Mathematical Programming (2022) 194:621–660
https://doi.org/10.1007/s10107-021-01640-3

FULL LENGTH PAPER

Series A

Optimization on flag manifolds

Ke Ye[1] · Ken Sze-Wai Wong[2] · Lek-Heng Lim[3]

Received: 8 August 2019 / Accepted: 10 March 2021 / Published online: 23 June 2021

Abstract

A flag is a sequence of nested subspaces. Flags are ubiquitous in numerical analysis, arising in finite elements, multigrid, spectral, and pseudospectral methods for numerical PDE; they arise in the form of Krylov subspaces in matrix computations, and as multiresolution analysis in wavelets constructions. They are common in statistics too—principal component, canonical correlation, and correspondence analyses may all be viewed as methods for extracting flags from a data set. The main goal of this article is to develop the tools needed for optimizing over a set of flags, which is a smooth manifold called the flag manifold, and it contains the Grassmannian as the simplest special case. We will derive closed-form analytic expressions for various differential geometric objects required for Riemannian optimization algorithms on the flag manifold; introducing various systems of extrinsic coordinates that allow us to parameterize points, metrics, tangent spaces, geodesics, distances, parallel transports, gradients, Hessians in terms of matrices and matrix operations; and thereby permitting us to formulate steepest descent, conjugate gradient, and Newton algorithms on the flag manifold using only standard numerical linear algebra.

KY is supported by NSFC Grant No. 11801548, NSFC Grant No. 11688101, and National Key R&D Program of China Grant No. 2018YFA0306702 and 2020YFA0712300. LHL is supported by DARPA D15AP00109 and HR00112190040, NSF IIS 1546413, DMS 1854831.

✉ Lek-Heng Lim
lekheng@uchicago.edu

Ke Ye
keyk@amss.ac.cn

Ken Sze-Wai Wong
kenwong@uchicago.edu

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

2 Department of Statistics, University of Chicago, Chicago, IL 60637-1514, USA

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

Keywords Flag manifold · Riemannian optimization · Manifold optimization · Multiscale · Multiresolution

Mathematics Subject Classification 62H12 · 14M15 · 90C30 · 62H10 · 68T10

1 Introduction

Launched around 20 years ago in a classic article of Edelman, Arias, and Smith [20], Riemannian *manifold optimization* is now entrenched as a mainstay of optimization theory [2,4,19,51]. While studies of optimization algorithms on Riemannian manifolds predate [20], the distinguishing feature of Edelman et al.'s approach is that their algorithms are built entirely and directly from standard algorithms in numerical linear algebra; in particular, they do not require numerical solutions of differential equations. For instance, the parallel transport of a vector in [20] is not merely discussed in the abstract but may be explicitly computed in efficient and numerically stable ways via closed-form analytic expressions involving QR and singular value decompositions of various matrices.

The requirement that differential geometric quantities appearing in a manifold optimization algorithms have analytic expressions in terms of standard matrix decompositions limits the type of Riemannian manifolds that one may consider. Aside from Euclidean spaces, we know of exactly three[1] Riemannian manifolds [4] on which one may define optimization algorithms in this manner:

(i) Stiefel manifold $\mathrm{V}(k, n)$,
(ii) Grassmann manifold $\mathrm{Gr}(k, n)$,
(iii) manifold of positive definite matrices $\mathbb{S}^n_{++}$.

The main contribution of this article is to furnish a fourth: *flag manifold*.

A flag in a finite-dimensional vector space $\mathbb{V}$ over $\mathbb{R}$ is a nested sequence of linear subspaces $\{\mathbb{V}_i\}_{i=1}^d$ of $\mathbb{V}$, i.e.,

$$\{0\} \subseteq \mathbb{V}_1 \subseteq \cdots \subseteq \mathbb{V}_d \subseteq \mathbb{V}.$$

For any increasing integer sequence of length d, $0 < n_1 < \cdots < n_d < n$, the set of all flags $\{\mathbb{V}_i\}_{i=1}^d$ with $\dim(\mathbb{V}_i) = n_i$, $i = 1, \ldots, d$, is a smooth manifold called a *flag manifold*, and denoted by $\mathrm{Flag}(n_1, \ldots, n_d; \mathbb{V})$. This is a generalization of the Grassmannian $\mathrm{Gr}(k, \mathbb{V})$ that parameterizes k-dimensional linear subspaces in $\mathbb{V}$ as flags of length one are just subspaces, i.e., $\mathrm{Flag}(k; \mathbb{V}) = \mathrm{Gr}(k, \mathbb{V})$. Flag manifolds, sometimes also called flag varieties, were first studied by Ehresmann [21] and saw rapid development in 1950's [11,12,14,17]. They are now ubiquitous in many areas of pure mathematics, and, as we will discuss next, they are also ubiquitous in applied mathematics, just hidden in plain sight.

The optimization algorithms on Grassmann and Stiefel manifolds originally proposed in [20] have found widespread applications: e.g., computer vision [49,50], shape

[1] Discounting manifolds that can be realized as products or open subsets of these manifolds, e.g., those considered in [1,31,35,44].

analysis [43,45], matrix computations [33,48], subspace tracking [8], and numerous other areas—unsurprising as subspaces and their orthonormal bases are ubiquitous in all areas of science and engineering. For the same reason, we expect optimization algorithms on flag manifolds to be similarly useful as flags are also ubiquitous—any multilevel, multiresolution, or multiscale phenomena likely involve flags, whether implicitly or explicitly. We will discuss some examples from numerical analysis and statistics.

1.1 Flags in numerical analysis

In numerical analysis, flags naturally arise in finite elements, multigrid, spectral and pseudospectral methods, wavelets, iterative matrix computations, etc, in several ways.

Example 1 *(Mesh refinement)* In multigrid, algebraic multigrid, finite element methods, we often consider a sequence of increasingly finer *grids* or *meshes* $G_1 \subsetneq G_2 \subsetneq G_3 \subsetneq \cdots$ on the domain of interest Ω. The vector space of real-valued functions

$$\mathbb{V}_k := \{f : G_k \to \mathbb{R}\}$$

gives us a flag $\mathbb{V}_1 \subsetneq \mathbb{V}_2 \subsetneq \mathbb{V}_3 \subsetneq \cdots$ of finite-dimensional vector spaces where $\dim \mathbb{V}_k = |G_k|$, the number of grid points in G_k. The aforementioned numerical methods are essentially different ways of extracting approximate solutions of increasing accuracy from the flag.

Example 2 *(Increasing order)* In spectral and pseudospectral methods, we consider a class of functions of increasing complexity determined by an *order* d, e.g., polynomial or trigonometric polynomial functions of degree d, on the domain of interest Ω. The vector space

$$\mathbb{V}_d := \{f : \Omega \to \mathbb{R} : \deg(f) \leq d\}$$

gives us a flag $\mathbb{V}_1 \subsetneq \mathbb{V}_2 \subsetneq \mathbb{V}_3 \subsetneq \cdots$ as d is increased. Again, these methods operate by extracting approximate solutions of increasing accuracy from the flag.

Example 3 *(Cyclic subspaces)* Given $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$, the subspace

$$K_k(A, b) := \operatorname{span}\{b, Ab, \dots, A^{k-1}b\}$$

is called the kth *Krylov subspace*. The gist behind Krylov subspace methods in numerical linear algebra, whether for computing solutions to linear systems, least squares problems, eigenvalue problems, matrix functions, etc, are all based on finding a sequence of increasingly better approximations from the flag $K_0(A, b) \subsetneq K_1(A, b) \subsetneq \cdots \subsetneq K_k(A, b)$, assuming that A has at least k distinct eigenvalues.

Example 4 *(Multiresolution)* A standard way to construct wavelets is to define a *multiresolution analysis*, i.e., a sequence of subspaces $\mathbb{V}_{k+1} \subsetneq \mathbb{V}_k$ defined by

$$f(t) \in \mathbb{V}_k \quad \Leftrightarrow \quad f(t/2) \in \mathbb{V}_{k+1}.$$

The convention in wavelet literature has the indexing in reverse order but this is a minor matter—a nested of sequence of subspaces is a flag regardless of how the subspaces in the sequence are labeled. So a multiresolution analysis is also a flag.

This is not an exhaustive list, flags also arise in numerical analysis in other ways, e.g., analysis of eigenvalue methods [6,28].

1.2 Flags in statistics

Although not usually viewed in this manner, classical multivariate data analysis techniques [34] may be cast as nested subspace-searching problems, i.e., constrained or unconstrained optimization problems on the flag manifold.

We let $\mathbb{1}$ denote a vector of all ones (of appropriate dimension). We assume that our data set is given in the form of a samples-by-variables design matrix $X \in \mathbb{R}^{n \times p}$, $n \geq p$, which we call a data matrix for short. Let $\overline{x} = \frac{1}{n} X^{\mathsf{T}} \mathbb{1} \in \mathbb{R}^p$ be its sample mean and $S_X = (X - \mathbb{1}\overline{x}^{\mathsf{T}})^{\mathsf{T}}(X - \mathbb{1}\overline{x}^{\mathsf{T}}) \in \mathbb{R}^{p \times p}$ be its sample covariance. For another data matrix $Y \in \mathbb{R}^{n \times q}$, $S_{XY} = (X - \mathbb{1}\overline{x}^{\mathsf{T}})^{\mathsf{T}}(Y - \mathbb{1}\overline{y}^{\mathsf{T}}) = S_{YX}^{\mathsf{T}} \in \mathbb{R}^{p \times q}$ denotes sample cross-covariance.

Example 5 *(Principal Component Analysis (PCA))* The kth *principal subspace* of X is $\operatorname{im}(Z_k)$, where Z_k is the $p \times k$ matrix given by

$$Z_k = \operatorname{argmax}\{\operatorname{tr}(Z^{\mathsf{T}} S_X Z) : Z \in \mathrm{V}(k, p)\}, \quad k = 1, \ldots, p. \tag{1}$$

So $\operatorname{im}(Z_k)$ is a k-dimensional linear subspace of $\mathbb{R}^p$ spanned by the orthonormal columns of Z_k. In an appropriate sense, the kth principal subspace captures the greatest variability in the data among all k-dimensional subspaces of $\mathbb{R}^p$. In principal component analysis, the data points, i.e., columns of X, are often projected onto $\operatorname{im}(Z_k)$ with $k = 2, 3$ for visualization or with other small values of k for dimension reduction. Clearly $\operatorname{im}(Z_k)$ is contained in $\operatorname{im}(Z_{k+1})$ and the flag

$$\operatorname{im}(Z_1) \subsetneq \operatorname{im}(Z_2) \subsetneq \cdots \subsetneq \operatorname{im}(Z_p)$$

explains an increasing amount of variance in the data.

In [42, Theorem 9], it is shown how one may directly define PCA as an optimization problem on a flag manifold, a powerful perspective that in turn allows one to generalize and extend PCA in various manners. Nevertheless what is lacking in [42] is an algorithm for optimization on flag manifolds, a gap that our article will fill.

Example 6 *(Canonical Correlation Analysis (CCA))* The kth pair of *canonical correlation loadings* $(a_k, b_k) \in \mathbb{R}^p \times \mathbb{R}^q$ is defined recursively by

$$(a_k, b_k) = \operatorname{argmax}\{a^\mathsf{T} S_{XY} b : a^\mathsf{T} S_X a = b^\mathsf{T} S_Y b = 1,\\ a^\mathsf{T} S_X a_j = a^\mathsf{T} S_{XY} b_j = b^\mathsf{T} S_{YX} a_j = b^\mathsf{T} S_Y b_j = 0,\ j = 1, \dots, k-1\}. \tag{2}$$

Let $A_k = [a_1, \dots, a_k] \in \mathbb{R}^{p \times k}$ and $B_k = [b_1, \dots, b_k] \in \mathbb{R}^{q \times k}$. Then the canonical correlation subspaces of X and Y are given by

$$\operatorname{im}(A_1) \subsetneq \cdots \subsetneq \operatorname{im}(A_p) \quad \text{and} \quad \operatorname{im}(B_1) \subsetneq \cdots \subsetneq \operatorname{im}(B_q),$$

which are flags in $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively. Collectively they capture how the shared variance between the two data sets increases with k.

Example 7 *(Correspondence Analysis (CA))* Let $t = \mathbb{1}^\mathsf{T} X \mathbb{1} \in \mathbb{R}$, $r = \frac{1}{t} X \mathbb{1} \in \mathbb{R}^n$, $c = \frac{1}{t} X^\mathsf{T} \mathbb{1} \in \mathbb{R}^p$ denote the total, row, and column weights of X respectively and set $D_r = \frac{1}{t} \operatorname{diag}(r) \in \mathbb{R}^{n \times n}$, $D_c = \frac{1}{t} \operatorname{diag}(c) \in \mathbb{R}^{p \times p}$. For $k = 1, \dots, p$, we seek matrices $U_k \in \mathbb{R}^{k \times n}$ and $V_k \in \mathbb{R}^{k \times p}$ such that

$$(U_k, V_k) = \operatorname{argmax}\Big\{\operatorname{tr}\big(U^\mathsf{T}(\tfrac{1}{t}X - rc^\mathsf{T})V\big) : U^\mathsf{T} D_r U = I = V^\mathsf{T} D_c V\Big\}. \tag{3}$$

The solution

$$\operatorname{im}(U_1) \subsetneq \cdots \subsetneq \operatorname{im}(U_p) \quad \text{and} \quad \operatorname{im}(V_1) \subsetneq \cdots \subsetneq \operatorname{im}(V_p)$$

are flags in $\mathbb{R}^n$ and $\mathbb{R}^p$ respectively and collectively they explain the increasing deviation from the independence of occurrence of two outcomes.

For reasons such as sensitivity of the higher-dimensional subspaces to noise in the data, in practice one relies on the first few subspaces in these flags to make various inference about the data. Nevertheless, we stress that the respective flags that solve (1), (2), (3) over all k will paint a complete picture showing the full profile of how variance, shared variance, or deviation from independence vary across dimensions.

Apart from PCA, CCA, and CA, flags arise in other multivariate data analytic techniques [34], e.g., factor analysis (FA), linear discriminant analysis (LDA), multidimensional scaling (MDS), etc, in much the same manner. One notable example is the independent subspace analysis proposed in [38,39], a generalization of independent component analysis.

1.3 Prior work and our contributions

Some elements of optimization theory on flag manifolds have been considered in [38], although optimization is not its main focus and only analytic expressions for tangent spaces and gradients have been obtained. In particular, no actual algorithm appears in [38]—note that a Riemannian steepest descent algorithm in the spirit of [20] would at least require analytic expressions for geodesics and, to the best of our knowledge, they have never been derived; in fact prior to this article it is not even known if such expressions exist.

The main contribution of our article is in providing all necessary ingredients for optimization algorithms on flag manifolds in full details, and from two different perspectives—representing a flag manifold as (i) a homogeneous space, where a flag is represented as an equivalence class of matrices; and as (ii) a compact submanifold of $\mathbb{R}^{n\times n}$, where every flag is uniquely represented by a matrix. We will provide four systems of extrinsic coordinates for representing a flag manifold that arise from (i) and (ii)—while modern differential geometry invariably adopts an intrinsic coordinate-free approach, we emphasize that such suitable extrinsic coordinate systems are indispensable for performing computations on manifolds.

In particular, the analytic expressions for various differential geometric objects and operations required for our optimization algorithms will rely on these coordinate systems. We will supply ready-to-use formulas and algorithms, rigorously proven but also made accessible to applied mathematicians and practitioners. For the readers' convenience, the following is a road map to the formulas and algorithms:

Object on flag manifold	Results
Point	Propositions 4, 12, 17, 21
Tangent vector	Propositions 6, 13, 18, 22, Corollary 2
Metric	Propositions 7, 14, 23
Geodesic	Propositions 8, 9, 15, 19
Arclength	Corollary 1, Proposition 15
Geodesic distance	Proposition 10
Parallel transport	Propositions 11, 16, 20
Gradient	Proposition 24
Hessian	Proposition 25
Steepest descent	Algorithm 1
Conjugate gradient	Algorithm 2

1.4 Outline

We begin by reviewing basic materials about Lie groups, Lie algebras, homogeneous spaces, and Riemannian manifolds (Sect. 2). We then proceed to describe the basic differential geometry of flag manifolds (Sect. 3), develop four concrete matrix representations of flag manifolds, and derive closed-form analytic expressions for various differential geometric objects in terms of standard matrix operations (Sects. 4, 5, 6). With these, standard nonlinear optimization algorithms can be ported to the flag manifold almost as an afterthought (Sect. 7). We illustrate using two numerical experiments with steepest descent on the flag manifold (Sect. 8).

2 Basic differential geometry of homogeneous spaces

We will need some rudimentary properties of homogeneous spaces not typically found in the manifold optimization literature, e.g., [4,20]. While these materials are certainly available in the standard differential geometry literature, e.g., [10,25,29], they are not presented in a form easily accessible to practitioners. This section provides a self-contained review, pared down to a bare minimum of just what we need later.

