THE UNIVERSITY OF CHICAGO

GEOMETRIC METHODS IN STATISTICS AND OPTIMIZATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY

SZE WAI WONG

CHICAGO, ILLINOIS

JUNE 2018

To my family

"A large part of mathematics which becomes useful developed with absolutely no desire to be useful, and in a situation where nobody could possibly know in what area it would become useful; and there were no general indications that it ever would be so."

John von Neumann

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I am grateful and fortunate to work with my dissertation advisor Lek-Heng Lim. Lek-Heng is an encouraging, enthusiastic, thoughtful and friendly advisor. His countless support and encouragement are an indispensable part of my graduate studies. Not only have I learned much from him intellectually, I have also grown much interpersonally and become a more outgoing and considerate person.

I extend my sincere gratitude to the other dissertation committee members, Rina Roygel Barber and Yali Amit, for their time and comments on this work.

I also like to thank for the discussions, typically academic and in restaurant setting, with supportive friends - Somak Dutta, Jessica Foo, Marc Goessling, Wooseok Ha, Clare Huang, Danny Ng, Vivak Patel, Ulises Pereira, Yoann Potiron, Yang Qi, Jose Rodriguz, Si Tang, Nathan Vos, Wei Wang, Liwen Zhang, and especially Ke Ye for the numerous days drawing on whiteboard together.

ABSTRACT

Statistical estimation problems in multivariate analysis and machine learning often seek linear relations among variables. This translates to finding an affine subspace from the sample data set that, in an appropriate sense, either best represents the data set or best separates it into components. In other words, statistical estimation problems are optimization problems on the affine Grassmannian, a noncompact smooth manifold that parameterizes all affine subspaces of a fixed dimension. The affine Grassmannian is a natural generalization of Euclidean space, points being 0-dimensional affine subspaces. The main objective of the first part of this work is to show that, like the Euclidean space, the affine Grassmannian can serve as a concrete computational platform for data analytic problems — points on the affine Grassmannian can be concretely represented and readily manipulated; distances, metrics, probability densities, geodesics, exponential maps, parallel transports, etc, all have closed-form expressions that can be easily calculated; and optimization algorithms, including steepest descent, Newton, conjugate gradient, have efficient affine Grassmannian analogues that use only standard numerical linear algebra.

We then extend the framework to a nest of linear subspaces, that represent the variables in different regimes. Diving into the multi-scale representation of the data revealed by these problems requires a systematic study of nest of linear subspaces, which form a compact Riemannian manifold called the flag manifold. The main goal of this work is to show that flag manifold can be represented by matrix groups concisely and computed easily, and optimization on flag manifold can be performed with matrix operations, which bridges the gap between algebra and geometry.

Lastly, we study the Yates's algorithm that was first proposed to exploit the structure of full factorial designed experiment to obtain least squares estimates for factor effects for all factors and their relevant interactions. In short it is an organized way to do iterative summation which avoids repeated computation. Many well-known algorithms including Fast Fourier transform and Walsh transform turned out to be special cases of Yates's method.

Here we show that Yates's algorithm is optimal in the sense of a contraction of tensors but may be improved when considered from the perspective of bilinear complexity. We also show that it is a projection of a tensor network and point out its relationship with tensor train and tree tensor network.

CHAPTER 1

INTRODUCTION

Geometry of data has a deep root in the history of statistics [6, 7, 8, 38, 40]. Information geometry has been proposed to study family of probability distributions through Riemannian metric – the Fisher information metric, with each probability distribution being a point on the manifold. In later days manifold learning was proposed as a nonlinear dimensionality reduction technique in unsupervised learning when the underlying data set lies on a manifold approximately. These manifold structures typically arise in image and video sets [74, 14, 76, 48, 23, 24, 25, 22, 28, 35, 36, 37, 42, 44, 45, 46, 49, 71, 73, 72, 55]. In this dissertation we are more interested in exploring some more fundamental geometric structures - affine subspaces and nest of linear subspaces, through the lens of linear algebra, differential geometry and algebraic geometry.

In statistics, we often seek a first-order relationship among variables which can be computed efficiently and interpreted effectively, whilst in continuous optimization we rely on linear methods for finding the next iterates. To achieve numerical efficiency we often turn to linear algebra as subroutines in the actual optimization algorithms. In this dissertation we show that although differential geometry involves nonlinear structures and nonlinear calculation, we can still develop machinery with the help of linear algebra to describe the problems and solve them efficiently.

We will begin by defining the two manifolds, the affine Grassmannian and the flag manifold, that parametrize affine subspaces and nest of linear subspaces respectively.

A k-dimensional affine subspace of $\mathbb{R}^n$, denoted $\mathbf{A} + b$, is a k-dimensional linear subspace $\mathbf{A} \subseteq \mathbb{R}^n$ translated by a displacement vector $b \in \mathbb{R}^n$. The set of all k-dimensional affine subspaces in $\mathbb{R}^n$ forms a smooth manifold called the *affine Grassmannian*, denoted $\mathrm{Graff}(k, n)$, an analogue of the usual Grassmannian $\mathrm{Gr}(k, n)$ that parameterizes k-dimensional linear subspaces in $\mathbb{R}^n$.

A flag in a finite-dimensional vector space $\mathbf{V}$ over $\mathbb{R}$ is an increasing sequence of linear

subspaces $\{\mathbf{V}_i\}_{i=1}^d$ of $\mathbf{V}$, satisfying $\mathbf{V}_1 \subsetneq \cdots \subsetneq \mathbf{V}_d \subset \mathbf{V}$. All flags $\{\mathbf{V}_i\}_{i=1}^d$ with fixed dimensions, i.e. $\dim(\mathbf{V}_i) = n_i, i = 1 \cdots d$, form a smooth manifold $\mathrm{Flag}(n_1, \cdots, n_d; \mathbf{V})$, that we name it *flag manifold*. This is a generalization of Grassmannian $\mathrm{Gr}(k, n)$ that parameterizes k-dimensional linear subspaces in $\mathbb{R}^n$, which is equivalent to $\mathrm{Flag}(k; \mathbb{R}^n)$.

The main impetus for the first part of this dissertation is the observation that many statistical estimation problems (see Examples 2.1.1–2.1.4, 3.1.1–3.1.3) involve a search for linear relations or hierarchies of linear relations among variables and may ultimately be formulated as a problem of finding one (or more) linear or affine subspace that either best represents a given data set or best separates two (or more) components of the data set. More precisely, the problem is one of optimization on the affine Grassmannian or flag manifold.

Parameterizing a data set by geometric structures has become a popular alternative to probabilistic modeling, particularly when the intrinsic dimension of the data set is low or when it satisfies obvious geometric constraints. The two most common geometric structures employed are (i) a mixture of affine spaces [24, 35, 37] and (ii) a manifold, which invariably reduces to (i) since in this context manifolds are by-and-large regarded as collections of tangent spaces [49].[1] This provides a third impetus for studying the geometric object that parameterizes all affine spaces of a fixed dimension. In fact, with mixtures of affine subspaces of *different* dimensions in mind, we will introduce the doubly infinite affine Grassmannian (Section 2.5), which parameterizes affine spaces of all dimensions.

Lastly, we analyze the complexity of Yates's algorithm from the perspective of tensor contraction and matrix multiplication. In full factorial designed experiments, Yates's algorithm was proposed to efficiently obtain least squares estimates for factor effects for all factors and their relevant interactions. In short, it does index contractions systematically to avoid repeated calculations in iterative summations. Many widely used algorithms including Fast Fourier transform and Walsh transform are special cases of Yates's algorithm. We show

1. The original manifold learning techniques ISOMAP [47], LLE [43], and Laplacian Eigenmap [10] are essentially different ways to approximate a manifold by a collection of its tangent spaces.

that Yates's algorithm is optimal in the sense of a contraction of tensors and provide an upper bound of its complexity. We also show that it is a projection of a tensor network and illustrates its relationship with tensor train and tree tensor network.

CHAPTER 2

STATISTICAL ESTIMATION AND AFFINE GRASSMANNIAN

2.1 Introduction

A k-dimensional affine subspace of $\mathbb{R}^n$, denoted $\mathbf{A} + b$, is a k-dimensional linear subspace $\mathbf{A} \subseteq \mathbb{R}^n$ translated by a displacement vector $b \in \mathbb{R}^n$. The set of all k-dimensional affine subspaces in $\mathbb{R}^n$ forms a smooth manifold called the *affine Grassmannian*, denoted $\mathrm{Graff}(k, n)$, an analogue of the usual Grassmannian $\mathrm{Gr}(k, n)$ that parameterizes k-dimensional linear subspaces in $\mathbb{R}^n$.

The affine Grassmannian is a relatively obscured object, especially when compared to its ubiquitous cousin, the Grassmannian. Nevertheless, $\mathrm{Graff}(k, n)$, which like $\mathbb{R}^n$, is a non-compact manifold that naturally generalizes Euclidean space (points are 0-dimensional affine subspaces and so $\mathrm{Graff}(0, n) = \mathbb{R}^n$). The non-compactness makes $\mathrm{Graff}(k, n)$ harder to study than $\mathrm{Gr}(k, n)$, which is compact. Our main objective is to develop the foundations for working with the affine Grassmannian, particularly distances (Sections 2.4 and 2.5), probability distributions (Section 2.6), and optimization algorithms (Section 2.7), with a view towards statistical estimation problems.

Our study of the affine Grassmannian will differ substantially from traditional studies of differential geometry in statistics, a topic with a long history [6, 7, 8, 38, 40]. We emphasize three key differences: (a) We do not view our manifold in an abstract fashion comprising charts glued together; instead we emphasize the use of global coordinates for efficient computations. (b) Our manifold arises not as the parameter space of a family of probability distributions but as a concrete computational platform (like $\mathbb{R}^n$) on which distances, metrics, probability densities, geodesics, exponential maps, parallel transports, optimization algorithms, etc, may all be efficiently computed using standard numerical linear algebra. (c) Algebraic geometry will play a role as important as differential geometry in our study.

The point (a) deserves special elaboration. A main reason for the widespread applicabil-

ity of the Grassmannian is the existence of several excellent choices of global coordinates, allowing subspaces to be represented as matrices and therefore the use of a vast range of algorithms in numerical linear algebra [1, 2, 3, 15]. Such concrete realizations of an abstract manifold is essential for application purposes. We will show that this also holds for the affine Grassmannian (Section 2.3), which leads to our next impetus: by providing a corresponding set of tools for the affine Grassmannian, we effectively extend the wide range of data analytic techniques that uses the Grassmannian as a model for linear subspaces [74, 14, 76, 48, 23, 24, 25, 22, 28, 35, 36, 37, 42, 44, 45, 46, 49, 71, 73, 72, 55] to affine subspaces.

We will begin by seeing how classical multivariate analysis and machine learning techniques may be cast as affine subspace-searching problems, i.e., constrained or unconstrained optimization problems on the affine Grassmannian.

Example 2.1.1 (Linear Regression)**.** Consider a linear regression problem with $X \in \mathbb{R}^{n\times p}$, a design matrix of explanatory variables, and $y \in \mathbb{R}^n$, a vector of response variables. Let $\mathbb{1} = [1, \dots, 1]^\mathsf{T} \in \mathbb{R}^n$ and $e_{p+1} = [0, \dots, 0, 1]^\mathsf{T} \in \mathbb{R}^{p+1}$. Set $\widetilde{X} = [X, \mathbb{1}] \in \mathbb{R}^{n\times(p+1)}$ and define the affine subspace $\left\{ \begin{bmatrix} z \\ \beta^\mathsf{T} z \end{bmatrix} \in \mathbb{R}^{p+1} : z \in \mathbb{R}^p \right\} + \beta_{p+1} e_{p+1}$, chosen so that $\widetilde{\beta} = [\beta, \beta_{p+1}]^\mathsf{T} \in \mathbb{R}^{n+1}$ minimizes the sum of squared residuals $\|\widetilde{X}\widetilde{\beta} - y\|^2$. Then $\beta \in \mathbb{R}^n$ is the vector of regression coefficients. The affine subspace may be written as $\operatorname{span}\left(\begin{bmatrix} I_p \\ \beta^\mathsf{T} \end{bmatrix} \right) + \beta_{p+1} e_{p+1}$ where I_p is the $p \times p$ identity matrix. It best represents the data (X, y) in the sense of linear regression. This description corresponds to how one usually pictures linear regression — drawing an affine hyperplane through a collection of n scattered data points $(x_i, y_i)^\mathsf{T} \in \mathbb{R}^p \times \mathbb{R} = \mathbb{R}^{p+1}$, where x_i is the ith row of X and y_i is the ith entry of y, $i = 1, \dots, n$.

Example 2.1.2 (Errors-in-Variables Regression)**.** In this example, we use the same notations as above. We concatenate the explanatory variables and response variable and assign them equal weights. The best-fitting affine subspace of the data set $\{(x_i, y_i)^\mathsf{T} \in \mathbb{R}^{p+1} : x_i \in \mathbb{R}^p,\ y_i \in \mathbb{R},\ i = 1, \dots, n\}$ in this case is given by $\operatorname{span}(w) + b$ where $w, b \in \mathbb{R}^{p+1}$ are the minimizer of the loss function $\sum_{i=1}^n \|(I - ww^\mathsf{T})((x_i, y_i)^\mathsf{T} - b)\|_F^2$ subject to $w^\mathsf{T} b = 0$, and

may be obtained by solving a total least squares problem.

Example 2.1.3 (Principal Component Analysis)**.** Let $\overline{x} = \frac{1}{n} X^{\mathsf{T}} \mathbb{1} \in \mathbb{R}^p$ be the sample mean of a data matrix $X \in \mathbb{R}^{n \times p}$ so that $\overline{X} = X - \mathbb{1}\overline{x}^{\mathsf{T}}$ is mean-centered. For $k \leq p$, the kth principal subspace is $\operatorname{span}(Z_k)$, a k-dimensional linear subspace of $\mathbb{R}^p$ such that $Z_k \in \mathbb{R}^{p \times k}$ maximizes $\operatorname{tr}(Z_k^{\mathsf{T}} \overline{X}^{\mathsf{T}} \overline{X} Z_k)$, subject to $Z_k^{\mathsf{T}} Z_k = I_k$. The affine subspace $\operatorname{span}(Z_k) + \overline{x}$ in $\mathbb{R}^p$ captures the greatest k-dimensional variability in the data X. The k largest principal components of X are defined successively for $k = 1, \ldots, p$ as orthonormal basis of $\operatorname{span}(Z_k)$.

Example 2.1.4 (Support Vector Machine)**.** Let $\{(x_i, y_i) : x_i \in \mathbb{R}^p,\ y_i = \pm 1,\ i = 1, \ldots, n\}$ be a training set for binary classification. The best separating hyperplane is given by $w^{\mathsf{T}} x - \beta = 0$, where $(w, \beta)^{\mathsf{T}} \in \mathbb{R}^p \times \mathbb{R}$ can be found by minimizing $\|w\|$ subject to $y_i(w^{\mathsf{T}} x_i - b) \geq 1$ for all $i = 1, \ldots, n$. In other words, the best separating hyperplane is the affine subspace $\ker(w^{\mathsf{T}}) + \beta \mathbb{1}$.

These four examples represent a sampling of the most rudimentary classical examples. It is straightforward to extend them to include more modern considerations. We may incorporate say, sparsity or robustness, by changing the objective function used; or have matrix variables in place of vector variables by considering affine subspaces within other vector spaces, e.g., $\mathbb{S}^n$ or $\mathbb{R}^{m \times n}$ in place of $\mathbb{R}^n$.

These simple examples may be solved in the usual manners with techniques in numerical linear algebra: least squares for linear regression, singular value decomposition for errors-in-variables regression, eigenvalue decomposition for principal component analysis, linear programming for support vector machines. Nevertheless, viewing them in their full generality as optimization problems on the affine Grassmannian allows us to treat them on equal footings and facilitates development of new multivariate statistics/machine learning techniques. More importantly, we argue that the prevailing approaches may be suboptimal. For instance, in Example 2.1.3 one circumvents the problem of finding a best-fitting affine subspace with a two-step heuristic: First find the empirical mean of the data set $\overline{x}$ and then

mean center to reduce the problem to one of finding a best-fitting linear subspace $\operatorname{span}(Z)$. There is no reason that $\operatorname{span}(Z) + \overline{x}$ would give the best-fitting affine subspace.

2.1.1 Outline

We introduce the affine Grassmannian formally in Section 2.2 and develop its basic properties. We then introduce systems of global coordinates in Section 2.3, allowing us to compute various quantities on $\operatorname{Graff}(k, n)$. The next three sections cover the two most fundamental aspects from an application perspective: distances and probability distributions. In Section 2.4, we show how one may define various notions of distances and metrics on $\operatorname{Graff}(k, n)$ and extend these in Section 2.5 so that one may define distances between an affine subspace in $\operatorname{Graff}(k, n)$ and another in $\operatorname{Graff}(l, n)$. This is followed by a discussion of natural probability distributions on $\operatorname{Graff}(k, n)$ in Section 2.6. We conclude by presenting optimization algorithms on $\operatorname{Graff}(k, n)$ in Section 2.7, and providing numerical experiments in Section 2.8. We summarize our major contributions in this paper below:

- parametrization of a point: Definition 2.3.1 and Definition 2.3.4;
- characterization of embedding $\operatorname{Graff}(k, n)$ into $\operatorname{Gr}(k + 1, n + 1)$: Theorem 2.2.5 and Theorem 2.4.4
- distance between affine subspaces: Theorem 2.5.2
- probability distribution on $\operatorname{Graff}(k, n)$: Definition 2.6.3
- optimization algorithms on $\operatorname{Graff}(k, n)$: Algorithms 2.7.1–2.7.5

2.2 Affine Grassmannian

The Grassmannian of affine subspaces or affine Grassmannian was first described in [32] but has received relatively little attention compared to the Grassmannian of linear subspaces

$\mathrm{Gr}(k,n)$.[1] Aside from a brief discussion in [41, Section 9.1.3], we are unaware of any systematic treatments. Nevertheless, given that it naturally parameterizes all k-dimensional affine subspaces in $\mathbb{R}^n$, it is evidently an important object that could rival the usual Grassmannian in practical applicability, particularly in statistical estimation.

As such we will establish some basic properties of the affine Grassmannian that elucidates its structure with a view towards practical applications. The results here are neither difficult nor surprising, and certainly routine to the experts, but to the best of our knowledge they have not appeared before elsewhere.

We remind the reader of some basic terminologies. A *k-plane* is a k-dimensional linear subspace and a *k-flat* is a k-dimensional affine subspace. A *k-frame* is an ordered basis of a k-plane and we will regard it as an $n \times k$ matrix whose columns $a_1, \dots, a_k$ are the basis vectors. A *flag* is a strictly increasing sequence of nested linear subspaces, $\mathbf{X}_0 \subset \mathbf{X}_1 \subset \mathbf{X}_2 \subset \cdots$. A flag is said to be *complete* if $\dim \mathbf{X}_k = k$, *finite* if $k = 0, 1, \dots, n$, and *infinite* if $k \in \mathbb{N} \cup \{0\}$. We write $\mathrm{Gr}(k,n)$ for the *Grassmannian* of k-planes in $\mathbb{R}^n$, $\mathrm{V}(k,n)$ for the *Stiefel manifold* of orthonormal k-frames, and $\mathrm{O}(n) \coloneqq \mathrm{V}(n,n)$ for the *orthogonal group*. We may regard $\mathrm{V}(k,n)$ as a homogeneous space,

$$\mathrm{V}(k,n) \cong \mathrm{O}(n)/\mathrm{O}(n-k), \tag{2.2.1}$$

or more concretely as the set of $n \times k$ matrices with orthonormal columns. There is a right action of the orthogonal group $\mathrm{O}(k)$ on $\mathrm{V}(k,n)$: For $Q \in \mathrm{O}(k)$ and $A \in \mathrm{V}(k,n)$, the action yields $AQ \in \mathrm{V}(k,n)$ and the resulting homogeneous space is $\mathrm{Gr}(k,n)$, i.e.,

$$\mathrm{Gr}(k,n) \cong \mathrm{V}(k,n)/\mathrm{O}(k) \cong \mathrm{O}(n)/\big(\mathrm{O}(n-k) \times \mathrm{O}(k)\big). \tag{2.2.2}$$

By (2.2.2), $\mathbf{A} \in \mathrm{Gr}(k,n)$ may be identified with the equivalence class of its orthonormal

1. The term 'affine Grassmannian" is now used far more commonly to refer to another very different object; see [4, 17, 34] In this article, it will always be used in the sense of Definition 2.2.1. If desired, 'Grassmannian of affine subspaces' may be used to avoid ambiguity.

k-frames $\{AQ \in \mathrm{V}(k,n) : Q \in \mathrm{O}(k)\}$. Note $\operatorname{span}(AQ) = \operatorname{span}(A)$ for $Q \in \mathrm{O}(k)$.

Definition 2.2.1 (Affine Grassmannian)**.** Let $k < n$ be positive integers. The *Grassmannian of k-dimensional affine subspaces* in $\mathbb{R}^n$ or Grassmannian of k-flats in $\mathbb{R}^n$, denoted by $\operatorname{Graff}(k,n)$, is the set of all k-dimensional affine subspaces of $\mathbb{R}^n$. For an abstract vector space V, we write $\operatorname{Graff}_k(V)$ for the set of k-flats in V.

This set-theoretic definition does not reveal much about the rich geometry behind $\operatorname{Graff}(k,n)$. We will examine it below as (i) a differential manifold, (ii) a vector bundle, (iii) a homogeneous space, and (iv) an algebraic variety.

Throughout this article, a boldfaced letter $\mathbf{A}$ will always denote a subspace and the corresponding normal typeface letter A will then denote a matrix of basis vectors (often but not necessarily orthonormal) of $\mathbf{A}$. We denote a k-dimensional affine subspace as $\mathbf{A} + b \in \operatorname{Graff}(k,n)$ where $\mathbf{A} \in \operatorname{Gr}(k,n)$ is a k-dimensional linear subspace and $b \in \mathbb{R}^n$ is the displacement of $\mathbf{A}$ from the origin. If $A = [a_1, \dots, a_k] \in \mathbb{R}^{n\times k}$ is a basis of $\mathbf{A}$, then

$$\mathbf{A} + b \coloneqq \{\lambda_1 a_1 + \cdots + \lambda_k a_k + b \in \mathbb{R}^n : \lambda_1, \dots, \lambda_k \in \mathbb{R}\}. \tag{2.2.3}$$

The notation $\mathbf{A} + b$ may be taken to mean a coset of the subgroup $\mathbf{A}$ in the additive group $\mathbb{R}^n$ or the Minkowski sum of the sets $\mathbf{A}$ and $\{b\}$ in the Euclidean space $\mathbb{R}^n$. The dimension of $\mathbf{A} + b$ is defined to be the dimension of the vector space $\mathbf{A}$. As one would expect of a coset representative, the displacement vector b is not unique: For any $a \in \mathbf{A}$, we have $\mathbf{A} + b = \mathbf{A} + (a + b)$.

We may choose an orthonormal basis for $\mathbf{A}$ so that $A \in \mathrm{V}(k,n)$ and choose b to be orthogonal to $\mathbf{A}$ so that $A^{\mathsf{T}} b = 0$. Hence we may always represent $\mathbf{A} + b \in \operatorname{Graff}(k,n)$ by a matrix $[A, b_0] \in \mathbb{R}^{n\times(k+1)}$ where $A^{\mathsf{T}} A = I$ and $A^{\mathsf{T}} b_0 = 0$; in this case we call $[A, b_0]$ *orthogonal affine coordinates*. A moment's thought would reveal that any two orthogonal affine coordinates $[A, b_0], [A', b_0'] \in \mathbb{R}^{n\times(k+1)}$ of the same affine subspace $\mathbf{A} + b$ must have $A' = AQ$ for some $Q \in \mathrm{O}(k)$ and $b_0' = b_0$.

We will not insist on using orthogonal affine coordinates at all times as they can be unnecessarily restrictive (especially in proofs). Without these orthogonality conditions, a matrix $[A, b_0] \in \mathbb{R}^{n\times(k+1)}$ that represents an affine subspace $\mathbf{A} + b$ in the sense of (2.2.3) is called its *affine coordinates*.

Proposition 2.2.2. $\mathrm{Graff}(k, n)$ *is a smooth manifold.*

Proof. Let $\mathbf{A}+b \in \mathrm{Graff}(k, n)$ be represented by affine coordinates $[A, b_0] = [a_1, a_2, \dots a_k, b_0] \in \mathbb{R}^{n\times(k+1)}$, where b_0 is chosen so that $b - b_0 \in \mathbf{A}$. Since A has rank k, without loss of generality, we may assume that the $k \times k$ leading principal minor of A is nonzero.

Let U be the set of all $\mathbf{X} + y \in \mathrm{Graff}(k, n)$ whose affine coordinates $[X, y_0]$ have nonzero $k \times k$ leading principal minors. Then U is an open subset of $\mathrm{Graff}(k, n)$ containing $\mathbf{A} + b$. Each $\mathbf{X} + y \in U$ has unique affine coordinates $[\hat{X}, \hat{y}] \in \mathbb{R}^{n\times(k+1)}$ of the form

$$[\hat{X}, \hat{y}] = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ \hat{x}_{k+1,1} & \hat{x}_{k+1,2} & \dots & \hat{x}_{k+1,k} & \hat{y}_{k+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{x}_{n,1} & \hat{x}_{n,2} & \dots & \hat{x}_{n,k} & \hat{y}_n \end{bmatrix}.$$

It is routine to verify that $\varphi : U \to \mathbb{R}^{(n-k)(k+1)}$, $\mathbf{X} + y \mapsto [\hat{X}, \hat{y}]$, is a homeomorphism and thus gives a local chart for U. We may likewise define other local charts by the nonvanishing of other $k \times k$ minors and verify that the transition functions $\varphi_1 \circ \varphi_2^{-1}$ are smooth for any two such local charts $\varphi_i : U_i \to \mathbb{R}^{(n-k)(k+1)}$, $i = 1, 2$. □

It turns out that $\mathrm{Graff}(k, n)$ may be viewed as a vector bundle over $\mathrm{Gr}(k, n)$. Recall that if S is a subbundle of a vector bundle E on a manifold M, then Q is called the *quotient bundle* on M of E by S if there is a short exact sequence of vector bundles

$$0 \to S \to E \to Q \to 0. \tag{2.2.4}$$

In the context of Grassmannians, there is a special vector bundle over $\mathrm{Gr}(k, n)$, called the

tautological bundle, whose fiber over $\mathbf{A} \in \mathrm{Gr}(k,n)$ is simply $\mathbf{A}$ itself. One may view this as a subbundle of the *trivial vector bundle* $\mathrm{Gr}(k,n) \times \mathbb{R}^n$. If S is the tautological bundle and E is the trivial bundle in (2.2.4), then the quotient bundle Q is called the *universal quotient bundle* of $\mathrm{Gr}(k,n)$ [21, 39].

Proposition 2.2.3. $\mathrm{Graff}(k,n)$ *is the universal quotient bundle on* $\mathrm{Gr}(k,n)$.

Proof. Let $p : \mathrm{Graff}(k,n) \to \mathrm{Gr}(k,n)$, $\mathbf{A} + b \mapsto \mathbf{A}$ be the map that translates an affine space back to the origin. In terms of affine coordinates, $p([a_1, \dots, a_k, b_0]) = [a_1, \dots, a_k]$ where a_i's and b_0 are chosen as in the proof of Proposition 2.2.2. Notice that the fiber $p^{-1}(\mathbf{A})$ for $\mathbf{A} \in \mathrm{Gr}(k,n)$ is simply $\mathbb{R}^n/\mathbf{A}$, a linear subspace of dimension $n-k$. Local trivializations of $\mathrm{Graff}(k,n)$ are obtained from local charts of $\mathrm{Gr}(k,n)$ by construction. Hence $\mathrm{Graff}(k,n)$ is a vector bundle over $\mathrm{Gr}(k,n)$. Moreover we have $q : \mathrm{Gr}(k,n) \times \mathbb{R}^n \to \mathrm{Graff}(k,n)$, $(\mathbf{A}, b) \mapsto \mathbf{A} + b$. It is straightforward to check that q is a surjective bundle map and the kernel of q is the tautological vector bundle S over $\mathrm{Gr}(k,n)$, i.e., we have an exact sequence

$$0 \to S \to \mathrm{Gr}(k,n) \times \mathbb{R}^n \to \mathrm{Graff}(k,n) \to 0.$$

This shows that $\mathrm{Graff}(k,n)$ is the universal quotient bundle. □

By either Proposition 2.2.2 or Proposition 2.2.3, we see that $\dim \mathrm{Graff}(k,n) = (n-k)(k+1)$. Unlike $\mathrm{Gr}(k,n)$, $\mathrm{Graff}(k,n)$ is non-compact: take a sequence in $\mathrm{Graff}(k,n)$ represented in coordinates by $[A, mb]$ with $m \in \mathbb{N}$, $A = [a_1, \dots, a_k] \in \mathrm{V}(k,n)$, and $0 \neq b \in \mathbb{R}^n$ such that $A^{\mathsf{T}} b = 0$; observe that it has no convergent subsequence.

The *group of orthogonal affine transformations* is denoted $\mathrm{E}(n)$ and is the set $\mathrm{O}(n) \times \mathbb{R}^n$ equipped with the group operation $(Q_1, c_1)(Q_2, c_2) = (Q_1 Q_2, c_1 + Q_1 c_2)$.[2] The *affine Stiefel manifold* is defined to be the product manifold $\mathrm{Vaff}(k,n) \coloneqq \mathrm{V}(k,n) \times \mathbb{R}^n$. It is a

2. As semidirect product, $\mathrm{E}(n) = \mathrm{O}(n) \ltimes_\vartheta \mathbb{R}^n$, where $\vartheta : \mathrm{O}(n) \to \mathrm{Aut}(\mathbb{R}^n) = \mathrm{GL}(n)$ as inclusion.

homogeneous space because of the following analogue of (2.2.1),

$$\operatorname{Vaff}(k,n) \cong \operatorname{E}(n)/\operatorname{O}(k).$$

Given that $\operatorname{E}(n)$ has wide-ranging applications in engineering [13], we provide the following description of $\operatorname{Graff}(k,n)$ as a quotient of $\operatorname{E}(n)$.

Proposition 2.2.4. $\operatorname{Graff}(k,n)$ *is a reductive homogeneous Riemannian manifold. In fact, we have the following analogue of* (2.2.2)*,*

$$\operatorname{Graff}(k,n) \cong \operatorname{Vaff}(k,n)/\operatorname{O}(n-k) \cong \operatorname{E}(n)/\big(\operatorname{O}(n-k) \times \operatorname{E}(k)\big).$$

Proof. Since $\operatorname{Graff}(k,n)$ can be identified with an open subset of $\operatorname{Gr}(k+1,n+1)$, the Riemannian metric g_e on $\operatorname{Gr}(k+1,n+1)$ induces a metric on $\operatorname{Graff}(k,n)$. With this induced metric equipped, $\operatorname{Graff}(k,n)$ is a Riemannian manifold. The group $\operatorname{E}(n)$ acts on $\operatorname{Graff}(k,n)$ by $(Q,c)\cdot(\mathbf{A}+b) = Q\cdot\mathbf{A}+Qb+c$, where $(Q,c)\in\operatorname{E}(n)=\operatorname{O}(n)\times\mathbb{R}^n$, $\mathbf{A}+b\in\operatorname{Graff}(k,n)$, and $Q\cdot\mathbf{A} \coloneqq \operatorname{span}(QA)$. It is easy to see that $\operatorname{E}(n)$ acts on $\operatorname{Graff}(k,n)$ transitively and so $\operatorname{Graff}(k,n) \cong \operatorname{E}(n)/\operatorname{Stab}_{\mathbf{A}+b}\big(\operatorname{E}(n)\big)$, where $\operatorname{Stab}_{\mathbf{A}+b}\big(\operatorname{E}(n)\big)$ is the stabilizer of any fixed affine linear subspace $\mathbf{A}+b\in\operatorname{Graff}(k,n)$ in $\operatorname{E}(n)$. Now $\operatorname{Stab}_{\mathbf{A}+b}\big(\operatorname{E}(n)\big)$ consists of two types of actions. The first action is the affine action inside the plane $\mathbf{A}$, which is $\operatorname{E}(k)$, while the second action is the rotation around the orthogonal complement of $\mathbf{A}$, which is $\operatorname{O}(n-k)$. Hence we obtain $\operatorname{Stab}_{\mathbf{A}+b}\big(\operatorname{E}(n)\big) \cong \operatorname{O}(n-k)\times\operatorname{E}(k)$, and the representation of $\operatorname{Graff}(k,n)$ as a homogeneous Riemannian manifold follows. □

We now turn to the algebraic geometric aspects of $\operatorname{Graff}(k,n)$. One of our main goals is to show that the vast array of optimization techniques [1, 2, 3, 15, 26] and any probability densities [12] defined on the usual Grassmannian may be adpated to the affine Grassmannian. In this regard, it is the following view of $\operatorname{Graff}(k,n)$ as an Zariski open dense subset of $\operatorname{Gr}(k+1,n+1)$ that will prove most useful. Our construction of this embedding is illustrated

in Figure 2.1.

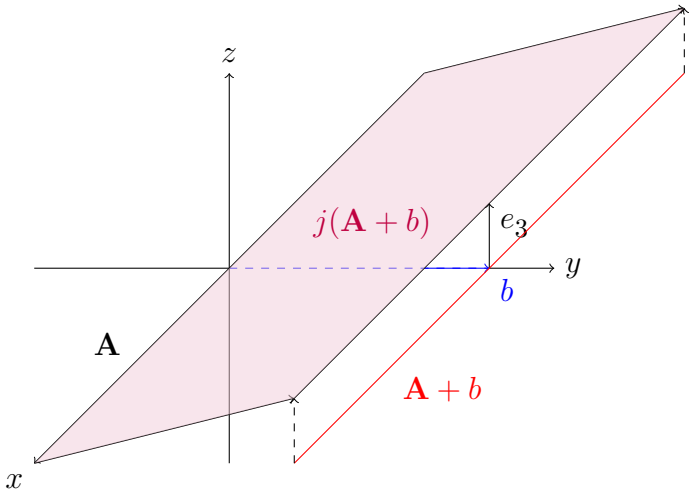

Figure 2.1: Here our linear subspace $\mathbf{A}$ is the x-axis. It is displaced by b along the y-axis to the affine subspace $\mathbf{A}+b$. The embedding $j : \mathrm{Graff}(k,n) \to \mathrm{Gr}(k+1,n+1)$ takes $\mathbf{A}+b$ to the smallest 2-plane containing $\mathbf{A}$ and $b+e_3$, where e_3 is a unit vector along the z-axis.

