Operator-Valued Formulas for Riemannian Gradient and Hessian and Families of Tractable Metrics in Riemannian Optimization

Du Nguyen[1]

Abstract

We provide explicit formulas for the Levi-Civita connection and Riemannian Hessian for a Riemannian manifold that is a quotient of a manifold embedded in an inner product space with a non-constant metric function. Together with a classical formula for projection, this allows us to evaluate Riemannian gradient and Hessian for several families of metrics on classical manifolds, including a family of metrics on Stiefel manifolds connecting both the constant and canonical ambient metrics with closed-form geodesics. Using these formulas, we derive Riemannian optimization frameworks on quotients of Stiefel manifolds, including flag manifolds, and a new family of complete quotient metrics on the manifold of positive-semidefinite matrices of fixed rank, considered as a quotient of a product of Stiefel and positive-definite matrix manifold with affine-invariant metrics. The method is procedural, and in many instances, the Riemannian gradient and Hessian formulas could be derived by symbolic calculus. The method extends the list of potential metrics that could be used in manifold optimization and machine learning.

Keywords Optimization · Riemannian Hessian · Stiefel · Positive-definite ·
Positive-semidefinite · Flag manifold · Machine learning

Mathematics Subject Classification 65K10 · 58C05 · 49Q12 · 53C25 · 57Z20 ·
57Z25 · 68T05

✉ Du Nguyen
  nguyendu@post.harvard.edu

[1] Darien, CT, USA

1 Introduction

In the foundational paper [9], the authors introduced Riemannian optimization frameworks for Stiefel and Grassmann manifolds, where the Riemannian gradient and Hessian are given globally using a projection operator and a Christoffel function, a bilinear operator on the matrix space containing the Stiefel manifolds. The introduction of these global operators allows efficient computation when the manifolds are defined by constraints, or as quotients of constrained manifolds, as opposed to computation using local charts in an abstract manifold. The frameworks help popularize the field of Riemannian optimization. Since then, similar frameworks have been introduced for many manifolds with several software packages [8, 16, 26] implementing aspects of these frameworks. The Riemannian Hessian calculation is in general difficult. It could be computed using the calculus of variation. "Doing so is tedious" [9], so the detailed calculations for Stiefel manifolds were not included in that paper, a generalization and full derivation recently appeared in [11].

In this article, we attempt to address the problem: given a manifold, described by equality constraints in a vector space, or quotient of such manifold, and a metric, also defined by an analytic formula, compute the gradient and Riemannian Hessian for a function on the manifold. Recall the Riemannian Hessian is computed using the Levi-Civita connection. By computing, we mean a procedural, not necessarily closed-form approach. We are looking for a sequence of equations, operators, and expressions to solve and evaluate, rather than starting from a distance minimizing problem. We believe that the approach we take, using a classical formula for projections together with an adaptation of the Christoffel symbol calculation to ambient space addresses the problem effectively for many manifolds encountered in applications. This method provides a very explicit and transparent procedure that we hope will be helpful to researchers in the field. The main feature is it can handle manifolds with non-constant embedded metrics, such as Stiefel manifolds with the canonical metric, or the manifold of positive-definite matrices with the affine-invariant metrics. The method allows us to compute Riemannian gradients and Hessian for several new families of metrics on manifolds often encountered in applications, including optimization and machine learning. While the effect of changing metrics on first-order optimization methods has been considered previously, we hope this will lead to future works on adapting metrics to second-order methods. The approach is also suitable in the case where the gradient formula is not of closed-form. We also provide several useful identities known in special cases.

As an application of the method developed here, we give a short derivation of the gradient and Hessian of a family of metrics studied recently in [11], extending both the canonical and embedded metric on the *Stiefel manifolds with a closed-form geodesics formula*. (We were ignorance of [11], which were done before we started this project in 2020). We derive the Riemannian framework for the induced metrics on certain quotients of Stiefel manifolds, including *flag manifolds*. We also give *complete metrics on the fixed-rank positive-semidefinite matrix manifolds*, with efficiently computable gradient, Hessian, and geodesics.

When the metric on a submanifold of a vector space (called the ambient space) is induced from the Euclidean metric on the vector space, the Riemannian gradient

and Hessian correspond to the *projected gradient* and the *projected Hessian* in the literature of constrained optimization [9, section 4.9], [10]. When the metric is given by an analytic expression, (which is a metric when restricted to the constrained set, but not necessarily on the whole vector space), using [22, Lemma 4.3], we can compute Riemannian Hessian by projecting the Riemannian Hessian on the ambient space, and [22, Lemma 7.45] give us a similar computation for a quotient manifold. Representing the metric as an operator defined on the quotient/constrained manifold, we derive directly formulas for the Levi-Civita connection and the Riemannian Hessian that are dependent on the operator values on the manifold alone, giving us a procedural and simple approach, as explained. This approach has also been advocated in [1]. However, its usage is dependent on an extension with a computable Levi-Civita connection in the ambient space. This last step seems to restrict its use, a tricky extension is sometimes used, for example, in the case of the Grassmann manifold in [1]. We address this issue by working directly with the metric expression, giving an explicit formula for the connection in Eq. (3.10). This formula generalizes the connection formula using the *Weingarten map* in [2] for constant metrics.

Our framework works for both real and complex cases. The paper [15] gave the original treatment of the Hessian for the unitary/complex case. The case of Stiefel manifolds was studied in [9, 11], we reprove the results using our framework and provide complete formulas for the Hessian. Optimization on flag manifolds was studied in [21, 28] but a second-order method had not been given in an efficient form, an expression for the Riemannian Hessian vector product was not given/implemented. We provide second-order methods for the flag manifolds with the full family of metrics of [11], including efficient formulas for the Riemannian Hessian.

The affine-invariant metric on positive-definite matrices was also widely studied, for example in [12, 23–25]. There are numerous metrics on the fixed-rank positive-semidefinite (PSD) manifolds, which we mentioned [7] that motivated our approach. Although working with the same product of Stiefel and positive-definite manifolds with the affine-invariant metric, that paper did not use the Riemannian submersion metric on the quotient and focused on first-order methods. We compute the Levi-Civita connection for second-order methods. In [13, 27], two different families of metrics on PSD manifolds are studied. They both require solving Lyapunov equations but have different behaviors on the positive-definite part. Articles [15, 17] discuss the effect of adapting metrics to optimization problems ([17] adapts ambient metrics to the objective function using first-order methods.)

In the next section, we provide some background and summarize notations. In Sect. 3, we formulate and prove the main theoretical results of the paper. We then identify the adjoints of common operators on matrix spaces. We apply the theory developed to the manifolds discussed above. We then discuss numerical results and implementation. We conclude with a discussion of future directions.

2 Preliminaries

First-order approximation of a function f on $\mathbb{R}^n$ relies on the computation of the gradient and the second-order approximation relies on the Hessian matrix or Hessian-vector

product. When a function f is defined on a Riemannian manifold $\mathcal{M}$, the relevant quantities are the Riemannian gradient, which provides a first-order approximation for f, and Riemannian Hessian, providing the second-order term.

When a manifold $\mathcal{M}$ is embedded in an inner product space $\mathcal{E}$ with the inner product denoted by $\langle \cdots \rangle_{\mathcal{E}}$, if we have a function g from $\mathcal{M}$ with values in the set of positive-definite operators operating on $\mathcal{E}$, we can define an inner product of two vectors $\omega_1, \omega_2 \in \mathcal{E}$ by $\langle \omega_1, \mathsf{g}(Y)\omega_2 \rangle_{\mathcal{E}}$ for $Y \in \mathcal{M}$. This induces an inner product on each tangent space $T_Y\mathcal{M}$ and hence a Riemannian metric on $\mathcal{M}$, assuming sufficient smoothness. In this setup, if f is extended to a function $\hat{f}$ on an open neighborhood of $\mathcal{M} \subset \mathcal{E}$, then the Riemannian gradient relates to the gradient of $\hat{f}$ through a projection to $T_Y\mathcal{M}$. In the theory of generalized least squares (GLS) [4], it is well-known that a projection to the nullspace of a full-rank matrix J in an inner product space equipped with a metric g (also represented by a matrix) is given by the formula $I_{\mathcal{E}} - g^{-1}J^{\mathsf{T}}(Jg^{-1}J^{\mathsf{T}})^{-1}J$ ($I_{\mathcal{E}}$ is the identity matrix/operator of $\mathcal{E}$). If the tangent space is the nullspace of an operator J, and an operator g is used to describe the metric instead of matrices J and g, we have a similar formula where the transposed matrix J^{T} is replaced by the adjoint operator J^{t}. J is the Jacobian of a full rank equality constraint. This projection formula is not often used in the literature, the projection is usually derived directly by minimizing the distance to the tangent space. It turns out when J is given by a matrix equation, J^{t} is simple to compute. For manifolds common in applications, $\mathsf{J}\,\mathsf{g}^{-1}\,\mathsf{J}^{\mathsf{t}}$ could often be inverted efficiently. Thus, this will be our main approach to computing the Riemannian gradient.

The Levi-Civita connection of the manifold, which allows us to take covariant derivatives of the gradient, is used to compute the Riemannian Hessian. A vector field ξ in our context could be considered as a $\mathcal{E}$-valued function from $\mathcal{M}$, such that $\xi(Y) \in T_Y\mathcal{M}$ for all $Y \in \mathcal{M}$. For two vector fields ξ, η on $\mathcal{M}$, the directional derivative $\mathrm{D}_\xi \eta$ is an $\mathcal{E}$-valued function but generally not a vector field (i.e. $(\mathrm{D}_\xi \eta)(Y) \in T_Y\mathcal{M}$ may not hold). A covariant derivative [22] (or connection) associates a vector field $\nabla_\xi \eta$ to two vector fields ξ, η on $\mathcal{M}$. The association is linear in ξ, $\mathbb{R}$-linear in η and satisfies the product rule

$$\nabla_\xi (f\eta) = f\nabla_\xi \eta + (\mathrm{D}_\xi f)\eta$$

for a function f on $\mathcal{M}$, where $\mathrm{D}_\xi f$ denotes the Lie derivative of f (the directional derivative of f along direction ξ_x at each $x \in \mathcal{M}$). For a Riemannian metric $\langle, \rangle_R$ on $\mathcal{M}$, the Levi-Civita connection is the unique connection that is compatible with metric, $\mathrm{D}_\xi \langle \eta, \phi \rangle_R = \langle \nabla_\xi \eta, \phi \rangle_R + \langle \eta, \nabla_\xi \phi \rangle_R$ (ϕ is another vector field), and torsion-free, $\nabla_\xi \eta - \nabla_\eta \xi = [\xi, \eta]$. If a coordinate chart of $\mathcal{M}$ is identified with an open subset of $\mathbb{R}^n$ and $\langle, \rangle_R$ is given by a positive-definite operator g_R, (i.e. $\langle \xi, \eta \rangle_R = \langle \xi, \mathsf{g}_R \eta \rangle_{\mathbb{R}^n}$)

$$\nabla_\xi \eta = \mathrm{D}_\xi \eta + \frac{1}{2}\mathsf{g}_R^{-1}((\mathrm{D}_\xi \mathsf{g}_R)\eta + (\mathrm{D}_\eta \mathsf{g}_R)\xi - \mathcal{X}(\xi, \eta)),$$

where $\mathcal{X}(\xi, \eta) \in \mathbb{R}^n$ (uniquely defined) satisfies $\langle \mathrm{D}_\phi \mathsf{g}_R \xi, \eta \rangle_{\mathbb{R}^n} = \langle \phi, \mathcal{X}(\xi, \eta) \rangle_{\mathbb{R}^n}$ for all vector field ϕ. The formula is valid for each coordinate chart, and it is often given in terms of Christoffel symbols in index notation ([22], proposition 3.13). We

will generalize this operator formula. The gradient and $\mathcal{X}$ are examples of index-raising, translating from a (multi)linear scalar function h to a (multi)linear vector-valued function of one less variable, that evaluates back to h using the inner product pairing.

The Riemannian Hessian could be provided in two formats, as a bilinear form $\mathsf{rhess}_f^{02}(\xi, \eta)$, returning a scalar function to every two vector fields ξ, η on the manifold, or a (Riemannian) Hessian vector product $\mathsf{rhess}_f^{11}\xi$, an operator returning a vector field given a vector field input ξ. In optimization, as we need to invert the Hessian in second-order methods, the Riemannian Hessian vector product form rhess_f^{11} is more practical. However, rhess_f^{02} is directly related to the Levi-Civita connection (see Eq. (3.12) below), and can be read from the geodesic equation: In [9], the authors showed the geodesic equation (for a Stiefel manifold) is given by $\ddot{Y} + \Gamma(\dot{Y}, \dot{Y}) = 0$ where the Christoffel function Γ (defined below) maps two vector fields to an ambient function and the bilinear form rhess_f^{02} is $\hat{f}_{YY}(\xi, \eta) - \langle \Gamma(\xi, \eta), \hat{f}_Y \rangle_{\mathcal{E}}$. Here, $\hat{f}_Y$ and $\hat{f}_{YY}$ are the ambient gradient and Hessian, see Sect. 3.

In application, we also work with quotient manifolds, an example of Riemannian submersion. Recall ([22], Definition 7.44) a Riemannian submersion $\pi : \mathcal{M} \to \mathcal{B}$ between two manifolds $\mathcal{M}$ and $\mathcal{B}$ is a smooth, onto mapping, such that the differential $d\pi$ is onto at every point $Y \in \mathcal{M}$, the fiber $\pi^{-1}(b)$, $b \in \mathcal{B}$ is a Riemannian submanifold of $\mathcal{M}$, and $d\pi$ preserves scalar products of vectors normal to fibers. An important example is the quotient space by a free and proper action of a group of isometries. At each point $Y \in \mathcal{M}$, the tangent space of $\pi^{-1}(\pi Y)$ is called the vertical space, and its orthogonal complement with respect to the Riemannian metric is called the horizontal space. The collection of horizontal spaces $\mathcal{H}_Y$ ($Y \in \mathcal{M}$) of a submersion is a subbundle $\mathcal{H}$. The *horizontal lift*, identifying a tangent vector ξ at $b = \pi(Y) \in \mathcal{B}$ with a horizontal tangent vector $\xi_{\mathcal{H}}$ at Y is a linear isometry between the tangent space $T_b\mathcal{B}$ and $\mathcal{H}_Y$, the horizontal space at Y. The Riemannian framework for a quotient of embedded manifolds is studied through the horizontal lifts, the focus is on the horizontal bundle $\mathcal{H}$ instead of the tangent bundle $T\mathcal{M}$.

The reader can consult [1, 9] for details of Riemannian optimization, including the basic algorithms once the Euclidean and Riemannian gradient and Hessian are computed. In essence, it has been recognized that many popular equation solving and optimization algorithms on Euclidean spaces can be extended to a manifold framework ([9, 10]). Steepest descent on real vector spaces could be extended to manifolds using the Riemannian gradient defined above together with a *retraction*. Here, a retraction R is a sufficiently smooth map mapping $X \in \mathcal{M}$, $\eta \in T_X\mathcal{M}$ to $R(X, \eta) \in \mathcal{M}$ for sufficiently small η. Also, using the Riemannian Hessian, second-order optimization methods, for example, Trust-Region ([1]), could be extended to manifold context. At the i-th iteration step, an optimization algorithm produces a tangent vector η_i to the manifold point Y_i, which will produce the next iteration point $Y_{i+1} = R(Y_i, \eta_i)$ ([3], chapter 4 of [1]). For manifolds considered in this article, computationally efficient retractions are available.

2.1 Notations

We will attempt to state and prove statements for both the real and Hermitian cases at the same time when there are parallel results, as discussed in Sect. 4.1. The base field $\mathbb{K}$ will be $\mathbb{R}$ or $\mathbb{C}$. We use the notation $\mathbb{K}^{n \times m}$ to denote the space of matrices of size $n \times m$ over $\mathbb{K}$. We consider both real and complex vector spaces as real vector spaces, and by $\mathrm{Tr}_{\mathbb{R}}$ we denote the trace of a matrix in the real case or the real part of the trace in the complex case. A real matrix space is a real inner product space with the Frobenius inner product $\mathrm{Tr}_{\mathbb{R}} \, ab^{\mathsf{T}}$, while a complex matrix space becomes a real inner product space with inner product $\mathrm{Tr}_{\mathbb{R}} \, ab^{\mathsf{H}}$ (see Sect. 4.1). We will use the notation t to denote the real adjoint T for a real vector space, and Hermitian adjoint H for a complex vector space, both for matrices and operators. We denote $\mathrm{sym}_{\mathsf{t}} X = \frac{1}{2}(X + X^{\mathsf{t}})$, $\mathrm{skew}_{\mathsf{t}} X = \frac{1}{2}(X - X^{\mathsf{t}})$. We denote by $\mathrm{Sym}_{\mathsf{t}, \mathbb{K}, n}$ the space of t-symmetric matrices $X \in \mathbb{K}^{n \times n}$ with $X^{\mathsf{t}} = X$. The t-antisymmetric space $\mathrm{Skew}_{\mathsf{t}, \mathbb{K}, n}$ is defined similarly. Symbols ξ, η are often used to denote tangent vector or vector fields, while ω is used to denote a vector on the ambient space. The directional derivative in direction v is denoted by D_v, it applies to scalar, vector, or operator-valued functions. If X is a vector field and f is a function, the Lie derivatives will be written as $\mathrm{D}_{\mathsf{X}} f$. We also apply Lie derivatives on scalar or operator-valued functions when X is a vector field, and write $\mathrm{D}_{\mathsf{X}} \mathsf{g}$ for example, where g is a metric operator. Because the vector field X may be a matrix, we prefer the notation $\mathrm{D}_{\mathsf{X}} \mathsf{g}$ to the usual Lie derivative notation $\mathsf{X} \mathsf{g}$ which may be ambiguous. By $\mathrm{U}_{\mathbb{K}, d}$ we denote the group of $\mathbb{K}^{d \times d}$ matrices U satisfying $U^{\mathsf{t}} U = I_d$ (called t-orthogonal), thus $\mathrm{U}_{\mathbb{K}, d}$ is the real orthogonal group $\mathrm{O}(d)$ when $\mathbb{K} = \mathbb{R}$ and unitary group $\mathrm{U}(d)$ when $\mathbb{K} = \mathbb{C}$.

