NUMERICAL ALGORITHMS ON THE AFFINE GRASSMANNIAN[*]

LEK-HENG LIM[†], KEN SZE-WAI WONG[‡], AND KE YE[§]

Abstract. The affine Grassmannian is a noncompact smooth manifold that parameterizes all affine subspaces of a fixed dimension. It is a natural generalization of Euclidean space, points being zero-dimensional affine subspaces. We will realize the affine Grassmannian as a matrix manifold and extend Riemannian optimization algorithms including steepest descent, Newton method, and conjugate gradient, to real-valued functions on the affine Grassmannian. Like their counterparts for the Grassmannian, they rely only on standard numerical linear algebra and are readily computable.

Key words. affine Grassmannian, affine subspaces, manifold optimization

AMS subject classifications. 14M15, 90C30

DOI. 10.1137/18M1169321

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

The affine Grassmannian is a relatively obscure object compared to its ubiquitous cousin, the Grassmannian. Nevertheless, it is $\mathrm{Graff}(k, n)$, which like $\mathbb{R}^n$ is a noncompact manifold, that is the natural generalization of Euclidean space—points are zero-dimensional affine subspaces and so $\mathrm{Graff}(0, n) = \mathbb{R}^n$. The noncompactness makes $\mathrm{Graff}(k, n)$ harder to study than $\mathrm{Gr}(k, n)$, which is compact. The two main objectives of our article are to (i) develop a concrete foundation for optimization algorithms on the affine Grassmannian in the spirit of [2, 8]; (ii) explicitly describe three such algorithms: steepest decent, conjugate gradient, and Newton method.

As in [1, 2, 8] we do not view our manifold in an abstract fashion comprising charts glued together; instead we emphasize the use of extrinsic coordinates in the form of matrices for efficient computations. The affine Grassmannian then becomes a concrete computational platform (like $\mathbb{R}^n$) on which geodesics, exponential maps, parallel transports, Riemannian gradient, and Hessian, etc., may all be efficiently computed using standard numerical linear algebra.

In fact, a main reason for the widespread applicability of the Grassmannian is the existence of several excellent choices of extrinsic matrix coordinates, allowing subspaces to be represented as matrices and thereby the use of a vast range of algorithms

[†]Computational and Applied Mathematics Initiative, Department of Statistics, University of Chicago, Chicago, IL 60637-1514 (lekheng@galton.uchicago.edu).

[‡]Department of Statistics, University of Chicago, Chicago, IL 60637-1514 (kenwong@uchicago.edu).

[§]KLMM, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China (keyk@amss.ac.cn).

in numerical linear algebra [1, 2, 3, 8]. Such concrete realizations of an abstract manifold are essential for application purposes. By providing a corresponding set of tools for the affine Grassmannian, we have effectively extended the wide range of data analytic techniques that use the Grassmannian as a model for linear subspaces [9, 13, 12, 16, 22, 24, 28, 29, 30] to affine subspaces.

Before this work, the affine Grassmannian, as used in the sense[1] of this article, i.e., the manifold that parameterizes k-dimensional affine subspaces in $\mathbb{R}^n$, has received scant attention in both pure and applied mathematics. To the best of our knowledge, this article and its companion [23] are the first to study it systematically. We have kept the technical materials in this article to a bare minimum, providing just enough to discuss the numerical optimization algorithms for affine Grassmannians. Readers are referred to the aforementioned companion article [23] for the more technical properties of the affine Grassmannian. We summarize our contributions:

- In section 2, we show that the affine Grassmannian is a Riemannian manifold that can be embedded as an open submanifold of the Grassmannian. We introduce some basic systems of extrinsic coordinates: affine coordinates, orthogonal affine coordinates, and projective affine coordinates. These simple coordinate systems are convenient in proofs but are inadequate when it comes to actual computations.
- In section 3, we introduce two more sophisticated systems of coordinates that will be critical to our optimization algorithms—Stiefel coordinates and projection coordinates—representing points on the affine Grassmannian as $(n+1) \times (k+1)$ matrices with orthonormal columns and as $(n+1) \times (n+1)$ orthogonal projection matrices, respectively. We establish a result that allows us to switch between these two systems of coordinates.
- In section 4, we describe the common differential geometric objects essential in our optimization algorithms—tangent spaces, exponential maps, geodesics, parallel transport, gradients, Hessians—concretely in terms of Stiefel coordinates and projection coordinates. In particular, we will see that once expressed as matrices in either coordinate system, these objects become readily computable via standard numerical linear algebra.
- In section 5, we describe (in pseudocodes) steepest descent, Newton method, and conjugate gradient in Stiefel coordinates and the first two in projection coordinates.
- In section 6, we report the results of our numerical experiments on two test problems: (a) a nonlinear nonconvex optimization problem that arises from a coupling of a symmetric eigenvalue problem with a quadratic fractional programming problem, and (b) the problem of computing Fréchet and Karcher means of two affine subspaces. These problems are judiciously chosen—they are nontrivial and yet their exact solutions may be determined in closed form, which in turn allows us to ascertain whether our algorithms indeed converge to their actual global optimizers. In the extensive tests we carried out on both problems, the iterates generated by our algorithms converge to the true solutions in every instance.

---

[1]We would like to caution the reader that the term "affine Grassmannian" is now used far more commonly to refer to another very different object; see [4, 10, 20]. In this article, it will be used exclusively in the sense of Definition 2.1. If desired, "Grassmannian of affine subspaces" may be used to avoid ambiguity.

The utility of these algorithms will become evident when one attempts to solve an optimization problem, say, the test problem in (a):

$$\text{minimize} \quad \text{tr}(X^\mathsf{T} A X) + \frac{1}{1 + \|y\|^2}(y^\mathsf{T} A y + 2 b^\mathsf{T} y + c)$$

$$\text{subject to} \quad X^\mathsf{T} X = I, \ X^\mathsf{T} y = 0$$

for $X \in \mathbb{R}^{n \times k}$ and $y \in \mathbb{R}^n$ as a general nonlinear optimization problem without exploiting the fact that it can be transformed into an optimization problem on the affine Grassmannian. As we reported in section 6, all standard nonlinear optimization methods that we tried failed without finding a *feasible point*, never mind the global minimizer; whereas both the steepest descent and conjugate gradient methods on the affine Grassmannian converge to the global minimizer rapidly over a wide range of values of n and k.

The key to our approach involves an embedding of $\text{Graff}(k, n)$ into $\text{Gr}(k+1, n+1)$. In general, the cost of manifold optimization algorithms depends on the dimension of the ambient space—if we embed $\text{Graff}(k, n)$ into a manifold M and perform the computations in M, the computational cost will be dependent on $\dim(M)$. So one cannot do better than an ambient space M that has the same dimension as $\text{Graff}(k, n)$, which is what our method does: as we will see later,

$$\dim \text{Graff}(k, n) = (n - k)(k + 1) = \dim \text{Gr}(k + 1, n + 1),$$

and thus there is no increase to the intrinsic dimension of $\text{Graff}(k, n)$. This we think is the most compelling reason for embedding $\text{Graff}(k, n)$ into $\text{Gr}(k + 1, n + 1)$.

2. Affine Grassmannian. The affine Grassmannian was first described in [19] but has received relatively little attention compared to the Grassmannian of linear subspaces $\text{Gr}(k, n)$. Aside from a brief discussion in [27, section 9.1.3], we are unaware of any systematic treatment. 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. To distinguish it from a different but identically named object, we may also refer to it as the *Grassmannian of affine subspaces*.

We will establish some basic properties of the affine Grassmannian with a view towards optimization algorithms in the style of [2, 8]. These results are neither difficult nor surprising, certainly routine to the experts, but have not appeared before elsewhere to the best of our knowledge.

We remind the reader of some basic terminologies. We write $\text{span}(X)$ for the linear span of a set X and $\text{im}(A)$ for the image of a matrix A regarded as a linear operator; so if $A = [a_1, \ldots, a_k] \in \mathbb{R}^{n \times k}$, where $a_1, \ldots, a_k \in \mathbb{R}^n$ are column vectors, then $\text{im}(A) = \text{span}\{a_1, \ldots, a_k\}$. 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 $a_1, \ldots, a_k$ are the basis vectors. A *flag* is a strictly increasing sequence of nested linear subspaces, $\mathbb{X}_0 \subset \mathbb{X}_1 \subset \mathbb{X}_2 \subset \cdots$. A flag is said to be *complete* if $\dim \mathbb{X}_k = k$, *finite* if $k = 0, 1, \ldots, n$, and *infinite* if $k \in \mathbb{N} \cup \{0\}$. We write $\text{Gr}(k, n)$ for the *Grassmannian* of k-planes in $\mathbb{R}^n$, $\text{V}(k, n)$ for the *Stiefel manifold* of orthonormal k-frames, and $\text{O}(n) := \text{V}(n, n)$ for the *orthogonal group*. We may regard $\text{V}(k, n)$ as a homogeneous space,

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

or more concretely as the set of $n \times k$ matrices with orthonormal columns; here $\mathrm{O}(n-k)$ is identified with a subgroup of $\mathrm{O}(n)$ by regarding $A \in \mathrm{O}(n-k)$ as $\begin{bmatrix} I & 0 \\ 0 & A \end{bmatrix} \in \mathrm{O}(n)$. There is a right action of the orthogonal group $\mathrm{O}(k)$ on $\mathrm{V}(k,n)$: For $Q \in \mathrm{O}(k)$ and $A \in \mathrm{V}(k,n)$, the action yields $AQ \in \mathrm{V}(k,n)$ and the resulting homogeneous space is $\mathrm{Gr}(k,n)$, i.e.,

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

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

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 $\mathrm{Graff}(k,n)$, is the set of all k-dimensional affine subspaces of $\mathbb{R}^n$.*

This set-theoretic definition reveals little about the rich geometry behind $\mathrm{Graff}(k,n)$, which we will see is a smooth Riemannian manifold intimately related to the Grassmannian $\mathrm{Gr}(k+1, n+1)$.