Theorem 2.2.5. (i) $\mathrm{Graff}(k,n)$ *is an algebraic variety that is irreducible and nonsingular.*

(ii) $\mathrm{Graff}(k,n)$ *may be embedded as a Zariski open subset of* $\mathrm{Gr}(k+1,n+1)$*,*

$$j : \mathrm{Graff}(k,n) \to \mathrm{Gr}(k+1,n+1), \quad \mathbf{A}+b \mapsto \mathrm{span}(\mathbf{A} \cup \{b+e_{n+1}\}), \tag{2.2.5}$$

where $e_{n+1} = (0,\dots,0,1)^{\mathsf{T}} \in \mathbb{R}^{n+1}$*. The image is open and dense in both the Zariski and manifold topologies.*

(iii) $\mathrm{Gr}(k+1,n+1)$ *may be regarded as the disjoint union of* $\mathrm{Gr}(k+1,n)$ *and* $\mathrm{Graff}(k,n)$*; more precisely,*

$$\mathrm{Gr}(k+1,n+1) = X \cup X^c, \quad X \cong \mathrm{Graff}(k,n), \quad X^c \cong \mathrm{Gr}(k+1,n).$$

Proof. Substituting 'smooth' with 'regular' and 'differential manifold' by 'algebraic variety' in the proof of Proposition 2.2.2, we see that $\mathrm{Graff}(k,n)$ is a nonsingular algebraic variety. Its irreducibility follows from Proposition 2.2.3 since $\mathrm{Gr}(k,n)$ is irreducible and all fibers of

$\mathrm{Graff}(k,n) \to \mathrm{Gr}(k,n)$ are irreducible and of the same dimension. We use 'algebraic variety' is used here in the sense of an abstract algebraic variety, i.e., $\mathrm{Graff}(k,n)$ is obtained by gluing together affine open subsets.

The embedding j takes k-flats in $\mathbb{R}^n$ to $(k+1)$-planes in $\mathbb{R}^{n+1}$, i.e., $\mathbb{R}^n \supseteq \mathbf{A} + b \mapsto \mathrm{span}(\mathbf{A} \cup \{b + e_{n+1}\}) \subseteq \mathbb{R}^{n+1}$. It maps $\mathbb{R}^n$ onto $E_n \coloneqq \mathrm{span}\{e_1, \dots, e_n\} \subset \mathbb{R}^{n+1}$ where $e_1, \dots, e_n, e_{n+1}$ are the standard basis vectors of $\mathbb{R}^{n+1}$. Linear subspaces $\mathbf{A} \subseteq \mathbb{R}^n$ are then mapped to $j(\mathbf{A}) \subseteq E_n$. Clearly j is an embedding.

We set $X \coloneqq j\big(\mathrm{Graff}(k,n)\big) \subseteq \mathrm{Gr}(k+1, n+1)$ and set X^c to be the set-theoretic complement of X in $\mathrm{Gr}(k+1,n+1)$. By (ii), $X \cong \mathrm{Graff}(k,n)$. By the definition of X^c, a $(k+1)$-plane $\mathbf{B} \in \mathrm{Gr}(k+1,n+1)$ is in X^c if and only if $\mathbf{B} \subseteq E_n$, which is to say that $X^c = \mathrm{Gr}_{k+1}(E_n) \cong \mathrm{Gr}(k+1,n)$. Lastly we see that X is Zariski open because its complement X^c, comprising $(k+1)$-planes in E_n, is clearly Zariski closed. □

In the proof we identified $\mathbb{R}^n$ with the subset $\{(x_1, \dots, x_n, 0)^\mathsf{T} \in \mathbb{R}^{n+1} : x_1, \dots, x_n \in \mathbb{R}\}$ to obtain a complete flag $\{0\} \subset \mathbb{R}^1 \subset \mathbb{R}^2 \subset \cdots \subset \mathbb{R}^n \subset \mathbb{R}^{n+1} \subset \cdots$. Given this, our choice of e_{n+1} in the embedding j in (2.2.5) is the most natural one.

It is sometimes desirable to represent elements of $\mathrm{Gr}(k,n)$ as actual matrices instead of equivalence classes of matrices. For example, we will see that this is the case when we discuss probability distributions on $\mathrm{Gr}(k,n)$ and $\mathrm{Graff}(k,n)$ in Section 2.6. The Grassmannian has a well-known representation [41, Example 1.2.20] as the set of rank-k orthogonal projection matrices, or, equivalently, the set of trace-k idempotent symmetric matrices:[3]

$$\mathrm{Gr}(k,n) \cong \{P \in \mathbb{R}^{n \times n} : P^\mathsf{T} = P^2 = P,\ \mathrm{tr}(P) = k\}. \tag{2.2.6}$$

Note that $\mathrm{rank}(P) = \mathrm{tr}(P)$ for an orthogonal projection matrix P. A straightforward affine

3. A projection matrix satisfies $P^2 = P$ and an orthogonal projection matrix is in addition symmetric, i.e., $P^\intercal = P$. An orthogonal projection matrix P is not an orthogonal matrix unless $P = I$.

analogue of (2.2.6) for $\operatorname{Graff}(k, n)$ is simply

$$\operatorname{Graff}(k, n) \cong \{[P, b] \in \mathbb{R}^{n\times(n+1)} : P^{\mathsf{T}} = P^2 = P,\ \operatorname{tr}(P) = k,\ Pb = 0\}, \tag{2.2.7}$$

where $\mathbf{A} + b \in \operatorname{Graff}(k, n)$ with orthogonal affine coordinates $[A, b_0] \in \mathbb{R}^{n\times(k+1)}$ is represented as the matrix $[AA^{\mathsf{T}}, b_0] \in \mathbb{R}^{n\times(n+1)}$.[4] We will call this the matrix of *projection affine coordinates* for $\mathbf{A} + b$.

2.3 Global coordinates for the affine Grassmannian

One reason for the wide applicability of the Grassmannian is the existence of several excellent choices of global coordinates, allowing subspaces to be represented as matrices and thereby facilitating the use of a vast range of algorithms in numerical linear algebra [1, 2, 3, 15]. Such concrete realizations of an abstract manifold is essential for applications purposes. We will show that this is also the case for the affine Grassmannian.

There are three particularly useful systems of global coordinates on the Grassmannian: points on $\operatorname{Gr}(k, n)$ can be represented as (i) an equivalence class of matrices $A \in \mathbb{R}^{n\times k}$ with linearly independent columns such that $A \sim AS$ for any $S \in \operatorname{GL}(k)$, (ii) an equivalence class of matrices $A \in \operatorname{V}(k, n)$ with orthonormal columns such that $A \sim AQ$ for any $Q \in \operatorname{O}(k)$, (iii) a projection matrix $P \in \mathbb{R}^{n\times n}$ satisfying $P^2 = P^{\mathsf{T}} = P$ and $\operatorname{tr}(P) = k$. These correspond to representing $\mathbf{A}$ by (i) bases of $\mathbf{A}$, (ii) orthonormal bases of $\mathbf{A}$, (iii) an orthogonal projection onto $\mathbf{A}$. The affine coordinates, orthogonal affine coordinates, and projection affine coordinates introduced in Section 2.2 are obvious analogues of (i), (ii), and (iii) respectively. In the following we will introduce two more.

For an affine subspace $\mathbf{A} + b \in \operatorname{Graff}(k, n)$, its orthogonal affine coordinates are $[A, b_0] \in \operatorname{V}(k, n) \times \mathbb{R}^n$ where $A^{\mathsf{T}} b_0 = 0$, i.e., b_0 is orthogonal to the columns of A. However as b_0 is in general not of unit norm, we may not regard $[A, b_0]$ as an element of $\operatorname{V}(n, k + 1)$. With this

4. If A is an orthonormal basis for the subspace $\mathbf{A}$, then AA^{T} is the orthogonal projection onto $\mathbf{A}$.

in mind, we introduce the notion of Stiefel coordinates, which is the most suitable system of coordinates for computations.

Definition 2.3.1. Let $\mathbf{A} + b \in \mathrm{Graff}(k, n)$ and $[A, b_0] \in \mathbb{R}^{n\times(k+1)}$ be its orthogonal affine coordinates, i.e., $A^{\mathsf{T}}A = I$ and $A^{\mathsf{T}}b_0 = 0$. The matrix of *Stiefel coordinates* for $\mathbf{A} + b$ is the $(n + 1) \times (k + 1)$ matrix with orthonormal columns,

$$Y_{\mathbf{A}+b} \coloneqq \begin{bmatrix} A & b_0/\sqrt{1 + \|b_0\|^2} \\ 0 & 1/\sqrt{1 + \|b_0\|^2} \end{bmatrix} \in \mathrm{V}(n + 1, k + 1).$$

Two orthogonal affine coordinates $[A, b_0], [A', b_0']$ of $\mathbf{A}+b$ give two corresponding matrices of Stiefel coordinates $Y_{\mathbf{A}+b}$, $Y'_{\mathbf{A}+b}$. By the remark after our definition of orthogonal affine coordinates, $A = A'Q'$ for some $Q' \in \mathrm{O}(k)$ and $b_0 = b_0'$. Hence

$$Y_{\mathbf{A}+b} = \begin{bmatrix} A & b_0/\sqrt{1 + \|b_0\|^2} \\ 0 & 1/\sqrt{1 + \|b_0\|^2} \end{bmatrix} = \begin{bmatrix} A' & b_0'/\sqrt{1 + \|b_0'\|^2} \\ 0 & 1/\sqrt{1 + \|b_0'\|^2} \end{bmatrix} \begin{bmatrix} Q' & 0 \\ 0 & 1 \end{bmatrix} = Y'_{\mathbf{A}+b}Q \tag{2.3.1}$$

where $Q \coloneqq \left[\begin{smallmatrix} Q' & 0 \\ 0 & 1 \end{smallmatrix}\right] \in \mathrm{O}(k + 1)$. Hence two different matrices of Stiefel coordinates for the same affine space differ by an orthogonal transformation.

Proposition 2.3.2. *Consider the equivalence class of matrices given by*

$$\begin{bmatrix} A & b \\ 0 & \gamma \end{bmatrix} \cdot \mathrm{O}(k + 1) \coloneqq \left\{ \begin{bmatrix} A & b \\ 0 & \gamma \end{bmatrix} Q \in \mathbb{R}^{(n+1)\times(k+1)} : Q \in \mathrm{O}(k + 1) \right\}.$$

The affine Grassmannian may be represented as a set of equivalence classes of $(n+1)\times(k+1)$

matrices with orthonormal columns,

$$\operatorname{Graff}(k,n) \cong \left\{ \begin{bmatrix} A & b \\ 0 & \gamma \end{bmatrix} \cdot \operatorname{O}(k+1) : \begin{bmatrix} A & b \\ 0 & \gamma \end{bmatrix} \in \operatorname{V}(k+1,n+1) \right\} \tag{2.3.2}$$

$$\subseteq \operatorname{V}(k+1,n+1)/\operatorname{O}(k+1) = \operatorname{Gr}(k+1,n+1). \tag{2.3.3}$$

An affine subspace $\mathbf{A}+b \in \operatorname{Graff}(k,n)$ is represented by the equivalence class $Y_{\mathbf{A}+b} \cdot \operatorname{O}(k+1)$ corresponding to its matrix of Stiefel coordinates.

Proof. The set of equivalence classes on the RHS of (2.3.2) is the set X in Theorem 2.2.5(iii) if $\operatorname{Gr}(k+1,n+1)$ is regarded as the homogeneous space in (2.3.3). □

The following lemma is easy to see from the definition of Stiefel coordinates and our discussion above. It will be useful for the optimization algorithms in Section 2.7, allowing us to check *feasibility*, i.e., whether a point represented as an $(n+1) \times (k+1)$ matrix is an element of the feasible set $j\big(\operatorname{Graff}(k,n)\big)$.

Lemma 2.3.3. (i) *Any matrix of the form $\left[\begin{smallmatrix} A & b \\ 0 & \gamma \end{smallmatrix}\right] \in \operatorname{V}(k+1,n+1)$, i.e.,*

$$A^{\mathsf{T}}A = I, \qquad A^{\mathsf{T}}b = 0, \qquad \|b\|^2 + \gamma^2 = 1,$$

is the matrix of Stiefel coordinates for some $\mathbf{A}+b \in \operatorname{Graff}(k,n)$.

(ii) *Two matrices of Stiefel coordinates $\left[\begin{smallmatrix} A & b \\ 0 & \gamma \end{smallmatrix}\right], \left[\begin{smallmatrix} A' & b' \\ 0 & \gamma' \end{smallmatrix}\right] \in \operatorname{V}(k+1,n+1)$ represent the same affine subspace iff there exists $\left[\begin{smallmatrix} Q' & 0 \\ 0 & 1 \end{smallmatrix}\right] \in \operatorname{O}(k+1)$ such that*

$$\begin{bmatrix} A & b \\ 0 & \gamma \end{bmatrix} = \begin{bmatrix} A' & b' \\ 0 & \gamma' \end{bmatrix} \begin{bmatrix} Q' & 0 \\ 0 & 1 \end{bmatrix}.$$

(iii) *If $\left[\begin{smallmatrix} A & b \\ 0 & \gamma \end{smallmatrix}\right] \in \operatorname{V}(k+1,n+1)$ is a matrix of Stiefel coordinates for $\mathbf{A}+b$, then every other*

matrix of Stiefel coordinates for $\mathbf{A}+b$ *belongs to the equivalence class* $\left[\begin{smallmatrix} A & b \\ 0 & \gamma \end{smallmatrix}\right] \cdot \mathrm{O}(k+1)$*, but not every matrix in* $\left[\begin{smallmatrix} A & b \\ 0 & \gamma \end{smallmatrix}\right] \cdot \mathrm{O}(k+1)$ *is a matrix of Stiefel coordinates for* $\mathbf{A}+b$*.*

The matrix of projection affine coordinates $[P, b] \in \mathbb{R}^{n\times(n+1)}$ in (2.2.7) is not an orthogonal projection matrix. With this in mind, we introduce the following notion.

Definition 2.3.4. Let $\mathbf{A}+b \in \mathrm{Graff}(k,n)$ and $[P,b] \in \mathbb{R}^{n\times(n+1)}$ be its projection affine coordinates. The matrix of *projection coordinates* for $\mathbf{A}+b$ is the orthogonal projection matrix

$$P_{\mathbf{A}+b} \coloneqq \begin{bmatrix} P + bb^{\mathsf{T}}/(\|b\|^2+1) & b/(\|b\|^2+1) \\ b^{\mathsf{T}}/(\|b|^2+1) & 1/(\|b\|^2+1) \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}.$$

Alternatively, in terms of orthogonal affine coordinates $[A, b_0] \in \mathbb{R}^{n\times(k+1)}$,

$$P_{\mathbf{A}+b} = \begin{bmatrix} AA^{\mathsf{T}} + b_0 b_0^{\mathsf{T}}/(\|b_0\|^2+1) & b_0/(\|b_0\|^2+1) \\ b_0^{\mathsf{T}}/(\|b_0|^2+1) & 1/(\|b_0\|^2+1) \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}.$$

It is straightforward to verify that $P_{\mathbf{A}+b}$ is indeed an orthogonal projection matrix, i.e., $P_{\mathbf{A}+b}^2 = P_{\mathbf{A}+b} = P_{\mathbf{A}+b}^{\mathsf{T}}$. Unlike Stiefel coordinates, projection coordinates of a given affine subspace are unique. As in Proposition 2.3.2, the next result gives a concrete description of the set $X = j\big(\mathrm{Graff}(k,n)\big)$ in Theorem 2.2.5(iii), but in terms of projection coordinates. With this description, $\mathrm{Graff}(k,n)$ may be regarded as a subvariety of $\mathbb{R}^{(n+1)\times(n+1)}$.

Proposition 2.3.5. *The affine Grassmannian may be represented as a set of* $(n+1)\times(n+1)$ *orthogonal projection matrices,*

$$\mathrm{Graff}(k,n) \cong \Bigg\{ \begin{bmatrix} P + bb^{\mathsf{T}}/(\|b\|^2+1) & b/(\|b\|^2+1) \\ b^{\mathsf{T}}/(\|b|^2+1) & 1/(\|b\|^2+1) \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)} : \\ P \in \mathbb{R}^{n\times n},\ P^{\mathsf{T}} = P^2 = P,\ \mathrm{tr}(P) = k,\ Pb = 0 \Bigg\}. \quad (2.3.4)$$

An affine subspace $\mathbf{A}+b \in \mathrm{Graff}(k,n)$ *is uniquely represented by its matrix of projection*

coordinates $P_{\mathbf{A}+b}$.

Proof. Let $\mathbf{A}+b \in \mathrm{Graff}(k,n)$ have orthogonal affine coordinates $[A, b_0]$. Since $P = AA^{\mathsf{T}} \in \mathbb{R}^{n\times n}$ is an orthogonal projection matrix that satisfies $Pb_0 = 0$, the map $\mathbf{A}+b \mapsto P_{\mathbf{A}+b}$ takes $\mathrm{Graff}(k,n)$ onto the set of matrices on the RHS of (2.3.4) with inverse given by $P_{\mathbf{A}+b} \mapsto \mathrm{im}(P) + b_0$. □

The following lemma allows us to check feasibility for the algorithms in Section 2.7 when we use projection coordinates.

Lemma 2.3.6. *An orthogonal projection matrix* $\left[\begin{smallmatrix} S & d \\ d^{\mathsf{T}} & \gamma \end{smallmatrix}\right] \in \mathbb{R}^{(n+1)\times(n+1)}$ *is the matrix of projection coordinates for some affine subspace in* $\mathbb{R}^n$ *iff* (i) $\gamma \neq 0$*;* (ii) $S - \gamma^{-1}dd^{\mathsf{T}} \in \mathbb{R}^{n\times n}$ *is an orthogonal projection matrix;* (iii) $Sd = 0$*. In addition,* $\left[\begin{smallmatrix} S & d \\ d^{\mathsf{T}} & \gamma \end{smallmatrix}\right] \in \mathbb{R}^{(n+1)\times(n+1)}$ *is the matrix of projection coordinates for* $\mathbf{A}+b \in \mathrm{Graff}(k,n)$ *iff* $S-\gamma^{-1}dd^{\mathsf{T}} = AA^{\mathsf{T}}$ *and* $\gamma^{-1}d = b_0$ *where* $[A, b_0] \in \mathbb{R}^{n\times(k+1)}$ *is an orthogonal affine coordinates of* $\mathbf{A}+b$*.*

The next lemma allows us to switch between Stiefel and projection coordinates.

Lemma 2.3.7. (i) *If* $Y_{\mathbf{A}+b} \in \mathrm{V}(k+1, n+1)$ *is a matrix of Stiefel coordinates for* $\mathbf{A}+b$*, then the matrix of projection coordinates for* $\mathbf{A}+b$ *is given by*

$$P_{\mathbf{A}+b} = Y_{\mathbf{A}+b}Y_{\mathbf{A}+b}^{\mathsf{T}} \in \mathbb{R}^{(n+1)\times(n+1)}.$$

(ii) *If* $P_{\mathbf{A}+b} \in \mathbb{R}^{(n+1)\times(n+1)}$ *is the matrix of projection coordinates for* $\mathbf{A}+b$*, then a matrix of Stiefel coordinates for* $\mathbf{A}+b$ *is given by any* $Y_{\mathbf{A}+b} \in \mathrm{V}(k+1, n+1)$ *whose columns form an orthonormal eigenbasis for the* 1*-eigenspace of* $P_{\mathbf{A}+b}$*.*

Proof. (i) follows from the observation that for any $Q \in \mathrm{O}(k+1)$,

$$\left(\begin{bmatrix} A & b/\sqrt{\|b\|^2+1} \\ 0 & 1/\sqrt{\|b\|^2+1} \end{bmatrix} Q\right)\left(\begin{bmatrix} A & b/\sqrt{\|b\|^2+1} \\ 0 & 1/\sqrt{\|b\|^2+1} \end{bmatrix} Q\right)^{\mathsf{T}} = \begin{bmatrix} AA^{\mathsf{T}} + bb^{\mathsf{T}}/\|b\|^2 + 1 & b/(\|b\|^2+1) \\ b^{\mathsf{T}}/(\|b|^2+1) & 1/(\|b\|^2+1) \end{bmatrix}.$$

For (ii), recall that the eigenvalues of an orthogonal projection matrix are 0's and 1's with multiplicities given by its nullity and rank respectively. Thus we have an eigenvalue decomposition of the form $P_{\mathbf{A}+b} = V \begin{bmatrix} I_{k+1} & \\ & 0_{n-k} \end{bmatrix} V^{\mathsf{T}} = V_{k+1} V_{k+1}^{\mathsf{T}}$, where the columns of $V_{k+1} \in \mathrm{V}(k+1, n+1)$ are the eigenvectors corresponding to the eigenvalue 1. Let $v \in \mathbb{R}^{k+1}$ be the last row of V_{k+1} and $Q \in \mathrm{O}(k+1)$ be a Householder matrix [19] such that $Q^{\mathsf{T}} v = \|v\| e_{k+1}$. Then $Y_{\mathbf{A}+b} = V_{k+1} Q$ has the form required in Lemma 2.3.3(i) for a matrix of Stiefel coordinates. □

The above proof also shows that projection coordinates are unique even though Stiefel coordinates are not. In principle, they are interchangeable via Lemma 2.3.7 but in reality, one form is usually more natural than the other for a specific use.

2.4 Geodesics and distances between affine subspaces

An important reason for the widespread applicability of the usual Grassmannian is that one has concrete, explicitly computable expressions for geodesics and distances on $\mathrm{Gr}(k, n)$. In [2, 15, 53], these expressions were obtained from a purely differential geometric perspective. One might imagine that a notion of distance between affine subspaces could be similarly obtained from the differential geometric structures on $\mathrm{Graff}(k, n)$ established in Propositions 2.2.2, 2.2.3, and 2.2.4. Surprisingly this is not the case.

A more careful examination of the arguments in [2, 15, 53] for obtaining an explicit expression for geodesics and geodesic distances on the $\mathrm{V}(k, n)$ and $\mathrm{Gr}(k, n)$ reveal that they depend on a somewhat obscure structure, namely, that of a *geodesic orbit space* [5, 20]. In general, if G is a compact semisimple Lie group and G/H is a reductive homogeneous space, then there is a standard metric induced by the restriction of the Killing form on $\mathfrak{g}/\mathfrak{h}$ where $\mathfrak{g}$ and $\mathfrak{h}$ are the Lie algebras of G and H respectively. With this standard metric, G/H is a geodesic orbit space, i.e., all geodesics are orbits of one parameter subgroups of G. In the case of $\mathrm{Gr}(k, n) = \mathrm{O}(n)/\big(\mathrm{O}(n-k) \times \mathrm{O}(k)\big)$ and Stiefel manifold $\mathrm{V}(k, n) = \mathrm{O}(n)/\mathrm{O}(n-k)$,

$\mathrm{O}(n)$ is a compact semisimple Lie group and the Riemannian metrics we use on $\mathrm{Gr}(k,n)$ and $V(k,n)$ are indeed the standard metrics. Hence they are geodesic orbit spaces. Moreover, for matrix Lie groups like $\mathrm{O}(n)$, we know that all their one parameter subgroups are given by the exponential maps. These observations allow us to write down geodesics on $\mathrm{Gr}(k,n)$ and $\mathrm{V}(k,n)$ explicitly.

In seeking an expression for the geodesic distance between affine subspaces, it might appear that we could just apply the same arguments to $\mathrm{Graff}(k,n)$, given that Proposition 2.2.4 guarantees the homogeneous space structure of $\mathrm{Graff}(k,n)$. The difficulty in this situation is that $\mathrm{Graff}(k,n) = \mathrm{E}(n)/\big(\mathrm{E}(n-k)\times\mathrm{O}(k)\big)$ might not be a geodesic orbit space since $\mathrm{E}(n) = \mathrm{O}(n)\times\mathbb{R}^n$ is not compact and therefore does not have a standard metric on $\mathrm{Graff}(k,n)$ as in the case of $\mathrm{Gr}(k,n)$ and $\mathrm{V}(k,n)$.

What about the vector bundle structure on $\mathrm{Graff}(k,n)$ then? If E is a vector bundle over a Riemannian manifold M, then there is always a metric induced on E by the metric on M, namely, the *pullback* of the metric on M. Nevertheless, this metric on E is evidently not very interesting — by definition, it disregards the fibers of the bundle. In the context of Proposition 2.2.3, this is akin to defining the distance between $\mathbf{A}+b$ and $\mathbf{B}+c \in \mathrm{Graff}(k,n)$ as the usual Grassmann distance between $\mathbf{A}$ and $\mathbf{B} \in \mathrm{Gr}(k,n)$, which ignores b and c totally.

In summary, the differential geometric structures on $\mathrm{Graff}(k,n)$ established in Propositions 2.2.2, 2.2.3, and 2.2.4 do not really help us define a distance between two affine subspaces. We will instead turn to the algebraic geometric properties of $\mathrm{Graff}(k,n)$ in Theorem 2.2.5 to provide the framework for defining such a distance. We will first describe the distance between two equidimensional affine subspaces and then extend it to affine subspaces of different dimensions in Section 2.5.

Recall that the Riemannian metric on $\mathrm{Gr}(k,n)$ yields the following well-known *Grassmann distance* between two subspaces $\mathbf{A},\mathbf{B} \in \mathrm{Gr}(k,n)$,

$$d_{\mathrm{Gr}(k,n)}(\mathbf{A},\mathbf{B}) = \Big(\sum\nolimits_{i=1}^{k} \theta_i^2\Big)^{1/2}, \tag{2.4.1}$$

where $\theta_1, \dots, \theta_k$ are the principal angles between $\mathbf{A}$ and $\mathbf{B}$. This distance is easily computable via SVD as $\theta_i = \cos^{-1} \sigma_i$, where σ_i is the ith singular value of the matrix $A^{\mathsf{T}}B$ for any orthonormal bases A and B of $\mathbf{A}$ and $\mathbf{B}$ [19, 54].

By Theorem 2.2.5(ii), we may identify $\operatorname{Graff}(k, n)$ with its image $j\big(\operatorname{Graff}(k, n)\big)$ in $\operatorname{Gr}(k+1, n+1)$. As a subset of $\operatorname{Gr}(k+1, n+1)$, $\operatorname{Graff}(k, n)$ inherits the Grassmann distance $d_{\operatorname{Gr}(k+1,n+1)}$ on $\operatorname{Gr}(k+1, n+1)$ and we obtain a distance as in Theorem 2.4.1 that can also be readily computed using SVD. We will show in Theorem 2.4.3 that this distance is in fact *intrinsic*.

Theorem 2.4.1. *For any two affine k-flats $\mathbf{A} + b$ and $\mathbf{B} + c \in \operatorname{Graff}(k, n)$,*

$$d_{\operatorname{Graff}(k,n)}(\mathbf{A} + b, \mathbf{B} + c) \coloneqq d_{\operatorname{Gr}(k+1,n+1)}\big(j(\mathbf{A} + b), j(\mathbf{B} + c)\big),$$

where j is the embedding in (2.2.5)*, defines a notion of distance consistent with the Grassmann distance. If*

$$Y_{\mathbf{A}+b} = \begin{bmatrix} A & b_0/\sqrt{1 + \|b_0\|^2} \\ 0 & 1/\sqrt{1 + \|b_0\|^2} \end{bmatrix}, \qquad Y_{\mathbf{B}+c} = \begin{bmatrix} B & c_0/\sqrt{1 + \|c_0\|^2} \\ 0 & 1/\sqrt{1 + \|c_0\|^2} \end{bmatrix}$$

are the matrices of Stiefel coordinates for $\mathbf{A} + b$ and $\mathbf{B} + c$ respectively, then

$$d_{\operatorname{Graff}(k,n)}(\mathbf{A} + b, \mathbf{B} + c) = \Big(\sum\nolimits_{i=1}^{k+1} \theta_i^2\Big)^{1/2}, \tag{2.4.2}$$

where $\theta_i = \cos^{-1} \sigma_i$ and σ_i is the ith singular value of $Y_{\mathbf{A}+b}^{\mathsf{T}} Y_{\mathbf{B}+c} \in \mathbb{R}^{(k+1)\times(k+1)}$.

Proof. Any nonempty subset of a metric space is a metric space. It remains to check that the definition does not depend on a choice of Stiefel coordinates. Let $Y_{\mathbf{A}+b}$ and $Y'_{\mathbf{A}+b}$ be two different matrices of Stiefel coordinates for $\mathbf{A} + b$ and $Y_{\mathbf{B}+c}$ and $Y'_{\mathbf{B}+c}$ be two different matrices of Stiefel coordinates for $\mathbf{B} + c$. By Lemma 2.3.3(ii), there exist $Q_1, Q_2 \in \operatorname{O}(k+1)$

such that $Y_{\mathbf{A}+b} = Y'_{\mathbf{A}+b}Q_1$, $Y_{\mathbf{B}+c} = Y'_{\mathbf{B}+c}Q_2$. The required result then follows from

$$\sigma_i(Y_{\mathbf{A}+b}^{\mathsf{T}}Y_{\mathbf{B}+c}) = \sigma_i(Q_1^{\mathsf{T}}Y_{\mathbf{A}+b}'^{\mathsf{T}}Y'_{\mathbf{B}+c}Q_2) = \sigma_i(Y_{\mathbf{A}+b}'^{\mathsf{T}}Y'_{\mathbf{B}+c}),$$

for $i = 1, \dots, k$. □

The proof above also shows that $\theta_1, \dots, \theta_{k+1}$ are independent of the choice of Stiefel coordinates. We will call θ_i the ith *affine principal angles* between the respective affine subspaces and denote it by $\theta_i(\mathbf{A} + b, \mathbf{B} + c)$. Consider the SVD,

$$Y_{\mathbf{A}+b}^{\mathsf{T}}Y_{\mathbf{B}+c} = U\Sigma V^{\mathsf{T}} \tag{2.4.3}$$

where $U, V \in \mathrm{O}(k+1)$ and $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_{k+1})$. Let

$$Y_{\mathbf{A}+b}U = [p_1, \dots, p_{k+1}], \qquad Y_{\mathbf{B}+c}V = [q_1, \dots, q_{k+1}].$$

We will call the pair of column vectors (p_i, q_i) the ith *affine principal vectors* between $\mathbf{A} + b$ and $\mathbf{B} + c$. These are clearly the affine analogues of principal angles and principal vectors of linear subspaces [11, 19, 54].

We will next show that the distance in Theorem 2.4.1 is the only possible distance on an affine Grassmannian compatible with the usual Grassmann distance on a Grassmannian. On any connected Riemannian manifold M with Riemannian metric g, there is an intrinsic distance function d_M on M with respect to g,

$$d_M(x, y) \coloneqq \inf\{L(\gamma) : \gamma \text{ is a piecewise smooth curve connecting } x \text{ and } y \text{ in } M\}.$$

For a connected submanifold of $N \subseteq M$, there is a natural Riemannian metric g_N on N

induced by g and therefore a corresponding intrinsic distance function,

$$d_N(x, y) \coloneqq \inf\{L(\gamma) : \gamma \text{ is a piecewise smooth curve connecting } x \text{ and } y \text{ in } N\}.$$

On the other hand, we may also define a distance function $d_M|_N$ on N by simply restricting the distance function d_M to N — note that this is what we have done in Theorem 2.4.1 with $M = \mathrm{Gr}(k+1, n+1)$ and $N = \mathrm{Graff}(k, n)$. In general, $d_M|_N \neq d_N$. For example, for $N = \mathbb{S}^2$ embedded as the unit sphere in $M = \mathbb{R}^3$, the two distance functions on $\mathbb{S}^2$ are obviously different. However, for our embedding of $\mathrm{Graff}(k, n)$ in $\mathrm{Gr}(k+1, n+1)$, the two distances on $\mathrm{Graff}(k, n)$ agree.

Proposition 2.4.2. *Let K be a closed submanifold of codimension at least two in M and let N be the complement of K in M. Then $d_M|_N = d_N$.*

Proof. We need to show that for any two distinct points $x, y \in N$, $d_M(x, y) = d_N(x, y)$. By definition of d_M and d_N it suffices to show that any piecewise smooth curve γ in M connecting x, y can be approximated by a piecewise smooth curve in N connecting x, y. The assumption on codimension implies that $x, y \in N$ is connected by a piecewise smooth curve in N. The transversality theorem [27, Theorem 2.4] then implies that γ can be approximated by curves in N connecting x and y. □

Theorem 2.4.3. *The distance $d_{\mathrm{Graff}(k,n)}$ in Theorem 2.4.1 is intrinsic with respect to the Riemannian metric on $\mathrm{Graff}(k, n)$ induced from that of $\mathrm{Gr}(k+1, n+1)$.*

Proof. By Theorem 2.2.5, the complement of $\mathrm{Graff}(k, n)$ in $\mathrm{Gr}(k+1, n+1)$ is $\mathrm{Gr}(k+1, n)$, which has codimension $k + 1 \geq 2$. Hence Proposition 2.4.2 applies. □

The next theorem plays an important role for our path-following algorithms in Section 2.7, showing that they will almost never lead to a point outside $\mathrm{Graff}(k, n)$.

