j} = 0$ if $i + j$ is odd.*

If we want to assemble the eigenvectors, we have to store the 3-Givens rotations applied on $\mathbf{T}_k$. The first operation is to apply the initialization of the bulge on the identity $\mathbf{I}_n$. Let us notice that the complete Francis's iteration produces 3-Givens rotations, each of them being shifted from one index to the right compared to the previous one in order to "follow" the bulge. Here are the important observations:

- Applying a 3-Givens rotation on a chessboard matrix preserves the chessboard structure.

- Let us number the 3-Givens rotations by the index k. Given a initial chessboard matrix $\mathbf{M}$, we denote $\mathbf{G}_k$ these rotations augmented by appropriate identities such that $\mathbf{G}_k \in \mathrm{SO}(n)$. Francis's iteration creates the new transformation $\mathbf{M}\mathbf{G}_1\mathbf{G}_2\mathbf{G}_3...\mathbf{G}_{n-2}$. If k is odd, $\mathbf{G}_k$ only modifies the odd columns, and conversely if k is even. Let us define $\mathbf{M}^{\mathrm{o}} \in \mathrm{SO}(\lceil \frac{n}{2} \rceil)$ and $\mathbf{M}^{\mathrm{e}} \in \mathrm{SO}(\lfloor \frac{n}{2} \rfloor)$ such that for all i, j with $i + j$ even, we have

$$\mathbf{M}_{i,j} = \begin{cases} \mathbf{M}^{\mathrm{o}}_{\lceil \frac{i}{2} \rceil, \lceil \frac{j}{2} \rceil} & \text{if } i, j \text{ odd} \\ \mathbf{M}^{\mathrm{e}}_{\frac{i}{2}, \frac{j}{2}} & \text{if } i, j \text{ even} \end{cases}.$$

  The nice property is that odd- and even-indexed rotations $\mathbf{G}_k$ are decoupled:

$$[\mathbf{M}\mathbf{G}_1\mathbf{G}_2\mathbf{G}_3...\mathbf{G}_{n-2}]_{i,j} = \begin{cases} [\mathbf{M}^{\mathrm{o}}\mathbf{G}^{\mathrm{o}}_1\mathbf{G}^{\mathrm{o}}_3\mathbf{G}^{\mathrm{o}}_5...]_{\lceil \frac{i}{2} \rceil, \lceil \frac{j}{2} \rceil} & \text{if } i, j \text{ odd} \\ [\mathbf{M}^{\mathrm{e}}\mathbf{G}^{\mathrm{e}}_2\mathbf{G}^{\mathrm{e}}_4\mathbf{G}^{\mathrm{e}}_6...]_{\frac{i}{2}, \frac{j}{2}} & \text{if } i, j \text{ even} \end{cases}.$$

  All the operations to retrieve the eigenvectors can thus can performed on 2 $\frac{n}{2} \times \frac{n}{2}$ matrices instead of a $n \times n$ matrix. The memory needed during the iteration is divided by two. The number of operations to perform is also divided by two. A non-negligible advantage is that the spatial locality of the operations is increased, dividing also the number of access to the main memory by a factor 2. Finally, recall that a Given rotation only require $\mathcal{O}(6 \cdot \frac{n}{2})$ flop to be applied.

Once the real Schur form of the tridiagonal matrix $\mathbf{T}$ is obtained, the eigenvalue decomposition is straightforward to obtain from the method of Theorem 1.3. The algorithms `eigen(A)` and `eigvals(A)` applied on `A::SkewHermTridiagonal` or `A::SkewHermitian` available through `SkewLinearAlgebra.jl` implement Francis's algorithm enhanced with the 3 previous observations.

