OPERATOR-VALUED FORMULAS FOR RIEMANNIAN GRADIENT AND HESSIAN AND FAMILIES OF TRACTABLE METRICS

DU NGUYEN

Abstract. We provide an explicit formula 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.

1. Introduction

In this article, we attempt to address the problem: given a manifold, described by constraint equations and a metric, also defined by an analytic formula, compute the Levi-Civita connection, Riemannian Hessian, and gradient for a function on the manifold. 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 metrics, 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

2020 Mathematics Subject Classification. 65K10, 58C05, 49Q12, 53C25, 57Z20, 57Z25, 68T05.

Key words and phrases. Optimization, Riemannian Hessian, Stiefel, Positive-definite, Positive-semidefinite, Flag manifold, Machine Learning.
is also suitable in the case where the gradient formula is not of closed-form. We also provide a number of 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 derive the Riemannian framework for the induced metrics on 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.

1.1. Background. In the foundational paper [9], the authors computed the geodesic equation for a Stiefel manifold $St_{\mathbb{R},p,n}$ of matrices $Y \in \mathbb{R}^{n \times p}$ satisfying $Y^TY = I_p$, with both the constant metric $\text{Tr}(\eta^T\eta)$ on a tangent vector $\eta$, and the canonical metric $\text{Tr}(\eta^T(I_p - \frac{1}{2}YY^T)\eta)$ using calculus of variation. "Doing so is tedious" ([9]), so the details of the calculations were not included in the paper ([11] recently provided a full derivation). Many examples in the literature usually start with a manifold with a known geodesic equation and construct new manifolds from there. In contrast, we will prove several general formulas for Riemannian gradient, Hessian, and geodesic equations applicable when a subspace of the tangent space of a manifold is identified as a subspace of a fixed inner product space. This subspace of the tangent space description often arises directly from the description of the manifold by constraint or by quotient formulation. For example, in the Stiefel case, the tangent space at a matrix $Y \in St_{\mathbb{R},p,n}$ is identified with the nullspace of the operator $J(Y) : \eta \mapsto Y^T\eta + \eta^TY$, ($\eta \in \mathbb{R}^{n \times p}$) which follows from the defining equation of the Stiefel manifold. The Grassmann manifold, the quotient of a Stiefel manifold by the right multiplication action of the orthogonal group, has its tangent space identified with the subspace $Y^T\eta = 0$ [9]. Given this operator $J(Y)$ and an algebraic formula for the metric, our procedure can produce both the Riemannian gradient and Hessian bypassing calculus of variation, with a relatively short calculation for commonly encountered manifolds. The formulas also suggest a procedural approach to compute the gradient or Hessian when they cannot be reduced to simple expressions. In this approach, the Christoffel symbols are replaced by an operator-valued function called the Christoffel function ($\Gamma$ called $\Gamma_c$ in [9]). In subsequent work, we show the Christoffel function could be used to compute Riemannian curvatures. Thus, this approach has further potentials both in theory and in practice.

Symbolic algebra and differentiation have played a major role in optimization problems arising from machine learning applications in recent years. For manifold applications, it has also been used in [27, 10]. A simple adoption of symbolic algebra for noncommutative variables is helpful in deriving formulas for projections and connections. While none of our results are dependent on symbolic calculus, it is helpful for sanity check and exploration. The seemingly complicated Christoffel symbols can be handled symbolically due to two facts: first, directional derivatives of matrix expressions can be evaluated rule-based; second, for the trace inner product, index raising is also a symbolic manipulation (e.g., gradients of $\text{Tr}(AxB)$ and $\text{Tr}(Ax^TB)$ with respect to $x$ are simple algebraic expressions). For relatively complex examples with non-constant ambient metrics, we found it could produce the correct gradients and Hessian mostly automatically when applying our procedure. To keep focus, we will not discuss this topic in further detail and only mention
that we derive Riemannian frameworks for several manifolds symbolically in several notebooks in our code repository [20].

Computing the Riemannian connection and the geodesic equation are important steps in understanding the geometry of a manifold in applied problems. We hope this paper provides a step in making this computation more accessible.

1.2. Riemannian gradient, Hessian and Levi-Civita connection. 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, a vector field providing first-order approximation for $f$ on the manifold, and Riemannian Hessian, which provides the second-order term.

When a manifold $\mathcal{M}$ is embedded in an inner product space $\mathcal{E}$ with the inner product denoted by $\langle \cdot, \cdot \rangle_{\mathcal{E}}$, 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 $\langle \omega_1, g(Y)\omega_2 \rangle_{\mathcal{E}}$ 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, the Riemannian gradient can be computed via a projection from $\mathcal{E}$ to $T_Y \mathcal{M}$ (we have the same picture for a horizontal subspace of $T_Y \mathcal{M}$). 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 $I_\mathcal{E} - g^{-1} J^T (J g^{-1} J^T)^{-1} J$ ($I_\mathcal{E}$ 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 $J^\dagger$. As mentioned, $J$ is often available explicitly. 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, $J^\dagger$ is simple to compute. For manifolds common in applications, $J g^{-1} J^\dagger$ could often be inverted efficiently, as we will see in several examples. Thus, this will be our main approach to compute 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 $\xi$ 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 $\xi, \eta$ on $\mathcal{M}$, the directional derivative $D_\xi \eta$ is an $\mathcal{E}$-valued function but generally not a vector field (i.e. not a $T \mathcal{M}$-valued function). A covariant derivative [23] (or connection) associates a vector field $\nabla_\xi \eta$ to two vector fields $\xi, \eta$ on $\mathcal{M}$. The association is linear in $\xi$, additive in $\eta$ and satisfies the product rule

$$\nabla_\xi (f \eta) = f \nabla_\xi \eta + (D_\xi f) \eta$$

for a function $f$ on $\mathcal{M}$, where $D_\xi f$ denotes the Lie derivative of $f$ (the directional derivative of $f$ along direction $\xi$, at each $x \in \mathcal{M}$). For a Riemannian metric $\langle \cdot, \cdot \rangle_R$ on $\mathcal{M}$, the Levi-Civita connection is the unique connection that is compatible with metric, $D_\xi (\eta, \phi)_R = \langle \nabla_\xi \eta, \phi \rangle_R + \langle \eta, \nabla_\xi \phi \rangle_R$ ($\phi$ 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. $\langle \xi, \eta \rangle_R = (\xi, g_R \eta)_{\mathbb{R}^n}$)

$$\nabla \xi \eta = D\xi \eta + \frac{1}{2} g_R^{-1}((D\xi g_R) \eta + (D\eta g_R) \xi - \mathcal{X}(\xi, \eta))$$

where $\mathcal{X}(\xi, \eta) \in \mathbb{R}^n$ (uniquely defined) satisfies $\langle D\phi g_R \xi, \eta \rangle_{\mathbb{R}^n} = \langle \phi, \mathcal{X}(\xi, \eta) \rangle_{\mathbb{R}^n}$ for all vector field $\phi$. The formula is valid for each coordinate chart, and it is often given in terms of Christoffel symbols in index notation ([23], 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 $f$ to a (multi)linear vector-valued function of one less variable, that evaluates back to $f$ using the inner product pairing.