2.1 Lie groups and Lie algebras

Let M be a smooth manifold and T^*M be its cotangent bundle. A *Riemannian metric* on M is a smooth section $g : M \to T^*M \otimes T^*M$ such that $g_x := g(x) \in T_x^*M \otimes T_x^*M$ is a positive definite symmetric bilinear form on the tangent space T_xM for every $x \in M$. Intuitively, a Riemannian metric gives an inner product on T_xM for every $x \in M$ and it varies smoothly with respect to $x \in M$. Let G be a group and let $m : G \times G \to G$ be the multiplication map $m(a_1, a_2) = a_1a_2$ and $i : G \to G$ be the inversion map $i(a) = a^{-1}$. Then G is a *Lie group* if it is a smooth manifold and the group operations m and i are smooth maps. The tangent space $\mathfrak{g}$ of G at the identity $e \in G$ is a *Lie algebra*, i.e., a vector space equipped with a *Lie bracket*, a bilinear map $[\cdot, \cdot] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ satisfying $[X, Y] = -[Y, X]$ (skew-symmetry) and $[X, [Y, Z]] + [Z, [X, Y]] + [Y, [Z, X]] = 0$ (Jacobi identity). For example, if G is the orthogonal group $\mathrm{O}(n)$ of all $n \times n$ real orthogonal matrices, then its Lie algebra $\mathfrak{so}(n)$ is the vector space of all $n \times n$ real skew-symmetric matrices.

For a Lie group G, we may define the *left* and *right translation* maps $L_a, R_a : G \to G$ by $L_a(x) = m(a, x) = ax$ and $R_a(x) = m(x, a) = xa$. We say that a Riemannian metric g on G is *left invariant* if for all $a \in G$,

$$g_{L_a(x)}\big((dL_a)_x(X), (dL_a)_x(Y)\big) = g_x(X, Y);$$

right invariant if for all $b \in G$,

$$g_{R_b(x)}\big((dR_b)_x(X), (dR_b)_x(Y)\big) = g_x(X, Y);$$

and *bi-invariant* if for all $a, b \in G$,

$$g_{R_b \circ L_a(x)}\big((d(R_b \circ L_a))_x(X), (d(R_b \circ L_a))_x(Y)\big) = g_x(X, Y)$$

over all $X, Y \in T_xM$. These notions are pertinent for Sect. 2.2: There are infinitely many Riemannian metrics that one could put on the flag manifold but there is a natural choice that is induced from the unique bi-invariant metric on $\mathrm{O}(n)$ and this is the one that we use.

2.2 Homogeneous spaces

We now recall some basic definitions and facts about homogeneous spaces. Throughout this article, we will use double brackets $[\![x]\!]$ to denote the equivalence class of x.

Definition 1 Let G be a Lie group acting on a smooth manifold M via $\varphi : G \times M \to M$. If the action φ is smooth and transitive, i.e., for any $x, y \in M$, there is some $a \in G$ such that $\varphi(a, x) = y$, then M is called a *homogeneous space* of the Lie group G.

For a point $x \in M$, the subgroup $G_x = \{a \in G : \varphi(a, x) = x\}$ is called the *isotropy group* of x. We write G/G_x for the quotient group of G by G_x and denote by $[\![a]\!] \in G/G_x$ the *coset* (or equivalence class) of $a \in G$. Since G acts on M transitively, we see that there is a one-to-one correspondence F between G/G_x and M given by

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

for any $x \in M$. In fact, F defines a diffeomorphism between the two smooth manifolds, which is the content of the following theorem [10, Theorems 9.2 and 9.3].

Theorem 1 *Let G be a Lie group acting on a smooth manifold M. For any $x \in M$, there exists a unique smooth structure on G/G_x such that the action*

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

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

The Grassmannian $\mathrm{Gr}(k, n)$ of k-dimensional subspaces in $\mathbb{R}^n$ is probably the best known example of a homogeneous space in manifold optimization. Indeed, $\mathrm{O}(n)$ acts transitively on $\mathrm{Gr}(k, n)$ and as any k-dimensional subspace $\mathbb{W} \subseteq \mathbb{R}^n$ has isotropy group isomorphic to $\mathrm{O}(k) \times \mathrm{O}(n-k)$, we obtain the well-known characterization of Grassmannian

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

that is crucial for manifold optimization. Throughout this article '$\cong$' will mean diffeomorphism.

Let G be a Lie group and M a homogeneous space of G with action $\varphi : G \times M \to M$. Fix any $x \in M$ and let H denote its isotropy group. By Theorem 1 we may identify M with G/H. The left translation map in Sect. 2.1 may be extended to this setting as $L_a : M \to M$, $L_a(y) = \varphi(a, y)$ for any $a \in G$. In particular, if $a \in H$, then $L_a(x) = x$, and we have a linear isomorphism

$$(dL_a)_x : T_x M \to T_x M.$$

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

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

As $M = G/H$, we have $T_{[\![e]\!]}M = \mathfrak{g}/\mathfrak{h}$ where $\mathfrak{g}$ and $\mathfrak{h}$ are the Lie algebras of G and H respectively. Here $e \in G$ is the identity element. This allows us to define the *adjoint representation* $\mathrm{Ad}_H : H \to \mathrm{GL}(\mathfrak{g}/\mathfrak{h})$, $a \mapsto d(L_a \circ R_{a^{-1}})_{[\![e]\!]}$; recall that $\mathrm{GL}(\mathbb{V})$ denotes the group of linear isomorphisms on the vector space $\mathbb{V}$ and here we take $\mathbb{V} = \mathfrak{g}/\mathfrak{h}$. In other words, for any $a \in H$ and $X \in \mathfrak{g}/\mathfrak{h}$,

$$\mathrm{Ad}_H(a)(X) = d(L_a \circ R_{a^{-1}})_{[\![e]\!]}(X).$$

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

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

We state an important result about their existence and construction [16, Proposition 3.16].

Proposition 1 *Let G be a connected Lie group and H a closed Lie subgroup with Lie algebras $\mathfrak{g}$ and $\mathfrak{h}$ respectively. If there is a subspace $\mathfrak{m}$ of $\mathfrak{g}$ such that $\mathfrak{g} = \mathfrak{m} \oplus \mathfrak{h}$ and $\mathrm{Ad}_H(\mathfrak{m}) \subseteq \mathfrak{m}$, then there is a one-to-one correspondence between G-invariant metrics on $M = G/H$ and Ad_H-invariant inner products on $\mathfrak{m}$.*

Proposition 1 says that if $\mathfrak{h} \subseteq \mathfrak{g}$ admits a complement $\mathfrak{m}$, then we may obtain a G-invariant metric g on M by an Ad_H-invariant inner product on $\mathfrak{m}$. Moreover, we may identify T_xM with $\mathfrak{m}$, implying that the metric g on M is essentially determined by g_x at a single arbitrary point $x \in M$.

Proposition 2 *If G is a compact Lie group, then G admits a bi-invariant metric and this metric induces a G-invariant metric g on $M = G/H$ for any closed subgroup $H \subseteq G$.*

If in addition G is *simple*, i.e., has no nontrivial connected normal subgroups, then G admits a unique bi-invariant metric called the *canonical metric* and in this case M is called a *normal homogeneous space*. A flag manifold is a normal homogeneous space and the metric we derive later in Propositions 7, 14, 23 comes from the unique bi-invariant metric on $\mathrm{O}(n)$.

2.3 Geodesic orbit spaces

Let $M = G/H$ be a homogeneous space of G. If M has a Riemannian metric g such that every geodesic in M is an orbit of a one-parameter subgroup of G, then we say that (M, g) is a *geodesic orbit space*. The following result [30] will allow us to construct several interesting examples.

Theorem 2 *Let G be a compact Lie group with a bi-invariant metric g and H be a subgroup such that $M = G/H$ is a smooth manifold (e.g., H is closed subgroup). Then $M = G/H$ together with the metric $\widetilde{g}$ induced by g is a geodesic orbit space.*

In general it is difficult if not impossible to determine closed-form analytic expressions for geodesics on a Riemannian manifold. But in the case of a geodesic orbit space, since its geodesics are simply orbits of one-parameter subgroups of G, the task reduces to determining the latter. The next result [23, Theorem 1.3.5] will be helpful towards this end.

Theorem 3 *If G is a matrix Lie group , then every one-parameter subgroup $\gamma(t)$ of G is of the form*

$$\gamma(t) = \exp(ta) := \sum_{k=0}^{\infty} \frac{t^k a^k}{k!}$$

for some $a \in \mathfrak{g}$.

So for example, every one-parameter subgroup of $\mathrm{SO}(n)$ must take the form $\gamma(t) = \exp(ta)$ for some skew-symmetric matrix $a \in \mathfrak{so}(n)$.

Fortuitously, as we will see in Sect. 4, a flag manifold is a geodesic orbit space G/H where G can be either $\mathrm{O}(n)$ or $\mathrm{SO}(n)$ with an appropriate choice of subgroup H. This is the key to obtaining closed-form analytic expressions for geodesics in Propositions 8, 9, 15, 19.

2.4 Riemannian notions

Although not specific to homogeneous or geodesic orbit spaces, we state the famous Hopf–Rinow theorem [16, Theorem 1.8] and recall the definitions of Riemannian gradient and Hessian [10,25,29] below for easy reference.

Theorem 4 (Hopf–Rinow) *Let (M, g) be a connected Riemannian manifold. Then the following statements are equivalent:*

(i) closed and bounded subsets of M are compact;
(ii) M is a complete metric space;
(iii) M is geodesically complete, i.e., the exponential map $\exp_x : T_x M \to M$ is defined on the whole $T_x M$ for all $x \in M$.

Furthermore, any one of these conditions guarantees that any two points x, y on M can be connected by a distance minimizing geodesic on M.

A flag manifold satisfies the Hopf–Rinow theorem, ensuring that there is a geodesic curve γ_x connecting any initial point x with any optimal point x_*. Various path-following algorithms may then be viewed as different ways of alternating between approximating the geodesic curve γ_x and updating the initial point x. Also, Theorem 4(iii) guarantees that the exponential map is well-defined in our steepest descent (Algorithm 1) and conjugate gradient (Algorithm 2) methods.

In the following we will write $\mathfrak{X}(M)$ for the set of all smooth vector fields on M.

Definition 2 (*Riemannian gradient and Hessian*) Let (M, g) be a Riemannian manifold. Let $f : M \to \mathbb{R}$ be a smooth function. The *Riemannian gradient* of f, denoted ∇f, is defined by

$$g(\nabla f, V) = V(f),$$

for any $V \in \mathfrak{X}(M)$. The *Riemannian Hessian* of f, denoted $\nabla^2 f$, is defined by

$$(\nabla^2 f)(U, V) = g\big(\nabla_U(\nabla f), V\big),$$

where $U, V \in \mathfrak{X}(M)$ and $\nabla_U V$ is the *covariant derivative*[2] of V along U, which is uniquely determined by the Riemannian metric g.

By their definitions, ∇f is a smooth vector field and $\nabla^2 f$ is a smooth field of *symmetric bilinear forms*. In particular, $\nabla^2 f$ is uniquely determined by its values at points of the form (V, V) over all $V \in \mathfrak{X}(M)$ because of bilinearity and symmetry, i.e.,

$$\nabla^2 f(U, V) = \frac{1}{2}\big(\nabla^2 f(U+V, U+V) - \nabla^2 f(U, U) - \nabla^2 f(V, V)\big), \tag{4}$$

for any $U, V \in \mathfrak{X}(M)$. Definition 2 is standard but not as useful for us as a *pointwise* definition—the Riemannian gradient $\nabla f(x)$ and Riemannian Hessian $\nabla^2 f(x)$ at a point $x \in M$ is given by

$$g_x(\nabla f(x), X) = \frac{df\big(\exp(tX)\big)}{dt}\bigg|_{t=0}, \quad \nabla^2 f(x)(X, X) = \frac{d^2 f\big(\exp(tX)\big)}{dt^2}\bigg|_{t=0}, \tag{5}$$

where $\exp(tX)$ is the geodesic curve emanating from x in the direction $X \in T_x M$. We may obtain (5) by Taylor expanding $f\big(\exp(tX)\big)$.

Given a specific function f, one may express (5) in terms of *local coordinates* on M but in general there are no global formulas for $\nabla f(x)$ and $\nabla^2 f(x)$, and without which it would be difficult if not impossible to do optimization on M. We will see in Sect. 6 that when M is a flag manifold, then the gradient and Hessian in (5) may be expressed globally in terms of extrinsic coordinates.

3 Basic differential geometry of flag manifolds

We will now define flags and flag manifolds formally and discuss some basic properties. Let n be a positive integer and $\mathbb{V}$ be an n-dimensional vector space over $\mathbb{R}$. We write $\mathrm{V}(k, \mathbb{V})$ for the Stiefel manifold [46] of orthonormal k-frames in $\mathbb{V}$ and $\mathrm{Gr}(k, \mathbb{V})$ for the Grassmannian [24] of k-dimensional subspaces in $\mathbb{V}$. If the choice of $\mathbb{V}$ is unimportant or if $\mathbb{V} = \mathbb{R}^n$, then we will just write $\mathrm{V}(k, n)$ and $\mathrm{Gr}(k, n)$.

[2] More precisely, the covariant derivative associated with the Levi-Civita connection on M.

Definition 3 Let $0 < n_1 < \cdots < n_d < n$ be an increasing sequence of d positive integers and $\mathbb{V}$ be an n-dimensional vector space over $\mathbb{R}$. A flag of type $(n_1, \ldots, n_d)$ in $\mathbb{V}$ is a sequence of subspaces

$$\mathbb{V}_1 \subseteq \mathbb{V}_2 \subseteq \cdots \subseteq \mathbb{V}_d, \quad \dim \mathbb{V}_i = n_i, \quad i = 1, \ldots, d.$$

We denote the set of such flags by $\operatorname{Flag}(n_1, \ldots, n_d; \mathbb{V})$ and call it the *flag manifold* of type $(n_1, \ldots, n_d)$. If $\mathbb{V}$ is unimportant or if $\mathbb{V} = \mathbb{R}^n$, then we will just write $\operatorname{Flag}(n_1, \ldots, n_d; n)$.

For notational convenience we will adopt the following convention throughout:

$$n_0 := 0, \qquad n_{d+1} := n, \qquad \mathbb{V}_0 := \{0\}, \qquad \mathbb{V}_{d+1} := \mathbb{V}.$$

We will see in Proposition 3 that flag manifolds are indeed manifolds. When $d = 1$, $\operatorname{Flag}(k; \mathbb{V})$ is the set of all k-dimensional subspaces of $\mathbb{V}$, which is the Grassmannian $\operatorname{Gr}(k, \mathbb{V})$. The other extreme case is when $d = n - 1$ and $n_i = i$, $i = 1, \ldots, n-1$, and in which case $\operatorname{Flag}(1, \ldots, n-1; \mathbb{V})$ comprises all *complete flags* of $\mathbb{V}$, i.e.,

$$\mathbb{V}_1 \subseteq \mathbb{V}_2 \subseteq \cdots \subseteq \mathbb{V}_{n-1}, \quad \dim \mathbb{V}_i = i, \quad i = 1, \ldots, n-1.$$

Like the Grassmannian, the flag manifold is not merely a set but has rich geometric structures. We will start with the most basic ones and defer other useful characterizations to Sects. 4 and 5.

Proposition 3 *Let $0 < n_1 < \cdots < n_d < n$ be integers and $\mathbb{V}$ be an n-dimensional real vector space. The flag manifold $\operatorname{Flag}(n_1, \ldots, n_d; \mathbb{V})$ is*

(i) a connected compact smooth manifold;
(ii) an irreducible affine variety;
(iii) a closed submanifold of $\operatorname{Gr}(n_1, \mathbb{V}) \times \operatorname{Gr}(n_2, \mathbb{V}) \times \cdots \times \operatorname{Gr}(n_d, \mathbb{V})$;
(iv) a closed submanifold of $\operatorname{Gr}(n_1, \mathbb{V}) \times \operatorname{Gr}(n_2 - n_1, \mathbb{V}) \times \cdots \times \operatorname{Gr}(n_d - n_{d-1}, \mathbb{V})$;
(v) a fiber bundle on $\operatorname{Gr}(n_d, \mathbb{V})$ whose fiber over $\mathbb{W} \in \operatorname{Gr}(n_d, \mathbb{V})$ is $\operatorname{Flag}(n_1, \ldots, n_{d-1}; \mathbb{W})$;
(vi) a smooth projective variety.

Proof Property (i) is well-known [13,36] but also follows from the characterization in Proposition 4 as a quotient of a compact connected Lie group by a closed subgroup. Property (ii) is a consequence of Propositions 17 and 21, where we give two different ways of representing $\operatorname{Flag}(n_1, \ldots, n_d; \mathbb{V})$ as an affine variety in $\mathbb{R}^m$, $m = (nd)^2$. Property (vi) is a consequence of (iii) or (iv), given that the Grassmannian is a projective variety.