4.4 The matrix exponential

In this section, we derive an efficient formula to compute the matrix exponential of a skew-symmetric matrix $\mathbf{A}$ given the eigenvalue decomposition $\mathbf{A} = \mathbf{V}i\mathbf{\Theta}\mathbf{V}^*$. We pose $\mathbf{V} = \mathbf{V}_{\mathrm{Re}} + i\mathbf{V}_{\mathrm{Im}}$ with $\mathbf{V}_{\mathrm{Re}}, \mathbf{V}_{\mathrm{Im}} \in \mathbb{R}^{n\times n}$. Computing directly $\mathbf{V}e^{i\mathbf{\Theta}}\mathbf{V}^*$ would involve costly complex arithmetic. However, we know that $e^{\mathbf{A}}$ is real by definition. Therefore, we must have $\mathrm{Im}\left(\mathbf{V}e^{i\mathbf{\Theta}}\mathbf{V}^*\right) = 0$ and

$e^{\mathbf{A}} = \mathrm{Re}\left(\mathbf{V}e^{i\mathbf{\Theta}}\mathbf{V}^*\right)$. This yields

$$\begin{aligned}
e^{\mathbf{A}} &= \mathrm{Re}\left(\mathbf{V}e^{i\mathbf{\Theta}}\mathbf{V}^*\right) \\
&= \mathrm{Re}\Big((\mathbf{V}_{\mathrm{Re}} + i\mathbf{V}_{\mathrm{Im}})[\cos(\mathbf{\Theta}) + i\sin(\mathbf{\Theta})](\mathbf{V}_{\mathrm{Re}} + i\mathbf{V}_{\mathrm{Im}})^*\Big) \\
&= [\mathbf{V}_{\mathrm{Re}}\cos(\mathbf{\Theta}) - \mathbf{V}_{\mathrm{Im}}\sin(\mathbf{\Theta})]\mathbf{V}_{\mathrm{Re}}^T + [\mathbf{V}_{\mathrm{Re}}\sin(\mathbf{\Theta}) + \mathbf{V}_{\mathrm{Im}}\cos(\mathbf{\Theta})]\mathbf{V}_{\mathrm{Im}}^T
\end{aligned}$$

This trick allows to obtain the exponential from the eigendecomposition with only two $n \times n$ real matrix products instead of the four needed initially. This formula is implemented in `exp(A::SkewHermitian)`. In the same fashion, `SkewLinearAlgebra.jl` provides the set of efficiently computed trigonometric functions `sin`, `cos`, `cis`, `tan`, `cot`, `sinh`, `cosh`, `tanh`, `coth` and `log`.

4.5 Performance analysis

In this section, we compare the implementation of the `SkewLinearAlgebra.jl` package against the `LinearAlgebra.jl` package. The latter package implements direct calls to `LAPACK` for the benchmarked functions. Therefore, we obtain a comparison of performance with `LAPACK`. We compare the algorithms in single threaded mode. The matrices used in the test are generated using normal random variables through the function `randn`. The tests are performed in double precision (64 bits) on a Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz 2.21 GHz processor.

We provide a little guidance to the reader to go through Tables 4.1 and 4.2. The difference between the “General” and “Skew-symmetric” columns represents the improvement of performance provided by the `SkewLinearAlgebra.jl` package. The “Symmetric” column provides a standard reference to compare the performance a skew-symmetric eigensolver.

- “General”: Type `A::Matrix`, `A = randn(n, n)`.
- “Symmetric”: Type `A::Symmetric`, `A = Symmetric(randn(n, n))`.
- “Skew-symmetric”:
  Type `A::SkewHermitian` (new type), `A = skewhermitian(randn(n, n))`.
- “Symmetric tridiagonal”:
  Type `A::SymTridiagonal`, `A = SymTridiagonal(randn(n), randn(n-1))`[1].
- “Skew-symmetric tridiagonal”:
  Type `A::SkewHermTridiagonal` (new type), `A = SkewHermTridiagonal(randn(n-1))`.

In `Julia`, the symbol `!` after a function specifies that the method is performed in-place.

Size n	10		100		500		1000	
Scale	$\cdot 10^{-6}s$		$\cdot 10^{-4}s$		$\cdot 10^{-2}s$		$\cdot 10^{-1}s$	
Type	Sym	Skew	Sym	Skew	Sym	Skew	Sym	Skew
`eigvals!`	0.232	0.258	0.176	0.0045	0.371	0.506	0.149	0.199
`eigen!`	11.2	0.520	7.87	2.97	2.35	1.67	0.984	1.50
`exp`	/	4.81	/	4.72	/	3.17	/	3.13

Table 4.1: Benchmark on tridiagonal matrices

[1] It would have been preferable for the comparison to initialize `A = SymTridiagonal(zeros(n), randn(n-1))`. However, we highlighted that such matrices leaded to the failure of `LAPACK`'s `stegr` routine through `LAPACKException (22)`.