The Riemannian Hessian could be provided in two formats, as a bilinear form $\text{rhess}^{(0)}(\xi, \eta)$, returning a scalar function to every two vector fields $\xi, \eta$ on the manifold, or a (Riemannian) Hessian vector product $\text{rhess}^{(1)} \xi$, an operator returning a vector field given a vector field input $\xi$. In optimization, as we need to invert the Hessian in second-order methods, the Riemannian Hessian vector product form $\text{rhess}^{(1)}$ is more practical. However, $\text{rhess}^{(2)}$ is directly related to the Levi-Civita connection (see eq. (2.12) below), and can be read from the geodesic equation: In [9], the authors showed the geodesic equation (for a Stiefel manifold) is given by $\dot{Y} + \Gamma(Y, Y) = 0$ where the Christoffel function $\Gamma$ (defined below) maps two vector fields to an ambient function and the bilinear form $\text{rhess}^{(2)}$ is $\hat{f}_{YY}(\xi, \eta) - (\Gamma(\xi, \eta), \hat{f}_Y)\xi$. Here, $\hat{f}_Y$ and $\hat{f}_{YY}$ are the ambient gradient and Hessian defined in section 2. We will provide an explicit formula for $\Gamma$.

1.3. Riemannian optimization. The reader can consult [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 ([10] [9]). Many first-order algorithms (steepest descent, conjugate gradient) on real vector spaces could be extended to manifolds using the Riemannian gradient defined above together with a retraction (some algorithm, for example, conjugate-gradient, requires more differential-geometric measures). 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 $\eta_i$ to the manifold point $Y_i$, which will produce the next iteration point $Y_{i+1}$ via a retraction ([3], chapter 4 of [1]). For manifolds considered in this article, effective retractions are available.

1.4. 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 section 1.3. 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 $\text{Tr}_{\mathbb{R}}$ 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 $\text{Tr}_{\mathbb{R}} ab^\top$, while a complex matrix space becomes a real inner product space with inner product $\text{Tr}_{\mathbb{R}} ab^H$ (see section 1.5 for details). 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 $\text{sym}_t X = \frac{1}{2}(X + X^t)$, $\text{skew}_t X = \frac{1}{2}(X - X^t)$. 

{\text{sym}_t X = \frac{1}{2}(X + X^t), \text{skew}_t X = \frac{1}{2}(X - X^t).}
We denote by $\text{Sym}_{t,K,n}$ the space of $t$-symmetric matrices $X \in K^{n \times n}$ with $X^t = X$. The $t$-antisymmetric space $\text{Skew}_{t,K,n}$ is defined similarly. Symbols $\xi, \eta$ are often used to denote tangent vector or vector fields, while $\omega$ 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 $D_X f$. We also apply Lie derivatives on scalar or operator-valued functions when $X$ is a vector field, and write $D_X g$ for example, where $g$ is a metric operator. Because the vector field $X$ may be a matrix, we prefer the notation $D_X g$ to the usual Lie derivative notation $\mathcal{L}_X g$ which may be ambiguous. By $U_{K,d}$ we denote the group of $K^{d \times d}$ matrices $U$ satisfying $U^t U = I_d$ (called $t$-orthogonal), thus $U_{K,d}$ is the real orthogonal group $O(d)$ when $K = \mathbb{R}$ and unitary group $U(d)$ when $K = \mathbb{C}$.

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 $J(Y)$, or the range of an operator $N(Y)$, 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 $\Pi_{\mathcal{H},g}$ to $\mathcal{H}$, the Christoffel metric term $K$, and their directional derivatives. We also use the symbols $\hat{f}_Y$ and $\hat{f}_{Y^*}$ 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 |
|----------|---------|
| $E$      | ambient space, $\mathcal{M}$ is embedded in $E$. (e.g. $K^{n \times p}$) |
| $\mathcal{E}_I, \mathcal{E}_N$ | inner product spaces, range of $J(Y)$ and domain of $N(Y)$ below. |
| $\mathcal{H}$ | A subbundle of $TM$. Either $TM$ or a horizontal bundle in practice. |
| $J(Y)$ | operator from $E$ onto $\mathcal{E}_I$, Null$(J(Y)) = \mathcal{H}_Y \subset T_Y \mathcal{M}$. (e.g. $Y^t \omega + \omega Y$) |
| $N(Y)$ | inject. oper. from $\mathcal{E}_N$ to $\mathcal{H}_Y \subset T_Y \mathcal{M}$ (e.g. $N(A,B) = YA + YL B$) |
| xtrace   | index-raising operator for the trace (Frobenius) inner product |
| $g(Y)$  | metric given as self-adjoint operator on $E$ (e.g. $(\alpha_1 Y Y^t + \alpha_2 Y L Y^t) \eta$) |
| $\Pi_{\mathcal{H},g}$ | projection to $\mathcal{H} \subset TM$ in proposition $[2]$. |
| $K(\xi,\eta)$ | Christoffel metric term $\frac{1}{2}(g(D_\xi \xi) + g(\xi) g - \text{xtrace}((\xi,\eta) \xi, \phi))$ |
| $\Gamma_{\mathcal{H}}(\xi,\eta)$ | Christoffel function $\Pi_{\mathcal{H}} \xi g^{-1} K(\xi,\eta) - (D_\xi \Pi_{\mathcal{H},g}) \eta$ |

1.5. Our contribution. We formally define a concept of ambient space with a metric operator to make precise the conditions for our approach. Under these conditions, given the operator-valued function $J$ defining a subbundle $\mathcal{H}$ of the tangent bundle (in practice, $\mathcal{H}$ is the full tangent bundle for an embedded manifold or the horizontal subbundle corresponding to a quotient manifold) and an operator-valued function $g$ describing the metric, if $\hat{f}_Y$ is the ambient (Euclidean) gradient, we can use the formula $g^{-1} \hat{f}_Y - g^{-1} J^t (J g J^t)^{-1} J g^{-1} \hat{f}_Y$ to evaluate the Riemannian gradient. We also provide explicit formulas for the Levi-Civita connection eq. $[2.9]$ and Riemannian Hessian eq. $[2.11]$. Among the results, in case $\mathcal{H} = TM$, identifying the tangent space of a manifold $\mathcal{M}$ with a subspace of an ambient space $E$, the Riemannian Hessian product of a function $f$ for a vector field $\xi$ is given by:

$$\text{rhes}^{11}_f \xi = \Pi_g g^{-1} (\hat{f}_Y \xi + g(D_\xi \Pi_g)(g^{-1} \hat{f}_Y) - (D_\xi g)(g^{-1} \hat{f}_Y) + K(\xi, \Pi_g g^{-1} \hat{f}_Y))$$
Here, $\Pi_g$ is the projection from the ambient space to the tangent space, $\hat{f}_Y$ and $\hat{f}_{YY}$ are the ambient gradient and Hessian (will be defined precisely below) and $K$ is the Christoffel metric term. The identification of the tangent space of a manifold $\mathcal{M}$ with a subspace of $\mathcal{E}$ makes $\Pi_g$ and $g$ operator-valued functions on $\mathcal{M}$, thus their directional derivatives are well-defined. So $Dg, D\Pi_g$ and $K$ are all globally defined operator-valued functions. In general, they can be computed from analytic expressions for the metric and the defining equations for the subbundle $\mathcal{H}$. For the first-order case, we also provide a formula for the gradient when the fiber $\mathcal{H}_Y$ of $\mathcal{H}$ at $Y$ is described as the range of a one-to-one map $N(Y)$ at each $Y$. The two ways to describe the subspace $\mathcal{H}_Y$, as the range of $N$ or as the nullspace of $J$, gives us two expressions for the gradient (hence the Hessian) that can be evaluated by linear operator algebra.