Throughout this article, a blackboard boldfaced letter $\mathbb{A}$ 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 $\mathbb{A}$. We denote a k-dimensional affine subspace as $\mathbb{A} + b \in \mathrm{Graff}(k,n)$, where $\mathbb{A} \in \mathrm{Gr}(k,n)$ is a k-dimensional linear subspace and $b \in \mathbb{R}^n$ is the displacement of $\mathbb{A}$ from the origin. If $A = [a_1, \ldots, a_k] \in \mathbb{R}^{n \times k}$ is a basis of $\mathbb{A}$, then

$$(2.3) \qquad \mathbb{A} + b := \{\lambda_1 a_1 + \cdots + \lambda_k a_k + b \in \mathbb{R}^n : \lambda_1, \ldots, \lambda_k \in \mathbb{R}\}.$$

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

Since a k-dimensional affine subspace of $\mathbb{R}^n$ may be described by a k-dimensional subspace of $\mathbb{R}^n$ and a displacement vector in $\mathbb{R}^n$, it might be tempting to guess that $\mathrm{Graff}(k,n)$ is identical to $\mathrm{Gr}(k,n) \times \mathbb{R}^n$. However, as we have seen, the representation of an affine subspace as $\mathbb{A} + b$ is not unique and we emphasize that

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

Although $\mathrm{Graff}(k,n)$ can be regarded as a quotient of $\mathrm{Gr}(k,n) \times \mathbb{R}^n$, this description is not used in our article. Instead we point interested readers to [23] for details on this and other quotient structures of $\mathrm{Graff}(k,n)$.

We may choose an orthonormal basis for $\mathbb{A}$ so that $A \in \mathrm{V}(k,n)$ and choose b to be orthogonal to $\mathbb{A}$ so that $A^\mathsf{T} b = 0$. Hence we may always represent $\mathbb{A} + b \in \mathrm{Graff}(k,n)$ 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 $\mathbb{A} + b$ must have $A' = AQ$ for some $Q \in \mathrm{O}(k)$ and $b_0' = b_0$.

We will not insist on using orthogonal affine coordinates at all times as they can be unnecessarily restrictive, especially in proofs. Without these orthogonality conditions, a matrix $[A, b_0] \in \mathbb{R}^{n \times (k+1)}$ that represents an affine subspace $\mathbb{A} + b$ in the sense of (2.3) is called its *affine coordinates*.

Our main goal is to show that the vast array of optimization techniques [1, 2, 3, 8, 14] may be adapted to the affine Grassmannian. In this regard, it is the following
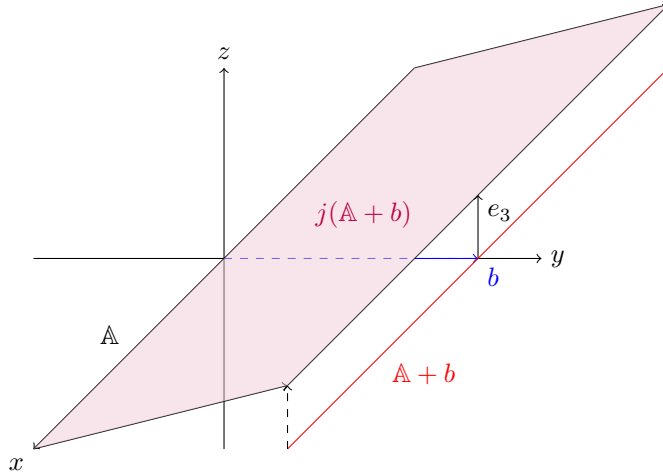
FIG. 1. *The affine subspace $\mathbb{A} + b$ is given by the x-axis $\mathbb{A}$ displaced by b along the y-axis. The embedding $j : \mathrm{Graff}(k, n) \to \mathrm{Gr}(k + 1, n + 1)$ takes $\mathbb{A} + b$ to the smallest 2-plane containing $\mathbb{A}$ and $b + e_3$, where e_3 is the standard unit vector along the z-axis.*

view of $\mathrm{Graff}(k, n)$ as an embedded open submanifold of $\mathrm{Gr}(k + 1, n + 1)$ that will prove most useful. Our construction of this embedding is illustrated in Figure 1 and formally stated in Theorem 2.2.

A Grassmannian comes equipped with a Radon measure [25, section 3.9] and all measure theoretic statements on $\mathrm{Gr}(k, n)$ will be with respect to this. We also remind the reader of two basic terminologies: Let M and N be smooth manifolds. Then N is called an *open submanifold* of M if there is a smooth injective map $f : N \to M$ such that $f(N)$ is an open subset of M. A smooth map $f : N \to M$ is called an *embedding* if f is injective and the differential map $df_x : \mathsf{T}_x(N) \to \mathsf{T}_{f(x)}(M)$ is also injective at any $x \in N$.

THEOREM 2.2. *Let $n \geq 2$ and $1 \leq k \leq n$. The affine Grassmannian $\mathrm{Graff}(k, n)$ is an open submanifold of $\mathrm{Gr}(k+1, n+1)$ whose complement has codimension at least two and measure zero. For concreteness, we will use the map*

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

where $e_{n+1} = (0, \ldots, 0, 1)^{\mathsf{T}} \in \mathbb{R}^{n+1}$ as our default embedding map.

Proof. We will first prove that j as defined in (2.4) is an injective map and its image is an open subset of $\mathrm{Graff}(k, n)$. Let $\mathbb{A} + b \in \mathrm{Graff}(k, n)$. First we observe that whenever

$$\mathrm{span}(\mathbb{A} \cup \{b + e_{n+1}\}) = \mathrm{span}(\mathbb{A}' \cup \{b' + e_{n+1}\}),$$

we have that $b' + e_{n+1} \in \mathrm{span}(\mathbb{A} \cup \{b + e_{n+1}\})$. This implies that $b' - b \in \mathbb{A}$ since $\mathbb{A}$ is a subspace of $\mathbb{R}^n$. So $\mathbb{A} = \mathbb{A}'$ and therefore $\mathbb{A} + b = \mathbb{A}' + b'$, i.e., the map j is injective.

Next, for the complement of $j\big(\mathrm{Graff}(k, n)\big)$ in $\mathrm{Gr}(k+1, n+1)$, note that a $(k+1)$-dimensional linear subspace $\mathbb{B}$ of $\mathbb{R}^{n+1}$ is an image of some $\mathbb{A} + b \in \mathrm{Graff}(k, n)$ under the map j iff the $(n + 1)$th coordinate of some vector in $\mathbb{B}$ is nonzero. So the complement consists of all $\mathbb{B}$ contained in the subspace $\mathbb{R}^n$ of $\mathbb{R}^{n+1}$, i.e., it is diffeomorphic to the Grassmannian $\mathrm{Gr}(k + 1, n)$, which has dimension $(k + 1)(n - k - 1) = (k + 1)(n - k) - (k + 1)$, thus codimension $k + 1 \geq 2$ and, thus, of measure

zero. Being diffeomorphic to $\mathrm{Gr}(k+1,n)$, the complement of $j\big(\mathrm{Graff}(k,n)\big)$ is a closed subset of $\mathrm{Gr}(k+1,n+1)$ and thus $j\big(\mathrm{Graff}(k,n)\big)$ is an open subset. □

In the proof we have identified $\mathbb{R}^n$ with the subset

$$\{(x_1,\ldots,x_n,0)^\mathsf{T} \in \mathbb{R}^{n+1} : x_1,\ldots,x_n \in \mathbb{R}\}$$

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 (2.4) is the most natural one. Henceforth we will often identify $\mathrm{Graff}(k,n)$ with its embedded image $j\big(\mathrm{Graff}(k,n)\big)$. Whenever we speak of $\mathrm{Graff}(k,n)$ as if it is a subset of $\mathrm{Gr}(k+1,n+1)$, we are implicitly assuming this identification. In this regard, we may view $\mathrm{Gr}(k+1,n+1)$ as a compactification of the noncompact manifold $\mathrm{Graff}(k,n)$. Note that any embedding of $\mathbb{R}^n$ into $\mathbb{R}^{n+1}$ (mapping $\mathbb{R}^n$ to an n-dimensional subspace of $\mathbb{R}^{n+1}$) will give a different embedding of $\mathrm{Graff}(k,n)$ into $\mathrm{Gr}(k+1,n+1)$. While we would argue that there is a natural simplicity with our preferred choice of embedding, we are open to the possibility that depending on the application under consideration, other embeddings could be useful as well. Fortunately, the subsequent discussions in this article can be readily adapted to a different choice of embedding via a linear change of coordinates.

