NUMERICAL ALGORITHMS ON THE AFFINE GRASSMANNIAN

LEK-HENG LIM, KEN SZE-WAI WONG, AND KE YE

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 this article 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.

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 $\operatorname{Graff}(k, n)$, an analogue of the usual Grassmannian $\operatorname{Gr}(k, n)$ that parameterizes k-dimensional linear subspaces in $\mathbb{R}^n$.

The main impetus for this article is the observation that many statistical estimation problems (see Examples 1.1–1.4) involve a search for linear relations among variables and may ultimately be formulated as a problem of finding one (or more) 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.

The affine Grassmannian is a relatively obscured object, especially when compared to its ubiquitous cousin, the Grassmannian. Nevertheless, it is $\operatorname{Graff}(k, n)$, which like $\mathbb{R}^n$ is a non-compact manifold, that is the natural generalization of Euclidean space (points are 0-dimensional affine subspaces and so $\operatorname{Graff}(0, n) = \mathbb{R}^n$). The non-compactness makes $\operatorname{Graff}(k, n)$ harder to study than $\operatorname{Gr}(k, n)$, which is compact. The main objective of our article is to develop the foundations for working with the affine Grassmannian, particularly distances (Sections 4 and 5), probability distributions (Section 6), and optimization algorithms (Section 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

---

2010 *Mathematics Subject Classification.* 62H12, 14M15, 90C30, 62H10, 68T10.

Key words and phrases. affine Grassmannian, affine Schubert variety, affine subspaces, Bingham/Langevin distribution, distances and metrics, manifold optimization, multivariate analysis.

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 applicability 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 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 [9, 14, 16, 48, 23, 24, 25, 22, 28, 35, 36, 37, 42, 44, 45, 46, 49, 50, 52, 51, 55] to affine subspaces.

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 such context manifolds are by-and-large regarded as collections of tangent spaces[1] [49]. 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 5), which parameterizes all affine spaces of all dimensions.

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 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 $\bigl\{ \bigl[\begin{smallmatrix} z \\ \beta^\mathsf{T} z\end{smallmatrix}\bigr] \in \mathbb{R}^{p+1} : z \in \mathbb{R}^p \bigr\} + \beta_{p+1}e_{p+1}$, 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 $\operatorname{span}\Bigl(\Bigl[\begin{smallmatrix} I_p \\ \beta^\mathsf{T}\end{smallmatrix}\Bigr]\Bigr) + \beta_{p+1}e_{p+1}$ 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 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 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 $k \le p$, 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 $k = 1, \dots, p$ as orthonormal basis of $\operatorname{span}(Z_k)$.

[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.

Example 1.4 (Support Vector Machine)**.** Let $\{(x_i, y_i) : x_i \in \mathbb{R}^p,\ y_i = \pm 1,\ i = 1, \dots, n\}$ be a training set for binary classification. The best separating hyperplane is given by $w^{\mathsf{T}} x - \beta = 0$, where $(w, \beta) \in \mathbb{R}^p \times \mathbb{R}$ can be found by minimizing $\|w\|$ subject to $y_i(w^{\mathsf{T}} x_i - b) \geq 1$ for all $i = 1, \dots, n$. 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 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. Affine Grassmannian

The Grassmannian of affine subspaces or affine Grassmannian[2] was first described in [32] but has received relatively little attention compared to the Grassmannian of linear subspaces $\operatorname{Gr}(k, n)$. 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 $\operatorname{Gr}(k, n)$ for the *Grassmannian* of k-planes in $\mathbb{R}^n$, $\operatorname{V}(k, n)$ for the *Stiefel manifold* of orthonormal k-frames, and $\operatorname{O}(n) := \operatorname{V}(n, n)$ for the *orthogonal group*. We may regard $\operatorname{V}(k, n)$ as a homogeneous space,

$$\operatorname{V}(k, n) \cong \operatorname{O}(n) / \operatorname{O}(n - k), \tag{2.1}$$

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

$$\operatorname{Gr}(k, n) \cong \operatorname{V}(k, n) / \operatorname{O}(k) \cong \operatorname{O}(n) / \big(\operatorname{O}(n - k) \times \operatorname{O}(k)\big). \tag{2.2}$$

By (2.2), $\mathbf{A} \in \operatorname{Gr}(k, n)$ may be identified with the equivalence class of its orthonormal k-frames $\{AQ \in \operatorname{V}(k, n) : Q \in \operatorname{O}(k)\}$. Note $\operatorname{span}(AQ) = \operatorname{span}(A)$ for $Q \in \operatorname{O}(k)$.

[2] 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 2.1. If desired, ‘Grassmannian of affine subspaces’ may be used to avoid ambiguity.

Definition 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

$$\mathbf{A} + b \coloneqq \{\lambda_1 a_1 + \dots + \lambda_k a_k + b \in \mathbb{R}^n : \lambda_1, \dots, \lambda_k \in \mathbb{R}\}. \tag{2.3}$$

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 $A \in \operatorname{V}(k,n)$ 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 $Q \in \operatorname{O}(k)$ 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 (2.3) is called its *affine coordinates*.

Proposition 2.2. $\operatorname{Graff}(k,n)$ *is a smooth manifold.*

Proof. Let $\mathbf{A} + b \in \operatorname{Graff}(k,n)$ 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 $\mathbf{X} + y \in \operatorname{Graff}(k,n)$ whose affine coordinates $[X, y_0]$ have nonzero $k \times k$ leading principal minors. Then U is an open subset of $\operatorname{Graff}(k,n)$ 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

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

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 $\varphi_i : U_i \to \mathbb{R}^{(n-k)(k+1)}$, $i = 1, 2$. $\square$

It turns out that $\operatorname{Graff}(k,n)$ may be viewed as a vector bundle over $\operatorname{Gr}(k,n)$. 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

$$0 \to S \to E \to Q \to 0. \tag{2.4}$$

In the context of Grassmannians, there is a special vector bundle over $\operatorname{Gr}(k,n)$, called the *tautological bundle*, whose fiber over $\mathbf{A} \in \operatorname{Gr}(k,n)$ is simply $\mathbf{A}$ itself. One may view this as a subbundle of the *trivial vector bundle* $\operatorname{Gr}(k,n) \times \mathbb{R}^n$. If S is the tautological bundle and E is the trivial bundle in (2.4), then the quotient bundle Q is called the *universal quotient bundle* of $\operatorname{Gr}(k,n)$ [21, 39].