Theorem 2.4.4. *Let* $\mathbf{A}+b, \mathbf{B}+c \in \operatorname{Graff}(k,n)$ *and let*

$$Y_{\mathbf{A}+b} = \begin{bmatrix} A & b_0/\sqrt{\|b_0\|^2+1} \\ 0 & 1/\sqrt{\|b_0\|^2+1} \end{bmatrix}, \qquad Y_{\mathbf{B}+c} = \begin{bmatrix} B & c_0/\sqrt{\|c_0\|^2+1} \\ 0 & 1/\sqrt{\|c\|^2+1} \end{bmatrix}$$

be their Stiefel coordinates. If $Y_{\mathbf{A}+b}^{\mathsf{T}} Y_{\mathbf{B}+c}$ *is invertible, then there is at most one point on the distance minimizing geodesic in* $\operatorname{Gr}(k+1,n+1)$ *connecting* $\mathbf{A}+b$ *and* $\mathbf{B}+c$ *which lies outside* $j\big(\operatorname{Graff}(k,n)\big)$*. Here* j *is the embedding in* (2.2.5)*.*

Proof. Let $U \in \mathrm{O}(k+1)$ and the diagonal matrix Σ be as in (2.4.3). We will write $\Theta \coloneqq \operatorname{diag}(\theta_1,\dots,\theta_{k+1}) = \cos^{-1}\Sigma$ for the diagonal matrix of affine principal angles. By [2] the geodesic $\gamma : [0,1] \to \operatorname{Gr}(k+1,n+1)$ connecting $j(\mathbf{A}+b)$ and $j(\mathbf{B}+c)$ is given by $\gamma(t) = \operatorname{span}\big(Y_{\mathbf{A}+b} U\cos(t\Theta) + Q\sin(t\Theta)\big)$, where $Q \in \mathrm{O}(k+1)$ is such that the RHS of

$$(I - Y_{\mathbf{A}+b}Y_{\mathbf{A}+b}^{\mathsf{T}})Y_{\mathbf{B}+c}(Y_{\mathbf{A}+b}^{\mathsf{T}}Y_{\mathbf{B}+c})^{-1} = Q(\tan(\Theta))U^{\mathsf{T}}$$

gives an SVD of the matrix on the LHS. Let the last row of U and Q as $[u_{k+1,1},\dots,u_{k+1,k+1}]^{\mathsf{T}}$ and $[q_{k+1,1},\dots,q_{k+1,k+1}]^{\mathsf{T}}$ respectively. Then $\gamma(t) \in \operatorname{Gr}(k+1,n+1) \setminus j\big(\operatorname{Graff}(k,n)\big)$ if and only if the entries on last row of $\gamma(t)$ are all zero, i.e.,

$$\frac{u_{k+1,i}\cos(t\theta_i)}{\sqrt{\|b\|^2+1}} + q_{k+1,i}\sin(t\theta_i) = 0,$$

for all $i = 1,\dots,k+1$. So at most one point on γ lies outside $\operatorname{Graff}(k,n)$. □

Corollary 2.4.5. *Let* $\mathbf{A}+b$ *and* $\mathbf{B}+c \in \operatorname{Graff}(k,n)$*. The distance minimizing geodesic* $\gamma : [0,1] \to \operatorname{Graff}(k,n)$ *connecting* $\mathbf{A}+b$ *and* $\mathbf{B}+c$ *is given by*

$$\gamma(t) = j^{-1}\big(\operatorname{span}(Y_{\mathbf{A}+b}U\cos t\Theta + Q\sin t\Theta)\big), \tag{2.4.4}$$

where $Q, U \in \mathrm{O}(k+1)$ and the diagonal matrix Θ are determined by the SVD

$$(I - Y_{\mathbf{A}+b} Y_{\mathbf{A}+b}^{\mathsf{T}}) Y_{\mathbf{B}+c} (Y_{\mathbf{A}+b}^{\mathsf{T}} Y_{\mathbf{B}+c})^{-1} = Q(\tan \Theta) U^{\mathsf{T}}.$$

The matrix U is the same as that in (2.4.3) *and $\Theta = \operatorname{diag}(\theta_1, \dots, \theta_{k+1})$ is the diagonal matrix of affine principal angles. γ attains the distance in* (2.4.2) *and its derivative at $t = 0$ is given by*

$$\gamma'(0) = j^{-1}\big(Q \Theta U^{\mathsf{T}}\big). \tag{2.4.5}$$

Table 2.1: Distances on $\operatorname{Graff}(k, n)$ in terms of affine principal angles and Stiefel coordinates.

	Affine principal angles	*Stiefel coordinates*
Asimov	$d^{\alpha}_{\operatorname{Graff}(k,n)}(\mathbf{A}+b, \mathbf{B}+c) = \theta_{k+1}$	$\cos^{-1} \lVert Y_{\mathbf{A}+b}^{\mathsf{T}} Y_{\mathbf{B}+c} \rVert_2$
Binet–Cauchy	$d^{\beta}_{\operatorname{Graff}(k,n)}(\mathbf{A}+b, \mathbf{B}+c) = \Big(1 - \prod_{i=1}^{k+1} \cos^2 \theta_i\Big)^{1/2}$	$(1 - (\det Y_{\mathbf{A}+b}^{\mathsf{T}} Y_{\mathbf{B}+c})^2)^{1/2}$
Chordal	$d^{\kappa}_{\operatorname{Graff}(k,n)}(\mathbf{A}+b, \mathbf{B}+c) = \Big(\sum_{i=1}^{k+1} \sin^2 \theta_i\Big)^{1/2}$	$\frac{1}{\sqrt{2}} \lVert Y_{\mathbf{A}+b} Y_{\mathbf{A}+b}^{\mathsf{T}} - Y_{\mathbf{B}+c} Y_{\mathbf{B}+c}^{\mathsf{T}} \rVert_F$
Fubini–Study	$d^{\phi}_{\operatorname{Graff}(k,n)}(\mathbf{A}+b, \mathbf{B}+c) = \cos^{-1}\Big(\prod_{i=1}^{k+1} \cos \theta_i\Big)$	$\cos^{-1} \lvert \det Y_{\mathbf{A}+b}^{\mathsf{T}} Y_{\mathbf{B}+c} \rvert$
Martin	$d^{\mu}_{\operatorname{Graff}(k,n)}(\mathbf{A}+b, \mathbf{B}+c) = \Big(\log \prod_{i=1}^{k+1} 1/\cos^2 \theta_i\Big)^{1/2}$	$(-2 \log \det Y_{\mathbf{A}+b}^{\mathsf{T}} Y_{\mathbf{B}+c})^{1/2}$
Procrustes	$d^{\rho}_{\operatorname{Graff}(k,n)}(\mathbf{A}+b, \mathbf{B}+c) = 2\Big(\sum_{i=1}^{k+1} \sin^2(\theta_i/2)\Big)^{1/2}$	$\lVert Y_{\mathbf{A}+b} U - Y_{\mathbf{B}+c} V \rVert_F$
Projection	$d^{\pi}_{\operatorname{Graff}(k,n)}(\mathbf{A}+b, \mathbf{B}+c) = \sin \theta_{k+1}$	$\lVert Y_{\mathbf{A}+b} Y_{\mathbf{A}+b}^{\mathsf{T}} - Y_{\mathbf{B}+c} Y_{\mathbf{B}+c}^{\mathsf{T}} \rVert_2$
Spectral	$d^{\sigma}_{\operatorname{Graff}(k,n)}(\mathbf{A}+b, \mathbf{B}+c) = 2 \sin(\theta_{k+1}/2)$	$\lVert Y_{\mathbf{A}+b} U - Y_{\mathbf{B}+c} V \rVert_2$

The Grassmann distance in (2.4.1) is the best known distance on the Grassmannian. But there are in fact several common distances on the Grassmannian [54, Table 2] and we may extend them to the affine Grassmannian by applying the embedding $j : \operatorname{Graff}(k, n) \to \operatorname{Gr}(k+1, n+1)$ and emulating our arguments in this section. We summarize these distances in Table 2.1. The matrices $U, V \in \mathrm{O}(k+1)$ in the right column of Table 2.1 are the ones in (2.4.3).

2.5 Distances between affine subspaces of different dimensions

The discussions in the previous section are about distances between affine subspaces of the *same* dimension. To provide a geometric framework for modeling mixtures of inequidimensional affine subspaces, we provide an extension to distance between affine subspaces of *different* dimensions.

For a long time, it is unclear how one might define a notion of distance between linear subspaces of different dimensions — not for the lack of proposed formulas for such a distance but that it is not clear what these ad hoc formulas measure, whether they depend on a choice of coordinates, or if they stay invariant when the dimension of the ambient space is changed. These issues have recently been resolved in [54]. Given a k-dimensional linear subspace $\mathbf{A} \in \operatorname{Gr}(k, n)$ and an l-dimensional linear subspace $\mathbf{B} \in \operatorname{Gr}(l, n)$ of $\mathbb{R}^n$, say $k < l$, it has been shown in [54] that if we take the set of all k-dimensional subspaces contained in $\mathbf{B}$, i.e., $\Omega_-(\mathbf{B}) \coloneqq \{\mathbf{X} \in \operatorname{Gr}(k, n) : \mathbf{X} \subseteq \mathbf{B}\}$, and define the required distance to be

$$d_{\operatorname{Gr}(k,n)}\big(\mathbf{A}, \Omega_-(\mathbf{B})\big) = \inf_{\mathbf{Y} \in \Omega_-(\mathbf{B})} d_{\operatorname{Gr}(k,n)}(\mathbf{A}, \mathbf{Y}), \tag{2.5.1}$$

and if we take the set of all l-dimensional subspaces containing $\mathbf{A}$, i.e., $\Omega_+(\mathbf{A}) \coloneqq \{\mathbf{Y} \in \operatorname{Gr}(l, n) : \mathbf{A} \subseteq \mathbf{Y}\}$, and define the required distance to be

$$d_{\operatorname{Gr}(l,n)}\big(\mathbf{B}, \Omega_+(\mathbf{A})\big) = \inf_{\mathbf{X} \in \Omega_+(\mathbf{A})} d_{\operatorname{Gr}(l,n)}(\mathbf{B}, \mathbf{X}), \tag{2.5.2}$$

the two results are identical. Their common value gives a natural notion of distance between subspaces of different dimensions, denoted by $\delta(\mathbf{A}, \mathbf{B})$. In addition this distance is independent of n, the dimension of the ambient space, and is easily computable via the SVD:

$$\delta(\mathbf{A}, \mathbf{B}) = \Big(\sum\nolimits_{i=1}^{\min(k,l)} \theta_i^2\Big)^{1/2}. \tag{2.5.3}$$

Moreover, this strategy applies to any of the distances in [54, Table 2], extending them to

subspaces of different dimensions [54, Theorem 12]. There is also an accompanying geometric insight — the sets $\Omega_+(\mathbf{A})$ and $\Omega_-(\mathbf{B})$ are well-studied objects called *Schubert varieties* and thus the distances (2.5.1) and (2.5.2) are distances from a point to a Schubert variety within the respective Grassmannians.

The goal of this section is to show that the framework described above may be adapted to define various notions of distances for affine subspaces of different dimensions. The proofs of Lemma 2.5.1, Theorems 2.5.2 and 2.5.3 are similar to those of their linear counterparts [54, Lemma 3, Theorems 7 and 12] and will be omitted.

We start by defining the *infinite affine Grassmannian*, a geometric object that parameterizes k-dimensional flats in $\mathbb{R}^n$ for *all* $n \geq k$. Formally, this is defined as

$$\operatorname{Graff}(k,\infty) \coloneqq \varinjlim \operatorname{Graff}(k,n),$$

where the *direct limit* is taken in the directed system given by the natural inclusions $\iota_n : \operatorname{Graff}(k,n) \to \operatorname{Graff}(k,n+1)$ for $n \geq k$. To be specific, if $[A,b] \in \mathbb{R}^{n\times(k+1)}$ is the affine coordinates of $\mathbf{A}+b \in \operatorname{Graff}(k,n)$, then $\iota_n(\mathbf{A}+b) = \mathbf{A}'+b'$ where $\mathbf{A}' = \operatorname{span}\left[\begin{smallmatrix} A \\ 0 \end{smallmatrix}\right]$, $b' = \left[\begin{smallmatrix} b \\ 0 \end{smallmatrix}\right]$, i.e., $\mathbf{A}'+b' \in \operatorname{Graff}(k,n+1)$ has affine coordinates $\left[\begin{smallmatrix} A & b \\ 0 & 0 \end{smallmatrix}\right] \in \mathbb{R}^{(n+1)\times(k+1)}$. Readers unfamiliar with direct limits may simply identify $[A,b]$ with $\left[\begin{smallmatrix} A & b \\ 0 & 0 \end{smallmatrix}\right]$ and thereby regard

$$\operatorname{Graff}(k,n) \subseteq \operatorname{Graff}(k,n+1) \qquad \text{and} \qquad \operatorname{Graff}(k,\infty) = \bigcup\nolimits_{n=k}^{\infty} \operatorname{Graff}(k,n).$$

Lemma 2.5.1. *The value $d_{\operatorname{Graff}(k,n)}(\mathbf{A}+b, \mathbf{B}+c)$ of two k-flats $\mathbf{A}+b$ and $\mathbf{B}+c \in \operatorname{Graff}(k,n)$ is independent of n, the dimension of their ambient space. Consequently, $d_{\operatorname{Graff}(k,n)}$ induces a distance $d_{\operatorname{Graff}(k,\infty)}$ on $\operatorname{Graff}(k,\infty)$.*

Let $\mathbf{A}+b \in \operatorname{Graff}(k,n)$ and $\mathbf{B}+c \in \operatorname{Graff}(l,n)$ where $k \leq l \leq n$. The *affine Schubert*

varieties of l-flats containing $\mathbf{A} + b$ and k-flats contained in $\mathbf{B} + c$ are

$$\Omega_+(\mathbf{A} + b) \coloneqq \big\{\mathbf{X} + y \in \operatorname{Graff}(l, n) : \mathbf{A} + b \subseteq \mathbf{X} + y\big\}, \tag{2.5.4}$$

$$\Omega_-(\mathbf{B} + c) \coloneqq \big\{\mathbf{Y} + z \in \operatorname{Graff}(k, n) : \mathbf{Y} + z \subseteq \mathbf{B} + c\big\}. \tag{2.5.5}$$

Theorem 2.5.2. *Let $k \leq l \leq n$. For any $\mathbf{A} + b \in \operatorname{Graff}(k, n)$ and $\mathbf{B} + c \in \operatorname{Graff}(l, n)$, the following distances are equal,*

$$d_{\operatorname{Graff}(k,n)}\big(\mathbf{A} + b, \Omega_-(\mathbf{B} + c)\big) = d_{\operatorname{Graff}(l,n)}\big(\mathbf{B} + c, \Omega_+(\mathbf{A} + b)\big), \tag{2.5.6}$$

and their common value $\delta(\mathbf{A} + b, \mathbf{B} + c)$ may be computed explicitly as

$$\delta(\mathbf{A} + b, \mathbf{B} + c) = \Big(\sum\nolimits_{i=1}^{\min(k,l)+1} \theta_i(\mathbf{A} + b, \mathbf{B} + c)^2\Big)^{1/2}. \tag{2.5.7}$$

The affine principal angles $\theta_1, \dots, \theta_{\min(h,l)+1}$ are as defined in Theorem 2.4.1 except that now they correspond to the singular values of a rectangular matrix

$$Y_{\mathbf{A}+b}^{\mathsf{T}} Y_{\mathbf{B}+c} = \begin{bmatrix} A & b_0/\sqrt{1 + \|b_0\|^2} \\ 0 & 1/\sqrt{1 + \|b_0\|^2} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} B & c_0/\sqrt{1 + \|c_0\|^2} \\ 0 & 1/\sqrt{1 + \|c_0\|^2} \end{bmatrix} \in \mathbb{R}^{(k+1)\times(l+1)}.$$

Like its counterpart for linear subspaces, δ defines a distance between the respective affine subspaces in the sense of a distance of a point to a set. It reduces to the usual Grassmann distance $d_{\operatorname{Gr}(k,n)}$ in (2.4.1) when $b = c = 0$ and $\dim \mathbf{A} = \dim \mathbf{B} = k$.

Another advantage of relying on an embedding of $\operatorname{Graff}(k, n)$ into $\operatorname{Gr}(k + 1, n + 1)$ for our definition of distance between affine subspaces is that $\operatorname{Graff}(k, n)$ automatically inherits the other distances on $\operatorname{Gr}(k + 1, n + 1)$, i.e., the distances in Table 2.1 may be extended to affine subspaces of different dimensions.

Theorem 2.5.3. *Let $k \leq l \leq n$. Let $\mathbf{A}+b \in \operatorname{Graff}(k,n)$, $\mathbf{B}+c \in \operatorname{Graff}(l,n)$. Then*

$$d^*_{\operatorname{Graff}(k,n)}\big(\mathbf{A}+b, \Omega_-(\mathbf{B}+c)\big) = d^*_{\operatorname{Graff}(l,n)}\big(\mathbf{B}+c, \Omega_+(\mathbf{A}+b)\big)$$

for $* = \alpha, \beta, \kappa, \mu, \pi, \rho, \sigma, \phi$. Their common value $\delta^*(\mathbf{A}+b, \mathbf{B}+c)$ is given by:

$$\begin{aligned}
&\delta^\alpha(\mathbf{A}+b, \mathbf{B}+c) = \theta_{k+1}, && \delta^\beta(\mathbf{A}+b, \mathbf{B}+c) = \Big(1 - \textstyle\prod_{i=1}^{k+1} \cos^2\theta_i\Big)^{1/2},\\
&\delta^\pi(\mathbf{A}+b, \mathbf{B}+c) = \sin\theta_{k+1}, && \delta^\mu(\mathbf{A}+b, \mathbf{B}+c) = \Big(\log \textstyle\prod_{i=1}^{k+1} \frac{1}{\cos^2\theta_i}\Big)^{1/2},\\
&\delta^\sigma(\mathbf{A}+b, \mathbf{B}+c) = 2\sin(\theta_{k+1}/2), && \delta^\phi(\mathbf{A}+b, \mathbf{B}+c) = \cos^{-1}\big(\textstyle\prod_{i=1}^{k+1} \cos\theta_i\big),\\
&\delta^\kappa(\mathbf{A}+b, \mathbf{B}+c) = \Big(\textstyle\sum_{i=1}^{k+1} \sin^2\theta_i\Big)^{1/2}, && \delta^\rho(\mathbf{A}+b, \mathbf{B}+c) = \Big(2\textstyle\sum_{i=1}^{k+1} \sin^2(\theta_i/2)\Big)^{1/2},
\end{aligned}$$

where $\theta_1, \dots, \theta_{k+1}$ are as defined above.

For the two affine Schubert varieties in (2.5.4), $\Omega_+(\mathbf{A}+b)$ may be viewed a Grassmannian whereas $\Omega_-(\mathbf{B}+c)$ may be viewed as an affine Grassmannian.

Proposition 2.5.4. *Let $\mathbf{A}+b \in \operatorname{Graff}(k,n)$ and $\mathbf{B}+c \in \operatorname{Graff}(l,n)$. Then*

$$\Omega_+(\mathbf{A}+b) \cong \operatorname{Gr}(n-l, n-k) \quad \text{and} \quad \Omega_-(\mathbf{B}+c) \cong \operatorname{Graff}(k,l)$$

as Riemannian manifolds and algebraic varieties.

Proof. We first observe that the map $\varphi : \Omega_+(\mathbf{A}+b) \to \Omega_+(\mathbf{A})$, $\mathbf{X}+y \mapsto \mathbf{X}+y-b$, is well-defined since $\mathbf{A} \subset \mathbf{X}+y-b$ by our choice of $\mathbf{X}+y$. Also, $\psi : \Omega_+(\mathbf{A}) \to \Omega_+(\mathbf{A}+b)$, $\mathbf{X} \mapsto \mathbf{X}+b$, is the inverse of φ and so it is an isomorphism. Together with [54, Proposition 21], we obtain the first isomorphism $\Omega_+(\mathbf{A}+b) \cong \Omega_+(\mathbf{A}) \cong \operatorname{Gr}(n-l, n-k)$. For the second isomorphism, consider $\varphi' : \Omega_-(\mathbf{B}+c) \to \operatorname{Graff}_k(\mathbf{B})$, $\mathbf{Y}+z \mapsto \mathbf{Y}+z-c$, which is well-defined since $\mathbf{Y}+z-c$ is an affine subspace of dimension k in $\mathbf{B}$. Its inverse is given by $\psi' : \operatorname{Graff}_k(\mathbf{B}) \to \Omega_-(\mathbf{B}+c)$, $\mathbf{Y}+z \mapsto \mathbf{Y}+z+c$, and so it is an isomorphism. The required isomorphism then follows from $\Omega_-(\mathbf{B}+c) \cong \operatorname{Graff}_k(\mathbf{B}) \cong \operatorname{Graff}(k,l)$. □

The reason for the asymmetry in Proposition 2.5.4 is as follows. $\Omega_+(\mathbf{A}+b)$ is a Grassmannian of *linear* subspaces since all affine subspaces containing $\mathbf{A}+b$ can be shifted back to the origin by the vector b. In the case of $\Omega_-(\mathbf{B}+c)$, shifting $\mathbf{B}+c$ back to the origin by c and then taking all affine subspaces contained in $\mathbf{B}$ still gives a Grassmannian of *affine* subspaces. We may also express them as

$$\Omega_+(\mathbf{A}+b) \cong \{[P,d] \in \mathbb{R}^{n\times(n+1)} : P^{\mathsf{T}} = P^2 = P,\ Pd = 0,\ \operatorname{tr}(P) = l,\ \mathbf{A} \subseteq \operatorname{im}(P)\},$$
$$\Omega_-(\mathbf{B}+c) \cong \{[P,d] \in \mathbb{R}^{n\times(n+1)} : P^{\mathsf{T}} = P^2 = P,\ Pd = 0,\ \operatorname{tr}(P) = k,\ \operatorname{im}(P) \subseteq \mathbf{B}\},$$

and regard $\operatorname{Graff}(k,n)$, $\operatorname{Graff}(l,n)$, $\Omega_+(\mathbf{A}+b)$, $\Omega_-(\mathbf{B}+c)$ as subsets of $\mathbb{R}^{n\times(n+1)}$.

As we mentioned earlier, δ in (2.5.7) and δ^* in Theorem 2.5.3 are distances in the sense of a distance from a point to a set, but they are not *metrics*. Nevertheless, it is not difficult to derive metrics from them. The *doubly infinite affine Grassmannian*, a geometric model for affine subspaces of all dimensions, is the disjoint union

$$\operatorname{Graff}(\infty,\infty) \coloneqq \bigsqcup\nolimits_{k=1}^{\infty} \operatorname{Graff}(k,\infty).$$

This is the affine analogue of $\operatorname{Gr}(\infty,\infty)$, the doubly infinite Grassmannian of linear subspaces of all dimensions, defined in [54, Section 5], where it is also shown to be metrizable with respect to any of the common distances between linear subspaces.

We will see how $\operatorname{Graff}(\infty,\infty)$ can likewise be metricized, i.e., how a metric can be defined between any pair of affine subspaces of arbitrary dimensions. The embedding $j : \operatorname{Graff}(k,n) \to \operatorname{Gr}(k+1,n+1)$ induces an embedding of sets $j_\infty : \operatorname{Graff}(\infty,\infty) \to \operatorname{Gr}(\infty,\infty)$. So $\operatorname{Graff}(\infty,\infty)$ may be identified with $j_\infty\big(\operatorname{Graff}(\infty,\infty)\big)$ and regarded as a subset of $\operatorname{Gr}(\infty,\infty)$. It inherits any metric on $\operatorname{Gr}(\infty,\infty)$: If $\mathbf{A}+b$ and $\mathbf{B}+c$ are affine subspaces of possibly different dimensions, we may define

$$d^*_{\operatorname{Graff}(\infty,\infty)}(\mathbf{A}+b, \mathbf{B}+c) \coloneqq d^*_{\operatorname{Gr}(\infty,\infty)}\big(j_\infty(\mathbf{A}+b), j_\infty(\mathbf{B}+c)\big),$$

for any choice of metric $d^*_{\mathrm{Gr}(\infty,\infty)}$ on $\mathrm{Gr}(\infty,\infty)$. For example, the following metrics correspond to Grassmann, chordal, and Procrustes distances.

Table 2.2: Metrics on $\mathrm{Graff}(\infty,\infty)$ in terms of affine principal angles and $k = \dim \mathbf{A}$, $l = \dim \mathbf{B}$.

Grassmann metric	$d_{\mathrm{Graff}(\infty,\infty)}(\mathbf{A}+b, \mathbf{B}+c) = \Big(\lvert k-l\rvert \pi^2/4 + \sum_{i=1}^{\min(k+1,l+1)} \theta_i^2\Big)^{1/2}$
Chordal metric	$d^{\kappa}_{\mathrm{Graff}(\infty,\infty)}(\mathbf{A}+b, \mathbf{B}+c) = \Big(\lvert k-l\rvert + \sum_{i=1}^{\min(k+1,l+1)} \sin^2 \theta_i\Big)^{1/2}$
Procrustes metric	$d^{\rho}_{\mathrm{Graff}(\infty,\infty)}(\mathbf{A}+b, \mathbf{B}+c) = \Big(\lvert k-l\rvert + 2\sum_{i=1}^{\min(k+1,l+1)} \sin^2(\theta_i/2)\Big)^{1/2}$

2.6 Probability densities on the affine Grassmannian

To do statistics on the affine Grassmannian, we will need reasonable notions of probability densities on them; we will introduce three in this section: uniform, Langevin (or von Mises–Fisher), and Langevin–Gaussian.

The Riemannian metric on $\mathrm{Gr}(k,n)$ that induces the Grassmann distance in (2.4.1) also induces a volume density $d\gamma_{k,n}$ on $\mathrm{Gr}(k,n)$ [41, Proposition 9.1.12] with

$$\mathrm{Vol}\big(\mathrm{Gr}(k,n)\big) = \int_{\mathrm{Gr}(k,n)} |d\gamma_{k,n}| = \binom{n}{k} \frac{\prod_{j=1}^{n} \omega_j}{\big(\prod_{j=1}^{k} \omega_j\big)\big(\prod_{j=1}^{n-k} \omega_j\big)}, \tag{2.6.1}$$

where $\omega_m \coloneqq \pi^{m/2}/\Gamma(1+m/2)$, volume of the unit ball in $\mathbb{R}^m$. A natural uniform probability density on $\mathrm{Gr}(k,n)$ is given by $d\mu_{k,n} \coloneqq \mathrm{Vol}\big(\mathrm{Gr}(k,n)\big)^{-1} |d\gamma_{k,n}|$.

By Theorem 2.2.5(ii), $\mathrm{Graff}(k,n)$ is a Zariski open dense subset in $\mathrm{Gr}(k+1,n+1)$ and we must have $\mu_{k+1,n+1}\big(\mathrm{Graff}(k,n)\big) = 1$. Therefore the restriction of $\mu_{k+1,n+1}$ to $\mathrm{Graff}(k,n)$ gives us a *uniform probability measure* on $\mathrm{Graff}(k,n)$. It has an interesting property: suppose $k \le l$ such that $k + l \ge n$ and $\mathrm{Graff}(k,n)$, $\mathrm{Graff}(l,n)$ are given their respective uniform probability measures. If we take two arbitrary affine subspaces of $\mathbb{R}^n$, $\mathbf{A} + b$ of dimension k and $\mathbf{B} + c$ of dimension l, the probability that a randomly chosen l-dimensional affine subspace contains $\mathbf{A} + b$ is equal to the probability that a randomly chosen k-dimensional

affine subspace is contained in $\mathbf{B}+c$.

Proposition 2.6.1. *Let $k \leq l \leq n$ be such that $k+l \geq n$. Let $\mathbf{A}+b \in \mathrm{Graff}(k,n)$ and $\mathbf{B}+c \in \mathrm{Graff}(l,n)$. The relative volume of $\Omega_+(\mathbf{A}+b)$ in $\mathrm{Graff}(l,n)$ and $\Omega_-(\mathbf{B}+c)$ in $\mathrm{Graff}(k,n)$ are identical. Furthermore, their common value does not depend on the choices of $\mathbf{A}+b$ and $\mathbf{B}+c$ but only on k,l,n and is given by*

$$\mu_{l+1,n+1}\big(\Omega_+(\mathbf{A}+b)\big) = \mu_{k+1,n+1}\big(\Omega_-(\mathbf{B}+c)\big) = \frac{(l+1)!(n-k)!\prod_{j=l-k+1}^{l+1}\omega_j}{(n+1)!(l-k)!\prod_{j=n-k+1}^{n+1}\omega_j}.$$

Proof. By Theorem 2.2.5(ii), we have

$$\mathrm{Vol}\big(\mathrm{Graff}(k,n)\big) = \mathrm{Vol}\big(\mathrm{Gr}(k+1,n+1)\big) = \binom{n+1}{k+1}\frac{\prod_{j=1}^{n+1}\omega_j}{\big(\prod_{j=1}^{k+1}\omega_j\big)\big(\prod_{j=1}^{n-k}\omega_j\big)}.$$

By Proposition 2.5.4, we have

$$\mathrm{Vol}\big(\Omega_+(\mathbf{A}+\mathbf{b})\big) = \mathrm{Vol}\big(\mathrm{Gr}(n-l,n-k)\big) = \binom{n-k}{n-l}\frac{\prod_{j=1}^{n-k}\omega_j}{\big(\prod_{j=1}^{n-l}\omega_j\big)\big(\prod_{j=1}^{l-k}\omega_j\big)},$$
$$\mathrm{Vol}\big(\Omega_-(\mathbf{B}+\mathbf{c})\big) = \mathrm{Vol}\big(\mathrm{Graff}(k,l)\big) = \binom{l+1}{k+1}\frac{\prod_{j=1}^{l+1}\omega_j}{\big(\prod_{j=1}^{k+1}\omega_j\big)\big(\prod_{j=1}^{l-k}\omega_j\big)}.$$

We divide $\mathrm{Vol}\big(\Omega_+(\mathbf{A}+\mathbf{b})\big)$ and $\mathrm{Vol}\big(\Omega_-(\mathbf{B}+\mathbf{c})\big)$ by $\mathrm{Vol}\big(\mathrm{Graff}(l,n)\big)$ and $\mathrm{Vol}\big(\mathrm{Graff}(k,n)\big)$ respectively and notice that $k+l \geq n$ to complete the proof. □

In the following we will use projection coordinates on $\mathrm{Graff}(n,k)$ as defined in Definition 2.3.4. By embedding $\mathrm{Graff}(k,n)$ as a subset $X = j\big(\mathrm{Graff}(k,n)\big) \subset \mathrm{Gr}(k+1,n+1)$ as in Theorem 2.2.5(ii) and noting that X is an open dense subset, we have that $\mu(X)=1$ for any Borel probability measure μ on $\mathrm{Gr}(k+1,n+1)$ (and that $\mu(X^c)=0$). Hence $\mathrm{Graff}(k,n)$ inherits any continuous probability distribution on $\mathrm{Gr}(k+1,n+1)$, in particular the Langevin distribution [12].

Definition 2.6.2. The *Langevin distribution*, also known as the *von Mises–Fisher distribu-*

tion, on $\operatorname{Graff}(k,n)$ is given by the probability density function

$$f_L(P_{\mathbf{A}+b} \mid S) \coloneqq \frac{1}{{}_1F_1\Big(\frac{1}{2}(k+1); \frac{1}{2}(n+1); S\Big)} \exp\big(\operatorname{tr}(SP_{\mathbf{A}+b})\big)$$

for any $\mathbf{A} + b \in \operatorname{Graff}(k,n)$. Here $S \in \mathbb{R}^{(n+1)\times(n+1)}$ is symmetric and ${}_1F_1$ is the confluent hypergeometric function of the first kind of a matrix argument [33].

${}_1F_1(a; b; S)$ has well-known expressions as series and integrals and may be characterized via functional equations and recurrence relations. However, its explicit expression is unimportant for this article — the only thing to note is that it can be efficiently evaluated [33] for any $a, b \in \mathbb{C}$ and symmetric $S \in \mathbb{C}^{(n+1)\times(n+1)}$.

Roughly speaking, the parameter $S \in \mathbb{R}^{(n+1)\times(n+1)}$ may be interpreted as a 'mean direction' and its eigendecomposition $S = V\Lambda V^{\mathsf{T}}$ gives an 'orientation' $V \in \mathrm{O}(n+1)$ with 'concentrations' $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_{n+1})$. In some sense, the Langevin distribution measures the first-order 'spread' on $\operatorname{Graff}(k,n)$. If $S = 0$, then the distribution reduces to the uniform distribution but if S is 'large' (i.e., $|\lambda_i|$'s are large), then the distribution concentrate about the orientation V. One might think that a *Bingham distribution* on $\operatorname{Graff}(k,n)$ that measures second-order 'spread' may be defined by

$$f_B(P_{\mathbf{A}+b} \mid S) \coloneqq \frac{1}{{}_1F_1\Big(\frac{1}{2}(k+1); \frac{1}{2}(n+1); S\Big)} \exp\big(\operatorname{tr}(P_{\mathbf{A}+b}SP_{\mathbf{A}+b})\big),$$

but this is identical to the Langevin distribution since $\operatorname{tr}(PSP) = \operatorname{tr}(SP^2) = \operatorname{tr}(SP)$ for any projection matrix P.

The Langevin distribution treats an affine subspace $\mathbf{A}+b \in \operatorname{Graff}(k,n)$ as a single object but there are occasions where it is desirable to distinguish between the linear subspace $\mathbf{A} \in \operatorname{Gr}(k,n)$ and the displacement vector $b \in \mathbb{R}^n$. We will next construct probability distributions on $\operatorname{Graff}(k,n)$ by amalgamating different probability distributions on $\operatorname{Gr}(k,n)$ and $\mathbb{R}^n$ (or rather, $\mathbb{R}^{n-k}$, as we will see).

In the following, we identify $\mathrm{Gr}(k,n)$ and $\mathrm{Graff}(k,n)$ with its projection affine coordinates, i.e., imposing equality in (2.2.6) and (2.2.7),

$$\mathrm{Gr}(k,n) = \{P \in \mathbb{R}^{n\times n} : P^{\mathsf{T}} = P^2 = P,\ \mathrm{tr}(P) = k\},$$
$$\mathrm{Graff}(k,n) = \{[P,b] \in \mathbb{R}^{n\times(n+1)} : P \in \mathrm{Gr}(k,n),\ Pb = 0\}.$$

We will define a marginal density on the linear subspaces, and then impose a conditional density on the displacement vectors in the orthogonal complement of the respective linear subspaces. We will use the Langevin distribution $f_L(P \mid S)$ on the linear spaces $P \in \mathrm{Gr}(n,k)$. Conditioning on P, we know there exists $Q \in \mathrm{O}(n)$ such that $\ker(P) = \{b \in \mathbb{R}^n : Pb = 0\} = QE_{n-k} \cong \mathbb{R}^{n-k}$, where $E_{n-k} \coloneqq \mathrm{span}\{e_1,\dots,e_{n-k}\} \subset \mathbb{R}^{n+1}$. We may use any probability distribution on $\ker(P) \cong \mathbb{R}^{n-k}$ but for concreteness, a natural choice is the *spherical Gaussian distribution* with probability density $f_G(x \mid \sigma^2) \coloneqq (2\pi\sigma^2)^{-(n-k)/2}\exp(-\|x\|^2/2\sigma^2)$. The conditional density on $\ker(P)$ is then

$$f_G(b \mid P,\sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^{n-k}}}\exp\left(-\frac{\|b\|^2}{2\sigma^2}\right) \tag{2.6.2}$$

for all $b \in \ker(P)$. Note that $Q^{\mathsf{T}}b = \begin{bmatrix} b' \\ 0 \end{bmatrix}$ where $b' \in \mathbb{R}^{n-k}$ and since $\|b\| = \|Q^{\mathsf{T}}b\| = \|b'\|$, it is fine to have b instead of b' appearing on the RHS of (2.6.2).