As an open subset of $\mathrm{Gr}(k+1,n+1)$, $\mathrm{Graff}(k,n)$ inherits a smooth structure from that on $\mathrm{Gr}(k+1,n+1)$ but one may also define a smooth structure on $\mathrm{Graff}(k,n)$ intrinsically, i.e., in a way that does not depend on the map j. Nevertheless these two smooth structures are identical. We refer interested readers to [23] for the details.

From Theorem 2.2, we derive a few other observations that will be important for our optimization algorithms. As we mentioned at the end of section 1, from the perspective of optimization, the most important feature of the embedding j is that it does not increase dimension; since the computational costs of optimization algorithms invariably depend on the dimension of the ambient space, it is ideal in this regard.

COROLLARY 2.3. *Let $n \geq 2$ and $1 \leq k \leq n$. $\mathrm{Graff}(k,n)$ is a path-connected Riemannian manifold with the canonical metric induced from that of $\mathrm{Gr}(k+1,n+1)$. In addition,*

(i) *the dimension of the ambient manifold $\mathrm{Gr}(k+1,n+1)$ is exactly the same as $\mathrm{Graff}(k,n)$, i.e.,*

$$\dim \mathrm{Graff}(k,n) = (n-k)(k+1) = \dim \mathrm{Gr}(k+1,n+1);$$

(ii) *the geodesic distance between two points $\mathbb{A}+b$ and $\mathbb{B}+c$ in $\mathrm{Graff}(k,n)$ is equal to that between $j(\mathbb{A}+b)$ and $j(\mathbb{B}+c)$ in $\mathrm{Gr}(k+1,n+1)$;*

(iii) *if $f : \mathrm{Graff}(k,n) \to \mathbb{R}$ is a continuous function that can be extended to*

$$\widetilde{f} : \mathrm{Gr}(k+1,n+1) \to \mathbb{R},$$

then the minimizer and maximizer of $\widetilde{f}$ are almost surely attained in $\mathrm{Graff}(k,n)$ with respect to the uniform probability measure on $\mathrm{Gr}(k+1,n+1)$.

Proof. Again, we will identify $\mathrm{Graff}(k,n)$ with its embedded image $j(\mathrm{Graff}(k,n))$ so that an expression like $\mathbb{A}+b \in \mathrm{Gr}(k+1,n+1)$ is understood to mean $j(\mathbb{A}+b) \in \mathrm{Gr}(k+1,n+1)$. It is a basic fact in differential geometry [21, Chapter 8] that every open subset of a Riemannian manifold is also a Riemannian manifold with the induced metric. Explicit expressions for the Riemannian metric on $\mathrm{Graff}(k,n)$ will be discussed later in Propositions 4.1(ii) and 4.4(ii). (i) follows from Theorem 2.2, i.e., $\mathrm{Graff}(k,n)$ is an open submanifold of $\mathrm{Gr}(k+1,n+1)$. Since the codimension of the complement of $\mathrm{Graff}(k,n)$ in $\mathrm{Gr}(k+1,n+1)$ is at least two, (ii) follows from the transversality theorem in differential topology [15], as we will show next. Let

$X := \mathrm{Gr}(k+1, n+1) \backslash \mathrm{Graff}(k, n)$, i.e., the complement of $\mathrm{Graff}(k, n)$ in $\mathrm{Gr}(k+1, n+1)$. As we noted in the proof of Theorem 2.2, X is diffeomorphic to $\mathrm{Gr}(k+1, n)$, which is of dimension $(k+1)(n-k-1)$, and thus

$$\dim(X) + 1 < \dim \mathrm{Gr}(k+1, n+1).$$

This implies that every curve γ on $\mathrm{Gr}(k+1, n+1)$ can be perturbed to some γ' so that $\gamma' \cap X = \varnothing$. Since geodesic distance between two points x_0 and x_1 on a manifold M is defined by

$$d_M(x_0, x_1) := \inf\{L(\gamma) : \gamma \text{ curve connecting } x_0 \text{ and } x_1\}$$

with $L(\gamma)$ the length of γ, it follows that

$$d_{\mathrm{Graff}(k,n)}(\mathbb{A} + b, \mathbb{B} + c) \geq L(\gamma_0),$$

where γ_0 is the shortest geodesic curve in $\mathrm{Gr}(k+1, n+1)$ connecting $\mathbb{A} + b$ and $\mathbb{B} + c$. Now, for any $\varepsilon > 0$, we may perturb γ_0 to γ_ε so that $\gamma_\varepsilon \cap X = \varnothing$, i.e., γ_ε lies entirely within $\mathrm{Graff}(k, n)$, and $L(\gamma_\varepsilon) = L(\gamma_0) + \varepsilon$. So we have

$$d_{\mathrm{Graff}(k,n)}(\mathbb{A} + b, \mathbb{B} + c) \leq L(\gamma_0) + \varepsilon,$$

and since $\varepsilon > 0$ can be made arbitrarily small,

$$d_{\mathrm{Graff}(k,n)}(\mathbb{A} + b, \mathbb{B} + c) = L(\gamma_0) = d_{\mathrm{Gr}(k+1,n+1)}(\mathbb{A} + b, \mathbb{B} + c),$$

where the second equality is by our choice of γ_0. Note that the existence of γ_ε, lying entirely within $\mathrm{Graff}(k, n)$ and connecting any pair of points $\mathbb{A} + b, \mathbb{B} + c \in \mathrm{Graff}(k, n)$, also shows that $\mathrm{Graff}(k, n)$ is path connected. For (iii), first note that the Radon measure on $\mathrm{Gr}(k+1, n+1)$ is finite on the whole space since $\mathrm{Gr}(k+1, n+1)$ is compact and thus we may normalize it to give a uniform probability measure. It also follows from the compactness of $\mathrm{Gr}(k+1, n+1)$ that $\widetilde{f}$ always attains its extrema on $\mathrm{Gr}(k+1, n+1)$. That the minimizer and maximizer lie in $\mathrm{Graff}(k, n)$ with probability one is then a consequence of the fact that its complement has null measure.   □

Note that (ii) and (iii) rely on Theorem 2.2 and do not hold in general for other embedded manifolds. For example, if B is the solid unit ball in $\mathbb{R}^3$ and M is the complement of B in $\mathbb{R}^3$, then (ii) and (iii) obviously fail to hold for M. Clearly, the uniform probability measure in (iii) may be replaced by any probability measure that is absolutely continuous with respect to it, e.g., those corresponding to the von Mises–Fisher or Langevin–Gaussian distributions in [23].

A word about why we have excluded $k = 0$ from Theorem 2.2 and Corollary 2.3: In this case, $\mathrm{Graff}(0, n) = \mathbb{R}^n$ and $\mathrm{Gr}(1, n+1) = \mathbb{RP}^n$, the n-dimensional real projective space; the map j in (2.4) is the classical embedding of $\mathbb{R}^n$ into $\mathbb{RP}^n$. If we regard $\mathbb{RP}^n$ as a disjoint union

$$\mathbb{RP}^n = \mathbb{R}^n \sqcup \mathbb{R}^{n-1} \sqcup \cdots \sqcup \mathbb{R}^0,$$

then the copy of $\mathbb{R}^n$ on the right is precisely $j(\mathbb{R}^n)$. Thus the complement of $j(\mathbb{R}^n)$ has codimension one in $\mathbb{RP}^n$ and Corollary 2.3(ii) and (iii) do not hold for $k = 0$ (although (i) does).

The Riemannian metric on $j(\mathbb{R}^n)$ induced from that on $\mathbb{RP}^n$ is also not the usual Euclidean metric on $\mathbb{R}^n$. Indeed, if we regard $\mathbb{RP}^n$ as $\mathbb{S}^n$ with antipodal points identified, then $j : \mathbb{R}^n \to \mathbb{RP}^n$ is the stereographic projection and $j(\mathbb{R}^n)$ is the upper

hemisphere[2] of $\mathbb{S}^n$ where each point in $j(\mathbb{R}^n)$ is identified with its antipodal point on the lower hemisphere. The Riemannian metric on $j(\mathbb{R}^n)$ is simply the one on $\mathbb{S}^n$ restricted to the upper hemisphere.