Proposition 2.3. *$\operatorname{Graff}(k,n)$ is the universal quotient bundle on $\operatorname{Gr}(k,n)$.*

Proof. Let $p : \operatorname{Graff}(k,n) \to \operatorname{Gr}(k,n)$, $\mathbf{A} + b \mapsto \mathbf{A}$ 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. Notice that the fiber $p^{-1}(\mathbf{A})$ for $\mathbf{A} \in \operatorname{Gr}(k,n)$ is simply $\mathbb{R}^n/\mathbf{A}$, a linear subspace of dimension $n-k$. Local trivializations of $\operatorname{Graff}(k,n)$ are obtained from local charts of $\operatorname{Gr}(k,n)$ by construction. Hence $\operatorname{Graff}(k,n)$ is a vector bundle over $\operatorname{Gr}(k,n)$. Moreover we have $q : \operatorname{Gr}(k,n) \times \mathbb{R}^n \to \operatorname{Graff}(k,n)$, $(\mathbf{A}, b) \mapsto \mathbf{A} + b$. It is straightforward to check that q is a surjective bundle map and the kernel of q is the tautological vector bundle S over $\operatorname{Gr}(k,n)$, i.e., we have an exact sequence

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

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

By either Proposition 2.2 or Proposition 2.3, we see that $\dim \operatorname{Graff}(k,n) = (n-k)(k+1)$. Unlike $\operatorname{Gr}(k,n)$, $\operatorname{Graff}(k,n)$ is non-compact: take a sequence in $\operatorname{Graff}(k,n)$ represented in coordinates by $[A, mb]$ with $m \in \mathbb{N}$, $A = [a_1,\dots,a_k] \in \operatorname{V}(k,n)$, 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 $\operatorname{E}(n)$ and is the set[3] $\operatorname{O}(n) \times \mathbb{R}^n$ equipped with the group operation $(Q_1,c_1)(Q_2,c_2) = (Q_1Q_2, c_1 + Q_1c_2)$. The *affine Stiefel manifold* is defined to be the product manifold $\operatorname{Vaff}(k,n) := \operatorname{V}(k,n) \times \mathbb{R}^n$. It is a homogeneous space because of the following analogue of (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 2.4. *$\operatorname{Graff}(k,n)$ is a reductive homogeneous Riemannian manifold. In fact, we have the following analogue of* (2.2),

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

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 $Q \cdot \mathbf{A} := \operatorname{span}(QA)$. It is easy to see that $\operatorname{E}(n)$ acts on $\operatorname{Graff}(k,n)$ transitively and so $\operatorname{Graff}(k,n) \cong \operatorname{E}(n)/\operatorname{Stab}_{\mathbf{A}+b}\bigl(\operatorname{E}(n)\bigr)$, where $\operatorname{Stab}_{\mathbf{A}+b}\bigl(\operatorname{E}(n)\bigr)$ is the stabilizer of any fixed affine linear subspace $\mathbf{A}+b \in \operatorname{Graff}(k,n)$ in $\operatorname{E}(n)$. Now $\operatorname{Stab}_{\mathbf{A}+b}\bigl(\operatorname{E}(n)\bigr)$ 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 $\operatorname{Stab}_{\mathbf{A}+b}\bigl(\operatorname{E}(n)\bigr) \cong \operatorname{O}(n-k) \times \operatorname{E}(k)$, and the representation of $\operatorname{Graff}(k,n)$ as a homogeneous Riemannian manifold follows. $\square$

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 1.

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

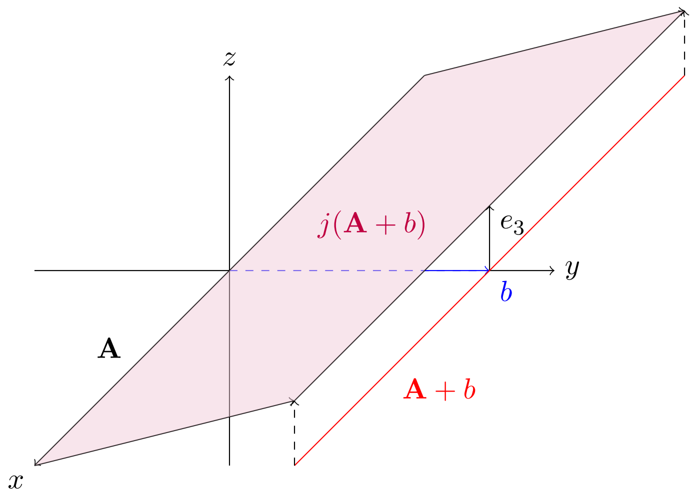

FIGURE 1. Here our linear subspace $\mathbf{A}$ is the x-axis. It is displaced by b along the y-axis to the affine subspace $\mathbf{A}+b$. The embedding $j : \operatorname{Graff}(k,n) \to \operatorname{Gr}(k+1,n+1)$ takes $\mathbf{A}+b$ to the smallest 2-plane containing $\mathbf{A}$ and $b+e_3$, where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

Theorem 2.5. (i) $\operatorname{Graff}(k,n)$ *is an algebraic variety that is irreducible and nonsingular.*

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

$$j : \operatorname{Graff}(k,n) \to \operatorname{Gr}(k+1,n+1), \quad \mathbf{A}+b \mapsto \operatorname{span}(\mathbf{A} \cup \{b+e_{n+1}\}), \tag{2.5}$$