Definition 2.6.3. The probability density function of the *Langevin–Gaussian distribution* on $\mathrm{Graff}(k,n)$ is $f_{LG}([P,b] \mid S,\sigma^2) \coloneqq f_L(P \mid S)f_G(b \mid P,\sigma^2)$, i.e.,

$$f_{LG}([P,b] \mid S,\sigma^2) = \frac{1}{{}_1F_1\left(\frac{1}{2}k;\frac{1}{2}n;S\right)\sqrt{(2\pi\sigma^2)^{n-k}}}\exp\left(\mathrm{tr}(SP) - \frac{\|b\|^2}{2\sigma^2}\right),$$

where $S \in \mathbb{R}^{n\times n}$ is symmetric, $\sigma^2 > 0$.

2.7 Optimization on the affine Grassmannian

The embedding of $\operatorname{Graff}(k,n)$ as an open smooth submanifold of $\operatorname{Gr}(k+1,n+1)$ (by Proposition 2.2.2 and Theorem 2.2.5(ii)) allows us to borrow the Riemannian optimization framework on Grassmannians in [1, 2, 3, 15] to develop optimization algorithms on the affine Grassmannian. Moreover, we can show that the iterates of our algorithms will not fall outside $\operatorname{Graff}(k,n)$ with probability one, with respect to, say, any of the probability distributions in Section 2.6. We will present various geometric notions and algorithms on $\operatorname{Graff}(k,n)$ in terms of both orthogonal affine and projection affine coordinates. The higher dimensions required by the projection affine coordinates generally makes them less preferable to the orthogonal affine coordinates.

Propositions 2.7.1 and 2.7.2 are respectively summaries of [15] and [26] adapted for the affine Grassmannian. We refer readers to the original sources for the proofs.

Proposition 2.7.1. *The following are basic differential geometric notions on* $\operatorname{Graff}(k,n)$ *expressed in Stiefel coordinates.*

(i) *Tangent space: The tangent space at* $\mathbf{A}+b \in \operatorname{Graff}(k,n)$ *has representation*

$$\mathsf{T}_{\mathbf{A}+b}\big(\operatorname{Graff}(k,n)\big) = \big\{\Delta \in \mathbb{R}^{(n+1)\times(k+1)} : Y_{\mathbf{A}+b}^{\mathsf{T}}\Delta = 0\big\}.$$

(ii) *Exponential map: The geodesic* $Y(t)$ *with* $Y(0) = Y_{\mathbf{A}+b}$ *and* $\dot{Y}(0) = H$ *in* $\operatorname{Graff}(k,n)$ *has expression*

$$Y(t) = \begin{bmatrix} Y_{\mathbf{A}+b}V & U \end{bmatrix} \begin{bmatrix} \cos(t\Sigma) \\ \sin(t\Sigma) \end{bmatrix} V^{\mathsf{T}},$$

where $H = U\Sigma V^{\mathsf{T}}$ *is a condensed* SVD.

(iii) *Parallel transport: The parallel transport of* $\Delta \in \mathsf{T}_{\mathbf{A}+b}\big(\operatorname{Graff}(k,n)\big)$ *along the geodesic*

given by H has expression

$$\tau\Delta(t) = \left(\begin{bmatrix} Y_{\mathbf{A}+b}V & U \end{bmatrix} \begin{bmatrix} -\sin(t\Sigma) \\ \cos(t\Sigma) \end{bmatrix} U^{\mathsf{T}} + (I - UU^{\mathsf{T}})\right)\Delta,$$

where $H = U\Sigma V^{\mathsf{T}}$ is a condensed SVD.

(iv) *Gradient: Let $f : \mathbb{R}^{(n+1)\times(k+1)} \to \mathbb{R}$ satisfy $f(YQ) = f(Y)$ for every Y with $Y^{\mathsf{T}}Y = I$ and $Q \in \mathrm{O}(k+1)$. The gradient of f at $Y = Y_{\mathbf{A}+b}$ is*

$$\nabla f = f_Y - YY^{\mathsf{T}}f_Y \in \mathsf{T}_{\mathbf{A}+b}\big(\mathrm{Graff}(k,n)\big),$$

where $f_Y \in \mathbb{R}^{(n+1)\times(k+1)}$ with $(f_Y)_{ij} = \frac{\partial f}{\partial y_{ij}}$.

(v) *Hessian: Let f be as in* (iv). *The Hessian of f at $Y = Y_{\mathbf{A}+b}$ is*

(a) *as bilinear form: $\nabla^2 f : \mathsf{T}_{\mathbf{A}+b}\big(\mathrm{Graff}(k,n)\big) \times \mathsf{T}_{\mathbf{A}+b}\big(\mathrm{Graff}(k,n)\big) \to \mathbb{R}$,*

$$\nabla^2 f(\Delta, \Delta') = f_{YY}(\Delta, \Delta') - \mathrm{tr}(\Delta^{\mathsf{T}}\Delta' Y^{\mathsf{T}} f_Y),$$

where $(f_{YY})_{ij,hl} = \frac{\partial^2 f}{\partial y_{ij}\partial y_{hl}}$ and $f_{YY}(\Delta, \Delta') = \sum_{i,j,h,l}(f_{YY})_{ij,hl}\delta_{ij}\delta'_{hl}$; or

(b) *as linear map: $\nabla^2 f : \mathsf{T}_{\mathbf{A}+b}\big(\mathrm{Graff}(k,n)\big) \to \mathsf{T}_{\mathbf{A}+b}\big(\mathrm{Graff}(k,n)\big)$,*

$$\nabla^2 f(\Delta) = \sum\nolimits_{i,j,h,l=1}^{n+1,k+1,n+1,k+1} (f_{YY})_{ij,hl}\delta_{ij}E_{hl} - \Delta f_Y^{\mathsf{T}} Y,$$

where $E_{hl} \in \mathbb{R}^{(n+1)\times(k+1)}$ has (h,l)th entry 1 and all other entries 0.

Proposition 2.7.2. *The following are basic differential geometric notions on* $\mathrm{Graff}(k,n)$ *expressed in projection coordinates. We write $[X,Y] = XY - YX$ for the commutator bracket and $\Lambda^2(\mathbb{R}^n)$ for the space of $n \times n$ skew symmetric matrices.*

(i) *Tangent space: The tangent space at* $\mathbf{A} + b \in \mathrm{Graff}(k, n)$ *has representation*

$$\mathsf{T}_{\mathbf{A}+b}\big(\mathrm{Graff}(k,n)\big) = \{[P_{\mathbf{A}+b}, \Omega] \in \mathbb{R}^{(n+1)\times(n+1)} : \Omega \in \Lambda^2(\mathbb{R}^{n+1})\}.$$

(ii) *Exponential map: Let* $P = P_{\mathbf{A}+b}$ *and* $\Theta \in \mathbb{R}^{(n+1)\times(n+1)}$ *be such that* $[[P, \Omega], P] = \Theta^{\mathsf{T}}\left[\begin{smallmatrix} 0 & Z \\ -Z^{\mathsf{T}} & 0 \end{smallmatrix}\right]\Theta$ *and* $P = \Theta^{\mathsf{T}}\left[\begin{smallmatrix} I_{k+1} & 0 \\ 0 & 0 \end{smallmatrix}\right]\Theta$. *The exponential map is given by*

$$\exp_{\mathbf{A}+b}([P, \Omega]) = \frac{1}{2} I_{n+1} + \Theta^T \begin{bmatrix} \frac{1}{2}\cos(2\sqrt{ZZ^{\mathsf{T}}}) & -\operatorname{sinc}(2\sqrt{ZZ^{\mathsf{T}}})Z \\ -Z^{\mathsf{T}}\operatorname{sinc}(2\sqrt{ZZ^{\mathsf{T}}}) & -\frac{1}{2}\sin(2\sqrt{Z^{\mathsf{T}}Z}) \end{bmatrix} \Theta.$$

(iii) *Gradient: Let* $f : \mathbb{R}^{(n+1)\times(n+1)} \to \mathbb{R}$. *The gradient of* f *at* $P = P_{\mathbf{A}+b}$ *is*

$$\nabla f = [P, [P, f_P]] \in \mathsf{T}_{\mathbf{A}+b}\big(\mathrm{Graff}(k,n)\big),$$

where $f_P \in \mathbb{R}^{(n+1)\times(n+1)}$ *with* $(f_P)_{ij} = \frac{\partial f}{\partial p_{ij}}$.

(iv) *Hessian: Let* f *and* f_P *be as in* (iii). *The Hessian of* f *at* $P = P_{\mathbf{A}+b}$ *is*

(a) *as bilinear form:* $\nabla^2 f : \mathsf{T}_{\mathbf{A}+b}\big(\mathrm{Graff}(k,n)\big) \times \mathsf{T}_{\mathbf{A}+b}\big(\mathrm{Graff}(k,n)\big) \to \mathbb{R}$,

$$\nabla^2 f(\Delta, \Delta') = \mathrm{tr}\Big(\big([P, [P, \textstyle\sum_{i,j,h,l}(f_{PP})_{ij,hl}\delta_{ij}E_{hl}]] - \frac{1}{2}[P, [\nabla f, \Delta]] - \frac{1}{2}[\nabla f, [P, \Delta]]\big)\Delta'\Big),$$

where $f_{PP} \in \mathbb{R}^{n^2\times n^2}$ *with* $(f_{PP})_{ij,hl} = \frac{\partial^2 f}{\partial p_{ij}\partial p_{hl}}$ *and* $E_{hl} \in \mathbb{R}^{(n+1)\times(n+1)}$ *has* (h,l)*th entry* 1 *and all other entries* 0*; or*

(b) *as linear map:* $\nabla^2 f : \mathsf{T}_{\mathbf{A}+b}\big(\mathrm{Graff}(k,n)\big) \to \mathsf{T}_{\mathbf{A}+b}\big(\mathrm{Graff}(k,n)\big)$,

$$\nabla^2 f(\Delta) = \big[P, \big[P, \textstyle\sum_{i,j,h,l}(f_{PP})_{ij,hl}\delta_{ij}E_{hl}\big]\big] - \frac{1}{2}[P, [\nabla f, \Delta]] - \frac{1}{2}[\nabla f, [P, \Delta]].$$

Both forms of the Hessians are needed — they are used in different settings for computing descent direction in Newton's method.

We now discuss the methods of steepest descent, Newton, and conjugate gradient on the affine Grassmannian. The steepest descent and Newton methods are given in both Stiefel coordinates (Algorithms 2.7.1 and 2.7.2) and projection coordinates (Algorithms 2.7.4 and 2.7.5) but conjugate gradient method is only given in Stiefel coordinates (Algorithm 2.7.3) as we do not have a closed form expression for parallel transport in projection coordinates.

We will rely on our embedding of $\mathrm{Graff}(k, n)$ into $\mathrm{Gr}(k+1, n+1)$ via Stiefel coordinates or projection coordinates as given by Propositions 2.3.2 and 2.3.5 respectively. We will then borrow the corresponding methods on the usual Grassmannian developed in [2, 15] in conjunction with Propositions 2.7.1 and 2.7.2.

There is one caveat: Algorithms 2.7.1–2.7.5 are formulated as *infeasible methods*. If we start from a point in $\mathrm{Graff}(k, n)$, regarded as a subset of $\mathrm{Gr}(k+1, n+1)$, the next iterate along the geodesic may become *infeasible*, i.e., fall outside $\mathrm{Graff}(k, n)$. By Theorem 2.4.4, this will almost never happen but even if it does,[5] the algorithms will still work fine as algorithms on $\mathrm{Gr}(k+1, n+1)$.

Algorithm 2.7.1 Steepest descent in Stiefel coordinates

Initialize $\mathbf{A}_0 + b_0 \in \mathrm{Graff}(k, n)$ in Stiefel coordinates $Y_0 \coloneqq Y_{\mathbf{A}_0 + b_0} \in \mathbb{R}^{(n+1)\times(k+1)}$.
for $i = 0, 1, \dots$ **do**
  set $G_i = f_Y(Y_i) - Y_i Y_i^{\mathsf{T}} f_Y(Y_i)$; ▷ gradient of f at Y_i
  compute $-G_i = U \Sigma V^{\mathsf{T}}$; ▷ condensed SVD
  minimize $f(Y(t)) = f(Y_i V \cos(t\Sigma) V^{\mathsf{T}} + U \sin(t\Sigma) V^{\mathsf{T}})$ over $t \in \mathbb{R}$; ▷ exact line search
  set $Y_{i+1} = Y(t_{\min})$;
end for

If desired, we may undertake a more careful *prediction–correction* approach. Instead of having the points Y_{i+1} (in Stiefel coordinates) or P_{i+1} (in projection coordinates) be the next iterates, they will be 'predictors' of the next iterates. We will then use Lemmas 2.3.3 or 2.3.6 to check if Y_{i+1} or P_{i+1} are in $\mathrm{Graff}(k, n)$. In the unlikely scenario when they do fall outside $\mathrm{Graff}(k, n)$, e.g., if we have $Y_{i+1} = \left[\begin{smallmatrix} A & b \\ 0 & \gamma \end{smallmatrix}\right]$ where $A^{\mathsf{T}} b \neq 0$ or $P_{i+1} = \left[\begin{smallmatrix} S & d \\ d^{\mathsf{T}} & \gamma \end{smallmatrix}\right]$

5. Occurs with probability zero when the problem has noise-free initial data and the algorithms are performed in exact arithmetic.

Algorithm 2.7.2 Newton's method in Stiefel coordinates

Initialize $\mathbf{A}_0 + b_0 \in \mathrm{Graff}(k,n)$ in Stiefel coordinates $Y_0 \coloneqq Y_{\mathbf{A}_0+b_0} \in \mathbb{R}^{(n+1)\times(k+1)}$.
for $i = 0, 1, \ldots$ **do**
  set $G_i = f_Y(Y_i) - Y_i Y_i^\mathsf{T} f_Y(Y_i)$; ▷ gradient of f at Y_i
  find Δ such that $Y_i^\mathsf{T}\Delta = 0$ and $\nabla^2 f(\Delta) - \Delta(Y_i^\mathsf{T} f_Y(Y_i)) = -G$;
  compute $\Delta = U\Sigma V^\mathsf{T}$; ▷ condensed SVD
  $Y_{i+1} = Y_i V\cos(t\Sigma)V^\mathsf{T} + U\sin(t\Sigma)V^\mathsf{T}$; ▷ arbitrary step size t
end for

Algorithm 2.7.3 Conjugate gradient in Stiefel coordinates

Initialize $\mathbf{A}_0 + b_0 \in \mathrm{Graff}(k,n)$ in Stiefel coordinates $Y_0 \coloneqq Y_{\mathbf{A}_0+b_0} \in \mathbb{R}^{(n+1)\times(k+1)}$.
Set $G_0 = f_Y(Y_0) - Y_0 Y_0^\mathsf{T} f_Y(Y_0)$ and $H_0 = -G_0$.
for $i = 0, 1, \ldots$ **do**
  compute $H_i = U\Sigma V^\mathsf{T}$; ▷ condensed SVD
  minimize $f(Y(t)) = f(Y_i V\cos(t\Sigma)V^\mathsf{T} + U\sin(t\Sigma)V^\mathsf{T})$ over $t \in \mathbb{R}$; ▷ exact line search
  set $Y_{i+1} = Y(t_{\min})$;
  set $G_{i+1} = f_Y(Y_{i+1}) - Y_{i+1}Y_{i+1}^\mathsf{T} f_Y(Y_{i+1})$; ▷ gradient of f at Y_{i+1}
  procedure DESCENT(Y_i, G_i, H_i) ▷ set new descent direction at Y_{i+1}
    $\tau H_i = (-Y_i V\sin(t_{\min}\Sigma) + U\cos(t_{\min}\Sigma))\Sigma V^\mathsf{T}$; ▷ parallel transport of H_i
    $\tau G_i = G_i - \big(Y_i V\sin(t_{\min}\Sigma) + U(I - \cos(t_{\min}\Sigma))\big)U^\mathsf{T} G_i$; ▷ parallel transport of G_i
    $\gamma_i = \mathrm{tr}((G_{i+1} - \tau G_i)^\mathsf{T} G_{i+1}) / \mathrm{tr}(G_i^\mathsf{T} G_i)$;
    $H_{i+1} = -G_{i+1} + \gamma_i \tau H_i$;
  end procedure
  reset $H_{i+1} = -G_{i+1}$ if $i + 1 \equiv 0 \mod (k+1)(n-k)$;
end for

Algorithm 2.7.4 Steepest descent in projection coordinates

Initialize $\mathbf{A}_0 + b_0 \in \mathrm{Graff}(k,n)$ in projection coordinates $P_0 \coloneqq P_{\mathbf{A}_0+b_0} \in \mathbb{R}^{(n+1)\times(n+1)}$.
for $i = 0, 1, \ldots$ **do**
  set $\nabla f(P_i) = [P_i, [P_i, f_P(P_i)]]$;
  find $\Theta \in \mathbb{R}^{(n+1)\times(n+1)}$ and $t > 0$ so that $P_i = \Theta^\mathsf{T}\begin{bmatrix} I_{k+1} & 0 \\ 0 & 0 \end{bmatrix}\Theta$ and $-t\nabla f(P_i) = \begin{bmatrix} 0 & Z \\ -Z^\mathsf{T} & 0 \end{bmatrix}$;
  set $P_{i+1} = \frac{1}{2} I_{n+1} + \Theta^T \begin{bmatrix} \frac{1}{2}\cos(2\sqrt{ZZ^\mathsf{T}}) & -\mathrm{sinc}(2\sqrt{ZZ^\mathsf{T}})Z \\ -Z^\mathsf{T}\,\mathrm{sinc}(2\sqrt{ZZ^\mathsf{T}}) & -\frac{1}{2}\sin(2\sqrt{Z^\mathsf{T}Z}) \end{bmatrix}\Theta$;
end for

Algorithm 2.7.5 Newton's method in projection coordinates

Initialize $\mathbf{A}_0 + b_0 \in \operatorname{Graff}(k, n)$ in projection coordinates $P_0 \coloneqq P_{\mathbf{A}_0+b_0} \in \mathbb{R}^{(n+1)\times(n+1)}$.
for $i = 0, 1, \dots$ **do**
find $\Omega_i \in \bigwedge^2(\mathbb{R}^{n+1})$ such that

$$[P_i, [P_i, \nabla^2 f([P_i, [P_i, \Omega_i]])]] - [P_i, [\nabla f(P_i), [P_i, \Omega_i]]] = -[P_i, [P_i, \nabla f(P_i)]];$$

find $\Theta_i \in \mathrm{SO}(n+1)$ such that $P_i = \Theta_i^{\mathsf{T}} \begin{bmatrix} I_{k+1} & 0 \\ 0 & 0 \end{bmatrix} \Theta_i$; ▷ QR factorization
compute $\Theta_i(I - [P_i, [P_i, t\Omega_i]])\Theta_i^{\mathsf{T}} = Q_i R_i$; ▷ QR factorization with positive diagonal in R_i
set $P_{i+1} = \Theta_i^{\mathsf{T}} Q_i \Theta_i P_i \Theta_i^{\mathsf{T}} Q_i^{\mathsf{T}} \Theta_i$;
end for

where $Sd \neq 0$, we will 'correct' the iterates to feasible points $\widetilde{Y}_{i+1}$ or $\widetilde{P}_{i+1}$ by an appropriate reorthogonalization.

2.8 Numerical Experiments

We will present various numerical experiments on two problems to illustrate the conjugate gradient and steepest descent algorithms in Section 2.7. These problems are deliberately chosen to be non-trivial and yet have closed form solutions — so that we may check whether our algorithms have converged to the true solutions of these problems.

2.8.1 Eigenvalue problem coupled with quadratic fractional programming

Let $A \in \mathbb{R}^{n\times n}$ be symmetric, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$. We would like to solve

$$\begin{aligned} &\text{minimize} \quad \operatorname{tr}(X^{\mathsf{T}} A X) + \frac{1}{1+\|y\|^2}(y^{\mathsf{T}} A y + 2b^{\mathsf{T}} y + c), \\ &\text{subject to} \quad X^{\mathsf{T}} X = I,\ X^{\mathsf{T}} y = 0, \end{aligned} \tag{2.8.1}$$

over all $X \in \mathbb{R}^{n\times k}$ and $y \in \mathbb{R}^n$. If we set $y = 0$ in (2.8.1), the resulting quadratic trace minimization problem with orthonormal constraints is essentially a *symmetric eigenvalue problem*; if we set $X = 0$ in (2.8.1), the resulting nonconvex optimization problem is a

quadratic fractional programming problem.

By rearranging terms, we see that (2.8.1) is equivalent to a minimization problem over an affine Grassmannian,

$$\min_{\mathbf{X}+y\in\mathrm{Graff}(k,n)} \operatorname{tr}\left(\begin{bmatrix} X & y/\sqrt{1+\|y\|^2} \\ 0 & 1/\sqrt{1+\|y\|^2} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} A & b \\ b^{\mathsf{T}} & c \end{bmatrix} \begin{bmatrix} X & y/\sqrt{1+\|y\|^2} \\ 0 & 1/\sqrt{1+\|y\|^2} \end{bmatrix}\right), \tag{2.8.2}$$

which shows that the problem (2.8.1) is in fact coordinate independent, depending on X and y only through the affine subspace $\operatorname{span}(X) + y = \mathbf{X} + y$. Formulated in this manner, we may determine a closed-form solution via the eigenvalue decomposition of $\left[\begin{smallmatrix} A & b \\ b^{\mathsf{T}} & c \end{smallmatrix}\right]$ — the optimum value is the sum of the $k+1$ smallest eigenvalues.

Figure 3.1 shows convergence trajectories of steepest descent and conjugate gradient in Stiefel coordinates, i.e., Algorithms 2.7.1 and 2.7.3, on $\mathrm{Graff}(3,6)$ for the problem (2.8.2). $\mathrm{Graff}(3,6)$ is a 12-dimensional manifold; we generate $A \in \mathbb{R}^{6\times 6}$, $b \in \mathbb{R}^6$, $c \in \mathbb{R}$ randomly with $\mathcal{N}(0,1)$ entries, and likewise pick a random initial point in $\mathrm{Graff}(3,6)$. The gradient of $f(Y) \coloneqq \operatorname{tr}\big(Y^{\mathsf{T}}\left[\begin{smallmatrix} A & b \\ b^{\mathsf{T}} & c \end{smallmatrix}\right]Y\big)$ is given by $\nabla f(Y) = \left[\begin{smallmatrix} A & b \\ b^{\mathsf{T}} & c \end{smallmatrix}\right]Y$. Both algorithms converge to the true solution but conjugate gradient converges twice as fast when measured by the number of iterations, taking around 20 iterations for near-zero error reduction as opposed to steepest descent's 40 iterations. The caveat is that each iteration of conjugate gradient is more involved and requires roughly twice the amount of time it takes for each iteration of steepest descent.

We perform more extensive experiments by taking the average of 100 instances of the problem (2.8.1) for various values of k and n to generate tables of timing and accuracy. Table 2.3 and 2.4 show the robustness of the algorithm with respect to different choices of k and n.

Table 2.5 shows a modest initial increase followed by a decrease in elapsed time to convergence as k increases — a reflection of the intrinsic dimension of the problem as $\dim\big(\mathrm{Graff}(k,100)\big) = (k+1)(100-k)$ first increases and then decreases. On the other

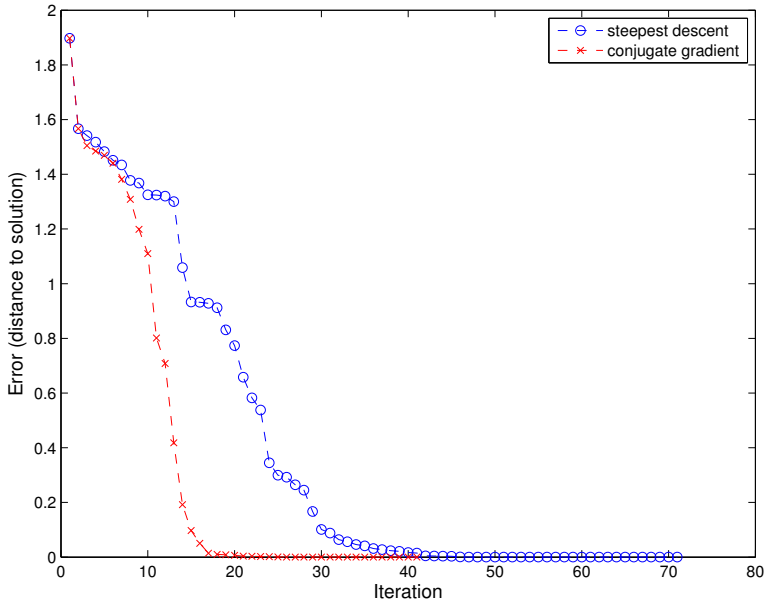

Figure 2.2: Convergence trajectories of steepest descent and conjugate gradient for a quadratic optimization problem on the affine Grassmannian $\mathrm{Graff}(3, 6)$.

k	10	21	32	43	54	65	76	87	98
Steepest descent ($\times 10^{-6}$)	0.61	3.1	1.5	1.7	2.9	6.8	1.2	0.25	0.1
Conjugate gradient ($\times 10^{-8}$)	0.77	1.5	1.9	2.4	2.3	2.9	3.1	3.5	3.3

Table 2.3: Accuracy (distance to true solution) of steepest descent and conjugate gradient for a quadratic optimization problem on $\mathrm{Graff}(k, 100)$.

hand, if we fix the dimension of ambient space, Table 2.6 shows that the elapsed time increases with k the dimension of the affine subspace we seek for. The results reflect the elapsed time increases with the dimension of the affine Grassmannian.

2.8.2 Fréchet mean and Karcher mean of affine subspaces

Let $d = d_{\mathrm{Graff}(k,n)}$, the geodesic distance on $\mathrm{Graff}(k, n)$ as defined in (2.4.2). We would like to solve for the minimizer $\mathbf{X} + y \in \mathrm{Graff}(k, n)$ in the sum-of-square-distances minimization problem:

$$\min_{\mathbf{X}+y \in \mathrm{Graff}(k,n)} \sum_{i=1}^{m} d^2(\mathbf{A}_i + b_i, \mathbf{X} + y), \tag{2.8.3}$$

n	7	17	27	37	47	57	67	77	87
Steepest descent ($\times 10^{-7}$)	4.4	4.8	4.4	4.7	4.7	4.7	4.3	4.7	4.1
Conjugate gradient ($\times 10^{-6}$)	0.83	0.98	1.0	1.3	1.2	1.3	1.5	1.6	1.5

Table 2.4: Accuracy (distance to true solution) of steepest descent and conjugate gradient for a quadratic optimization problem on $\mathrm{Graff}(6, n)$.

k	10	21	32	43	54	65	76	87	98
Steepest descent	0.6	0.89	1.4	1.4	1.8	1.9	2.0	2.0	1.3
Conjugate gradient	0.18	0.26	0.35	0.39	0.49	0.48	0.51	0.51	0.41

Table 2.5: Elapsed time (in seconds) of steepest descent and conjugate gradient for a quadratic optimization problem on $\mathrm{Graff}(k, 100)$.

where $\mathbf{A}_i + b_i \in \mathrm{Graff}(k, n)$, $i = 1, \ldots, m$. The Riemannian gradient of the objective function

$$f_m(\mathbf{X} + y) = \sum\nolimits_{i=1}^{m} d^2(\mathbf{A}_i + b_i, \mathbf{X} + y) \tag{2.8.4}$$

is given by [31]

$$\nabla f_m(\mathbf{X} + y) = \frac{1}{2} \sum\nolimits_{i=1}^{m} \log_{\mathbf{X}+y}(\mathbf{A}_i + b_i),$$

where $\log_{\mathbf{X}+y}(\mathbf{A} + b)$ denotes the derivative of the geodesic $\gamma(t)$ connecting $\mathbf{X} + y$ and $\mathbf{A} + b$ at $t = 0$, with an explicit expression given by (2.4.5).

The global minimizer of this problem is called the *Fréchet mean* and a local minimizer is called a *Karcher mean*. For the case $m = 2$, they coincide and is given by the midpoint of the geodesic connecting $\mathbf{A}_1 + b_1$ and $\mathbf{A}_2 + b_2$, which has a closed form expression given by (2.4.4) with $t = 1/2$.

We will take the $\mathrm{Graff}(7, 19)$, a 96-dimensional manifold, as our specific example. Our objective function is $f_2(\mathbf{X} + y) = d^2(\mathbf{A}_1 + b_1, \mathbf{X} + y) + d^2(\mathbf{A}_2 + b_2, \mathbf{X} + y)$ and we set our initial point as one of the two affine subspaces.

The result, depicted in Figure 3.2, shows that steepest descent outperforms conjugate gradient in this specific example. The departure from the example we considered in Section 2.8.1 can be explained by the fact that the function in (2.8.4) is geodesically convex

n	7	17	27	37	47	57	67	77	87
Steepest descent	0.67	0.96	0.94	1.1	1.2	1.3	1.4	1.4	1.5
Conjugate gradient	0.23	0.29	0.3	0.34	0.33	0.38	0.39	0.39	0.42

Table 2.6: Elapsed time (in seconds) of steepest descent and conjugate gradient for a quadratic optimization problem on $\mathrm{Graff}(6, n)$

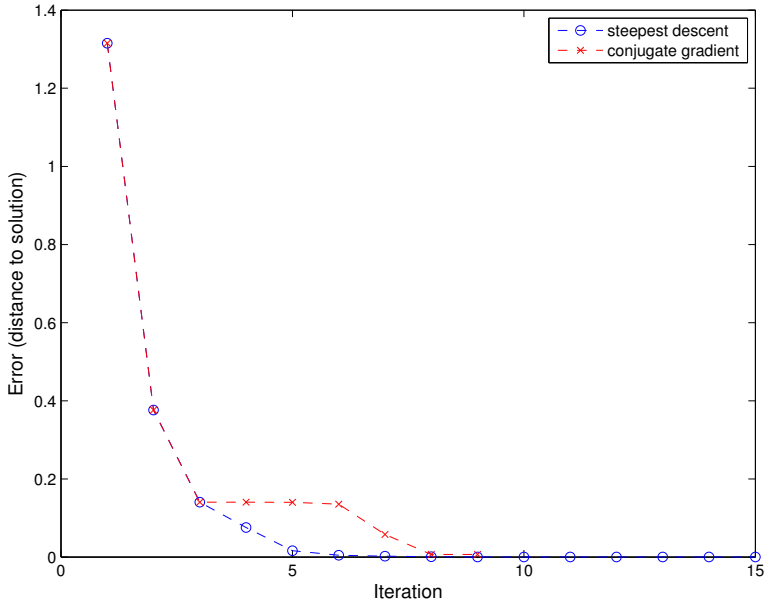

Figure 2.3: Convergence trajectories of steepest descent and conjugate gradient for Fréchet/Karcher mean on the affine Grassmannian $\mathrm{Graff}(7, 19)$

on the affine Grassmannian, whereas the function in (2.8.2) is not. In fact, when we find the Karcher mean of $m > 2$ affine subspaces by extending f_m to the objective function in (2.8.3), we see faster convergence (as measured by actual elapsed time) in conjugate gradient instead. Note that f_2 is geodesically convex in a neighborhood of the entire geodesic connecting $\mathbf{A}_1 + b_1$ and $\mathbf{A}_2 + b_2$, but f_m for $m > 2$ is only geodesically convex in neighborhoods of its local minimizers, i.e., the Karcher means.

k	1	2	3	4	5	6	7	8	9
Steepest descent ($\times 10^{-7}$)	5.3	5.1	4.6	4.8	4.4	4.9	4.7	4.6	5.0
Conjugate gradient ($\times 10^{-1}$)	0.5	2.6	1.5	1.6	2.7	2.0	2.0	2.5	19.0

Table 2.7: Accuracy (distance to true solution) of steepest descent and conjugate gradient for Fréchet/Karcher mean on $\mathrm{Graff}(k, 10)$.

n	7	8	9	10	11	12	13	14	15
Steepest descent ($\times 10^{-7}$)	4.4	4.8	4.4	4.7	4.7	4.7	4.3	4.7	4.1
Conjugate gradient ($\times 10^{-2}$)	0.36	1.6	1.3	1.3	1.2	1.5	1.5	1.4	1.6

Table 2.8: Accuracy (distance to true solution) of steepest descent and conjugate gradient for Fréchet/Karcher mean on $\mathrm{Graff}(6, n)$.

k	1	2	3	4	5	6	7	8	9
Steepest descent ($\times 10^{-2}$)	4.0	4.6	4.9	5.1	5.1	5.5	5.3	5.1	5.4
Conjugate gradient ($\times 10^{-2}$)	3.6	4.5	4.9	5.0	5.4	5.4	5.3	4.5	12.0

Table 2.9: Elapsed time (in seconds) of steepest descent and conjugate gradient for Fréchet/Karcher mean on $\mathrm{Graff}(k, 10)$.

n	7	8	9	10	11	12	13	14	15
Steepest descent ($\times 10^{-1}$)	3.1	3.0	3.5	3.3	3.7	4.1	3.8	4.1	4.3
Conjugate gradient ($\times 10^{-1}$)	17.0	2.1	2.8	3.1	3.4	3.8	3.9	3.6	3.6

Table 2.10: Elapsed time (in seconds) of steepest descent and conjugate gradient for Fréchet/Karcher mean on $\mathrm{Graff}(6, n)$.

More extensive numerical experiments indicate that steepest descent and conjugate gradient are about equally fast for minimizing (2.8.4), cf. Tables 2.9 and 2.10, but that steepest descent is more accurate by orders of magnitude compared to conjugate gradient, cf. Tables 2.7 and 2.8. Again we believe the reason is that steepest descent is better suited for geodesically convex problems compared to conjugate gradient.

2.9 Conclusion

We introduce the affine Grassmannian, study its algebraic and differential geometric properties and develop concrete systems of global coordinates, computable distances and metrics, natural families of probability densities, and optimization algorithms on $\mathrm{Graff}(k, n)$. We demonstrated the correctness of two of these algorithms through relatively extensive numerical experiments on two nontrivial problems with closed-form solutions. One of our main

goals is to lay the foundations for its systematic use in statistics and machine learning, where estimation problems can often be formulated as optimization problems on $\mathrm{Graff}(k, n)$ although we hope that the material developed here would be useful in other areas as well.