Size n	10			100			500			1000		
Scale	$\cdot 10^{-5}s$			$\cdot 10^{-4}s$			$\cdot 10^{-2}s$			$\cdot 10^{-1}s$		
Type	Gen	Sym	Skew	Gen	Sym	Skew	Gen	Sym	Skew	Gen	Sym	Skew
`hessenberg!`	0.258	0.233	0.421	2.94	1.81	1.89	1.93	0.833	1.18	1.73	0.583	1.18
`eigvals!`	1.42	0.524	0.632	29.1	3.41	3.66	13.1	1.13	1.55	7.02	0.683	1.19
`eigen!`	1.17	1.45	1.01	54.6	9.34	7.23	19.5	3.73	4.66	10.8	2.30	4.15
`exp`	0.732	1.52	1.12	8.73	10.2	8.67	8.29	4.56	6.07	7.49	2.82	5.05

Table 4.2: Benchmark on dense matrices

Table 4.1 shows that the tridiagonal skew-symmetric eigensolver features the same performances as the symmetric one. In particular, we can notice that it is faster to obtain the complete eigendecomposition when the matrices are of size n less than 500. This is a result of the "chessboard" structure of the Schur decomposition we observed. Table 4.2 shows the result on dense matrices. In all the cases, the skew-symmetric algorithms are faster than the general ones, asserting the utility of `SkewLinearAlgebra.jl`. The difference with the symmetric case essentially resides in the gap created during the tridiagonal reduction. For n of the order of 100, we even observe that the skew-symmetric algorithms are faster because the tridiagonal reduction is not yet dominating.

4.6 Conclusion

The `SkewLinearAlgebra.jl` package is an attempt to offer specialized tools for dense skew-symmetric matrices. These tools aim at providing performances close to the performances of the algorithms available in `LinearAlgebra.jl` for their symmetric counterparts. Famous algorithms such as tridiagonalization and the QR algorithm are specialized to take advantage of the skew-symmetry. We have seen that the tridiagonalization can be easily adapted and that the QR algorithm reduces to very cheap iterations in term of both computational cost and memory storage. Either for the computation of the eigenvalues or trigonometric functions, we cared to keep the arithmetic real and thus efficient. In the last section of this chapter, we witnessed that the code of `SkewLinearAlgebra.jl` outperformed the algorithms for general matrices and competed with the symmetric implementations.

SkewLinearAlgebra.jl

Conclusion

Since its popularization in the seminal paper [15], the interest for the Stiefel manifold has steadily increased due to the large variety of problems where it appears and can be used. From clinical applications [32] to neural networks [24] and data processing [39], research questions involving the Stiefel manifold are flourishing.

The goal of this master's thesis was to improve the current characterization of the cut locus on the Stiefel manifold. To begin with, this objective lead us to have more interest about a one-parameter family of Riemannian metrics, the β-metrics. In particular, we highlighted that all the β-metrics were bilipschitz equivalent and we provided the best lipschitz constants between them. Afterwards, we established relations with the easy-to-compute Frobenius distance. Our demonstrations were based on the fact that Stiefel can be seen as the subset of a hypersphere and our intuitions were based on simple visual concepts, true in low dimension, that we generalized to higher dimensions. These ideas gave us both a lower and an upper bound on the geodesic distance by the Frobenius distance. As an aside, it showed that the β-distances are bilipschitz equivalent to the Frobenius distance as well. These bounds allow to compute upper bounds on the *cut time* by solving a nonlinear equation. Finally, we exploited the properties of the canonical geodesics to obtain geometric properties.

Although this work provides new results on the geodesic distances on the Stiefel manifold, it does not solve the geodesic endpoint problem. Hence, many questions are left to be answered. Can better bounds on the cut time be obtained? Can we generalize the diameter under the β-metric to $\pi\sqrt{\beta p}$? Can we build better quasi-geodesic paths [3]? Can cheaper or more precise algorithms be developed? To this last question, we provided an improvement through the development of `SkewLinearAlgebra.jl`.