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

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

Proof. Substituting 'smooth' with 'regular' and 'differential manifold' by 'algebraic variety' in the proof of Proposition 2.2, we see that $\operatorname{Graff}(k,n)$ is a nonsingular algebraic variety. Its irreducibility follows from Proposition 2.3 since $\operatorname{Gr}(k,n)$ is irreducible and all fibers of $\operatorname{Graff}(k,n) \to \operatorname{Gr}(k,n)$ are irreducible and of the same dimension. We use 'algebraic variety' is used here in the sense of an abstract algebraic variety, i.e., $\operatorname{Graff}(k,n)$ 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., $\mathbb{R}^n \supseteq \mathbf{A}+b \mapsto \operatorname{span}(\mathbf{A} \cup \{b+e_{n+1}\}) \subseteq \mathbb{R}^{n+1}$. It maps $\mathbb{R}^n$ onto $E_n \coloneqq \operatorname{span}\{e_1,\dots,e_n\} \subset \mathbb{R}^{n+1}$ 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 $X \coloneqq j\bigl(\operatorname{Graff}(k,n)\bigr) \subseteq \operatorname{Gr}(k+1,n+1)$ and set X^c to be the set-theoretic complement of X in $\operatorname{Gr}(k+1,n+1)$. By (ii), $X \cong \operatorname{Graff}(k,n)$. By the definition of X^c, a $(k+1)$-plane $\mathbf{B} \in \operatorname{Gr}(k+1,n+1)$ is in X^c if and only if $\mathbf{B} \subseteq E_n$, which is to say that $X^c = \operatorname{Gr}_{k+1}(E_n) \cong \operatorname{Gr}(k+1,n)$. 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 (2.5) is the most natural one.

It is sometimes desirable to represent elements of $\operatorname{Gr}(k,n)$ 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 6. The Grassmannian has a well-known representation [41, Example 1.2.20] as the set of rank-k orthogonal projection[4] matrices, or, equivalently, the set of trace-k idempotent symmetric matrices:

$$\mathrm{Gr}(k,n) \cong \{P \in \mathbb{R}^{n\times n} : P^{\mathsf{T}} = P^2 = P,\ \mathrm{tr}(P) = k\}. \tag{2.6}$$

Note that $\mathrm{rank}(P) = \mathrm{tr}(P)$ for an orthogonal projection matrix P. A straightforward affine analogue of (2.6) for $\mathrm{Graff}(k,n)$ is simply

$$\mathrm{Graff}(k,n) \cong \{[P,b] \in \mathbb{R}^{n\times(n+1)} : P^{\mathsf{T}} = P^2 = P,\ \mathrm{tr}(P) = k,\ Pb = 0\}, \tag{2.7}$$

where $\mathbf{A} + b \in \mathrm{Graff}(k,n)$ with orthogonal affine coordinates $[A, b_0] \in \mathbb{R}^{n\times(k+1)}$ is represented as the matrix[5] $[AA^{\mathsf{T}}, b_0] \in \mathbb{R}^{n\times(n+1)}$. We will call this the matrix of *projection affine coordinates* for $\mathbf{A} + b$.

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 $\mathrm{Gr}(k,n)$ 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 $S \in \mathrm{GL}(k)$, (ii) an equivalence class of matrices $A \in \mathrm{V}(k,n)$ with orthonormal columns such that $A \sim AQ$ for any $Q \in \mathrm{O}(k)$, (iii) a projection matrix $P \in \mathbb{R}^{n\times n}$ satisfying $P^2 = P^{\mathsf{T}} = P$ and $\mathrm{tr}(P) = k$. 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 are obvious analogues of (i), (ii), and (iii) respectively. In the following we will introduce two more.

For an affine subspace $\mathbf{A}+b \in \mathrm{Graff}(k,n)$, its orthogonal affine coordinates are $[A, b_0] \in \mathrm{V}(k,n)\times\mathbb{R}^n$ 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 $\mathrm{V}(n, k+1)$. With this in mind, we introduce the notion of Stiefel coordinates, which is the most suitable system of coordinates for computations.

Definition 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,

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

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

$$Y_{\mathbf{A}+b} = \begin{bmatrix} A & b_0/\sqrt{1+\|b_0\|^2} \\ 0 & 1/\sqrt{1+\|b_0\|^2} \end{bmatrix} = \begin{bmatrix} A' & b_0'/\sqrt{1+\|b_0'\|^2} \\ 0 & 1/\sqrt{1+\|b_0'\|^2} \end{bmatrix} \begin{bmatrix} Q' & 0 \\ 0 & 1 \end{bmatrix} = Y'_{\mathbf{A}+b}Q \tag{3.1}$$

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

[4] A projection matrix satisfies $P^2 = P$ and an orthogonal projection matrix is in addition symmetric, i.e., $P^{\mathsf{T}} = P$. An orthogonal projection matrix P is not an orthogonal matrix unless $P = I$.

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

Proposition 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,

$$\operatorname{Graff}(k,n) \cong \left\{ \begin{bmatrix} A & b \\ 0 & \gamma \end{bmatrix} \cdot \mathrm{O}(k+1) : \begin{bmatrix} A & b \\ 0 & \gamma \end{bmatrix} \in \mathrm{V}(k+1,n+1) \right\} \tag{3.2}$$

$$\subseteq \mathrm{V}(k+1,n+1)/\,\mathrm{O}(k+1) = \operatorname{Gr}(k+1,n+1). \tag{3.3}$$

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

Proof. The set of equivalence classes on the RHS of (3.2) is the set X in Theorem 2.5(iii) if $\operatorname{Gr}(k+1,n+1)$ is regarded as the homogeneous space in (3.3). $\square$

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 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 $j\bigl(\operatorname{Graff}(k,n)\bigr)$.

Lemma 3.3. (i) *Any matrix of the form $\left[\begin{smallmatrix} A & b \\ 0 & \gamma \end{smallmatrix}\right] \in \mathrm{V}(k+1,n+1)$, 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 $\left[\begin{smallmatrix} A & b \\ 0 & \gamma \end{smallmatrix}\right], \left[\begin{smallmatrix} A' & b' \\ 0 & \gamma' \end{smallmatrix}\right] \in \mathrm{V}(k+1,n+1)$ represent the same affine subspace iff there exists $\left[\begin{smallmatrix} Q' & 0 \\ 0 & 1 \end{smallmatrix}\right] \in \mathrm{O}(k+1)$ 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 $\left[\begin{smallmatrix} A & b \\ 0 & \gamma \end{smallmatrix}\right] \in \mathrm{V}(k+1,n+1)$ 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 (2.7) is not an orthogonal projection matrix. With this in mind, we introduce the following notion.