In the following, let $\{\mathbb{V}_i\}_{i=1}^d \in \operatorname{Flag}(n_1, \ldots, n_d; \mathbb{V})$, i.e., $\dim \mathbb{V}_i = n_i$, $i = 1, \ldots, d$. For (iii), the map

$$\begin{aligned} \varepsilon : \operatorname{Flag}(n_1, \ldots, n_d; \mathbb{V}) &\to \operatorname{Gr}(n_1, \mathbb{V}) \times \cdots \times \operatorname{Gr}(n_d, \mathbb{V}), \\ \{\mathbb{V}_i\}_{i=1}^d &\mapsto (\mathbb{V}_1, \mathbb{V}_2, \ldots, \mathbb{V}_d) \end{aligned} \tag{6}$$

is clearly an embedding. Its image is closed since if $(\mathbb{V}_1, \dots, \mathbb{V}_d) \notin \varepsilon\bigl(\mathrm{Flag}(n_1, \dots, n_d; \mathbb{V})\bigr)$, then there exists some $i \in \{1, \dots, d-1\}$ such that $\mathbb{V}_i \nsubseteq \mathbb{V}_{i+1}$; so if $\mathbb{V}'_i \in \mathrm{Gr}(n_i, \mathbb{V})$ and $\mathbb{V}'_{i+1} \in \mathrm{Gr}(n_{i+1}, \mathbb{V})$ are in some small neighborhood of $\mathbb{V}_i$ and $\mathbb{V}_{i+1}$ respectively, then $\mathbb{V}'_i \nsubseteq \mathbb{V}'_{i+1}$.

For (iv), choose and fix an inner product on $\mathbb{V}$. Let $\mathbb{V}_i^{\perp}$ denote the orthogonal complement of $\mathbb{V}_i$ in $\mathbb{V}_{i+1}$, $i = 1, \dots, d-1$. The map

$$\begin{aligned} \varepsilon' : \mathrm{Flag}(n_1, \dots, n_d; \mathbb{V}) &\to \mathrm{Gr}(n_1, \mathbb{V}) \times \mathrm{Gr}(n_2 - n_1, \mathbb{V}) \times \cdots \times \mathrm{Gr}(n_d - n_{d-1}, \mathbb{V}), \\ \{\mathbb{V}_i\}_{i=1}^d &\mapsto (\mathbb{V}_1, \mathbb{V}_1^{\perp}, \dots, \mathbb{V}_{d-1}^{\perp}) \end{aligned} \tag{7}$$

is clearly an embedding. That the image of ε' is closed follows from the same argument used for ε.

For (v), consider the map

$$\rho : \mathrm{Flag}(n_1, \dots, n_d; \mathbb{V}) \to \mathrm{Gr}(n_d, \mathbb{V}), \quad \{\mathbb{V}_i\}_{i=1}^d \mapsto \mathbb{V}_d,$$

which is clearly surjective and smooth. For any $\mathbb{W} \in \mathrm{Gr}(n_d, \mathbb{V})$, $\rho^{-1}(\mathbb{W})$ consists of flags of the form

$$\mathbb{V}'_1 \subseteq \mathbb{V}'_2 \subseteq \cdots \subseteq \mathbb{V}'_{d-1} \subseteq \mathbb{W}, \quad \dim \mathbb{V}'_i = n_i, \quad i = 1, \dots, d-1.$$

In other words, the fiber $\rho^{-1}(\mathbb{W}) \cong \mathrm{Flag}(n_1, \dots, n_{d-1}; \mathbb{W})$. $\square$

The fiber bundle structure in Proposition 3(v) may be recursively applied to get

$$\begin{gathered} \mathrm{Flag}(n_1, \dots, n_{d-1}; n_d) \to \mathrm{Flag}(n_1, \dots, n_d; n) \to \mathrm{Gr}(n_d, n), \\ \mathrm{Flag}(n_1, \dots, n_{d-2}; n_{d-1}) \to \mathrm{Flag}(n_1, \dots, n_{d-1}; n_d) \to \mathrm{Gr}(n_{d-1}, n), \end{gathered}$$

and so on, ending in the well-known characterization of the Stiefel manifold as a principal bundle over the Grassmannian

$$\mathrm{O}(k) \to \mathrm{V}(k, n) \to \mathrm{Gr}(k, n).$$

In the next two sections, we will see how the flag manifold may be equipped with extrinsic matrix coordinates and be represented as either homogeneous spaces of matrices (Sect. 4) or manifolds of matrices (Sect. 5) that in turn give closed-form analytic expressions for various differential geometric objects and operations needed for optimization algorithms.

4 Flag manifolds as matrix homogeneous spaces

We will discuss three representations of the flag manifold as *matrix homogeneous spaces*, i.e., where a flag is represented as an equivalence class of matrices:

$$\begin{aligned}\operatorname{Flag}(n_1, \dots, n_d; n) &\cong \mathrm{O}(n)/\big(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1)\\ &\qquad \times \dots \times \mathrm{O}(n_d - n_{d-1}) \times \mathrm{O}(n - n_d)\big),\end{aligned} \tag{8}$$

$$\operatorname{Flag}(n_1, \dots, n_d; n) \cong \mathrm{V}(n_d, n)/\big(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \dots \times \mathrm{O}(n_d - n_{d-1})\big), \tag{9}$$

$$\begin{aligned}\operatorname{Flag}(n_1, \dots, n_d; n) &\cong \mathrm{SO}(n)/\operatorname{S}\big(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1)\\ &\qquad \times \dots \times \mathrm{O}(n_d - n_{d-1}) \times \mathrm{O}(n - n_d)\big).\end{aligned} \tag{10}$$

The characterization (8) is standard [13,36] and generalizes the well-known characterization of the Grassmannian as $\operatorname{Gr}(k, n) \cong \mathrm{O}(n)/\big(\mathrm{O}(k) \times \mathrm{O}(n-k)\big)$ whereas the characterization (9) generalizes another well-known characterization of the Grassmannian as $\operatorname{Gr}(k, n) \cong \mathrm{V}(k, n)/\mathrm{O}(k)$. In the last characterization (10), $\mathrm{SO}(n) = \{Q \in \mathrm{O}(n) : \det(Q) = 1\}$ is the special orthogonal group and $\operatorname{S}\big(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \dots \times \mathrm{O}(n_d - n_{d-1}) \times \mathrm{O}(n - n_d)\big)$, formally defined in Proposition 4, is the group of unit-determinant block diagonal matrices with orthogonal blocks.

Nevertheless, we will soon see that it is desirable to describe $\operatorname{Flag}(n_1, \dots, n_d; n)$ as a homogeneous space G/H where G is a connected Lie group—note that $\mathrm{O}(n)$ is not connected whereas $\mathrm{V}(n_d, n)$ is not a group, so (8) and (9) do not meet this criterion. With this in mind, we state and prove (10) formally.

Proposition 4 *Let $0 < n_1 < \dots < n_d < n$ be d positive integers. The flag manifold $\operatorname{Flag}(n_1, \dots, n_d; n)$ is diffeomorphic to the homogeneous space*

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

where $\operatorname{S}\big(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \dots \times \mathrm{O}(n_d - n_{d-1}) \times \mathrm{O}(n - n_d)\big)$ is the subgroup of unit-determinant block diagonal matrices with orthogonal blocks, i.e.,

$$\begin{bmatrix} Q_1 & 0 & \cdots & 0\\ 0 & Q_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & Q_{d+1}\end{bmatrix} \in \mathrm{O}(n), \quad Q_i \in \mathrm{O}(n_i - n_{i-1}), \quad i = 1, \dots, d+1, \quad \prod_{i=1}^{d+1} \det(Q_i) = 1.$$

Proof We start with the characterization (8), i.e., in this proof we assume '=' in place of '≅' in (8). We claim that the required diffeomorphism τ is given as in the commutative diagram below:

$$\begin{array}{ccc} \mathrm{SO}(n) & \xrightarrow{\quad\quad j \quad\quad} & \mathrm{O}(n) \\ {\scriptstyle \pi'}\downarrow & & \downarrow{\scriptstyle \pi} \\ \mathrm{SO}(n)/\operatorname{S}\big(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \dots \times \mathrm{O}(n - n_d)\big) & \xrightarrow{\tau} & \operatorname{Flag}(n_1, \dots, n_d; n) \end{array}$$

Here j is the inclusion of $\mathrm{SO}(n)$ in $\mathrm{O}(n)$, π and π' the respective quotient maps, and τ the induced map. Since

$$\begin{aligned}&\mathrm{SO}(n) \cap \big(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n - n_d)\big)\\&\quad = \mathrm{S}\big(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n - n_d)\big),\end{aligned}$$

τ is injective. To show that it is surjective, let $\{\mathbb{V}_i\}_{i=1}^d \in \mathrm{Flag}(n_1, \ldots, n_d; n)$ be a flag represented by some $A \in \mathrm{O}(n)$, i.e., $\pi(A) = \{\mathbb{V}_i\}_{i=1}^d$. If $\det(A) = 1$, then we already have $\tau(\pi'(A)) = \{\mathbb{V}_i\}_{i=1}^d$ by commutativity of the diagram. If $\det(A) = -1$, take any $A_1 \in \mathrm{O}(n_1)$ with $\det(A_1) = -1$, set

$$B = A \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & I_{n_2-n_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I_{n-n_d} \end{bmatrix} \in \mathrm{SO}(n),$$

and observe that $\tau(\pi'(B)) = \pi(B) = \pi(A)$. ◻

4.1 Orthogonal coordinates for the flag manifold

An immediate consequence of Proposition 4 is that the flag manifold is connected. The characterization (10) says that a point on $\mathrm{Flag}(n_1, \ldots, n_d; n)$ may be represented by the equivalence class of matrices

$$[\![Q]\!] = \left\{ Q \begin{bmatrix} Q_1 & 0 & \cdots & 0 \\ 0 & Q_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_{d+1} \end{bmatrix} : Q_i \in \mathrm{O}(n_i - n_{i-1}),\ i = 1, \ldots, d+1,\ \prod_{i=1}^{d+1} \det Q_i = 1 \right\} \tag{11}$$

for some $Q \in \mathrm{SO}(n)$. We will call such a representation *orthogonal coordinates* for the flag manifold.

The Lie algebra of $\mathrm{S}\big(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n - n_d)\big)$ is simply $\mathfrak{so}(n_1) \times \mathfrak{so}(n_2 - n_1) \times \cdots \times \mathfrak{so}(n - n_d)$, which we will regard as a Lie subalgebra of block diagonal matrices,

$$\mathfrak{h} = \left\{ \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{d+1} \end{bmatrix} \in \mathfrak{so}(n) : \begin{array}{r} A_1 \in \mathfrak{so}(n_1),\, A_2 \in \mathfrak{so}(n_2 - n_1), \ldots \\ \ldots, A_{d+1} \in \mathfrak{so}(n - n_d) \end{array} \right\}. \tag{12}$$

Let $\mathfrak{m}$ be the natural complement of $\mathfrak{h}$ in $\mathfrak{so}(n)$,

$$\mathfrak{m} = \left\{ \begin{bmatrix} 0 & B_{1,2} & \cdots & B_{1,d+1} \\ -B_{1,2}^{\mathsf{T}} & 0 & \cdots & B_{2,d+1} \\ \vdots & \vdots & \ddots & \vdots \\ -B_{1,d+1}^{\mathsf{T}} & -B_{2,d+1}^{\mathsf{T}} & \cdots & 0 \end{bmatrix} \in \mathfrak{so}(n) : \begin{array}{r} B_{ij} \in \mathbb{R}^{(n_i - n_{i-1}) \times (n_j - n_{j-1})}, \\ 1 \le i < j \le d+1 \end{array} \right\}. \tag{13}$$

In particular, we have the direct sum decomposition $\mathfrak{so}(n) = \mathfrak{h} \oplus \mathfrak{m}$ as vector spaces.

The groups $\mathrm{O}(n)$ and $\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n - n_d)$ have the same Lie algebras as $\mathrm{SO}(n)$ and $\mathrm{S}\big(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n - n_d)\big)$, namely, $\mathfrak{so}(n)$ and $\mathfrak{so}(n_1) \times \mathfrak{so}(n_2 - n_1) \times \cdots \times \mathfrak{so}(n - n_d)$ respectively. If $e \in G$ is the identity element of G, then the tangent space of a homogeneous space G/H at any point is a translation of the tangent space at $[\![e]\!] \in G/H$, which depends only on the Lie algebras $\mathfrak{g}$ and $\mathfrak{h}$ of G and H respectively, as

$$T_{[\![e]\!]} G/H \simeq \mathfrak{g}/\mathfrak{h},$$

a fact that we will use in the proof of Proposition 6.

As such we do not need to distinguish the two homogeneous space structures (8) and (10) when we discuss geometric quantities associated with tangent spaces, e.g., geodesic, gradient, Hessian, parallel transport. In the sequel we will make free use of this flexibility in switching between (8) and (10).

Proposition 5 *Let* $\mathfrak{h}$ *and* $\mathfrak{m}$ *be as in (12) and (13) and* $H = \mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n - n_d)$. *Then the subspace* $\mathfrak{m}$ *is* Ad_H*-invariant, i.e.,* $\mathrm{Ad}(a)(X) \in \mathfrak{m}$ *for every* $a \in H$ *and* $X \in \mathfrak{m}$.

Proof We need to show that $\mathrm{Ad}(a)(X) \in \mathfrak{m}$ whenever $a \in H$ and $X \in \mathfrak{m}$. For notational simplicity, we assume $d = 2$. Let

$$a = \begin{bmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ 0 & 0 & A_3 \end{bmatrix} \qquad \text{and} \qquad X = \begin{bmatrix} 0 & B_{1,2} & B_{1,3} \\ -B_{1,2}^{\mathsf{T}} & 0 & B_{2,3} \\ -B_{1,3}^{\mathsf{T}} & -B_{2,3}^{\mathsf{T}} & 0 \end{bmatrix},$$

where $A_i \in \mathrm{O}(n_i - n_{i-1}), i = 1, 2, 3,$ and $B_{ij} \in \mathbb{R}^{(n_i - n_{i-1}) \times (n_j - n_{j-1})}, 1 \le i < j \le 3$. Then $\mathrm{Ad}(a)(X) = aXa^{-1} = aXa^{\mathsf{T}}$ since a is an orthogonal matrix; and we have

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

as required. □

We now have all the ingredients necessary for deriving closed-form analytic expressions for the tangent space, metric, geodesic, geodesic distance, and parallel transport on a flag manifold in orthogonal coordinates. We begin with the representation of a tangent space as a vector space of matrices.

Proposition 6 (Tangent space I) *Let* $[\![Q]\!] \in \mathrm{Flag}(n_1, \dots, n_d; n) = \mathrm{O}(n)/\big(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n_d - n_{d-1}) \times \mathrm{O}(n - n_d)\big)$ *be represented by* $Q \in \mathrm{O}(n)$. *Its* *tangent space at* $[\![Q]\!]$ *is given by*

$$\begin{aligned} T_{[\![Q]\!]} \operatorname{Flag}(n_1, \dots, n_d; n) &= \{QB \in \mathbb{R}^{n\times n} : B \in \mathfrak{m}\} \\ &= \left\{ Q \begin{bmatrix} 0 & B_{1,2} & \cdots & B_{1,d+1} \\ -B_{1,2}^{\mathsf{T}} & 0 & \cdots & B_{2,d+1} \\ \vdots & \vdots & \ddots & \vdots \\ -B_{1,d+1}^{\mathsf{T}} & -B_{2,d+1}^{\mathsf{T}} & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{n\times n} : \begin{array}{r} B_{i,j} \in \mathbb{R}^{(n_i - n_{i-1})\times(n_j - n_{j-1})}, \\ 1 \le i < j \le d+1 \end{array} \right\}. \end{aligned}$$

In particular, the dimension of a flag manifold is given by

$$\dim \operatorname{Flag}(n_1, \dots, n_d; n) = \sum_{i=1}^{d} (n_i - n_{i-1})(n - n_i).$$

Proof Let $M = \operatorname{Flag}(n_1, \dots, n_d; n)$. For $Q = I$, the identity matrix, this follow from $T_{[\![I]\!]}M \simeq \mathfrak{g}/\mathfrak{h} \simeq \mathfrak{m}$. For Q arbitrary, the left translation $L_Q : M \to M$ is a diffeomorphism, which means that $(dL_Q)_{[\![I]\!]} : T_{[\![I]\!]}M \to T_{[\![Q]\!]}M$ is an isomorphism. The result then follows from $(dL_Q)_{[\![I]\!]}(X) = QX$ for all $X \in T_{[\![I]\!]}M$. □

There are several ways to equip $\operatorname{Flag}(n_1, \dots, n_d; n)$ with a Riemannian metric but there is a distinguished choice that is given by a negative multiple of the *Killing form* of $\mathfrak{so}(n)$, although we will not need to introduce this concept.

Proposition 7 (Riemannian metric I) *The metric* g *on* $\operatorname{Flag}(n_1, \dots, n_d; n)$ *defined by*

$$g_{[\![Q]\!]}(X, Y) = \frac{1}{2} \operatorname{tr}(X^{\mathsf{T}} Y) \tag{14}$$

for all $X, Y \in T_{[\![Q]\!]} \operatorname{Flag}(n_1, \dots, n_d; n)$ *is an* $\mathrm{SO}(n)$*-invariant metric. If we write*

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

where $B_{ij}, C_{ij} \in \mathbb{R}^{(n_i - n_{i-1})\times(n_j - n_{j-1})}$, $1 \le i < j \le d$*, then* g *may be expressed as*

$$g_{[\![Q]\!]}(X, Y) = \sum_{1 \le i < j \le d+1} \operatorname{tr}(B_{ij}^{\mathsf{T}} C_{ij}). \tag{15}$$

Proof We will first need to establish an $\mathrm{Ad}_{\mathrm{SO}(n)}$-invariant inner product on $\mathfrak{so}(n)$. It is a standard fact [41] that bi-invariant metrics on a Lie group G are in one-to-one correspondence with Ad_G-invariant inner products on its Lie algebra $\mathfrak{g}$. In our case, $G = \mathrm{SO}(n)$, $\mathfrak{g} = \mathfrak{so}(n)$, and $\mathrm{Ad}_{\mathrm{SO}(n)} : \mathrm{SO}(n) \to \mathrm{GL}(\mathfrak{so}(n))$. Since $\mathrm{SO}(n)$ is compact, by Proposition 2 it has a bi-invariant metric, which corresponds to an $\mathrm{Ad}_{\mathrm{SO}(n)}$-invariant inner product on $\mathfrak{so}(n)$.