In our approach, a subspace $\mathcal{H}_Y$ of the tangent space at a point Y on a manifold $\mathcal{M}$ is defined as either the nullspace of an operator $\mathrm{J}(Y)$, or the range of an operator $\mathrm{N}(Y)$, both are operator-valued functions on $\mathcal{M}$. Since we most often work with one manifold point Y at a time, we sometimes drop the symbol Y to make the expressions less lengthy. Other operator-valued functions defined in this paper include the ambient metric g, the projection $\Pi_{\mathcal{H}, \mathsf{g}}$ to $\mathcal{H}$, the Christoffel metric term K, and their directional derivatives. We also use the symbols $\hat{f}_Y$ and $\hat{f}_{YY}$ to denote the ambient gradient and Hessian (Y is the manifold variable). We summarize below symbols and concepts related to the main ideas in the paper, with the Stiefel case as an example (details explained in sections below).

Symbol	Concept
$\mathcal{E}$	Ambient space, $\mathcal{M}$ is embedded in $\mathcal{E}$. (e.g. $\mathbb{K}^{n \times p}$).
$\mathcal{E}_{\mathrm{J}}, \mathcal{E}_{\mathrm{N}}$	Inner product spaces, range of J(Y) and domain of N(Y) below.
$\mathcal{H}$	A subbundle of $T\mathcal{M}$. Either $T\mathcal{M}$ or a horizontal bundle in practice.
J(Y)	Operator from $\mathcal{E}$ onto $\mathcal{E}_{\mathrm{J}}$, Null(J(Y)) $= \mathcal{H}_Y \subset T_Y\mathcal{M}$. (e.g. $Y^{\mathrm{t}}\omega + \omega^{\mathrm{t}}Y$).
N(Y)	Inject. oper. from $\mathcal{E}_{\mathrm{N}}$ to $\mathcal{E}$ onto $\mathcal{H}_Y \subset T_Y\mathcal{M}$ (e.g. N$(A, B) = YA + Y_\perp B$).
xtrace	Index-raising operator for the trace (Frobenius) inner product.
g(Y)	Metric given as self-adjoint operator on $\mathcal{E}$ (e.g. $(\alpha_1 YY^{\mathrm{t}} + \alpha_0 Y_\perp Y_\perp^{\mathrm{t}})\eta$).
$\Pi_{\mathrm{g}}, \Pi_{\mathcal{H},\mathrm{g}}$	Projection to $\mathcal{H} \subset T\mathcal{M}$ in Proposition 3.2.
K(ξ, η)	Chrisfl. metric term $\frac{1}{2}((\mathrm{D}_\xi\mathrm{g})\eta + (\mathrm{D}_\eta)\mathrm{g}\xi - \mathrm{xtrace}(\langle(\mathrm{D}_\phi\mathrm{g})\xi, \eta\rangle_{\mathcal{E}}, \phi))$.
$\Gamma_{\mathcal{H}}(\xi, \eta)$	Chrisfl. function $\Pi_{\mathcal{H},\mathrm{g}}\mathrm{g}^{-1}\mathrm{K}(\xi, \eta) - (\mathrm{D}_\xi\Pi_{\mathcal{H},\mathrm{g}})\eta$.

3 Ambient Space and Optimization on Riemannian Manifolds

If $\hat{f}$ is a scalar function from an open subset $\mathcal{U}$ of a Euclidean space $\mathcal{E}$, its gradient $\mathrm{grad}\,\hat{f}$ satisfies $\langle \hat{\eta}, \mathrm{grad}\,\hat{f}\rangle_{\mathcal{E}} = \mathrm{D}_{\hat{\eta}}\hat{f}$ for all vector fields $\hat{\eta}$ on $\mathcal{U}$ where $\mathrm{D}_{\hat{\eta}}\hat{f}$ is the Lie derivative of $\hat{f}$ with respect to $\hat{\eta}$. As well-known [1, 9], the Riemannian gradient and Hessian product of a function f on a submanifold $\mathcal{M} \subset \mathcal{E}$ could be computed from the Euclidean gradient and Hessian, which are evaluated by extending f to a function $\hat{f}$ on a region of $\mathcal{E}$ near $\mathcal{M}$. The process is independent of the extension $\hat{f}$.

Definition 3.1 We call an inner product (Euclidean) space $(\mathcal{E}, \langle, \rangle_{\mathcal{E}})$ an embedded ambient space of a Riemannian manifold $\mathcal{M}$ if there is a differentiable (not necessarily Riemannian) embedding $\mathcal{M} \subset \mathcal{E}$.

Let f be a function on $\mathcal{M}$ and $\hat{f}$ be an extension of f to an open neighborhood of $\mathcal{E}$ containing $\mathcal{M}$. We call $\mathrm{grad}\,\hat{f}$ an ambient gradient of f. It is a vector-valued function from $\mathcal{M}$ to $\mathcal{E}$ such that for all vector fields η on $\mathcal{M}$

$$\langle \eta(Y), \mathrm{grad}\,\hat{f}(Y)\rangle_{\mathcal{E}} = (\mathrm{D}_{\eta(Y)}f)(Y) \text{ for all } Y \in \mathcal{M} \tag{3.1}$$

or equivalently $\langle \eta, \mathrm{grad}\,\hat{f}\rangle_{\mathcal{E}} = \mathrm{D}_\eta f$. Given an ambient gradient $\mathrm{grad}\,\hat{f}$ with continuous derivatives, we define the ambient Hessian to be the map $\mathrm{hess}\,\hat{f}$ associating to a vector field ξ on $\mathcal{M}$ the derivative $\mathrm{D}_\xi\mathrm{grad}\,\hat{f}$. We define the ambient Hessian bilinear form $\mathrm{hess}\,\hat{f}^{02}(\xi, \eta)$ to be $\langle(\mathrm{D}_\xi\mathrm{grad}\,\hat{f}), \eta\rangle_{\mathcal{E}}$. If $Y \in \mathcal{M}$ is considered as a variable, we also use the notation $\hat{f}_Y$ for $\mathrm{grad}\,\hat{f}$ and $\hat{f}_{YY}$ for $\mathrm{hess}\,\hat{f}$.

By the Whitney embedding theorem, any manifold has an ambient space. Coordinate charts could be considered as a collection of compatible local ambient spaces.

From the embedding $\mathcal{M} \subset \mathcal{E}$, the tangent space of $\mathcal{M}$ at each point $Y \in \mathcal{M}$ is considered as a subspace of $\mathcal{E}$. Thus, a vector field on $\mathcal{M}$ could be considered as an $\mathcal{E}$-valued function on $\mathcal{M}$ and we can take its directional derivatives. This derivative is dependent on the embedding and hence not intrinsic. For a function f and two vector fields ξ, η on $\mathcal{M} \subset \mathcal{E}$ we have:

$$\hat{f}_{YY}^{02}(\xi, \eta) = \mathrm{hess}\,\hat{f}^{02}(\xi, \eta) = \mathrm{D}_\xi(\mathrm{D}_\eta f) - \langle\mathrm{D}_\xi\eta, \hat{f}_Y\rangle_{\mathcal{E}}. \tag{3.2}$$

This follows from $D_\xi \langle \eta, \hat{f}_Y \rangle_\mathcal{E} = \langle D_\xi \eta, \hat{f}_Y \rangle_\mathcal{E} + \langle \eta, D_\xi(\hat{f}_Y) \rangle_\mathcal{E} = D_\xi(D_\eta f)$ by taking directional derivatives of Eq. (3.1), thus $\langle \eta, D_\xi(\hat{f}_Y) \rangle_\mathcal{E}$ can be evaluated by Eq. (3.2).

We begin with a standard result of inner product spaces. Recall that the adjoint of a linear map A between two inner product spaces V and W is the map A^t such that $\langle Av, w \rangle_W = \langle v, A^t w \rangle_V$ where $\langle, \rangle_V, \langle, \rangle_W$ denote the inner products on V and W, respectively. If A is represented by a matrix also called A in two orthogonal bases in V and W respectively, then A^t is represented by its transpose A^T. A projection from an inner product space V to a subspace W is a linear operator Π_W on V such that $\Pi_W v \in W$ and $\langle v, w \rangle_V = \langle \Pi_W v, w \rangle_V$ for all $w \in W$, $v \in V$. It is well-known a projection always exists and unique, and $\Pi_W v$ minimizes the distance from v to W.

Proposition 3.1 *Let $\mathcal{E}$ be a vector space with an inner product $\langle, \rangle_\mathcal{E}$. Let g be a self-adjoint positive-definite operator on $\mathcal{E}$, thus $\langle \mathsf{g}e_1, e_2 \rangle_\mathcal{E} = \langle e_1, \mathsf{g}e_2 \rangle_\mathcal{E}$. The operator g defines a new inner product on $\mathcal{E}$ by $\langle e_1, e_2 \rangle_{\mathcal{E},\mathsf{g}} := \langle e_1, \mathsf{g}e_2 \rangle_\mathcal{E}$. If $W = \mathrm{Null}(\mathrm{J})$ for a map J from $\mathcal{E}$ onto an inner product space $\mathcal{E}_\mathrm{J}$, the projection $\Pi_\mathsf{g} = \Pi_{\mathsf{g},W}$ from $\mathcal{E}$ to W under the inner product $\langle, \rangle_{\mathcal{E},\mathsf{g}}$ is given by $\Pi_\mathsf{g} e = e - \mathsf{g}^{-1} \mathrm{J}^t (\mathrm{J}\mathsf{g}^{-1}\mathrm{J}^t)^{-1} \mathrm{J}e$, where J^t is the adjoint map of J for all $e \in \mathcal{E}$.*

Alternatively, if N is a one-to-one map from an inner product space $\mathcal{E}_\mathrm{N}$ to $\mathcal{E}$ such that $W = \mathrm{N}(\mathcal{E}_\mathrm{N})$, then the projection to W could be given by $\Pi_\mathsf{g} e = \mathrm{N}(\mathrm{N}^t\mathsf{g}\mathrm{N})^{-1}\mathrm{N}^t\mathsf{g}e$.

The operators $\mathsf{g}\Pi_\mathsf{g}$ and $\Pi_\mathsf{g}\mathsf{g}^{-1}$ are self-adjoint under $\langle, \rangle_\mathcal{E}$.

Proof The assumption that J is onto $\mathcal{E}_\mathrm{J}$ shows J^t is injective (as $\mathrm{J}^t a = 0$ implies $\langle a, \mathrm{J}\omega \rangle_{\mathcal{E}_\mathrm{J}} = 0$ for all $\omega \in \mathcal{E}$, and since J is onto this implies $a = 0$). This in turn implies $\mathrm{J}\mathsf{g}^{-1}\mathrm{J}^t$ is invertible as if $\mathrm{J}\mathsf{g}^{-1}\mathrm{J}^t a = 0$, then $\langle \mathrm{J}\mathsf{g}^{-1}\mathrm{J}^t a, a \rangle_\mathcal{E} = 0$, so $\langle \mathsf{g}^{-1}\mathrm{J}^t a, \mathrm{J}^t a \rangle_\mathcal{E} = 0$ and hence $\mathrm{J}^t a = 0$ as g is positive-definite. We can show $\mathrm{N}^t\mathsf{g}\mathrm{N}$ is invertible similarly.

For the first case, if $e_W \in W = \mathrm{Null}(\mathrm{J})$ and $e \in \mathcal{E}$,

$$\langle \mathsf{g}^{-1}\mathrm{J}^t(\mathrm{J}\mathsf{g}^{-1}\mathrm{J}^t)^{-1}\mathrm{J}e, \mathsf{g}e_W \rangle_\mathcal{E} = \langle \mathrm{J}^t(\mathrm{J}\mathsf{g}^{-1}\mathrm{J}^t)^{-1}\mathrm{J}e, e_W \rangle_\mathcal{E} = \langle (\mathrm{J}\mathsf{g}^{-1}\mathrm{J}^t)^{-1}\mathrm{J}e, \mathrm{J}e_W \rangle_\mathcal{E},$$

where the last term is zero because $e_W \in \mathrm{Null}(\mathrm{J})$, so $\langle \Pi_\mathsf{g}e, \mathsf{g}e_W \rangle_\mathcal{E} = \langle e, \mathsf{g}e_W \rangle_\mathcal{E}$. For the second case, assuming $e_W = \mathrm{N}(e_\mathrm{N})$ for $e_\mathrm{N} \in \mathcal{E}_\mathrm{N}$ then (Using $(AB)^t = B^t A^t$):

$$\langle \mathrm{N}(\mathrm{N}^t\mathsf{g}\mathrm{N})^{-1}\mathrm{N}^t\mathsf{g}e, \mathsf{g}\mathrm{N}(e_\mathrm{N}) \rangle_\mathcal{E} = \langle \mathsf{g}e, \mathrm{N}(\mathrm{N}^t\mathsf{g}\mathrm{N})^{-1}\mathrm{N}^t\mathsf{g}\mathrm{N}(e_\mathrm{N}) \rangle_\mathcal{E} = \langle \mathsf{g}e, \mathrm{N}(e_\mathrm{N}) \rangle_\mathcal{E}.$$

The last statement follows from the defining equations of Π_g. $\qquad\square$

Recall the Riemannian gradient of a function f on a manifold $\mathcal{M}$ with Riemannian metric $\langle, \rangle_R$ is the vector field rgrad_f such that $\langle \mathrm{rgrad}_f(Y), \xi(Y) \rangle_R = (D_\xi f)(Y)$ for any point $Y \in \mathcal{M}$, and any vector field ξ. Let $\mathcal{H}$ be a subbundle of the tangent bundle $T\mathcal{M}$, we recall this means $\mathcal{H}$ is a collection of subspaces (fibers) $\mathcal{H}_Y \subset T_Y\mathcal{M}$ for $Y \in \mathcal{M}$ such that $\mathcal{H}$ is itself a vector bundle on $\mathcal{M}$, i.e. $\mathcal{H}$ is locally a product of a vector space and an open subset of $\mathcal{M}$, together with a linear coordinate change condition (see [22], definition 7.24 for details). We can define the $\mathcal{H}$-Riemannian gradient $\mathrm{rgrad}_{\mathcal{H},f}$ of f as the unique $\mathcal{H}$-valued vector field such that $\langle \mathrm{rgrad}_{\mathcal{H},f}, \xi_\mathcal{H} \rangle_R = D_{\xi_\mathcal{H}}f$ for any $\mathcal{H}$-valued vector field $\xi_\mathcal{H}$. Uniqueness follows from nondegeneracy of the inner product restricted to $\mathcal{H}$. Clearly, when $\mathcal{H} = T\mathcal{M}$, $\mathrm{rgrad}_{\mathcal{H},f} = \mathrm{rgrad}_{T\mathcal{M}}f$ is the usual Riemannian gradient. We have:

Proposition 3.2 *Let $(\mathcal{E}, \langle, \rangle_{\mathcal{E}})$ be an embedded ambient space of a manifold $\mathcal{M}$ as in definition 3.1. Let g be a smooth operator-valued function associating each $Y \in \mathcal{M}$ a self-adjoint positive-definite operator $\mathsf{g}(Y)$ on $\mathcal{E}$. Thus, each $\mathsf{g}(Y)$ defines an inner product on $\mathcal{E}$, which induces an inner product on $T_Y\mathcal{M}$ and hence g induces a Riemannian metric on $\mathcal{M}$. Let $\mathcal{H}$ be a subbundle of $T\mathcal{M}$. Define $\Pi_{\mathcal{H},\mathsf{g}}$ to be the operator-valued function such that $\Pi_{\mathcal{H},\mathsf{g}}(Y)$ is the projection associated with $\mathsf{g}(Y)$ from $\mathcal{E}$ to the fiber $\mathcal{H}_Y$, and for the case $\mathcal{H} = T\mathcal{M}$, define $\Pi_{\mathcal{M},\mathsf{g}} = \Pi_{T\mathcal{M},\mathsf{g}}$. For an ambient gradient $\mathsf{grad}\,\hat{f}$ of f, the $\mathcal{H}$-Riemannian gradient of f can be evaluated as:*

$$\mathsf{rgrad}_{\mathcal{H},f} = \Pi_{\mathcal{H},\mathsf{g}}\mathsf{g}^{-1}\mathsf{grad}\,\hat{f}. \tag{3.3}$$

If there is an inner product space $\mathcal{E}_{\mathrm{J}}$ and a map J from $\mathcal{M}$ to the space $\mathfrak{L}(\mathcal{E}, \mathcal{E}_{\mathrm{J}})$ of linear maps from $\mathcal{E}$ to $\mathcal{E}_{\mathrm{J}}$, such that for each $Y \in \mathcal{M}$, the range of $\mathrm{J}(Y)$ is precisely $\mathcal{E}_{\mathrm{J}}$, and its nullspace is $\mathcal{H}_Y$ then $\Pi_{\mathcal{H},\mathsf{g}}(Y)e$ for $e \in \mathcal{E}$ could be given by:

$$\Pi_{\mathcal{H},\mathsf{g}}(Y)e = e - \mathsf{g}^{-1}\,\mathrm{J}^{\mathsf{t}}(\mathrm{J}\,\mathsf{g}^{-1}\,\mathrm{J}^{\mathsf{t}})^{-1}\,\mathrm{J}\,e; \ \textit{all are evaluated at } Y. \tag{3.4}$$

If there is an inner product space $\mathcal{E}_{\mathrm{N}}$ and a map N from $\mathcal{M}$ to the space $\mathfrak{L}(\mathcal{E}_{\mathrm{N}}, \mathcal{E})$ of linear maps from $\mathcal{E}_{\mathrm{N}}$ to $\mathcal{E}$ such that for each $Y \in \mathcal{M}$, $\mathrm{N}(Y)$ is one-to-one, with its range is precisely $\mathcal{H}_Y$ then:

$$\Pi_{\mathcal{H},\mathsf{g}}e = \mathrm{N}(\mathrm{N}^{\mathsf{t}}\mathsf{g}\mathrm{N})^{-1}\mathrm{N}^{\mathsf{t}}\mathsf{g}e; \ \textit{all are evaluated at } Y. \tag{3.5}$$

Proof For any $\mathcal{H}$-valued vector field $\xi_{\mathcal{H}}$, we have:

$$\langle \Pi_{\mathcal{H},\mathsf{g}}\mathsf{g}^{-1}\mathsf{grad}\,\hat{f}, \mathsf{g}\xi_{\mathcal{H}}\rangle_{\mathcal{E}} = \langle \mathsf{grad}\,\hat{f}, \Pi_{\mathcal{H},\mathsf{g}}\mathsf{g}^{-1}\mathsf{g}\xi_{\mathcal{H}}\rangle_{\mathcal{E}} = \langle \hat{f}_Y, \xi_{\mathcal{H}}\rangle_{\mathcal{E}} = \mathrm{D}_{\xi_{\mathcal{H}}}f$$

because $\Pi_{\mathcal{H},\mathsf{g}}\mathsf{g}^{-1}$ is self-adjoint and the projection is idempotent. The remaining statements are just a parametrized version of Proposition 3.1. $\qquad\square$

Note, we are not making any smoothness assumption on J or N yet, although $\Pi_{\mathcal{H},\mathsf{g}}$ is assumed to be sufficiently smooth. In fact, N is often not smooth. J is usually smooth as it is constructed from a smooth constraint on $\mathcal{M}$, or on the horizontal requirements of a vector field.