Definition 3.4. Let $\mathbf{A}+b \in \operatorname{Graff}(k,n)$ 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 $[A,b_0] \in \mathbb{R}^{n\times(k+1)}$,

$$P_{\mathbf{A}+b} = \begin{bmatrix} AA^{\mathsf{T}} + b_0b_0^{\mathsf{T}}/(\|b_0\|^2+1) & b_0/(\|b_0\|^2+1) \\ b_0^{\mathsf{T}}/(\|b_0|^2+1) & 1/(\|b_0\|^2+1) \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}.$$

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 3.2, the next result gives a concrete description of the set $X = j\bigl(\operatorname{Graff}(k,n)\bigr)$ in Theorem 2.5(iii), but in terms of projection coordinates. With this description, $\operatorname{Graff}(k,n)$ may be regarded as a subvariety of $\mathbb{R}^{(n+1)\times(n+1)}$.

Proposition 3.5. *The affine Grassmannian may be represented as a set of $(n+1)\times(n+1)$ orthogonal projection matrices,*

$$\operatorname{Graff}(k,n) \cong \left\{ \begin{bmatrix} P + bb^{\mathsf{T}}/(\|b\|^2+1) & b/(\|b\|^2+1) \\ b^{\mathsf{T}}/(\|b|^2+1) & 1/(\|b\|^2+1) \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)} : \right.$$
$$\left. P \in \mathbb{R}^{n\times n},\ P^{\mathsf{T}} = P^2 = P,\ \operatorname{tr}(P) = k,\ Pb = 0 \right\}. \tag{3.4}$$

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

Proof. Let $\mathbf{A}+b \in \operatorname{Graff}(k,n)$ 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 $\operatorname{Graff}(k,n)$ onto the set of matrices on the RHS of (3.4) with inverse given by $P_{\mathbf{A}+b} \mapsto \operatorname{im}(P) + b_0$. $\square$

The following lemma allows us to check feasibility for the algorithms in Section 7 when we use projection coordinates.

Lemma 3.6. *An orthogonal projection matrix $\begin{bmatrix} S & d \\ d^{\mathsf{T}} & \gamma \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}$ 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, $\begin{bmatrix} S & d \\ d^{\mathsf{T}} & \gamma \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}$ is the matrix of projection coordinates for $\mathbf{A}+b \in \operatorname{Graff}(k,n)$ 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 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 3.3(i) for a matrix of Stiefel coordinates. $\square$

The above proof also shows that projection coordinates are unique even though Stiefel coordinates are not. In principle, they are interchangeable via Lemma 3.7 but in reality, one form is usually more natural than the other for a specific use.

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, 2.3, and 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 $\mathrm{Gr}(k,n) = \mathrm{O}(n)/\bigl(\mathrm{O}(n-k)\times\mathrm{O}(k)\bigr)$ 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 2.4 guarantees the homogeneous space structure of $\mathrm{Graff}(k,n)$. The difficulty in this situation is that $\mathrm{Graff}(k,n) = \mathrm{E}(n)/\bigl(\mathrm{E}(n-k)\times\mathrm{O}(k)\bigr)$ 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 2.3, 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, 2.3, and 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 2.5 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 5.

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

$$d_{\mathrm{Gr}(k,n)}(\mathbf{A},\mathbf{B}) = \Bigl(\sum\nolimits_{i=1}^{k} \theta_i^2\Bigr)^{1/2}, \tag{4.1}$$

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 2.5(ii), we may identify $\mathrm{Graff}(k,n)$ with its image $j\bigl(\mathrm{Graff}(k,n)\bigr)$ in $\mathrm{Gr}(k+1,n+1)$. As a subset of $\mathrm{Gr}(k+1,n+1)$, $\mathrm{Graff}(k,n)$ inherits the Grassmann distance $d_{\mathrm{Gr}(k+1,n+1)}$ on $\mathrm{Gr}(k+1,n+1)$ and we obtain a distance as in Theorem 4.1 that can also be readily computed using SVD. We will show in Theorem 4.3 that this distance is in fact *intrinsic*.

Theorem 4.1. *For any two affine k-flats $\mathbf{A}+b$ and $\mathbf{B}+c \in \mathrm{Graff}(k,n)$,*

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

where j is the embedding in (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

$$d_{\mathrm{Graff}(k,n)}(\mathbf{A}+b, \mathbf{B}+c) = \Bigl(\sum\nolimits_{i=1}^{k+1} \theta_i^2\Bigr)^{1/2}, \tag{4.2}$$

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 3.3(ii), there exist $Q_1, Q_2 \in \mathrm{O}(k+1)$ 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

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

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,

$$Y_{\mathbf{A}+b}^{\mathsf{T}} Y_{\mathbf{B}+c} = U\Sigma V^{\mathsf{T}} \tag{4.3}$$

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 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 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 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. $\square$

Theorem 4.3. *The distance $d_{\operatorname{Graff}(k,n)}$ in Theorem 4.1 is intrinsic with respect to the Riemannian metric on $\operatorname{Graff}(k,n)$ induced from that of $\operatorname{Gr}(k+1,n+1)$.*

Proof. By Theorem 2.5, the complement of $N = \operatorname{Graff}(k,n)$ in M is $\operatorname{Gr}(k+1,n)$ has codimension $k+1 \geq 2$. Hence Proposition 4.2 applies. $\square$

The next theorem plays an important role for our path-following algorithms in Section 7, showing that they will almost never lead to a point outside $\operatorname{Graff}(k,n)$.

Theorem 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 $j\bigl(\operatorname{Graff}(k,n)\bigr)$. Here j is the embedding in (2.5).