It is sometimes desirable to represent elements of $\mathrm{Gr}(k, n)$ as actual matrices instead of equivalence classes of matrices. The Grassmannian has a well-known representation [27, Example 1.2.20] as rank-k orthogonal projection[3] matrices or, equivalently, trace-k idempotent symmetric matrices:

$$(2.5) \qquad \mathrm{Gr}(k, n) \cong \{P \in \mathbb{R}^{n \times n} : P^\mathsf{T} = P^2 = P, \ \mathrm{tr}(P) = k\}.$$

Note that $\mathrm{rank}(P) = \mathrm{tr}(P)$ for an orthogonal projection matrix P. A straightforward affine analogue of (2.5) for $\mathrm{Graff}(k, n)$ is simply

$$(2.6) \qquad \mathrm{Graff}(k, n) \cong \{[P, b] \in \mathbb{R}^{n \times (n+1)} : P^\mathsf{T} = P^2 = P, \ \mathrm{tr}(P) = k, \ Pb = 0\},$$

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

There are three particularly useful systems of matrix coordinates on the Grassmannian: A point on $\mathrm{Gr}(k, n)$ may 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 $S \in \mathrm{GL}(k)$, the group of invertible $k \times k$ matrices; (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 $\mathbb{A}$ by (i) bases of $\mathbb{A}$, (ii) orthonormal bases of $\mathbb{A}$, (iii) an orthogonal projection onto $\mathbb{A}$. The affine coordinates, orthogonal affine coordinates, and projection affine coordinates introduced in this section are analogues of (i), (ii), and (iii), respectively.

These relatively simplistic extrinsic coordinates are inadequate in computations. As we will see in sections 4 and 5, explicit representations of tangent space vectors and geodesics, effective computations of exponential maps, parallel transports, gradients, and Hessians, require more sophisticated systems of extrinsic matrix coordinates. In section 3 we will introduce two of these.

Nevertheless, the simpler coordinate systems in this section do serve a valuable role—they come in handy in proofs, where the more complicated systems of coordinates in section 3 can be unnecessarily cumbersome. The bottom line is that different coordinates are good for different purposes[5] and having several choices makes the affine Grassmannian a versatile platform in applications.

3. Matrix coordinates for the affine Grassmannian. One reason for the wide applicability of the Grassmannian is the existence of several excellent choices of extrinsic coordinates in terms of matrices, 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, 8]. Here we will introduce two systems of extrinsic coordinates,

---

[2]Here hemispheres are open hemispheres, i.e., the equator is excluded.

[3]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 $P = I$.

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

[5]This is also the case for Grassmannians: Orthonormal or projection matrix coordinates may be invaluable for computations as in [8, 14] but they obscure mathematical properties evident in, say, Plücker coordinates [26, Chapter 14].

representing a point on $\mathrm{Graff}(n,k)$ as an $(n+1) \times (k+1)$ orthonormal matrix and as an $(n+1) \times (n+1)$ projection matrix, respectively.

For an affine subspace $\mathbb{A} + b \in \mathrm{Graff}(k,n)$, 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 $\mathrm{V}(k+1, n)$. 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* $\mathbb{A} + b \in \mathrm{Graff}(k,n)$ *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* $\mathbb{A} + b$ *is the* $(n+1) \times (k+1)$ *matrix with orthonormal columns,*

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

Two orthogonal affine coordinates $[A, b_0], [A', b_0']$ of $\mathbb{A} + b$ give two corresponding matrices of Stiefel coordinates $Y_{\mathbb{A}+b}$, $Y'_{\mathbb{A}+b}$. 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

$$(3.1) \quad Y_{\mathbb{A}+b} = \begin{bmatrix} A & b_0/\sqrt{1 + \|b_0\|^2} \\ 0 & 1/\sqrt{1 + \|b_0\|^2} \end{bmatrix} = \begin{bmatrix} A' & b_0'/\sqrt{1 + \|b_0'\|^2} \\ 0 & 1/\sqrt{1 + \|b_0'\|^2} \end{bmatrix} \begin{bmatrix} Q' & 0 \\ 0 & 1 \end{bmatrix} = Y'_{\mathbb{A}+b}Q,$$

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

PROPOSITION 3.2. *Consider the equivalence class of matrices given by*

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

The affine Grassmannian may be represented as a set of equivalence classes of $(n+1) \times (k+1)$ *matrices with orthonormal columns,*

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

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

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

Proof. The set of equivalence classes on the right-hand side (RHS) of (3.2) is the set $j\big(\mathrm{Graff}(k,n)\big)$, j as defined in (2.4), if $\mathrm{Gr}(k+1, n+1)$ is regarded as the homogeneous space in (3.3). ☐

The following lemma is easy to see from the definition of Stiefel coordinates and our discussion above. It will be useful for the optimization algorithms in section 5, allowing us to check *feasibility*, i.e., whether a point represented as an $(n+1) \times (k+1)$ matrix is in the feasible set $j\big(\mathrm{Graff}(k,n)\big)$.

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 $\mathbb{A} + b \in \mathrm{Graff}(k, n)$.

(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* $\mathbb{A} + b$, *then every other matrix of Stiefel coordinates for* $\mathbb{A} + b$ *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* $\mathbb{A} + b$.

The matrix of projection affine coordinates $[P, b] \in \mathbb{R}^{n \times (n+1)}$ in (2.6) is not an orthogonal projection matrix. With this in mind, we introduce the following notion.

Definition 3.4. *Let* $\mathbb{A} + b \in \mathrm{Graff}(k, n)$ *and* $[P, b] \in \mathbb{R}^{n \times (n+1)}$ *be its projection affine coordinates. The matrix of* projection coordinates *for* $\mathbb{A} + b$ *is the orthogonal projection matrix*

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

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

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

It is straightforward to verify that $P_{\mathbb{A}+b}$ is indeed an orthogonal projection matrix, i.e., $P_{\mathbb{A}+b}^2 = P_{\mathbb{A}+b} = P_{\mathbb{A}+b}^{\mathsf{T}}$. 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 $j\big(\mathrm{Graff}(k, n)\big)$, j as defined in (2.4), but in terms of projection coordinates. With this description, $\mathrm{Graff}(k, n)$ may be regarded as a subvariety of $\mathbb{R}^{(n+1) \times (n+1)}$.

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

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

An affine subspace $\mathbb{A} + b \in \mathrm{Graff}(k, n)$ *is uniquely represented by its projection coordinates* $P_{\mathbb{A}+b}$.

Proof. Let $\mathbb{A} + b \in \mathrm{Graff}(k, n)$ have orthogonal affine coordinates $[A, b_0]$. Since $P = AA^{\mathsf{T}} \in \mathbb{R}^{n \times n}$ is an orthogonal projection matrix that satisfies $Pb_0 = 0$, the map $\mathbb{A} + b \mapsto P_{\mathbb{A}+b}$ takes $\mathrm{Graff}(k, n)$ onto the set of matrices on the RHS of (3.4) with inverse given by $P_{\mathbb{A}+b} \mapsto \mathrm{im}(P) + b_0$. □

The next lemma allows feasibility checking in 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* $\mathbb{A}+b \in$ 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* $\mathbb{A} + b$ *in orthogonal affine coordinates.*

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

LEMMA 3.7.
  (i) *If* $Y_{\mathbb{A}+b} \in \mathrm{V}(k+1, n+1)$ *is a matrix of Stiefel coordinates for* $\mathbb{A}+b$, *then the matrix of projection coordinates for* $\mathbb{A}+b$ *is given by*
$$P_{\mathbb{A}+b} = Y_{\mathbb{A}+b}Y_{\mathbb{A}+b}^\mathsf{T} \in \mathbb{R}^{(n+1)\times(n+1)}.$$

  (ii) *If* $P_{\mathbb{A}+b} \in \mathbb{R}^{(n+1)\times(n+1)}$ *is the matrix of projection coordinates for* $\mathbb{A}+b$, *then there exists a matrix of Stiefel coordinates* $Y_{\mathbb{A}+b}$ *for* $\mathbb{A}+b$ *such that* $P_{\mathbb{A}+b} = Y_{\mathbb{A}+b}Y_{\mathbb{A}+b}^\mathsf{T}$.

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 $P_{\mathbb{A}+b} = V\begin{bmatrix} I_{k+1} & \\ & 0_{n-k} \end{bmatrix}V^\mathsf{T} = V_{k+1}V_{k+1}^\mathsf{T}$, where the columns of $V_{k+1} \in \mathrm{V}(k+1, n+1)$ are the eigenvectors corresponding to the eigenvalue 1, which form an orthonormal basis of $j(\mathbb{A}+b) = \mathrm{span}(\mathbb{A}\cup\{b+e_{n+1}\})$. 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 [11] such that $Q^\mathsf{T}v = \|v\|e_{k+1}$. Then $Y_{\mathbb{A}+b} = V_{k+1}Q$ has the form required in Lemma 3.3(i) for a matrix of Stiefel coordinates, as
$$Y_{\mathbb{A}+b}^\mathsf{T}Y_{\mathbb{A}+b} = Q^\mathsf{T}V_{k+1}^\mathsf{T}V_{k+1}Q = I. \qquad \square$$

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. Tangent space, exponential map, geodesic, parallel transport, gradient, and Hessian on the affine Grassmannian. The embedding of Graff(k,n) as an open smooth submanifold of Gr$(k+1, n+1)$ by Theorem 2.2 and Corollary 2.3 allows us to borrow the Riemannian optimization framework on Grassmannians in [1, 2, 3, 8] to develop optimization algorithms on the affine Grassmannian. We will present various geometric notions and algorithms on Graff(k,n) in terms of both Stiefel and projection coordinates. The higher dimensions required by projection coordinates generally make them less desirable than Stiefel coordinates.