Definition 3.2 A triple $(\mathcal{M}, \mathsf{g}, \mathcal{E})$ with $\mathcal{E}$ an inner product space, $\mathcal{M} \subset \mathcal{E}$ a differentiable manifold submersion, and g is a positive-definite operator-valued-function from $\mathcal{M}$ to $\mathfrak{L}(\mathcal{E}, \mathcal{E})$ is called an embedded ambient structure of $\mathcal{M}$. $\mathcal{M}$ is a Riemannian manifold with the metric induced by g.

From the definition of Lie brackets, for an embedded ambient space $\mathcal{E}$ of $\mathcal{M}$ we have

$$\mathrm{D}_{\xi}\eta - \mathrm{D}_{\eta}\xi = [\xi, \eta] \text{ for all vector fields } \xi, \eta \text{ on } \mathcal{M}. \tag{3.6}$$

Recall if $\mathcal{M}, \langle, \rangle_R$ is a Riemannian manifold with the Levi-Civita connection ∇, the Riemannian Hessian (vector product) of a function f is the operator sending a tangent vector ξ to the tangent vector $\mathsf{rhess}_f^{11}\xi = \nabla_{\xi}\mathsf{rgrad}_f$. The Riemannian Hessian

bilinear form is the map evaluating on two vector fields ξ, η as $\langle \nabla_\xi \mathrm{rgrad}_f, \eta \rangle_R$. For a subbundle $\mathcal{H}$ of $T\mathcal{M}$ and a $\mathcal{H}$-valued vector field $\xi_\mathcal{H}$, we define the $\mathcal{H}$-Riemannian Hessian similarly as $\mathrm{rhess}^{11}_{\mathcal{H},f} \xi_\mathcal{H} = \Pi_{\mathcal{H},\mathsf{g}} \nabla_{\xi_\mathcal{H}} \mathrm{rgrad}_{\mathcal{H},f}$ and we call $\mathrm{rhess}^{02}_{\mathcal{H},f}(\xi_\mathcal{H}, \eta_\mathcal{H}) = \langle \Pi_{\mathcal{H},\mathsf{g}} \nabla_{\xi_\mathcal{H}} \mathrm{rgrad}_{\mathcal{H},f}, \eta_\mathcal{H} \rangle_R = \langle \nabla_{\xi_\mathcal{H}} \mathrm{rgrad}_{\mathcal{H},f}, \eta_\mathcal{H} \rangle_R$ the $\mathcal{H}$-Riemannian Hessian bilinear form. The next theorem shows how to compute the Riemannian connection and the associated Riemannian Hessian.

Theorem 3.1 *Let $(\mathcal{M}, \mathsf{g}, \mathcal{E})$ be an embedded ambient structure of a Riemannian manifold $\mathcal{M}$. There exists an $\mathcal{E}$-valued bilinear form $\mathcal{X}$ sending a pair of vector fields (ξ, η) to $\mathcal{X}(\xi, \eta) \in \mathcal{E}$ such that for any vector field ξ_0:*

$$\langle \mathcal{X}(\xi, \eta), \xi_0 \rangle_\mathcal{E} = \langle \xi, (\mathrm{D}_{\xi_0} \mathsf{g}) \eta \rangle_\mathcal{E}. \tag{3.7}$$

Let $\Pi_{\mathcal{M},\mathsf{g}}$ be the projection from $\mathcal{E}$ to the tangent bundle of $\mathcal{M}$. Then $\Pi_{\mathcal{M};\mathsf{g}} \mathsf{g}^{-1} \mathcal{X}(\xi, \eta)$ is uniquely defined given ξ, η and $\mathcal{X}(\xi, \eta)$ is also unique if we require $\mathcal{X}(\xi(Y), \eta(Y))$ to be in $T_Y\mathcal{M}$ for all $Y \in \mathcal{M}$. For two vector fields ξ, η on $\mathcal{M}$, define

$$\begin{aligned} \mathrm{K}(\xi, \eta) &:= \frac{1}{2}((\mathrm{D}_\xi \mathsf{g})\eta + (\mathrm{D}_\eta \mathsf{g})\xi - \mathcal{X}(\xi, \eta)) \in \mathcal{E}, \\ \hat{\nabla}_\xi \eta &:= \mathrm{D}_\xi \eta + \mathsf{g}^{-1} \mathrm{K}(\xi, \eta), \\ \nabla_\xi \eta &:= \Pi_{\mathcal{M};\mathsf{g}} \hat{\nabla}_\xi \eta = \Pi_{\mathcal{M};\mathsf{g}}(\mathrm{D}_\xi \eta + \mathsf{g}^{-1} \mathrm{K}(\xi, \eta)). \end{aligned} \tag{3.8}$$

Then $\nabla_\xi \eta$ is the covariant derivative associated with the Levi-Civita connection. It could be written using the Christoffel function Γ:

$$\begin{aligned} \Gamma(\xi, \eta) &:= -(\mathrm{D}_\xi \Pi_{\mathcal{M};\mathsf{g}})\eta + \Pi_{\mathcal{M};\mathsf{g}} \mathsf{g}^{-1} \mathrm{K}(\xi, \eta), \\ \nabla_\xi \eta &= \mathrm{D}_\xi \eta + \Gamma(\xi, \eta). \end{aligned} \tag{3.9}$$

If $\mathcal{H}$ is a subbundle of $T\mathcal{M}$, and $\xi_\mathcal{H}, \eta_\mathcal{H}$ are two $\mathcal{H}$-valued vector fields, we have:

$$\begin{aligned} \Pi_{\mathcal{H},\mathsf{g}} \nabla_{\xi_\mathcal{H}} \eta_\mathcal{H} &= \mathrm{D}_{\xi_\mathcal{H}} \eta_\mathcal{H} + \Gamma_\mathcal{H}(\xi_\mathcal{H}, \eta_\mathcal{H}) \text{ with} \\ \Gamma_\mathcal{H}(\xi_\mathcal{H}, \eta_\mathcal{H}) &:= -(\mathrm{D}_{\xi_\mathcal{H}} \Pi_{\mathcal{H};\mathsf{g}})\eta_\mathcal{H} + \Pi_{\mathcal{H};\mathsf{g}} \mathsf{g}^{-1} \mathrm{K}(\xi_\mathcal{H}, \eta_\mathcal{H}). \end{aligned} \tag{3.10}$$

If f is a function on $\mathcal{M}$, $\hat{f}_Y$ is an ambient gradient of f and $\hat{f}_{YY}$ is the ambient Hessian operator, then $\mathrm{rhess}^{11}_{\mathcal{H},f} \xi_\mathcal{H} := \Pi_{\mathcal{H};\mathsf{g}} \nabla_{\xi_\mathcal{H}} \mathrm{rgrad}_{\mathcal{H},f}$ and $\mathrm{rhess}^{02}_{\mathcal{H},f}$ are given by:

$$\begin{aligned} \mathrm{rhess}^{11}_{\mathcal{H},f} \xi_\mathcal{H} &= \Pi_{\mathcal{H};\mathsf{g}} \mathsf{g}^{-1}(\hat{f}_{YY} \xi_\mathcal{H} + \mathsf{g}(\mathrm{D}_{\xi_\mathcal{H}}(\Pi_{\mathcal{H},\mathsf{g}} \mathsf{g}^{-1})) \hat{f}_Y + \mathrm{K}(\xi_\mathcal{H}, \Pi_{\mathcal{H},\mathsf{g}} \mathsf{g}^{-1} \hat{f}_Y)) \\ &= \Pi_{\mathcal{H};\mathsf{g}} \mathsf{g}^{-1}(\hat{f}_{YY} \xi_\mathcal{H} + \mathsf{g}(\mathrm{D}_{\xi_\mathcal{H}} \Pi_{\mathcal{H},\mathsf{g}}) \mathsf{g}^{-1} \hat{f}_Y \\ &\quad -(\mathrm{D}_{\xi_\mathcal{H}} \mathsf{g}) \mathsf{g}^{-1} \hat{f}_Y + \mathrm{K}(\xi_\mathcal{H}, \Pi_{\mathcal{H},\mathsf{g}} \mathsf{g}^{-1} \hat{f}_Y)), \\ \mathrm{rhess}^{02}_{\mathcal{H},f}(\eta_\mathcal{H}, \xi_\mathcal{H}) &= \langle \nabla_{\xi_\mathcal{H}} \Pi_{\mathcal{H},\mathsf{g}} \mathsf{g}^{-1} \hat{f}_Y, \mathsf{g} \eta_\mathcal{H} \rangle_\mathcal{E} \\ &= \hat{f}_{YY}(\xi_\mathcal{H}, \eta_\mathcal{H}) - \langle \Gamma_\mathcal{H}(\xi_\mathcal{H}, \eta_\mathcal{H}), \hat{f}_Y \rangle_\mathcal{E}. \end{aligned} \tag{3.11, 3.12}$$

The form $\Gamma(\xi, \eta)$ appeared in [9] and was computed for the case of a Stiefel manifold, and was called a Christoffel function. It includes the Christoffel metric term K and the derivative of $\Pi_{\mathcal{M},\mathsf{g}}$. Evaluated at $Y \in \mathcal{M}$, it depends only on the tangent vectors $\eta(Y)$ and $\xi(Y)$, not on the whole vector fields. Equation (2.57) in that reference is the expression of rhess_f^{02} in terms of Γ above. In [9], $\Gamma_{\mathcal{H}}$ was computed for a Grassmann manifold. The formulation for subbundles allows us to extend the result to Riemannian submersions and quotient manifolds. Equation 3.11 generalizes the *Weingarten map* formula in [2, equations 7,10] when $\mathsf{g} = I_{\mathcal{E}}$, since by product rule

$$\Pi_{\mathcal{H};\mathsf{g}}(\mathrm{D}_{\xi_{\mathcal{H}}}\Pi_{\mathcal{H},\mathsf{g}})\hat{f}_Y = (\mathrm{D}_{\xi_{\mathcal{H}}}(\Pi_{\mathcal{H},\mathsf{g}}^2))\hat{f}_Y - (\mathrm{D}_{\xi_{\mathcal{H}}}\Pi_{\mathcal{H},\mathsf{g}})\Pi_{\mathcal{H};\mathsf{g}}\hat{f}_Y,$$

and $\mathsf{rhess}_{\mathcal{H},f}^{11}\xi_{\mathcal{H}}$ becomes

$$\Pi_{\mathcal{H};\mathsf{g}}(\hat{f}_{YY}\xi_{\mathcal{H}}) + \Pi_{\mathcal{H};\mathsf{g}}(\mathrm{D}_{\xi_{\mathcal{H}}}\Pi_{\mathcal{H},\mathsf{g}})\hat{f}_Y = \Pi_{\mathcal{H};\mathsf{g}}(\hat{f}_{YY}\xi_{\mathcal{H}}) + (\mathrm{D}_{\xi_{\mathcal{H}}}\Pi_{\mathcal{H},\mathsf{g}})((I_{\mathcal{E}} - \Pi_{\mathcal{H};\mathsf{g}})\hat{f}_Y).$$

Here, $V := (I_{\mathcal{E}} - \Pi_{\mathcal{H};\mathsf{g}})\hat{f}_Y$ is vertical ($\Pi_{\mathcal{H};\mathsf{g}}V = 0$) and $(\mathrm{D}_{\xi_{\mathcal{H}}}\Pi_{\mathcal{H},\mathsf{g}})V$ is horizontal.

Proof $\mathcal{X}$ is the familiar index-raising term: for $Y \in \mathcal{M}$ and $v_0, v_1, v_2 \in T_Y\mathcal{M}$, as $\langle v_1, (\mathrm{D}_{v_0}\mathsf{g})v_2\rangle_{\mathcal{E}}$ is a tri-linear function on $T_Y\mathcal{M}$ and the Riemannian inner product on $T_Y\mathcal{M}$ is nondegenerate, the index-raising bilinear form $\tilde{\mathcal{X}}$ with value in $T_Y\mathcal{M}$ is uniquely defined, so $\mathcal{X}(\xi(Y), \eta(Y)) = \tilde{\mathcal{X}}(\xi(Y), \eta(Y))$ satisfies Eq. (3.7), where we consider $T_Y\mathcal{M}$ as a subspace of $\mathcal{E}$. Thus, we have proved the existence of $\mathcal{X}$. If we take another $\mathcal{E}$-valued function $\mathcal{X}_1$ satisfying the same condition but not necessarily in the tangent space, the expression $\Pi_{\mathcal{M},\mathsf{g}}\mathsf{g}^{-1}\mathcal{X}_1$, hence $\Pi_{\mathcal{M},\mathsf{g}}\mathsf{g}^{-1}\mathrm{K}$ is independent of the choice of $\mathcal{X}_1$, as for three vector fields ξ_0, ξ, η

$$\langle\mathsf{g}\xi_0, \Pi_{\mathcal{M},\mathsf{g}}\mathsf{g}^{-1}\mathcal{X}_1(\xi, \eta)\rangle_{\mathcal{E}} = \langle\mathsf{g}\xi_0, \mathsf{g}^{-1}\mathcal{X}_1(\xi, \eta)\rangle_{\mathcal{E}} = \langle\xi, \mathrm{D}_{\xi_0}\eta\rangle_{\mathcal{E}}.$$

We can verify directly that $\nabla_\xi\eta$ satisfies the conditions of a covariant derivative: linear in ξ and satisfying the product rule with respect to η. Similar to the calculation with coordinate charts, we can show ∇ is compatible with metric: for two vector fields η, ξ, $2\langle\nabla_\xi\eta, \mathsf{g}\eta\rangle_{\mathcal{E}} = 2\langle\Pi_{\mathcal{M},\mathsf{g}}\hat{\nabla}_\xi\eta, \mathsf{g}\eta\rangle_{\mathcal{E}}$, which is $2\langle\mathrm{D}_\xi\eta + \mathsf{g}^{-1}\mathrm{K}(\xi, \eta), \mathsf{g}\eta\rangle_{\mathcal{E}}$ by definition and by property of the projection. Expanding the last expression and use $\langle\mathcal{X}(\xi, \eta), \eta\rangle_{\mathcal{E}} = \langle(\mathrm{D}_\eta\mathsf{g})\xi, \eta\rangle_{\mathcal{E}}$

$$2\langle\mathrm{D}_\xi\eta, \mathsf{g}\eta\rangle_{\mathcal{E}} + 2\langle\mathsf{g}^{-1}\mathrm{K}(\xi, \eta), \mathsf{g}\eta\rangle_{\mathcal{E}}$$
$$= 2\langle\mathrm{D}_\xi\eta, \mathsf{g}\eta\rangle_{\mathcal{E}} + \langle(\mathrm{D}_\xi\mathsf{g})\eta + (\mathrm{D}_\eta\mathsf{g})\xi - \mathcal{X}(\xi, \eta), \eta\rangle_{\mathcal{E}}$$
$$= 2\langle\mathrm{D}_\xi\eta, \mathsf{g}\eta\rangle_{\mathcal{E}} + \langle(\mathrm{D}_\xi\mathsf{g})\eta, \eta\rangle_{\mathcal{E}} = \mathrm{D}_\xi\langle\eta, \mathsf{g}\eta\rangle_{\mathcal{E}}.$$

Torsion-free follows from the fact that K is symmetric and Eq. (3.6):

$$\nabla_\xi\eta - \nabla_\eta\xi = \Pi_{\mathcal{M},\mathsf{g}}(\mathrm{D}_\xi\eta - \mathrm{D}_\eta\xi) = [\xi, \eta].$$

For Eq. (3.9), we note $\Pi_{\mathcal{M},\mathsf{g}}\mathrm{D}_\xi\eta = \mathrm{D}_\xi(\Pi_{\mathcal{M},\mathsf{g}}\eta) - (\mathrm{D}_\xi\Pi_{\mathcal{M},\mathsf{g}})\eta$ so

$$\nabla_\xi\eta = \mathrm{D}_\xi(\eta) - (\mathrm{D}_\xi\Pi_{\mathcal{M},\mathsf{g}})\eta + \Pi_{\mathcal{M},\mathsf{g}}\mathsf{g}^{-1}\mathrm{K}(\xi, \eta).$$