The difference with the traditional embedded manifold approach ([1], Proposition 5.32, or [17] for a non-constant metric example), is in those cases, a Riemannian metric for the ambient space is provided, while we only define the operator-valued metric function for points on the manifold in our approach. Therefore, we can use a formula such as $\text{Tr}_g \eta^t(I - \frac{1}{2}YY^t)\eta$ for a metric on a Stiefel manifold, while that formula is not a metric for an arbitrary matrix $Y$. To use the traditional approach for a nonconstant metric, we need to extend the metric to an ambient manifold, then project the ambient connection back to the embedded manifold. There are many ways to extend the metric, but they all produce the same connection on the embedded manifold, thus, it is plausible that there is a connection formula involving only the metric operator. We provide this formula. The approach could be extended to quotients of embedded manifolds easily. Differentiating the metric operator is straightforward for classical manifolds, while extending the metric by an analytic formula to a large ambient space shifts the problem to evaluating the ambient connection, which is may also be complicated (especially if we want to use a metric formula that works on the full ambient space. In contrast, for a Stiefel manifold, our approach essentially extends the formula $\text{Tr}_g \eta^t(I - \frac{1}{2}YY^t)\eta$ to an ambient neighborhood of the manifold but not the full ambient space). We demonstrate this approach in several examples:

1. For a Stiefel manifold $\text{St}_{k,d,n}$ and two positive numbers $\alpha_0, \alpha_1$, we consider a family of metrics of the form $\text{Tr}(\alpha_0\eta^t\eta + (\alpha_1 - \alpha_0)\eta^tYY^t\eta)$ for a tangent vector $\eta$ at $Y \in \text{St}_{k,d,n}$, generalizing both the constant ambient metric (with $\alpha_0 = \alpha_1 = 1$) and the canonical metric ($\alpha_0 = 2\alpha_1 = 1$) defined in [9]. These metrics have been studied recently in [11], with the geodesic equation derived by calculus of variation. We show its gradient and Hessian could also be derived easily in our framework. In [19], we give computationally efficient geodesic formulas generalizing those in [9], and propose an algorithm for the geodesic distance of these metrics.

2. We show it is simple to obtain a Riemannian optimization framework for quotients of a Stiefel manifold $\text{St}_{k,d,n}$ by a group of isometries. We consider the case where this group is a block-diagonal group. This case includes the flag manifolds [22, 29]. We treat the full family of metrics in [11]. The formulas obtained have the same form as those of Stiefel manifolds, and we can provide the Riemannian Hessian operator and second-order methods, an improvement over previous works. In [19] we also propose an algorithm for geodesic distance on flag manifolds.
3. We apply our formula to the manifold $S^n_{+}$ of positive-definite matrices in $\mathbb{K}^{n \times n}$ with the affine-invariant metrics. We recover the known expressions for Riemannian gradient and Hessian (see section 1.6 for related works).

4. We consider the quotient metric for the manifold of positive-semidefinite matrices of rank $d$ in $\mathbb{K}^{n \times n}$, $S^+_d$, considered as $(St_{\mathbb{K},d,n} \times S^+_d)/U_{\mathbb{K},d}$, the quotient by an orthogonal group of the product of a Stiefel manifold equipped with a metric in the family introduced above and the positive-definite matrix manifold with the affine-invariant metrics. We show the Riemannian gradient and Hessian could be computed with the help of a Lyapunov-type equation eq. (8.8) (solvable at cost $O(d^3)$ adapting the Bartels-Stewart algorithm). The affine-invariant behavior of the $S^+_d$ part distinguishes this metric from others considered in the literature. While it shares some features with a metric considered in [7], this metric comes from a Riemannian submersion, allowing us to compute the Hessian by our framework. With this metric, $S^+_d$ is a complete manifold with known geodesics. It could be shown (not presented here) an approximate geodesic considered in [7] could be considered as a limiting geodesic, where our Riemannian metric becomes degenerate.

5. We implement the operators in our approach in Python that can be used with a manifold optimization package, extending the base class Manifold in the package Pymanopt (based on Manopt, [27], [8]), see [20]. Note that our implementation can take metric parameters, while most implementations typically handle one metric per module. With our approach, it is simple to test the torsion-free and metric-compatibility properties of the Levi-Civita connection numerically as the derivatives involved are of operator-valued functions. We tested the implementation with the Trust-Region algorithm for a few benchmark problems.

Our analysis offers 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 above 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 Riemannian Hessian expressions (maybe tedious algebraically), 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 ([26]).

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 $Jg^{-1}J^t$ (or $NgN$). While in our examples these operators are reduced to simple ones that could be inverted efficiently, otherwise, $Jg^{-1}J^t$ could be solved by a conjugate gradient (CG)-method (which has been used for Riemannian optimization, see [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.
1.6. Related works and outline. The formulas for the Hessian (theorem 1) have mostly been used with a constant metric on an ambient manifold. For example, formula (7) of [2] is the special case for our Hessian formula for constant ambient metrics, or section 4.9 of [9] also discusses the Christoffel function for an embedded manifold. As discussed, we provide a formalism allowing us to compute this function via operators on ambient space. The paper [15] gave the original treatment of the Hessian for the unitary/complex case. The formulas for Stiefel manifolds overlap with those in [9, 11], obtained by different methods. Optimization on flag manifolds was studied in [22, 29]. The affine-invariant metric on positive-definite matrices was also widely studied, for example in [25, 24, 26, 12]. There are numerous metrics on the fixed-rank positive-semidefinite (PSD) manifolds, 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 [28, 13], 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, 18] discuss the effect of adapting metrics to optimization problems ([18] adapts ambient metrics to the objective function using first-order methods.)