Propositions 4.1, Theorem 4.2, and Proposition 4.4 are, respectively, summaries of [8] and [14] adapted for the affine Grassmannian. We will only give a sketch of the proof, referring readers to the original sources for more details.

PROPOSITION 4.1. *The following are basic differential geometric notions on* Graff(k,n) *expressed in Stiefel coordinates.*

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

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

(ii) *Riemannian metric: The Riemannian metric g on $\operatorname{Graff}(k, n)$ is given by*

$$g_{\mathbb{A}+b}(\Delta_1, \Delta_2) = \operatorname{tr}(\Delta_1^{\mathsf{T}}\Delta_2)$$

for $\Delta_1, \Delta_2 \in \mathsf{T}_{\mathbb{A}+b}\bigl(\operatorname{Graff}(k, n)\bigr)$, i.e., $Y_{\mathbb{A}+b}^{\mathsf{T}}\Delta_i = 0$, $i = 1, 2$.

(iii) *Exponential map: The geodesic with $Y(0) = Y_{\mathbb{A}+b}$ and $\dot{Y}(0) = H$ in $\operatorname{Graff}(k, n)$ is given by*

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

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

(iv) *Parallel transport: The parallel transport of $\Delta \in \mathsf{T}_{\mathbb{A}+b}\bigl(\operatorname{Graff}(k, n)\bigr)$ along the geodesic given by H has expression*

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

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

(v) *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 $Q \in \operatorname{O}(k + 1)$. The gradient of f at $Y = Y_{\mathbb{A}+b}$ is*

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

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

(vi) *Hessian: Let f be as in (v). The Hessian of f at $Y = Y_{\mathbb{A}+b}$ is,*
   (a) *as a bilinear form, $\nabla^2 f : \mathsf{T}_{\mathbb{A}+b}\bigl(\operatorname{Graff}(k, n)\bigr) \times \mathsf{T}_{\mathbb{A}+b}\bigl(\operatorname{Graff}(k, n)\bigr) \to \mathbb{R}$,*

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

where $f_{YY} \in \mathbb{R}^{(n+1)(k+1)\times(n+1)(k+1)}$ with $(f_{YY})_{ij,hl} = \frac{\partial^2 f}{\partial y_{ij}\partial y_{hl}}$ and

$$f_{YY}(\Delta, \Delta') = \sum_{i,j,h,l=1}^{n+1,k+1,n+1,k+1} (f_{YY})_{ij,hl}\delta_{ij}\delta'_{hl};$$

   (b) *as a linear map, $\nabla^2 f : \mathsf{T}_{\mathbb{A}+b}\bigl(\operatorname{Graff}(k, n)\bigr) \to \mathsf{T}_{\mathbb{A}+b}\bigl(\operatorname{Graff}(k, n)\bigr)$,*

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

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

Sketch of proof. These essentially follow from the corresponding formulas for the Grassmannian in [8, 14]. For instance, the Riemannian metric g is induced by the canonical Riemannian metric on $\operatorname{Gr}(k + 1, n + 1)$ [8, section 2.5], the geodesic $X(t)$ on $\operatorname{Gr}(k, n)$ starting at $X(0) = X_{\mathbb{A}}$ in the direction $\dot{X}(0) = H$ is given in [8, Equation (2.65)] as

$$X(t) = \begin{bmatrix} X_{\mathbb{A}}V & U \end{bmatrix} \begin{bmatrix} \cos(t\Sigma) \\ \sin(t\Sigma) \end{bmatrix} V^{\mathsf{T}},$$

where $X_{\mathbb{A}}$ is the matrix representation of $X(0)$ and $H = U\Sigma V^{\mathsf{T}}$ is a condensed SVD of H. The displayed formula in (iii) for a geodesic in $\mathrm{Graff}(k, n)$ is then obtained by taking the inverse image of the corresponding geodesic in $\mathrm{Gr}(k + 1, n + 1)$ under the embedding j. Other formulas may be similarly obtained by the same procedure from their counterparts on the Grassmannian.   □

As the complement of $\mathrm{Graff}(k, n)$ in $\mathrm{Gr}(k + 1, n + 1)$ has codimension at least two, in Proposition 4.1(iii) the situation $Y(t) \notin \mathrm{Graff}(k, n)$ occurs with probability zero. For an analogue, one may think of a geodesic in $\mathbb{R}^3$ with the x-axis removed. This guarantees that Algorithms 5.1–5.5 will almost never lead to a point outside $\mathrm{Graff}(k, n)$; and even if that happens, it does not matter since these algorithms are formulated as *infeasible methods* (see discussion after Algorithm 5.1).

Since the distance-minimizing geodesic connecting two points on $\mathrm{Gr}(k+1, n+1)$ is not necessarily unique,[6] it is possible that there is more than one geodesic on $\mathrm{Graff}(k, n)$ connecting two given points. However, distance-minimizing geodesics can all be parameterized as in Proposition 4.1 even if they are not unique. In fact, we may explicitly compute the geodesic distance between any two points on $\mathrm{Graff}(k, n)$. Before stating the result, we first recall from Corollary 2.3(ii) that for any two affine k-flats $\mathbb{A} + b$ and $\mathbb{B} + c \in \mathrm{Graff}(k, n)$,

$$d_{\mathrm{Graff}(k,n)}(\mathbb{A} + b, \mathbb{B} + c) := d_{\mathrm{Gr}(k+1,n+1)}\big(j(\mathbb{A} + b), j(\mathbb{B} + c)\big),$$

where j is the embedding in (2.4), defines a notion of distance consistent with the geodesic distance on a Grassmannian.

THEOREM 4.2. *Let $\mathbb{A} + b$ and $\mathbb{B} + c \in \mathrm{Graff}(k, n)$. If*

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

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

$$(4.1) \qquad\qquad d_{\mathrm{Graff}(k,n)}(\mathbb{A} + b, \mathbb{B} + c) = \left(\sum_{i=1}^{k+1} \theta_i^2 \right)^{1/2},$$

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

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 $Y_{\mathbb{A}+b}$ and $Y'_{\mathbb{A}+b}$ be two different matrices of Stiefel coordinates for $\mathbb{A} + b$ and $Y_{\mathbb{B}+c}$ and $Y'_{\mathbb{B}+c}$ be two different matrices of Stiefel coordinates for $\mathbb{B} + c$. By Lemma 3.3(ii), there exist $Q_1, Q_2 \in \mathrm{O}(k + 1)$ such that $Y_{\mathbb{A}+b} = Y'_{\mathbb{A}+b}Q_1$, $Y_{\mathbb{B}+c} = Y'_{\mathbb{B}+c}Q_2$. The required result then follows from

$$\sigma_i(Y_{\mathbb{A}+b}^{\mathsf{T}}Y_{\mathbb{B}+c}) = \sigma_i(Q_1^{\mathsf{T}}Y'^{\mathsf{T}}_{\mathbb{A}+b}Y'_{\mathbb{B}+c}Q_2) = \sigma_i(Y'^{\mathsf{T}}_{\mathbb{A}+b}Y'_{\mathbb{B}+c}), \qquad i = 1, \ldots, k. \qquad □$$

The proof above also shows that $\theta_1, \ldots, \theta_{k+1}$ 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 $\theta_i(\mathbb{A} + b, \mathbb{B} + c)$. Consider the SVD,

$$(4.2) \qquad\qquad Y_{\mathbb{A}+b}^{\mathsf{T}}Y_{\mathbb{B}+c} = U\Sigma V^{\mathsf{T}},$$

---

[6]For example, there are two distance-minimizing geodesics on $\mathrm{Gr}(1, 2) \simeq \mathbb{S}^1$ for any pair of antipodal points.

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

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

We will call the pair of column vectors (p_i, q_i) the ith *affine principal vectors* between $\mathbb{A} + b$ and $\mathbb{B} + c$. These are the affine analogues of principal angles and vectors of linear subspaces [6, 11, 32].

This expression for a geodesic in Proposition 4.1(iii) assumes that we are given an initial point and an initial direction; the following gives an alternative expression for a geodesic in $\mathrm{Graff}(k,n)$ that connects two given points.