For Eq. (3.10), $\Pi_{\mathcal{H},g}\Pi_{\mathcal{M},g} = \Pi_{\mathcal{H},g}$ if $Y \in \mathcal{M}$, $\mathcal{H}_Y \subset T_Y\mathcal{M}$. Hence, Eq. (3.8) implies

$$\Pi_{\mathcal{H},g}\nabla_{\xi_{\mathcal{H}}}\eta_{\mathcal{H}} = \Pi_{\mathcal{H};g}(D_\xi\eta_{\mathcal{H}} + g^{-1}K(\xi_{\mathcal{H}}, \eta_{\mathcal{H}})),$$

and as before, we use $\Pi_{\mathcal{H},g}D_{\xi_{\mathcal{H}}}\eta_{\mathcal{H}} = D_{\xi_{\mathcal{H}}}(\Pi_{\mathcal{H},g}\eta_{\mathcal{H}}) - (D_{\xi_{\mathcal{H}}}\Pi_{\mathcal{H},g})\eta_{\mathcal{H}}$ and $\Pi_{\mathcal{H},g}\eta_{\mathcal{H}} = \eta_{\mathcal{H}}$. The first line of Eq. (3.11) is by definition and $D_\xi(\Pi_{\mathcal{H},g}g^{-1}\hat{f}_Y) = D_\xi(\Pi_{\mathcal{H},g}g^{-1})\hat{f}_Y + \Pi_{\mathcal{H},g}g^{-1}\hat{f}_{YY}$. Expand, note $\Pi_{\mathcal{H},g}g^{-1}(g\Pi_{\mathcal{H},g}D_{\xi_{\mathcal{H}}}g^{-1})\hat{f}_Y = -\Pi_{\mathcal{H},g}g^{-1}(D_{\xi_{\mathcal{H}}}g)g^{-1}\hat{f}_Y$ (as $\Pi_{\mathcal{H},g}$ is idempotent), we have the second line. For Eq. (3.12):

$$\begin{aligned}
\langle\nabla_{\xi_{\mathcal{H}}}\Pi_{\mathcal{H},g}(g^{-1}\hat{f}_Y), g\eta_{\mathcal{H}}\rangle_{\mathcal{E}} &= D_{\xi_{\mathcal{H}}}\langle\Pi_{\mathcal{H},g}(g^{-1}\hat{f}_Y), g\eta_{\mathcal{H}}\rangle_{\mathcal{E}} - \langle\Pi_{\mathcal{H},g}(g^{-1}\hat{f}_Y), g\nabla_\xi\eta_{\mathcal{H}}\rangle_{\mathcal{E}} \\
&= D_{\xi_{\mathcal{H}}}\langle\hat{f}_Y, \eta_{\mathcal{H}}\rangle_{\mathcal{E}} - \langle\hat{f}_Y, \Pi_{\mathcal{H},g}g^{-1}g\Pi_{\mathcal{H},g}\hat{\nabla}_{\xi_{\mathcal{H}}}\eta_{\mathcal{H}}\rangle_{\mathcal{E}} \\
&= D_{\xi_{\mathcal{H}}}\langle\hat{f}_Y, \eta_{\mathcal{H}}\rangle_{\mathcal{E}} - \langle\hat{f}_Y, \Pi_{\mathcal{H},g}\hat{\nabla}_{\xi_{\mathcal{H}}}\eta_{\mathcal{H}}\rangle_{\mathcal{E}} \\
&= D_{\xi_{\mathcal{H}}}\langle\hat{f}_Y, \eta_{\mathcal{H}}\rangle_{\mathcal{E}} - \langle\hat{f}_Y, D_{\xi_{\mathcal{H}}}\eta_{\mathcal{H}}\rangle_{\mathcal{E}} - \langle\hat{f}_Y, \Gamma_{\mathcal{H}}(\xi_{\mathcal{H}}, \eta_{\mathcal{H}})\rangle_{\mathcal{E}} \\
&= \hat{f}_{YY}(\xi_{\mathcal{H}}, \eta_{\mathcal{H}}) - \langle\hat{f}_Y, \Gamma_{\mathcal{H}}(\xi_{\mathcal{H}}, \eta_{\mathcal{H}})\rangle_{\mathcal{E}},
\end{aligned}$$

from compatibility with metric, idempotency of $\Pi_{\mathcal{H},g}$, Eqs. (3.10) and (3.2).   □

When the projection is given in terms of J, and J is sufficiently smooth we have:

Proposition 3.3 *If J as in Proposition 3.2 is of class C^2 then:*

$$\Gamma_{\mathcal{H}}(\xi_{\mathcal{H}}, \eta_{\mathcal{H}}) = g^{-1}J^t(Jg^{-1}J^t)^{-1}(D_{\xi_{\mathcal{H}}}J)\eta_{\mathcal{H}} + \Pi_{\mathcal{H},g}g^{-1}K(\xi_{\mathcal{H}}, \eta_{\mathcal{H}}) \quad (3.13)$$

for two $\mathcal{H}$-valued tangent vectors $\xi_{\mathcal{H}}, \eta_{\mathcal{H}}$ at $Y \in \mathcal{M}$. We can evaluate $\mathrm{rhess}^{11}_{\mathcal{H},f}\xi_{\mathcal{H}}$ by setting $\omega = \hat{f}_Y$ in the following formula, which is valid for all $\mathcal{E}$-valued function ω:

$$\begin{aligned}
\nabla_{\xi_{\mathcal{H}}}\Pi_{\mathcal{H},g}g^{-1}\omega = &\Pi_{\mathcal{H},g}g^{-1}D_{\xi_{\mathcal{H}}}\omega - \Pi_{\mathcal{H},g}g^{-1}(D_{\xi_{\mathcal{H}}}g)g^{-1}\omega \\
&- \Pi_{\mathcal{H},g}(D_{\xi_{\mathcal{H}}}(g^{-1}J^t))(Jg^{-1}J^t)^{-1}Jg^{-1}\omega + \Pi_{\mathcal{H},g}g^{-1}K(\xi_{\mathcal{H}}, \Pi_{\mathcal{H},g}(g^{-1}\omega)).
\end{aligned}$$
$$(3.14)$$

Proof The first expression follows by expanding $D_{\xi_{\mathcal{H}}}\Pi_{\mathcal{H},g}$ in terms of J, noting $J\eta_{\mathcal{H}} = 0$. For the second, expand $\Pi_{\mathcal{H},g}\hat{\nabla}_{\xi_{\mathcal{H}}}(\mathcal{H}, \Pi_g g^{-1}\omega) = \Pi_{\mathcal{H},g}D_{\xi_{\mathcal{H}}}(\Pi_{\mathcal{H},g}g^{-1}\omega) + \Pi_{\mathcal{H},g}g^{-1}K(\xi_{\mathcal{H}}, \Pi_{\mathcal{H},g}g^{-1}\omega)$, then expand the first term and use $\Pi_{\mathcal{H},g}g^{-1}J^t = 0$.   □

The following proposition allows us to apply the results so far in familiar contexts:

Proposition 3.4 *Let $(\mathcal{M}, g, \mathcal{E})$ be an embedded ambient structure.*

1. *Fix an orthogonal basis e_i of $\mathcal{E}$, let f be a function on $\mathcal{M}$, which is a restriction of a function $\hat{f}$ on $\mathcal{E}$, define $\hat{f}_Y$ to be the function from $\mathcal{M}$ to $\mathcal{E}$, having the i-th component the directional derivative $D_{e_i}\hat{f}$, then $\hat{f}_Y$ is an ambient gradient. If $\mathcal{M}$ is defined by the equation $C(Y) = 0$ $(Y \in \mathcal{M})$ with a full rank Jacobian, then the nullspace of the Jacobian $J_C(Y)$ is the tangent space of $\mathcal{M}$ at Y, hence $J_C(Y)$ could be used as the operator $J(Y)$.*

2. *(Riemannian submersion) Let $(\mathcal{M}, \mathsf{g}, \mathcal{E})$ be an embedded ambient structure. Let $\pi : \mathcal{M} \rightarrow \mathcal{B}$ be a Riemannian submersion, with $\mathcal{H}$ the corresponding horizontal subbundle of $T\mathcal{M}$. If ξ, η are two vector fields on $\mathcal{B}$ with $\xi_\mathcal{H}, \eta_\mathcal{H}$ their horizontal lifts, then the Levi-Civita connection $\nabla^\mathcal{B}_\xi \eta$ on $\mathcal{B}$ lifts to $\Pi_{\mathcal{H},\mathsf{g}} \nabla_{\xi_\mathcal{H}} \eta_\mathcal{H}$, hence Eq. (3.10) applies. Also, Riemannian gradients and Hessians on $\mathcal{B}$ lift to $\mathcal{H}$-Riemannian gradients and Hessians on $\mathcal{M}$.*

Proof The construction of $\hat{f}_Y$ ensures $\langle \hat{f}_Y, e_i \rangle_\mathcal{E} = \mathrm{D}_{e_i} f$. The statement about the Jacobian is simply the implicit function theorem. Isometry of horizontal lift and [22], Lemma 7.45, item 3, gives us Statement 2. □

In practice, $\hat{f}_Y$ is computed by index-raising the directional derivative. For clarity, so far we use the subscript $\mathcal{H}$ to indicate the relation to a subbundle $\mathcal{H}$. For the rest of the paper, we will drop the subscripts $\mathcal{H}$ on vector fields (referring to ξ instead of $\xi_\mathcal{H}$) as it will be clear from the context if we discuss a vector field in $\mathcal{H}$, or just a regular vector field.

Remark 3.1 The results of this section offer two theoretical insights in deciding potential metric candidates in optimization problems:

(1.) Non-constant ambient metrics may have the same big-O time picture as the constant one. This is the case with the examples in this paper when the constraint and the metrics are given in matrix polynomials or inversion. If the ambient Hessian could be computed efficiently, in many cases the (maybe tedious) Riemannian Hessian expressions, could be computed by operator composition with the same order-of-magnitude time complexity as the Riemannian gradient. This suggests non-constant metrics may be competitive if the improvement in convergence rate is significant. For certain problems involving positive-definite matrices, a non-constant metric is a better option([25]).

(2.) There is a theoretical bound for the cost of computing the gradient, assuming that the metric g is easy to invert. If the complexity of computing g and J is known, it remains to estimate the cost of inverting $\mathsf{J}\,\mathsf{g}^{-1}\,\mathsf{J}^\mathsf{t}$ (or $\mathsf{N}^\mathsf{t}\mathsf{g}\mathsf{N}$). While in our examples these operators are reduced to simple ones that could be inverted efficiently, otherwise, $\mathsf{J}\,\mathsf{g}^{-1}\,\mathsf{J}^\mathsf{t}$ could be solved by a conjugate gradient (CG)-method (one example is in [14]). In that case, the time cost is proportional to the rank of J (or N)) times the cost of each CG step, which can be estimated depending on the problem.

Example 3.1 Let $\mathcal{M}$ be a submanifold of $\mathcal{E}$, defined by a system of equations $C(x) = 0$, where C is a map from $\mathcal{E}$ to $\mathbb{R}^k$ ($x \in \mathcal{M}$). In this case, $\mathsf{J}_C = C_x$ is the Jacobian of C, assumed to be of full rank. The projection of $\omega \in \mathcal{E}$ to the tangent space $T_x\mathcal{M}$ given by

$$\Pi_\mathsf{g}\omega = \omega - C_x^\mathsf{T}(C_x C_x^\mathsf{T})^{-1} C_x \omega \qquad (3.15)$$

and the covariant derivative is given by $\nabla_\xi \eta = \mathrm{D}_\xi \eta + C_x^\mathsf{T}(C_x C_x^\mathsf{T})^{-1}(\mathrm{D}_\xi C_x)\eta$ for two vector fields ξ, η. With $\Gamma(\xi, \eta) = C_x^\mathsf{T}(C_x C_x^\mathsf{T})^{-1}(\mathrm{D}_\xi C_x)\eta$, the Riemannian Hessian

bilinear form is computed from Eq. (3.12), and the Riemannian Hessian operator is:

$$\Pi_{\mathfrak{g}}(\hat{f}_{xx}\xi - (D_{\xi}C_x)^{\mathsf{T}}(C_x C_x^{\mathsf{T}})^{-1}C_x \hat{f}_x).$$

The expression $(C_x C_x^{\mathsf{T}})^{-1}C_x \hat{f}_x$ is often used as an estimate for the Lagrange multiplier, this result was discussed in section 4.9 of [9]. When $C(x) = x^{\mathsf{T}}x - 1$ (the unit sphere) $J_C \omega = x^{\mathsf{T}}\omega$, the Riemannian connection is thus $\nabla_{\xi}\eta = D_{\xi}\eta + x\xi^{\mathsf{T}}\eta$, a well-known result.

Our main interest is to study matrix manifolds. As seen, we need to compute N^{t} or J^{t}. We will review adjoint operators for basic matrix operations.

4 Matrix Manifolds: Inner Products and Adjoint Operators

4.1 Matrices and Adjoints

We will use the trace (Frobenius) inner product on matrix vector spaces considered here. Again, the base field $\mathbb{K}$ is either $\mathbb{R}$ or $\mathbb{C}$. We use the letters m, n, p to denote the dimensions of vector spaces. We will prove results for both the real and complex cases together, as often there is a complex result using the Hermitian transpose corresponding to a real result using the real transpose. The reason is when $\mathbb{C}^{n \times m}$, as a real vector space, is equipped with the real inner product $\operatorname{Re}\operatorname{Tr}(ab^{\mathsf{H}})$ (for $a, b \in \mathbb{C}^{n \times m}$, H is the Hermitian transpose), then the adjoint of the scalar multiplication operator by a complex number c, is the multiplication by the conjugate $\bar{c}$. T o fix some notations, we use the symbol $\operatorname{Tr}_{\mathbb{R}}$ to denote the real part of the trace, so for a matrix $a \in \mathbb{K}^{n \times n}$, $\operatorname{Tr}_{\mathbb{R}} a = \operatorname{Tr} a$ if $\mathbb{K} = \mathbb{R}$ and $\operatorname{Tr}_{\mathbb{R}} a = \operatorname{Re}(\operatorname{Tr} a)$ if $\mathbb{K} = \mathbb{C}$. The symbol t will be used on either an operator, where it specifies the adjoint with respect to these inner products, or to a matrix, where it specifies the corresponding adjoint matrix. When $\mathbb{K} = \mathbb{R}$, we take t to be the real transpose, and when $\mathbb{K} = \mathbb{C}$ we take t to be the hermitian transpose. The inner product of two matrices a, b is $\operatorname{Tr}_{\mathbb{R}}(ab^{\mathsf{t}})$. Recall that we denote by $\operatorname{Sym}_{\mathsf{t},\mathbb{K},n}$ the space of all t-symmetric matrices ($A^{\mathsf{t}} = A$), and $\operatorname{Skew}_{\mathsf{t},\mathbb{K},n}$ the space of all t-antisymmetric matrices ($A^{\mathsf{t}} = -A$). We consider both $\operatorname{Sym}_{\mathsf{t},\mathbb{K},n}$ and $\operatorname{Skew}_{\mathsf{t},\mathbb{K},n}$ inner product spaces under $\operatorname{Tr}_{\mathbb{R}}$. We defined the symmetrizer $\operatorname{sym}_{\mathsf{t}}$ and antisymmetrizer $\operatorname{skew}_{\mathsf{t}}$ in Sect. 2.1, with the usual meaning.

Proposition 4.1 *With the above notations, let A_i, B_i, X be matrices such that the functional $L(X) = \sum_{i=1}^{k} \operatorname{Tr}_{\mathbb{R}}(A_i X B_i) + \operatorname{Tr}_{\mathbb{R}}(C_i X^{\mathsf{t}} D_i)$ is well-formed. We have:*
1. The matrix $\operatorname{xtrace}(L, X) = \sum_{i=1}^{k} A_i^{\mathsf{t}} B_i^{\mathsf{t}} + D_i C_i$ is the unique matrix L_1 such that $\operatorname{Tr}_{\mathbb{R}} L_1 X^{\mathsf{t}} = L(X)$ for all $X \in \mathbb{K}^{n \times m}$ (this is the gradient of L).
2. The matrix $\operatorname{xtrace}^{\operatorname{sym}}(L, X) = \operatorname{sym}_{\mathsf{t}}(\sum_{i=1}^{k} A_i^{\mathsf{t}} B_i^{\mathsf{t}} + D_i C_i)$ is the unique matrix $L_2 \in \operatorname{Sym}_{\mathsf{t},\mathbb{K},n}$ satisfying $\operatorname{Tr}_{\mathbb{R}}(L_2 X^{\mathsf{t}}) = L(X)$ for all $X \in \operatorname{Sym}_{\mathsf{t},\mathbb{K},n}$.
3. The matrix $\operatorname{xtrace}^{\operatorname{skew}}(L, X) = \operatorname{skew}_{\mathsf{t}}(\sum_{i=1}^{k} A_i^{\mathsf{t}} B_i^{\mathsf{t}} + D_i C_i)$ is the unique matrix $L_3 \in \operatorname{Skew}_{\mathsf{t},\mathbb{K},n}$ satisfying $\operatorname{Tr}_{\mathbb{R}}(L_3 X^{\mathsf{t}}) = L(X)$ for all $X \in \operatorname{Skew}_{\mathsf{t},\mathbb{K},n}$.

There is an abuse of notation as xtrace(L, X) is not a function of two variables, but X should be considered a (symbolic) variable and L is a function in X, however, this notation is convenient in symbolic implementation.

Proof We have $\mathrm{Tr}_{\mathbb{R}}(\mathrm{xtrace}(L)X^{\mathrm{t}}) = L(X)$ from

$$\mathrm{Tr}_{\mathbb{R}}(A_i X B_i) = \mathrm{Tr}_{\mathbb{R}}(B_i^{\mathrm{t}} X^{\mathrm{t}} A_i^{\mathrm{t}}) = \mathrm{Tr}_{\mathbb{R}}(A_i^{\mathrm{t}} B_i^{\mathrm{t}} X^{\mathrm{t}})$$

and $\mathrm{Tr}_{\mathbb{R}}(C_i X^{\mathrm{t}} D_i) = \mathrm{Tr}_{\mathbb{R}}(D_i C_i X^{\mathrm{t}})$. Uniqueness follows from the fact that $\mathrm{Tr}_{\mathbb{R}}$ is a non-degenerate bilinear form. The last two statements follow from

- $\mathrm{Tr}_{\mathbb{R}}(\mathrm{xtrace}(L)X^{\mathrm{t}}) = \mathrm{Tr}_{\mathbb{R}}(\mathrm{xtrace}(L)^{\mathrm{t}})X$ if $X^{\mathrm{t}} = X$.
- $\mathrm{Tr}_{\mathbb{R}}(\mathrm{xtrace}(L)^{\mathrm{t}}X) = -\mathrm{Tr}_{\mathbb{R}}(\mathrm{xtrace}(L)^{\mathrm{t}})X$ if $X^{\mathrm{t}} = -X$.

□

Remark 4.1 The index-raising operation/gradient xtrace could be implemented as a symbolic operation on matrix trace expressions, as it involves only linear operations, matrix transpose, and multiplications. It could be used to compute an ambient gradient, for example. For another application, let $\mathcal{M}$ be a manifold with ambient space $\mathbb{K}^{n \times m}$, recall $\mathrm{rhess}_f^{02}(\xi, \eta) = \hat{f}_{YY}(\xi, \eta) - \Gamma(\xi, \eta)$. Assume $\langle \Gamma(\xi, \eta), \hat{f}_Y \rangle_{\mathcal{E}} = \sum_i \mathrm{Tr}_{\mathbb{R}}(A_i \eta B_i) + \mathrm{Tr}_{\mathbb{R}}(C_i \eta^{\mathrm{t}} D_i)$ with A_i, B_i, C_i, D_i are not dependent on η, and identify tangent vectors with their images in $\mathbb{K}^{n \times m}$, we have:

$$\mathrm{rhess}_f^{11}\xi = \Pi_{\mathsf{g}} \mathsf{g}^{-1} \, \mathrm{xtrace}(\mathrm{rhess}_f^{02}(\xi, \eta), \eta),$$

as the inner product of the right-hand side with η is $\mathrm{rhess}_f^{02}(\xi, \eta)$, and the projection ensures it is in the tangent space. If the ambient space is identified with $\mathrm{Sym}_{\mathrm{t}, \mathbb{K}, n}$, $\mathrm{Skew}_{\mathrm{t}, \mathbb{K}, n}$ or a direct sum of matrix spaces, we also have similar statements.