In the next section, 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. Ambient space and optimization on Riemannian manifolds

While an abstract manifold is defined in terms of coordinate charts, manifolds appearing in applications are usually subspaces of an inner product space $E$ with inner product $\langle \cdot, \cdot \rangle_E$, or quotient space of such subspaces, defined by constraints and symmetries ([9, 1]). For a manifold $M$ embedded in $E$ differentiably, its tangent spaces are also identified with subspaces of $E$, the identification is parametrized smoothly by points of $M$. For example, if $M$ is a unit sphere in $E = \mathbb{R}^m$, the tangent space at a point $x$ is the subspace of vectors $\eta$ such that $x^T \eta = 0$. In the quotient manifold situation (reviewed below), for $Y \in M$, we focus on a subspace $H_Y$ of $T_Y M$, the horizontal subspace with respect to a group action. The space $H_Y$ is also considered as a subspace of $E$, allowing us to treat quotient manifolds in the same setting.

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

Definition 1. We call an inner product (Euclidean) space $(E, \langle \cdot, \cdot \rangle_E)$ an embedded ambient space of a Riemannian manifold $M$ if there is a differentiable (not necessarily Riemannian) embedding $M \subset E$. 
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 $\hat{\text{grad}}f$ an ambient gradient of $f$. It is a vector-valued function from $\mathcal{M}$ to $\mathcal{E}$ such that for all vector fields $\eta$ on $\mathcal{M}$

\begin{equation}
\langle \eta(Y), \hat{\text{grad}}f(Y) \rangle_{\mathcal{E}} = (D_{\eta(Y)}f)(Y) \quad \text{for all } Y \in \mathcal{M}
\end{equation}

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

By the Whitney embedding theorem, any manifold has an ambient space. Coordinate charts could be considered as a collection of compatible local ambient spaces. However, globally defined ambient spaces are more effective in computations.

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.

**Lemma 1.** For a function $f$ and two vector fields $\xi, \eta$ on $\mathcal{M} \subset \mathcal{E}$ we have:

\begin{equation}
\hat{f}_{Y\xi Y}^{\mathcal{E}}(\xi, \eta) = \text{hess}^{\mathcal{E}}(\xi, \eta) = D_{\xi}(D_{\eta}f) - \langle D_{\xi}\eta, \hat{f}_{Y}\rangle_{\mathcal{E}}
\end{equation}

**Proof.** We have $D_{\xi}\langle \eta, \hat{f}_{Y}\rangle_{\mathcal{E}} = \langle D_{\xi}\eta, \hat{f}_{Y}\rangle_{\mathcal{E}} + \langle \eta, D_{\xi}(\hat{f}_{Y}) \rangle_{\mathcal{E}} = D_{\xi}(D_{\eta}f)$ by taking directional derivatives of eq. (2.1), thus $\langle \eta, D_{\xi}(\hat{f}_{Y}) \rangle_{\mathcal{E}}$ can be evaluated by eq. (2.2). $\square$

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 $A^\dagger$ such that $\langle Av, w \rangle_W = \langle v, A^\dagger w \rangle_V$ 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 $A^\dagger$ is represented by its transpose $A^\dagger$. A projection from an inner product space $V$ to a subspace $W$ is a linear map $\Pi_W$ such that $\langle v, w \rangle_V = \langle \Pi_W v, w \rangle_V$ 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 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 $\langle ge_1, e_2 \rangle_{\mathcal{E}} = \langle e_1, ge_2 \rangle_{\mathcal{E}}$. The operator $g$ defines a new inner product on $\mathcal{E}$ by $\langle e_1, e_2 \rangle_{g, \mathcal{E}} := \langle e_1, ge_2 \rangle_{\mathcal{E}}$. If $W = \text{Null}(J)$ for a map $J$ from $\mathcal{E}$ onto an inner product space $\mathcal{E}_J$, the projection $\Pi_{gJ} = \Pi_{g,J}W$ from $\mathcal{E}$ to $W$ under the inner product $\langle , \rangle_{g, \mathcal{E}}$ is given by $\Pi_{gJ} e = e - g^{-1}J^\dagger(Jg^{-1}J^\dagger)^{-1}J e$, where $J^\dagger$ 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 $\mathcal{E}_N$ to $\mathcal{E}$ such that $W = N(\mathcal{E}_N)$, then the projection to $W$ could be given by $\Pi_{gJ} e = N(N^\dagger gN)^{-1}N^\dagger ge$.

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

**Proof.** The assumption that $J$ is onto $\mathcal{E}_J$ shows $J^\dagger$ is injective (as $J^\dagger a = 0$ implies $\langle a, J\omega \rangle_{\mathcal{E}_J} = 0$ for all $\omega \in \mathcal{E}$, and since $J$ is onto this implies $a = 0$). This in turn implies $Jg^{-1}J^\dagger$ is invertible as if $Jg^{-1}J^\dagger a = 0$, then $\langle Jg^{-1}J^\dagger a, a \rangle_{\mathcal{E}} = 0$, so $\langle g^{-1}J^\dagger a, J^\dagger a \rangle_{\mathcal{E}} = 0$ and hence $J^\dagger a = 0$ as $g$ is positive-definite. We can show $N^\dagger gN$ is invertible similarly.
For the first case, if \( e_W \in W = \text{Null}(J) \) and \( e \in \mathcal{E} \),
\[
\langle \mathbf{g}^{-1} J \mathbf{1} (J \mathbf{g}^{-1} J) \mathbf{1} J e, \mathbf{g} e_W \rangle_{\mathcal{E}} = \langle J (J \mathbf{g}^{-1} J) \mathbf{1} J e, e_W \rangle_{\mathcal{E}} = \langle (J \mathbf{g}^{-1} J) \mathbf{1} J e, J e_W \rangle_{\mathcal{E}}
\]
where the last term is zero because \( e_W \in \text{Null}(J) \), so
\[
\langle \Pi_{\mathcal{E}} e, \mathbf{g} e_W \rangle_{\mathcal{E}} = \langle e, \mathbf{g} e_W \rangle_{\mathcal{E}}. \]
For the second case, assuming \( e_W = N(e_N) \) for \( e_N \in \mathcal{E}_N \) then
USING \((AB)^t = B^t A^t\): 
\[
\langle N(N^t \mathbf{g} N)^{-1} N^t \mathbf{g} e, \mathbf{g} N(e_N) \rangle_{\mathcal{E}} = \langle \mathbf{g} e, N(N^t \mathbf{g} N)^{-1} N^t \mathbf{g} N(e_N) \rangle_{\mathcal{E}} = \langle \mathbf{g} e, N(e_N) \rangle_{\mathcal{E}}
\]
The last statement follows from the defining equations of \( \Pi_{\mathcal{E}} \). □ □ □