There are germs of ideas that we have introduced with a view towards future work. An example is the very rudimentary "affine Schubert calculus" in Section 2.5 that would allow one to model a mixture of affine subspaces of different dimensions. One of our future goals is to use this framework to aggregate inequidimensional affine subspaces estimated from different datasets in a way that best summarizes them. We believe statistical estimation should take into account the intrinsic geometry of the data, and the deviation from the underlying geometric structures may then be used as a measure of accuracy of the statistical model.

CHAPTER 3

OPTIMIZATION ON FLAG MANIFOLD

3.1 Introduction

A flag in a finite-dimensional vector space $\mathbf{V}$ over $\mathbb{R}$ is an increasing sequence of linear subspaces $\{\mathbf{V}_i\}_{i=1}^d$ of $\mathbf{V}$, satisfying $\mathbf{V}_1 \subsetneq \cdots \subsetneq \mathbf{V}_d \subset \mathbf{V}$. All flags $\{\mathbf{V}_i\}_{i=1}^d$ with fixed dimensions, i.e. $\dim(\mathbf{V}_i) = n_i, i = 1 \cdots d$, form a smooth manifold $\mathrm{Flag}(n_1 \cdots n_d; \mathbf{V})$, that we name flag manifold, a generalization of Grassmannian $\mathrm{Gr}(k, n)$ that parameterizes k-dimensional linear subspaces in $\mathbb{R}^n$, which is equivalent to $\mathrm{Flag}(k; \mathbb{R}^n)$.

In Principal Component Analysis (PCA), we choose a dimension $k < n$ and find a linear subspace of dimension k that minimizes the squared sum of distances from the points their projections onto the subspace, which we call the PCA subspace of dimension k. Flag naturally appears here in the sense that PCA subspace of dimension k is contained in PCA subspace of dimension $k+1$. The flag of PCA subspaces represents a hierarchy of structures that explains an increasing amount of variance among the data set. This prompts us to consider an alternative view of the flag manifold as a mixture of independent subspaces, as the underlying vector space $\mathbf{V}$ can be decomposed as $\bigoplus_{i=0}^d (\mathbf{V}_{i+1} \setminus \mathbf{V}_i)$. if we denote $\mathbf{V}_{d+1} = \mathbf{V}$ and $\mathbf{V}_0 = \emptyset$.

Optimization on manifold has been proposed to solve problems in signal processing and computer vision [73, 72], for instance in matrix completion [75], independent subspace analysis [67] and subspace tracking [74]. On another hand, flag manifold has been used in the analysis of eigenvalue methods [77, 78]. Optimization on manifold is analogous to optimization in Euclidean space, with the exception that instead of adding the descent direction to current iterate in the later setting, we move the iterate along the geodesic by exponential map to ensure the iterates stay on the manifold. The nonlinear structure of manifold adds complication to optimization, motivating us to develop closed-form expressions of corresponding

machinery that can be computed with standard numerical linear algebra in this work. Our hope is to apply this optimization framework to exploit multi-scale representation of data sets to find hidden latent structure.

We will begin by seeing how classical multivariate analysis techniques may be cast as nested subspace-searching problems, i.e., constrained or unconstrained optimization problems on the flag manifold.

Example 3.1.1 (Principal Component Analysis)**.** Let $\overline{x} = \frac{1}{n} X^{\mathsf{T}} \mathbb{1} \in \mathbb{R}^p$ be the sample mean of a data matrix $X \in \mathbb{R}^{n \times p}$ so that $\overline{X} = X - \mathbb{1}\overline{x}^{\mathsf{T}}$ is mean-centered. For $k \leq p$, the kth principal subspace is $\mathrm{im}(Z_k)$, a k-dimensional linear subspace of $\mathbb{R}^p$ such that $Z_k \in \mathbb{R}^{p \times k}$ maximizes $\mathrm{tr}(Z_k^{\mathsf{T}} \overline{X}^{\mathsf{T}} \overline{X} Z_k)$, subject to $Z_k^{\mathsf{T}} Z_k = I_k$. The principal subspace of dimension k $\mathrm{im}(Z_k)$ in $\mathbb{R}^p$ captures the greatest k-dimensional variability in the data $\overline{X}$, and the principal subspace of dimension k is contained in the principal subspace of dimension $k+1$. Hence the nested sequence of subspaces $\{\mathrm{im}(Z_i)\}_{i=1}^p$, which explains an increasing amount of variance among $\overline{X}$, forms a flag in $\mathrm{Flag}(1, \cdots, p; \mathbb{R}^n)$.

Example 3.1.2 (Canonical Correlation Analysis)**.** Let $X = (X_1, \cdots, X_p)^{\mathsf{T}}$ and $Y = (Y_1, \cdots, Y_q)^{\mathsf{T}}$ be vectors of random variables with finite second moments. Denote $r = \min(p, q)$. Let $\Sigma_{XX} = \mathrm{cov}(X, X), \Sigma_{XY} = \mathrm{cov}(X, Y)$, and $\Sigma_{YY} = \mathrm{cov}(Y, Y)$. We seek two projection directions $a_1 \in \mathbb{R}^p$ and $b_1 \in \mathbb{R}^q$ such that $\mathrm{corr}\left(a_1^{\mathsf{T}} X, b_1^{\mathsf{T}} Y\right)$ is maximized. In other words, the optimization problem can be formulated as

$$\max_{a_1 \in \mathbb{R}^p, b_1 \in \mathbb{R}^q} \mathrm{corr}\left(a_1^{\mathsf{T}} X, b_1^{\mathsf{T}} Y\right) = \frac{a_1^{\mathsf{T}} \Sigma_{XY} b_1}{\sqrt{a_1^{\mathsf{T}} \Sigma_{XX} a_1} \sqrt{b_1^{\mathsf{T}} \Sigma_{YY} b_1}}. \tag{3.1.1}$$

The random variables $a_1^{\mathsf{T}} X$ and $b_1^{\mathsf{T}} Y$ are denoted as first pair of canonical covariates, while a_1 and b_1 are called the first pair of canonical loadings. The subsequent canonical loadings can be defined by requiring the k-th canonical covariates are uncorrelated with the previous canonical covariates respectively. In other words, the k-th pair of canonical loadings $a_k \in \mathbb{R}^p$

and $b_k \in \mathbb{R}^q$ can be found by maximizing $\operatorname{corr}\left(a_k^\mathsf{T} X, b_k^\mathsf{T} Y\right)$, subject to $\operatorname{cov}\left(a_k^\mathsf{T} X, a_i^\mathsf{T} X\right) = a_k^\mathsf{T} \Sigma_{XX} a_i = 0$ and $\operatorname{cov}\left(b_k^\mathsf{T} Y, b_i^\mathsf{T} Y\right) = b_k^\mathsf{T} \Sigma_{YY} b_i = 0$ for $i = 1, \cdots, k-1$. Therefore $\{\mathbf{V}_i\}_{i=1}^r$ and $\{\mathbf{W}_i\}_{i=1}^r$, defined by $\mathbf{V}_i = \operatorname{span}(a_1, \cdots, a_i)$ and $\mathbf{W}_i = \operatorname{span}(b_1, \cdots, b_i)$ for $i = 1, \cdots, k$, form two flags in $\operatorname{Flag}(1, \cdots, r; \mathbb{R}^n)$.

Example 3.1.3 (Correspondence Analysis)**.** Let $X \in \mathbb{R}^{n \times p}$ be a data matrix that comes from a contingency table. Let $n_X = \sum_{i=1}^n \sum_{j=1}^p X_{ij}$ be the total number of observations. We compute the row weights $w_n = \frac{1}{n_X} X \mathbb{1}$ and column weights $w_p = \frac{1}{n_X} \mathbb{1}^\mathsf{T} X$, and formulate the matrix $M = \frac{1}{n_X} X - w_n w_p$ that represents the deviation from independence of occurrence of the two outcomes. Correspondence Analysis seeks a generalized singular value decomposition, where the generalized singular vectors can be found by maximizing $\operatorname{tr}(U^\mathsf{T} M V)$ subject to $U^\mathsf{T} \operatorname{diag}(w_n) U = I$ and $V^\mathsf{T} \operatorname{diag}(w_p) V = I$. Hence $\{\mathbf{U}_i\}_{i=1}^m$ and $\{\mathbf{V}_i\}_{i=1}^p$, defined by $\mathbf{U}_i = \operatorname{span}(U_1, \cdots, U_i)$ for $i = 1, \cdots, m$ and $\mathbf{V}_i = \operatorname{span}(V_1, \cdots, V_j)$ for $j = 1, \cdots, p$, form two flags in $\operatorname{Flag}(1, \cdots, m; \mathbb{R}^n)$ and $\operatorname{Flag}(1, \cdots, p; \mathbb{R}^p)$ respectively.

We pursue several objectives in this work. We will begin by reviewing the algebra and geometry behind the optimization framework, in particular the basics of Lie algebra and homogeneous space. Next we proceed to unfold the differential geometry of flag manifold and develop the concrete matrix representations of flag manifolds and the associated geometric quantities. We then move on to illustrate optimization algorithms on flag manifold which only utilize matrix multiplication and matrix exponential, and provide numerical experiments to demonstrate the convergence of these algorithms. Optimization algorithms on flag manfifolds are considered in [67] to solve the invariant subspace analysis (ISA), which is a generalization of independent component analysis (ICA). However, only tangent spaces and gradient functions are explicitly computed in [67], using the homogeneous space structure of flag manifolds. No actual algorithm is given in [67], although authors provide some numerical experiments. The main contribution of this paper is to calculate necessary ingredients for optimization algorithms on flag manifolds in full details, from different perspectives. All our calculations are validated rigorously and we also provide ready-to-use formulae and al-

gorithms, which are easily accessible to applied mathematicians and other practitioners. For readers' convenience, we summarize important formulae and algorithms in this paper below:

- parametrization of a point: Proposition 3.3.12;
- tangent vector: Proposition 3.3.6 and Proposition 3.3.13;
- geodesic: Proposition 3.3.7, Proposition 3.3.9 and Proposition 3.3.14;
- geodesic distance: Proposition 3.3.8;
- parallel transport: Proposition 3.3.11 and Proposition **??**;
- gradient: Proposition 3.3.16;
- Hessian: Proposition 3.3.17;
- steepest descent algorithm: Algorithm 3.4.1;
- conjugate gradient algorithm: Algorithm 3.4.2.

3.2 Preliminaries

The purpose of this section is to provide a self-contained review of the differential geometry of homogeneous spaces. Readers can find details in standard textbooks in differential geometry, for example, [56, 57, 58].

3.2.1 Lie groups and Lie algebras

Let M be a smooth manifold and let T^*M be its cotangent bundle. A *Riemannian metric* on M is a smooth section $g : M \to T^*M \otimes T^*M$ such that $g_x \coloneqq g(x) \in T^*_xM \otimes T^*_xM$ is a non-degenerate symmetric bilinear form on the tangent space T_xM for every $x \in M$. Intuitively, a Riemannian metric gives an inner product on T_xM for every $x \in M$ and it varies smoothly with respect to $x \in M$. Let G be a group and let $m : G \times G \to G$ be the

multiplication map $m(a_1, a_2) = a_1 a_2$ and let $\mathrm{inv}\, G :\to G$ be the inversion map $\mathrm{inv}(a) = a^{-1}$. The group G is called a *Lie group* if G is also a smooth manifold and group operations m and inv are smooth. The tangent space $\mathfrak{g}$ of G at the identity $e \in G$ is a *Lie algebra*.[1] For example, if G is the orthogonal group $\mathrm{O}(n)$ consisting of all $n \times n$ real orthogonal matrices, then its Lie algebra is $\mathfrak{so}(n)$, consisting of all $n \times n$ real skew-symmetric matrices.

3.2.2 Homogeneous spaces

We will review some basic properties of homogeneous spaces.

Definition 3.2.1. Let G be a Lie group, acting on a smooth manifold M via $\varphi : G \times M \to M$. If the action φ is smooth and transitive,[2] then M is called a homogeneous space of the Lie group G.

Let $x \in M$ be a point, we call the subgroup $G_x = \{a \in G : \varphi(a, x) = x\}$ the *isotropy group* of x. We denote by G/G_x the quotient of G by G_x and we denote by $[a] \in G/G_x$ the coset (or equivalence class) represented by $a \in G$. Since G acts on M transitively, we see that there exists a one to one correspondence F between G/G_x and M for any $x \in M$, defined by

$$F : G/G_x \to M, \quad F([a]) = \varphi(a, x).$$

In fact, this point-set correspondence turns out to be a diffeomorphism between two smooth manifolds. This is the content of the following theorem:

Theorem 3.2.2. *[58, Theorem 9.2, 9.3] Let G be a Lie group acting on a smooth manifold M and let x be a point in M. There exists a unique smooth structure on G/G_x such that the action*

$$\psi : G \times G/G_x \to G, \quad \psi(a, [a']) = [aa']$$

1. A Lie algebra $\mathfrak{g}$ is a vector space equipped with a bilinear map (Lie bracket) $[\cdot, \cdot] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ satisfying $[X, Y] = -[Y, X]$ (skew-symmetry) and $[X, [Y, Z]] + [Z, [X, Y]] + [Y, [Z, X]] = 0$ (Jacobi identity).

2. transitive means that for every pair of elements $m, m' \in M$, there exits some $a \in G$ such that $\varphi(a, m) = m'$.

is smooth. Moreover, the map $F : G/G_x \to M$ *sending* $[a]$ *to* $\varphi(a, x)$ *is a* G*-equivariant diffeomorphism, i.e.,* F *is a diffeomorphism such that* $F(\psi(a, [a'])) = \varphi(a, F([a']))$.

The Grassmannian $\mathrm{Gr}(k, n)$ of k-dimensional subspace of $\mathbb{R}^n$ is the most important example of homogeneous spaces. Indeed, $\mathrm{O}(n)$ acts transitively on $\mathrm{Gr}(k, n)$ and for a given k-dimensional subspace $\mathbf{W}$ of $\mathbf{V}$, its isotropy group is isomorphic to $\mathrm{O}(k) \times \mathrm{O}(n-k)$ hence we have $\mathrm{Gr}(k, n) \simeq \mathrm{O}(n)/(O(k) \times O(n-k))$.

Let G be a Lie group and let M be a homogeneous space of G with action $\varphi : G \times M \to M$. By Theorem 3.2.2 we may identify M with G/G_x for any $x \in M$. For simplicity, we fix $x \in M$ and denote by H the isotropy group G_x of x. We define a map $L_a(y) = \varphi(a, y) \in M$. In particular, if $a \in H$ then $L_a(x) = x$ and hence we have a linear isomorphism

$$dL_a|_x : T_xM \to T_xM.$$

Let $g : M \to T^*M \otimes T^*M$ be a Riemannian metric on M. We say that g is a G*-invariant metric* if for every $y \in M$ and $a \in G$,

$$g_{L_a(y)}(dL_a|_y X, dL_a|_y Y) = g_y(X, Y), \quad X, Y \in T_yM.$$

Moreover, $M = G/H$ and hence $T_xM = \mathfrak{g}/\mathfrak{h}$ where $\mathfrak{g}, \mathfrak{h}$ are Lie algebras of G and H respectively. This implies that we have a representation

$$\mathrm{Ad}_H : H \to \mathrm{GL}(\mathfrak{g}/\mathfrak{h}), \quad \mathrm{Ad}_H(a)(X) = dL_a|_x(X), \quad X \in \mathfrak{g}/\mathfrak{h}.$$

An inner product η on the vector space $\mathfrak{g}/\mathfrak{h}$ is said to be Ad_H*-invariant* if for every $a \in H$,

$$\eta(\mathrm{Ad}_H(a)(X), \mathrm{Ad}_H(a)(Y)) = \eta(X, Y), \quad X, Y \in \mathfrak{g}/\mathfrak{h}.$$

Proposition 3.2.3. *[63, Proposition* 3.16*] Let* G *be a connected Lie group and let* H *be its*

closed Lie subgroup. Suppose that $\mathfrak{g}$ *and* $\mathfrak{h}$ *are Lie algebras of* G *and* H*, respectively. If there exists a subspace* $\mathfrak{m}$ *of* $\mathfrak{g}$ *such that* $\mathfrak{g} = \mathfrak{m} \oplus \mathfrak{h}$ *and* $\mathrm{Ad}_H(\mathfrak{m}) \subseteq \mathfrak{m}$*, then there is a one to one correspondence between* G*-invariant metrics on* $M = G/H$ *and* Ad_H*-invariant inner product on the vector space* $\mathfrak{m}$*. In particular, if* G *is compact, then* G *admits a bi-invariant metric.*

Proposition 3.2.3 says that once $\mathfrak{h} \subseteq \mathfrak{g}$ admits a complement $\mathfrak{m}$, we may induce a G-invariant metric g on M by some inner product on $\mathfrak{m}$. Moreover, we may identify T_xM with $\mathfrak{m}$. This implies that the metric g on M is essentially determined by g_x.

Let $M = G/H$ be a homogeneous space of G. If M has a Riemannian metric g such that every geodesic in M is an orbit of a one-parameter subgroup of G, then we say that (M, g) is a *geodesic orbit space (or GO-space)*. If M is a GO-space, then by definition, all its geodesics are simply orbits of one-parameter subgroups of G.

Theorem 3.2.4. *[61, Theorem* 1.3.5*] If* G *is a matrix Lie group equipped with a bi-invariant metric, every one-parameter subgroup* $\gamma(t)$ *of* G *is of the form*

$$\gamma(t) = \exp(ta) \coloneqq \sum_{k=0}^{\infty} \frac{t^k a^k}{k!}, a \in \mathfrak{g}.$$

In particular, if $G = \mathrm{SO}(n)$ then we see that every one-parameter subgroup is of the form

$$\gamma(t) = \exp(ta), a \in \mathfrak{so}(n).$$

Theorem 3.2.5. *[62] If* G *is a compact Lie group with a bi-invariant metric* g*, then* $M = G/H$ *together with the metric* $\tilde{g}$ *induced by* g *is a GO-space.*

We will also need the famous Hopf-Rinow theorem to calculate the Riemannian distance on flag manifolds.

Theorem 3.2.6 (Hopf-Rinow Theorem)**.** *[63, Theorem* 1.8*] Let* (M, g) *be a connected Riemannian manifold. Then the following statements are equivalent:*

- *The closed and bounded subsets of M are compact;*
- *M is a complete metric space;*
- *M is geodesically complete, i.e., the exponential map $\exp_x : T_xM \to M$ is defined on the whole T_xM for all $x \in M$.*

Furthermore, any one the above three implies: Any two points x, y on M can be connected by a distance minimizing geodesic on M.

3.3 Differential geometry of flag manifolds

3.3.1 Definitions and basic properties

Let n be a positive integer and let $\mathbf{V}$ be an n-dimensional vector space over $\mathbb{R}$. We first define flags in $\mathbf{V}$ and flag manifolds.

Definition 3.3.1. Let $0 < n_1 < \cdots < n_d < n$ be an increasing sequence of d positive integers. A flag of type $(n_1, \ldots, n_d)$ in $\mathbf{V}$ is a sequence of subspaces

$$\mathbf{V}_1 \subsetneq \cdots \subsetneq \mathbf{V}_d, \quad \dim \mathbf{V}_i = n_i, i = 1, \ldots, d.$$

We denote the set of all flags in $\mathbf{V}$ of type $(n_1, \ldots, n_d)$ by $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$. If $\mathbf{V}$ is understood, we simply denote $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ by $\mathrm{Flag}(n_1, \ldots, n_d)$.

Flag manifolds are direct generalizations of Grassmannian manifolds. For example, $\mathrm{Flag}(k)$ is simply the set of all k-dimensional subspaces of $\mathbf{V}$, which is the Grassmannian $\mathrm{Gr}(k, n)$. Another extreme case is when $d = n - 1$ and $n_i = i, i = 1, \ldots, n - 1$, where $\mathrm{Flag}(1, \ldots, n - 1; \mathbf{V})$ is consisting of all complete flags:

$$\mathbf{V}_1 \subsetneq \cdots \subsetneq \mathbf{V}_{n-1}, \quad \dim \mathbf{V}_i = i.$$

It turns out that, like Grassmannian manifolds, $\mathrm{Flag}(n_1, \ldots, n_d)$ is not merely a set. Indeed, it has various geometric structures. We summarize some of them in the following proposition.

Proposition 3.3.2. *[59, 60] Let $0 < n_1 < \cdots < n_d < n$ be an increasing sequence of d positive integers and let $\mathbf{V}$ be an n-dimensional vector space over $\mathbb{R}$. The flag manifold $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ is*

- *a connected, compact smooth manifold;*
- *a homogeneous space:*

$$\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V}) \simeq \mathrm{O}(n)/(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n_d - n_{d-1}) \times \mathrm{O}(n - n_d)).$$

Proposition 3.3.3. *The flag manifold $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ is*

- *a closed sub-manifold of $\mathrm{Gr}(n_1, n) \times \cdots \mathrm{Gr}(n_d, n)$;*
- *a closed sub-manifold of $\mathrm{Gr}(n_1, n) \times \mathrm{Gr}(n_2 - n_1, n) \times \cdots \times \mathrm{Gr}(n_d - n_{d-1}, n) \times \mathrm{Gr}(n - n_d, n)$;*
- *a fiber bundle on $\mathrm{Gr}(n_d, n)$ whose fiber over $\mathbf{W} \in \mathrm{Gr}(n_d, n)$ is $\mathrm{Flag}(n_1, \ldots, n_{d-1}; \mathbf{W})$.*

Proof. By definition an element F in $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ is a sequence of linear subspaces of $\mathbf{V}$:

$$F : \mathbf{V}_1 \subsetneq \mathbf{V}_2 \subsetneq \cdots \subsetneq \mathbf{V}_d, \quad \dim \mathbf{V}_i = n_i, i = 1, \ldots, d.$$

This gives us a map

$$j : \mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V}) \to \mathrm{Gr}(n_1, n) \times \cdots \times \mathrm{Gr}(n_d, n), \quad j(F) = (\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_d).$$

It is easy to verify that j is an embedding. Moreover, the image of j is a closed subset of $\mathrm{Gr}(n_1, n) \times \cdots \times \mathrm{Gr}(n_d, n)$. Indeed, if

$$(\mathbf{V}_1 \times \mathbf{V}_2 \times \cdots \times \mathbf{V}_d) \notin j(\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})),$$

then there exists some $i = 1, \ldots, d-1$ such that $\mathbf{V}_i \not\subseteq \mathbf{V}_{i+1}$. This implies that if $\mathbf{V}'_i \in \mathrm{Gr}(n_i, n)$ and $\mathbf{V}'_{i+1} \in \mathrm{Gr}(n_{i+1}, n)$ are in a small neighborhood of $\mathbf{V}_i$ and $\mathbf{V}_{i+1}$ respectively, then $\mathbf{V}'_i \not\subseteq \mathbf{V}'_{i+1}$. Next, we choose and fix an inner product on $\mathbf{V}$. Let $\mathbf{W}$ be a subspace of $\mathbf{V}$. For a subspace $\mathbf{U}$ of $\mathbf{W}$ we denote by $\mathbf{U}^{\perp}_{\mathbf{W}}$ the orthogonal complement of $\mathbf{W}$ in $\mathbf{V}$. If $F : \mathbf{V}_1 \subsetneq \mathbf{V}_2 \subsetneq \cdots \subsetneq \mathbf{V}_d, \quad \dim \mathbf{V}_i = n_i, i = 1, \ldots, d$ is an element in $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$, then for each $i = 2, \ldots, d$ we have $\mathbf{V}_i^{\perp}{}_{\mathbf{V}_{i+1}}$ and hence we obtain a map

$$i : \mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V}) \to \mathrm{Gr}(n_1, n) \times \mathrm{Gr}(n_2 - n_1, n) \times \cdots \times \mathrm{Gr}(n_d - n_{d-1}, n) \times \mathrm{Gr}(n - n_d, n)$$
$$\mathbf{V}_1 \subsetneq \mathbf{V}_2 \subsetneq \cdots \subsetneq \mathbf{V}_d \mapsto (\mathbf{V}_1, \mathbf{V}_1^{\perp}{}_{\mathbf{V}_2}, \ldots, \mathbf{V}_{d-1}^{\perp}{}_{\mathbf{V}_d}, \mathbf{V}_d^{\perp}{}_{\mathbf{V}}).$$

The map i is clearly an embedding. The closedness of the image of i in $\mathrm{Gr}(n_1, n) \times \mathrm{Gr}(n_2 - n_1, n) \times \cdots \times \mathrm{Gr}(n_d - n_{d-1}, n) \times \mathrm{Gr}(n - n_d, n)$ follows from a similar argument for that of the image of j. Lastly, we consider the map

$$\rho : \mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V}) \to \mathrm{Gr}(n_d, n), \quad \rho(\mathbf{V}_1 \subsetneq \cdots \subsetneq \mathbf{V}_d) = \mathbf{V}_d.$$

It is clear that ρ is surjective, smooth and $\rho^{-1}(\mathbf{V}_d)$ consists of flags of the form

$$\mathbf{V}'_1 \subsetneq \mathbf{V}'_2 \subsetneq \cdots \subsetneq \mathbf{V}'_{d-1} \subsetneq \mathbf{V}_d, \quad \dim \mathbf{V}'_i = n_i, i = 1, \ldots, d-1.$$

This implies that the fiber $\rho^{-1}(\mathbf{V}_d) \simeq \mathrm{Flag}(n_1, \ldots, n_{d-1}; \mathbf{V}_d)$. □

3.3.2 Flag manifolds as homogeneous spaces

Because of Proposition 3.3.2, from now on, we call the set $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ the *flag manifold* of type $(n_1, \ldots, n_d)$. In particular, if $d = 1$, then $\mathrm{Gr}(k, n) = \mathrm{Flag}(k; \mathbf{V}) \simeq \mathrm{O}(n)/(\mathrm{O}(k) \times \mathrm{O}(n-k))$.

Proposition 3.3.4. *Let $0 < n_1 < \cdots < n_d < n$ be an increasing sequence of d positive integers between 0 and n and let $\mathbf{V}$ be an n-dimensional vector space over $\mathbb{R}$. The Flag*

manifold $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ *is homeomorphic to the homogeneous space*

$$\mathrm{SO}(n)/S(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n_d - n_{d-1}) \times \mathrm{O}(n - n_d)).$$

Here $S(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n_d - n_{d-1}) \times \mathrm{O}(n - n_d))$ *is the subgroup of* $\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n_d - n_{d-1}) \times \mathrm{O}(n - n_d)$ *consisting of all block diagonal matrices*

$$\begin{bmatrix} Q_1 & 0 & \cdots & 0 \\ 0 & Q_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_{d+1} \end{bmatrix},$$

where $Q_i \in \mathrm{O}(n_i), i = 1, \ldots, d+1$ *and* $\prod_{i=1}^{d+1} \det(Q_i) = 1$.

Proof. By Proposition 3.3.2, we see that

$$\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V}) \simeq \mathrm{O}(n)/(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n - n_d)).$$

□

Let Q be an element in $\mathrm{SO}(n)$. We denote the equivalence class represented by Q by

$$[Q] = \left\{ Q \begin{bmatrix} Q_1 & 0 & \cdots & 0 \\ 0 & Q_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_{d+1} \end{bmatrix} : Q_i \in \mathrm{O}(n_i - n_{i-1}), i = 1, \ldots, d+1, \prod_{i=1}^{d+1} \det Q_i = 1 \right\}. \tag{3.3.1}$$

Noticing that the Lie algebra of $S(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n_d - n_{d-1}) \times \mathrm{O}(n - n_d))$ is simply $\mathfrak{so}(n_1) \times \mathfrak{so}(n_2 - n_1) \times \cdots \times \mathfrak{so}(n - n_d)$. We may embed the Lie algebra $\mathfrak{so}(n_1) \times$

$\mathfrak{so}(n_2 - n_1) \times \cdots \times \mathfrak{so}(n - n_d)$ via the Lie algebra homomorphism

$$i : \mathfrak{so}(n_1) \times \mathfrak{so}(n_2 - n_1) \times \cdots \times \mathfrak{so}(n - n_d) \to \mathfrak{so}(n)$$

defined by

$$i(A_1, \ldots, A_{d+1}) = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{d+1} \end{bmatrix} \in \mathfrak{so}(n).$$

We denote by $\mathfrak{h}$ the image $i(\mathfrak{so}(n_1) \times \mathfrak{so}(n_2 - n_1) \times \cdots \times \mathfrak{so}(n - n_d))$. A natural complement $\mathfrak{m}$ of $\mathfrak{h}$ in $\mathfrak{g} \coloneqq \mathfrak{so}(n)$ is the set of matrices of the form

$$\begin{bmatrix} 0 & B_{1,2} & \cdots & B_{1,d+1} \\ -B_{1,2}^{\mathsf{T}} & 0 & \cdots & B_{2,d+1} \\ \vdots & \vdots & \ddots & \vdots \\ -B_{1,d+1}^{\mathsf{T}} & -B_{2,d+1}^{\mathsf{T}} & \cdots & 0 \end{bmatrix} \in \mathfrak{so}(n), \quad B_{ij} \in \mathbb{R}^{(n_i - n_{i-1}) \times (n_j - n_{j-1})}, 1 \leq i < j \leq d+1.$$

Proposition 3.3.5. *Let $\mathfrak{g}, \mathfrak{h}$ and $\mathfrak{m}$ be as above. The subspace $\mathfrak{m}$ is Ad_H-invariant, i.e., $Ad(a)(X) \in \mathfrak{m}$ for every $a \in H$ and $X \in \mathfrak{m}$.*

Proof. It remains to show that the commutator $Ad(a)(X) \in \mathfrak{m}$ whenever $a \in H$ and $X \in \mathfrak{m}$. For simplicity, we only calculate the case for $d = 2$. The general case is proved in a similar way. First we write

$$a = \begin{bmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ 0 & 0 & A_3 \end{bmatrix}, \quad A_i \in \mathrm{O}(n_i - n_{i-1}), i = 1, 2, 3,$$

and

$$X = \begin{bmatrix} 0 & B_{1,2} & B_{1,3} \\ -B_{1,2}^{\mathsf{T}} & 0 & B_{2,3} \\ -B_{1,3}^{\mathsf{T}} & -B_{2,3}^{\mathsf{T}} & 0 \end{bmatrix}, \quad B_{ij} \in \mathbb{R}^{(n_i - n_{i-1}) \times (n_j - n_{j-1})}, 1 \leq i < j \leq 3.$$

Then we have $Ad(a)(X) = aXa^{-1} = aXa^{\mathsf{T}}$ because a is an orthogonal matrix. Lastly we have

$$\begin{aligned} aXa^{\mathsf{T}} &= \begin{bmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ 0 & 0 & A_3 \end{bmatrix} \begin{bmatrix} 0 & B_{1,2} & B_{1,3} \\ -B_{1,2}^{\mathsf{T}} & 0 & B_{2,3} \\ -B_{1,3}^{\mathsf{T}} & -B_{2,3}^{\mathsf{T}} & 0 \end{bmatrix} \begin{bmatrix} A_1^{\mathsf{T}} & 0 & 0 \\ 0 & A_2^{\mathsf{T}} & 0 \\ 0 & 0 & A_3^{\mathsf{T}} \end{bmatrix} \\ &= \begin{bmatrix} 0 & A_1 B_{1,2} & A_1 B_{1,3} \\ -A_2 B_{1,2}^{\mathsf{T}} & 0 & A_2 B_{2,3} \\ -A_3 B_{1,3}^{\mathsf{T}} & -A_3 B_{2,3}^{\mathsf{T}} & 0 \end{bmatrix} \begin{bmatrix} A_1^{\mathsf{T}} & 0 & 0 \\ 0 & A_2^{\mathsf{T}} & 0 \\ 0 & 0 & A_3^{\mathsf{T}} \end{bmatrix} \\ &= \begin{bmatrix} 0 & A_1 B_{1,2} A_2^{\mathsf{T}} & A_1 B_{1,3} A_3^{\mathsf{T}} \\ -A_2 B_{1,2}^{\mathsf{T}} A_1^{\mathsf{T}} & 0 & A_2 B_{2,3} A_3^{\mathsf{T}} \\ -A_3 B_{1,3}^{\mathsf{T}} A_1^{\mathsf{T}} & -A_3 B_{2,3}^{\mathsf{T}} A_2^{\mathsf{T}} & 0 \end{bmatrix} \in \mathfrak{m} \end{aligned}$$

and this completes the proof. □

Proposition 3.3.6 (Tangent spaces I)**.** *Let $[Q]$ be the point in $M \coloneqq \mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ represented by $Q \in \mathrm{O}(n)$. The tangent space $T_{[Q]}M$ of M is isomorphic to the vector space consisting of matrices of the form*

$$X = Q \begin{bmatrix} 0 & B_{1,2} & \cdots & B_{1,d+1} \\ -B_{1,2}^{\mathsf{T}} & 0 & \cdots & B_{2,d+1} \\ \vdots & \vdots & \ddots & \vdots \\ -B_{1,d+1}^{\mathsf{T}} & -B_{2,d+1}^{\mathsf{T}} & \cdots & 0 \end{bmatrix}, B_{i,j} \in \mathbb{R}^{(n_i - n_{i-1}) \times (n_j - n_{j-1})}, 1 \leq i < j \leq d+1.$$

In particular, the dimension of $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ *is*

$$\dim \mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V}) = \sum_{1 \leq i < j \leq d+1} (n_i - n_{i-1})(n_j - n_{j-1}).$$

Proof. When $Q = \mathrm{Id}_n$, the identity matrix, this is straightforward from the identification $\mathfrak{g}/\mathfrak{h} \simeq \mathfrak{m}$. The general case follows from the fact that the left translation $L_Q : M \to M$ by Q is a diffeomorphism, which induces an isomorphism $dL_Q|_{[\mathrm{Id}]}$ between $T_{[\mathrm{Id}]}M$ and $T_{[Q]}M$. By definition, $dL_Q|_{[\mathrm{Id}]}(X) = QX, X \in T_{[\mathrm{Id}]}M$. □

There are several ways to equip $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ a Riemannian metric. We will concentrate on the one induced by the unique (up to scalar) bi-invariant metric on $G \coloneqq \mathrm{SO}(n)$. [3] First we define an inner product $(\cdot, \cdot)$ on $\mathfrak{so}(n)$ by

$$(A, B) = \frac{1}{2} \operatorname{tr}(A^{\mathsf{T}} B), A, B \in \mathfrak{so}(n).$$

Since $\mathfrak{m}$ is a subspace of $\mathfrak{so}(n)$, the restriction $(\cdot, \cdot)|_{\mathfrak{m}}$ is an inner product on $\mathfrak{m}$. We denote by H the subgroup $S(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n_d - n_{d-1}) \times \mathrm{O}(n - n_d))$ of G. It is easy to verify that $(\cdot, \cdot)|_{\mathfrak{m}}$ is Ad_H-invariant and this together with Proposition 3.3.5 and Proposition 3.2.3 implies that $(\cdot, \cdot)|_{\mathfrak{m}}$ uniquely determines a G-invariant metric g on $G/H = \mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$.[4] Indeed, the metric g is simply given by the formula

$$g(X, Y) = \frac{1}{2} \operatorname{tr}(X^{\mathsf{T}} Y), \quad X, Y \in T_{[Q]} \mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V}).$$

3. There is a one to one correspondence between bi-invariant metric on G and Ad_G-invariant metric on $\mathfrak{g}$ (see for e.g., [68]). We may consider the representation $\mathrm{Ad} : \mathrm{SO}(n) \to GL(\mathfrak{so}(n))$. If $n \neq 2, 4$, $\mathfrak{so}(n)$ is a simple Lie Algebra. This implies that there is a unique (up to scalars) Ad_G-invariant non-degenerate bilinear form on $\mathfrak{so}(n)$. Indeed, this unique bilinear form is a negative multiple of the Killing form of $\mathfrak{so}(n)$. Moreover, it is straightforward to verify that the metric g is the pull-back of $\sum_{i=1}^{d} g_i$ via the embedding of $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ into $\mathrm{Gr}(n_1, n) \times \cdots \mathrm{Gr}(n_d, n)$ described in Proposition 3.3.3, where g_i is the standard metric on $\mathrm{Gr}(n_i, n), i = 1, \ldots, d$.