When $n \neq 2, 4$, $\mathfrak{so}(n)$ is a simple Lie algebra and so the $\mathrm{Ad}_{\mathrm{SO}(n)}$-invariant inner product is unique up to a scalar multiple. When $n = 2$, SO(2) is one-dimensional and thus abelian, so the bi-invariant metric on SO(2) is unique up to a scalar. When $n = 4$, $\mathrm{SO}(4) \simeq \mathrm{SO}(3) \times \mathrm{SO}(3)$ as Lie groups, so it has a two-dimensional family of bi-invariant metrics. For all values of n, we may take our $\mathrm{Ad}_{\mathrm{SO}(n)}$-invariant inner product (the choice is unique for all $n \neq 4$) as

$$\langle X, Y\rangle := \frac{1}{2}\operatorname{tr}(X^{\mathsf{T}}Y) \tag{16}$$

for all $X, Y \in \mathfrak{so}(n)$.

Let $G = \mathrm{SO}(n)$ and $H = \mathrm{S}\big(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n - n_d)\big)$. We will use the characterization of a flag manifold in (10), i.e., $\mathrm{Flag}(n_1, \dots, n_d; n) = G/H$. Since $\mathfrak{m}$ is a subspace of $\mathfrak{so}(n)$, the restriction of $\langle\cdot,\cdot\rangle$ in (16) to $\mathfrak{m}$, denoted by $\langle\cdot,\cdot\rangle_{\mathfrak{m}}$, is an inner product on $\mathfrak{m}$. It is easy to verify that $\langle\cdot,\cdot\rangle_{\mathfrak{m}}$ is Ad_H-invariant. Taken together with Propositions 1, 2, and 5, we have that $\langle\cdot,\cdot\rangle_{\mathfrak{m}}$ uniquely determines a G-invariant metric g on G/H, as required. □

Unsurprisingly the metric g in Proposition 7 coincides with the canonical metric on Grassmannian ($d = 1$) introduced in [20]. It also follows from Theorem 2 that, with this metric g, $\mathrm{Flag}(n_1, \dots, n_d; n)$ is not merely a Riemannian manifold but also a geodesic orbit space. In fact, g is the only choice of a metric that makes $\mathrm{Flag}(n_1, \dots, n_d; n)$ into a geodesic orbit space [5]. We will next derive explicit analytic expressions for geodesic (Propositions 8 and 9), arclength (Corollary 1), geodesic distance (Proposition 10), and parallel transport (Proposition 11).

Proposition 8 (Geodesic I) *Let* $[\![Q]\!] \in \mathrm{Flag}(n_1, \dots, n_d; n) = \mathrm{O}(n)/\big(\mathrm{O}(n_1) \times \cdots \times \mathrm{O}(n - n_d)\big)$ *and g be the metric in (15). Every geodesic on* $\mathrm{Flag}(n_1, \dots, n_d; n)$ *passing through* $[\![Q]\!]$ *takes the form*

$$[\![Q(t)]\!] = \left\{ Q\exp(tB)\begin{bmatrix} Q_1 & 0 & \cdots & 0\\ 0 & Q_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & Q_{d+1}\end{bmatrix} \in \mathrm{O}(n) : Q_i \in \mathrm{O}(n_i - n_{i-1}),\ i = 1, \dots, d+1 \right\},$$

for some direction

$$B = \begin{bmatrix} 0 & B_{1,2} & \cdots & B_{1,d+1}\\ -B_{1,2}^{\mathsf{T}} & 0 & \cdots & B_{2,d+1}\\ \vdots & \vdots & \ddots & \vdots\\ -B_{1,d+1}^{\mathsf{T}} & -B_{2,d+1}^{\mathsf{T}} & \cdots & 0\end{bmatrix} \in \mathbb{R}^{n\times n}, \quad \begin{matrix} B_{ij} \in \mathbb{R}^{(n_i - n_{i-1})\times(n_j - n_{j-1})},\\ 1 \le i < j \le d+1 \end{matrix}. \tag{17}$$

Proof Since $\mathrm{Flag}(n_1, \dots, n_d; n)$ with the metric g in Proposition 7 is a geodesic orbit space, the result follows immediately from Theorem 3. □

Corollary 1 (Arclength I) *The arclength of a geodesic* $\gamma(t) = [\![Q(t)]\!]$ *passing through* Q *in the direction* B *is given by*

$$\|\gamma(t)\| = t\Big[\sum\nolimits_{1\le i<j\le d+1} \operatorname{tr}(B_{ij}^{\mathsf{T}}B_{ij})\Big]^{1/2} = t\sqrt{\frac{\operatorname{tr}(B^{\mathsf{T}}B)}{2}},$$

where B is as in (17).

Proof This follows from the definition of arclength

$$\|\gamma(t)\| := \int_0^t \sqrt{g_{[\![Q]\!]}(\gamma'(x), \gamma'(x))}\, dx$$

and the expressions for g in (14) and (15). □

Proposition 9 (Geodesic II) *Let γ be a geodesic in* $\mathrm{Flag}(n_1, \dots, n_d; n) = \mathrm{O}(n)/\big(\mathrm{O}(n_1) \times \cdots \times \mathrm{O}(n - n_d)\big)$ *with $\gamma(0) = [\![Q]\!]$ for some $Q \in \mathrm{O}(n)$ and $\gamma'(0) = H \in T_{[\![Q]\!]}\mathrm{Flag}(n_1, \dots, n_d; n)$. Let $Q^\top H = VDV^\top$ with $V \in \mathrm{O}(n)$ and*

$$D = \operatorname{diag}\left(\begin{bmatrix} 0 & -\lambda_1 \\ \lambda_1 & 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 & -\lambda_r \\ \lambda_r & 0 \end{bmatrix}, 0_{n-2r}\right) \in \mathfrak{so}(n), \tag{18}$$

where $2r = \operatorname{rank}(Q^\top H)$ and $\lambda_1, \dots, \lambda_r$ are positive real numbers. Then $\gamma(t) = [\![U\Sigma(t)V^\top]\!]$ where $U = QV \in \mathrm{O}(n)$ and

$$\Sigma(t) = \operatorname{diag}\left(\begin{bmatrix} \cos t\lambda_1 & -\sin t\lambda_1 \\ \sin t\lambda_1 & \cos t\lambda_1 \end{bmatrix}, \dots, \begin{bmatrix} \cos t\lambda_r & -\sin t\lambda_r \\ \sin t\lambda_r & \cos t\lambda_r \end{bmatrix}, I_{n-2r}\right) \in \mathrm{O}(n). \tag{19}$$

Proof By Proposition 8, the geodesic γ takes the form $\gamma(t) = [\![Q\exp(tB)]\!]$ for some $B \in \mathfrak{so}(n)$ and $Q \in \mathrm{O}(n)$ representing $\gamma(0)$. Hence we have $H = \gamma'(0) = QB$ and $Q^\top H = B$. Since B is a skew-symmetric and thus a normal matrix, by the spectral theorem [7, Theorem 7.25], $B = VDV^\top$ for some $V \in \mathrm{O}(n)$ and D of the form in (18), with $2r = \operatorname{rank}(B) = \operatorname{rank}(Q^\top H)$ and $\lambda_1, \dots, \lambda_r$ are positive reals as they are singular values of B. Therefore,

$$Q\exp(tB) = U\Sigma(t)V^\top,$$

where $U = QV$ and $\Sigma(t)$ is as in (19). □

Proposition 10 (Geodesic distance) *The geodesic distance with respect to the metric g between $[\![P]\!], [\![Q]\!] \in \mathrm{Flag}(n_1, \dots, n_d; n) = \mathrm{O}(n)/\big(\mathrm{O}(n_1) \times \cdots \times \mathrm{O}(n - n_d)\big)$ is*

$$d([\![P]\!], [\![Q]\!]) = \sqrt{\sum\nolimits_{i=1}^r \lambda_i^2}, \tag{20}$$

where $\lambda_1, \dots, \lambda_r$ are positive real numbers such that $P^\top Q = V\Sigma V^\top$ with $V \in \mathrm{O}(n)$ and

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

Proof By Proposition 3(iii), we may regard $\operatorname{Flag}(n_1, \dots, n_d; n)$ as a closed, and therefore compact, submanifold of $\operatorname{Gr}(n_1, n) \times \cdots \times \operatorname{Gr}(n_d, n)$. By Theorem 4, there is a distance minimizing geodesic $[\![P \exp(tB)]\!]$ connecting $[\![P]\!]$ and $[\![Q]\!]$. By Corollary 1, we get (20) with $\lambda_1, \lambda_1, \dots, \lambda_r, \lambda_r$ the nonzero singular values of B. Lastly, by Proposition 9, we get the decomposition $P^{\mathsf{T}} Q = V \Sigma V^{\mathsf{T}}$ for some $V \in \mathrm{O}(n)$. □

Let $\mathfrak{m}$ be as in (13). For $B \in \mathfrak{m}$, we define a map

$$\varphi_B : \mathfrak{m} \to \mathfrak{m}, \quad X \mapsto \frac{1}{2}[B, X]_{\mathfrak{m}} := \frac{1}{2} \operatorname{proj}_{\mathfrak{m}}([B, X]), \tag{21}$$

where $\operatorname{proj}_{\mathfrak{m}} : \mathfrak{so}(n) \to \mathfrak{m}$ is the projection from $\mathfrak{so}(n) = \mathfrak{h} \oplus \mathfrak{m}$ to $\mathfrak{m}$. For example, if $d = 2$ and

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

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

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

Proposition 11 (Parallel transport I) *Let $B, X \in T_{[\![I]\!]} \operatorname{Flag}(n_1, \dots, n_d; n) \cong \mathfrak{m}$ and $[\![Q]\!] \in \operatorname{Flag}(n_1, \dots, n_d; n)$. The parallel transport of $QX \in T_{[\![Q]\!]} \operatorname{Flag}(n_1, \dots, n_d; n)$ along the geodesic $[\![Q \exp(tB)]\!]$ is*

$$X(t) = Q \exp(tB) e^{-\varphi_{tB}}(X), \tag{22}$$

where $e^{-\varphi_B} : \mathfrak{m} \to \mathfrak{m}$, for φ_B as in (21), is defined by

$$e^{-\varphi_B} = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \varphi_B^k. \tag{23}$$

Proof This follows from applying [47, Lemma 3.1] to $\operatorname{Flag}(n_1, \dots, n_d; n)$. □

For the $d = 1$ case, i.e., $\operatorname{Flag}(k; n) = \operatorname{Gr}(k, n)$, it is straightforward to verify that $[B, X]_{\mathfrak{m}} = 0$ for all $B, X \in \mathfrak{m}$. So the expression for parallel transport in (22) reduces to $X(t) = Q \exp(tB) X$, which is the well-known expression for parallel transport on the Grassmannian [20].

4.2 Stiefel coordinates for the flag manifold

We next discuss the characterization of a flag manifold as a quotient of the Stiefel manifold (9) and discuss its consequences. This characterization will give our coordinates of choice for use in our optimization algorithms (see Sect. 6).

Proposition 12 *Let* $0 < n_1 < \cdots < n_d < n$ *be* d *positive integers. The flag manifold* $\operatorname{Flag}(n_1, \ldots, n_d; n)$ *is diffeomorphic to the homogeneous space*

$$\mathrm{V}(n_d, n)/\big(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n_d - n_{d-1})\big) \tag{24}$$

where $\mathrm{V}(n_d, n)$ *is the Stiefel manifold of orthonormal* n_d*-frames in* $\mathbb{R}^n$ *as described in Sect. 3.*

Proof This follows from the standard characterization of $\mathrm{V}(n_d, n)$ is a homogeneous space of $O(n)$, $\mathrm{V}(n_d, n) \cong \mathrm{O}(n)/\,\mathrm{O}(n - n_d)$, together with (8). $\square$

For the rest of this article, we will regard the Stiefel manifold $\mathrm{V}(k, n)$ as the set of all $n \times k$ matrices whose column vectors are orthonormal. With this identification, Proposition 12 allows us to represent a flag $\{\mathbb{V}_i\}_{i=1}^d \in \operatorname{Flag}(n_1, \ldots, n_d; n)$ by a matrix $Y = [y_1, \ldots, y_{n_d}] \in \mathbb{R}^{n \times n_d}$ with orthonormal $y_1, \ldots, y_{n_d} \in \mathbb{R}^n$ and where the first n_i of them span the subspace $\mathbb{V}_i$, $i = 1, \ldots, d$. This representation is not unique but if $Y' \in \mathbb{R}^{n \times n_d}$ is another such matrix, then

$$Y' = Y \begin{bmatrix} Q_1 & 0 & \cdots & 0 \\ 0 & Q_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_d \end{bmatrix}, \quad Q_i \in \mathrm{O}(n_i - n_{i-1}), \quad i = 1, \ldots, d. \tag{25}$$

Hence $\{\mathbb{V}_i\}_{i=1}^d \in \operatorname{Flag}(n_1, \ldots, n_d; n)$ may be represented by the equivalence class of matrices

$$[\![Y]\!] = \left\{ Y \begin{bmatrix} Q_1 & 0 & \cdots & 0 \\ 0 & Q_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_d \end{bmatrix} \in \mathbb{R}^{n \times n_d} : \begin{array}{r} Y \in \mathrm{V}(n_d, n), \ \operatorname{span}\{y_1, \ldots, y_{n_i}\} = \mathbb{V}_i, \\ Q_i \in \mathrm{O}(n_i - n_{i-1}), \ i = 1, \ldots, d \end{array} \right\}. \tag{26}$$

We will call such a representation *Stiefel coordinates* for the flag manifold.

In the following, for any $k < n$, we write

$$I_{n,k} := \begin{bmatrix} I_k \\ 0 \end{bmatrix} \in \mathbb{R}^{n \times k},$$

i.e., the $n \times k$ matrix comprising the first k columns of the $n \times n$ identity matrix I_n. Thus for any $A = [a_1, \ldots, a_n] \in \mathbb{R}^{n \times n}$, $AI_{n,k} = [a_1, \ldots, a_k] \in \mathbb{R}^{n \times k}$ gives us the first k columns of A.

For a flag $\{\mathbb{V}_i\}_{i=1}^d$, it is easy to convert between its orthogonal coordinates, i.e., $[\![Q]\!]$ in (11) with $Q \in \mathrm{O}(n)$, and its Stiefel coordinates, i.e., $[\![Y]\!]$ in (26) with $Y \in \mathrm{V}(n_d, n)$. Given $Q \in \mathrm{O}(n)$, one just takes its first n_d columns to get $Y = QI_{n,n_d}$; note that QI_{n,n_i} is automatically an orthonormal basis for the subspace $\mathbb{V}_i$, $i = 1, \ldots, d$. Given $Y \in \mathrm{V}(n_d, n)$, take any orthonormal basis $Y^{\perp} \in \mathrm{V}(n - n_d, n)$ of the orthogonal complement of $\operatorname{im}(Y)$ to get $Q = [Y, Y^{\perp}] \in \mathrm{O}(n)$.

We now derive expressions for tangent space, metric, arclength, geodesic, and parallel transport in Stiefel coordinates.

Proposition 13 (Tangent space II) *Let* $[\![Y]\!] \in \operatorname{Flag}(n_1, \dots, n_d; n) = \operatorname{V}(n_d, n)/\big(\operatorname{O}(n_1) \times \operatorname{O}(n_2 - n_1) \times \cdots \times \operatorname{O}(n_d - n_{d-1})\big)$ *be represented by* $Y \in \operatorname{V}(n_d, n)$. *Its tangent space at* $[\![Y]\!]$ *is given by*

$$T_{[\![Y]\!]} \operatorname{Flag}(n_1, \dots, n_d; n) = \{[Y, Y^{\perp}] B I_{n,n_d} \in \mathbb{R}^{n \times n_d} : B \in \mathfrak{m}\},$$

where $Y^{\perp} \in \operatorname{V}(n - n_d, n)$ *is such that* $[Y, Y^{\perp}] \in \operatorname{O}(n)$ *and* $\mathfrak{m}$ *is as in (13).*

Proof The proof essentially follows from differentiating a curve $\tau(t)$ in $\operatorname{Flag}(n_1, \dots, n_d; n)$ with $\tau(0) = [\![Y]\!]$ and noting that the tangent vector $\tau'(0)$ is perpendicular to $\mathfrak{h}$ in (12), whose orthogonal complement is precisely $\mathfrak{m}$. □

The description of $T_{[\![Y]\!]} \operatorname{Flag}(n_1, \dots, n_d; n)$ in Proposition 13 is a parametric one (like the description of the unit circle as $\{(\cos\theta, \sin\theta) : \theta \in [0, 2\pi)\}$). We may also derive an implicit description of $T_{[\![Y]\!]} \operatorname{Flag}(n_1, \dots, n_d; n)$ (like the description of the unit circle as $\{(x, y) : x^2 + y^2 = 1\}$).