COROLLARY 4.3. *Let $\mathbb{A}+b$ and $\mathbb{B}+c \in \mathrm{Graff}(k,n)$. Let $\gamma : [0,1] \to \mathrm{Gr}(k+1, n+1)$ be the curve*

$$(4.3) \qquad \gamma(t) = \mathrm{span}(Y_{\mathbb{A}+b}U\cos(t\Theta)U^{\mathsf{T}} + Q\sin(t\Theta)U^{\mathsf{T}}),$$

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

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

The orthogonal matrix U is the same as that in (4.2) and $\Theta = \mathrm{diag}(\theta_1, \ldots, \theta_{k+1})$ is the diagonal matrix of affine principal angles. Then γ has the following properties:
  (i) *γ is a distance-minimizing curve connecting $j(\mathbb{A}+b)$ and $j(\mathbb{B}+c)$, i.e., attains (4.1);*
 (ii) *the derivative of γ at $t = 0$ is given by*

$$(4.4) \qquad\qquad\qquad\qquad \gamma'(0) = Q\Theta U^{\mathsf{T}};$$

(iii) *there is at most one value of $t \in (0,1)$ such that $\gamma(t) \notin j\big(\mathrm{Graff}(k,n)\big)$.*

Sketch of proof. The expression in [8, Theorem 2.3] for a distance-minimizing geodesic connecting two points in $\mathrm{Gr}(k+1, n+1)$ gives (4.3). By Theorem 2.2, $\mathrm{Graff}(k,n)$ is embedded in $\mathrm{Gr}(k+1, n+1)$ as an open submanifold whose complement has measure zero. Since the complement of $\mathrm{Graff}(k,n)$ in $\mathrm{Gr}(k+1, n+1)$ comprises points with coordinates $\begin{bmatrix} A \\ 0 \end{bmatrix} \in \mathbb{R}^{(n+1)\times(k+1)}$, where $A \in \mathbb{R}^{n\times(k+1)}$ and $A^{\mathsf{T}}A = I$, a simple calculation shows γ has at most one point not contained in $\mathrm{Graff}(k,n)$. ☐

As a reminder, the situation $\gamma(t) \notin \mathrm{Graff}(k,n)$ occurs with probability zero as the complement of $\mathrm{Graff}(k,n)$ in $\mathrm{Gr}(k+1, n+1)$ has codimension at least two. We conclude this section with the analogue of Proposition 4.1 in projection coordinates.

PROPOSITION 4.4. *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 $\mathbb{A} + b \in \mathrm{Graff}(k,n)$ has representation*

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

 (ii) *Riemannian metric: The Riemannian metric g on $\mathrm{Graff}(k,n)$ is given by*

$$g_{\mathbb{A}+b}(\Delta_1, \Delta_2) = \mathrm{tr}(\Delta_1^{\mathsf{T}}\Delta_2),$$

where $\Delta_1, \Delta_2 \in \mathsf{T}_{\mathbb{A}+b}\big(\mathrm{Graff}(k,n)\big)$, i.e., $\Delta_i = [P_{\mathbb{A}+b}, \Omega_i]$ for some $\Omega_i \in \Lambda^2(\mathbb{R}^{n+1})$, $i = 1, 2$.

(iii) *Exponential map: Let $P = P_{\mathbb{A}+b}$ 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*

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

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

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

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

(v) *Hessian: Let f and f_P be as in* (iv). *The Hessian of f at $P = P_{\mathbb{A}+b}$ is,*
   (a) *as a bilinear form, $\nabla^2 f : \mathsf{T}_{\mathbb{A}+b}\big(\operatorname{Graff}(k, n)\big) \times \mathsf{T}_{\mathbb{A}+b}\big(\operatorname{Graff}(k, n)\big) \to \mathbb{R}$,*

$$\nabla^2 f(\Delta, \Delta')$$
$$= \operatorname{tr}\left(\left(\left[P, \left[P, \sum_{i,j,h,l=1}^{n+1} (f_{PP})_{ij,hl}\delta_{ij}E_{hl}\right]\right] - \frac{1}{2}[P, [\nabla f, \Delta]] - \frac{1}{2}[\nabla f, [P, \Delta]]\right)\Delta'\right),$$

where $f_{PP} \in \mathbb{R}^{(n+1)^2\times(n+1)^2}$ 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;
   (b) *as a linear map, $\nabla^2 f : \mathsf{T}_{\mathbb{A}+b}\big(\operatorname{Graff}(k, n)\big) \to \mathsf{T}_{\mathbb{A}+b}\big(\operatorname{Graff}(k, n)\big)$,*

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

Sketch of proof. Again, these formulas follow from their counterparts on the Grassmannian in [14] by applying j^{-1}, as we did in the proof of Proposition 4.1. $\qquad\square$

A notable omission in Proposition 4.4 is a formula for parallel transport. While parallel transport on $\operatorname{Graff}(k, n)$ in Stiefel coordinates takes a relatively simple form in Proposition 4.1, its explicit expression in projection coordinates is extremely complicated, and as a result unilluminating and error prone. We do not recommend computing parallel transport in projection coordinates—one should instead change projection coordinates to Stiefel coordinates by Lemma 3.7, compute parallel transport in Stiefel coordinates using Proposition 4.1(iv), and then transform the result back to projection coordinates by Lemma 3.7 again.

5. Steepest descent, conjugate gradient, and Newton method on the affine Grassmannian. We now describe the methods of steepest descent, conjugate gradient, and Newton on the affine Grassmannian. The steepest descent and Newton methods are given in both Stiefel coordinates (Algorithms 5.1 and 5.3) and projection coordinates (Algorithms 5.4 and 5.5) but the conjugate gradient method is only given in Stiefel coordinates (Algorithm 5.2) as we do not have a closed-form expression for parallel transport in projection coordinates.

We will rely on our embedding of $\operatorname{Graff}(k, n)$ into $\operatorname{Gr}(k + 1, n + 1)$ via Stiefel or projection coordinates as given by Propositions 3.2 and 3.5, respectively. We then borrow the corresponding method on the Grassmannian in [2, 8] in conjunction with Propositions 4.1 and 4.4.

There is one caveat: Algorithms 5.1–5.5 are formulated as infeasible methods. If we start from a point in $\operatorname{Graff}(k, n)$, regarded as a subset of $\operatorname{Gr}(k + 1, n + 1)$, the next

---

Algorithm 5.1 Steepest descent in Stiefel coordinates.

Initialize $\mathbb{A}_0 + b_0 \in \mathrm{Graff}(k, n)$ in Stiefel coordinates $Y_0 := Y_{\mathbb{A}_0+b_0} \in \mathbb{R}^{(n+1)\times(k+1)}$.
for $i = 0, 1, \ldots$ **do**
    set $G_i = f_Y(Y_i) - Y_i Y_i^\mathsf{T} f_Y(Y_i)$;   ▷ gradient of f at Y_i
    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 5.2 Conjugate gradient in Stiefel coordinates.

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

Algorithm 5.3 Newton's method in Stiefel coordinates.

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

---

iterate along the geodesic may become *infeasible*, i.e., fall outside $\mathrm{Graff}(k, n)$. By Theorem 2.2, this will occur with probability zero but even if it does, the algorithms will still work fine as algorithms on $\mathrm{Gr}(k + 1, n + 1)$.

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

Algorithm 5.4 Steepest descent in projection coordinates.

Initialize $\mathbb{A}_0 + b_0 \in \mathrm{Graff}(k, n)$ in projection coordinates $P_0 := P_{\mathbb{A}_0 + b_0} \in \mathbb{R}^{(n+1) \times (n+1)}$.

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

    set $\nabla f(P_i) = [P_i, [P_i, f_P(P_i)]]$;

    find $\Theta \in \mathbb{R}^{(n+1) \times (n+1)}$ and $t > 0$ so that $P_i = \Theta^\mathsf{T} \left[\begin{smallmatrix} I_{k+1} & 0 \\ 0 & 0 \end{smallmatrix} \right] \Theta$ and $-t \nabla f(P_i) = \left[\begin{smallmatrix} 0 & Z \\ -Z^\mathsf{T} & 0 \end{smallmatrix} \right]$;

    set $P_{i+1} = \frac{1}{2} I_{n+1} + \Theta^T \left[\begin{array}{cc} \frac{1}{2} \cos(2\sqrt{ZZ^\mathsf{T}}) & -\operatorname{sinc}(2\sqrt{ZZ^\mathsf{T}})Z \\ -Z^\mathsf{T} \operatorname{sinc}(2\sqrt{ZZ^\mathsf{T}}) & -\frac{1}{2} \sin(2\sqrt{Z^\mathsf{T}Z}) \end{array} \right] \Theta$;

end for

Algorithm 5.5 Newton's method in projection coordinates.

Initialize $\mathbb{A}_0 + b_0 \in \mathrm{Graff}(k, n)$ in projection coordinates $P_0 := P_{\mathbb{A}_0 + b_0} \in \mathbb{R}^{(n+1) \times (n+1)}$.

for $i = 0, 1, \ldots$ **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 $P_i = \Theta_i^\mathsf{T} \left[\begin{smallmatrix} I_{k+1} & 0 \\ 0 & 0 \end{smallmatrix} \right] \Theta_i$;   ▷ QR factorization

    compute $\Theta_i(I - [P_i, [P_i, t\Omega_i]])\Theta_i^\mathsf{T} = Q_i R_i$;   ▷ QR factorization with positive diagonal in R_i

    set $P_{i+1} = \Theta_i^\mathsf{T} Q_i \Theta_i P_i \Theta_i^\mathsf{T} Q_i^\mathsf{T} \Theta_i$;

end for

We recognize that the univariate optimization problem for finding the step size t may not be easily solvable for a given choice of objective function f. The "exact line search" step is intended to be a placeholder; it may be substituted for with any other reasonable methodology for choosing step size.