Proposition 4.2 *With the same notations as Proposition* 4.1*:*
1. The adjoint of the left multiplication operator by a matrix $A \in \mathbb{K}^{m \times n}$, sending $X \in \mathbb{K}^{n \times p}$ to $AX \in \mathbb{K}^{m \times p}$ is the left multiplication by A^{t}, sending $Y \in \mathbb{K}^{m \times p}$ to $A^{\mathrm{t}}Y \in \mathbb{K}^{m \times n}$.
2. The adjoint of the right multiplication operator by a matrix $A \in \mathbb{K}^{m \times n}$ from $\mathbb{K}^{p \times m}$ to $\mathbb{K}^{p \times n}$ is the right multiplication by A^{t}.
3. The adjoint of the operator sending $X \mapsto X^{\mathrm{t}}$ for $X \in \mathbb{K}^{m \times m}$ is again the operator $Y \mapsto Y^{\mathrm{t}}$ for $Y \in \mathbb{K}^{m \times m}$. Adjoint is additive, and $(F \circ G)^{\mathrm{t}} = G^{\mathrm{t}} \circ F^{\mathrm{t}}$ for two linear operators F and G.
4. The adjoint of the left multiplication operator by $A \in \mathbb{K}^{p \times n}$ sending $X \in \mathrm{Sym}_{\mathrm{t}, \mathbb{K}, p}$ to $AX \in \mathbb{K}^{p \times n}$ is the operator sending $Y \mapsto \frac{1}{2}(A^{\mathrm{t}}Y + Y^{\mathrm{t}}A)$ for $Y \in \mathbb{K}^{p \times n}$. Conversely, the adjoint of the operator $Y \mapsto \frac{1}{2}(A^{\mathrm{t}}Y + Y^{\mathrm{t}}A) \in \mathrm{Sym}_{\mathrm{t}, \mathbb{K}, p}$ is the operator $X \mapsto AX$.
5. The adjoint of the left multiplication operator by $A \in \mathbb{K}^{p \times n}$ sending $X \in \mathrm{Skew}_{\mathrm{t}, \mathbb{K}, p}$ to $AX \in \mathbb{K}^{p \times n}$ is the operator sending $Y \mapsto \frac{1}{2}(A^{\mathrm{t}}Y - Y^{\mathrm{t}}A)$ for $Y \in \mathbb{K}^{p \times n}$. Conversely, the adjoint of the operator $Y \mapsto \frac{1}{2}(A^{\mathrm{t}}Y - Y^{\mathrm{t}}A) \in \mathrm{Skew}_{\mathrm{t}, \mathbb{K}, p}$ is the operator $X \mapsto AX$.
6. Adjoint is linear on the space of operators. If F_1 and F_2 are two linear operators from a space V to two spaces W_1 and W_2, then the adjoint of the direct sum operator

(operator sending X to $\begin{bmatrix} F_1X & F_2X \end{bmatrix}$) is the map sending $\begin{bmatrix} A \\ B \end{bmatrix}$ to $F_1^{\mathsf{t}}A + F_2^{\mathsf{t}}B$. Adjoint of the map sending $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ to FX_1 is the map $Y \mapsto \begin{bmatrix} F^{\mathsf{t}}Y \\ 0 \end{bmatrix}$, and more generally a map sending a row block X_i of a matrix X to FX_i is the map sending Y to a matrix where the i-th block is $F^{\mathsf{t}}Y$, and zero outside of this block.

Most of the proof is just a simple application of trace calculus. For the first statement, the real case follows from $\operatorname{Tr}(Aab^{\mathsf{T}}) = \operatorname{Tr}(a(A^{\mathsf{T}}b)^{\mathsf{T}})$, and $\operatorname{Tr}_{\mathbb{R}}(Aab^{\mathsf{H}}) = \operatorname{Tr}_{\mathbb{R}}(a(A^{\mathsf{H}}b)^{\mathsf{H}})$ gives us the complex case. Statement 2. is proved similarly, statement 4 is standard. Statements 4. and 5. are checked by direct substitution, and 6. is just the operator version of the corresponding matrix statement, observing for example:

$$\operatorname{Tr}_{\mathbb{R}}(F_1XA^{\mathsf{t}} + F_2XB^{\mathsf{t}}) = \operatorname{Tr}_{\mathbb{R}}((F_1^{\mathsf{t}}A + F_2^{\mathsf{t}}B)X^{\mathsf{t}}).$$

5 Application to Stiefel Manifold

The Stiefel manifold $\operatorname{St}_{\mathbb{K},d,n}$ on $\mathcal{E} = \mathbb{K}^{n\times d}$ is defined by the equation $Y^{\mathsf{t}}Y = I_d$, where the tangent space at a point Y consists of matrices η satisfying $\eta^{\mathsf{t}}Y + Y\eta^{\mathsf{t}} = 0$. When $\mathbb{K} = \mathbb{R}$ we assume $d < n$, so the manifold is connected. We apply the results of Sect. 3 for the full tangent bundle $\mathcal{H} = T\operatorname{St}_{\mathbb{K},d,n}$. We can consider an ambient metric:

$$\mathsf{g}(Y)\omega = \alpha_0\omega + (\alpha_1 - \alpha_0)YY^{\mathsf{t}}\omega = \alpha_0(I_n - YY^{\mathsf{t}})\omega + \alpha_1 YY^{\mathsf{t}}\omega \tag{5.1}$$

for $\omega \in \mathcal{E} = \mathbb{K}^{n\times d}$. It is easy to see $\omega_0 - YY^{\mathsf{t}}\omega_0$ is an eigenvector of $\mathsf{g}(Y)$ with eigenvalue α_0, and $YY^{\mathsf{t}}\omega_1$ is an eigenvector with eigenvalue α_1, for any $\omega_0, \omega_1 \in \mathcal{E}$, and these are the only eigenvalues and vectors. Hence, $\mathsf{g}(Y)^{-1}\omega = \alpha_0^{-1}(I_n - YY^{\mathsf{t}})\omega + \alpha_1^{-1}YY^{\mathsf{t}}\omega$ and g is a Riemannian metric if α_0, α_1 are positive. We can describe the tangent space as a nullspace of $\mathsf{J}(Y)$ with $\mathsf{J}(Y)\omega = \omega^{\mathsf{t}}Y + Y^{\mathsf{t}}\omega \in \mathcal{E}_{\mathsf{J}} := \operatorname{Sym}_{\mathsf{t},\mathbb{K},d}$. We will evaluate everything at Y, so we will write J and g instead of $\mathsf{J}(Y)$ and $\mathsf{g}(Y)$, etc. By Proposition 4.2, $\mathsf{J}^{\mathsf{t}}a = (aY^{\mathsf{t}})^{\mathsf{t}} + Ya = 2Ya$ for $a \in \mathcal{E}_{\mathsf{J}}$. We have $\mathsf{g}^{-1}\mathsf{J}^{\mathsf{t}}a = \alpha_0^{-1}2Ya + (\alpha_1^{-1} - \alpha_0^{-1})2Ya = 2\alpha_1^{-1}Ya$. Thus $\mathsf{J}\,\mathsf{g}^{-1}\mathsf{J}^{\mathsf{t}}a = \mathsf{J}(2\alpha_1^{-1}Ya) = 4\alpha_1^{-1}a$ and by Proposition 3.1:

$$\Pi_{\mathsf{g}}v = \Pi_{\mathsf{g}}v = v - 2\alpha_1^{-1}\frac{\alpha_1}{4}(Yv^{\mathsf{t}}Y + YY^{\mathsf{t}}v) = v - \frac{1}{2}(Yv^{\mathsf{t}}Y + YY^{\mathsf{t}}v). \tag{5.2}$$

In this case, the ambient gradient $\hat{f}_Y$ is the matrix of partial derivatives of an extension of f on the ambient space $\mathbb{K}^{n\times d}$. More conveniently, using the eigenspaces of g, $\Pi_{\mathsf{g}}v = (I_n - YY^{\mathsf{t}})v + Y\operatorname{skew}_{\mathsf{t}}Y^{\mathsf{t}}v$ and $\mathsf{g}^{-1}\hat{f}_Y = \alpha_0^{-1}(I_n - YY^{\mathsf{t}})\hat{f}_Y + \alpha_1^{-1}YY^{\mathsf{t}}\hat{f}_Y$

$$\begin{aligned}
\Pi_{\mathsf{g}}\mathsf{g}^{-1}\hat{f}_Y &= \alpha_0^{-1}(I_n - YY^{\mathsf{t}})\hat{f}_Y + \alpha_1^{-1}Y\operatorname{skew}_{\mathsf{t}}(Y^{\mathsf{t}}\hat{f}_Y) \\
&= \alpha_0^{-1}\hat{f}_Y + \frac{\alpha_1^{-1} - 2\alpha_0^{-1}}{2}YY^{\mathsf{t}}\hat{f}_Y - \frac{\alpha_1^{-1}}{2}Y\hat{f}_Y^{\mathsf{t}}Y.
\end{aligned}$$

If ξ and η are vector fields, $(D_\xi g)\eta = (\alpha_1 - \alpha_0)(\xi Y^t + Y\xi^t)\eta$. Using Proposition 4.1, we can take the cross term $\mathcal{X}(\xi, \eta) = (\alpha_1 - \alpha_0)(\xi\eta^t + \eta\xi^t)Y$, thus:

$$K(\xi, \eta) = \frac{\alpha_1 - \alpha_0}{2}((\xi Y^t\eta + \eta Y^t\xi) + Y(\xi^t\eta + \eta^t\xi) - (\xi\eta^t + \eta\xi^t)Y).$$

By the tangent condition, $(\xi Y^t\eta + \eta Y^t\xi) = -(\xi\eta^t + \eta\xi^t)Y$, hence $K(\xi, \eta) = \frac{\alpha_1 - \alpha_0}{2} F$ with $F = Y(\xi^t\eta + \eta^t\xi) - 2(\xi^t\eta + \eta^t\xi)Y$, we see $Y^t F$ is symmetric so $\text{skew}_t Y^t F = 0$, therefore

$$\Pi_g g^{-1} F = \alpha_0^{-1}(I_n - YY^t)F = -2\alpha_0^{-1}(I_n - YY^t)(\xi^t\eta + \eta^t\xi)Y,$$

$$\Pi_g g^{-1} K(\xi, \eta) = \frac{\alpha_0 - \alpha_1}{\alpha_0}(I_n - YY^t)(\xi\eta^t + \eta\xi^t)Y. \tag{5.3}$$

Using $g^{-1} J^t (J g^{-1} J^t)^{-1}(D_\xi J)\eta = \frac{1}{2}Y(\xi^t\eta + \eta^t\xi)$ to evaluate Eq. (3.13), the connection for two vector fields ξ, η is:

$$\nabla_\xi \eta = D_\xi \eta + \frac{1}{2}Y(\xi^t\eta + \eta^t\xi) + \frac{\alpha_0 - \alpha_1}{\alpha_0}(I_n - YY^t)(\xi\eta^t + \eta\xi^t)Y. \tag{5.4}$$

With $\Pi_0 = (I_n - YY^t)$ and let $\hat{f}_{YY}$ be the ambient Hessian, from Eq. (3.12):

$$\text{rhess}_f^{02}(\xi, \eta) = \hat{f}_{YY}(\xi, \eta) - \text{Tr}_\mathbb{R}(\hat{f}_Y^t\{\frac{1}{2}Y(\xi^t\eta + \eta^t\xi) + \frac{\alpha_0 - \alpha_1}{\alpha_0}\Pi_0(\xi\eta^t + \eta\xi^t)Y\}). \tag{5.5}$$

By Remark 4.1, $\text{rhess}_f^{11}\xi$ is $\Pi_g g^{-1}$ xtrace($\text{rhess}_{\xi,\eta}^{02}, \eta$), thus

$$\text{rhess}_f^{11}\xi = \Pi_{Y,g} g^{-1}(\hat{f}_{YY}\xi - \frac{1}{2}\xi(\hat{f}_Y^t Y + Y^t\hat{f}_Y) - \frac{\alpha_0 - \alpha_1}{\alpha_0}(\Pi_0 \hat{f}_Y Y^t + Y\hat{f}_Y^t \Pi_0)\xi). \tag{5.6}$$

We note the term inside $\Pi_{Y,g} g^{-1}$ can be modified by any expression sent to zero by $\Pi_{Y,g} g^{-1}$. The case $\alpha_0 = 1, \alpha_1 = \frac{1}{2}$ correspond to the canonical metric on a Stiefel manifold, where the connection is given by formula 2.49 of [9], in a slightly different form, but we could show they are the same by noting $YY^t(\xi\eta^t + \eta\xi^t)Y = Y(\xi^t YY^t\eta + \eta^t YY^t\xi)$ using the tangent constraint. The case $\alpha_0 = \alpha_1 = 1$ corresponds to the constant trace metric where we do not need to compute K. This family of metrics has been studied in [11], where a closed-form geodesic formula is provided. In [20] we also provide efficient closed-form geodesic formulas similar to those in [9].

6 Quotients of a Stiefel Manifold and Flag Manifolds

Continuing with the setup in the previous section, consider a Stiefel manifold $St_{\mathbb{K},d,n}$ (we will assume $0 < d < n$). The metric induced by the operator g in Eq. (5.1),

$\alpha_0 \operatorname{Tr}_{\mathbb{R}} \omega_1^t \omega_2 + (\alpha_1 - \alpha_0) \operatorname{Tr} \omega_1^t Y Y^t \omega_2$ with $Y \in \operatorname{St}_{\mathbb{K},d,n}, \omega_1, \omega_2 \in \mathcal{E}$ is preserved if we replace Y, ω_1, ω_2 by $YU, \omega_1 U, \omega_2 U$, for $U^t U = I_d$, or if we define the t-orthogonal group by $U_{\mathbb{K},d} := \{U \in \mathbb{K}^{d \times d} | U^t U = I_d\}$ then this is a group of isometries of g. Therefore, any subgroup G of $U_{\mathbb{K},d}$ acts on $\operatorname{St}_{\mathbb{K},d,n}$ by right-multiplication also preserves the metric, and if G is compact, we can consider the quotient manifold $\operatorname{St}_{\mathbb{K},d,n}/G$, identifying $Y \in \operatorname{St}_{\mathbb{K},d,n}$ with YU for $U \in G$.

If the cost function f on the Stiefel manifold is invariant when applying transformations by a group G, it may be advantageous to consider optimization on a quotient manifold. The case of the Rayleigh quotient cost function $\operatorname{Tr}_{\mathbb{R}} Y^t A Y$ for $Y \in \operatorname{St}_{\mathbb{K},d,n}$ and A is a positive-definite matrix is well-known. As the cost function is invariant if we replace Y by YU for $U \in G = U_{\mathbb{K},d}$, we can optimize over the Grassmann manifolds $\operatorname{St}_{\mathbb{K},d,n}/U_{\mathbb{K},d}$. When $Y \in \operatorname{St}_{\mathbb{K},d,n}$ is divided into two column blocks $Y = [Y_1|Y_2]$ of $d_1 + d_2 = d$ columns, the cost function $f_1 = \operatorname{Tr}_{\mathbb{R}} Y_1^t A_1 Y_1 + \operatorname{Tr}_{\mathbb{R}} Y_2^t A_2 Y_2$ for two positive-definite matrices $A_1, A_2 \in \mathbb{K}^{n \times n}$ is invariant if we replace Y_1, Y_2 by $Y_1 U_1, Y_2 U_2$ for $U_1 \in U_{\mathbb{K},d_1}, U_2 \in U_{\mathbb{K},d_2}$, while the cost function $f_2 = \operatorname{Tr}_{\mathbb{R}} Y_1^t A_1 Y_1 + \operatorname{Tr}_{\mathbb{R}} Y_2 B_2 Y_2^t$ for positive-definite matrices $A_1 \in \mathbb{K}^{n \times n}, B_2 \in \mathbb{K}^{d_2 \times d_2}$ is invariant if we replace Y_1 by $Y_1 U_1$ only, for a generic B_2. In the first case, we can optimize over $\operatorname{St}_{\mathbb{K},d,n}/(U_{\mathbb{K},d_1} \times U_{\mathbb{K},d_2})$ and in the second case we can optimize over $\operatorname{St}_{\mathbb{K},d,n}/U_{\mathbb{K},d_1}$. We will define an optimization framework for a quotient of $\operatorname{St}_{\mathbb{K},d,n}$ by a group G, $\{I_d\} \subset G \subset U_{\mathbb{K},d}$, where G consists of $q + 1$ blocks diagonal blocks, $q \geq 0$, with at most one block could be trivial. The case $G = \{I_d\}$ corresponds to the Stiefel manifold, $G = U_{\mathbb{K},d}$ corresponds to the Grassmann manifold and the intermediate case includes flag manifolds. Background materials for optimization on flag manifolds are in [21, 28], but the examples just described and the review below should be sufficient to understand the setup and the results. We generalize the formula for rhess^{02} in [28, Proposition 25] to the full family of metrics in [11] and provide a formula for rhess^{11}.