Corollary 2 (Tangent space III) *Let* $[\![Y]\!] \in \operatorname{Flag}(n_1, \dots, n_d; n) = \operatorname{V}(n_d, n)/\big(\operatorname{O}(n_1) \times \operatorname{O}(n_2 - n_1) \times \cdots \times \operatorname{O}(n_d - n_{d-1})\big)$ *be represented by* $Y \in \operatorname{V}(n_d, n)$. *Let* Y *be partition as*

$$Y = [Y_1, \dots, Y_d], \quad Y_i \in \operatorname{V}(n_i - n_{i-1}, n), \quad i = 1, \dots, d.$$

Then its tangent space at $[\![Y]\!]$ *is given by*

$$\begin{aligned} &T_{[\![Y]\!]} \operatorname{Flag}(n_1, \dots, n_d; n) = \{[X_1, \dots, X_d] \in \mathbb{R}^{n \times n_d} : X_i \in \mathbb{R}^{n \times (n_i - n_{i-1})}, \\ &Y_i^{\mathsf{T}} X_j + X_i^{\mathsf{T}} Y_j = 0,\ Y_i^{\mathsf{T}} X_i = 0,\ 1 \le i, j \le d\}. \end{aligned} \tag{27}$$

Equivalently, the matrix $[X_1, \dots, X_d]$ *can be expressed as*

$$[X_1, \cdots, X_d] = [Y_1, \cdots, Y_d, Y^{\perp}] \begin{bmatrix} 0 & B_{1,2} & \cdots & B_{1,d} \\ -B_{1,2}^{\mathsf{T}} & 0 & \cdots & B_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ -B_{1,d}^{\mathsf{T}} & -B_{2,d}^{\mathsf{T}} & \cdots & 0 \\ -B_{1,d+1}^{\mathsf{T}} & -B_{2,d+1}^{\mathsf{T}} & \cdots & -B_{d,d+1}^{\mathsf{T}} \end{bmatrix},$$

where $Y^{\perp} \in \operatorname{V}(n - n_d, n)$ *is such that* $[Y, Y^{\perp}] \in \operatorname{O}(n)$ *and* $B_{ij} \in \mathbb{R}^{(n_i - n_{i-1}) \times (n_j - n_{j-1})}$, $1 \le i < j \le d + 1$.

Proof The calculation is straightforward and details may be found in [38]. Since $[Y_1, \dots, Y_d] \in \operatorname{V}(n_d, n)$ and $Y_i \in \operatorname{V}(n_i - n_{i-1}, n)$, the Y_i's are characterized by

$$Y_i^{\mathsf{T}} Y_i = I_{n_i - n_{i-1}}, \quad Y_i^{\mathsf{T}} Y_j = 0, \quad i \ne j = 1, \dots, d. \tag{28}$$

Differentiating (28) gives us the first relation in (27). On the other hand, by Proposition 13 we notice that a tangent vector in $T_{[\![Y]\!]} \operatorname{Flag}(n_1, \dots, n_d; n)$ is written as $[Y, Y^{\perp}] B I_{n,n_d}$ for some $B \in \mathfrak{m}$, from which we may easily verify the second relation in (27). □

The *implicit* characterization of tangent vectors in (27), which also appears in [38], is of course mathematically equivalent to the *explicit* characterizations in Propositions 6 and 13. Nevertheless, for the purpose of practical computations we require explicit expressions and consequently we do not use (27) anywhere in our algorithms.

Comparing Propositions 6 and 13, for a tangent vector $QB \in T_{[\![Q]\!]}\operatorname{Flag}(n_1, \dots, n_d; n)$ in orthogonal coordinates $Q \in \mathrm{O}(n)$, its corresponding tangent vector in Stiefel coordinates $Y = QI_{n,n_d} \in \mathrm{V}(n_d, n)$ is simply given by $QBI_{n,n_d} \in T_{[\![Y]\!]}\operatorname{Flag}(n_1, \dots, n_d; n)$. Conversely, $[Y, Y^{\perp}]BI_{n,n_d} \in T_{[\![Y]\!]}\operatorname{Flag}(n_1, \dots, n_d; n)$ in Stiefel coordinates corresponds to $QB \in T_{[\![Q]\!]}\operatorname{Flag}(n_1, \dots, n_d; n)$ in orthogonal coordinates where $Q = [Y, Y^{\perp}]$. Note that from the matrix BI_{n,n_d}, i.e., just the first n_d columns of $B \in \mathfrak{m}$, the full matrix B can be easily and uniquely recovered by its skew symmetry.

The straightforward translation between orthogonal and Stiefel coordinate representations of points and tangent vectors on a flag manifold allows us to immediately deduce analogues of Propositions 7, 8, 11, and Corollary 1.

Proposition 14 (Riemannian metric II) *The metric g at a point $[\![Y]\!] \in \operatorname{Flag}(n_1, \dots, n_d; n) = \mathrm{V}(n_d, n)/\big(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n_d - n_{d-1})\big)$ is given by*

$$g_{[\![Y]\!]}(W, Z) = \sum_{1 \le i < j \le d+1} \operatorname{tr}(B_{ij}^{\mathsf{T}} C_{ij}), \tag{29}$$

where $W, Z \in T_{[\![Y]\!]}\operatorname{Flag}(n_1, \dots, n_d; n)$ are

$$W = [Y, Y^{\perp}]\begin{bmatrix} 0 & B_{1,2} & \cdots & B_{1,d} \\ -B_{1,2}^{\mathsf{T}} & 0 & \cdots & B_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ -B_{1,d}^{\mathsf{T}} & -B_{2,d}^{\mathsf{T}} & \cdots & 0 \\ -B_{1,d+1}^{\mathsf{T}} & -B_{2,d+1}^{\mathsf{T}} & \cdots & -B_{d,d+1}^{\mathsf{T}} \end{bmatrix},$$

$$Z = [Y, Y^{\perp}]\begin{bmatrix} 0 & C_{1,2} & \cdots & C_{1,d} \\ -C_{1,2}^{\mathsf{T}} & 0 & \cdots & C_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ -C_{1,d}^{\mathsf{T}} & -C_{2,d}^{\mathsf{T}} & \cdots & 0 \\ -C_{1,d+1}^{\mathsf{T}} & -C_{2,d+1}^{\mathsf{T}} & \cdots & -C_{d,d+1}^{\mathsf{T}} \end{bmatrix} \in \mathbb{R}^{n \times n_d}.$$

Proposition 15 (Arclength II, Geodesics III) *Let $[\![Y]\!] \in \operatorname{Flag}(n_1, \dots, n_d; n) = \mathrm{V}(n_d, n)/\big(\mathrm{O}(n_1) \times \mathrm{O}(n_2 - n_1) \times \cdots \times \mathrm{O}(n_d - n_{d-1})\big)$ and g be the metric in (29). Every geodesic γ on $\operatorname{Flag}(n_1, \dots, n_d; n)$ passing through $[\![Y]\!]$ takes the form*

$$\gamma(t) = [\![Y(t)]\!] = \left\{ [Y, Y^{\perp}]\exp(tB)\begin{bmatrix} Q_1 & 0 & \cdots & 0 \\ 0 & Q_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_d \\ 0 & 0 & \cdots & 0 \end{bmatrix} \in \mathrm{V}(n_d, n) : \begin{array}{r} Q_i \in \mathrm{O}(n_i - n_{i-1}), \\ i = 1, \dots, d \end{array} \right\},$$

where $[Y, Y^{\perp}] \in \mathrm{O}(n)$ and $B \in \mathfrak{m}$. In particular, the arclength of $\gamma(t)$ is

$$\|\gamma(t)\| = t\Big[\sum\nolimits_{1 \le i < j \le d+1} \operatorname{tr}(B_{ij}^{\mathsf{T}} B_{ij})\Big]^{1/2}.$$

Proposition 16 (Parallel transport II) *Let* $[\![Y]\!] \in \operatorname{Flag}(n_1, \dots, n_d; n)$ *and*

$$[Y, Y^{\perp}]BI_{n,n_d},\ [Y, Y^{\perp}]XI_{n,n_d} \in T_{[\![Y]\!]} \operatorname{Flag}(n_1, \dots, n_d; n).$$

The parallel transport of $[Y, Y^{\perp}]XI_{n,n_d}$ *along the geodesic* $[\![[Y, Y^{\perp}]\exp(tB)I_{n,n_d}]\!]$ *is given by*

$$X(t) = [Y, Y^{\perp}]\exp(tB)e^{-\varphi_{tB}}(X)I_{n,n_d}, \tag{30}$$

with $e^{-\varphi_{tB}}$ *defined as in (21) and (23).*

While it is also straightforward to obtain analogues of Propositions 9 and 10 in Stiefel coordinates, we omit them as the expressions are more involved and we will not need them in the sequel.

5 Flag manifolds as matrix manifolds

By Proposition 3(iii) and (iv), we see that a flag manifold may be regarded as a submanifold of a product of Grassmannians. Since a Grassmannian can be represented as a subset of matrices in $\mathbb{R}^{n\times n}$ [37, Example 1.2.20],

$$\operatorname{Gr}(k, n) \cong \{P \in \mathbb{R}^{n\times n} : P^2 = P = P^{\mathsf{T}},\ \operatorname{tr}(P) = k\}, \tag{31}$$

so can a flag manifold; and we will discuss two different ways do this, corresponding to (iii) and (iv) in Proposition 3:

$$\begin{aligned}
&\operatorname{Flag}(n_1, \dots, n_d; n) \subseteq \operatorname{Gr}(n_1, n) \times \operatorname{Gr}(n_2, n) \times \cdots \times \operatorname{Gr}(n_d, n),\\
&\operatorname{Flag}(n_1, \dots, n_d; n) \subseteq \operatorname{Gr}(n_1, n) \times \operatorname{Gr}(n_2 - n_1, n) \times \cdots \times \operatorname{Gr}(n_d - n_{d-1}, n).
\end{aligned}$$

The correspondence in (31) is given by a map that takes a k-dimensional subspace $\mathbb{W} \in \operatorname{Gr}(k, n)$ to its orthogonal projector,

$$\varepsilon : \operatorname{Gr}(k, n) \to \mathbb{R}^{n\times n}, \quad \mathbb{W} \mapsto WW^{\mathsf{T}}, \tag{32}$$

where $W \in \mathbb{R}^{n\times k}$ is any orthonormal basis of $\mathbb{W}$. Note that if W' is another such $n \times k$ matrix, then $W' = WQ$ for some $Q \in \mathrm{O}(k)$ and so $W'W'^{\mathsf{T}} = WW^{\mathsf{T}}$ and the map ε is well-defined. It is also injective and its image is precisely the set on the right of (31).

5.1 Projection coordinates for the flag manifold

We will construct our first analogue of (32) for the flag manifold. Let

$$\varepsilon : \operatorname{Flag}(n_1, \dots, n_d; n) \to \mathbb{R}^{nd\times nd}, \quad \{\mathbb{V}_i\}_{i=1}^d \mapsto \operatorname{diag}(V_1V_1^{\mathsf{T}}, \dots, V_dV_d^{\mathsf{T}}), \tag{33}$$

where $V_i \in \mathbb{R}^{n\times n_i}$ is an orthonormal basis of $\mathbb{V}_i$, $i = 1, \dots, d$, and the image is a block-diagonal matrix in $\mathbb{R}^{nd\times nd}$ with d blocks $V_1V_1^{\mathsf{T}}, \dots, V_dV_d^{\mathsf{T}} \in \mathbb{R}^{n\times n}$. In fact, the map in (33) is essentially the map in (6) that we used to establish Proposition 3(iii) except that we identify the Grassmannians with sets of projection matrices as in (31).

Proposition 17 *The flag manifold* $\operatorname{Flag}(n_1, \dots, n_d; n)$ *is diffeomorphic to*

$$\{P = \operatorname{diag}(P_1, \dots, P_d) \in \mathbb{R}^{nd \times nd} : P_i^2 = P_i = P_i^{\mathsf{T}},\ \operatorname{tr}(P_i) = n_i,\ P_j P_i = P_i,\ i < j\}. \tag{34}$$

Proof One may check that ε in (33) has its image contained in the set (34); and the map that takes $P = \operatorname{diag}(P_1, \dots, P_d)$ to the flag $\{\operatorname{im}(P_i)\}_{i=1}^d \in \operatorname{Flag}(n_1, \dots, n_d; n)$ is its inverse. □

We will call the representation in Proposition 17 *projection coordinates* for the flag manifold. Unlike the orthogonal and Stiefel coordinates introduced earlier, which are not unique, projection coordinates are unique. Let $\{\mathbb{V}_i\}_{i=1}^d \in \operatorname{Flag}(n_1, \dots, n_d; n)$ with

(a) orthogonal coordinates $[\![Q]\!]$ for some $Q \in \mathrm{O}(n)$;
(b) Stiefel coordinates $[\![Y]\!]$ for some $Y \in \mathrm{V}(n_d, n)$;
(c) projection coordinates P as in (34).

We have seen how we may easily convert between orthogonal and Stiefel coordinates after (26), we now see how they may be interchanged with projection coordinates just as easily:

(a)→(c): Given $Q = [q_1, \dots, q_n] \in \mathrm{O}(n)$, let $Q_i = [q_1, \dots, q_{n_i}] \in \mathrm{V}(n_i, n)$; then $P_i = Q_i Q_i^{\mathsf{T}}$, $i = 1, \dots, d$.
(b)→(c): Given $Y = [y_1, \dots, y_{n_d}] \in \mathrm{V}(n_d, n)$, let $Y_i = [y_1, \dots, y_{n_i}] \in \mathrm{V}(n_i, n)$; then $P_i = Y_i Y_i^{\mathsf{T}}$, $i = 1, \dots, d$.
(c)→(b): Given $P = \operatorname{diag}(P_1, \dots, P_d)$, let $y_1, \dots, y_{n_i}$ be an orthonormal basis of $\operatorname{im}(P_i)$; then $Y_i = [y_1, \dots, y_{n_i}] \in \mathrm{V}(n_i, n)$, $i = 1, \dots, d$.
(c)→(a): As above but appending an orthonormal basis $y_{n_d+1}, \dots, y_n$ of $\operatorname{im}(P_d)^{\perp}$ gives us $Q = [y_1, \dots, y_{n_d}, y_{n_d+1}, \dots, y_n] \in \mathrm{O}(n)$.

As is the case for the Grassmannian, the flag manifold has several extrinsic coordinates systems with which differential geometric objects and operations have closed-form analytic expressions and where one coordinate representation can be transformed to another with relative ease. This flexibility to switch between coordinate systems can be exploited in computations but as we will see next, it can also be exploited in deriving the requisite analytic expressions.

Proposition 18 (Tangent spaces IV) *Let* $P = \operatorname{diag}(P_1, \dots, P_d) \in \operatorname{Flag}(n_1, \dots, n_d; n)$ *as represented in (34). Then the tangent space is given by*

$$\begin{aligned} T_P \operatorname{Flag}(n_1, \dots, n_d; n) = \{Z = \operatorname{diag}(Z_1, \dots, Z_d) \in \mathbb{R}^{nd \times nd} &: Z_i P_i + P_i Z_i = Z_i = Z_i^{\mathsf{T}}, \\ \operatorname{tr}(Z_i) = 0,\ Z_j P_i + P_j Z_i = Z_i,\ i < j,\ i, j = 1, \dots, d.\} \end{aligned} \tag{35}$$

Proof Let $\gamma(t)$ be a curve in $\operatorname{Flag}(n_1, \dots, n_d; n)$ as characterized by (34), i.e., $\gamma : (-1, 1) \to \mathbb{R}^{nd \times nd}$, $t \mapsto \operatorname{diag}\bigl(P_1(t), \dots, P_d(t)\bigr)$ where

$$\begin{aligned} P_i(t)^2 = P_i(t),\ P_i(t)^{\mathsf{T}} = P_i(t),\ \operatorname{tr}(P_i(t)) = n_i,\ P_j(t) P_i(t) \\ = P_i(t),\ i < j,\ i, j = 1, \dots, d, \end{aligned} \tag{36}$$

for all $t \in (-1, 1)$. Taking derivatives of these relations at $t = 0$ gives the required description. □

Again the ease of translation from orthogonal and Stiefel coordinates to projection coordinates yields counterparts of Proposition 7–11 readily. We will just provide expressions for geodesic and parallel transport as examples.