4. If G is a compact Lie group, then G admits a bi-invariant metric and this metric induces a G-invariant metric g on $M = G/H$ for any closed subgroup $H \subseteq G$. The metric g is called the canonical metric on M and (M, g) is called a normal homogeneous space.

Equivalently, if we write

$$X = Q\begin{bmatrix} 0 & B_{1,2} & \cdots & B_{1,d+1} \\ -B_{1,2}^{\mathsf{T}} & 0 & \cdots & B_{2,d+1} \\ \vdots & \vdots & \ddots & \vdots \\ -B_{1,d+1}^{\mathsf{T}} & -B_{2,d+1}^{\mathsf{T}} & \cdots & 0 \end{bmatrix}, \quad \text{and} \quad Y = Q\begin{bmatrix} 0 & C_{1,2} & \cdots & C_{1,d+1} \\ -C_{1,2}^{\mathsf{T}} & 0 & \cdots & C_{2,d+1} \\ \vdots & \vdots & \ddots & \vdots \\ -C_{1,d+1}^{\mathsf{T}} & -C_{2,d+1}^{\mathsf{T}} & \cdots & 0 \end{bmatrix},$$

where $B_{ij}, C_{ij} \in \mathbb{R}^{(n_i - n_{i-1}) \times (n_j - n_{j-1})}, 1 \leq i < j \leq d$, then the metric g is simply

$$g(X, Y) = \sum_{1 \leq i < j \leq d+1} \operatorname{tr}(B_{ij}^{\mathsf{T}} C_{ij}).$$

This metric coincides with canonical metrics on Stiefel manifold ($d = 0$) and Grassmannian manifold ($d = 1$) introduced in [65]. According to Theorem 3.2.5, we see that $(\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V}), g)$ is a GO-space.[5] In particular, we have

Proposition 3.3.7 (Geodesics I)**.** *Let* $[Q]$ *be an element in* $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V}) = \mathrm{O}(n)/(\mathrm{O}(n_1) \times \cdots \times \mathrm{O}(n - n_d))$. *Every geodesic on the Riemannian manifold* $(\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V}), g)$ *passing through* $[Q]$ *is of the form*

$$[Q(t)] = \left\{ Q(t) \begin{bmatrix} Q_1 & 0 & \cdots & 0 \\ 0 & Q_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_{d+1} \end{bmatrix} : Q_i \in \mathrm{O}(n_i - n_{i-1}), i = 1, \ldots, d+1, \quad \prod_{i=1}^{d+1} \det(Q_i) = 1 \right\},$$

5. In fact, the metric g is the only choice for us to make $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ a GO-space [69].

where $Q(t) = Q\exp(tB)$,

$$B = \begin{bmatrix} 0 & B_{1,2} & \cdots & B_{1,d+1} \\ -B_{1,2}^{\mathsf{T}} & 0 & \cdots & B_{2,d+1} \\ \vdots & \vdots & \ddots & \vdots \\ -B_{1,d+1}^{\mathsf{T}} & -B_{2,d+1}^{\mathsf{T}} & \cdots & 0 \end{bmatrix}, B_{ij} \in \mathbb{R}^{(n_i - n_{i-1})\times(n_j - n_{j-1})}, 1 \leq i < j \leq d+1. \tag{3.3.2}$$

Let $Q(t)$ be a geodesic in $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$, then by Proposition 3.3.7, we are able to compute the arc-length of $[Q(t)]$.

Proposition 3.3.8 (Arc-length of geodesics)**.** *Let* $\gamma(t) = [Q(t)]$ *and* B *be as in Proposition 3.3.9. The arc-length* $\|\gamma(t)\|$ *of a geodesic* $\gamma(t)$ *is*

$$\int_0^t \sqrt{g(\gamma'(x), \gamma'(x))}dx = t\sqrt{\sum_{1\leq i<j\leq d+1} \mathrm{tr}(B_{ij}^{\mathsf{T}} B_{ij})} = t\sqrt{\sum_{1\leq i<j\leq d+1} s_{ij}} = \frac{t}{2}\,\mathrm{tr}(B^{\mathsf{T}}B),$$

where s_{ij} *is the square sum of singular values of* $B_{ij}, 1 \leq i < j \leq d+1$.

Proposition 3.3.9 (Geodesics II)**.** *Let* $\gamma(t)$ *be a geodesic in* $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ $= \mathrm{O}(n)/(\mathrm{O}(n_1) \times \cdots \times \mathrm{O}(n - n_d))$ *and let* $Q \in \mathrm{O}(n)$ *representing* $\gamma(0)$ *and* $Q'(0) = H$. *Suppose that* $Q^{\mathsf{T}}H = VDV^{\mathsf{T}}$ *with* $V \in \mathrm{O}(n)$ *and*

$$D = \mathrm{diag}\left\{ \begin{bmatrix} 0 & -\lambda_1 \\ \lambda_1 & 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 & -\lambda_r \\ \lambda_r & 0 \end{bmatrix}, 0_{n-2r} \right\}, \tag{3.3.3}$$

where $2r$ *is the rank of the matrix* $Q^{\mathsf{T}}H$ *and* $\lambda_1, \ldots, \lambda_r$ *are positive real numbers. Then* $\gamma(t) = [U\Sigma(t)V^{\mathsf{T}}]$, *where* $U = QV \in \mathrm{O}(n)$ *and*

$$\Sigma(t) = \mathrm{diag}\left\{ \begin{bmatrix} \cos t\lambda_1 & -\sin t\lambda_1 \\ \sin t\lambda_1 & \cos t\lambda_1 \end{bmatrix}, \ldots, \begin{bmatrix} \cos t\lambda_r & -\sin t\lambda_r \\ \sin t\lambda_r & \cos t\lambda_r \end{bmatrix}, I_{n-2r} \right\}. \tag{3.3.4}$$

Proof. Since $\gamma(t)$ is a geodesic in $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$, we may write it as $\gamma(t) = [Q \exp(tB)]$ for some $B \in \mathfrak{so}(n)$ and $Q \in \mathrm{O}(n)$ representing $\gamma(0)$, by Proposition 3.3.7. Hence we have $H = Q'(0) = QB$ and $Q^{\mathsf{T}}H = B$. Since B is skew-symmetric and hence is a normal matrix, by the Spectral theorem [66, Theorem 7.25], we see that B can be decomposed as $B = VDV^{\mathsf{T}}$, with $V \in \mathrm{O}(n)$ and

$$D = \mathrm{diag}\left\{ \begin{bmatrix} 0 & -\lambda_1 \\ \lambda_1 & 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 & -\lambda_r \\ \lambda_r & 0 \end{bmatrix}, 0_{n-2r} \right\},$$

where $2r$ is the rank of the matrix $Q^{\mathsf{T}}H$ and $\lambda_1, \ldots, \lambda_r$ are positive reals numbers.[6] A direct computation shows that we have a decomposition

$$Q \exp(tB) = U\Sigma(t)V^{\mathsf{T}},$$

where $U = QV$ and $\Sigma(t) \in \mathrm{O}(n)$ is the required block diagonal matrix. □

Proposition 3.3.10 (Riemannian distance on flag manifolds)**.** *Suppose that Q and P are $n \times n$ orthogonal matrices. The Riemannian distance (w.r.t. the metric g) between $[Q]$ and $[P]$ is*

$$d([Q], [P]) = \sqrt{\sum_{i=1}^{r} \lambda_i^2},$$

where $\lambda_1, \ldots, \lambda_r$ are positive real numbers such that

$$Q^{\mathsf{T}}P = V\Sigma V^{\mathsf{T}},$$

6. In fact, positive real numbers $\lambda_1, \lambda_1, \lambda_2, \lambda_2, \ldots, \lambda_r, \lambda_r$ are singular values of the matrix B.

where $V \in \mathrm{O}(n)$ *and*

$$\Sigma = \mathrm{diag}\left\{ \begin{bmatrix} \cos\lambda_1 & -\sin\lambda_1 \\ \sin\lambda_1 & \cos\lambda_1 \end{bmatrix}, \ldots, \begin{bmatrix} \cos\lambda_r & -\sin\lambda_r \\ \sin\lambda_r & \cos\lambda_r \end{bmatrix}, 0_{n-2r} \right\}.$$

Proof. As $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ can be embedded as a closed sub-manifold of $\mathrm{Gr}(n_1, n) \times \cdots \mathrm{Gr}(n_d, n)$, its closed and bounded sub-manifolds must be compact. By Theorem 3.2.6, there is a distance minimizing geodesic $[Q\exp(tB)]$ connecting $[Q]$ and $[P]$. By Proposition 3.3.8, we may obtain the formula

$$d([Q],[P]) = \sqrt{\sum_1^r \lambda_i^2},$$

where $\lambda_1, \lambda_1, \ldots, \lambda_r, \lambda_r$ are singular values of B. Lastly, according to Proposition 3.3.9, we see that we have the decomposition $Q^\mathsf{T}P = V\Sigma V^\mathsf{T}$ for some $V \in \mathrm{O}(n)$. □

For $B \in \mathfrak{m}$, we consider the map $\varphi_B : \mathfrak{m} \to \mathfrak{m}$ defined by $\varphi_B(X) = \frac{1}{2}[B, X]_m$, the half of the projection of the $[B, X] \in \mathfrak{g}$ onto $\mathfrak{m}$. For example, if $d = 2$ and

$$B = \begin{bmatrix} 0 & B_{12} & B_{13} \\ -B_{12}^\mathsf{T} & 0 & B_{23} \\ -B_{13}^\mathsf{T} & -B_{23}^\mathsf{T} & 0 \end{bmatrix} \in \mathfrak{m}, \quad X = \begin{bmatrix} 0 & X_{12} & X_{13} \\ -X_{12}^\mathsf{T} & 0 & X_{23} \\ -X_{13}^\mathsf{T} & -X_{23}^\mathsf{T} & 0 \end{bmatrix} \in \mathfrak{m},$$

where $B_{ij}, X_{ij} \in \mathbb{R}^{(n_i - n_{i-1}) \times (n_j - n_{j-1})}, 1 \le i < j \le 3$, then

$$\varphi_B(X) = \begin{bmatrix} 0 & -B_{12}X_{23}^\mathsf{T} + X_{12}B_{23}^\mathsf{T} & B_{11}X_{23} - X_{11}B_{23} \\ X_{23}B_{12}^\mathsf{T} - B_{23}X_{12}^\mathsf{T} & 0 & -B_{11}X_{12}^\mathsf{T} + X_{11}B_{12}^\mathsf{T} \\ -X_{23}^\mathsf{T}B_{11}^\mathsf{T} + B_{23}^\mathsf{T}X_{11}^\mathsf{T} & X_{12}B_{11}^\mathsf{T} - B_{12}X_{11}^\mathsf{T} & 0 \end{bmatrix} \in \mathfrak{m}.$$

We also define a map $e^{-\varphi_B} = \sum_{k=0}^\infty \frac{(-1)^k}{k!}\varphi_B^k : \mathfrak{m} \to \mathfrak{m}$, which can be used to compute the

parallel transport of a vector field on $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$.

Proposition 3.3.11 (Parallel transport I)**.** *Let $B, X \in T_{[\mathrm{Id}]}\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V}) \simeq \mathfrak{m}$ and let $[Q]$ be a point in $M \coloneqq \mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$. The parallel transport of $QX \in T_{[Q]}M$ along the geodesic $[Q\exp(tB)]$ on M is $X(t) = Q\exp(tB)e^{-\varphi_{tB}}(X)$.*

Proof. This is a direct result of [64, Lemma 3.1] applied to flag manifolds. □

We remark here that if $d = 1$ (in this case, $\mathrm{Flag}(k; \mathbf{V}) = \mathrm{Gr}(k, n)$), then it is straightforward to verify that $[B, X]_{\mathfrak{m}} = 0$ whenever $B, X \in \mathfrak{m}$. This implies that the parallel transport of Y along $[Q\exp(tB)]$ is $X(t) = Q\exp(tB)X$, which coincides with the computation in [65].

3.3.3 Other descriptions of flag manifolds

By Proposition 3.3.3, we see that $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ can be regarded as a subset of $\mathrm{Gr}(n_1, n) \times \cdots \times \mathrm{Gr}(n_d, n)$. On the other hand, $\mathrm{Gr}(k, n)$ can be regarded as a subset of $\mathbb{R}^{n\times n}$. Namely, for a given $\mathbf{W} \in \mathrm{Gr}(k, n)$, we let W be a $n \times k$ matrix whose column vectors are orthonormal and span the space $\mathbf{W}$. If W' is another such $n \times k$ matrix, then $W' = WQ$ where $Q \in \mathrm{O}(k)$. We consider a map

$$s : \mathrm{Gr}(k, n) \to \mathbb{R}^{n\times}, \quad f(\mathbf{W}) = WW^{\mathsf{T}}.$$

It is clear that s is well-defined and that s is injective.[7] The map f gives us a global way to describe $\mathrm{Gr}(k, n)$. Therefore, we have a global way to describe the flag manifold $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$:

$$f : \mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V}) \to (\mathbb{R}^{n\times n})^{\times d} \coloneqq \underbrace{\mathbb{R}^{n\times n} \times \cdots \times \mathbb{R}^{n\times n}}_{d \text{ copies}}$$

$$\mathbf{V}_1 \subsetneq \cdots \subsetneq \mathbf{V}_d \mapsto (V_1V_1^{\mathsf{T}}, \ldots, V_dV_d^{\mathsf{T}}),$$

7. WW^{T} is in fact the projection operator from $\mathbb{R}^n$ to its subspace $\mathbf{W}$.

where V_i is a $n \times n_i$ matrix whose column vectors are orthonormal and span the subspace $\mathbf{V}_i, i = 1, \ldots, d$.

Moreover, let $\mathrm{St}(n_d; \mathbf{V})$ be the set of all $n \times n_d$ matrices whose column vectors are orthonormal. The set $\mathrm{St}(n_d; \mathbf{V})$ is called the *Stiefel manifold of orthonormal n_d-frames in* $\mathbf{V}$. In fact, $\mathrm{St}(n_d; \mathbf{V})$ is a homogeneous space of $O(n)$, i.e.,

$$\mathrm{St}(n_d; \mathbf{V}) \simeq \mathrm{O}(n)/(\mathrm{O}(n - n_d)).$$

Details of Stiefel manifolds can be found in [70, 65]. The homogeneous structure of Stiefel manifolds together with Proposition 3.3.2 imply that we may regard flag manifolds as quotients of Stiefel manifolds:

$$\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V}) \simeq \mathrm{St}(n_d; \mathbf{V})/(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n_d - n_{d-1})).$$

Hence we may represent a flag $F : \mathbf{V}_1 \subsetneq \cdots \subsetneq \mathbf{V}_d$ by an $n \times n_d$ matrix A whose column vectors are orthonormal and the first n_i of them span the subspace $\mathbf{V}_i, i = 1, \ldots, d$. This representation is clearly not unique. Indeed, if A' is another such $n \times n_d$ matrix, then

$$A' = A \begin{bmatrix} Q_1 & 0 & \cdots & 0 \\ 0 & Q_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_d \end{bmatrix},$$

where $Q_i \in \mathrm{O}(n_i - n_{i-1}), i = 1, \ldots, d$. Let $Q \in O(n)$ be an orthogonal matrix representing an element $[Q]$ in $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$, then taking the first n_d columns of Q gives an $n \times n_d$ matrix A whose column vectors are orthonormal and the first n_i of them span the subspace $\mathbf{V}_i, i = 1, \ldots, d$. To summarize, we have

Proposition 3.3.12. *The flag manifold* $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ *is*

- *isomorphic to*

$$\{(P_1,\dots,P_d)\in(\mathbb{R}^{n\times n})^{\times d} : P_i^2=P_i,\quad P_i^{\mathsf{T}}=P_i,\quad \operatorname{tr}(P_i)=n_i,\quad P_jP_i=P_i,\\ i<j,\quad i,j=1,\dots,d\}. \quad (3.3.5)$$

- *isomorphic to the set of equivalence classes* $[A]$*, where*

$$[A]=\left\{ A\begin{bmatrix} Q_1 & 0 & \cdots & 0\\ 0 & Q_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & Q_d\end{bmatrix} : \begin{array}{l} \textit{column vectors of } A \textit{ are orthonormal and}\\ \textit{the first } n_i \textit{ of them span the subspace } \mathbf{V}_i\\ Q_i\in \mathrm{O}(n_i-n_{i-1}), i=1,\dots,d\end{array}\right\} \quad (3.3.6)$$

- *Let* $F:\mathbf{V}_1\subsetneq\cdots\subsetneq\mathbf{V}_d$ *be an element in* $\mathrm{Flag}(n_1,\dots,n_d;\mathbf{V})$*. Let* $Q\in\mathrm{O}(n)$ *be an* $n\times n$ *orthogonal matrix such that* $[Q]=F$ *and let* $P=(P_1,\dots,P_d)\in(\mathbb{R}^{n\times n})^{\times d}$ *be the* d*-tuple of matrices representing the flag* F*. Let* A *be an* $n\times n_d$ *matrix such that* $[A]=F$*. We take* $v_1,\dots,v_{n_d}$ *to be the first* n_d *columns of* Q *and take* $a_1,\dots,a_{n_d}$ *to be the column vectors of* A*. Then* Q,P *and* A *are related as follows:*

  1. $P_i=V_iV_i^{\mathsf{T}}$*, where* $V_i=[v_1,\dots,v_{n_i}], i=1,\dots,d$*;*
  2. $[A]=[V_d]$ *where* $V_d=[v_1,\dots,v_{n_d}]$*;*
  3. $\mathbf{V}_i=\{P_i(v):v\in\mathbb{R}^n\}, i=1,\dots,d$*;*
  4. *if* $w_1,\dots,w_{n_d}$ *is a set of orthonormal vectors in* $\mathbb{R}^n$ *such that* $w_1,\dots,w_{n_i}$ *is an orthonormal basis of* $\{P_i(v):v\in\mathbb{R}^n\}, i=1,\dots,d$*, then* $[A]=[W]$ *where* $W=[w_1,\dots,w_{n_d}]$*;*
  5. *let* $v_{n_d+1},\dots,v_n$ *be vectors such that* $v_1,\dots,v_n$ *are orthonormal basis of* $\mathbb{R}^n$*, then* $[Q]=[Q']$ *where* $Q'=[v_1,\dots,v_n]$*;*
  6. $P_i=V_iV_i^{\mathsf{T}}$*, where* $V_i=[a_1,\dots,a_{n_i}], i=1,\dots,d$*;*

Because of Proposition 3.3.12, we may study the Riemannian geometry of flag manifolds using various descriptions of them. Let

$$\gamma(t) = (P_1(t), \dots, P_d(t)) \tag{3.3.7}$$

be a curve in $(\mathbb{R}^{n\times n})^{\times d}$ such that $P_i(t)^2 = P_i(t), P_i(t)^{\mathsf{T}} = P_i(t), \operatorname{tr}(P_i(t)) = n_i, P_j(t)P_i(t) = P_i(t), i < j, i, j = 1, \dots, d, t \in (-1, 1)$. We denote$P_i(0)$ by P_i. Moreover, if A is an $n \times n_d$ matrix representing an element $[A] \in \operatorname{Flag}(n_1, \dots, n_d; \mathbf{V})$, we may write A as a block matrix $A = [A_1, \cdots, A_d]$, where $A_i \in \mathbb{R}^{n\times(n_i - n_{i-1})}$ and $A_i^{\mathsf{T}} A_j = 0, i \neq j$, $A_i^{\mathsf{T}} A_i = \operatorname{Id}_{n_i - n_{i-1}}$, $i = 1, \dots, d$. A curve in $\operatorname{Flag}(n_1, \dots, n_d; \mathbf{V})$ passing through $[A]$ can be written as $[\tau(t)]$ where $\tau(t) = [A_1(t), \cdots, A_d(t)]$ such that

$$A_i(t)^{\mathsf{T}} A_j(t) = 0, i \neq j, \quad A_i(t)^{\mathsf{T}} A_i(t) = \operatorname{Id}_{n_i - n_{i-1}}, i = 1, \dots, d, \quad \text{and} \quad \tau(0) = A. \tag{3.3.8}$$

We may obtain the following descriptions of tangent spaces of flag manifolds by differentiating curves $\gamma(t)$ and $[\tau(t)]$.

Proposition 3.3.13 (Tangent spaces II)**.** *The tangent space* $T_{(P_1,\dots,P_d)} \operatorname{Flag}(n_1, \dots, n_d; \mathbf{V})$ *is the vector space consisting of points* $(X_1, \dots, X_d) \in (\mathbb{R}^{n\times n})^{\times d}$ *such that*

$$X_i P_i + P_i X_i = X_i, \quad X_i^{\mathsf{T}} = X_i, \quad \operatorname{tr}(X_i) = 0, \quad X_j P_i + P_j X_i = X_i, \quad i < j, \quad i, j = 1, \dots, d.$$

Moreover, let A *be an* $n \times n_d$ *matrix such that column vectors of* A *are orthonormal. The tangent space* $T_{[A]} \operatorname{Flag}(n_1, \dots, n_d; \mathbf{V})$ *of* $\operatorname{Flag}(n_1, \dots, n_d; \mathbf{V})$ *at* $[A]$ *consists of matrices* $[X_1, \dots, X_d]$ *where* $X_i \in \mathbb{R}^{n\times(n_i - n_{i-1})}, i = 1, \dots, d$ *satisfy the conditions:*

$$A_i^{\mathsf{T}} X_j + X_i^{\mathsf{T}} A_j = 0 \quad \textit{and} \quad A_i^{\mathsf{T}} X_i = 0, \quad i, j = 1, \dots, d. \tag{3.3.9}$$

Equivalently, $[X_1, \ldots, X_d]$ *can be parametrized as*

$$[X_1, \ldots, X_d] = [A_1, \ldots, A_d, A^\perp] \begin{bmatrix} 0 & B_{1,2} & \cdots & B_{1,d} \\ -B_{1,2}^\mathsf{T} & 0 & \cdots & B_{2,d} \\ \vdots & \cdots & \ddots & \vdots \\ -B_{1,d}^\mathsf{T} & -B_{2,d}^\mathsf{T} & \cdots & 0 \\ -B_{1,d+1}^\mathsf{T} & -B_{2,d+1}^\mathsf{T} & \cdots & -B_{d,d+1}^\mathsf{T} \end{bmatrix},$$

where $A^\perp$ *is an* $n \times (n - n_d)$ *matrix such that* $[A_1, \ldots, A_d, A^\perp] \in \mathrm{O}(n)$ *and* $B_{ij} \in \mathbb{R}^{(n_i - n_{i-1}) \times (n_j - n_{j-1})}, 1 \leq i < j \leq d+1$.

Proof. The first part follows from differentiating a curve $\gamma(t)$ which satisfies the equations defining $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ in $(\mathbb{R}^{n \times n})^{\times d}$. The second part follows from differentiating the curve $\tau(t)$ and noticing the fact that the tangent vector of $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ at a point $[A]$ is perpendicular to the vector space consisting of matrices B of the form

$$B = \begin{bmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_d \end{bmatrix}, \quad B_i \in \mathfrak{so}(n_i - n_{i-1}), i = 1, \ldots, d.$$

[8] The last assertion is obtained from Proposition 3.3.6 and Proposition 3.3.12. □

Proposition 3.3.14 (Geodesics III)**.** *Geodesics on* $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ *passing through the point* $[A]$ *are of the form* $[Q \exp(tB) I_{n,n_d}]$. *Here* $Q = \begin{bmatrix} A & X \end{bmatrix} \in \mathrm{O}(n)$ *for some* $n \times (n - n_d)$ *matrix* X. *Furthermore, geodesics passing through* $[A]$ *can be written as* $[U \Sigma(t) V^\mathsf{T} I_{n,n_d}]$ *for*

8. The calculation is straightforward but details can be found in [67].

some $U, V \in \mathrm{O}(n)$ *and* $\Sigma(t)$ *is of the form*

$$\Sigma(t) = \mathrm{diag}\left\{ \begin{bmatrix} \cos t\lambda_1 & -\sin t\lambda_1 \\ \sin t\lambda_1 & \cos t\lambda_1 \end{bmatrix}, \ldots, \begin{bmatrix} \cos t\lambda_r & -\sin t\lambda_r \\ \sin t\lambda_r & \cos t\lambda_r \end{bmatrix}, \underbrace{0}_{(n-2r)\times(n-2r)} \right\}.$$

Proof. Let Q be an $n \times n$ orthogonal matrix and let $I_{n,n_d} = \begin{bmatrix} \mathrm{Id}_{n_d} \\ 0 \end{bmatrix}$ be the $n \times n_d$ matrix. The $n \times n_d$ matrix $A = QI_{n,n_d}$ represents the class $[A] \in \mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$. By Proposition 3.3.7, we see that geodesics passing through $[Q]$ must be of the form $[Q\exp(tB)]$. The second part follows from Proposition 3.3.9. □

Proposition 3.3.15 (Parallel transport II)**.** *Let* $B, X \in T_{[\mathrm{Id}]}\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ *and let* $[A]$ *be a point in* $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$*. Let* Q *be an* $n \times n$ *orthogonal matrix such that* $A = QI_{n,n_d}$*. The parallel transport of* $AX \in T_{[A]}M$ *along the geodesic* $[Q\exp(tB)I_{n,n_d}]$ *on* $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ *is* $X(t) = Q\exp(tB)e^{-\varphi_{tB}}(X)I_{n,n_d}$*.*

3.3.4 The gradient and Hessian of a function

In this subsection, we will discuss the gradient and Hessian of a function on flag manifolds, which are main ingredients of optimization algorithms.

Proposition 3.3.16 (The gradient of a function)**.** *Let* $F : \mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V}) \to \mathbb{R}$ *be a smooth function on* $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ *and let* D *be the* $n \times n_d$ *matrix whose* (i, j)*-th entry is* $\frac{\partial F}{\partial x_{ij}}$*, where* $\{x_{ij} : i = 1, \ldots, n, j = 1, \ldots, n_d\}$ *is a coordinate on* $Flag(n_1, \ldots, n_d; \mathbf{V})$*. Let* $[A]$ *be a point in* $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$*. We write* D *as* $[D_1, \ldots, D_d]$*, where* D_i *is an* $n \times (n_i - n_{i-1})$ *matrix and we write* $A = [A_1, \ldots, A_d]$ *where* $A_i \in \mathbb{R}^{n\times(n_i - n_{i-1})}, i = 1, \ldots, d$*. The gradient* ∇F *of* F *on* $\mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ *is given by* $\nabla F([A]) = [\Delta_1, \cdots, \Delta_d]$*, where*

$$\Delta_i = D_i - (A_i A_i^{\mathsf{T}} D_i + \sum_{j \neq i} A_j D_j^{\mathsf{T}} A_i). \tag{3.3.10}$$

Proof 1. By definition of ∇F, for every $Y \in T_{[A]}\operatorname{Flag}(n_1, \ldots, n_d; \mathbf{V})$, we have

$$\operatorname{tr}(\nabla F, Y) = g(\nabla F, Y) = \operatorname{tr}(D^{\mathsf{T}} Y).$$

Hence we may conclude that ∇F is the projection of D onto $T_{[A]}\operatorname{Flag}(n_1, \ldots, n_d; \mathbf{V})$, i.e., $D = \nabla F + Z$ where Z is perpendicular to $T_{[A]}\operatorname{Flag}(n_1, \ldots, n_d; \mathbf{V})$. We write $Z = [Z_1, \ldots, Z_d]$ and take

$$Z_i = A_i A_i^{\mathsf{T}} D_i + \sum_{j \neq i} A_j D_j^{\mathsf{T}} A_i, \quad i = 1, \ldots, d.$$

It is straightforward to verify that $D - Z \in T_{[A]}\operatorname{Flag}(n_1, \ldots, n_d; \mathbf{V})$ and that Z is perpendicular to $T_{[A]}\operatorname{Flag}(n_1, \ldots, n_d; \mathbf{V})$, by relations (3.3.9) in Proposition 3.3.13. □

Proof 2. The gradient ∇F can also be found by solving an optimization problem, see [67] for details. □

In particular, if $d = 1$, the gradient of a function F on $\operatorname{Flag}(k; \mathbf{V}) = \operatorname{Gr}(k, n)$ is $\nabla F(A) = \Delta = D - AA^{\mathsf{T}} D$. This coincides with the computation in [65].

Next we compute the Hessian of a function F on $\operatorname{Flag}(n_1, \ldots, n_d; \mathbf{V})$. To do this, we recall that the Hessian $\operatorname{Hess}(F)_{[A]}$ at a point $[A] \in \operatorname{Flag}(n_1, \ldots, n_d; \mathbf{V})$ is a symmetric bilinear form on the tangent space $T_{[A]}\operatorname{Flag}(n_1, \ldots, n_d; \mathbf{V})$. Moreover, if $\gamma(t)$ is a geodesic such that $\gamma'(0) = X$ and $\gamma(0) = [A]$, then we have

$$\operatorname{Hess}(F)_{[A]}(X, X) = \frac{d^2}{dt^2}\Big|_{t=0} F(\gamma(t)). \tag{3.3.11}$$

We remark here that since $\operatorname{Hess}(F)_{[A]}$ is a symmetric bilinear form, $\operatorname{Hess}(F)_{[A]}(x, y)$ can be determined by the formula

$$\operatorname{Hess}(F)_{[A]}(X, Y) = \frac{1}{2}(\operatorname{Hess}(F)_{[A]}(X + Y, X + Y) - \operatorname{Hess}(F)_{[A]}(X, X) - \operatorname{Hess}(F)_{[A]}(Y, Y)).$$

By Proposition 3.3.14, we see that the geodesic $\gamma(t)$ such that $\gamma(0) = [A]$ and $\gamma'(0) = X$

is $[Q\exp(tB)I_{n,n_d}]$, where $Q \in \mathrm{O}(n)$ such that $A = QI_{n,n_d}$ and $X = QBI_{n,n_d}$. First we compute $\frac{d}{dt}F(\gamma(t))$ by chain rule:

$$\frac{d}{dt}F(\gamma(t)) = \mathrm{tr}(F_{\gamma(t)}^{\mathsf{T}}\gamma'(t)) \tag{3.3.12}$$

Then we have

$$\frac{d^2}{dt^2}F(\gamma(t)) = \frac{d}{dt}(\mathrm{tr}(F_{\gamma(t)}^{\mathsf{T}}\gamma'(t))) = \mathrm{tr}(\gamma'(t)^{\mathsf{T}}F_{\gamma(t),\gamma(t)}^{\mathsf{T}}\gamma'(t)) + \mathrm{tr}(F_{\gamma(t)}^{\mathsf{T}}\gamma''(t)). \tag{3.3.13}$$

Now we evaluate $\frac{d^2}{dt^2}F(\gamma(t))$ at $t = 0$ to obtain:

$$\mathrm{Hess}(F)_{[A]}(X,X) = \frac{d^2}{dt^2}\Big|_{t=0} F(\gamma(t)) = F_{A,A}(X,X) - \mathrm{tr}(F_A^{\mathsf{T}}QB^{\mathsf{T}}Q^{\mathsf{T}}X), \tag{3.3.14}$$

where $F_{A,A}$ is the bilinear form determined by the Hessian matrix $(\frac{\partial^2 F}{\partial y_{ij}y_{kl}})$.[9] We may combine Equations (3.3.11) and (3.3.14) to obtain:

Proposition 3.3.17 (Hessian of a function)**.** *Let $F : \mathrm{Flag}(n_1,\ldots,n_d;\mathbf{V}) \to \mathbb{R}$ be a smooth function on $\mathrm{Flag}(n_1,\ldots,n_d;\mathbf{V})$ and let A be an $n \times n_d$ matrix representing a point in $\mathrm{Flag}(n_1,\ldots,n_d;\mathbf{V})$. The Hessian of F, $\mathrm{Hess}(F)_{[A]}(X,Y)$, is*

$$F_{A,A}(X,Y) + \frac{1}{2}(\mathrm{tr}(F_A^{\mathsf{T}}QB^{\mathsf{T}}Q^{\mathsf{T}}X) + \mathrm{tr}(F_A^{\mathsf{T}}QC^{\mathsf{T}}Q^{\mathsf{T}}Y) - \mathrm{tr}(F_A^{\mathsf{T}}Q(B+C)^{\mathsf{T}}Q^{\mathsf{T}}(X+Y))), \tag{3.3.15}$$

where $X, Y \in T_{[A]}\mathrm{Flag}(n_1,\ldots,n_d;\mathbf{V})$ and B, C are unqiue skew symmetric matrices such that $X = BI_{n,n_d}$ and $Y = CI_{n,n_d}$, respectively.