Recall the Riemannian gradient of a function \( f \) on a manifold \( \mathcal{M} \) with Riemannian metric \( \langle \cdot, \cdot \rangle_R \) is the vector field \( \mathbf{rgrad}_f \) such that
\[
\langle \mathbf{rgrad}_f (Y) \rangle (\xi)_{\mathcal{E}} = (D\xi f)(Y)
\]
for any point \( Y \in \mathcal{M} \), and any vector field \( \xi \). 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 a vector product of a vector space and an open subset of \( \mathcal{M} \), together with a linear coordinate change condition (see [24], definition 7.24 for details). We can define the \( \mathcal{H} \)-Riemannian gradient \( \mathbf{rgrad}_{\mathcal{H}, f} \) of \( f \) as the unique \( \mathcal{H} \)-valued vector field such that 
\[
\langle \mathbf{rgrad}_{\mathcal{H}, f}, \xi_{\mathcal{H}} \rangle_R = (D\xi f)(Y)
\]
for any \( \mathcal{H} \)-valued vector field \( \xi_{\mathcal{H}} \). Uniqueness follows from nondegeneracy of the inner product restricted to \( \mathcal{H} \). Clearly, when \( \mathcal{H} = T\mathcal{M} \), \( \mathbf{rgrad}_{\mathcal{H}, f} = \mathbf{rgrad} \) is the usual Riemannian gradient. We have:

**Proposition 2.** Let \( (\mathcal{E}, \langle \cdot, \cdot \rangle_{\mathcal{E}}) \) be an embedded ambient space of a manifold \( \mathcal{M} \) as in definition 7. Let \( \mathbf{g} \) be a smooth operator-valued function associating each \( Y \in \mathcal{M} \) a self-adjoint positive-definite operator \( \mathbf{g}(Y) \) on \( \mathcal{E} \). Thus, each \( \mathbf{g}(Y) \) defines an inner product on \( \mathcal{E} \), which induces an inner product on \( T_Y \mathcal{M} \) and hence \( \mathbf{g} \) induces a Riemannian metric on \( \mathcal{M} \). Let \( \mathcal{H} \) be a subbundle of \( T\mathcal{M} \). Define \( \Pi_{\mathcal{H}, \mathbf{g}} \) to be the operator-valued function such that \( \Pi_{\mathcal{H}, \mathbf{g}}(Y) \) is the projection associated with \( \mathbf{g}(Y) \) from \( \mathcal{E} \) to the fiber \( \mathcal{H}_Y \), and for the case \( \mathcal{H} = T\mathcal{M} \), define \( \Pi_{\mathcal{E}, \mathbf{g}} = \Pi_{T\mathcal{M}, \mathbf{g}} \). For an ambient gradient \( \mathbf{grad} \) of \( f \), the \( \mathcal{H} \)-Riemannian gradient of \( f \) can be evaluated as:

\[
\mathbf{rgrad}_{\mathcal{H}, f} = \Pi_{\mathcal{H}, \mathbf{g}} \mathbf{g}^{-1} \mathbf{grad} \tag{2.3}
\]
If there is an inner product space \( \mathcal{E}_1 \) and a map \( J \) from \( \mathcal{M} \) to the space \( \mathcal{L}(\mathcal{E}, \mathcal{E}_1) \) of linear maps from \( \mathcal{E} \) to \( \mathcal{E}_1 \), such that for each \( Y \in \mathcal{M} \), the range of \( J(Y) \) is precisely \( \mathcal{E}_1 \), and its nullspace is \( \mathcal{H}_Y \) then \( \Pi_{\mathcal{H}, \mathbf{g}}(Y) e \) for \( e \in \mathcal{E} \) could be given by:

\[
\Pi_{\mathcal{H}, \mathbf{g}}(Y) e = e - \mathbf{g}^{-1} J (J^{-1} \mathbf{g}^{-1} J) \mathbf{1} J e; \text{ all are evaluated at } Y \tag{2.4}
\]
If there is an inner product space \( \mathcal{E}_N \) and a map \( N \) from \( \mathcal{M} \) to the space \( \mathcal{L}(\mathcal{E}_N, \mathcal{E}) \) of linear maps from \( \mathcal{E}_N \) to \( \mathcal{E} \) such that for each \( Y \in \mathcal{M} \), \( N(Y) \) is one-to-one, with its range is precisely \( \mathcal{H}_Y \) then:

\[
\Pi_{\mathcal{H}, \mathbf{g}} e = N(N^t \mathbf{g} N)^{-1} N^t \mathbf{g} e; \text{ all are evaluated at } Y \tag{2.5}
\]
Proof. For any \( \mathcal{H} \)-valued vector field \( \xi_{\mathcal{H}} \), we have:

\[
\langle \Pi_{\mathcal{H}, \mathbf{g}}^{-1} \mathbf{grad} \rangle, \xi_{\mathcal{H}} \rangle_{\mathcal{E}} = \langle \mathbf{grad} \rangle, \Pi_{\mathcal{H}, \mathbf{g}}^{-1} \mathbf{g} \xi_{\mathcal{H}} \rangle_{\mathcal{E}} = \langle \mathbf{grad} \rangle, D\xi_{\mathcal{H}} f \rangle_{\mathcal{E}}
\]
because \( \Pi_{\mathcal{H}, \mathbf{g}}^{-1} \) is self-adjoint and the projection is idempotent. The remaining statements are just a parametrized version of proposition [1]. □ □ □

Note, we are not making any smoothness assumption on \( J \) or \( N \) yet, although \( \Pi_{\mathcal{H}, \mathbf{g}} \) 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 2. A triple \((\mathcal{M}, g, \mathcal{E})\) 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 \(\mathcal{L}(\mathcal{E}, \mathcal{E})\) 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

\[ D_\xi \eta - D_\eta \xi = [\xi, \eta] \text{ for all vector fields } \xi, \eta \text{ on } \mathcal{M}. \]

Recall if \(\mathcal{M}, \langle \cdot, \cdot \rangle_R\) is a Riemannian manifold with the Levi-Civita connection \(\nabla\), the Riemannian Hessian (vector product) of a function \(f\) is the operator sending a tangent vector \(\xi\) to the tangent vector \(\text{rhes}_f^{11}\xi = \nabla_\xi \text{grad}_f\). The Riemannian Hessian bilinear form is the map evaluating on two vector fields \(\xi, \eta\) as \(\langle \nabla_\xi \text{grad}_f, \eta \rangle_R\). 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 \(\text{rhes}^{11}_f(\xi_\mathcal{H}, \eta_\mathcal{H}) = \Pi_{\mathcal{H}, g} \nabla_\xi \text{grad}_f, \eta_\mathcal{H} = \langle \nabla_\xi \text{grad}_f, \eta_\mathcal{H} \rangle_R\) the \(\mathcal{H}\)-Riemannian Hessian bilinear form. The next theorem shows how to compute the Riemannian connection and the associated Riemannian Hessian.