6. Numerical experiments. We will present various numerical experiments on two test problems to illustrate the conjugate gradient and steepest descent algorithms in section 5. These problems are deliberately chosen to be nontrivial and yet have closed-form solutions—so that we may check whether our algorithms have converged to the true solutions of these problems. The measured "error" in any of these problems will be taken to mean the distance between a true solution and the corresponding computed solution, with distance as defined in (4.1). We implemented Algorithms 5.1 and 5.2 in MATLAB and PYTHON and used a combination of (i) the Frobenius norm of the Riemannian gradient, (ii) distance between successive iterates, and (iii) number of iterations, for our stopping conditions.

6.1. 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

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

(6.1)

over all $X \in \mathbb{R}^{n \times k}$ and $y \in \mathbb{R}^n$. If we set $y = 0$ in (6.1), the resulting quadratic trace minimization problem with orthonormal constraints is essentially a *symmetric*
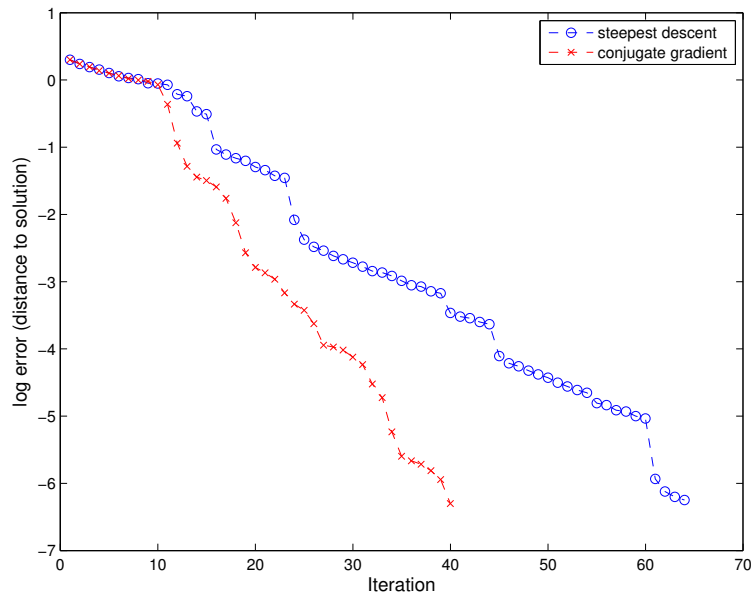
FIG. 2. *Convergence trajectories of steepest descent and conjugate gradient for a quadratic optimization problem on the affine Grassmannian* $\mathrm{Graff}(3,6)$.

eigenvalue problem; if we set $X = 0$ in (6.1), the resulting nonconvex optimization problem is called *quadratic fractional programming*.

By rearranging terms, (6.1) transforms into a minimization problem over an affine Grassmannian,

$$(6.2) \qquad \min_{\mathbb{X}+y\in\mathrm{Graff}(k,n)} \mathrm{tr}\left(\begin{bmatrix} X & y/\sqrt{1+\|y\|^2} \\ 0 & 1/\sqrt{1+\|y\|^2} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} A & b \\ b^{\mathsf{T}} & c \end{bmatrix} \begin{bmatrix} X & y/\sqrt{1+\|y\|^2} \\ 0 & 1/\sqrt{1+\|y\|^2} \end{bmatrix}\right),$$

which shows that the problem (6.1) is in fact coordinate independent, depending on X and y only through the affine subspace $\mathrm{im}(X) + y = \mathbb{X} + y$. 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 5.1 and 5.2, on $\mathrm{Graff}(3,6)$ for the problem (6.2). $\mathrm{Graff}(3,6)$ is a 12-dimensional manifold; we generate $A \in \mathbb{R}^{6\times6}$, $b \in \mathbb{R}^6$, $c \in \mathbb{R}$ randomly with $\mathcal{N}(0,1)$ entries, and likewise pick a random initial point in $\mathrm{Graff}(3,6)$. The gradient of $f(Y) \coloneqq \mathrm{tr}\left(Y^{\mathsf{T}}\begin{bmatrix} A & b \\ b^{\mathsf{T}} & c \end{bmatrix}Y\right)$ 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 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 (6.1) for various values of k and n to generate tables of timing and accuracy. Tables 1 and 3 are from the same set of numerical experiments; ditto for Tables 2 and 4.

TABLE 1
Accuracy (distance to true solution) of steepest descent and conjugate gradient for a quadratic optimization problem on Graff$(k, 100)$.

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

TABLE 2
Accuracy (distance to true solution) of steepest descent and conjugate gradient for a quadratic optimization problem on Graff$(6, n)$.

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 3
Elapsed time (in seconds) of steepest descent and conjugate gradient for a quadratic optimization problem on Graff$(k, 100)$.

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 4
Elapsed time (in seconds) of steepest descent and conjugate gradient for a quadratic optimization problem on Graff$(6, n)$.

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

Tables 1 and 2 show the robustness of the algorithm with respect to different choices of k and n. Table 3 shows a modest initial increase followed by a decrease in elapsed time to convergence as k increases—a reflection of the intrinsic dimension of the problem as $\dim\big(\mathrm{Graff}(k, 100)\big) = (k + 1)(100 - k)$ first increases and then decreases. On the other hand, if we fix the dimension of ambient space, Table 4 shows that the elapsed time increases with k. The results indicates that the elapsed time increases moderately with the dimension of the affine Grassmannian. The takeaway from Tables 1–4 is the lack of any conspicuous trend—speed and accuracy for both algorithms remain relatively flat over the range of values of n and k compared.

For comparison, we attempted to solve (6.1) as a general nonlinear constrained optimization problem using the MATLAB Optimization Toolbox. Even for values of n and k as small as $n = 4$ and $k = 2$, every available method in the toolbox—interior point, trust region, sequential quadratic programming, active set—failed without finding a feasible point. There is not a single case that converged even with minor adjustments of tolerance and maximum iterations to nondefault values.

6.2. Fréchet and Karcher means of affine subspaces. Let $d = d_{\mathrm{Graff}(k,n)}$ be the geodesic distance on Graff(k, n) as defined in (4.1). We would like to solve for the minimizer $\mathbb{X} + y \in \mathrm{Graff}(k, n)$ in the sum-of-square-distances minimization
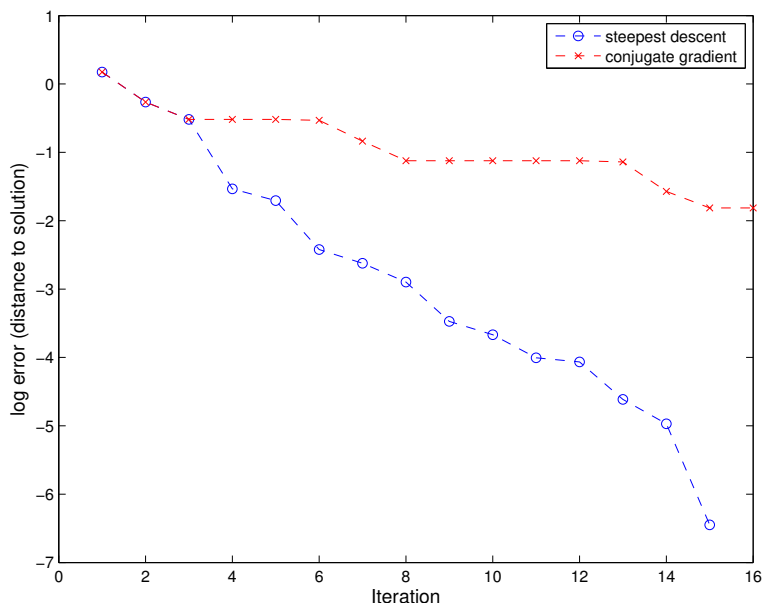
FIG. 3. *Convergence trajectories of steepest descent and conjugate gradient for Fréchet mean on the affine Grassmannian* $\operatorname{Graff}(7, 19)$.

problem:

$$(6.3) \qquad \min_{\mathbb{X}+y\in\operatorname{Graff}(k,n)} \sum_{i=1}^{m} d^2(\mathbb{A}_i + b_i, \mathbb{X} + y),$$

where $\mathbb{A}_i + b_i \in \operatorname{Graff}(k,n)$, $i = 1, \ldots, m$. The Riemannian gradient [17] of the objective function

$$(6.4) \qquad f_m(\mathbb{X} + y) = \sum_{i=1}^{m} d^2(\mathbb{A}_i + b_i, \mathbb{X} + y)$$

is given by

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

where $\log_{\mathbb{X}+y}(\mathbb{A} + b)$ denotes the derivative of the geodesic $\gamma(t)$ connecting $\mathbb{X} + y$ and $\mathbb{A} + b$ at $t = 0$ with an explicit expression given by (4.4).