Let us describe the group G of block-diagonal matrices considered here. Assume there is a sequence of positive integers $\hat{d} = \{d_1, \cdots, d_q\}, d_i > 0$ for $1 \leq i \leq q$, such that $\sum_{i=1}^q d_i \leq d$. Set $d_{q+1} = d - \sum_{i=1}^q d_i$, thus $d_{q+1} \geq 0$. This sequence allows a partition of a matrix $A \in \mathbb{K}^{d \times d}$ to $(q + 1) \times (q + 1)$ blocks $A_{[ij]} \in \mathbb{K}^{d_i \times d_j}$, $1 \leq i, j \leq q + 1$. The right-most or bottom blocks correspond to i or j equals to $q+1$ are empty when $d_{q+1} = 0$. Consider the subgroup $G = U_{\mathbb{K},\hat{d}} = U_{\mathbb{K},d_1} \times U_{\mathbb{K},d_2} \cdots \times U_{\mathbb{K},d_q} \times \{I_{d_{q+1}}\}$ of $U_{\mathbb{K},d}$ of block-diagonal matrices U, with the i-th diagonal block from the top $U_{[ii]} \in U_{\mathbb{K},d_i}, 1 \leq i \leq q$, and $U_{[q+1,q+1]} = I_{d_{q+1}}$. An element $U \in U_{\mathbb{K},\hat{d}}$ has the form

$$U = \operatorname{diag}(U_{[11]}, \cdots, U_{[qq]}, I_{d_{q+1}}) \text{ for } U_{[ii]} \in U_{\mathbb{K},d_i}, 1 \leq i \leq q$$

and transforms $Y = [Y_1| \cdots |Y_q|Y_{q+1}]$ to $[Y_1 U_1| \cdots |Y_q U_q|Y_{q+1}]$. When $q = 0$, we define $\hat{d} = \emptyset$ and $U_{\mathbb{K},\hat{d}} = \{I_d\}$. We will consider the manifold $\operatorname{St}_{\mathbb{K},d,n}/G$ for $G = U_{\mathbb{K},\hat{d}}$. Thus, when $\hat{d} = \emptyset$, this quotient is the Stiefel manifold $\operatorname{St}_{\mathbb{K},d,n}$ itself, when $\hat{d} = \{d\}$, it is the Grassmann manifold. When $d_{q+1} = 0$ i.e. $\sum_{i=1}^q d_i = d$, the quotient is called a flag manifold, denoted by $\operatorname{Flag}(d_1, \cdots, d_q; n, \mathbb{K})$. In the example above for $q = 2$, if $d_1 + d_2 = d$, $\operatorname{St}_{\mathbb{K},d,n}/(U_{\mathbb{K},d_1} \times U_{\mathbb{K},d_2})$ is a flag manifold (we do not

have a special name if $d_1 + d_2 < d$). Therefore, these quotients could be considered as intermediate objects between a Stiefel and a Grassmann manifold, as we will soon see more clearly.

Define the operator symf acting on $\mathbb{K}^{d \times d}$, sending $A \in \mathbb{K}^{d \times d}$ to A_{symf} such that $(A_{\text{symf}})_{[ij]} = \frac{1}{2}(A_{[ij]} + A_{[ji]}^{\mathsf{t}})$ if $1 \leq i \neq j \leq q + 1$ or $i = j = q + 1$, and $(A_{\text{symf}})_{[ii]} = A_{[ii]}$ if $1 \leq i \leq q$. Thus, symf preserves the diagonal blocks for $1 \leq i \leq q$, but symmetrizes the off-diagonal blocks and the $q + 1$-th diagonal block. The following illustrates the operation when $q = 2$ for $A = (A_{[ij]}) \in \mathbb{K}^{d \times d}$.

$$A_{\text{symf}} = \frac{1}{2} \begin{bmatrix} 2A_{[11]} & A_{[12]} + A_{[21]}^{\mathsf{t}} & A_{[13]} + A_{[31]}^{\mathsf{t}} \\ A_{[21]} + A_{[12]}^{\mathsf{t}} & 2A_{[22]} & A_{[23]} + A_{[32]}^{\mathsf{t}} \\ A_{[31]} + A_{[13]}^{\mathsf{t}} & A_{[32]} + A_{[23]}^{\mathsf{t}} & A_{[33]} + A_{[33]}^{\mathsf{t}} \end{bmatrix}.$$

For the case $\hat{d} = \varnothing$ of the full Stiefel manifold, symf is just sym_{t} and for the case $\hat{d} = \{d\}$ of the Grassmann manifold, symf is the identity map. We show these quotients share similar Riemannian optimization settings.

Theorem 6.1 *With the metric in Eq. (5.1), the horizontal space $\mathcal{H}_Y$ at $Y \in \text{St}_{\mathbb{K},d,n}$ of the quotient $\text{St}_{\mathbb{K},d,n} \to \text{St}_{\mathbb{K},d,n}/U_{\mathbb{K},\hat{d}}$ consists of matrices $\omega \in \mathcal{E} := \mathbb{K}^{n \times d}$ such that*

$$(Y^{\mathsf{t}}\omega)_{\text{symf}} = 0, \tag{6.1}$$

or equivalently, $Y^{\mathsf{t}}\omega$ is t-antisymmetric with first q diagonal blocks $((Y^{\mathsf{t}}\omega)_{\text{symf}})_{[ii]}$ vanish for $1 \leq i \leq q$. For $\omega \in \mathcal{E} = \mathbb{K}^{n \times d}$, the projection $\Pi_{\mathcal{H}}$ from $\mathcal{E}$ to $\mathcal{H}_Y$ and the Riemannian gradient are given by

$$\Pi_{\mathcal{H}}\omega = \omega - Y(Y^{\mathsf{t}}\omega)_{\text{symf}}, \tag{6.2}$$

$$\Pi_{\mathcal{H}}\mathsf{g}^{-1}\hat{f}_Y = \alpha_0^{-1}\hat{f}_Y + (\alpha_1^{-1} - \alpha_0^{-1})YY^{\mathsf{t}}\hat{f}_Y - \alpha_1^{-1}Y(Y^{\mathsf{t}}\hat{f}_Y)_{\text{symf}}. \tag{6.3}$$

Let $\Pi_0 = I_n - YY^{\mathsf{t}}$. For two vector fields ξ, η, the horizontal lift of the Levi-Civita connection and Riemannian Hessians are given by

$$\Pi_{\mathcal{H}}\nabla_\xi\eta = D_\xi\eta + Y(\xi^{\mathsf{t}}\eta)_{\text{symf}} + \frac{\alpha_0 - \alpha_1}{\alpha_0}\Pi_0(\xi\eta^{\mathsf{t}} + \eta\xi^{\mathsf{t}})Y, \tag{6.4}$$

$$\text{rhess}_f^{02}(\xi, \eta) = \hat{f}_{YY}(\xi, \eta) - \text{Tr}_{\mathbb{R}}(\hat{f}_Y^{\mathsf{t}}\{Y(\xi^{\mathsf{t}}\eta)_{\text{symf}} + \frac{\alpha_0 - \alpha_1}{\alpha_0}\Pi_0(\xi\eta^{\mathsf{t}} + \eta\xi^{\mathsf{t}})Y\}), \tag{6.5}$$

$$\text{rhess}_f^{11}\xi = \Pi_{\mathcal{H}}\mathsf{g}^{-1}(\hat{f}_{YY}\xi - \xi(Y^{\mathsf{t}}\hat{f}_Y)_{\text{symf}} - \frac{\alpha_0 - \alpha_1}{\alpha_0}(\Pi_0\hat{f}_Y Y^{\mathsf{t}} + Y\hat{f}_Y^{\mathsf{t}}\Pi_0)\xi). \tag{6.6}$$

Proof First we note that symf is a self-adjoint operator, as both the identity operator on the first q diagonal blocks and symmetrize operator on the remaining blocks are self-adjoint. The orbit of $Y \in \text{St}_{\mathbb{K},d,n}$ under the action of $U_{\mathbb{K},\hat{d}}$ is $YU_{\mathbb{K},\hat{d}}$, thus

the vertical space consists of matrices of the form YD with D is block-diagonal, t-skewsymmetric and $D_{[(q+1),(q+1)]} = 0$. Since $\mathsf{g}YD = \alpha_1 YD$, a horizontal vector ω satisfies $\text{skew}_{\mathsf{t}}(Y^{\mathsf{t}}\omega) = 0$ and $\text{Tr}_{\mathbb{R}}(\omega^{\mathsf{t}}YD) = 0$. This shows the first q diagonal blocks of $Y^{\mathsf{t}}\omega$ are zero, hence $(Y^{\mathsf{t}}\omega)_{\text{symf}} = 0$.

For the projection, we proceed like the Stiefel case, with the map $\mathsf{J}\,\omega = (Y^{\mathsf{t}}\omega)_{\text{symf}}$, mapping $\mathcal{E}$ to $\mathcal{E}_{\mathsf{J}} = \{A \in \mathbb{K}^{d \times d} | A_{[ij]} = A_{[ji]}^{\mathsf{t}}, 1 \leq i \neq j \leq q+1 \text{ or } i = j = q+1\}$. Since symf is self-adjoint, $\mathsf{J}^{\mathsf{t}} A = Y A_{\text{symf}} = Y A$ for $A \in \mathcal{E}_{\mathsf{J}}$. From here we get $(\mathsf{J}\,\mathsf{g}^{-1}\mathsf{J}^{\mathsf{t}})A = \alpha_1^{-1} A$ and Eq. (6.2) follows. Equation 6.3 is a substitution of $\mathsf{g}^{-1}\hat{f}_Y$ to Eq. (6.2), noting $(Y^{\mathsf{t}}(\mathsf{g}^{-1}\hat{f}_Y))_{\text{symf}} = \alpha_1^{-1}(Y^{\mathsf{t}}\hat{f}_Y)_{\text{symf}}$, using the eigen decomposition of g.

For the Levi-Civita connection, we use Eq. (3.10). For two horizontal vector fields ξ, η, $(\mathrm{D}_\xi \Pi_{\mathcal{H}})\eta = -\xi(Y^{\mathsf{t}}\eta)_{\text{symf}} - Y(\xi^{\mathsf{t}}\eta)_{\text{symf}} = -Y(\xi^{\mathsf{t}}\eta)_{\text{symf}}$ and $\Pi_{\mathcal{H}} = \Pi_{\mathcal{H}}\Pi$, where Π is the Stiefel projection Eq. (5.2), hence $\Pi_{\mathcal{H}}\mathsf{g}^{-1}\mathrm{K}(\xi, \eta) = \Pi\mathsf{g}^{-1}\mathrm{K}(\xi, \eta) - Y(Y^{\mathsf{t}}\Pi\mathsf{g}^{-1}\mathrm{K}(\xi, \eta))_{\text{symf}}$. The last term vanishes from Eq. (5.3), and we have Eq. (6.4).

Equation (6.5) follows from Eq. (3.12). We derive Eq. (6.6) from remark 4.1 and self-adjointness of symf, expanding $\text{Tr}_{\mathbb{R}}\,\hat{f}_Y Y(\xi^{\mathsf{t}}\eta)_{\text{symf}}$ to

$$\text{Tr}_{\mathbb{R}}(Y^{\mathsf{t}}\hat{f}_Y)^{\mathsf{t}}(\xi^{\mathsf{t}}\eta)_{\text{symf}} = \text{Tr}_{\mathbb{R}}(Y^{\mathsf{t}}\hat{f}_Y)_{\text{symf}}(\xi^{\mathsf{t}}\eta)^{\mathsf{t}} = \text{Tr}_{\mathbb{R}}\,\xi(Y^{\mathsf{t}}\hat{f}_Y)_{\text{symf}}\eta^{\mathsf{t}}.$$

$\square$

7 Positive-Definite Matrices

Consider the manifold $\mathrm{S}_{\mathbb{K},n}^+$ of t-symmetric positive-definite matrices in $\mathbb{K}^{n \times n}$. In our approach, we take $\mathcal{E} = \mathbb{K}^{n \times n}$ with its Frobenius inner product. The metric g is $\langle \xi, \mathsf{g}\eta \rangle_{\mathcal{E}} = \text{Tr}(\xi Y^{-1}\eta Y^{-1})$, with the metric operator $\mathsf{g} : \eta \mapsto Y^{-1}\eta Y^{-1}$ for two vector fields ξ, η. The full tangent bundle $\mathcal{H} = T\mathrm{S}_{\mathbb{K},n}^+$ is identified fiber-wise with the nullspace of the operator $\mathsf{J} : \eta \mapsto \mathsf{J}\eta = \eta - \eta^{\mathsf{t}}$, with $\mathcal{E}_{\mathsf{J}} = \text{Skew}_{\mathsf{t},\mathbb{K},n}$. By item 5 in Proposition 4.2, we have $\mathsf{J}^{\mathsf{t}} a = 2a$ where a is a t-antisymmetric matrix. From here $\mathsf{J}\,\mathsf{g}^{-1}\mathsf{J}^{\mathsf{t}} a = 4YaY$ and (write Π for Π_{g}):

$$\Pi\eta = \eta - \mathsf{g}^{-1}\mathsf{J}^{\mathsf{t}}(\mathsf{J}\,\mathsf{g}^{-1}\mathsf{J}^{\mathsf{t}})^{-1}\mathsf{J}\eta = \eta - 2Y(\frac{1}{4}Y^{-1}(\eta - \eta^{\mathsf{t}})Y^{-1})Y = \frac{1}{2}(\eta + \eta^{\mathsf{t}})$$
$$= \text{sym}_{\mathsf{t}}\eta.$$

Thus, the Riemannian gradient is $\Pi\mathsf{g}^{-1}\hat{f}_Y = \frac{1}{2}Y(\hat{f}_Y + \hat{f}_Y^{\mathsf{t}})Y$. Next, we compute $(\mathrm{D}_\xi \mathsf{g})\eta = \mathrm{D}_{\xi|\eta \text{ constant}}(Y^{-1}\eta Y^{-1}) = -Y^{-1}\xi Y^{-1}\eta Y^{-1} - Y^{-1}\eta Y^{-1}\xi Y^{-1}$, where we keep η constant in the derivative, as we evaluate $\mathrm{D}_\xi \mathsf{g}$ as an operator-valued function. From here, $(\mathrm{D}_\eta \mathsf{g})\xi = (\mathrm{D}_\xi \mathsf{g})\eta$. We note for three vector fields ξ, η, ξ_0

$$\text{Tr}_{\mathbb{R}}(Y^{-1}\xi_0 Y^{-1}\eta Y^{-1} + Y^{-1}\eta Y^{-1}\xi_0 Y^{-1})\xi = \text{Tr}_{\mathbb{R}}\,\xi_0(Y^{-1}\eta Y^{-1}\xi Y^{-1} + Y^{-1}\xi Y^{-1}\eta Y^{-1}).$$

Thus, we can take $\mathcal{X}(\xi, \eta) = -Y^{-1}\eta Y^{-1}\xi Y^{-1} - Y^{-1}\xi Y^{-1}\eta Y^{-1}$ and

$$\Gamma(\xi, \eta) = -(\mathrm{D}_\xi \Pi)\eta - \frac{1}{2}Y(Y^{-1}\eta Y^{-1}\xi Y^{-1} + Y^{-1}\xi Y^{-1}\eta Y^{-1})Y,$$

$$\nabla_\xi \eta = \mathrm{D}_\xi \eta - \frac{1}{2}(\xi Y^{-1}\eta + \eta Y^{-1}\xi). \tag{7.1}$$

Hence, the Riemannian Hessian bilinear form $\mathsf{rhess}^{02}(\xi, \eta)$ is

$$\hat{f}_{YY}(\xi, \eta) + \frac{1}{2}\mathrm{Tr}_{\mathbb{R}}((\xi Y^{-1}\eta + \eta Y^{-1}\xi)\hat{f}_Y) = \hat{f}_{YY}(\xi, \eta) + \frac{1}{2}\mathrm{Tr}_{\mathbb{R}}((\hat{f}_Y \xi Y^{-1} + Y^{-1}\xi \hat{f}_Y)\eta). \tag{7.2}$$

Using a symmetric version of remark 4.1, $\mathsf{rhess}_f^{11}\xi = \Pi_{\mathsf{g}}\mathsf{g}^{-1}\mathsf{sym}_{\mathsf{t}}(\hat{f}_{YY}\xi + \frac{1}{2}(\hat{f}_Y \xi Y^{-1} + Y^{-1}\xi \hat{f}_Y))$. We get the following formula, as in [8]:

$$\mathsf{rhess}_f^{11}\xi = Y\mathsf{sym}_{\mathsf{t}}(\hat{f}_{YY}\xi)Y + \mathsf{sym}_{\mathsf{t}}(\xi\mathsf{sym}_{\mathsf{t}}(\hat{f}_Y)Y). \tag{7.3}$$

8 A Family of Metrics for the Manifold of Positive-Semidefinite Matrices of Fixed Rank

In [7], the authors defined a family of metrics on the manifold $\mathrm{S}_{\mathbb{K},p,n}^+$ of positive-semidefinite matrices of size n and rank p for the case $\mathbb{K} = \mathbb{R}$. Each such matrix will have the form YPY^{t} with $Y \in \mathrm{St}_{\mathbb{K},p,n}$ ($Y^{\mathsf{t}}Y = I_p$) and P is positive-definite of size $p \times p$, up to the equivalent relation $(Y, P) \sim (YU, U^{\mathsf{t}}PU)$ for a matrix $U \in \mathrm{U}_{\mathbb{K},p}$, (that means $U \in \mathbb{K}^{p \times p}$ and $UU^{\mathsf{t}} = I_p$). So the manifold $\mathrm{S}_{\mathbb{K},p,n}^+$ could be identified with the quotient space $(\mathrm{St}_{\mathbb{K},p,n} \times \mathrm{S}_{\mathbb{K},p}^+)/\mathrm{U}_{\mathbb{K},p}$ of the product of the Stiefel manifold $\mathrm{St}_{\mathbb{K},p,n}$ and the manifold of positive-definite matrices $\mathrm{S}_{\mathbb{K},p}^+$ over the t-orthogonal group $\mathrm{U}_{\mathbb{K},p}$. (The paper actually uses $R = P^{\frac{1}{2}}$ to parametrize the space.) From our point of view, the ambient space is $\mathcal{E} = \mathbb{K}^{n \times p} \times \mathbb{K}^{p \times p}$, and the tangent space is identified with the image of the operator N_1 from $\mathbb{K}^{(n-p) \times p} \times \mathrm{Sym}_{\mathsf{t},\mathbb{K},p}$ to $\mathcal{E}$, $\mathrm{N}_1 : (B, D) \mapsto (Y_\perp B, D)$, where the matrix $(Y|Y_\perp)$ is t-orthogonal. On the tangent space, [7] uses the metric $\mathrm{Tr}(BB^{\mathsf{t}}) + k\,\mathrm{Tr}(DP^{-1}DP^{-1})$ for a positive number k. The action of the group gives us the vertical vectors $(Y\mathsf{q}, P\mathsf{q} - \mathsf{q}P)$ for a t-antisymmetric matrix q such that $\mathsf{q} + \mathsf{q}^{\mathsf{t}} = 0$. In the paper, the image of N_1 transverses but not orthogonal to the vertical vectors, and no second-order method is provided. We modify this approach, using a Riemannian quotient metric to provide a second-order method. In the following, the projection to the horizontal space is denoted by $\Pi_{\mathcal{H}}$. The horizontal lift of the Levi-Civita connection (denote by $\nabla^{\mathcal{H}}$) is $\Pi_{\mathcal{H}}\nabla$.