Geodesic curves on matrix manifolds of orthogonal matrices are usually expressed by using the matrix exponential of a skew-symmetric matrix. Popular programming languages do not provide standard specialized methods to leverage the skew-symmetric structure of these matrices. `SkewLinearAlgebra.jl` implements specialized methods that outperform classical algorithms and have performances close to the symmetric implementations. However, the use of the QR algorithm is not very appropriate for multi-threading. To enhance the package's performance in the context of parallel computing, divide-and-conquer methods could be developed, following the approach of the state-of-the-art routine `stemr` from `LAPACK` [2].

In conclusion, this work does not close the door to research on the geodesic endpoint problem on the Stiefel manifold. On the contrary, it raises many research questions that we hope will find answers in the future years.

Appendix A: The sampling of matrices on $\mathrm{St}(n,p)$

Sampling matrices over $\mathrm{St}(n,p)$ is an essential task in this master's thesis. In a first time, it is needed in order to get good intuitions and ideas of properties that could be proven. In a second time, it is needed to verify that the theoretical results obtained are respected by the minimizing geodesic computation algorithms available.

A quantitative analysis: Suppose $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathrm{St}(n,p)$ are two matrices generated by applying a Gram-Schmidt process on matrices such that each element follows a standard normal distribution ($\sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$). Hence, each element $\mathbf{U}_{i,j}$ and $\widetilde{\mathbf{U}}_{i,j}$ is a random variable, and so is every element $\alpha_{i,j} := \left(\mathbf{U}_{i,j} - \widetilde{\mathbf{U}}_{i,j}\right)^2$. Thus, $\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}^2 = \sum_{i,j} \alpha_{i,j}$ is a sum of random variables. If we make the simplifying assumption $\mathbf{U}_{i,j} \sim \mathcal{N}\left(0, \frac{1}{n}\right)$, we obtain $\mathbf{U}_{i,j} - \widetilde{\mathbf{U}}_{i,j} \sim \mathcal{N}\left(0, \frac{2}{n}\right)$ and $\alpha_{i,j} \sim \frac{2}{n}\chi_1^2$ (chi-squared distribution). Thus we have $\mu_\alpha = \frac{2}{n}$ and $\sigma_\alpha = \frac{2\sqrt{2}}{n}$. Using the Central Limit Theorem, for n and p large, we have

$$
\begin{aligned}
\sqrt{np}\left(\frac{\sum_{i,j} \alpha_{i,j}}{np} - \frac{2}{n}\right) &\to \mathcal{N}\left(0, 4\frac{2}{n^2}\right) \\
\iff \quad \|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}^2 = \sum_{i,j} \alpha_{i,j} &\to \mathcal{N}\left(2p, 4\frac{p2}{n}\right) \\
\iff \quad \frac{\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}^2}{4p} &\to \mathcal{N}\left(\frac{1}{2}, \frac{2}{4np}\right)
\end{aligned}
$$

Experiments show that the scalar factor 2 should not be present. Recall that $\mathbf{U}_{i,j} \sim \mathcal{N}\left(0, \frac{1}{n}\right)$ is a rough approximation of the reality. Anyway, the goal here is only to understand how the standard deviation of $\frac{\|\widetilde{\mathbf{U}} - \mathbf{U}\|_{\mathrm{F}}^2}{4p}$ evolves as n and p grow. From Figure 4.2, we see that our analysis catches the right behaviour: the standard deviation is proportional to $\frac{1}{\sqrt{np}}$.

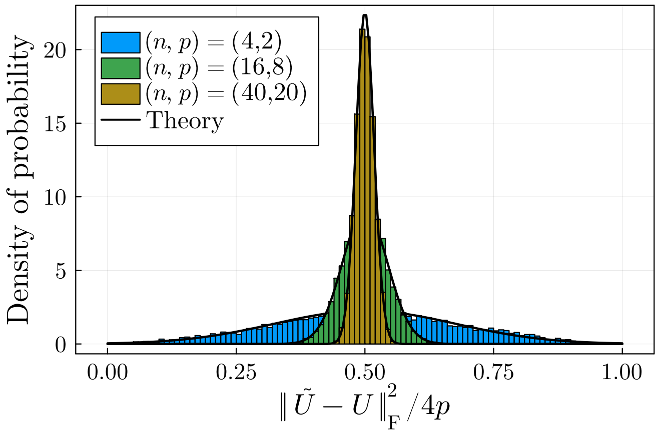

Figure 4.2: Approximated density probability to sample the Frobenius distance between random samples of $\mathrm{St}(n,p)$.