Theorem 1. Let \((\mathcal{M}, g, \mathcal{E})\) 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 \((\xi, \eta)\) to \(\mathcal{X}(\xi, \eta) \in \mathcal{E}\) such that for any vector field \(\xi_0\):

\[ (\mathcal{X}(\xi, \eta), \xi_0, \mathcal{E}) = \langle \xi, (D_\xi g)\eta \rangle \mathcal{E} \]

Let \(\Pi_{\mathcal{M}, g}\) be the projection from \(\mathcal{E}\) to the tangent bundle of \(\mathcal{M}\). Then \(\Pi_{\mathcal{M}, g}^{-1}\mathcal{X}(\xi, \eta)\) is uniquely defined given \(\xi, \eta\) 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 \(\xi, \eta\) on \(\mathcal{M}\), define

\[ K(\xi, \eta) := \frac{1}{2}((D_\xi g)\eta + (D_\eta g)\xi - \mathcal{X}(\xi, \eta)) \in \mathcal{E} \]

\[ \nabla_\xi \eta := D_\xi \eta + g^{-1}K(\xi, \eta) \]

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

\[ \Gamma(\xi, \eta) := -(D_\xi \Pi_{\mathcal{M}, g})\eta + \Pi_{\mathcal{M}, g}^{-1}K(\xi, \eta) \]

\[ \nabla_\xi \eta := D_\xi \eta + \Gamma(\xi, \eta) \]

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:

\[ \Pi_{\mathcal{H}, g} \nabla_\xi \eta_\mathcal{H} = D_\xi \eta_\mathcal{H} + \Gamma(\xi, \eta) \]

If \(f\) is a function on \(\mathcal{M}\), \(\tilde{f}_Y\) is an ambient gradient of \(f\) and \(f_{YY}\) is the ambient Hessian operator, we have \(\text{rhes}_f^{11}(\xi_\mathcal{H}) = \Pi_{\mathcal{H}, g} \nabla_\xi \text{grad}_f, \eta_\mathcal{H}\) and \(\text{rhes}_f^{02}(\eta_\mathcal{H})\) are given by:

\[ \text{rhes}_f^{11}(\xi_\mathcal{H}) = \Pi_{\mathcal{H}, g}^{-1}(\tilde{f}_{YY} \xi_\mathcal{H} + g(D_\xi \Pi_{\mathcal{H}, g}^{-1}))\tilde{f}_Y + K(\xi_\mathcal{H}, \Pi_{\mathcal{H}, g}^{-1}\tilde{f}_Y) \]

\[ = \Pi_{\mathcal{H}, g}^{-1}(\tilde{f}_{YY} \xi_\mathcal{H} + g(D_\xi \Pi_{\mathcal{H}, g})g^{-1}\tilde{f}_Y - (D_\xi g)g^{-1}\tilde{f}_Y + K(\xi_\mathcal{H}, \Pi_{\mathcal{H}, g}^{-1}\tilde{f}_Y) \]

\[ \text{rhes}_f^{02}(\eta_\mathcal{H}, \xi_\mathcal{H}) = (\nabla_\xi \Pi_{\mathcal{H}, g}^{-1}\tilde{f}_Y, \mathcal{E}) = \tilde{f}_{YY}(\xi_\mathcal{H}, \eta_\mathcal{H}) - \langle \nabla(\xi_\mathcal{H}, \eta_\mathcal{H}), \tilde{f}_Y \rangle \mathcal{E} \]
For eq. (2.10), for
For eq. (2.9), we note
We can verify directly that
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 \(\Pi_{M,g}\). Evaluated at \(Y \in M\), 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 \(\text{rhess}^{(2)}\) in terms of \(\Gamma\) above. In [9], \(\Gamma_{H}\) was computed for a Grassmann manifold. The formulation for subbundles allows us to extend the result to Riemannian submersions and quotient manifolds.

Proof. \(X\) is the familiar index-raising term: for \(Y \in M\) and \(v_0, v_1, v_2 \in T_Y M\), as
We can take another \(T\)-valued function \(X_1\) satisfying the same condition but not necessarily in the tangent space, the expression \(\Pi_{M,g}g^{-1}X_1\), hence \(\Pi_{M,g}g^{-1}K\) is independent of the choice of \(X_1\), as for three vector fields \(\xi_0, \xi, \eta\)
We can verify directly that \(\nabla_{\xi} \eta\) satisfies the conditions of a covariant derivative: linear in \(\xi\) and satisfying the product rule with respect to \(\eta\). Similar to the calculation with coordinate charts, we can show \(\nabla\) is compatible with metric: for two vector fields \(\xi, \eta\),
Torsion-free follows from the fact that \(K\) is symmetric and eq. (2.6):
For eq. (2.5), we note \(\Pi_{M,g}D_\xi \eta = D_\xi (\Pi_{M,g} \eta) - (D_\xi \Pi_{M,g}) \eta\) so
For eq. (2.9), for \(Y \in M\), \(H_Y \subset T_Y M\) implies \(\Pi_{H,g} \Pi_{M,g} = \Pi_{H,g}\). Therefore, eq. (2.8) implies
and as before, we use \(\Pi_{H,g}D_\xi \eta_\xi = D_\xi (\Pi_{H,g} \eta_\xi) - (D_\xi \Pi_{H,g}) \eta_\xi \) and \(\Pi_{H,g} \eta_\xi = \eta_\xi\). The first line of eq. (2.11) is by definition and \(D_\xi (\Pi_{H,g} g^{-1} f_Y) = D_\xi (\Pi_{H,g} g^{-1} f_Y) + \Pi_{H,g} g^{-1} f_Y Y\). Expand, note \(\Pi_{H,g} g^{-1} (g \Pi_{H,g} D_\xi g^{-1}) f_Y = -\Pi_{H,g} g^{-1} (D_\xi g) g^{-1} f_Y\) (as \(\Pi_{H,g}\) is idempotent), we have the second line. For eq. (2.12):

\[
v_{\xi, H}(g^{-1} f_Y, \eta_H) \varepsilon = D_{\xi, H}(g^{-1} f_Y, \eta_H) \varepsilon - \langle \Pi_{H,g} g^{-1} f_Y, g v_H \rangle \varepsilon = D_{\xi, H}(f_Y, H, \eta_H) \varepsilon - (f_Y, \Pi_{H,g} g^{-1} \Pi_{H,g} D_\xi \eta_H) \varepsilon = D_{\xi, H}(f_Y, \eta_H) \varepsilon - (f_Y, \Pi_{H,g} D_\xi \eta_H) \varepsilon = (f_Y, \Pi_{H,g} D_\xi \eta_H) \varepsilon - (f_Y, \Pi_{H,g} D_\xi \eta_H) \varepsilon.
\]

\[
\varepsilon = (f_Y, \Pi_{H,g} D_\xi \eta_H) \varepsilon - (f_Y, \Pi_{H,g} D_\xi \eta_H) \varepsilon.
\]
from compatibility with metric, idempotency of \( \Pi_{\mathcal{H},g} \), eq. (2.10) and eq. (2.2).