Proposition 19 (Geodesics IV) *Let $P = \operatorname{diag}(P_1, \dots, P_d) \in \operatorname{Flag}(n_1, \dots, n_d; n)$ be as represented in (34) and $Z = \operatorname{diag}(Z_1, \dots, Z_d) \in T_P \operatorname{Flag}(n_1, \dots, n_d; n)$ be as represented in (35). Then there exist $Y \in \mathrm{V}(n_d, n)$ and skew-symmetric $B \in \mathbb{R}^{n\times n}$ such that for $Y_i = Y I_{n,n_i}$, $B_i = B I_{n,n_i} \in \mathbb{R}^{n\times n_i}$,*

$$P_i = Y_i Y_i^{\mathsf{T}}, \quad Z_i = Y_i B_i^{\mathsf{T}} + B_i Y_i^{\mathsf{T}}, \quad i = 1, \dots, d; \tag{37}$$

and a geodesic $P(t)$ passing through P in the direction Z takes the form

$$\{\operatorname{diag}\bigl(P_1(t), \dots, P_d(t)\bigr) \in \mathbb{R}^{nd\times nd} : P_i(t) = Y_i(t) Y_i(t)^{\mathsf{T}},\ Y_i(t) = [Y, Y^{\perp}] \exp(tB) I_{n,n_i}\}. \tag{38}$$

Proof The matrix Y is just P in Stiefel coordinates and may be obtained from (c)→(b) above. By Proposition 15, in Stiefel coordinates, the geodesic through $[\![Y]\!]$ in direction $[Y, Y^{\perp}] B I_{n,n_d}$ is

$$[\![[Y, Y^{\perp}] \exp(tB) I_{n,n_d}]\!].$$

By (b)→(c), Y and P are related by $P_i = Y_i Y_i^{\mathsf{T}}$, $i = 1, \dots, d$, which upon differentiation gives $Z_i = Y_i B_i^{\mathsf{T}} + B_i Y_i^{\mathsf{T}}$. The required expression (38) then follows. □

The observant reader might have noticed that B_{d+1} does not appear in (37)—the reason is that since B is skew-symmetric, B_{d+1} is uniquely determined by $B_1, \dots, B_d$.

Proposition 20 (Parallel transport III) *Let P, Z, Y, B, and $P(t)$ be as in Proposition 19. Let $Y^{\perp} \in \mathrm{V}(n - n_d, n)$ be such that $[Y, Y^{\perp}] \in \mathrm{O}(n)$ and set*

$$Y_i(t) = [Y, Y^{\perp}] \exp(tB) I_{n,n_i}, \quad X_i(t) = [Y, Y^{\perp}] \exp(tB) e^{-\varphi_{tB}}(X) I_{n,n_i}, \quad i = 1, \dots, d. \tag{39}$$

Then the parallel transport of the tangent vector Z along the geodesic $P(t)$ is given by

$$Z(t) = \operatorname{diag}\bigl(Z_1(t), \dots, Z_d(t)\bigr), \quad Z_i(t) = Y_i(t) X_i(t)^{\mathsf{T}} + X_i(t) Y_i(t)^{\mathsf{T}}, \quad i = 1, \dots, d. \tag{40}$$

Proof As in the proof of Proposition 19, we obtain the corresponding projection coordinates $P = \operatorname{diag}(P_1, \dots, P_d)$, $P_i = Y_i Y_i^{\mathsf{T}}$, $Y_i = Y I_{n,n_i}$, $i = 1, \dots, d$. Differentiating these relations give a tangent vector $Z = \operatorname{diag}(Z_1, \dots, Z_d) \in T_P \operatorname{Flag}(n_1, \dots, n_d; n)$ in projection coordinates as

$$Z_i = Y_i X_i^{\mathsf{T}} + X_i Y_i^{\mathsf{T}}, \quad i = 1, \dots, d,$$

where $X \in T_{[\![Y]\!]}\operatorname{Flag}(n_1, \ldots, n_d; n)$ is the expression of the same tangent vector in Stiefel coordinates as in Proposition 13 and $X_i = X I_{n,n_i}$, $i = 1, \ldots, d$. The required expressions (39) and (40) then follow from the expression (30) for parallel transport in terms of Stiefel coordinates Y. □

5.2 Reduced projection coordinates for the flag manifold

We discuss a variation of projection coordinates on flag manifolds based on Proposition 3(iv). As in Sect. 5.1, if we identify the Grassmannians as sets of projection matrices as in (31), then the map in (7) becomes

$$\varepsilon' : \operatorname{Flag}(n_1, \ldots, n_d; n) \to \mathbb{R}^{nd \times nd}, \quad \{\mathbb{V}_i\}_{i=1}^d \mapsto \operatorname{diag}(W_1 W_1^{\mathsf{T}}, \ldots, W_d W_d^{\mathsf{T}}), \quad (41)$$

where column vectors of $W_i \in \mathbb{R}^{n \times (n_i - n_{i-1})}$ form an orthonormal basis of $\mathbb{V}_{i-1}^{\perp}$, the orthogonal complement of $\mathbb{V}_{i-1}$ in $\mathbb{V}_i$, $i = 1, \ldots, d$. This gives us another description of $\operatorname{Flag}(n_1, \ldots, n_d; n)$ as a matrix manifold, an analogue of Proposition 17.

Proposition 21 *The flag manifold* $\operatorname{Flag}(n_1, \ldots, n_d; n)$ *is diffeomorphic to*

$$\{R = \operatorname{diag}(R_1, \ldots, R_d) \in \mathbb{R}^{nd \times nd} : R_i^2 = R_i = R_i^{\mathsf{T}},\ \operatorname{tr}(R_i) = n_i - n_{i-1},\ R_i R_j = 0,\ i < j\}. \quad (42)$$

We call the representation in Proposition 21 *reduced projection coordinates* on the flag manifold $\operatorname{Flag}(n_1, \ldots, n_d; n)$. Again, it is straightforward to translate between the other three coordinates and reduced projection coordinates. This readily yields expressions for metric, tangent space, geodesic, and parallel transport in reduced projection coordinates as before. We will state those for tangent space and metric as examples.

Proposition 22 (Tangent spaces V) *Let* $R = \operatorname{diag}(R_1, \ldots, R_d) \in \operatorname{Flag}(n_1, \ldots, n_d; n)$ *be as represented in (42). Then the tangent space is given by*

$$\begin{aligned} T_R \operatorname{Flag}(n_1, \ldots, n_d; n) = \{Z = \operatorname{diag}(Z_1, \ldots, Z_d) \in \mathbb{R}^{nd \times nd} : R_i Z_i + Z_i R_i = Z_i = Z_i^{\mathsf{T}},& \\ \operatorname{tr}(Z_i) = 0,\ Z_i R_j + R_i Z_j = 0,\ 1 \le i < j \le d\}.& \end{aligned} \quad (43)$$

Propositions 21 and 22 give an alternative way to obtain the metric g in Proposition 7. Let g_i be the standard metric on $\operatorname{Gr}(n_i - n_{i-1}, n)$, $i = 1, \ldots, d$. Then it is straightforward to verify that g is the pull-back of $\sum_{i=1}^d g_i$ via the embedding (7) in Proposition 3(iv). This also gives us an expression for the metric in terms of reduced projection coordinates.

Proposition 23 (Riemannian metric III) *Let* $R = \operatorname{diag}(R_1, \ldots, R_d) \in \operatorname{Flag}(n_1, \ldots, n_d; n)$ *be as in (42). Let* $W = \operatorname{diag}(W_1, \ldots, W_d)$, $Z = \operatorname{diag}(Z_1, \ldots, Z_d) \in$

$T_R \operatorname{Flag}(n_1, \dots, n_d; n)$ be as in (43). Then there exist $V_i, A_i, B_i \in \mathbb{R}^{n \times (n_i - n_{i-1})}$, $i = 1, \dots, d$, such that

$$\begin{aligned} V_i V_i^{\mathsf{T}} &= R_i, \; V_i^{\mathsf{T}} V_i = I_{n_i - n_{i-1}}, \; V_i A_i^{\mathsf{T}} + A_i V_i^{\mathsf{T}} = W_i, \\ V_i^{\mathsf{T}} A_i &= 0, \; V_i B_i^{\mathsf{T}} + B_i V_i^{\mathsf{T}} = Z_i, \; V_i^{\mathsf{T}} B_i = 0, \end{aligned} \tag{44}$$

and the metric g is given by

$$g_R(W, Z) = \sum_{i=1}^{d} \operatorname{tr}(A_i^{\mathsf{T}} B_i).$$

Proof As the Grassmannian is just a flag manifold with $d = 1$, all our earlier discussions about Stiefel and projection coordinates also apply to it. So for $W_i, Z_i \in T_{R_i} \operatorname{Gr}(n_i - n_{i-1}, n)$ in projection coordinates, there exist $V_i, A_i, B_i \in \mathbb{R}^{n \times (n_i - n_{i-1})}$ satisfying (44). The standard Riemannian metric g_i on $\operatorname{Gr}(n_i - n_{i-1}, n)$ at R_i is then given by $g_i(W_i, Z_i) = \operatorname{tr}(A_i^{\mathsf{T}} B_i)$ and thus we have

$$g_R(W, Z) = \sum_{i=1}^{d} g_i(W_i, Z_i) = \sum_{i=1}^{d} \operatorname{tr}(A_i^{\mathsf{T}} B_i).$$

□

6 Riemannian Gradient and Hessian over the flag manifold

We will derive expressions for the Riemannian gradient and Riemannian Hessian of a real-valued function on a flag manifold, the main ingredients of optimization algorithms. Although in principle we may use any of the four extrinsic coordinate systems introduced in the last two sections—orthogonal (as $n \times n$ orthogonal matrices), Stiefel (as $n \times n_d$ orthonormal matrices), projection or reduced projection (as d-tuples of $n \times n$ projection matrices) coordinates—Stiefel coordinates give the most economical representation and we will use this as our coordinates of choice. So in the following we will identify

$$\operatorname{Flag}(n_1, \dots, n_d; n) = \operatorname{V}(n_d, n) / \big(\operatorname{O}(n_1) \times \operatorname{O}(n_2 - n_1) \times \cdots \times \operatorname{O}(n_d - n_{d-1})\big). \tag{45}$$

Our expressions for gradient and Hessian in Stiefel coordinates may of course be converted to other coordinates—straightforward although the results may be notationally messy.

Proposition 24 (Riemannian gradient) *Let $f : \operatorname{Flag}(n_1, \dots, n_d; n) \to \mathbb{R}$ be a smooth function expressed in Stiefel coordinates $Y \in \operatorname{V}(n_d, n)$. Define the $n \times n_d$ matrix of partial derivatives,*

$$f_Y := \left[\frac{\partial f}{\partial y_{ij}} \right]_{i,j=1}^{n,n_d}. \tag{46}$$

Write $Y = [Y_1, \ldots, Y_d]$ *where* $Y_i \in \mathbb{R}^{n\times(n_i-n_{i-1})}$ *and* $f_Y = [f_{Y_1}, \ldots, f_{Y_d}]$ *where* f_{Y_i} *is the* $n \times (n_i - n_{i-1})$ *submatrix,* $i = 1, \ldots, d$. *Then its Riemannian gradient at* $[\![Y]\!] \in \mathrm{Flag}(n_1, \ldots, n_d; n)$ *is given by* $\nabla f([\![Y]\!]) = [\Delta_1, \ldots, \Delta_d]$ *where*

$$\Delta_i = f_{Y_i} - \Big(Y_i Y_i^{\mathsf{T}} f_{Y_i} + \sum\nolimits_{j\neq i} Y_j f_{Y_j}^{\mathsf{T}} Y_i\Big), \quad i = 1, \ldots, d. \tag{47}$$

Proof For any $X \in T_{[\![Y]\!]}\,\mathrm{Flag}(n_1, \ldots, n_d; n)$, let $X_a \in \mathbb{R}^{n\times n}$ be the unique skew-symmetric matrix such that $X = QX_a I_{n,n_d}$, where $Q \in \mathrm{O}(n)$ is such that $Y = QI_{n,n_d}$. Since the metric expressed in Stiefel coordinates (29) and expressed in orthogonal coordinates (14) must be equal,

$$g_{[\![Y]\!]}\big(\nabla f([\![Y]\!]), X\big) = g_{[\![Q]\!]}\big(\nabla f([\![Y]\!])_a, X_a\big) = \frac{1}{2}\operatorname{tr}(\nabla f([\![Y]\!])_a^{\mathsf{T}} X_a). \tag{48}$$

By definition of Riemannian gradient (5), we also have

$$g_{[\![Y]\!]}(\nabla f([\![Y]\!]), X) = \frac{1}{2}\operatorname{tr}((Q^{\mathsf{T}} f_Y)^{\mathsf{T}} X_a). \tag{49}$$

Comparing (48) and (49), we see that $\nabla f([\![Y]\!])$ is the projection of f_Y onto $T_{[\![Y]\!]}\,\mathrm{Flag}(n_1, \ldots, n_d; n)$, i.e., $f_Y = \nabla f([\![Y]\!]) + Z$ for some Z orthogonal to $T_{[\![Y]\!]}\,\mathrm{Flag}(n_1, \ldots, n_d; n)$. We may take $Z = [Z_1, \ldots, Z_d]$ to be

$$Z_i := Y_i Y_i^{\mathsf{T}} f_{Y_i} + \sum\nolimits_{j\neq i} Y_j f_{Y_j}^{\mathsf{T}} Y_i, \quad i = 1, \ldots, d,$$

and verify that because of (27), we indeed have $f_Y - Z \in T_{[\![Y]\!]}\,\mathrm{Flag}(n_1, \ldots, n_d; n)$ and thus Z is orthogonal to $T_{[\![Y]\!]}\,\mathrm{Flag}(n_1, \ldots, n_d; n)$. □

The Riemannian gradient ∇f may also be derived by solving an optimization problem as in [38]. Note that if $d = 1$, (47) becomes $\nabla f([\![Y]\!]) = \Delta = f_Y - YY^{\mathsf{T}} f_Y$, the well-known expression for Riemannian gradient of Grassmannian in [20].

Proposition 25 (Riemannian Hessian) *Let* $f : \mathrm{Flag}(n_1, \ldots, n_d; n) \to \mathbb{R}$ *be a smooth function expressed in Stiefel coordinates* $Y \in \mathrm{V}(n_d, n)$ *and let* f_Y *be as in (46). Then its Riemannian Hessian* $\nabla^2 f([\![Y]\!])$ *at* $[\![Y]\!] \in \mathrm{Flag}(n_1, \ldots, n_d; n)$ *is the symmetric bilinear form given by*

$$\nabla^2 f([\![Y]\!])(X, X') = f_{Y,Y}(X, X') - \frac{1}{2}\big[\operatorname{tr}(f_Y^{\mathsf{T}} QB^{\mathsf{T}} Q^{\mathsf{T}} X') + \operatorname{tr}(f_Y^{\mathsf{T}} QC^{\mathsf{T}} Q^{\mathsf{T}} X)\big], \tag{50}$$

for $X, X' \in T_{[\![Y]\!]}\,\mathrm{Flag}(n_1, \ldots, n_d; n)$*, where*

$$f_{Y,Y}(X, X') := \sum_{i,k=1}^{n} \sum_{j,l=1}^{n_d} \frac{\partial^2 f}{\partial y_{ij}\partial y_{kl}} x_{ij} x'_{kl}, \tag{51}$$

$Q \in \mathrm{O}(n)$ is such that $QI_{n,n_d} = Y$, and $B, C \in \mathbb{R}^{n\times n}$ are the unique skew-symmetric matrices such that $X = QBI_{n,n_d}$, $X' = QCI_{n,n_d}$ respectively.

Proof By Proposition 15, a geodesic γ with $\gamma'(0) = X$ and $\gamma(0) = [\![Y]\!]$ takes the form $\gamma(t) = [\![Q\exp(tB)I_{n,n_d}]\!]$ where $Q \in \mathrm{O}(n)$ is such that $Y = QI_{n,n_d}$ and $X = QBI_{n,n_d}$. Applying chain rule,

$$\frac{d}{dt} f(\gamma(t)) = \operatorname{tr}\big(f_Y^{\mathsf{T}}\gamma'(t)\big), \qquad \frac{d^2}{dt^2} f(\gamma(t)) = \operatorname{tr}\big(\gamma'(t)^{\mathsf{T}} f_{\gamma(t),\gamma(t)}^{\mathsf{T}}\gamma'(t)\big) + \operatorname{tr}\big(f_Y^{\mathsf{T}}\gamma''(t)\big);$$

followed by evaluating at $t = 0$ gives

$$\nabla^2 f([\![Y]\!])(X, X) = \frac{d^2}{dt^2} f(\gamma(t))\Big|_{t=0} = f_{Y,Y}(X, X) - \operatorname{tr}(f_Y^{\mathsf{T}} QB^{\mathsf{T}} Q^{\mathsf{T}} X).$$

The required expression (50) then follows from (5) and (4). ◻

If $d = 1$, (50) reduces to the well-known expression for Riemannian Hessian of the Grassmannian [20, Sect. 2.5.4] since

$$\begin{aligned}\nabla^2 f([\![Y]\!])(X, X') &= f_{Y,Y}(X, X') - \operatorname{tr}(f_Y^{\mathsf{T}} QB^{\mathsf{T}} Q^{\mathsf{T}} X') \\ &= f_{Y,Y}(X, X') - \operatorname{tr}(f_Y^{\mathsf{T}} Y(X')^{\mathsf{T}} X) = f_{Y,Y}(X, X') - \operatorname{tr}(X^{\mathsf{T}} X' Y^{\mathsf{T}} f_Y).\end{aligned}$$

There is slight inconsistency in our definitions of f_Y and f_{YY} to make these expressions easily portable into computer codes. To be consistent with (51), we could define f_Y as a linear form:

$$f_Y(X) = \sum_{i,j=1}^{n,n_d} \frac{\partial f}{\partial y_{ij}} x_{ij}$$

for $X \in T_{[\![Y]\!]}\operatorname{Flag}(n_1, \dots, n_d; n)$. Alternatively, to be consistent with (46), we could define f_{YY} to be a hypermatrix of partials (this is not a 4-tensor, just a convenient way to represent a 2-tensor):

$$f_{YY} = \left[\frac{\partial^2 f}{\partial y_{ij}\partial y_{kl}}\right]_{i,j,k,l=1}^{n,n_d,n,n_d}.$$

7 Optimization algorithms on flag manifolds

With analytic expressions for points, tangent vectors, metric, geodesics, parallel transports, Riemannian gradients and Hessians in place, Riemannian manifold optimization algorithms are straightforward to derive from the usual ones. For example, for steepest descent, instead of adding the negative of the gradient to the current iterate, we move the current iterate along the geodesic with initial velocity vector given by the negative of the gradient. Again, we may do this in any of the four coordinates system we have introduced although for the same reason in Sect. 6, we prefer the Stiefel coordinates. Thus here we will assume the identification (45) as before.