Sampling almost uniformly the Frobenius distances: For n and p are large, Figure 4.2 shows that it is hopeless to obtain variance in Frobenius distances by sampling $\mathbf{U}, \widetilde{\mathbf{U}}$ uniformly on $\mathrm{St}(n,p)$. Let $M \in \mathbb{R}^{n \times n}$ with $M_{i,j} \sim \mathcal{N}(0,1)$ for all $(i,j) \in \{1,\ldots,n\}^2$. The idea is to control the Frobenius distance by building

$$\widetilde{\mathbf{U}} = \pm \exp(\mathrm{skew}(M)\kappa)\mathbf{U}, \tag{4.10}$$

for $\kappa \in \mathbb{R}$. For κ sufficiently small, we can control the evolution of the Frobenius distance. Figure 4.3 shows the branches for the "+" and "−" case. We understand that controlling κ allows to sample almost uniformly the Frobenius distance. Let say we sample N pairs of matrices and we take $N_1 = \frac{\sqrt{2}}{2}N$ pairs for the "+" branch and $N_2 = (1 - \frac{\sqrt{2}}{2})N$ pairs for the "−" branch. The factor $\frac{\sqrt{2}}{2}$ is introduced so that each branch is proportionally equally sampled. We index the samples by i for each branch and we set arbitrarily $\kappa_i = \beta \frac{i}{N} \log(i)$ for $\beta > 0$, then we obtain the result of Figure 4.4. For our purpose, this result is sufficiently close to a uniform sampling.

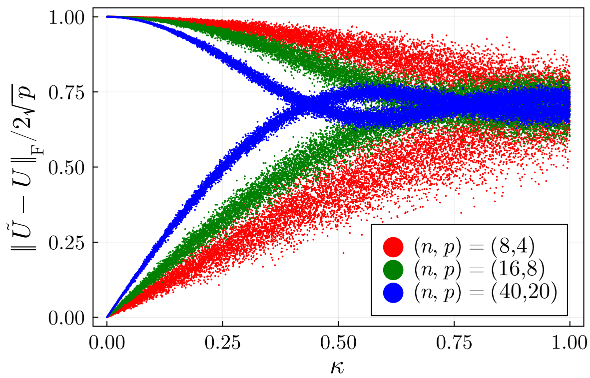

Figure 4.3: Frobenius distances of random samples computed by (4.10).

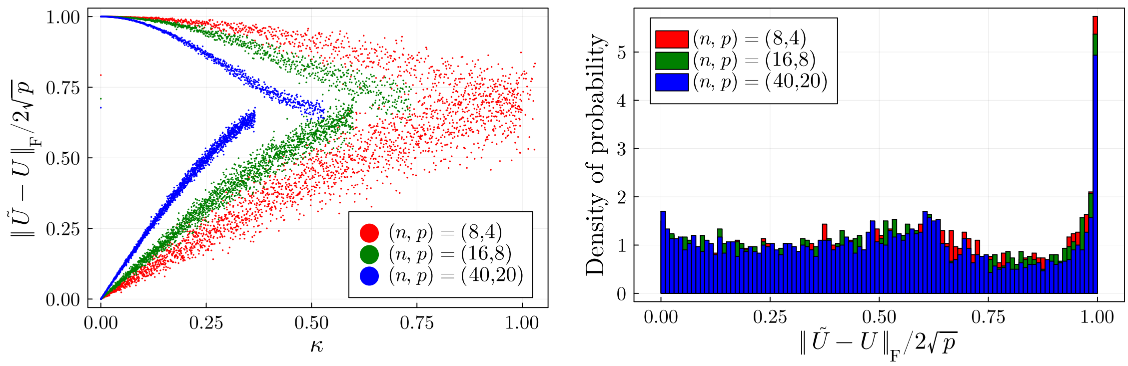

Figure 4.4: Frobenius distances of random samples computed by (4.10) with adapted $\{\kappa_i\}_N$.