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

**Proposition 3.** If \( J \) as in proposition \( \square \) is of class \( C^2 \) then:

\[
\Gamma_{\mathcal{H}}(\xi_{\mathcal{H}},\eta_{\mathcal{H}}) = g^{-1}J(\xi_{\mathcal{H}},\eta_{\mathcal{H}})\Pi_{\mathcal{H},g}^1 - \Pi_{\mathcal{H},g}^1K(\xi_{\mathcal{H}},\eta_{\mathcal{H}})
\]

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

\[
\nabla_{\xi_{\mathcal{H}}}\Pi_{\mathcal{H},g}^1\omega = \Pi_{\mathcal{H},g}^1D_{\xi_{\mathcal{H}}}\omega - \Pi_{\mathcal{H},g}^1(D_{\xi_{\mathcal{H}}}g^{-1}\omega) - \Pi_{\mathcal{H},g}^1(\xi_{\mathcal{H}},\Pi_{\mathcal{H},g}^1\omega) + \Pi_{\mathcal{H},g}^1K(\xi_{\mathcal{H}},\Pi_{\mathcal{H},g}^1\omega)
\]

**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 \( \Pi_{\mathcal{H},g}^1\nabla_{\xi_{\mathcal{H}}}J(\mathcal{H},\Pi_{\mathcal{H},g}^1\omega) = \Pi_{\mathcal{H},g}^1D_{\xi_{\mathcal{H}}}\Pi_{\mathcal{H},g}^1\omega + \Pi_{\mathcal{H},g}^1K(\xi_{\mathcal{H}},\Pi_{\mathcal{H},g}^1\omega) \), then expand the first term and use \( \Pi_{\mathcal{H},g}^1J = 0 \). \( \square \)

Recall (29, Definition 7.44) a Riemannian submersion \( \pi : \mathcal{M} \rightarrow \mathcal{B} \) between two manifolds \( \mathcal{M} \) and \( \mathcal{B} \) is a smooth, onto map, such that the differential \( d\pi \) is onto at every point \( Y \in \mathcal{M} \), the fiber \( \pi^{-1}(b) \) 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 \( \xi \) 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 following proposition allows us to apply the results so far in familiar contexts:

**Proposition 4.** Let \( (\mathcal{M},g,E) \) be an embedded ambient structure.

1. Fix an orthogonal basis \( e_i \) of \( E \), let \( f \) be a function on \( \mathcal{M} \), which is a restriction of a function \( \hat{f} \) on \( E \), define \( \hat{f}_Y \) to be the function from \( \mathcal{M} \) to \( E \), 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 \( (\mathcal{M},g,E) \) 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 \( \xi,\eta \) are two vector fields on \( \mathcal{B} \) with \( \xi_{\mathcal{H}},\eta_{\mathcal{H}} \) their horizontal lifts, then the Levi-Civita connection \( \nabla^g_{\xi_{\mathcal{H}}} \eta \) on \( \mathcal{B} \) lifts to \( \Pi_{\mathcal{H},g}^1\nabla^g_{\xi_{\mathcal{H}}}\eta \), hence eq. (2.10) 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 \( (\hat{f}_Y,e_i)_E = D_{e_i}\hat{f} \). The statement about the Jacobian is simply the implicit function theorem. Isometry of horizontal lift and [29], Lemma 7.45, item 3, gives us Statement 2. \( \square \)
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 $\xi$ 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.

3. An example

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, $J_C = C_x$ 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

$$
\Pi_\mathcal{E} \omega = \omega - C_x^T(C_x C_x^T)^{-1} C_x \omega
$$

and the covariant derivative is given by $\nabla_{\xi} \eta = D_{\xi} \eta + C_x^T(C_x C_x^T)^{-1}(D_x C_x) \eta$ for two vector fields $\xi, \eta$. With $\Gamma(\xi, \eta) = C_x^T(C_x C_x^T)^{-1}(D_x C_x) \eta$, the Riemannian Hessian bilinear form is computed from eq. (2.12), and the Riemannian Hessian operator is:

$$
\Pi_\mathcal{E} (\hat{f}_{xx} \xi - (D_x C_x)^T(C_x C_x^T)^{-1} C_x \hat{f}_x)
$$

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

To fix some notations, we use the symbol $\text{Tr}_\mathbb{R}$ to denote the real part of the trace, so for a matrix $a \in \mathbb{K}^{n \times n}$, $\text{Tr}_\mathbb{R} a = \text{Tr} a$ if $\mathbb{K} = \mathbb{R}$ and $\text{Tr}_\mathbb{R} a = \text{Re}(\text{Tr} a)$ 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 complex transpose. The inner product of two matrices $a, b$ is $\text{Tr}_\mathbb{R}(a b^H)$. Recall that we denote by $\text{Sym}_{\mathbb{K}}$, the space of all real symmetric matrices $(A^t = A)$, and $\text{Skew}_{\mathbb{K}}$ the space of all real antisymmetric matrices $(A^t = -A)$. We consider both $\text{Sym}_{\mathbb{K}}$ and $\text{Skew}_{\mathbb{K}}$ inner product spaces under $\text{Tr}_\mathbb{R}$. We defined the symmetrizer $\text{sym}_t$ and antisymmetrizer $\text{skew}_t$ in section 1.4 with the usual meaning.
Proposition 5. With the above notations, let $A_i, B_i, X$ be matrices such that the functional $L(X) = \sum_{i=1}^{k} Tr(B_i^t X B_i) + Tr(C_i X^t D_i)$ is well-formed. We have:

1. The matrix $xtrace(L, X) = \sum_{i=1}^{k} A_i^t B_i + D_i C_i$ is the unique matrix $L_1$ such that $Tr A_1 X^t = L(X)$ for all $X \in \mathbb{R}^{n \times m}$ (this is the gradient of $L$).

2. The matrix $xtrace^{sym}(L, X) = sym_i(\sum_{i=1}^{k} A_i^t B_i + D_i C_i)$ is the unique matrix $L_2 \in \text{Sym}_{k, \mathbb{R}}$ satisfying $Tr(L_2 X^t) = L(X)$ for all $X \in \text{Sym}_{k, \mathbb{R}}$.

3. The matrix $xtrace^{skew}(L, X) = skew_i(\sum_{i=1}^{k} A_i^t B_i + D_i C_i)$ is the unique matrix $L_3 \in \text{Skew}_{k, \mathbb{R}}$ satisfying $Tr(L_3 X^t) = L(X)$ for all $X \in \text{Skew}_{k, \mathbb{R}}$.

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. Since $Tr_k(A_i X B_i) = Tr_k(B_i^t X A_i^t) = Tr_k(A_i^t B_i^t X)$ and $Tr_k(C_i X^t D_i) = Tr_k(D_i C_i X^t)$, we have $Tr_k(xtrace(L) X^t) = L(X)$. Uniqueness follows from the fact that $Tr_k$ is a non-degenerate bilinear form. The last two statements follow from $Tr_k(xtrace(L) X^t) = Tr_k(xtrace(L) X)$ if $X^t = X$ and $Tr_k(xtrace(L) X) = -Tr_k(xtrace(L)^t X)$ if $X^t = -X$. □

Remark 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 $M$ be a manifold with ambient space $\mathbb{R}^{n \times m}$, recall $\text{rhess}_f^{(2)}(\xi, \eta) = f_{\text{Y}}(\xi, \eta) - \Gamma_\eta(\xi, \eta)$. Assume $\langle \Gamma_\eta(\xi, \eta), f_{\text{Y}} \rangle = \sum_i Tr_k(A_i \eta B_i + Tr_k(C_i \eta D_i)$ with $A_i, B_i, C_i, D_i$ are not dependent on $\eta$, and identify tangent vectors with their images in $\mathbb{R}^{n \times m}$, we have:

$$\text{rhess}_f^{(1)}\xi = \Pi g^{-1} xtrace(\text{rhess}_f^{(2)}(\xi, \eta))$$

as the inner product of the right-hand side with $\eta$ is $\text{rhess}_f^{(2)}(\xi, \eta)$, and the projection ensures it is in the tangent space. If the ambient space is identified with $\text{Sym}_{k, \mathbb{R}}$, $\text{Skew}_{k, \mathbb{R}}$ or a direct sum of matrix spaces, we also have similar statements.