According to equation (3.3.15), if $d = 1$, then the Hessian of F on $\mathrm{Flag}(n_1,\ldots,n_d;\mathbf{V})$ is given by

$$\mathrm{Hess}(F)_{[A]}(X,Y) = F_{A,A}(X,Y) - \mathrm{tr}(F_A^{\mathsf{T}}AY^{\mathsf{T}}X),$$

9. To be more precise, $F_{A,A}(X,Y) = \sum_{i,j,k,l}\frac{\partial^2 F}{\partial y_{ij}y_{kl}}X_{ij}Y_{kl}$

which is consistent with the computation in [65]. The Newton's method requires us to determine the tangent vector $Y \in T_{[A]} \operatorname{Flag}(n_1, \ldots, n_d; \mathbf{V})$ such that

$$\operatorname{Hess}(F)_{[A]}(Y, X) = g(-\nabla F, X) = -\operatorname{tr}(D([A])^{\mathsf{T}} X), \tag{3.3.16}$$

for every $X \in T_{[A]} \operatorname{Flag}(n_1, \ldots, n_d; \mathbf{V})$. Equation (3.3.15) implies that the tangent vector Y can be obtained by solving the linear system (3.3.16) for any fixed A and F.

3.4 Optimization algorithms

Riemannian optimization is analogous to optimization in Euclidean space. For instance, in gradient descent, instead of adding the negative multiple of the gradient by Proposition 3.3.16 to the current iterate, we move along the geodesic by Proposition 3.3.9 with initial velocity vector opposite to the gradient. The only subtley is to change of representation for the $\operatorname{Flag}(n_1, \ldots, n_d; \mathbf{V})$.

To be specific, let $[A] \in \operatorname{Flag}(n_1, \ldots, n_d; \mathbf{V})$ with $A = [Y_1, \cdots, Y_d]$, where $Y_i \in \mathbb{R}^{n \times (n_i - n_{i-1})}$, $i = 1, \cdots, d$ and $A^{\mathsf{T}} A = I$. First of all we need to find $\tilde{A} \in \mathbb{R}^{n \times (n - n_d)}$ to complement the subspace orthogonal to $\operatorname{im}(A)$, i.e. $\begin{bmatrix} A & \tilde{A} \end{bmatrix} \in \mathrm{O}(n)$, that corresponds to the same point in homogeneous space representation. After that, we compute the gradient $\nabla F \in \mathbb{R}^{n \times n_d}$ by Proposition 3.3.16 and set $G = \nabla F$. We can get the exponential map direction B in (3.3.2) by computing $\begin{bmatrix} A & \tilde{A} \end{bmatrix}^{\mathsf{T}} G \in \mathbb{R}^{n \times n_d}$ and use the skew-symmetric structure of B to convert it back. Lastly, we can find the next iterate by Proposition 3.3.14.

The algorithm for conjugate gradient is much involved. On top of updating the iterate, we need to update tangent vector by parallel transport simultaneously. However there is no neat equation for parallel transport without involving matrix exponential of Lie bracket. To remedy this, we can provide estimated tangent vector update by using the first few terms, and then projecting it back to the new tangent plane by enforcing the skew-symmetric structure.

Algorithm 3.4.1 Steepest descent

Initialize $[A_0] \in \mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ with $A_0 = [Y_1, \cdots, Y_d]$, where $Y_i \in \mathbb{R}^{n \times (n_i - n_{i-1})}, i = 1, \cdots, d$ and $A_0^\mathsf{T} A_0 = I$.
find $\widetilde{A_0} \in \mathbb{R}^{n \times (n-n_d)}$ such that $\begin{bmatrix} A_0 & \widetilde{A_0} \end{bmatrix} \in \mathrm{O}(n)$. ▷ find $\ker(A_0^\mathsf{T})$
set $X_0 = \begin{bmatrix} A_0 & \widetilde{A_0} \end{bmatrix}$.
for $i = 0, 1, \ldots$ **do**
  set $G_i = -\nabla F$; ▷ ∇F as in (3.3.10); gradient of f at A_i
  compute $\hat{B} = X_i^\mathsf{T} G_i$;
  set $B \in \mathbb{R}^{n \times n}$ by $B_{ij} = \hat{B}_{ij}$ for $j \leq n_d$; $B_{ij} = -\hat{B}_{ji}$ for $j \geq n_d$ and $i \leq n_d$; and $B_{ij} = 0$ otherwise
  minimize $f(X(t)) = f(X_i \exp(tB) I_{n,n_d})$ over $t \in \mathbb{R}$; ▷ exact line search
  set $X_{i+1} = X_i \exp(t_{\min} B)$;
end for

3.5 Numerical Experiments

We will firstly demonstrate the correctness of steepest descent by numerical experiment on dominant invariant subspace, which has closed-form solution that we can compare with. We the proceed to solve a variation of the same problem that cannot be solved with other means and illustrate the convergence with increasing function values and decline in norm of gradient.

3.5.1 Dominant invariant subspace

Let $A \in \mathbb{R}^{n \times n}$ be symmetric. We would like to solve

$$\text{maximize} \quad \mathrm{tr}(X^\mathsf{T} A X), \tag{3.5.1}$$

where $X \in \mathrm{Flag}(n_1, \ldots, n_d; \mathbb{R}^n)$. This is essentially equivalent to Principal Component Analysis (PCA), and solving this yields a flag of PCA subspaces of dimensions $n_1, \ldots, n_d$. An advantage of the viewpoint is that we usually do not know the intrinsic dimension of the data

Algorithm 3.4.2 Conjugate gradient

Initialize $[A_0] \in \mathrm{Flag}(n_1, \ldots, n_d; \mathbf{V})$ with $A_0 = [Y_1, \cdots, Y_d]$, where $Y_i \in \mathbb{R}^{n \times (n_i - n_{i-1})}, i = 1, \cdots, d$ and $A_0^\mathsf{T} A_0 = I$.

find $\widetilde{A_0} \in \mathbb{R}^{n \times (n-n_d)}$ such that $\begin{bmatrix} A_0 & \widetilde{A_0} \end{bmatrix} \in \mathrm{O}(n)$. ▷ find $\ker(A_0^\mathsf{T})$

set $X_0 = \begin{bmatrix} A_0 & \widetilde{A_0} \end{bmatrix}$.

set $G_0 = -\nabla F(A_0)$ and $H_0 = -G_0$; ▷ ∇F as in (3.3.10); gradient of f at A_0

for $i = 0, 1, \ldots$ **do**

compute $\hat{B} = X_i^\mathsf{T} G_i$;

set $B \in \mathbb{R}^{n \times n}$ by $B_{ij} = \hat{B}_{ij}$ for $j \leq n_d$; $B_{ij} = -\hat{B}_{ji}$ for $j \geq n_d$ and $i \leq n_d$; and $B_{ij} = 0$ otherwise

minimize $f(X(t)) = f(X_i \exp(tB) I_{n,n_d})$ over $t \in \mathbb{R}$; ▷ exact line search

set $A_{i+1} = A_i \exp(t_{\min} B)$;

set $\widetilde{A_{i+1}} = \widetilde{A_i} \exp(t_{\min} B)$;

set $X_{i+1} = \begin{bmatrix} A_{i+1} & \widetilde{A_{i+1}} \end{bmatrix}$.

set $G_{i+1} = -\nabla F(A_{i+1})$; ▷ gradient of f at A_{i+1}

procedure DESCENT(A_i, A_{i+1}, G_i, H_i) ▷ set new descent direction at A_{i+1}

$\tau H_i = A_{i+1} e^{-\varphi_{t_{\min} B}} (A_i^\mathsf{T} H_i)$; ▷ parallel transport of H_i; Proposition 3.3.11

$\tau G_i = A_{i+1} e^{-\varphi_{t_{\min} B}} (A_i^\mathsf{T} G_i)$; ▷ parallel transport of G_i

$\gamma_i = \mathrm{tr}((G_{i+1} - \tau G_i)^\mathsf{T} G_{i+1}) / \mathrm{tr}(G_i^\mathsf{T} G_i)$;

$H_{i+1} = -G_{i+1} + \gamma_i \tau H_i$;

end procedure

reset $H_{i+1} = -G_{i+1}$ if $i + 1 \equiv 0 \mod (k+1)(n-k)$;

end for

set, and by proposing a sequence of dimensions, we can efficiently compute the flag of PCA subspaces, and the spanning vectors, with this formulation. Moreover, we can also interpret this as finding subspaces of dimensions $n_1, n_2 - n_1, \ldots, n_d - n_{d-1}$ that are independent and explain different levels of the variance among the data set.

Figure 3.1 shows convergence trajectories of steepest descent, i.e. Algorithm 3.4.1, on $\mathrm{Flag}(3, 7, 12; \mathbb{R}^{60})$ for the problem (3.5.2). $\mathrm{Flag}(3, 7, 12; \mathbb{R}^{60})$ is a 623-dimensional manifold; we generate symmetric matrix $A \in \mathbb{R}^{60\times 60}$ randomly with $\mathcal{N}(0, 1)$ entries. The Euclidean gradient of $f(X) = \mathrm{tr}(X^\mathsf{T} AX)$ is given by $\nabla f(X) = 2AX$. The optimal value of this problem is the sum of the k largest eigenvalues which can be found by eigenvalue decomposition of A. We can see that steepest descent converges to the true solution, taking around 30 iterations. The convergence can also be seen from the diminishing marginal increase in function value, vanishing of the Riemannian gradient, and small distance between successive iterates. Notice that the vanishing of the Riemannian gradient can serve as an optimality condition. For implementation we use 2-norm of Riemannian gradient, distance between successive iterates, and number of iterations as stopping criteria of the algorithm.

We perform more extensive experiments by taking the average of 100 instances of the problem (3.5.2) for various increasing sequence of positive integers,which represent the dimensions of subspaces and the ambient space, to generate tables of elapsed time and accuracy. Table 3.1 and 3.3 shows the algorithm is robust to different choices of dimension of the ambient space and also number of subspaces. Moreover, Table 3.2 and 3.4 show that the elapsed time increases with the dimension of the manifold and the algorithm can work with a.

k	30	40	50	60	70	80	90	100
Accuracy ($\times 10^{-4}$)	2	8	64	32	4	87	20	15

Table 3.1: Accuracy (distance to true solution) of steepest descent for dominant invariant subspace on $\mathrm{Flag}(3, 9, 21; \mathbb{R}^k)$.

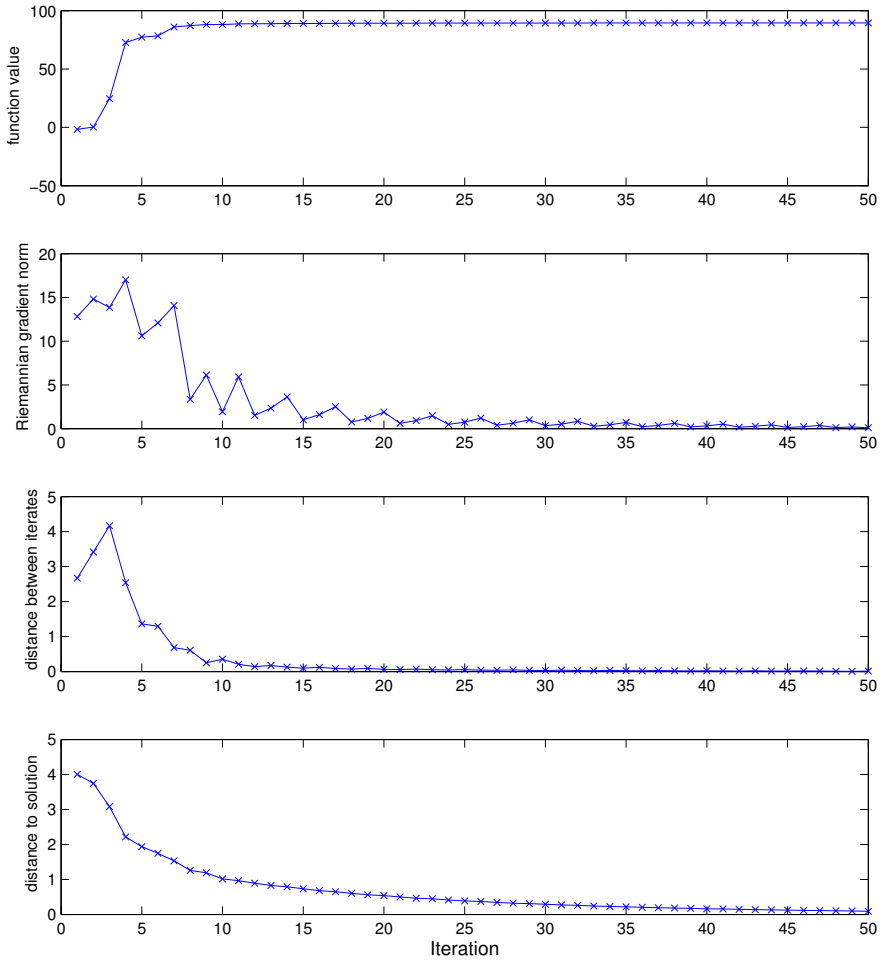

Figure 3.1: Convergence trajectories of steepest descent for dominant invariant subspace on $\mathrm{Flag}(3, 7, 12; \mathbb{R}^{60})$.

k	30	40	50	60	70	80	90	100
Elapsed Time	0.38	0.40	0.67	0.93	1.71	2.27	3.08	4.07

Table 3.2: Elapsed time of steepest descent for dominant invariant subspace on $\mathrm{Flag}(3, 9, 21; \mathbb{R}^k)$.

k	1	2	3	4	5	6	7	8	9	10
Accuracy ($\times 10^{-4}$)	1.4	3.4	3.4	8.6	2.8	18	19	5.1	9.3	11

Table 3.3: Accuracy (distance to true solution) of steepest descent for dominant invariant subspace on $\mathrm{Flag}(2, \cdots, 2k; \mathbb{R}^{60}), k = 1, \cdots, 10$.

3.5.2 Eigenvalue problem with nonlinear transform

Let $A \in \mathbb{R}^{n\times n}$ be symmetric. We would like to solve

$$\text{maximize} \quad \textstyle\sum_{i=1}^{d} \operatorname{tr}^2(X_i^\mathsf{T} A X_i), \tag{3.5.2}$$

over all $[X_1, \cdots, X_d] \in \mathrm{Flag}(n_1, \ldots, n_d; \mathbb{R}^n)$. This is a variation of the dominant invariant subspace problem and maximizes sum of nonlinear transforms of sum of eigenvalues of each independent subspaces instead. Our algorithms can indeed be applied to function $\sum_{i=1}^{d} g_i(\operatorname{tr}(X_i^\mathsf{T} A X_i))$, where $g_i \in C^1(\mathbb{R})$ for $i = 1, \cdots, d$.

We can see the convergence trajectories in Figure 3.2. Notice that we cannot solve this problem with eigenvalue decomposition and hence do not know the global optimum. It is possible that the solution we found is a local minimum, as demonstrated by the vanishing of the Riemannian gradient. We also observe that convergence speed is similar in this generalization when we compare Figure 3.1 and 3.2, where we fix the symmetric matrix A and the search space $\mathrm{Flag}(3, 7, 12; \mathbb{R}^{60})$.

3.6 Conclusion

We introduce the flag manifold, study its algebraic and geometric properties and develop concrete matrix representation of coordinates and tangent space, and computing exponential

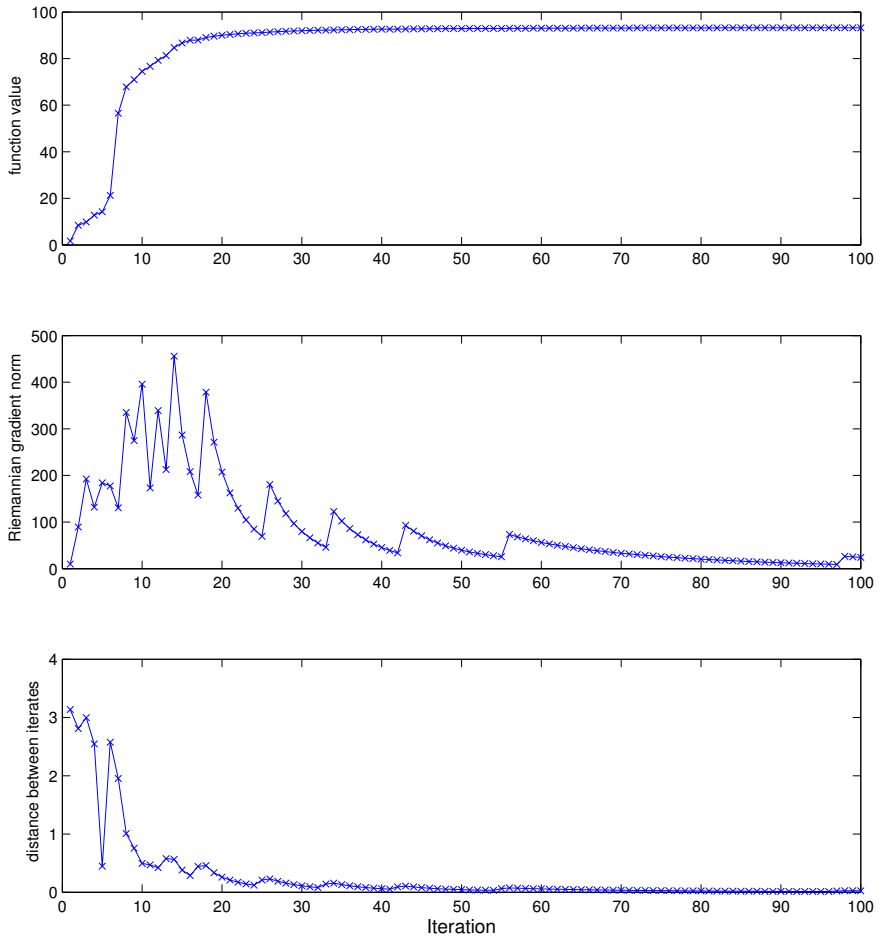

Figure 3.2: Convergence trajectories of steepest descent for eigenvalue problem on $\mathrm{Flag}(3, 7, 12; \mathbb{R}^{60})$.

k	1	2	3	4	5	6	7	8	9	10
Elapsed Time	0.54	0.81	0.79	0.96	1.05	0.91	1.20	1.06	1.18	1.12

Table 3.4: Elapsed time of steepest descent for dominant invariant subspace on $\mathrm{Flag}(2, \cdots, 2k; \mathbb{R}^{60}), k = 1, \cdots, 10$.

k	1	2	3	4	5	6	7	8	9	10
Accuracy ($\times 10^{-4}$)	1.4	3.4	3.4	8.6	2.8	18	19	5.1	9.3	11

Table 3.5: Accuracy (distance to true solution) of steepest descent for dominant invariant subspace on $\mathrm{Flag}(2, \cdots, 2k; \mathbb{R}^{60}), k = 1, \cdots, 10$.
map, parallel transport, gradient and Hessian of function using linear algebra operation alone. We demonstrate the correctness of the algorithm on symmetric eigenvalue problem which has a closed-form solution, and extend it to a new class of optimization problem that optimize nonlinear sum of eigenvalues of symmetric matrix. One of our future goal is to apply this framework to study mixture of subspaces model with orthogonality constraint. We believe optimization on flag manifold can help us discover a finer representation of data and achieve better convergence.

CHAPTER 4

TENSOR RANK AND COMPLEXITY OF YATES'S METHOD

4.1 Introduction

Yates's method was first proposed to exploit the structure of full factorial designed experiment to obtain least squares estimates for factor effects for all factors and their relevant interactions. In short it is an organized way to do iterative summation which avoids repeated computation. Many well-known algorithms including Fast Fourier transform and Walsh transform turned out to be special cases of Yates's method. Here we show that Yates's algorithm is optimal in the sense of a contraction of n tensors but may be improved when considered from the perspective of bilinear complexity. We also show that it is a projection of a tensor network and in particular has special relations with tensor train and tree tensor network. We want to evaluate

$$f(x_1, x_2, \ldots, x_n) := \sum_{y_1=0}^{m_1} \cdots \sum_{y_n=0}^{m_n} g_1(x_1, x_2, \ldots, x_n, y_1) g_2(x_2, \ldots, x_n, y_2) \cdots g_n(x_n, y_n)$$
$$h(y_1, y_2, \ldots, y_n) \quad (4.1.1)$$

for given real-valued functions $g_1, \ldots, g_n, h$ on discrete variables $x_1, \ldots, x_n, y_1, \ldots, y_n$. Yates's method does

$$\begin{aligned}
f_0(y_1, y_2, y_3, \ldots, y_n) &= h(y_1, y_2, y_3, \ldots, y_n), \\
f_1(x_n, y_1, y_2, \ldots, y_{n-1}) &= \sum_{y_n=0}^{m_n} g_n(x_n, y_n) f_0(y_1, y_2, y_3, \ldots, y_n), \\
f_2(x_{n-1}, x_n, y_1, \ldots, y_{n-2}) &= \sum_{y_{n-1}=0}^{m_{n-1}} g_{n-1}(x_{n-1}, x_n, y_{n-1}) f_1(x_n, y_1, y_2, \ldots, y_{n-1}), \\
&\vdots \\
f_{n-1}(x_2, x_3, \ldots, x_n, y_1) &= \sum_{y_2=0}^{m_2} g_2(x_2, \ldots, x_n, y_2) f_{n-2}(x_3, x_4, \ldots, x_{n-1}, y_1, y_2), \\
f_n(x_1, x_2, x_3, \ldots, x_n) &= \sum_{y_1=0}^{m_1} g_1(x_1, x_2, \ldots, x_n, y_1) f_{n-1}(x_2, x_3, , \ldots, x_n, y_1), \\
f(x_1, x_2, \ldots, x_n) &= f_n(x_1, x_2, \ldots, x_n).
\end{aligned} \tag{4.1.2}$$

- n-dimensional discrete Fourier transform on an $m_1 \times \cdots \times m_n$ grid.

$$f(x_1, x_2, \ldots, x_n) := \sum_{y_1=0}^{m_1} \cdots \sum_{y_n=0}^{m_n} \exp\left(2\pi i \left(\frac{x_1 y_1}{m_1} + \cdots + \frac{x_n y_n}{m_n}\right)\right) h(y_1, y_2, \ldots, y_n).$$

- 1-dimensional discrete Fourier transform on a grid with 2^n points. Set $m_1 = \cdots = m_n = 2$ and

$$\begin{aligned}
f(x_1, x_2, \ldots, x_n) &= f(2^{n-1} x_1 + \cdots + 2x_{n-1} + x_n), \\
h(y_1, y_2, \ldots, y_n) &= h(2^{n-1} y_n + \cdots + 2y_2 + y_1), \\
g_k(x_k, \ldots, x_n, y_k) &= \exp\left(2\pi i [2^{k-1} y_k (x_n + 2x_{n-1} + \cdots + 2^{n-k} x_k)] \Big/ 2^n\right), \\
& k = 1, \ldots, n.
\end{aligned}$$

Yates's method reduces to FFT.

- Walsh transform. Set $x_1, \ldots, x_n, y_1, \ldots, y_n \in \{0, 1\}$ and

$$f(x_1, x_2, \ldots, x_n) := \sum_{y_1=0}^{1} \cdots \sum_{y_n=0}^{1} (-1)^{x_1 y_1 + \cdots + x_n y_n} h(y_1, y_2, \ldots, y_n).$$

  Write $f_{x_1 x_2 x_3} = f(x_1, x_2, x_3)$ and $h_{y_1 y_2 y_3} = h(y_1, y_2, y_3)$. Yates's method does

Input	Step 1	Step 2	Step 3	Output
h_{000}	$h_{000} + h_{001}$	$h_{000} + h_{001} + h_{010} + h_{011}$	$h_{000} + h_{001} + h_{010} + h_{011} + h_{100} + h_{101} + h_{110} + h_{111}$	f_{000}
h_{001}	$h_{010} + h_{011}$	$h_{100} + h_{101} + h_{110} + h_{111}$	$h_{000} - h_{001} + h_{010} - h_{011} + h_{100} - h_{101} + h_{110} - h_{111}$	f_{001}
h_{010}	$h_{100} + h_{101}$	$h_{000} - h_{001} + h_{010} - h_{011}$	$h_{000} + h_{001} - h_{010} - h_{011} + h_{100} + h_{101} - h_{110} - h_{111}$	f_{010}
h_{011}	$h_{110} + h_{111}$	$h_{100} - h_{101} + h_{110} - h_{111}$	$h_{000} - h_{001} - h_{010} + h_{011} + h_{100} - h_{101} - h_{110} + h_{111}$	f_{011}
h_{100}	$h_{000} - h_{001}$	$h_{000} + h_{001} - h_{010} - h_{011}$	$h_{000} + h_{001} + h_{010} + h_{011} - h_{100} - h_{101} - h_{110} - h_{111}$	f_{100}
h_{101}	$h_{010} - h_{011}$	$h_{100} + h_{101} - h_{110} - h_{111}$	$h_{000} - h_{001} + h_{010} - h_{011} - h_{100} + h_{101} - h_{110} + h_{111}$	f_{101}
h_{110}	$h_{100} - h_{101}$	$h_{000} - h_{001} - h_{010} + h_{011}$	$h_{000} + h_{001} - h_{010} - h_{011} - h_{100} - h_{101} + h_{110} + h_{111}$	f_{110}
h_{111}	$h_{110} - h_{111}$	$h_{100} - h_{101} - h_{110} + h_{111}$	$h_{000} - h_{001} - h_{010} + h_{011} - h_{100} + h_{101} + h_{110} - h_{111}$	f_{111}

- Wavelet packet transform. In Haar wavelet packet transform, we compute the wavelet coefficients in each scale by taking the weighted sum and difference of each pair of coefficients in the previous layer. Set $x_1, \ldots, x_n, y_1, \ldots, y_n \in \{0, 1\}$ and

$$f(x_1, x_2, \ldots, x_n) := \sum_{y_1=0}^{1} \cdots \sum_{y_n=0}^{1} \frac{(-1)^{x_1 y_1 + \cdots + x_n y_n}}{2^{n/2}} h(y_1, y_2, \ldots, y_n).$$

  Write $f_{x_1 x_2 x_3} = f(x_1, x_2, x_3)$ and $h_{y_1 y_2 y_3} = h(y_1, y_2, y_3)$. Yates's method does

Input	Step (Scale) 1	Step 2	Step 3	Output
h_{000}	$(h_{000} + h_{001})/\sqrt{2}$	$(h_{000} + h_{001} + h_{010} + h_{011})/2$	$(h_{000} + h_{001} + h_{010} + h_{011} + h_{100} + h_{101} + h_{110} + h_{111})/2^{3/2}$	f_{000}
h_{001}	$(h_{010} + h_{011})/\sqrt{2}$	$(h_{100} + h_{101} + h_{110} + h_{111})/2$	$(h_{000} - h_{001} + h_{010} - h_{011} + h_{100} - h_{101} + h_{110} - h_{111})/2^{3/2}$	f_{001}
h_{010}	$(h_{100} + h_{101})/\sqrt{2}$	$(h_{000} - h_{001} + h_{010} - h_{011})/2$	$(h_{000} + h_{001} - h_{010} - h_{011} + h_{100} + h_{101} - h_{110} - h_{111})/2^{3/2}$	f_{010}
h_{011}	$(h_{110} + h_{111})/\sqrt{2}$	$(h_{100} - h_{101} + h_{110} - h_{111})/2$	$(h_{000} - h_{001} - h_{010} + h_{011} + h_{100} - h_{101} - h_{110} + h_{111})/2^{3/2}$	f_{011}
h_{100}	$(h_{000} - h_{001})/\sqrt{2}$	$(h_{000} + h_{001} - h_{010} - h_{011})/2$	$(h_{000} + h_{001} + h_{010} + h_{011} - h_{100} - h_{101} - h_{110} - h_{111})/2^{3/2}$	f_{100}
h_{101}	$(h_{010} - h_{011})/\sqrt{2}$	$(h_{100} + h_{101} - h_{110} - h_{111})/2$	$(h_{000} - h_{001} + h_{010} - h_{011} - h_{100} + h_{101} - h_{110} + h_{111})/2^{3/2}$	f_{101}
h_{110}	$(h_{100} - h_{101})/\sqrt{2}$	$(h_{000} - h_{001} - h_{010} + h_{011})/2$	$(h_{000} + h_{001} - h_{010} - h_{011} - h_{100} - h_{101} + h_{110} + h_{111})/2^{3/2}$	f_{110}
h_{111}	$(h_{110} - h_{111})/\sqrt{2}$	$(h_{100} - h_{101} - h_{110} + h_{111})/2$	$(h_{000} - h_{001} - h_{010} + h_{011} - h_{100} + h_{101} + h_{110} - h_{111})/2^{3/2}$	f_{111}

- Matrix-vector product when the matrix has Kronecker structure. For $A \in \mathbb{R}^{m \times m}$,

$B \in \mathbb{R}^{n\times n}$ and $v \in \mathbb{R}^{mn}$, we denote $u = (A \otimes B)v$. Set $x_i, y_j \in [m]$ and $x_k, y_l \in [n]$.

$$u_{x_i n+x_k+1} = \sum_{y_j=0}^{m-1} \sum_{y_l=0}^{n-1} a_{(x_i+1)(y_j+1)} b_{(x_k+1)(y_l+1)} v_{y_j n+y_l+1}$$

Set $m_1 = m$ and $m_2 = n$ and

$$\begin{aligned} f(x_i, x_k) &= u_{x_i n+x_k+1} \\ g_1(x_i, x_k, y_j) &= a_{(x_i+1)(y_j+1)} \\ g_2(x_k, y_l) &= b_{(x_k+1)(y_l+1)} \\ h(y_j, y_l) &= v_{y_j n+y_l+1} \end{aligned}$$

4.2 Yates's method as a multilinear map

We consider real vector spaces $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ of dimensions

$$\dim X_i = l_i + 1, \dim Y_i = m_i + 1, i = 1, 2, \ldots, n.$$

We fix a basis for each of X_i and Y_i hence we have identifications

$$\begin{aligned} X_i &\cong \mathbb{R}^{l_i+1}, \\ Y_i &\cong \mathbb{R}^{m_i+1}. \end{aligned}$$

Therefore, for a tensor $g_i \in X_i \otimes \cdots X_n \otimes Y_i$, we can view g_i as a hypermatrix, i.e.,

$$g_i = (g_i(x_i, \ldots, x_n, y_i)),$$

Here $(x_i, \ldots, x_n, y_i)$ is an element of $[l_i] \times \cdots \times [l_n] \times [m_i]$ where $[k]$ is the set $\{0, 1, \ldots, k\}$ and $g_{i,I}$ is a real number. Similarly, a tensor $h \in Y_1^* \otimes \cdots \otimes Y_n^*$ can be viewed as a hypermatrix

$$h = (h(y_1, \ldots, y_n)),$$

where $(y_1, \ldots, y_n)$ is an element of $[m_1] \times \cdots \times [m_n]$ and $h(y_1, \ldots, y_n)$ is a real number.

We consider a linear map

$$T : \bigotimes_{i=1}^{n} (X_i \otimes \cdots X_n \otimes Y_i) \otimes Y_1^* \otimes \cdots \otimes Y_n^* \to \bigotimes_{i=1}^{n} X_i$$

defined by

$$T(g_1 \otimes \cdots \otimes g_n \otimes h) = (f(x_1, \ldots, x_n)), \tag{4.2.1}$$

where $x_i \in [l_i]$, $i = 1, 2, \ldots, n$ and $f(x_1, \ldots, x_n)$ is the function

$$f(x_1, x_2, \ldots, x_n) := \sum_{y_1=0}^{m_1} \cdots \sum_{y_n=0}^{m_n} g_1(x_1, x_2, \ldots, x_n, y_1) g_2(x_2, \ldots, x_n, y_2) \cdots g_n(x_n, y_n)$$
$$h(y_1, y_2, \ldots, y_n). \tag{4.2.2}$$

We can think of T as a machine that whenever we input functions $g_1, \ldots, g_n$ and h, it outputs the function f.