The dual problem of sampling geodesic distances: We have seen that we could sample quite easily all possible Frobenius distances. However, at fixed Frobenius distance, many geodesic distances are possible. For the canonical metric typically, this geodesic distance is function of the eigenvalues of the rotation we apply on $\mathbf{U} \in \mathrm{St}(n, p)$. We have seen in this master's thesis that upper and lower bounds were reached when the eigenvalues of the rotation characterizing the geodesic achieved a high algebraic multiplicity. As shown in [37], the distribution of the eigenvalues of random skew-symmetric matrices repulses these eigenvalues from each other. This prevents high multiplicity. Due to this second curse, we observe on Figure 4.3 that, at fixed κ, the samples becomes very clustered as n and p grow. Therefore, the lower and upper bounds on the geodesic distance as a function of the Frobenius distance obtained in this master's thesis are particularly difficult to sample randomly. A way to tackle this repulsive behaviour of the eigenvalues of M (from Equation (4.10)) is to manually change the eigenvalues to a more favorable distribution, e.g. sampled as i.i.d uniform random variables between a maximum and a minimum value for example.

Bibliography

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[2] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. SIAM, Philadelphia, Pennsylvania, USA, third edition, 1999.

[3] Thomas Bendokat and Ralf Zimmermann. Efficient Quasi-Geodesics on the Stiefel manifold. In *Lecture Notes in Computer Science*, pages 763–771. Springer International Publishing, 2021.

[4] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

[5] Christian Bischof and Charles Van Loan. The WY Representation for Products of Householder Matrices. *Society for Industrial and Applied Mathematics.SIAM Journal on Scientific and Statistical Computing*, 8(1):1, 01 1987. Copyright - Copyright] © 1987 Society for Industrial and Applied Mathematics.

[6] Ake Bjoerck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. In *Milestones in Matrix Computation*, 1971.

[7] L Susan Blackford, Antoine Petitet, Roldan Pozo, Karin Remington, R Clint Whaley, James Demmel, Jack Dongarra, Iain Duff, Sven Hammarling, Greg Henry, et al. An updated set of Basic Linear Algebra Subprograms (BLAS). *ACM Transactions on Mathematical Software*, 28(2):135–151, 2002.

[8] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(42):1455–1459, 2014.

[9] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

[10] Darshan Bryner. Endpoint Geodesics on the Stiefel Manifold Embedded in Euclidean space. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1139–1159, 2017.

[11] Rudrasis Chakraborty and Baba C. Vemuri. Statistics on the Stiefel manifold: Theory and applications. *The Annals of Statistics*, 47(1):415 – 438, 2019.

[12] Inderjit S. Dhillon and Beresford N. Parlett. Multiple representations to compute orthogonal eigenvectors of symmetric tridiagonal matrices. *Linear Algebra and its Applications*, 387:1–28, 2004.

[13] Alberto Dolcetti and Donato Pertici. Skew symmetric logarithms and geodesics on On(r). *Advances in Geometry*, 18(4):495–507, 2018.

[14] J. Dongarra and F. Sullivan. Guest editors introduction to the top 10 algorithms. *Computing in Science and Engineering*, 2(1):22–23, 2000.

[15] Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[16] Lars Eldén and Haesun Park. A Procrustes problem on the Stiefel manifold. *Numerische Mathematik*, 82:599–619, 1999.

[17] Enrico Facca, Luca Berti, Francesco Fassó, and Mario Putti. Computing the Cut Locus of a Riemannian Manifold via Optimal Transport. 2021.

[18] J. G. F. Francis. The QR Transformation a Unitary Analogue to the LR Transformation - part 1. *Comput. J.*, 4:265–271, 1961.

[19] J. G. F. Francis. The QR transformation—part 2. *The Computer Journal*, 4(4):332–345, 01 1962.

[20] Bin Gao, Simon Vary, Pierre Ablin, and P.-A. Absil. Optimization flows landing on the Stiefel manifold. *IFAC-PapersOnLine*, 55(30):25–30, 2022. 25th IFAC Symposium on Mathematical Theory of Networks and Systems MTNS 2022.

[21] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, third edition, 1996.

[22] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[23] Roger A. Horn and Charles R. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.

[24] Lei Huang, Xianglong Liu, Bo Lang, Adams Yu, Yongliang Wang, and Bo Li. Orthogonal Weight Normalization: Solution to Optimization over multiple dependent Stiefel Manifolds in Deep Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[25] Knut Hüper and Florian Ullrich. Real Stiefel manifolds: An extrinsic point of view. pages 13–18, 2018.

[26] The MathWorks Inc. MATLAB version: 9.13.0 (r2022b), 2022.