Theorem 8.1 *Let* $\mathcal{M} = \mathrm{St}_{\mathbb{K},p,n} \times \mathrm{S}_{\mathbb{K},p}^+ \subset \mathcal{E} := \mathbb{K}^{n \times p} \times \mathbb{K}^{p \times p}$ *be an embedded ambient space of* $\mathcal{M}$. *Identifying the manifold* $\mathrm{S}_{\mathbb{K},p,n}^+$ *of positive-semidefinite matrices with* $\mathcal{B} = (\mathrm{St}_{\mathbb{K},p,n} \times \mathrm{S}_{\mathbb{K},p}^+)/\mathrm{U}_{\mathbb{K},p}$, *where the pair* $S = (Y, P)$ *represents the matrix*

YPY^t with $Y \in \mathrm{St}_{\mathbb{K},p,n}$, $P \in \mathrm{S}^+_{\mathbb{K},p}$, and the action of $U \in \mathrm{U}_{\mathbb{K},p}$ sends (Y, P) to $(YU, U^t PU)$. The self-adjoint metric operator

$$\mathsf{g}(S)(\omega_Y, \omega_P) = \mathsf{g}(\omega_Y, \omega_P) = (\alpha_0 \omega_Y + (\alpha_1 - \alpha_0)YY^t \omega_P, \beta P^{-1}\omega_P P^{-1}) \quad (8.1)$$

for $\omega = (\omega_Y, \omega_p) \in \mathcal{E} = \mathbb{K}^{n \times p} \times \mathbb{K}^{p \times p}$ defines the inner product $\mathrm{Tr}_{\mathbb{R}}(\alpha_0 \omega_Y^t \omega_Y + (\alpha_1 - \alpha_0)\omega_Y^t YY^t \omega_Y + \beta \omega_P P^{-1}\omega_P P^{-1})$ on $\mathcal{E}$, which induces a metric on $\mathrm{St}_{\mathbb{K},p,n} \times \mathrm{S}^+_{\mathbb{K},p}$, invariant under the action of $\mathrm{U}_{\mathbb{K},p}$ and induces a quotient metric on $\mathrm{S}^+_{\mathbb{K},p,n}$. Its tangent bundle $T\mathrm{S}^+_{\mathbb{K},p,n}$ lifts to the subbundle $\mathcal{H} \subset T(\mathrm{St}_{\mathbb{K},p,n} \times \mathrm{S}^+_{\mathbb{K},p})$ horizontal to the group action, where a vector $\eta = (\eta_Y, \eta_P) \in \mathcal{E}$ is a horizontal tangent vector at $S = (Y, P)$ if and only if it satisfies:

$$\alpha_1 Y^t \eta_Y + \beta \eta_P P^{-1} - \beta P^{-1}\eta_P = 0. \quad (8.2)$$

$\mathcal{H}_S = \mathcal{H}_{(Y,P)}$ could be identified as the range of the one-to-one operator $\mathrm{N}(S)$ from $\mathcal{E}_{\mathrm{N}} = \mathbb{K}^{(n-p) \times p} \times \mathrm{Sym}_{t,\mathbb{K},p}$ to $\mathcal{E}$, mapping $(B, D) \in \mathcal{E}_{\mathrm{N}}$ to:

$$\mathrm{N}(S)(B, D) = (\mathrm{N}_Y(B, D), \mathrm{N}_P(B, D)) := (\beta Y(P^{-1}D - DP^{-1}) + Y_\perp B, \alpha_1 D), \quad (8.3)$$

where $Y_\perp$ is orthogonal complement matrix to Y, $(Y|Y_\perp) \in \mathrm{U}_{\mathbb{K},n}$. The projection of $\omega = (\omega_Y, \omega_P) \in \mathcal{E} = \mathbb{K}^{n \times p} \times \mathbb{K}^{p \times p}$ to the horizontal space $\mathcal{H}_{(Y,P)}$ is given by

$$\Pi_{\mathcal{H}}(S)(\omega_Y, \omega_P) = (\beta Y(P^{-1}\mathcal{D} - \mathcal{D}P^{-1}) + \omega_Y - YY^t \omega_Y, \alpha_1 \mathcal{D})$$
$$\text{with } \mathcal{D} = \mathcal{D}(P)\omega := L(P)^{-1}\mathrm{sym}_t(\omega_P + Y^t \omega_Y P - PY^t \omega_Y) \quad (8.4)$$
$$\text{where } L(P)X := (\alpha_1 - 2\beta)X + \beta(PXP^{-1} + P^{-1}XP).$$

The operator $L(P)$ could be inverted by Proposition 8.1. The Riemannian Hessian could be evaluated by Eq. (3.11), and the lift of the Levi-Civita connection is given by

$$\nabla^{\mathcal{H}}_\xi \eta := \Pi_{\mathcal{H}} \nabla_\xi \eta = \mathrm{D}_\xi \eta - (\mathrm{D}_\xi \Pi_{\mathcal{H}})\eta + \Pi_{\mathcal{H}} \mathsf{g}^{-1}\mathrm{K}(\xi, \eta), \quad (8.5)$$

for horizontal lifts $\xi = (\xi_Y, \xi_P)$, $\eta = (\eta_Y, \eta_P)$ of tangent vectors of $\mathrm{S}^+_{\mathbb{K},p,n}$ with

$$\mathrm{K}(\xi, \eta)_Y = \frac{\alpha_1 - \alpha_0}{2}(Y(\eta_Y^t \xi_Y + \xi_Y^t \eta_Y) - 2(\eta_Y \xi_Y^t + \xi_Y \eta_Y^t)Y),$$
$$\mathrm{K}(\xi, \eta)_P = -\frac{\beta}{2}(P^{-1}\eta_P P^{-1}\xi_P P^{-1} + P^{-1}\xi_P P^{-1}\eta_P P^{-1}), \quad (8.6)$$
$$(\mathrm{D}_\xi \Pi_{\mathcal{H}})\omega = (((\mathrm{D}_\xi \Pi_{\mathcal{H}})\omega)_Y, \alpha_1 \mathring{\mathcal{D}}),$$
$$((\mathrm{D}_\xi \Pi_{\mathcal{H}})\omega)_Y = \beta \xi_Y(P^{-1}\mathcal{D} - \mathcal{D}P^{-1}) + \beta Y(P^{-1}\mathring{\mathcal{D}} - \mathring{\mathcal{D}}P^{-1}$$
$$+ \mathcal{D}P^{-1}\xi_P P^{-1} - P^{-1}\xi_P P^{-1}\mathcal{D}) - (\xi_Y Y^t + Y\xi_Y^t)\omega_Y,$$
$$\text{with } \mathring{\mathcal{D}} := (\mathrm{D}_\xi \mathcal{D})\omega = L(P)^{-1}\{\mathrm{sym}_t(\xi_Y^t \omega_Y P - P\xi_Y^t \omega_Y + Y^t \omega_Y \xi_P - \xi_P Y^t \omega_Y)$$

$$-\beta(\xi_P \mathcal{D} P^{-1} + P^{-1}\mathcal{D}\xi_P - P\mathcal{D}P^{-1}\xi_P P^{-1} - P^{-1}\xi_P P^{-1}\mathcal{D}P)\}. \tag{8.7}$$

Thus, the Hessian could be evaluated at $O(np^2)$-complexity, by operator composition using Eq. (3.11). Note that $(D_\xi \Pi_\mathcal{H})\eta$ could be further simplified if $\eta = (\eta_Y, \eta_P)$ is a horizontal tangent vector, as $\mathcal{D} = \frac{1}{\alpha_1}\eta_P$ in that case.

Proof We have $\mathrm{Tr}_\mathbb{R}(\alpha_1 \eta_Y^t Y\mathsf{q} + \beta(\eta_P^t P^{-1}(P\mathsf{q} - \mathsf{q}P)P^{-1})) = 0$ for a tangent vector (η_Y, η_P) and a t-antisymmetric matrix q from the horizontal condition. Using Proposition 4.1 this means $\mathrm{skew}_t(\alpha_1 \eta_Y^t Y + \beta P^{-1}\eta_P^t - \beta\eta_P^t P^{-1}) = 0$. Using the fact that $\eta_P^t = \eta_P$ and $\eta_Y^t Y$ is t-antisymmetric, we have Eq. (8.2).

It is clear that $\mathrm{N}(B, D)$ satisfies this equation and is one-to-one: if $\mathrm{N}(B, D) = 0$ then immediately $D = 0$, and $B = 0$ since $Y_\perp^t Y_\perp = I$. It is onto the tangent space by a dimension count. The adjoint $\mathrm{N}^t = (\mathrm{N}_B^t, \mathrm{N}_D^t)$ has two components corresponding to the B and D factors. By Proposition 4.2 we have

$$\mathrm{N}^t(\omega_Y, \omega_P)_B = Y_\perp^t \omega_Y,$$
$$\mathrm{N}^t(\omega_Y, \omega_P)_D = \mathrm{sym}_t(\alpha_1 \omega_P + \beta P^{-1}Y^t\omega_Y - \beta Y^t\omega_Y P^{-1}),$$
$$\mathrm{N}^t\mathsf{g}(\omega_Y, \omega_P)_B = \alpha_0 Y_\perp^t \omega_Y,$$
$$\mathrm{N}^t\mathsf{g}(\omega_Y, \omega_P)_D = \alpha_1\beta\,\mathrm{sym}_t(P^{-1}\omega_P P^{-1} + P^{-1}Y^t\omega_Y - Y^t\omega_Y P^{-1}),$$
$$(\mathrm{N}^t\mathsf{g}\mathrm{N}(B, D))_B = \alpha_0 B,$$
$$(\mathrm{N}^t\mathsf{g}\mathrm{N}(B, D))_D = \alpha_1\beta\,\mathrm{sym}_t(P^{-1}(\alpha_1 D)P^{-1}$$
$$+ P^{-1}Y^t(\beta Y(P^{-1}D - DP^{-1}) + Y_\perp B)$$
$$- Y^t(\beta Y(P^{-1}D - DP^{-1}) + Y_\perp B)P^{-1})$$
$$= \alpha_1\beta\,\mathrm{sym}_t(\alpha_1 P^{-1}DP^{-1} + \beta P^{-2}D - \beta P^{-1}DP^{-1}$$
$$- \beta P^{-1}DP^{-1} + \beta DP^{-2})$$
$$= \alpha_1\beta P^{-1}((\alpha_1 - 2\beta)D + \beta PDP^{-1} + \beta P^{-1}DP)P^{-1}$$
$$= \alpha_1\beta P^{-1}L(P)DP^{-1}.$$

Hence

$$(\mathrm{N}^t\mathsf{g}\mathrm{N})^{-1}(B, D) = (\alpha_0^{-1}B, (\alpha_1\beta)^{-1}L(P)^{-1}(PDP)),$$
$$(\mathrm{N}^t\mathsf{g}\mathrm{N})^{-1}\mathrm{N}^t\mathsf{g}(\omega_Y, \omega_P)$$
$$= (Y_\perp^t \omega_Y, \mathcal{D}).$$

The projection formula Eq. (8.4) is just $\mathrm{N}(\mathrm{N}^t\mathsf{g}\mathrm{N})^{-1}\mathrm{N}^t\mathsf{g}$. The formulas for K follow from the corresponding Stiefel and positive-definite manifold formulas. For $D_\xi \Pi_\mathcal{H}$, we take directional derivative of Eq. (8.4) using standard matrix calculus, the only difficulty is $\mathring{\mathcal{D}} = (D_\xi \mathcal{D})\omega$. We evaluate it by evaluating $D_\xi(L(P)\mathcal{D})\omega$ as

$$(\alpha_1 - 2\beta)(\mathring{\mathcal{D}} + \beta(P\mathring{\mathcal{D}}P^{-1} + P^{-1}\mathring{\mathcal{D}}P) + \beta(\xi_P \mathcal{D}P^{-1} + P^{-1}\mathcal{D}\xi_P)$$
$$- \beta(P\mathcal{D}P^{-1}\xi_P P^{-1} + P^{-1}\xi_P P^{-1}\mathcal{D}P)$$

$$= \mathrm{sym}_{\mathfrak{t}}(\xi_Y^{\mathfrak{t}} \omega_Y P - P \xi_Y^{\mathfrak{t}} \omega_Y + Y^{\mathfrak{t}} \omega_Y \xi_P - \xi_P Y^{\mathfrak{t}} \omega_Y)$$

by differentiating the equation for $\mathcal{D}$. From here, we get the equation for $\mathring{\mathcal{D}}$. $\qquad\square$

To solve for $\mathcal{D}$ and $\mathring{\mathcal{D}}$, we need an extension of the symmetric Lyapunov equation:

Proposition 8.1 *Let $P \in \mathrm{Sym}_{\mathfrak{t},\mathbb{K},p}$ be an $\mathfrak{t}$-symmetric matrix having the eigenvalue decomposition $P = U \Lambda U^{\mathfrak{t}}$ with eigenvalues $(\Lambda_i)_{i=1}^p$ and $U U^{\mathfrak{t}} = I$. Let the set of coefficients $(c_{st})_{-k \le s, t \le k}$ be such that $M_{ij} := \sum_{s=-k, t=-k}^{s=k, t=k} c_{st} \Lambda_i^s \Lambda_j^t$ is not zero for all pairs (Λ_i, Λ_j) of eigenvalues of P, then the equation*

$$\sum_{s=-k, t=-k}^{s=k, t=k} c_{st} P^s X P^t = B \tag{8.8}$$

has the following unique solution that could be computed at $O(p^3)$ complexity:

$$X = U\{(U^{\mathfrak{t}} B U)/M\} U^{\mathfrak{t}} \tag{8.9}$$

with $M = (M_{ij})_{i=1, j=1}^{i=p, j=p}$ and $/$ denotes the by-entry division. In particular, for a positive-definite matrix P and positive scalars (α, β), the equation

$$(\alpha - 2\beta)X + \beta(P^{-1} X P + P X P^{-1}) = B$$

has a unique solution with $M_{ij} = \alpha + \beta(\Lambda_i^{-1} \Lambda_j + \Lambda_i \Lambda_j^{-1} - 2)$.

Proof We follow the idea of [5, 6] but use the eigenvalue decomposition in place of the Schur decomposition. Substitute $P = U \Lambda U^{\mathfrak{t}}$ to Eq. (8.8) and multiply $U^{\mathfrak{t}}$ and U on the left-hand and right-hand sides of that equation, we get

$$\sum_{s=-k, t=-k}^{s=k, t=k} c_{st} \Lambda^s U^{\mathfrak{t}} X U \Lambda^t = U^{\mathfrak{t}} B U,$$

which is equivalent to $(U^{\mathfrak{t}} X U)_{ij} M_{ij} = (U^{\mathfrak{t}} B U)_{ij}$ or $U^{\mathfrak{t}} X U = (U^{\mathfrak{t}} B U)/M$, and we have Eq. (8.9). If α and β are positive then $\alpha_1 + \beta(\Lambda_i^{-1} \Lambda_j + \Lambda_i \Lambda_j^{-1} - 2) > 0$ by the *AGM*-inequality. The eigenvalue decomposition has $O(p^3)$ cost. $\qquad\square$

Horizontal lifts of geodesics on $\mathrm{S}_{\mathbb{K}, p, n}^+$ are geodesics on $\mathrm{St}_{\mathbb{K}, p, n} \times \mathrm{S}_{\mathbb{K}, p}^+$. Recall a *complete manifold* has geodesics that extend indefinitely along any direction. Under this quotient metric, $\mathrm{S}_{\mathbb{K}, p, n}^+$ is a complete Riemannian manifold, as both factors above are. The metric in [27] is complete, while the metric in [13] is not. If (η_Y, η_P) is a horizontal tangent vector at $S = (Y, P)$, a horizontal geodesic γ with $\gamma(0) = S, \dot{\gamma}(0) = (\eta_Y, \eta_P)$ is of the form $(Y(t), P^{1/2} \exp(t P^{-1/2} \eta_P P^{-1/2}) P^{1/2})$, with $Y(t)$ is the geodesic of the metric in Sect. 5, described in [11, 20].

9 Implementation and Data Availability

We developed a *Python* package based on *Pymanopt and Manopt* [8, 26] implementing the manifolds with metrics considered in this paper, for both the real and complex cases in the package [18]. To reuse the optimization code in [8, 26], our *NullRangeManifold* class implements templates of the methods *egrad2rgrad* and *ehess2rhess* based on the formulas in Propositions 3.2 and 3.3 and Theorem 3.1. For a new manifold, the user needs to implement the constraint operator J, its transpose and derivative, the metric operator g, its inverse, derivative, and the third Christoffel term $\mathcal{X}$, as well as a method to solve $\mathsf{J}\,\mathsf{g}^{-1}\,\mathsf{J}^{\mathsf{t}}$ (defaulted to use a conjugate-gradient solver otherwise). A retraction is also required. *NullRangeManifold* automatically provides the projection, Riemannian gradient, and Hessian derived in this paper. Besides the numerical implementation, we also include notebooks showing symbolic derivations of the formulas, and numerical tests, including geodesics in most cases.

We also implement real and complex Stiefel manifolds and positive-semidefinite manifolds with metric parameters. For each manifold, we provide a manifold class to support optimization problems. For flag manifolds, [18] contains a larger class of metric, the results in Sect. 6) is implemented separately in [19], the numerical result is shown here. For testing, we numerically verify the projection satisfying the nullspace condition, metric compatibility, and torsion-freeness of the connection. As the manifolds considered here are constructed from Stiefel or positive-definite matrix manifolds, we use their typical retractions. Since we focus on methodology in this paper, we will not discuss formal numerical experiments. However, we have tested each manifold with a quadratic cost problem including matrices with one size of 1000 dimensions with a Trust-Region solver, ([26], [1, section 7.2.2], summarized in Algorithm 1). Here, an approximate solution is sufficient in step 3 with $\eta \in T_{x_k}\mathcal{M}$, this step is the inner iteration versus the outer iteration in step 2; $||_{x_k}$ and $\langle\rangle_k$ denote norm/inner product at x_k; a terminal condition is imposed).