The global minimizer of this problem is called the *Fréchet mean* and a local minimizer is called a *Karcher mean* [18]. They are not equal in general [5] even for the case $m = 2$, although in this case a Fréchet mean is the midpoint, i.e., $t = 1/2$, of the geodesic connecting $\mathbb{A}_1 + b_1$ and $\mathbb{A}_2 + b_2$ given by the closed-form expression (4.3).

We will take the $\operatorname{Graff}(7, 19)$, a 96-dimensional manifold, as our specific example. Our objective function is $f_2(\mathbb{X} + y) = d^2(\mathbb{A}_1 + b_1, \mathbb{X} + y) + d^2(\mathbb{A}_2 + b_2, \mathbb{X} + y)$ 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, unlike the example we considered in section 6.1,

TABLE 5

Accuracy (distance to true solution) of steepest descent and conjugate gradient for Fréchet mean on $\operatorname{Graff}(k, 10)$.

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 6

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

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 7

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

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 8

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

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

which shows the opposite. So each algorithm serves a purpose for a different type of problem. On the other hand, when we attempted to find the Karcher mean of $m > 2$ affine subspaces by extending f_m to the objective function in (6.3), we see faster convergence (as measured by actual elapsed time) in conjugate gradient instead.

More extensive numerical experiments indicate that steepest descent is more accurate than conjugate gradient by orders of magnitude in minimizing (6.4) (see Tables 5 and 6), but that both are about equally fast (see Tables 7 and 8). While these numerical experiments are intended for testing our algorithms, we would like to point out their potential application to *model averaging*, i.e., aggregating affine subspaces estimated from different datasets.

7. Conclusion. We introduce the affine Grassmannian $\operatorname{Graff}(k, n)$, study its basic differential geometric properties, and develop several concrete systems of coordinates—three simple ones that are handy in proofs and two more sophisticated ones intended for computations; the latter two we called Stiefel and projection coordinates, respectively. We show that when expressed in terms of Stiefel or projection coordinates, basic geometric objects on $\operatorname{Graff}(k, n)$ may be readily represented as matrices and manipulated with standard routines in numerical linear algebra. With these in place, we ported the three standard Riemannian optimization algorithms on the Grassmannian—steepest descent, conjugate gradient, and Newton method—to the affine Grassmannian. We demonstrated the efficacy of the first two algorithms

through extensive numerical experiments on two nontrivial problems with closed-form solutions, which allows us to ascertain the correctness of our results. The encouraging outcomes in these experiments provide a positive outlook towards further potential applications of our framework. Our hope is that numerical algorithms on the affine Grassmannian could become a mainstay in statistics and machine learning, where estimation problems may often be formulated as optimization problems on $\mathrm{Graff}(k, n)$.

We end with a few words on future directions. In order to ascertain the correctness of our algorithms, we have restricted our numerical experiments to problems for which true solutions can be determined analytically. Our next goal is to use the algorithms developed here to solve realistic problems for which solutions are not known in advance. A more ambitious goal would be to develop the theory and algorithms for *convex optimization* on the affine Grassmannian: Note that every continuous geodesically convex function on $\mathrm{Gr}(k, n)$ is constant as it is compact [31], putting a damper on "convex optimization on the Grassmannian." On the other hand, one could potentially define a nontrivial class of continuous geodesically convex functions on $\mathrm{Graff}(k, n)$, a noncompact manifold, and thus "convex optimization on the affine Grassmannian" may be a possibility—in fact we already have a well-known example, namely, convex optimization on $\mathrm{Graff}(0, n) = \mathbb{R}^n$.

REFERENCES

[1] P.-A. ABSIL, R. MAHONY, AND R. SEPULCHRE (2008), *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ.

[2] P.-A. ABSIL, R. MAHONY, AND R. SEPULCHRE (2004), *Riemannian geometry of Grassmann manifolds with a view on algorithmic computation*, Acta Appl. Math., 80, pp. 199–220.

[3] P.-A. ABSIL, R. MAHONY, R. SEPULCHRE, AND P. VAN DOOREN (2002), *A Grassmann–Rayleigh quotient iteration for computing invariant subspaces*, SIAM Rev., 44, pp. 57–73.

[4] P. N. ACHAR AND L. RIDER (2015), *Parity sheaves on the affine Grassmannian and the Mirković–Vilonen conjecture*, Acta Math., 215, pp. 183–216.

[5] B. AFSARI, R. TRON, AND R. VIDAL (2013), *On the convergence of gradient descent for finding the Riemannian center of mass*, SIAM J. Control Optim., 51, pp. 2230–2260.

[6] Å. BJÖRCK AND G. H. GOLUB (1973), *Numerical methods for computing angles between linear subspaces*, Math. Comp., 27, pp. 579–594.

[7] N. BOUMAL, B. MISHRA, P.-A. ABSIL, AND R. SEPULCHRE (2014), *MANOPT, a MATLAB toolbox for optimization on manifolds*, J. Mach. Learn. Res., 15, pp. 1455–1459, http://www.manopt.org.

[8] A. EDELMAN, T. A. ARIAS, AND S. T. SMITH (1998), *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20, pp. 303–353.

[9] E. ELHAMIFAR AND R. VIDAL (2013), *Sparse subspace clustering: Algorithm, theory, and applications*, IEEE Trans. Pattern Anal. Mach. Intell., 35, pp. 2765–2781.

[10] E. FRENKEL AND D. GAITSGORY (2009), *Localization of $\mathfrak{g}$-modules on the affine Grassmannian*, Ann. of Math. (2), 170, pp. 1339–1381.

[11] G. GOLUB AND C. VAN LOAN (2013), *Matrix Computations*, 4th ed., John Hopkins University Press, Baltimore, MD.

[12] J. HAMM AND D. D. LEE (2008), *Grassmann discriminant analysis: A unifying view on subspace-based learning*, Proceedings of the International Conference on Machine Learning (ICML), ACM, New York, pp. 376–383.

[13] G. HARO, G. RANDALL, AND G. SAPIRO (2007), *Stratification learning: Detecting mixed density and dimensionality in high dimensional point clouds*, Advances in Neural Information Processing Systems (NIPS), MIT Press, Cambridge, MA, pp. 553–560.

[14] U. Helmke, K. Hüper, and J. Trumpf (2007), *Newton's Method on Graßmann Manifolds*, preprint, https://arxiv.org/abs/0709.2205.

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

[16] K. Hüper, U. Helmke, and S. Herzberg (2010), *On the computation of means on Grassmann manifolds*, International Symposium on Mathematical Theory of Networks and Systems (MTNS), pp. 2439–2441.

[17] H. Karcher (1977), *Riemannian center of mass and mollifier smoothing*, Comm. Pure Appl. Math., 30, pp. 509–541.

[18] H. Karcher (2014), *Riemannian Center of Mass and So-Called Karcher Mean*, preprint, https://arxiv.org/abs/1407.2087.

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

[20] T. Lam (2008), *Schubert polynomials for the affine Grassmannian*, J. Amer. Math. Soc., 21, pp. 259–281.

[21] J. M. Lee (2003), *Introduction to Smooth Manifolds*, Springer, New York.

[22] G. Lerman and T. Zhang (2011), *Robust recovery of multiple subspaces by geometric l_p minimization*, Ann. Statist., 39, pp. 2686–2715.

[23] L.-H. Lim, K. S.-W. Wong, and K. Ye (2018), *The Grassmannian of Affine Subspaces*, preprint, https://arxiv.org/abs/1807.10883.

[24] Y. Ma, A. Yang, H. Derksen, and R. Fossum (2008), *Estimation of subspace arrangements with applications in modeling and segmenting mixed data*, SIAM Rev., 50, pp. 413–458.

[25] P. Mattila (1995), *Geometry of sets and measures in Euclidean spaces*, Cambridge Stud. Adv. Math. 44, Cambridge University Press, Cambridge.

[26] E. Miller and B. Sturmfels (2005), *Combinatorial Commutative Algebra*, Grad. Texts in Math. 227, Springer, New York.

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

[28] B. St. Thomas, L. Lin, L.-H. Lim, and S. Mukherjee (2014), *Learning Subspaces of Different Dimensions*, preprint, arXiv:1404.6841.

[29] H. Tyagi, E. Vural, and P. Frossard (2013), *Tangent space estimation for smooth embeddings of Riemannian manifolds*, Inf. Inference, pp. 69–114.

[30] R. Vidal, Y. Ma, and S. Sastry (2005), *Generalized principal component analysis*, IEEE Trans. Pattern Anal. Mach. Intell., 27, pp. 1945–1959.

[31] S. T. Yau (1974), *Non-existence of continuous convex functions on certain Riemannian manifolds*, Math. Ann., 207 (1974), pp. 269–270.

[32] K. Ye and L.-H. Lim (2016), *Schubert varieties and distances between subspaces of different dimensions*, SIAM J. Matrix Anal. Appl., 37, pp. 1176–1197.