[27] Hüper Knut, Markina Irina, and Leite Fátima Silva. A Lagrangian approach to extremal curves on Stiefel manifolds. *Journal of Geometric Mechanics*, 13(1):55–72, 2021.

[28] Yui Man Lui, J. Ross Beveridge, and Michael Kirby. Canonical Stiefel Quotient and its application to generic face recognition in illumination spaces. In *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–8, 2009.

[29] David Mayers and Endre Süli. *An introduction to numerical analysis.* Cambridge Univ. Press, Cambridge, 2003.

[30] Nina Miolane, Nicolas Guigui, Alice Le Brigant, Johan Mathe, Benjamin Hou, Yann Thanwerdas, Stefan Heyder, Olivier Peltre, Niklas Koep, Hadi Zaatiti, Hatem Hajri, Yann Cabanes, Thomas Gerald, Paul Chauchat, Christian Shewmake, Daniel Brooks, Bernhard Kainz, Claire Donnat, Susan Holmes, and Xavier Pennec. Geomstats: A Python Package for Riemannian Geometry in Machine Learning. *Journal of Machine Learning Research*, 21(223):1–9, 2020.

[31] Cleve Moler and Charles Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *Society for Industrial and Applied Mathematics*, 45:3–49, 03 2003.

[32] Clément Pealat, Guillaume Bouleux, and Vincent Cheutet. Extracting most impacting emergency department patient flow by embedding laboratory-confirmed and clinical diagnosis on the Stiefel manifold. In *2019 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4, 2019.

[33] Henri Poincare. Sur les lignes geodesiques des surfaces convexes. *Transactions of the American Mathematical Society*, 6(3):237–274, 1905.

[34] Rentmeesters Quentin. *Algorithms for data fitting on some common homogeneous spaces.* PhD thesis, Catholic University of Louvain (UCLouvain), 2013.

[35] T. Sakai. *Riemannian Geometry.* Fields Institute Communications. American Mathematical Society, 1996.

[36] Robert Schreiber and Charles Van Loan. A Storage-Efficient WY Representation for Products of Householder Transformations. *SIAM Journal on Scientific and Statistical Computing*, 10, 02 1989.

[37] Philippe Sosoe and Uzy Smilansky. On the spectrum of random anti-symmetric and tournament matrices. *Random Matrices: Theory and Applications*, 05(03):1650010, jul 2016.

[38] G. W. Stewart. Block Gram–Schmidt Orthogonalization. *SIAM Journal on Scientific Computing*, 31(1):761–775, 2008.

[39] Fabian Theis, Thomas Cason, and P Absil. Soft dimension reduction for ICA by Joint Diagonalization on the Stiefel manifold. pages 354–361, 03 2009.

[40] Stanimire Tomov, Rajib Nath, and Jack Dongarra. Accelerating the reduction to upper Hessenberg, tridiagonal, and bidiagonal forms through hybrid GPU-based computing. *Parallel Computing*, 36(12):645–654, 2010.

[41] Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra.* SIAM, 1997.

[42] R C Ward and L J Gray. Eigensystem computation for skew-symmetric matrices and a class of symmetric matrices. [subroutines trizd, imzd, and tbakzd in Fortran for IBM 360/91 computer]. 5 1976.

[43] David S. Watkins. Francis's algorithm. *The American Mathematical Monthly*, 118(5):pp. 387–403, 2011.

[44] J. H. C. Whitehead. On the covering of a complete space by the geodesics through a point. *Annals of Mathematics*, 36(3):679–704, 1935.

[45] J.H. Wilkinson. Householder's method for symmetric matrices. (4):354–361, 1962.

[46] Ming Yin, Junbin Gao, and Yi Guo. Nonlinear low-rank representation on Stiefel manifolds. *Electronics Letters*, 51(10):749–751, 2015.

[47] Ralf Zimmermann. A Matrix-Algebraic Algorithm for the Riemannian Logarithm on the Stiefel Manifold under the Canonical Metric. *SIAM Journal on Matrix Analysis and Applications*, 38(2):322–342, 2017.

[48] Ralf Zimmermann and Knut Hüper. Computing the Riemannian Logarithm on the Stiefel Manifold: Metrics, Methods, and Performance. *SIAM Journal on Matrix Analysis and Applications*, 43(2):953–980, 2022.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**