Proof. Let $U \in \mathrm{O}(k+1)$ and the diagonal matrix Σ be as in (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 $\gamma(t) = \operatorname{span}\bigl(Y_{\mathbf{A}+b} U \cos(t\Theta) + Q \sin(t\Theta)\bigr)$, 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 $\gamma(t) \in \operatorname{Gr}(k+1,n+1) \setminus j\bigl(\operatorname{Graff}(k,n)\bigr)$ 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 $\operatorname{Graff}(k,n)$. $\square$

Corollary 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*

$$\gamma(t) = j^{-1}\bigl(\operatorname{span}(Y_{\mathbf{A}+b} U \cos t\Theta + Q \sin t\Theta)\bigr), \tag{4.4}$$

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 (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* (4.2) *and its derivative at $t = 0$ is given by*

$$\gamma'(0) = j^{-1}\bigl(Q\Theta U^{\mathsf{T}}\bigr). \tag{4.5}$$

The Grassmann distance in (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 1. The matrices $U, V \in \mathrm{O}(k+1)$ in the right column of Table 1 are the ones in (4.3).

TABLE 1. Distances on $\operatorname{Graff}(k,n)$ in terms of affine principal angles and Stiefel coordinates.

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

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., $\Omega_-(\mathbf{B}):=\{\mathbf{X}\in\operatorname{Gr}(k,n):\mathbf{X}\subseteq\mathbf{B}\}$, and define the required distance to be

$$d_{\operatorname{Gr}(k,n)}\big(\mathbf{A},\Omega_-(\mathbf{B})\big)=\inf_{\mathbf{Y}\in\Omega_-(\mathbf{B})}d_{\operatorname{Gr}(k,n)}(\mathbf{A},\mathbf{Y}),\tag{5.1}$$

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

$$d_{\operatorname{Gr}(l,n)}\big(\mathbf{B},\Omega_+(\mathbf{A})\big)=\inf_{\mathbf{X}\in\Omega_+(\mathbf{A})}d_{\operatorname{Gr}(l,n)}(\mathbf{B},\mathbf{X}),\tag{5.2}$$

the two results are identical. Their comon 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:

$$\delta(\mathbf{A},\mathbf{B})=\Big(\sum_{i=1}^{\min(k,l)}\theta_i^2\Big)^{1/2}.\tag{5.3}$$

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 (5.1) and (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 5.1, Theorems 5.2 and 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* $n \ge k$. 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 $n \ge k$. 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 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 $k \le l \le n$. The *affine Schubert varieties* of l-flats containing $\mathbf{A} + b$ and k-flats contained in $\mathbf{B} + c$ are

$$\Omega_+(\mathbf{A} + b) \coloneqq \bigl\{\mathbf{X} + y \in \operatorname{Graff}(l, n) : \mathbf{A} + b \subseteq \mathbf{X} + y\bigr\}, \tag{5.4}$$

$$\Omega_-(\mathbf{B} + c) \coloneqq \bigl\{\mathbf{Y} + z \in \operatorname{Graff}(k, n) : \mathbf{Y} + z \subseteq \mathbf{B} + c\bigr\}. \tag{5.5}$$

Theorem 5.2. *Let $k \le l \le n$. For any $\mathbf{A} + b \in \operatorname{Graff}(k, n)$ and $\mathbf{B} + c \in \operatorname{Graff}(l, n)$, the following distances are equal,*

$$d_{\operatorname{Graff}(k,n)}\bigl(\mathbf{A} + b, \Omega_-(\mathbf{B} + c)\bigr) = d_{\operatorname{Graff}(l,n)}\bigl(\mathbf{B} + c, \Omega_+(\mathbf{A} + b)\bigr), \tag{5.6}$$

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

$$\delta(\mathbf{A} + b, \mathbf{B} + c) = \Bigl(\sum\nolimits_{i=1}^{\min(k,l)+1} \theta_i(\mathbf{A} + b, \mathbf{B} + c)^2\Bigr)^{1/2}. \tag{5.7}$$

The affine principal angles $\theta_1, \dots, \theta_{\min(h,l)+1}$ are as defined in Theorem 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 (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 1 may be extended to affine subspaces of different dimensions.

Theorem 5.3. *Let $k \le l \le n$. Let $\mathbf{A} + b \in \operatorname{Graff}(k, n)$, $\mathbf{B} + c \in \operatorname{Graff}(l, n)$. Then*

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

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

$$\delta^\alpha(\mathbf{A}+b,\mathbf{B}+c) = \theta_{k+1}, \qquad \delta^\beta(\mathbf{A}+b,\mathbf{B}+c) = \Bigl(1-\prod_{i=1}^{k+1}\cos^2\theta_i\Bigr)^{1/2},$$

$$\delta^\pi(\mathbf{A}+b,\mathbf{B}+c) = \sin\theta_{k+1}, \qquad \delta^\mu(\mathbf{A}+b,\mathbf{B}+c) = \Bigl(\log\prod_{i=1}^{k+1}\frac{1}{\cos^2\theta_i}\Bigr)^{1/2},$$

$$\delta^\sigma(\mathbf{A}+b,\mathbf{B}+c) = 2\sin(\theta_{k+1}/2), \qquad \delta^\phi(\mathbf{A}+b,\mathbf{B}+c) = \cos^{-1}\Bigl(\prod_{i=1}^{k+1}\cos\theta_i\Bigr),$$

$$\delta^\kappa(\mathbf{A}+b,\mathbf{B}+c) = \Bigl(\sum_{i=1}^{k+1}\sin^2\theta_i\Bigr)^{1/2}, \qquad \delta^\rho(\mathbf{A}+b,\mathbf{B}+c) = \Bigl(2\sum_{i=1}^{k+1}\sin^2(\theta_i/2)\Bigr)^{1/2},$$

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

For the two affine Schubert varieties in (5.4), $\Omega_+(\mathbf{A}+b)$ may be viewed a Grassmannian whereas $\Omega_-(\mathbf{B}+c)$ may be viewed as an affine Grassmannian.