For flag manifolds, we optimize the function $f(Y) = \mathrm{Tr}_{\mathbb{R}}((Y\varLambda Y^{\mathsf{t}}A)^2)$ over matrices $Y \in \mathrm{St}_{\mathbb{R},d,n}$. Here, A is a positive-definite matrix, $d < n$ are two positive integers, $\hat{\boldsymbol{d}} = (d_1, \cdots, d_q)$ is a partition of d with $\sum_{i=1}^{q} d_i = d$, $\varLambda = \mathrm{diag}(\lambda_1 I_{d_1}, \cdots, \lambda_q I_{d_q})$ for positive numbers $\lambda_1 \cdots \lambda_q$. This cost function is invariant under the action of $\mathrm{U}_{\mathbb{R},\hat{\boldsymbol{d}}}$, thus could be considered as a function on the flag manifold $\mathrm{St}_{\mathbb{R},d,n}/\mathrm{U}_{\mathbb{R},\hat{\boldsymbol{d}}}$. The Euclidean gradient is $4(AY\varLambda)Y^{\mathsf{t}}(AY\varLambda)$, and the Euclidean Hessian is computed by routine matrix calculus. For testing, we consider $d = 60$, $\hat{\boldsymbol{d}} = (30, 20, 10)$, $n = 1000$ and use a trust-region solver. In this case, there is no noticeable variation in time when different values of α are used, typically a few seconds on a free colab engine in the notebook colab/SimpleFlag.ipynb in [19]. Convergence is achieved after typically 16 trust-region iterations, with small alpha requiring more outer iterations. The convergence is superlinear as seen in Fig. 1.

For another test of the Riemannian optimization framework, we consider a nonlinear weighted PCA (principal component analysis) problem, which could be solved by optimizing over the positive-semidefinite matrix manifold. Given a symmetric matrix


Algorithm 1 Riemannian Trust-Region minimizing a function f on a manifold $\mathcal{M}$

1: Input: $x_0 \in \mathcal{M}$; Parameters: $\bar{\Delta} > 0$, Δ_0, ρ'; Require a retraction R

2: **for** $k = 0, 1, 2 \cdots$ **do**

3:     Solve $\eta_k \leftarrow \underset{|\eta|_{x_k} \leq \Delta_k}{\operatorname{argmin}} \{\hat{m}_{x_k}(\eta) := f(x_k) + \langle \operatorname{rgrad} f(x_k), \eta \rangle_k + \frac{1}{2} \langle \operatorname{rhess}_f^{11}(x_k)\eta, \eta \rangle_k\}$;

4:     Calculate $\rho_k \leftarrow \frac{f(x_k) - f(R(x_k, \eta_k))}{\hat{m}_{x_k}(0_{x_k}) - \hat{m}_{x_k}(\eta_k)}$;

5:     **if** $\rho_k < \frac{1}{4}$ **then**

6:         $\Delta_{k+1} \leftarrow \frac{1}{4}\Delta_k$;

7:     **else if** $\rho_k > \frac{3}{4}$ and $|\eta_k| = \Delta_k$ **then**

8:         $\Delta_{k+1} \leftarrow \min(2\Delta_k, \bar{\Delta})$;

9:     **else**

10:         $\Delta_{k+1} \leftarrow \Delta_k$;

11:    **end if**

12:    **if** $\rho_k > \rho'$ **then**

13:        $x_{k+1} \leftarrow R(x_k, \eta_k)$;

14:    **else**

15:        $x_{k+1} \leftarrow x_k$;

16:   **end if**   ▷ Terminal condition is verified after this step

17: **end for**
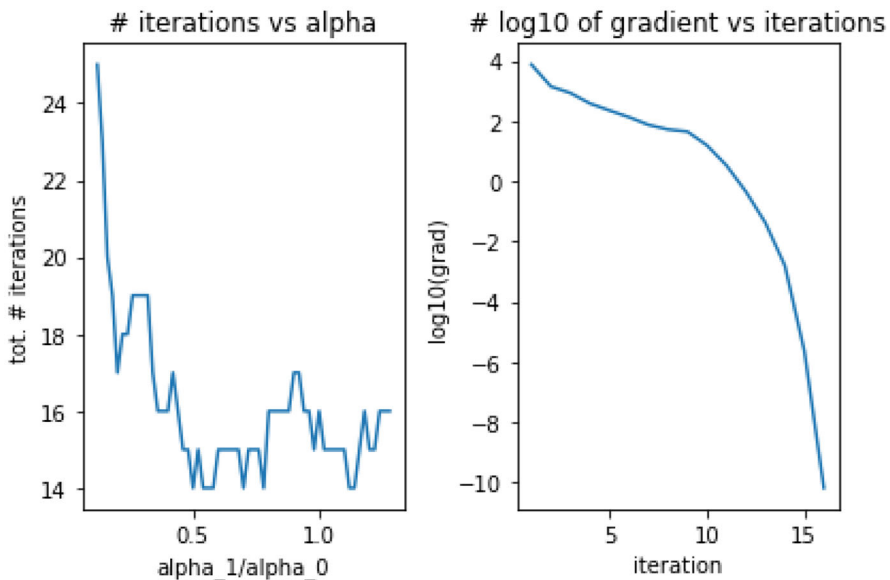
Fig. 1 Optimization on flag manifold Flag(30, 20, 10; 1000, $\mathbb{R}$) using a trust-region solver. Left, number of total iterations versus $\alpha = \alpha_1/\alpha_0$. Right, $\log_{10}$ of gradient for $\alpha = 1$

$A \in \operatorname{Sym}_{\mathsf{T},\mathbb{R},n}$ and a weight vector $W \in \mathbb{R}^n$, take the cost function to be

$$\operatorname{Tr}(A - YPY^{\mathsf{T}}) \operatorname{diag}(W)(A - YPY^{\mathsf{T}}) = \operatorname{Tr} W_d(A^2 - AYPY^{\mathsf{T}} - YPY^{\mathsf{T}}A + YP^2Y^{\mathsf{T}})$$

of a positive-semidefinite matrix $S = YPY^{\mathsf{T}} \in \mathsf{S}_{\mathbb{R},p,n}^+$, with $Y \in \operatorname{St}_{\mathbb{R},p,n}$, $P \in \mathsf{S}_{\mathbb{R},p}^+$. Here, W_d denotes the diagonal matrix $\operatorname{diag}(W)$. When W has identical weight λ,

Weighted PCA using Positive-semidefinite manifold optimization
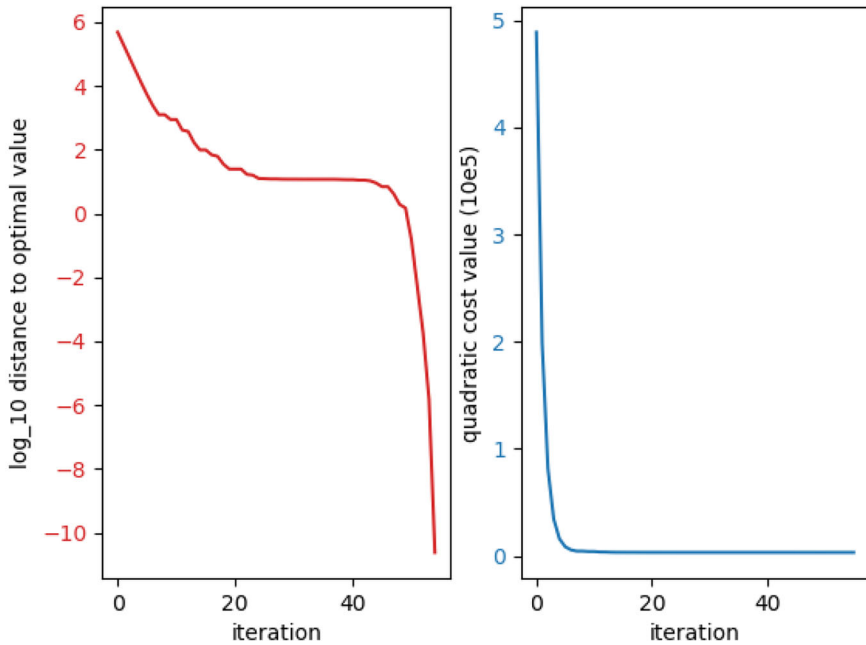
Fig. 2 Weighted PCA using Positive-semidefinite manifold optimization with a trust-region solver. Left, $\log_{10}$ of distance to the optimal value. Right, quadratic cost by iteration

$W_d = \lambda I_n$, expanding the cost function, we need to minimize $\operatorname{Tr} P^2 - 2 \operatorname{Tr} Y^T A Y P$ in Y and P, which implies $P = Y^T A Y$. Thus, the problem is optimizing $- \operatorname{Tr}(Y^T A Y)^2$ over the Stiefel manifold (actually over the Grassmann manifold as the function is invariant when Y is multiplied on the right by an orthogonal matrix), which could be considered as a quadratic PCA problem. When W has nonidentical weights, we optimize over the positive-semidefinite manifold, with a trust-region solver.

The cost function from $\mathcal{M} = \operatorname{St}_{\mathbb{R},p,n}, \times S^+_{\mathbb{R},p}$ extends to $\mathcal{E} = \mathbb{R}^{n \times p} \times \mathbb{R}^{p \times p}$, and is denoted by $\hat{f}(Y, P)$. For a horizontal tangent vector $\xi = (\xi_Y, \xi_P)$ at $(Y, P) \in \mathcal{M}$

$$
\begin{aligned}
(D_\xi \hat{f})(Y, P) = \operatorname{Tr} W_d (&-A\xi_Y P Y^\mathsf{T} - A Y \xi_P Y^\mathsf{T} - A Y P \xi_Y^\mathsf{T} \\
&-\xi_Y P Y^\mathsf{T} A - Y \xi_P Y^\mathsf{T} A - Y P \xi_Y^\mathsf{T} A + \xi_Y P^2 Y^\mathsf{T} \\
&+ Y(\xi_P P + P \xi_P) Y^\mathsf{T} + Y P^2 \xi_Y^\mathsf{T}).
\end{aligned}
$$

We have $\operatorname{Tr} W_d(-A\xi_Y P Y^\mathsf{T} - A Y P \xi_Y^\mathsf{T} - \xi_Y P Y^\mathsf{T} A - Y P \xi_Y^\mathsf{T} A) = -2 \operatorname{Tr}(A W_d + W_d A) Y P \xi_Y^\mathsf{T}$, $\operatorname{Tr} W_d(\xi_Y P^2 Y^\mathsf{T} + Y P^2 \xi_Y^\mathsf{T}) = 2 \operatorname{Tr} W_d Y P^2 \xi_Y^\mathsf{T}$ and similar equalities for ξ_P give us

$$
\operatorname{grad} \hat{f} = (-4 \operatorname{sym}_\mathsf{T}(A W_d) Y P + 2 W_d Y P^2, -2 \operatorname{sym}_\mathsf{T}(Y^\mathsf{T} W_d(A Y - Y P))).
$$

The ambient Hessian $\mathsf{hess}\,\hat{f}(\xi)$ follows from a directional derivative calculation

$$
\begin{aligned}
\mathsf{hess}\,\hat{f}(\xi) = {}& -4\mathsf{sym}_{\mathsf{T}}(AW_d)(\xi_Y P + Y\xi_P) + 2W_d\xi_Y P^2 + 2W_d Y(\xi_P P + P\xi_P), \\
& -2\mathsf{sym}_{\mathsf{T}}(\xi_Y^{\mathsf{T}} W_d(AY - YP)) \\
& -2\mathsf{sym}_{\mathsf{T}}(Y^{\mathsf{T}} W_d(A\xi_Y - \xi_Y P - Y\xi_P))).
\end{aligned}
$$

The Riemannian gradient is computed as $\Pi_{\mathcal{H},\mathsf{g}}\mathsf{g}^{-1}\mathsf{grad}\,\hat{f}$, with $\Pi_{\mathcal{H},\mathsf{g}}$ is given by Eq. (8.4), the Riemannian Hessian is computed from Eq. (3.11) and Theorem 8.1.

In our experiment (implemented in the notebook colab/WeightedPCA.ipynb in [18]), we take $n = 1000$, $p = 50$, with A and W generated randomly. To find the optimum $S = YPY^{\mathsf{T}}$, we optimize with $\alpha_0 = \alpha_1 = 1$, with β is 0.1 for the first 20 iterations, $\beta = 10$ for the next 20 and $\beta = 30$ for the remaining iterations. This choice of β comes from our limited experiments, we find varying β has a strong effect on the speed of convergence, and updating β as such gives better convergence rates than a static β. Philosophically, the small starting β could be thought of as focusing first on aligning the subspace. The convergence graph is summarized in Fig. 2. We hope to revisit the topic with a more systematic study in future works.

10 Conclusion

In this paper, we have proposed a framework to compute the Riemannian gradient, Levi-Civita connection, and the Riemannian Hessian effectively when the constraints, the symmetry of the problem, and the metrics are given analytically and have applied the framework to several manifolds important in applications. We look to apply the results in this paper to several problems in optimization, machine learning and computational statistics. We hope the research community will find the method useful in future works.

Declarations

Conflict of interest The author declares that he has no conflict of interest. No funds, grants, or other support were received.

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2007)
2. Absil, P.A., Mahony, R., Trumpf, J.: An extrinsic look at the Riemannian Hessian. In: F. Nielsen, F. Barbaresco (eds.) Geometric Science of Information–First International Conference, GSI 2013, Paris, France, Proceedings, Lecture Notes in Computer Science, Vol. 8085, pp. 361–368. Springer, Berlin (2013)

3. Adler, R.L., Dedieu, J., Margulies, J.Y., Martens, M., Shub, M.: Newton's method on Riemannian manifolds and a geometric model for the human spine. IMA J. Numer. Anal. **22**(3), 359–390 (2002)
4. Aitken, A.C.: On least squares and linear combination of observations. Proc. R. Soc. Edinb. **55**, 42–48 (1936)
5. Bartels, R.H., Stewart, G.W.: Solution of the matrix equation ax + xb = c [f4]. Commun. ACM **15**(9), 820–826 (1972)
6. Bhatia, R., Rosenthal, P.: How and why to solve the operator equation AX - XB = Y. Bull. Lond. Math. Soc. **29**, 1–21 (1997)
7. Bonnabel, S., Sepulchre, R.: Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. SIAM J. Matrix Anal. Appl. **31**(3), 1055–1070 (2010)
8. Boumal, N., Mishra, B., Absil, P.A., Sepulchre, R.: Manopt, a Matlab toolbox for optimization on manifolds. J. Mach. Learn. Res. **15**, 1455–1459 (2014)
9. 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)
10. Gabay, D.: Minimizing a differentiable function over a differential manifold. J. Optim. Theory Appl. **37**(2), 177–219 (1982)
11. Hüper, K., Markina, I., Silva Leite, F.: A Lagrangian approach to extremal curves on Stiefel manifolds. J. Geom. Mech. **13**(1), 55–72 (2021). https://doi.org/10.3934/jgm.2020031
12. Jeuris, B., Vandebril, R., Vandereycken, B.: A survey and comparison of contemporary algorithms for computing the matrix geometric mean. Electron. Trans. Numer. Anal. **39**, 379–402 (2012)
13. Journée, M., Bach, F., Absil, P.A., Sepulchre, R.: Low-rank optimization on the cone of positive semidefinite matrices. SIAM J. Optim. **20**(5), 2327–2351 (2010)
14. Kasai, H., Mishra, B.: Low-rank tensor completion: a Riemannian manifold preconditioning approach. In: M.F. Balcan, K.Q. Weinberger (eds.) Proceedings of The 33rd International Conference on Machine Learning, Proceedings of Machine Learning Research, Vol. 48, pp. 1012–1021. New York (2016). http://proceedings.mlr.press/v48/kasai16.html
15. Manton, J.H.: Optimization algorithms exploiting unitary constraints. IEEE Trans. Signal Process. **50**(3), 635–650 (2002)
16. Miolane, N., Brigant, A.L., Mathe, J., Hou, B., Guigui, N., Thanwerdas, Y., Heyder, S., Peltre, O., Koep, N., Zaatiti, H., Hajri, H., Cabanes, Y., Gerald, T., Chauchat, P., Shewmake, C., Kainz, B., Donnat, C., Holmes, S., Pennec, X.: Geomstats: A Python package for Riemannian geometry in machine learning (2020). arXiv:2004.04667
17. Mishra, B., Sepulchre, R.: Riemannian preconditioning. SIAM J. Optim. **26**(1), 635–660 (2016). https://doi.org/10.1137/140970860
18. Nguyen, D.: Project ManNullRange. https://github.com/dnguyend/ManNullRange (2020)
19. Nguyen, D.: Project SimpleFlag. https://github.com/dnguyend/SimpleFlag (2021)
20. Nguyen, D.: Closed-form geodesics and optimization for Riemannian logarithms of Stiefel and flag Manifolds. J. Optim. Theory Appl. **194**, 142–166 (2022). https://doi.org/10.1007/s10957-022-02012-3
21. Nishimori, Y., Akaho, S., Plumbley, M.D.: Riemannian optimization method on the flag manifold for independent subspace analysis. In: J. Rosca, D. Erdogmus, J.C. Príncipe, S. Haykin (eds.) ICA 2006: Independent Component Analysis and Blind Signal Separation, Proceedings, Lecture Notes in Computer Science, vol. 3889, pp. 295–302. Springer (2006)
22. O'Neill, B.: Semi-Riemannian Geometry with Applications to Relativity, Pure and Applied Mathematics, vol. 103. Academic Press Inc, New York, NY (1983)
23. Pennec, X., Fillard, P., Ayache, N.: A Riemannian framework for tensor computing. Int. J. Comput. Vis. **66**, 41–66 (2006)
24. Smith, S.T.: Covariance, subspace, and intrinsic Cramér-Rao bounds. IEEE Trans. Signal Process. **53**(5), 1610–1630 (2005)
25. Sra, S., Hosseini, R.: Conic geometric optimization on the manifold of positive definite matrices. SIAM J. Optim. **25**(1), 713–739 (2015)
26. Townsend, J., Koep, N., Weichwald, S.: Pymanopt:a Python toolbox for optimization on manifolds using automatic differentiation. J. Mach. Learn. Res. **17**(137), 1–5 (2016)
27. Vandereycken, B., Absil, P.A., Vandewalle, S.: A Riemannian geometry with complete geodesics for the set of positive semidefinite matrices of fixed rank. IMA J. Numer. Anal. **33**(2), 481–514 (2012)
28. Ye, K., Wong, K.S.W., Lim, L.H.: Optimization on flag manifolds. Math. Program. **194**, 621–660 (2022). https://doi.org/10.1007/s10107-021-01640-3