Notice that we have isomorphism of linear spaces

$$\begin{aligned}
&\bigotimes_{i=1}^{n} (X_i \otimes \cdots \otimes X_n \otimes Y_i) \otimes Y_1^* \otimes \cdots \otimes Y_n^* \\
&\quad \cong \bigotimes_{i=1}^{n} (X_i \otimes \cdots \otimes X_n) \otimes (Y_1 \otimes Y_1^*) \otimes \cdots \otimes (Y_n \otimes Y_n^*) \\
&\quad \cong \bigotimes_{i=1}^{n} X_i^{\otimes i} \otimes (Y_1 \otimes Y_1^*) \otimes \cdots \otimes (Y_n \otimes Y_n^*)
\end{aligned}$$

Hence we can also view T as a multilinear map

$$T : \bigotimes_{i=1}^{n} X_i^{\otimes i} \otimes (Y_1 \otimes Y_1^*) \otimes \cdots \otimes (Y_n \otimes Y_n^*) \to \bigotimes_{i=1}^{n} X_i.$$

By definition of T we can factor T as

$$\begin{aligned}
&\bigotimes_{i=1}^{n} X_i^{\otimes i} \otimes (Y_1 \otimes Y_1^*) \otimes \cdots \otimes (Y_n \otimes Y_n^*) \\
\xrightarrow{\text{Proj}_n} \quad &\bigotimes_{i=1}^{n-1} X_i^{\otimes i} \otimes X_n \otimes (Y_1 \otimes Y_1^*) \otimes \cdots \otimes (Y_n \otimes Y_n^*) \\
\xrightarrow{\text{Proj}_{n-1}} \quad &\bigotimes_{i=1}^{n-2} X_i^{\otimes i} \otimes X_{n-1} \otimes X_n \otimes (Y_1 \otimes Y_1^*) \otimes \cdots \otimes (Y_n \otimes Y_n^*) \\
&\vdots \\
\xrightarrow{\text{Proj}_1} \quad &\bigotimes_{i=1}^{n} X_i \otimes (Y_1 \otimes Y_1^*) \otimes \cdots \otimes (Y_n \otimes Y_n^*) \\
\xrightarrow{\text{Con}_n} \quad &\bigotimes_{i=1}^{n} X_i \otimes (Y_1 \otimes Y_1^*) \otimes \cdots \otimes (Y_{n-1} \otimes Y_{n-1}^*) \\
\xrightarrow{\text{Con}_{n-1}} \quad &\bigotimes_{i=1}^{n} X_i \otimes (Y_1 \otimes Y_1^*) \otimes \cdots \otimes (Y_{n-2} \otimes Y_{n-2}^*) \\
&\vdots \\
\xrightarrow{\text{Con}_1} \quad &\bigotimes_{i=1}^{n} X_i
\end{aligned}$$

where $\text{Proj}_i : X_i^{\otimes i} \to X_i$ is the projection map defined by

$$\text{Proj}_i(a) = (a(j, \ldots, j)),$$

with $a \in X_i^{\otimes i}$ and $j = 0, 2, \ldots, l_i$ and $\text{Con}_i : Y_i^* \otimes Y_i \to \mathbb{R}$ is the contraction map defined by

$$\text{Con}_i(v, \alpha) = \alpha(v).$$

Proposition 4.2.1. *The Yates's method* (4.1.2) *corresponds to the factorization*

$$T = \mathrm{Con}_1 \circ \cdots \mathrm{Con}_n \circ \mathrm{Proj}_1 \circ \cdots \circ \mathrm{Proj}_n \,. \tag{4.2.3}$$

Proof. The evaluation of the function

$$f(x_1, x_2, \ldots, x_n) := \sum_{y_1=0}^{m_1} \cdots \sum_{y_n=0}^{m_n} g_1(x_1, x_2, \ldots, x_n, y_1) g_2(x_2, \ldots, x_n, y_2) \cdots g_n(x_n, y_n)$$
$$h(y_1, y_2, \ldots, y_n) \tag{4.2.4}$$

is the same as the evaluation of the multilinear map T. The effect of projection maps Proj_i is to evaluate $g_1, \ldots, g_n$ at the same x_i. It is clear that the evaluation of the contraction map Con_i is the same as the calculation of f_i in (4.1.2). □

It is obvious that Con_i and Proj_j are commutative, i.e., $\mathrm{Con}_i \circ \mathrm{Proj}_j = \mathrm{Proj}_j \circ \mathrm{Con}_i$ for all $1 \leq i, j, \leq n$ hence we are allowed to permute Con_i's and Proj_j's in (4.2.3), for example, we can also write the tensor T as

$$T = \mathrm{Proj}_1 \circ \cdots \circ \mathrm{Proj}_n \circ \mathrm{Con}_1 \circ \cdots \mathrm{Con}_n \,. \tag{4.2.5}$$

4.3 Yates's method as a tensor network state

A *matrix product state*(MPS) is a tensor $a(x_1, \ldots, x_n) \in \mathbb{R}^{l_1+1} \otimes \cdots \mathbb{R}^{l_n+1}$, $x_i \in [l_i]$ which can be expressed as

$$a(x_1, \ldots, x_n) = \sum_{y_1=0}^{m_1} \cdots \sum_{y_{n-1}=0}^{m_{n-1}} b_1(y_n, x_1, y_1) b_2(y_1, x_2, y_2) \cdots b_{n-1}(y_{n-2}, x_{n-1}, y_{n-1})$$
$$b_n(y_{n-1}, x_n, y_n), \tag{4.3.1}$$

for some functions $b_j(y_{j-1}, x_j, y_j)$, $j = 1, 2, \ldots, n$. Algebraically, an MPS is an element in the image of

$$\mathrm{Con} : \bigotimes_{j=1}^{n} (\mathbb{R}^{m_{j-1}+1} \otimes \mathbb{R}^{l_j+1} \otimes \mathbb{R}^{*m_j+1}) \to \bigotimes_{j=1}^{n} \mathbb{R}^{l_j+1}.$$

where Con is the tensor product of contractions $\mathrm{Con}_j : \mathbb{R}^{*m_j+1} \otimes \mathbb{R}^{m_j+1} \to \mathbb{R}$. Here we adopt the convention that $\mathbb{R}^{m_0+1} = \mathbb{R}^{m_n+1}$.

A *tensor train*(TT) is a tensor $a(x_1, \ldots, x_n) \in \mathbb{R}^{l_1+1} \otimes \cdots \mathbb{R}^{l_n+1}$, $x_i \in [l_i]$ which can be expressed as

$$a(x_1, \ldots, x_n) = \sum_{y_1=0}^{m_1} \cdots \sum_{y_{n-1}=0}^{m_{n-1}} b_1(x_1, y_1) b_2(y_1, x_2, y_2) \cdots b_{n-1}(y_{n-2}, x_{n-1}, y_{n-1}) b_n(y_{n-1}, x_n),$$

for some functions $b_j(y_{j-1}, x_j, y_j)$, $j = 1, 2, \ldots, n$. Here we adopt the convention that $x_j = y_j = 0$ if $j < 0$ or $j > n$. Algebraically, a tensor train is an element in the image of

$$\mathrm{Con} : \bigotimes_{j=1}^{n} (\mathbb{R}^{m_{j-1}+1} \otimes \mathbb{R}^{l_j+1} \otimes \mathbb{R}^{*m_j+1}) \to \bigotimes_{j=1}^{n} \mathbb{R}^{l_j+1},$$

where Con is the tensor product of contractions $\mathrm{Con}_j : \mathbb{R}^{*m_j+1} \otimes \mathbb{R}^{m_j+1} \to \mathbb{R}$. Again, we adopt the convention that $\mathbb{R}^{m_0+1} = \mathbb{R}$.

Assume that $h \in \bigotimes_{i=1}^{n} Y_i$ can be expressed as

$$h = h_1 \otimes \cdots \otimes h_n,$$

where $h_i \in Y_i$, i.e., h is a rank one tensor.

On the one hand, we can write

$$\begin{aligned} T(f_1,\dots,f_n,h) &= \sum_{y_1=0}^{m_1}\cdots\sum_{y_n=0}^{m_n} h(y_1,\dots,y_n)\prod_{i=1}^{n} g_i(x_i,x_{i+1},\dots,x_n,y_i) \\ &= \sum_{y_1=0}^{m_1}\cdots\sum_{y_n=0}^{m_n}(\prod_{i=1}^{n} h_{i-1}(y_{i-1})g_i(x_i,x_{i+1},\dots,x_n,y_i)) \end{aligned}$$

Here we take $h_0 = h_n$ and $y_0 = y_n$. We set $\tilde{x_i} = (x_i, x_{i+1}, \dots, x_n)$ and

$$b_i(y_{i-1}, \tilde{x_i}, y_i) = h_{i-1}(y_{i-1})g_i(x_i, x_{i+1}, \dots, x_n, y_i),$$

then

$$\sum_{y_1=0}^{m_1}\cdots\sum_{y_n=0}^{m_n}(\prod_{i=1}^{n} h_{i-1}(y_{i-1})g_i(x_i,x_{i+1},\dots,x_n,y_i)) = \sum_{y_1=0}^{m_1}\cdots\sum_{y_n=0}^{m_n}\prod_{i=1}^{n} b_i(y_{i-1},\tilde{x_i},y_i) \tag{4.3.2}$$

On the other hand, we also have

$$\begin{aligned} T(f_1,\dots,f_n,h) &= \sum_{y_1=0}^{m_1}\cdots\sum_{y_n=0}^{m_n} h(y_1,\dots,y_n)\prod_{i=1}^{n} g_i(x_i,x_{i+1},\dots,x_n,y_i) \\ &= \sum_{y_1=0}^{m_1}\cdots\sum_{y_{n-1}=0}^{m_{n-1}}(\prod_{i=1}^{n-1} h_{i-1}(y_{i-1})g_i(x_i,x_{i+1},\dots,x_n,y_i)) \\ &\qquad (h_{n-1}(y_{n-1})\sum_{y_n=0}^{m_n} h_n(y_n)g_n(x_n,y_n)) \end{aligned}$$

If we set $\tilde{x_i} = (x_i, x_{i+1}, \ldots, x_n)$ and take

$$b_i(y_{i-1}, \tilde{x_i}, y_i) = \begin{cases} h_i(y_{i-1}) g_i(x_i, x_{i+1}, \ldots, x_n, y_i), \text{ if } 1 \leq i \leq n-1, \\ h_{n-1}(y_{n-1}) \sum_{y_n=0}^{m_n} h_n(y_n) g_n(x_n, y_n), \text{ if } i = n. \end{cases}$$

Then

$$\sum_{y_1=0}^{m_1} \cdots \sum_{y_{n-1}=0}^{m_{n-1}} \prod_{i=1}^{n} h_i(y_{i-1}) g_i(x_i, x_{i+1}, \ldots, x_n, y_i) = \sum_{y_1=0}^{m_1} \cdots \sum_{y_{n-1}=0}^{m_{n-1}} \prod_{i=1}^{n} b_i(y_{i-1}, \tilde{x_i}, y_i) \tag{4.3.3}$$

By equations (4.3.3) and (4.3.2) we have

Proposition 4.3.1. *The tensor $T(g_1, \ldots, g_n, h)$ is a projection of a linear combination of tensor trains and matrix product states, respectively.*

Proof. We will give a proof for tensor trains. The proof for matrix product states is essentially the same. For a given $h \in \bigotimes_{i=1} Y_i$ we can write h as a linear combination of rank one tensors in $\bigotimes_{i=1} Y_i$ hence it is sufficient to prove the statement for rank one tensor $h \in \bigotimes_{i=1} Y_i$. Assume that $h = h_1 \otimes \cdots \otimes h_n \in \bigotimes_{i=1} Y_i$ with $h_i \in Y_i$. We define

$$b_i(y_{i-1}, \tilde{x_i}, y_i) = \begin{cases} h_i(y_{i-1}) g_i(x_i^i, x_{i+1}^i, \ldots, x_n^i, y_i), \text{ if } 1 \leq i \leq n-1, \\ h_{n-1}(y_{n-1}) \sum_{y_n=0}^{m_n} h_n(y_n) g_n(x_n^n, y_n), \text{ if } i = n, \end{cases}$$

where $\tilde{x_i} = (x_i^i, \ldots, x_n^i)$ and $x_j^i, 1 \leq i, j \leq n$ are all independent variables. We define the tensor train

$$TT = \sum_{y_1=0}^{m_1} \cdots \sum_{y_{n-1}=0}^{m_{n-1}} \prod_{i=1}^{n} b_i(y_{i-1}, \tilde{x_i}, y_i).$$

It is clear from (4.3.3) that T is the image $\mathrm{Proj}_1 \circ \cdots \circ \mathrm{Proj}_n(TT)$. □

Corollary 4.3.2. *If $h \in \bigotimes_{i=1} Y_i$ is a rank one tensor then $T(g_1, \ldots, g_n, h)$ is a projection of a tensor train.*

4.4 Tensor rank of Yates's method

Since $T : \bigotimes_{i=1}^{n} X_i^{\otimes i} \otimes (Y_1 \otimes Y_1^*) \otimes \cdots \otimes (Y_n \otimes Y_n^*) \to \bigotimes_{i=1}^{n} X_i$ defined by (4.2.1) is a multilinear map, it corresponds to a tensor

$$\mu_T \in \bigotimes_{i=1}^{n} (X_i^{*\otimes i} \otimes X_i) \otimes (Y_1 \otimes Y_1^*) \otimes \cdots \otimes (Y_n \otimes Y_n^*),$$

defined by $\mu_T(f_1, \ldots, f_n, h, \alpha_1, \ldots, \alpha_n) = T(f_1, \ldots, f_n, h)(x_1, \ldots, x_n)$ where $\alpha_i \in X_i^* \cong \mathbb{R}^{*l_i+1}$ is the projection from $\mathbb{R}^{l_i+1}$ to its x_i-th component. We will determine the tensor rank of μ_T in this section.

We have seen in (4.2.5) that T has factorization

$$T = \mathrm{Proj}_1 \circ \cdots \circ \mathrm{Proj}_n \circ \mathrm{Con}_1 \circ \cdots \mathrm{Con}_n,$$

where $\mathrm{Proj}_i : X_i^{\otimes i} \to X_i$ is the projection and $\mathrm{Con}_i : Y_i^* \otimes Y_i \to \mathbb{R}$ is the contraction map for all $i = 1, 2, \ldots, n$. Since both Proj_i and Con_i are bilinear maps, they correspond to tensors $\mu_{\mathrm{Proj}_i} \in X_i^{*\otimes i} \otimes X_i$ and $\mu_{\mathrm{Con}_i} \in Y_i^* \otimes Y_i$.

Lemma 4.4.1. *We can write μ_T as*

$$\mu_T = \bigotimes_{i=1}^{n} \mu_{\mathrm{Proj}_i} \otimes \mu_{\mathrm{Con}_i}$$

Proof. This is straightforward by the definition of $\mu_T, \mu_{\mathrm{Proj}_i}$ and μ_{Con_i}. □

Lemma 4.4.2. *[79] If $F : U_1 \times \cdots \times U_n \to W$ is a multilinear map such that the image of F spans W then*

$$\mathrm{rank}(\mu_F) \geq \dim W$$

We can write Proj_i and Con_i explicitly. Let $v_0, \ldots, v_{l_i}$ (resp. $w_0, \ldots, w_{m_i}$) be the fixed basis of $X_i \cong \mathbb{R}^{l_i+1}$ (resp. $Y_i \cong \mathbb{R}^{m_i+1}$) and let $\alpha_0, \ldots, \alpha_{l_i}$ (resp. $\beta_0, \ldots, \beta_{m_i}$) be the dual

basis of X_i^* (resp. Y_i^*). Then we have

$$\begin{aligned} \mu_{\mathrm{Proj}_i} &= \sum_{j=0}^{l_i} \alpha_j^{\otimes i} \otimes v_j, \text{ and} \\ \mu_{\mathrm{Con}_i} &= \sum_{j=0}^{m_i} \beta_j \otimes w_j. \end{aligned} \tag{4.4.1}$$

Therefore,

$$\begin{aligned} \mathrm{rank}(\mu_{\mathrm{Proj}_i}) &\leq l_i + 1, \\ \mathrm{rank}(\mu_{\mathrm{Con}_i}) &\leq m_i + 1. \end{aligned}$$

Apply Lemma 4.4.2 to Proj_i and $\widehat{\mathrm{Con}_i}$ where $\widehat{\mathrm{Con}_i} : Y_i \to Y_i$ is the identity map and notice that $\mu_{\widehat{\mathrm{Con}_i}} = \mu_{\mathrm{Con}_i}$, we see that

$$\begin{aligned} \mathrm{rank}(\mu_{\mathrm{Proj}_i}) &= l_i + 1, \\ \mathrm{rank}(\mu_{\mathrm{Con}_i}) &= m_i + 1. \end{aligned}$$

By Lemma 4.4.1 we obtain

$$\mathrm{rank}(\mu_T) \leq \prod_{i=1}^{n} (l_i + 1)(m_i + 1).$$

Indeed, we can prove that the equality holds.

Theorem 4.4.3. *The rank of μ_T is equal to $\prod_{i=1}^{n}(l_i + 1)(m_i + 1)$.*

Proof. We consider the multilinear map $\hat{T} : \bigotimes_{i=1}^{n} X_i^{\otimes i} \otimes Y_i \to \bigotimes_{i=1}^{n} X_i \otimes Y_i$ defined by

$$\hat{T} = \bigotimes_{i=1}^{n} \mathrm{Proj}_i \otimes \widehat{\mathrm{Con}_i}.$$

It is clear that $\mu_{\hat{T}} = \mu_T$ and that $\hat{T}$ is surjective. Hence we obtain that

$$\operatorname{rank}(\mu_T) \geq \dim \bigotimes_{i=1}^{n} X_i \otimes Y_i = \prod_{i=1}^{n} (l_i + 1)(m_i + 1).$$

□

We remark here that the tensor rank of μ_T measures the least number of multiplications needed to evaluate $T(g_1, \ldots, g_n, h)$. Indeed, a rank one tensor in

$$\bigotimes_{i=1}^{n} (X_i^{*\otimes i} \otimes X_i) \otimes (Y_1 \otimes Y_1^*) \otimes \cdots \otimes (Y_n \otimes Y_n^*)$$

costs $\binom{n+1}{2} + n - 1$ multiplications to evaluate when it is regarded as a multilinear map

$$\bigotimes_{i=1}^{n} {X_i}^{\otimes i} \otimes (Y_1 \otimes Y_1^*) \otimes \cdots \otimes (Y_n \otimes Y_n^*) \to \bigotimes_{i=1}^{n} X_i.$$

For example, if $T := \alpha \otimes \beta \otimes w \in U^* \otimes V^* \otimes W$ is a rank one tensor, then when we regard T as a bilinear map $F : U \otimes V \to W$, we see that to evaluate $F(u, v)$ it is the same as evaluating $T(u, v) = \alpha(u)\beta(v)w$ which only costs one multiplication $\alpha(u)\beta(v)$. Hence we may restate Theorem 4.4.3 as

Theorem 4.4.4. *The multilinear map* $T : \bigotimes_{i=1}^{n} X_i^{\otimes i} \otimes (Y_1 \otimes Y_1^*) \otimes \cdots \otimes (Y_n \otimes Y_n^*) \to \bigotimes_{i=1}^{n} X_i$ *defined by* (4.2.1) *requires*

$$N := (\binom{n+1}{2} + n - 1) \prod_{i=1}^{n} (l_i + 1)(m_i + 1)$$

multiplications to evaluate.

Notice that Theorem 4.4.4 does not imply that the function $f(x_1, \ldots, x_n)$ requires N multiplications to evaluate for fixed $g_1, \ldots, g_n, h$. Readers should keep in mind that when we evaluate T, our inputs are functions $g_1, \ldots, g_n, h$ and our output is the tensor $(f(x_1, \ldots, x_n))$

where $x_i \in [l_i]$.

4.5 Yates's method as Matrix Multiplications

By looking at (4.1.2) closely, Yates's method is equivalent to performing n matrix multiplications. We assume $(x_i, \ldots, x_n) \in [l_i] \times \cdots \times [l_n]$ as in Section 4.2.

The first matrix multiplication is

$$\begin{bmatrix} f_1(0,0,\ldots,0) & \cdots & f_1(0,m_1,\ldots,m_{n-1}) \\ \vdots & \ddots & \vdots \\ f_1(l_n,0,\ldots,0) & \cdots & f_1(l_n,m_1,\ldots,m_{n-1}) \end{bmatrix}$$

$$= \begin{bmatrix} g_n(0,0) & \cdots & g_n(0,m_n) \\ \vdots & \ddots & \vdots \\ g_n(l_n,0) & \cdots & g_n(l_n,m_n) \end{bmatrix} \begin{bmatrix} f_0(0,\ldots,0,0) & \cdots & f_0(m_1,\ldots,m_{n-1},0) \\ \vdots & \ddots & \vdots \\ f_0(0,\ldots,0,m_n) & \cdots & f_0(m_1,\ldots,m_{n-1},m_n) \end{bmatrix},$$

with second matrix multiplication being

$$\begin{bmatrix} f_2(0,0,0\ldots,0) & \cdots & f_2(0,0,m_1,\ldots,m_{n-2}) \\ \vdots & \ddots & \vdots \\ f_2(l_{n-1},l_n,0,\ldots,0) & \cdots & f_2(l_{n-1},l_n,m_1,\ldots,m_{n-2}) \end{bmatrix}$$

$$= \begin{bmatrix} g_{n-1}(0,0,0) & \cdots & g_{n-1}(0,0,m_{n-1}) \\ \vdots & \ddots & \vdots \\ g_{n-1}(l_{n-1},l_n,0) & \cdots & g_{n-1}(l_{n-1},l_n,m_{n-1}) \end{bmatrix} \cdot$$

$$\begin{bmatrix} f_1(0,\ldots,0,0) & \cdots & f_1(l_n,m_1\ldots,m_{n-2},0) \\ \vdots & \ddots & \vdots \\ f_1(0,\ldots,0,m_{n-1}) & \cdots & f_1(l_n,m_1\ldots,m_{n-2},m_{n-1}) \end{bmatrix}. \quad (4.5.1)$$

The subsequent matrix operations are defined similarly.

Hence we have multiplication of a $(l_n+1)\times(m_n+1)$ matrix with a $(m_n+1)\times(m_1+1)\ldots(m_{n-1}+1)$ matrix, and then multiplication of a $(l_{n-1}+1)\times(m_{n-1}+1)(m_n+1)$ matrix with a $(m_{n-1}+1)(m_n+1)\times(l_n+1)(m_1+1)\ldots(m_{n-2}+1)$ matrix, and so on.

In general, in the k-th step, we have multiplication of a $(l_{n-k+1}+1)\times(m_{n-k+1}+1)\ldots(m_n+1)$ matrix with a $(m_{n-k+1}+1)\ldots(m_n+1)\times(l_{n-k+2}+1)\ldots(l_n+1)(m_1+1)\ldots(m_{n-k}+1)$ matrix. Therefore the time complexity of the Yates's method can be computed as the aggregation of the complexity of these n matrix multiplications.

For the case of $m_1=\cdots=m_n=l_1=\cdots=l_n=a-1$, the k-th step amounts to perform multiplication of an $a\times a^k$ matrix with a $a^k\times a^{n-1}$ matrix. By results in [80], Yates's method can be done in $O(a^{n+1+o(1)}\sum_{k=1}^{n}a^{\frac{k}{0.30298}})$ time.

REFERENCES

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

[2] Absil, P.-A., Mahony, R., & Sepulchre, R. (2004) Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Appl. Math.*, **80**, no. 2, pp. 199–220.

[3] Absil, P.-A., Mahony, R., Sepulchre, R., & Van Dooren, P. (2002) A Grassmann–Rayleigh quotient iteration for computing invariant subspaces. *SIAM Rev.*, **44**, no. 1, pp. 57–73.

[4] Achar, P. N. & Rider, L. (2015) Parity sheaves on the affine Grassmannian and the Mirković–Vilonen conjecture. *Acta Math.*, **215**, no. 2, pp. 183–216.

[5] Alekseevsky, D. & Arvanitoyeorgos, A. (2007) Riemannian flag manifolds with homogeneous geodesics. *Trans. Amer. Math. Soc.*, **359**, no. 8, pp. 3769–3789.

[6] Amari, S.-I. (2016) *Information Geometry and its Applications*, Applied Mathematical Sciences, **194**, Springer, Japan.

[7] Amari, S.-I., Barndorff-Nielsen, O. E., Kass, R. E., Lauritzen, S. L., & Rao, C. R. (1987) *Differential Geometry in Statistical Inference*, Institute of Mathematical Statistics Lecture Notes Monograph Series, **10**, IMS, Hayward, CA.

[8] Amari, S.-I. & Nagaoka, H. (2000) *Methods of Information Geometry*, Translations of Mathematical Monographs, **191**, American Mathematical Society, Providence, RI.

[9] Balzano, L., Nowak, R., & Recht, B. (2010) Online identification and tracking of subspaces from highly incomplete information. *Annual Allerton Conf. Commun. Control Comput.*, **48**, pp. 704–711.

[10] Belkin, M. & Niyogi, P. (2001) Laplacian eigenmaps and spectral techniques for embedding and clustering. *Proc. Adv. Neural Inform. Process. Systems*, **14**, pp. 586–691.

[11] Björck, Å. & Golub, G. H. (1973) Numerical methods for computing angles between linear subspaces. *Math. Comp.*, **27** (1973), no. 123, pp. 579–594.

[12] Chikuse, Y. (2012) *Statistics on Special Manifolds*, Lecture Notes in Statistics, **174**, Springer, New York, NY.

[13] Chirikjian, G. S. & Kyatkin, A. B. (2001) *Engineering Applications of Noncommutative Harmonic Analysis*, CRC Press, Boca Raton, FL.

[14] Da Silva, N. P. & Costeira, J. P. (2009) The normalized subspace inclusion: Robust clustering of motion subspaces. *Proc. IEEE Int. Conf. Comput. Vis.* (ICCV), **12**, pp. 1444–1450.

[15] Edelman, A., Arias, T., & Smith, S. T. (1999) The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, **20**, no. 2, pp. 303–353.

[16] Elhamifar, E. & Vidal, R. (2013) Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, **35**, no. 11, pp. 2765–2781.

[17] Frenkel, E. & Gaitsgory, D. (2009) Localization of $\mathfrak{g}$-modules on the affine Grassmannian. *Ann. of Math.*, **170**, no. 3, pp. 1339–1381.

[18] Gelman, A. & Meng, X.-L. (1998) Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Stat. Sci.*, **13**, no. 2, pp. 163–185.

[19] Golub, G. & Van Loan, C. (2013) *Matrix Computations*, 4th Ed., John Hopkins University Press, Baltimore, MD.

[20] Gordon, C. S. (1996) Homogeneous Riemannian manifolds whose geodesics are orbits. pp. 155–174, S. Gindikin, Ed., *Topics in Geometry*, Progress in Nonlinear Differential Equations and their Applications, **20**, Birkhäuser, Boston, MA.

[21] Griffiths, P. & Harris, J. (1994) *Principles of Algebraic Geometry*, John Wiley, New York, NY.

[22] Hamm, J. & Lee, D. D. (2008) Grassmann discriminant analysis: A unifying view on subspace-based learning. *Proc. Internat. Conf. Mach. Learn.* (ICML), **25**, pp. 376–383.

[23] Hansen, T. F. & Houle, D. (2008) Measuring and comparing evolvability and constraint in multivariate characters. *J. Evolution. Biol.*, **21**, no. 5, pp. 1201–1219.

[24] Haro, G., Randall, G., & Sapiro, G. (2006) Stratification learning: Detecting mixed density and dimensionality in high dimensional point clouds. *Proc. Adv. Neural Inform. Process. Syst.* (NIPS), **26**, pp. 553–560.

[25] He, Q., Kong, F. & Yan, R. (2007) Subspace-based gearbox condition monitoring by kernel principal component analysis. *Mech. Systems Signal Process.*, **21**, no. 4, pp. 1755–1772.

[26] Helmke, U., Hüper, K., & Trumpf, J. (2007) Newton's method on Graßmann manifolds. *preprint*, `https://arxiv.org/abs/0709.2205`.

[27] Hirsch, M. (1976) *Differential Topology*, Springer, New York, NY.

[28] Hüper, K., Helmke, U., & Herzberg, S. (2010) On the computation of means on Grassmann manifolds. *Proc. Int. Symp. Math. Theory Networks Syst.* (MTNS), **19**, pp. 2439–2441.

[29] Hyvärinen, A. (2005) Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, **6**, pp. 695–709.

[30] Jost, J. (2011) *Riemannian Geometry and Geometric Analysis*, 6th Ed., Springer, Heidelberg.

[31] Karcher, H. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, **30.5**, pp. 509-541

[32] Klain, D. A. & Rota, G.-C. (1997) *Introduction to Geometric Probability*, Lezioni Lincee, Cambridge University Press, Cambridge.

[33] Koev, P. & Edelman, A. (2006) The efficient evaluation of the hypergeometric function of a matrix argument. *Math. Comp.*, **75**, no. 254, pp. 833–846.

[34] Lam, T. (2008) Schubert polynomials for the affine Grassmannian. *J. Amer. Math. Soc.*, **21**, no. 1, pp. 259–281.

[35] Lerman, G. & Zhang, T. (2011) Robust recovery of multiple subspaces by geometric l_p minimization. *Ann. Statist.*, **39**, no. 5, pp. 2686–2715.

[36] Li, H. & Li, A. (2012) Utilizing improved Bayesian algorithm to identify blog comment spam. *Proc. IEEE Symp. Robot. Appl.* (ISRA), **1**, pp. 423–426.

[37] Ma, Y., Yang, A., Derksen, H., & Fossum, R. (2008) Estimation of subspace arrangements with applications in modeling and segmenting mixed data. *SIAM Rev.*, **50**, no. 3, pp. 413–458.

[38] Marriott, P. & Salmon, M. (Eds) (2000) *Applications of Differential Geometry to Econometrics*, Cambridge University Press, Cambridge.

[39] Milnor, J. W. & Stasheff, J. D. (1974) *Characteristic Classes*, Annals of Mathematics Studies, **76**, Princeton University Press, Princeton, NJ.

[40] Murray, M. K. & Rice, J. W. (1993) *Differential Geometry and Statistics*, Monographs on Statistics and Applied Probability, **48**, Chapman & Hall, London.

[41] Nicolaescu, L. I. (2007) *Lectures on the Geometry of Manifolds*, 2nd Ed., World Scientific, Hackensack, NJ.

[42] Peng, M., Bu, D., & Wang, Y. (2010) The measure of income mobility in vector space. *Physics Procedia*, **3**, no. 5, pp. 1725–1732.

[43] Roweis, S. T. & Saul, L. K. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**, no. 5500, pp. 2323–2326.

[44] Sharafuddin, E., Jiang, N., Jin, Y., & Zhang, Z.-L. (2010) Know your enemy, know yourself: Block-level network behavior profiling and tracking. *Proc. IEEE Global Telecomm. Conf.* (GLOBECOM), **53**, pp. 1–6.

[45] St. Thomas, B., Lin, L., Lim, L.-H., & Mukherjee, S. (2014) Learning subspaces of different dimensions. Preprint arxiv: 1404.6841.

[46] Sun, X., Wang, L., & Feng, J. (2007) Further results on the subspace distance. *Pattern Recognition*, **40**, no. 1, pp. 328–329.

[47] Tenenbaum, J., De Silva, V., & Langford, J. (2000) A global geometric framework for nonlinear dimensionality reduction, *Science*, **290**, no. 5500, pp. 2319–2323.

[48] Tomé, A. M., Teixeira, A. R., Figueiredo, N., Santos, I. M., Georgieva, P., & Lang, E. W. (2010) SSA of biomedical signals: A linear invariant systems approach. *Stat. Interface*, **3**, no. 3, pp. 345–355.

[49] Tyagi, H., Vural, E., & Frossard, P. (2013) Tangent space estimation for smooth embeddings of Riemannian manifolds. *Inf. Inference*, **2**, no. 1, pp. 69–114

[50] Vidal, R., Ma, Y., & Sastry, S. (2005) Generalized principal component analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, no. 12, pp. 1945–1959.

[51] Wang, R., Shan, S., Chen, X., & Gao, W. (2008) Manifold-manifold distance with application to face recognition based on image set. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (CVPR), **26**, pp. 1–8.

[52] Wang, L., Wang, X., & Feng, J. (2006) Subspace distance analysis with application to adaptive Bayesian algorithm for face recognition. *Pattern Recognition*, **39**, no. 3, pp. 456–464.

[53] Wong, Y.-C. (1967) Differential geometry of Grassmann manifolds. *Proc. Nat. Acad. Sci.*, **57**, no. 3, pp. 589–594.

[54] Ye, K. & Lim, L.-H. (2016) Schubert varieties and distances between linear spaces of different dimensions. *SIAM J. Matrix Anal. Appl.*, (2016), to appear, `https://arxiv.org/abs/1407.0900`.

[55] Zuccon, G., Azzopardi, L. A., & van Rijsbergen, C. J. (2009) Semantic spaces: Measuring the distance between different subspaces. *Quantum Interaction*, Lecture Notes in Artificial Intelligence, **5494**, Springer, Berlin, 2009, pp 225-236.

[56] S. Helgason, *Differential geometry, Lie groups, and symmetric spaces*, Acad. Press (1978) .

[57] S. Kobayashi and K. Nomizu, *Foundations of differential geometry* volume 1 and 2, Interscience (1969).

[58] W. M. Boothby, "An introduction to differentiable manifolds and Riemannian geometry", Academic Press New York-London (1975).

[59] D. Monk, "The geometry of flag manifolds", , *Proc. London Math. Soc.* (1959), **3**, no. 9, pp. 253–286.

[60] A. Borel, *Linear algebraic groups*, Graduate Texts in Mathematics, second edition, **126**, Springer-Verlag, New York (1991).

[61] R. Goodman and N. R. Wallach, *Symmetry, Representations, and Invariants*, Graduate Texts in Mathematics, **255**, Springer-Verlag, New York (2009).

[62] O. Kowalski and J. Szenthe, "On the existence of homogeneous geodesics in homogeneous Riemannian manifolds", *Geometriae Dedicata* (2000), **81**, no. 1, pp. 209–214.

[63] J. Cheeger and D. G. Ebin, *Comparison theorems in Riemannian geometry*, Revised reprint of the 1975 original, AMS Chelsea Publishing, Providence, RI (2008).

[64] K. Tojo, "Totally geodesic submanifolds of naturally reductive homogeneous spaces", *Tsukuba Journal of Mathematics* (1996), **20**, no. 1 pp. 181–90.

[65] A. Edelman, T. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, **20** (1999), no. 2, pp. 303–353.

[66] A. Sheldon, *Linear algebra done right*, Undergraduate Texts in Mathematics, Springer Science & Business Media (1997).

[67] Y. Nishimori, A. Shotaro and M. D. Plumbley, "Riemannian optimization method on the flag manifold for independent subspace analysis", *Independent component analysis and blind signal separation: 6th international conference, ICA 2006, Charleston, SC, USA, March 5-8, 2006, Proceedings*, Springer Berlin Heidelberg.

[68] B. ONeill, *Semi-Riemannian geometry*, Academic Press (1983).

[69] D. Alekseevsky and A. Arvanitoyeorgos, "Riemannian flag manifolds with homogeneous geodesics", *Transactions of the American Mathematical Society* (2007), **359**, no. 8, pp. 3769-3789.

[70] E. Stiefel,"Richtungsfelder und Fernparallelismus in -dimensionalen Mannigfaltigkeiten", *Comm. Math. Helv.* (19351936), **8**, no. 4, pp. 305-353.

[71] Vidal, R., Ma, Y., & Sastry, S. (2005) Generalized principal component analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, no. 12, pp. 1945–1959.

[72] Wang, R., Shan, S., Chen, X., & Gao, W. (2008) Manifold-manifold distance with application to face recognition based on image set. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (CVPR), **26**, pp. 1–8.

[73] Wang, L., Wang, X., & Feng, J. (2006) Subspace distance analysis with application to adaptive Bayesian algorithm for face recognition. *Pattern Recognition*, **39**, no. 3, pp. 456–464.

[74] Balzano, L., Nowak, R., & Recht, B. (2010) Online identification and tracking of subspaces from highly incomplete information. *Annual Allerton Conf. Commun. Control Comput.*, **48**, pp. 704–711.

[75] Vandereycken, B. (2013) Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, pp. 1214-1236.

[76] Elhamifar, E. & Vidal, R. (2013) Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, **35**, no. 11, pp. 2765–2781.

[77] Ammar, G. & and Clyde M. (1986) The geometry of matrix eigenvalue methods *Acta Applicandae Mathematica* **5.3** (1986), pp. 239-278.

[78] Jordan, J. & Uwe H. (2002) Controllability of the QR-algorithm on Hessenberg flags *Proc. of the Inter. Sym. on Mathematical Theory of Network and Systems* (MTNS)

[79] K. Ye and L. H. Lim, "Fast structured matrix computations: tensor rank and Cohn-Umans method", preprint (2015).

[80] F. Le Gall, "Faster Algorithms for Rectangular Matrix Multiplication", Foundations of Computer Science (FOCS), 2012.