Proposition 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 $\Omega_-(\mathbf{B}+c) \cong \operatorname{Graff}_k(\mathbf{B}) \cong \operatorname{Graff}(k,l)$. $\square$

The reason for the asymmetry in Proposition 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 (5.7) and δ^* in Theorem 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

$$\operatorname{Graff}(\infty,\infty) := \bigsqcup_{k=1}^{\infty} \operatorname{Graff}(k,\infty).$$

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 $j_\infty\bigl(\operatorname{Graff}(\infty,\infty)\bigr)$ 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

$$d^*_{\operatorname{Graff}(\infty,\infty)}(\mathbf{A}+b,\mathbf{B}+c) := d^*_{\operatorname{Gr}(\infty,\infty)}\bigl(j_\infty(\mathbf{A}+b), j_\infty(\mathbf{B}+c)\bigr),$$

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

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

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 (4.1) also induces a volume density $d\gamma_{k,n}$ on $\mathrm{Gr}(k,n)$ [41, Proposition 9.1.12] with

$$\mathrm{Vol}\bigl(\mathrm{Gr}(k,n)\bigr) = \int_{\mathrm{Gr}(k,n)} \lvert d\gamma_{k,n}\rvert = \binom{n}{k}\frac{\prod_{j=1}^{n}\omega_j}{\bigl(\prod_{j=1}^{k}\omega_j\bigr)\bigl(\prod_{j=1}^{n-k}\omega_j\bigr)}, \tag{6.1}$$

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

By Theorem 2.5(ii), $\mathrm{Graff}(k,n)$ is a Zariski open dense subset in $\mathrm{Gr}(k+1,n+1)$ and we must have $\mu_{k+1,n+1}\bigl(\mathrm{Graff}(k,n)\bigr) = 1$. 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 6.1. *Let $k \le l \le n$ be such that $k+l \ge n$. 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*

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

Proof. By Theorem 2.5(ii), we have

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

By Proposition 5.4, we have

$$\mathrm{Vol}\bigl(\Omega_+(\mathbf{A}+\mathbf{b})\bigr) = \mathrm{Vol}\bigl(\mathrm{Gr}(n-l,n-k)\bigr) = \binom{n-k}{n-l}\frac{\prod_{j=1}^{n-k}\omega_j}{\bigl(\prod_{j=1}^{n-l}\omega_j\bigr)\bigl(\prod_{j=1}^{l-k}\omega_j\bigr)},$$

$$\mathrm{Vol}\bigl(\Omega_-(\mathbf{B}+\mathbf{c})\bigr) = \mathrm{Vol}\bigl(\mathrm{Graff}(k,l)\bigr) = \binom{l+1}{k+1}\frac{\prod_{j=1}^{l+1}\omega_j}{\bigl(\prod_{j=1}^{k+1}\omega_j\bigr)\bigl(\prod_{j=1}^{l-k}\omega_j\bigr)}.$$

We divide $\operatorname{Vol}\bigl(\Omega_+(\mathbf{A}+\mathbf{b})\bigr)$ and $\operatorname{Vol}\bigl(\Omega_-(\mathbf{B}+\mathbf{c})\bigr)$ by $\operatorname{Vol}\bigl(\operatorname{Graff}(l,n)\bigr)$ and $\operatorname{Vol}\bigl(\operatorname{Graff}(k,n)\bigr)$ respectively and notice that $k+l \ge n$ to complete the proof. $\square$

In the following we will use projection coordinates on $\operatorname{Graff}(n,k)$ as defined in Definition 3.4. By embedding $\operatorname{Graff}(k,n)$ as a subset $X = j\bigl(\operatorname{Graff}(k,n)\bigr) \subset \operatorname{Gr}(k+1,n+1)$ as in Theorem 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 $\operatorname{Gr}(k+1,n+1)$ (and that $\mu(X^c)=0$). Hence $\operatorname{Graff}(k,n)$ inherits any continuous probability distribution on $\operatorname{Gr}(k+1,n+1)$, in particular the Langevin distribution [12].

Definition 6.2. The *Langevin distribution*, also known as the *von Mises–Fisher distribution*, on $\operatorname{Graff}(k,n)$ is given by the probability density function

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

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

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

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 $\operatorname{Gr}(k,n)$ and $\operatorname{Graff}(k,n)$ with its projection affine coordinates, i.e., imposing equality in (2.6) and (2.7),

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

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 $P \in \operatorname{Gr}(n,k)$. 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 $E_{n-k} \coloneqq \operatorname{span}\{e_1,\dots,e_{n-k}\} \subset \mathbb{R}^{n+1}$. 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

$$f_G(b \mid P,\sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^{n-k}}}\exp\left(-\frac{\|b\|^2}{2\sigma^2}\right) \tag{6.2}$$

for all $b \in \ker(P)$. Note that $Q^\mathsf{T} b = \left[\begin{smallmatrix} b' \\ 0 \end{smallmatrix}\right]$ 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 (6.2).

Definition 6.3. The probability density function of the *Langevin–Gaussian distribution* on $\operatorname{Graff}(k, n)$ is $f_{LG}([P, b] \mid S, \sigma^2) := f_L(P \mid S) f_G(b \mid P, \sigma^2)$, i.e.,

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

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

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 and Theorem 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 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 7.1 and 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 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*

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

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

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

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 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* $[[P,\Omega],P] = \Theta^{\mathsf{T}}\begin{bmatrix} 0 & Z \\ -Z^{\mathsf{T}} & 0\end{bmatrix}\Theta$ *and* $P = \Theta^{\mathsf{T}}\begin{bmatrix} I_{k+1} & 0 \\ 0 & 0\end{bmatrix}\Theta$*. 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:* $\nabla^2 f : \mathsf{T}_{\mathbf{A}+b}\big(\mathrm{Graff}(k,n)\big) \times \mathsf{T}_{\mathbf{A}+b}\big(\mathrm{Graff}(k,n)\big) \to \mathbb{R}$,

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

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

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

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 7.1 and 7.2) and projection coordinates (Algorithms 7.4 and 7.5) but conjugate gradient method is only given in Stiefel coordinates (Algorithm 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 3.2 and 3.5 respectively. We will then borrow the corresponding methods on the usual Grassmannian developed in [2, 15] in conjunction with Propositions 7.1 and 7.2.

There is one caveat: Algorithms 7.1–7.5 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 4.4, this will almost never[6] happen but even if it does, the algorithms will still work fine as algorithms on $\mathrm{Gr}(k+1,n+1)$.