We note that standard local convergence results in nonlinear optimization on Euclidean space, e.g., those in [40], extend verbatim to Riemannian manifolds [22, Theorems 4.3–4.7]. The difficulty of optimization on Riemannian manifolds is not in establishing such convergence results, it is in getting closed-form computable expressions for the various differential geometric operations involved in the optimization algorithms.

The algorithms discussed below may be customized in many ways. For example, while we will state Algorithms 1 and 2 with exact line search, any reasonable strategy for choosing step size can be used in practice.

7.1 Steepest descent over a flag manifold

We describe this in Algorithm 1. A point $[\![Y]\!] \in \operatorname{Flag}(n_1, \dots, n_d; n)$ is represented in Stiefel coordinates, i.e., as a matrix $Y \in \mathbb{R}^{n \times n_d}$, $Y^{\mathsf{T}} Y = I$. As usual, $Y^{\perp} \in \mathbb{R}^{n \times (n - n_d)}$ is such that $X := [Y, Y^{\perp}] \in \mathrm{O}(n)$. The Riemannian gradient $\nabla f \in \mathbb{R}^{n \times n_d}$ is given by Proposition 24 and we set $G = -\nabla f$ to be the search direction. The exponential map direction $B \in \mathbb{R}^{n \times n}$ is uniquely obtained from $B I_{n,n_d} = [Y, Y^{\perp}]^{\mathsf{T}} G$, i.e., B is the unique skew-symmetric matrix whose first n_d columns is $[Y, Y^{\perp}]^{\mathsf{T}} G$. The next iterate is then found along the geodesic determined by the current iterate and the direction as in Proposition 15. Note in particular that Algorithm 1 does not involve parallel transport.

Algorithm 1 Steepest descent in Stiefel coordinates

Require: $[\![Y_0]\!] \in \operatorname{Flag}(n_1, \dots, n_d; n)$ with $Y_0 \in \mathbb{R}^{n \times n_d}$ and $Y_0^{\mathsf{T}} Y_0 = I$;
find $Y_0^{\perp} \in \mathbb{R}^{n \times (n - n_d)}$ such that $[Y_0, Y_0^{\perp}] \in \mathrm{O}(n)$;
set $X_0 = [Y_0, Y_0^{\perp}]$;
for $i = 0, 1, \dots$ **do**
  set $G_i = -\nabla f([\![Y_i]\!])$; ▷ gradient at $[\![Y_i]\!]$ as in (47)
  set $X_i = [Y_i, Y_i^{\perp}]$;
  compute $\widehat{B} = X_i^{\mathsf{T}} G_i$;
  set $B \in \mathbb{R}^{n \times n}$ as $B_{ij} = \widehat{B}_{ij}$ for $j \le n_d$;
    $B_{ij} = -\widehat{B}_{ji}$ for $j \ge n_d$ and $i \le n_d$;
    $B_{ij} = 0$ otherwise;
  minimize $f\big(X_i \exp(tB) I_{n,n_d}\big)$ over $t \in \mathbb{R}$; ▷ $t_{\min}$ from exact line search
  set $X_{i+1} = X_i \exp(t_{\min} B)$;
end for
Ensure: $[\![Y_{\mathrm{opt}}]\!] = [\![X_{\mathrm{opt}} I_{n,n_d}]\!]$

7.2 Conjugate gradient over a flag manifold

We present the conjugate gradient method in Algorithm 2. Unlike steepest descent, conjugate gradient requires that we construct our new descent direction from the

$(k-1)$th and kth iterates, i.e., one needs to compare tangent vectors at two different points on the manifold and the only way to do this is to parallel transport the two tangent vectors to the same point. There is no avoiding parallel transport in conjugate gradient.

As the expression for parallel transport in (30) indicates, we will need to compute

$$e^{-\varphi_{tB}}(X) = \sum_{k=1}^{\infty} \frac{(-1)^k}{k!} \varphi_{tB}^k(X), \quad \varphi_{tB}(X) = \frac{t}{2}[B, X]_{\mathfrak{m}}.$$

The rapid decay of the exponential series allows us to to replace it by a finite sum, reducing the task to recursively computing the iterated brackets and projection onto $\mathfrak{m}$:

$$\varphi_{tB}^k(X) = \varphi_{tB} \circ \cdots \circ \varphi_{tB}(X) = \left(\frac{t}{2}\right)^k [B, [B, \ldots, [B, X]_{\mathfrak{m}} \ldots]_{\mathfrak{m}}]_{\mathfrak{m}}.$$

As we had pointed out at the end of Sect. 4.1, this step is unnecessary for the Grassmannian as $[B, X]_{\mathfrak{m}} = 0$ if $d = 1$, i.e., for $\mathrm{Flag}(k; n) = \mathrm{Gr}(k, n)$. A careful treatment of the computation of $e^{-\varphi_{tB}}(X)$ requires more details than we could go into here and is deferred to [32].

Algorithm 2 Conjugate gradient in Stiefel coordinates

Require: $[\![Y_0]\!] \in \mathrm{Flag}(n_1, \ldots, n_d; n)$ with $Y_0 \in \mathbb{R}^{n \times n_d}$ and $Y_0^{\mathsf{T}} Y_0 = I$;
find $Y_0^{\perp} \in \mathbb{R}^{n \times (n-n_d)}$ such that $[Y_0, Y_0^{\perp}] \in \mathrm{O}(n)$;
set $X_0 = [Y_0, Y_0^{\perp}]$;
set $G_0 = -\nabla f([\![Y_0]\!])$ and $H_0 = -G_0$; ▷ gradient at $[\![Y_0]\!]$ as in (47)
for $i = 0, 1, \ldots$ **do**
  compute $\widehat{B} = X_i^{\mathsf{T}} H_i$;
  set $B \in \mathbb{R}^{n \times n}$ as $B_{ij} = \widehat{B}_{ij}$ for $j \le n_d$;
    $B_{ij} = -\widehat{B}_{ji}$ for $j \ge n_d$ and $i \le n_d$;
    $B_{ij} = 0$ otherwise;
  minimize $f\big(X_i \exp(tB) I_{n,n_d}\big)$ over $t \in \mathbb{R}$; ▷ $t_{\min}$ from exact line search
  set $X_{i+1} = X_i \exp(t_{\min} B)$;
  set $Y_{i+1} = X_{i+1} I_{n,n_d}$;
  set $Y_{i+1}^{\perp} = X_{i+1} I_{n,n-n_d}$;
  set $G_{i+1} = -\nabla f([\![Y_{i+1}]\!])$; ▷ gradient at $[\![Y_{i+1}]\!]$ as in (47)
  find $\widetilde{G}_{i+1} \in \mathbb{R}^{n \times (n-n_d)}$ such that $\widehat{G}_{i+1} = [G_{i+1}, \widetilde{G}_{i+1}]$ is skew-symmetric;
  procedure DESCENT($[\![Y_i]\!], [\![Y_{i+1}]\!], G_i, H_i$) ▷ new descent direction at $[\![Y_{i+1}]\!]$
    $\tau H_i = X_i \exp(t_{\min} B) e^{-\varphi_{t_{\min} B}}(B) I_{n,n_d}$; ▷ parallel transport of H_i as in (30)
    $\tau G_i = X_i \exp(t_{\min} B) e^{-\varphi_{t_{\min} B}}(\widehat{G}_i) I_{n,n_d}$; ▷ parallel transport of G_i as in (30)
    $\gamma_i = g_{[\![Y_{i+1}]\!]}(G_{i+1} - \tau G_i, G_{i+1}) / g_{[\![Y_i]\!]}(G_i, G_i)$; ▷ g as in (29)
    $H_{i+1} = -G_{i+1} + \gamma_i \tau H_i$;
  end procedure
  reset $H_{i+1} = -G_{i+1}$ if $i + 1 \equiv 0 \bmod (k+1)(n-k)$;
end for
Ensure: $[\![Y_{\mathrm{opt}}]\!] = [\![X_{\mathrm{opt}} I_{n,n_d}]\!]$

7.3 Newton and other algorithms over a flag manifold

The closed-form analytic expressions derived in this article permit one to readily extend other optimization algorithms on Euclidean spaces to flag manifolds. For example, the *Newton search direction* is given by the tangent vector $X \in T_{[\![Y]\!]} \operatorname{Flag}(n_1, \dots, n_d; n)$ such that

$$\nabla^2 f([\![Y]\!])(X, X') = g_{[\![Y]\!]}\bigl(-\nabla f([\![Y]\!]), X'\bigr),$$

for every $X' \in T_{[\![Y]\!]} \operatorname{Flag}(n_1, \dots, n_d; n)$, which gives us a system of linear equations upon plugging in the expressions for Riemannian gradient in (47) and Riemannian Hessian in (50).

Using the Newton search direction for G_i in Algorithm 1 then gives us Newton method on the flag manifold. In a similar vein, one may derive other standard algorithms for unconstrained optimization, e.g., quasi-Newton method, accelerated gradient descent, stochastic gradient descent, trust region methods, etc, for the flag manifold. Nevertheless, given that the goal of our article is to develop foundational material, we will leave these to future work [32].

8 Numerical experiments

We will test our algorithm for steepest descent on the flag manifold numerically. As we explained in Sect. 7.2, the experiments for conjugate gradient algorithm is more involved and is deferred to [32]. We run our numerical experiments on two problems: (i) the principal flag problem in Sect. 8.1 is one for which the solution may be determined in closed-form analytically, and thus it serves to demonstrate the correctness of our algorithm, i.e., converges to the true solution; (ii) a variation of the previous problem with a more complicated objective function to show that the convergence behavior remains unchanged. In addition, neither problem can be solved by simply treating them as nonlinear optimization problems with equality constraints and applying standard nonlinear optimization algorithms.

In the following we will assume the identification in (45) and use Stiefel coordinates throughout.

8.1 Principal flags

Let $M \in \mathbb{R}^{n \times n}$ be symmetric. We seek the solution to

$$\begin{array}{ll} \text{maximize} & \operatorname{tr}(Y^{\mathsf{T}} M Y) \\ \text{subject to} & [\![Y]\!] \in \operatorname{Flag}(n_1, \dots, n_d; n). \end{array} \tag{52}$$

Here $Y \in \mathbb{R}^{n \times n_d}$, $Y^{\mathsf{T}} Y = I_n$, and the objective function is well-defined as a function on the flag manifold: If we have Y and Y' with $[\![Y]\!] = [\![Y']\!]$, then they must be related as in (25) and thus $\operatorname{tr}(Y^{\mathsf{T}} M Y) = \operatorname{tr}(Y'^{\mathsf{T}} M Y')$.

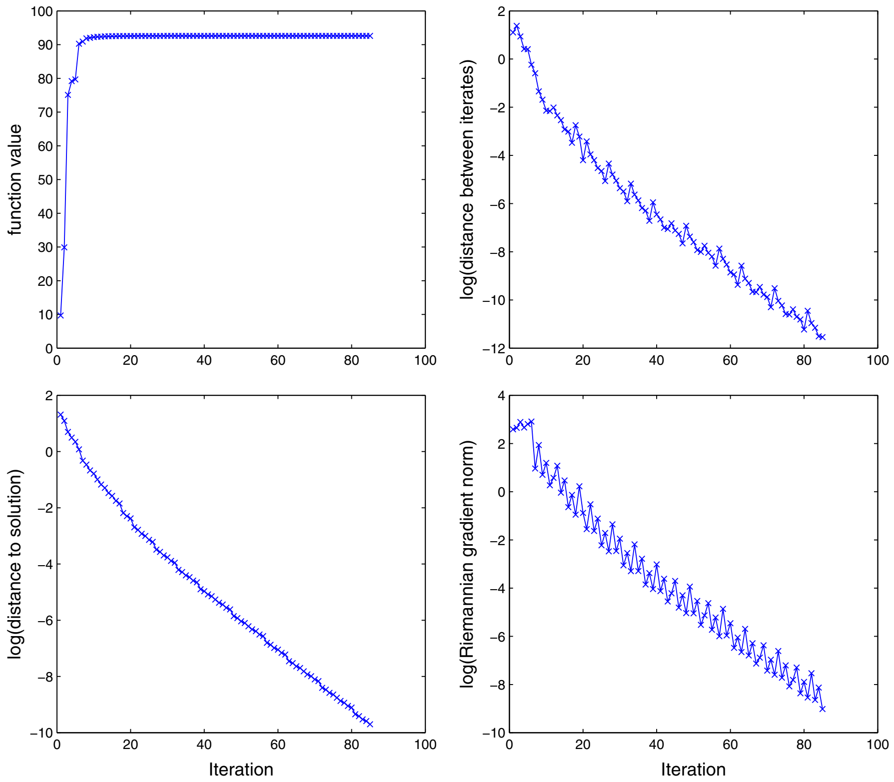

Fig. 1 Convergence trajectories for (52) on Flag(3, 7, 12; 60)

As we saw in Example 5, when M is a sample covariance matrix, the solution to (52) is equivalent to PCA when we seek a complete flag, i.e., $d = n - 1$ and $n_i = i$, $i = 1, \dots, n-1$. An advantage proffered by this approach is that if we do not know the intrinsic dimension of the data set a priori, then finding the flag as opposed to any particular subspace gives us the entire profile, showing how increasing dimension accounts for an increasingly amount of variance. The problem in (52) is thus a generalization of PCA, allowing us to seek any flag, not necessarily a complete one. It may also be interpreted as finding subspaces of dimensions $n_1, n_2 - n_1, \dots, n_d - n_{d-1}$ that are independent and explain different levels of variance in the data set.

Figure 1 shows the convergence trajectories of steepest descent, i.e., Algorithm 1, on the flag manifold Flag(3, 7, 12; 60), a 623-dimensional manifold. The symmetric matrix $M \in \mathbb{R}^{60\times 60}$ is generated randomly with standard normal entries. Since the true solution of (52) may be determined in closed form—it is the sum of the k largest eigenvalues of M—we may therefore conclude that Algorithm 1 converges to the true solution in around 80 iterations. Indeed the function values stabilize after as few as 10 iterations. At least for this problem, we see that the vanishing of the Riemannian gradient serves as a viable stopping condition. In our implementation, our stopping

Table 1 Distance to true solution for (52) on Flag(3, 9, 21; k)

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

Table 2 Elapsed time for (52) on Flag(3, 9, 21; k)

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

Table 3 Distance to true solution for (52) on Flag(2, ..., 2k; 60)

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

Table 4 Elapsed time for (52) on Flag(2, ..., 2k; 60)

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

conditions are determined by (i) Frobenius norm of Riemannian gradient, (ii) distance between successive iterates, and (iii) number of iterations.

We perform extensive experiments beyond that in Fig. 1 by taking average of 100 instances of the problem (52) for various values of $n_1, \dots, n_d$. We tabulate our results showing accuracy and speed in Tables 1–4. Tables 1 and 3 show that Algorithm 1 is robust across all dimensions of flags and ambient spaces that we tested. Tables 2 and 4 show that elapsed time taken for Algorithm 1 increases roughly linearly with the dimension of the flag manifold (Table 5).

We emphasize that the two problems below:

$$\begin{array}{ll} \text{maximize} & \operatorname{tr}(Y^{\mathsf{T}}MY) \\ \text{subject to} & [\![Y]\!] \in \operatorname{Flag}(3, 7, 12; 60), \end{array} \qquad \begin{array}{ll} \text{maximize} & \operatorname{tr}(Y^{\mathsf{T}}MY) \\ \text{subject to} & [\![Y]\!] \in \operatorname{Gr}(12, 60) \end{array} \tag{53}$$

will have entirely different optimizers even though $\operatorname{tr}(Y^{\mathsf{T}}MY)$ takes the same value on both $\operatorname{Gr}(12, 60)$ and Flag(3, 7, 12; 60) in Stiefel coordinates. The first thing to observe is that even if a point on Flag(3, 7, 12; 60) and a point on $\operatorname{Gr}(12, 60)$ are both represented by the same 60×12 orthonormal matrix, they will have entirely different geometric meanings—the former is a sequence of three nested subspaces of

Table 5 Distance to true solution for (52) on Flag(2, ..., 2k; 60)

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

$\mathbb{R}^{60}$ whereas the latter is merely a 12-dimensional subspace of $\mathbb{R}^{60}$. The flag manifold optimization problem captures this nested structure, which is much finer than the subspace structure captured by the Grassmannian optimization problem.