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

Algorithm 7.1 Steepest descent in Stiefel coordinates

Initialize $\mathbf{A}_0 + b_0 \in \operatorname{Graff}(k, n)$ in Stiefel coordinates $Y_0 := Y_{\mathbf{A}_0+b_0} \in \mathbb{R}^{(n+1)\times(k+1)}$.
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

Algorithm 7.2 Newton's method in Stiefel coordinates

Initialize $\mathbf{A}_0 + b_0 \in \operatorname{Graff}(k, n)$ in Stiefel coordinates $Y_0 := Y_{\mathbf{A}_0+b_0} \in \mathbb{R}^{(n+1)\times(k+1)}$.
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
  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 7.3 Conjugate gradient in Stiefel coordinates

Initialize $\mathbf{A}_0 + b_0 \in \operatorname{Graff}(k, n)$ in Stiefel coordinates $Y_0 := Y_{\mathbf{A}_0+b_0} \in \mathbb{R}^{(n+1)\times(k+1)}$.
Set $G_0 = f_Y(Y_0) - Y_0 Y_0^{\mathsf{T}} f_Y(Y_0)$ and $H_0 = -G_0$.
for $i = 0, 1, \dots$ **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
    $\tau G_i = G_i - \bigl(Y_i V \sin(t_{\min}\Sigma) + U(I - \cos(t_{\min}\Sigma))\bigr)U^{\mathsf{T}} G_i$; ▷ parallel transport of G_i
    $\gamma_i = \operatorname{tr}((G_{i+1} - \tau G_i)^{\mathsf{T}} G_{i+1}) / \operatorname{tr}(G_i^{\mathsf{T}} G_i)$;
    $H_{i+1} = -G_{i+1} + \gamma_i \tau H_i$;
  end procedure
  reset $H_{i+1} = -G_{i+1}$ if $i + 1 \equiv 0 \mod (k+1)(n-k)$;
end for

Algorithm 7.4 Steepest descent in projection coordinates

Initialize $\mathbf{A}_0 + b_0 \in \operatorname{Graff}(k, n)$ in projection coordinates $P_0 := P_{\mathbf{A}_0+b_0} \in \mathbb{R}^{(n+1)\times(n+1)}$.
for $i = 0, 1, \dots$ **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 $P_{i+1} = \frac{1}{2} I_{n+1} + \Theta^T \begin{bmatrix} \frac{1}{2}\cos(2\sqrt{ZZ^{\mathsf{T}}}) & -\operatorname{sinc}(2\sqrt{ZZ^{\mathsf{T}}})Z \\ -Z^{\mathsf{T}}\operatorname{sinc}(2\sqrt{ZZ^{\mathsf{T}}}) & -\frac{1}{2}\sin(2\sqrt{Z^{\mathsf{T}}Z}) \end{bmatrix} \Theta$;
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 3.3 or 3.6 to check if Y_{i+1}

Algorithm 7.5 Newton's method in projection coordinates

Initialize $\mathbf{A}_0 + b_0 \in \operatorname{Graff}(k,n)$ in projection coordinates $P_0 := P_{\mathbf{A}_0+b_0} \in \mathbb{R}^{(n+1)\times(n+1)}$.
for $i = 0, 1, \dots$ **do**
  find $\Omega_i \in \Lambda^2(\mathbb{R}^{n+1})$ 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 $\Theta_i(I - [P_i,[P_i,t\Omega_i]])\Theta_i^\mathsf{T} = Q_iR_i$; ▷ 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

or P_{i+1} are in $\operatorname{Graff}(k,n)$. In the unlikely scenario when they do fall outside $\operatorname{Graff}(k,n)$, e.g., if we have $Y_{i+1} = \begin{bmatrix} A & b \\ 0 & \gamma\end{bmatrix}$ where $A^\mathsf{T} b \neq 0$ or $P_{i+1} = \begin{bmatrix} S & d \\ d^\mathsf{T} & \gamma\end{bmatrix}$ 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.

8. Numerical Experiments

We will present various numerical experiments on two problems to illustrate the conjugate gradient and steepest descent algorithms in Section 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.

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

$$\begin{aligned} &\text{minimize} \quad \operatorname{tr}(X^\mathsf{T} AX) + \frac{1}{1+\|y\|^2}(y^\mathsf{T} Ay + 2b^\mathsf{T} y + c), \\ &\text{subject to} \quad X^\mathsf{T} X = I,\ X^\mathsf{T} y = 0, \end{aligned} \tag{8.1}$$

over all $X \in \mathbb{R}^{n\times k}$ and $y \in \mathbb{R}^n$. If we set $y = 0$ in (8.1), the resulting quadratic trace minimization problem with orthonormal constraints is essentially a *symmetric eigenvalue problem*; if we set $X = 0$ in (8.1), the resulting nonconvex optimization problem is a *quadratic fractional programming* problem.

By rearranging terms, we see that (8.1) is equivalent to a minimization problem over an affine Grassmannian,

$$\min_{\mathbf{X}+y\in\operatorname{Graff}(k,n)} \operatorname{tr}\left(\begin{bmatrix} X & y/\sqrt{1+\|y\|^2} \\ 0 & 1/\sqrt{1+\|y\|^2}\end{bmatrix}^\mathsf{T} \begin{bmatrix} A & b \\ b^\mathsf{T} & c\end{bmatrix} \begin{bmatrix} X & y/\sqrt{1+\|y\|^2} \\ 0 & 1/\sqrt{1+\|y\|^2}\end{bmatrix}\right), \tag{8.2}$$

which shows that the problem (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 $\begin{bmatrix} A & b \\ b^\mathsf{T} & c\end{bmatrix}$ — the optimum value is the sum of the $k+1$ smallest eigenvalues.