Any optimization algorithm that correctly models these structures will produce different results for the two problems in (53) because the initial point to the flag manifold optimization problem will have a nested subspace structure and any such algorithm will preserve this structure in its iterates; on the other hand an optimization algorithm for the Grassmannian neither requires its initial point to have a nested subspace structure nor preserve such a structure in its iterates. Mathematically, Flag(3, 7, 12; 60) is a fiber bundle over Gr(12, 60) with fiber Flag(3, 7; 12), giving us a submersion

$$\pi : \operatorname{Flag}(3, 7, 12; 60) \to \operatorname{Gr}(12, 60)$$

that collapses different equivalence classes in Flag(3, 7, 12; 60) into the same equivalence class in Gr(12, 60).

8.2 Nonlinear eigenflags

This is a variation of the principal flag problem (52):

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{d} \operatorname{tr}(Y_i^{\mathsf{T}} M Y_i)^2 \\ \text{subject to} & [\![Y_1, \dots, Y_d]\!] \in \operatorname{Flag}(n_1, \dots, n_d; n). \end{array} \tag{54}$$

Again $M \in \mathbb{R}^{n \times n}$ is a symmetric matrix and the flag is given in Stiefel coordinates $Y = [Y_1, \dots, Y_d] \in \mathbb{R}^{n \times n_d}$, $Y^{\mathsf{T}} Y = I$, but partitioned into submatrices $Y_i \in \mathbb{R}^{n \times (n_i - n_{i-1})}$, $Y_i^{\mathsf{T}} Y_i = I, i = 1, \dots, d$. More generally, the objective function in (54) may be replaced by $\sum_{i=1}^{d} f_i\bigl(\operatorname{tr}(Y_i^{\mathsf{T}} M Y_i)\bigr)$ with $f_1, \dots, f_d \in C^2(\mathbb{R})$. Choosing $f_1(x) = \cdots = f_d(x) = x$ gives us (52) and choosing $f_1(x) = \cdots = f_d(x) = x^2$ gives us (54). Note that it will take considerable effort to formulate a problem like (54) as a constrained optimization problem in Euclidean space.

The convergence trajectories for Algorithm 1 applied to (54) are shown in Fig. 2. The nonlinearity imposes a cost—it takes around 390 iterations to satisfy one of the our stopping criteria, although the function values stabilize after around 60 iterations. The jagged spikes seen in Fig. 2 are a result of iterates moving along a geodesic and then jumping to another geodesic. So this is indicative of steepest descent following a path that comprises multiple geodesics. A caveat is that unlike the principal flag problem (52), we do not have a closed-form solution for (54) and thus we may only guarantee convergence to a local minimizer, which is reflected in Fig. 2.

9 Conclusion

We end our article with a few parting thoughts and pointers for future work.

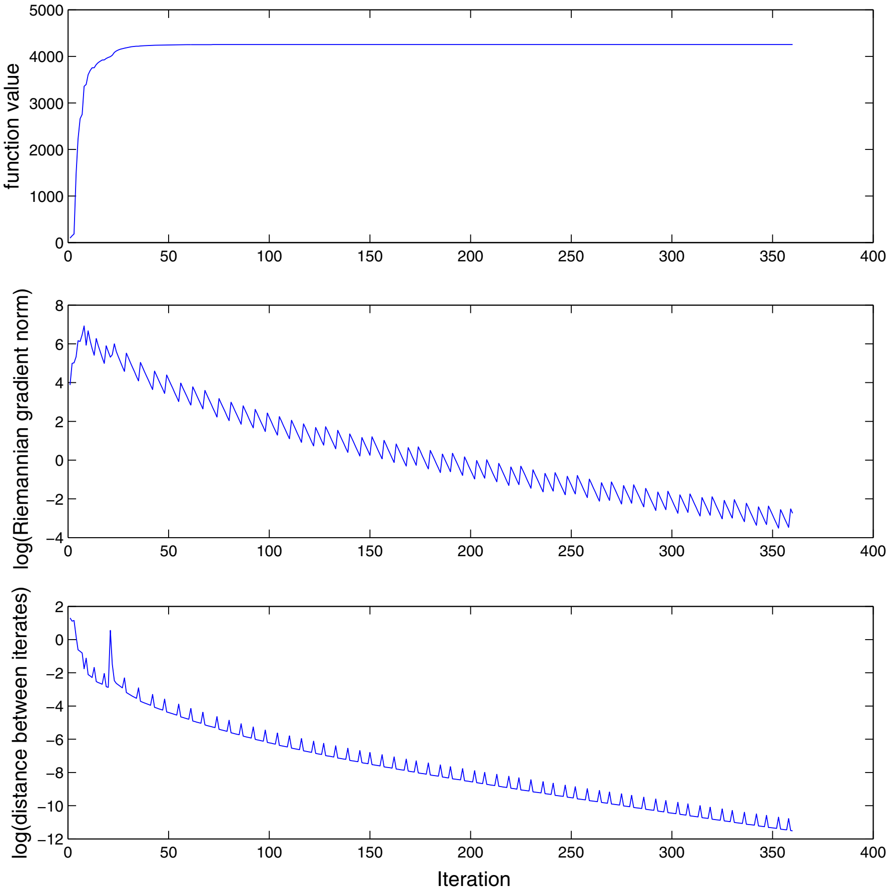

Fig. 2 Convergence trajectories for (54) on Flag(3, 7, 12; 60)

9.1 Coordinates on a flag manifold

We regard the four coordinate systems introduced in Sects. 4 and 5 as one of our main contributions and would like to add a few words about them.

Even in $\mathbb{R}^n$, one benefits from having a multitude of coordinate systems (polar, bipolar, cylindrical, spherical, parabolic, etc) other than the Cartesian one. Integrals difficult to evaluate in one coordinate system may be easily evaluated in another; differential equations impossible to solve in one coordinate system may be readily solved in another. In convex optimization, geometric programming problems are usually formulated in posynomial forms but algorithms are invariably applied to their convex forms [15]; interchanging between two coordinate systems $(y_1, \dots, y_n) = (e^{x_1}, \dots, e^{x_n})$ is not merely desirable but essential. In manifold optimization, algorithms for optimizing a function over a Grassmannian come in at least three different coordinate systems, where points are represented as equivalence classes of orthogonal matrices [20], as equivalence classes of full-rank matrices [3], or as projection matrices [26].

Even though we have only presented our algorithms in Sect. 7 in terms of Stiefel coordinates, we expect every coordinate system introduced in this article to be useful in its own way. For example, orthogonal coordinates allow us to represent flags as equivalence classes of matrices in $\mathrm{O}(n)$ or $\mathrm{SO}(n)$, which has a group structure not found in other coordinate systems, and in turn allow us to develop techniques (e.g., [27]) that we cannot easily do with other coordinate systems.

Projection and reduced projection coordinates allow us to represent flags as matrices as opposed to equivalence classes of matrices and this is useful if we want to define a probability density on the flag manifold—from an optimization perspective this is a first step towards probabilistic analysis of algorithms or randomized algorithms. Take the simplest flag manifold, i.e., a Grassmannian, for example, the *Langevin* or *von Mises–Fisher distribution* on $\mathrm{Gr}(k, n)$ [18] is given by the probability density function

$$f(P \mid S) := \frac{1}{{}_1F_1\big(\frac{1}{2}(k+1); \frac{1}{2}(n+1); S\big)} \exp\big(\operatorname{tr}(SP)\big),$$

where $S \in \mathbb{R}^{n\times n}$ is a symmetric matrix, ${}_1F_1$ is the confluent hypergeometric function of the first kind of a matrix argument, and $P \in \mathbb{R}^{n\times n}$ is a projection matrix as in (34) or (42) with $d = 1$. It is not clear how this can even be written down in the other two coordinate systems.

9.2 Optimization on a flag manifold

For most of its history, continuous optimization has been concerned with optimizing functions over the Euclidean space $\mathbb{R}^n$; but this has begun to change with the advent of semidefinite programming [9] and orthogonality-constrained optimization [20], where objective functions are naturally defined over the positive definite cone $\mathbb{S}^n_{++}$, the Stiefel manifold $\mathrm{V}(k, n)$, and the Grassmannian $\mathrm{Gr}(k, n)$. These developments have provided us with the capacity to optimize over not just vectors but also covariances, frames, and subspaces. The work in this article extends such capabilities to flags, which capture nested structures in multilevel, multiresolution, or multiscale phenomena.

In future work [32], we hope to compare the performance of algorithms in different coordinate systems on flag manifolds and investigate other computational issues that have been deferred from this first study.

Acknowledgements We thank the two anonymous referees for their exceptionally helpful suggestions and comments. KY is partially supported by National Key Research and Development Program of China No. 2018YFA0306702 and National Key Research and Development Program of China No. 2020YFA0712300, NSFC Grant No. 11801548 and NSFC Grant No. 11688101. LHL is supported by DARPA D15AP00109, HR00112190040, NSF IIS 1546413, DMS 1854831.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Absil, P.A., Amodei, L., Meyer, G.: Two Newton methods on the manifold of fixed-rank matrices endowed with Riemannian quotient geometries. Comput. Stat. **29**(3–4), 569–590 (2014)
2. Absil, P.A., Baker, C.G., Gallivan, K.A.: Trust-region methods on Riemannian manifolds. Found. Comput. Math. **7**(3), 303–330 (2007)
3. Absil, P.A., Mahony, R., Sepulchre, R.: Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. Acta Appl. Math. **80**(2), 199–220 (2004)
4. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton, NJ (2008)
5. Alekseevsky, D., Arvanitoyeorgos, A.: Riemannian flag manifolds with homogeneous geodesics. Trans. Am. Math. Soc. **359**(8), 3769–3789 (2007)
6. Ammar, G., Martin, C.: The geometry of matrix eigenvalue methods. Acta Appl. Math. **5**(3), 239–278 (1986)
7. Axler, S.: Linear Algebra Done Right. Undergraduate Texts in Mathematics, 3rd edn. Springer, Cham (2015)
8. Balzano, L., Nowak, R., Recht, B.: Online identification and tracking of subspaces from highly incomplete information. In: 2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 704–711. IEEE (2010)
9. Bellman, R., Fan, K.: On systems of linear inequalities in Hermitian matrix variables. In: Proceedings on Symposium Pure Mathematics, vol. VII, pp. 1–11. American Mathematical Society, Providence, RI (1963)
10. Boothby, W.M.: An Introduction to Differentiable Manifolds and Riemannian Geometry. Pure and Applied Mathematics, vol. 120, 2nd edn. Academic Press Inc, Orlando, FL (1986)
11. Borel, A.: La cohomologie mod 2 de certains espaces homogènes. Comment. Math. Helv. **27**, 165–197 (1953)
12. Borel, A.: Sur la cohomologie des espaces fibrés principaux et des espaces homogènes de groupes de Lie compacts. Ann. Math. **2**(57), 115–207 (1953)
13. Borel, A.: Linear Algebraic Groups. Graduate Texts in Mathematics, vol. 126, 2nd edn. Springer-Verlag, New York (1991)
14. Borel, A., Serre, J.P.: Groupes de Lie et puissances réduites de Steenrod. Am. J. Math. **75**, 409–448 (1953)
15. Boyd, S., Kim, S.J., Vandenberghe, L., Hassibi, A.: A tutorial on geometric programming. Optim. Eng. **8**(1), 67–127 (2007)
16. Cheeger, J., Ebin, D.G.: Comparison Theorems in Riemannian Geometry. AMS Chelsea Publishing, Providence, RI (2008)
17. Chern, S.S.: On the characteristic classes of complex sphere bundles and algebraic varieties. Am. J. Math. **75**, 565–597 (1953)
18. Chikuse, Y.: Statistics on Special Manifolds. Lecture Notes in Statistics, vol. 174. Springer-Verlag, New York (2003)
19. Curtef, O., Dirr, G., Helmke, U.: Riemannian optimization on tensor products of Grassmann manifolds: applications to generalized Rayleigh-quotients. SIAM J. Matrix Anal. Appl. **33**(1), 210–234 (2012)
20. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20**(2), 303–353 (1999)
21. Ehresmann, C.: Sur la topologie de certains espaces homogènes. Ann. Math. (2) **35**(2), 396–443 (1934)
22. Gabay, D.: Minimizing a differentiable function over a differential manifold. J. Optim. Theory Appl. **37**(2), 177–219 (1982)
23. Goodman, R., Wallach, N.R.: Symmetry, Representations, and Invariants, Graduate Texts in Mathematics, vol. 255. Springer, Dordrecht (2009)
24. Grassmann, H.: A New Branch of Mathematics. Open Court Publishing Co., Chicago, IL (1995)
25. Helgason, S.: Differential Geometry, Lie Groups, and Symmetric Spaces, Graduate Studies in Mathematics, vol. 34. American Mathematical Society, Providence, RI (2001)
26. Helmke, U., Hüper, K., Trumpf, J.: Newton's method on Graßmann manifolds. Preprint (2007). arxiv:0709.2205
27. Howe, R., Lee, S.T.: Spherical harmonics on Grassmannians. Colloq. Math. **118**(1), 349–364 (2010)

28. Jordan, J., Helmke, U.: Controllability of the QR-algorithm on Hessenberg flags. In: Proceeding of the Fifteenth International Symposium on Mathematical Theory of Network and Systems (MTNS 2002) (2002)
29. Kobayashi, S., Nomizu, K.: Foundations of Differential Geometry. Vols. I, II. Wiley Classics Library. Wiley, New York (1996)
30. Kowalski, O., Szenthe, J.: Erratum: "On the existence of homogeneous geodesics in homogeneous Riemannian manifolds" [Geom. Dedicata **81** (2000), no. 1-3, 209–214; MR1772203 (2001f:53104)]. Geom. Dedicata **84**(1–3), 331–332 (2001)
31. Lim, L.H., Wong, K.S.W., Ye, K.: Numerical algorithms on the affine Grassmannian. SIAM J. Matrix Anal. Appl. **40**(2), 371–393 (2019)
32. Lim, L.H., Ye, K.: Numerical algorithms on the flag manifold. Preprint (2019)
33. Lundström, E., Eldén, L.: Adaptive eigenvalue computations using Newton's method on the Grassmann manifold. SIAM J. Matrix Anal. Appl. **23**(3), 819–839 (2001/02)
34. Mardia, K.V., Kent, J.T., Bibby, J.M.: Multivariate Analysis. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Academic Press, London (1979)
35. Massart, E., Absil, P.A.: Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices. SIAM J. Matrix Anal. Appl. **41**(1), 171–198 (2020)
36. Monk, D.: The geometry of flag manifolds. Proc. Lond. Math. Soc. **3**(9), 253–286 (1959)
37. Nicolaescu, L.I.: Lectures on the Geometry of Manifolds. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ (2007)
38. Nishimori, Y., Akaho, S., Plumbley, M.D.: Riemannian optimization method on the flag manifold for independent subspace analysis. In: International Conference on Independent Component Analysis and Signal Separation, pp. 295–302. Springer (2006)
39. Nishimori, Y., Akaho, S., Plumbley, M.D.: Natural conjugate gradient on complex flag manifolds for complex independent subspace analysis. In: International Conference on Artificial Neural Networks, pp. 165–174. Springer (2008)
40. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer Series in Operations Research and Financial Engineering, 2nd edn. Springer, New York (2006)
41. O'Neill, B.: Semi-Riemannian Geometry, Pure and Applied Mathematics, vol. 103. Academic Press Inc, New York (1983)
42. Pennec, X.: Barycentric subspace analysis on manifolds. Ann. Stat. **46**(6A), 2711–2746 (2018)
43. Ring, W., Wirth, B.: Optimization methods on Riemannian manifolds and their application to shape space. SIAM J. Optim. **22**(2), 596–627 (2012)
44. Savas, B., Lim, L.H.: Quasi-Newton methods on Grassmannians and multilinear approximations of tensors. SIAM J. Sci. Comput. **32**(6), 3352–3393 (2010)
45. Schulz, V.H.: A Riemannian view on shape optimization. Found. Comput. Math. **14**(3), 483–501 (2014)
46. Stiefel, E.: Richtungsfelder und Fernparallelismus in n-dimensionalen Mannigfaltigkeiten. Comment. Math. Helv. **8**(1), 305–353 (1935)
47. Tojo, K.: Totally geodesic submanifolds of naturally reductive homogeneous spaces. Tsukuba J. Math. **20**(1), 181–190 (1996)
48. Vandereycken, B.: Low-rank matrix completion by Riemannian optimization. SIAM J. Optim. **23**(2), 1214–1236 (2013)
49. Wang, L., Wang, X., Feng, J.: Subspace distance analysis with application to adaptive Bayesian algorithm for face recognition. Pattern Recogn. **39**(3), 456–464 (2006)
50. Wang, R., Shan, S., Chen, X., Gao, W.: Manifold-manifold distance with application to face recognition based on image set. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, pp. 1–8. IEEE (2008)
51. Wen, Z., Yin, W.: A feasible method for optimization with orthogonality constraints. Math. Program. **142**(1–2, Ser. A), 397–434 (2013)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.