Figure 2 shows convergence trajectories of steepest descent and conjugate gradient in Stiefel coordinates, i.e., Algorithms 7.1 and 7.3, on $\operatorname{Graff}(3,6)$ for the problem (8.2). $\operatorname{Graff}(3,6)$ 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 $\operatorname{Graff}(3,6)$. The gradient of $f(Y) := \operatorname{tr}\bigl(Y^\mathsf{T}\begin{bmatrix} A & b \\ b^\mathsf{T} & c\end{bmatrix}Y\bigr)$ is given by $\nabla f(Y) = \begin{bmatrix} A & b \\ b^\mathsf{T} & c\end{bmatrix}Y$. 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

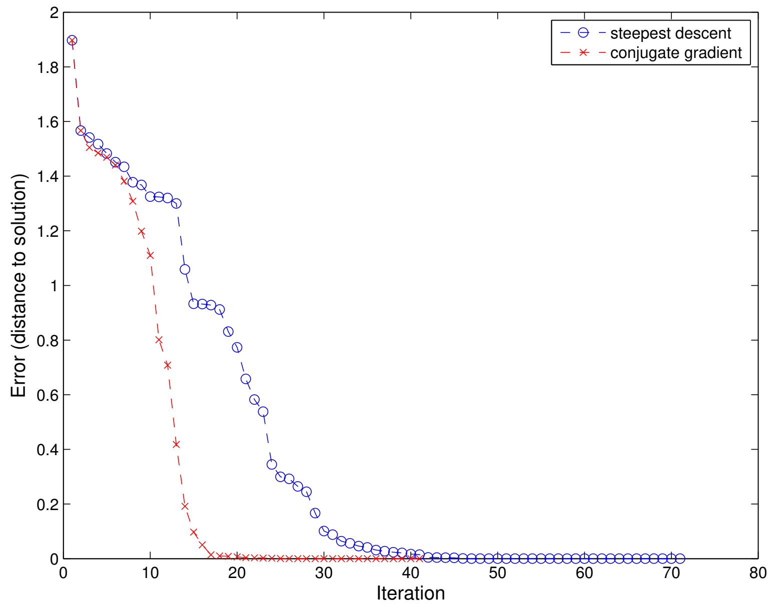

FIGURE 2. Convergence trajectories of steepest descent and conjugate gradient for a quadratic optimization problem on the affine Grassmannian $\mathrm{Graff}(3,6)$.

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 (8.1) for various values of k and n to generate tables of timing and accuracy. Table 3 and 4 show the robustness of the algorithm with respect to different choices of k and n.

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

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

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

Table 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 $\dim\bigl(\mathrm{Graff}(k, 100)\bigr) = (k+1)(100-k)$ first increases and then decreases. On the other hand, if we fix the dimension of ambient space, Table 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.

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 (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:

$$\min_{\mathbf{X}+y\in\mathrm{Graff}(k,n)} \sum\nolimits_{i=1}^{m} d^2(\mathbf{A}_i + b_i, \mathbf{X} + y), \tag{8.3}$$

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

$$f_m(\mathbf{X} + y) = \sum\nolimits_{i=1}^{m} d^2(\mathbf{A}_i + b_i, \mathbf{X} + y) \tag{8.4}$$

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 (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 (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, shows that steepest descent outperforms conjugate gradient in this specific example. The departure from the example we considered in Section 8.1 can be explained by the fact that the function in (8.4) is geodesically convex on the affine Grassmannian, whereas the function in (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 (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 7. Accuracy (distance to true solution) of steepest descent and conjugate gradient for Fréchet/Karcher mean on $\mathrm{Graff}(k, 10)$.

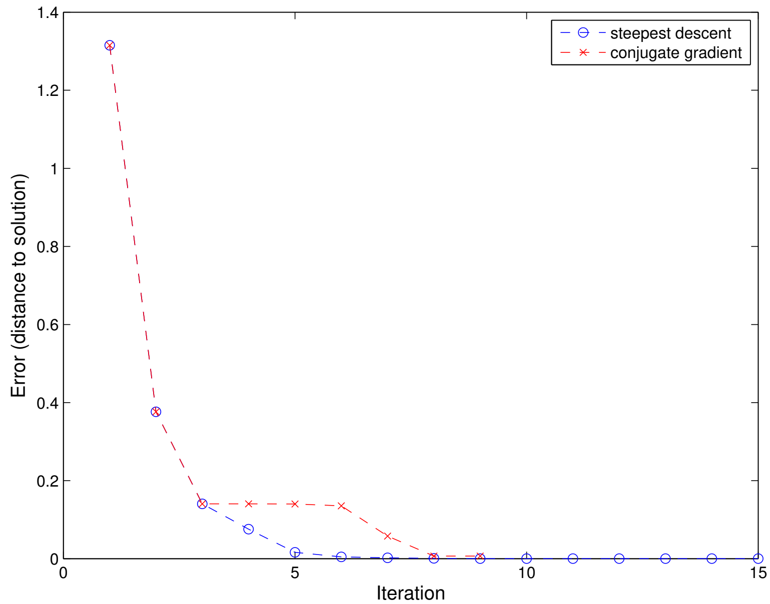

FIGURE 3. Convergence trajectories of steepest descent and conjugate gradient for Fréchet/Karcher mean on the affine Grassmannian $\operatorname{Graff}(7, 19)$

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

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

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

More extensive numerical experiments indicate that steepest descent and conjugate gradient are about equally fast for minimizing (8.4), cf. Tables 9 and 10, but that steepest decent is more accurate by orders of magnitude compared to conjugate gradient, cf. Tables 7 and 8. Again we believe the reason is that steepest descent is better suited for geodesically convex problems compared to conjugate gradient.

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 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.

Acknowledgment

The authors thank Pierre-Antoine Absil and Tingran Gao for very helpful discussions. The work in this article is generously supported by AFOSR FA9550-13-1-0133, DARPA D15AP00109, NSF IIS 1546413, DMS 1209136, and DMS 1057064. In addition, LHL's work is supported by a DARPA Director's Fellowship and KY's work is supported by the Hundred Talents Program of the Chinese Academy of Sciences.

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.*, **37**, no. 3, pp. 1176–1197.

[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.

Computational and Applied Mathematics Initiative, Department of Statistics, University of Chicago, Chicago, IL 60637-1514.

E-mail address: `lekheng@galton.uchicago.edu`

Department of Statistics, University of Chicago, Chicago, IL 60637-1514.

E-mail address: `kenwong@uchicago.edu`

KLMM, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

E-mail address: `keyk@amss.ac.cn`