Proposition 6. With the same notations as proposition 3.

1. The adjoint of the left multiplication operator by a matrix $A \in \mathbb{R}^{m \times n}$, sending $X \in \mathbb{R}^{n \times p}$ to $AX \in \mathbb{R}^{m \times p}$ is the left multiplication by $A^t$, sending $Y \in \mathbb{R}^{m \times p}$ to $A^t Y \in \mathbb{R}^{m \times n}$.

2. The adjoint of the right multiplication operator by a matrix $A \in \mathbb{R}^{m \times n}$ from $\mathbb{R}^{p \times m}$ to $\mathbb{R}^{p \times n}$ is the right multiplication by $A^t$.

3. The adjoint of the operator sending $X \mapsto X^t$ for $X \in \mathbb{R}^{m \times n}$ is again the operator $Y \mapsto Y^t$ for $Y \in \mathbb{R}^{m \times n}$. Adjoint is additive, and $(F \circ G)^t = G^t \circ F^t$ for two linear operators $F$ and $G$.

4. The adjoint of the left multiplication operator by $A \in \mathbb{R}^{p \times n}$ sending $X \in \text{Sym}_{k, \mathbb{R}}$ to $AX \in \mathbb{R}^{p \times n}$ is the operator sending $Y \mapsto \frac{1}{2}(A^t Y + Y^t A)$ for $Y \in \mathbb{R}^{p \times n}$. Conversely, the adjoint of the operator $Y \mapsto \frac{1}{2}(A^t Y + Y^t A) \in \text{Sym}_{k, \mathbb{R}}$ is the operator $X \mapsto AX$.

5. The adjoint of the left multiplication operator by $A \in \mathbb{R}^{p \times n}$ sending $X \in \text{Skew}_{k, \mathbb{R}}$ to $AX \in \mathbb{R}^{p \times n}$ is the operator sending $Y \mapsto \frac{1}{2}(A^t Y - Y^t A)$ for $Y \in \mathbb{R}^{p \times n}$. Conversely, the adjoint of the operator $Y \mapsto \frac{1}{2}(A^t Y - Y^t A) \in \text{Skew}_{k, \mathbb{R}}$ 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 $[F_1X \quad F_2X]$) is the map sending $\begin{bmatrix} A \\ B \end{bmatrix}$ to $F_1A + F_2B$.

Adjoint of the map sending $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ to $FX_1$ is the map $Y \mapsto \begin{bmatrix} F^1Y \\ 0 \end{bmatrix}$, and more generally a map sending a row block $X_1$ of a matrix $X$ to $FX_1$ is the map sending $Y$ to a matrix where the $i$-th block is $F^iY$, 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 $\text{Tr}(Ab^T) = \text{Tr}(a(A^Tb)^T)$, and $\text{Tr}_R(ab^T) = \text{Tr}_R(a(A^Tb)^T)$ 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:

$$\text{Tr}_R(F_1XA^t + F_2XB^t) = \text{Tr}_R((F_1A + F_2B)X^t)$$

5. APPLICATION TO STIEFEL MANIFOLD

The Stiefel manifold $\text{St}_{k,d,n}$ on $E = \mathbb{K}^{n \times d}$ is defined by the equation $Y^tY = I_d$, where the tangent space at a point $Y$ consists of matrices $\eta$ satisfying $\eta Y + Y \eta^t = 0$. We apply the results of section 2 for the full tangent bundle $\mathcal{H} = T\text{St}_{k,d,n}$. We can consider an ambient metric:

$$g(Y)\omega = \alpha_0 \omega + (\alpha_1 - \alpha_0) YY^t \omega = \alpha_0 (I_n - YY^t) \omega + \alpha_1 YY^t \omega$$

for $\omega \in \mathcal{E} = \mathbb{K}^{n \times d}$. It is easy to see $\omega_0 - YY^t \omega_0$ is an eigenvector of $g(Y)$ with eigenvalue $\alpha_0$, and $YY^t \omega_1$ is an eigenvector with eigenvalue $\alpha_1$, for any $\omega_0, \omega_1 \in \mathcal{E}$, and these are the only eigenvalues and vectors. Hence, $g(Y)^{-1} \omega = \alpha_0^{-1} (I_n - YY^t) \omega + \alpha_1^{-1} YY^t \omega$ and $g$ is a Riemannian metric if $\alpha_0, \alpha_1$ are positive. We can describe the tangent space as a nullspace of $J(Y)$ with $J(Y)\omega = \omega^tY + Y^t \omega \in \mathcal{E}_1 := \text{Sym}_{t,\mathbb{K},d}$. We will evaluate everything at $Y$, so we will write $J$ and $g$ instead of $J(Y)$ and $g(Y)$, etc. By proposition 6, $J^t a = (a Y^t)Y + Y a = 2 Ya$ for $a \in \mathcal{E}_1$. We have $g^{-1} J^t a = \alpha_0^{-1} 2 Ya + (\alpha_1 - \alpha_0^{-1}) 2 Ya = 2 \alpha_1^{-1} Ya$. Thus $J g^{-1} J^t a = J(2 \alpha_1^{-1} Ya) = 4 \alpha_1^{-1} a$ and by proposition 1:

$$\Pi g \nu = \nu - 2 \alpha_1^{-1} \frac{\alpha_1}{4} (Y \nu Y + YY^t \nu) = \nu - \frac{1}{2} (Y \nu Y + YY^t \nu)$$

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 $g$, $\Pi g \nu = (I_n - YY^t) \nu + \text{skew}_Y Y \nu$ and $g^{-1} \hat{f}_Y = \alpha_0^{-1} (I_n - YY^t) \hat{f}_Y + \alpha_1^{-1} YY^t \hat{f}_Y$

$$\Pi g \alpha_0^{-1} \hat{f}_Y = \alpha_0^{-1} (I_n - YY^t) \hat{f}_Y + \alpha_1^{-1} YY^t \hat{f}_Y$$

$$\alpha_0^{-1} \hat{f}_Y + \alpha_1^{-1} - \frac{2 \alpha_0^{-1}}{2} YY^t \hat{f}_Y - \frac{\alpha_1^{-1}}{2} Y \hat{f}_Y$$

If $\xi$ and $\eta$ are vector fields, $(D g) \eta = (\alpha_1 - \alpha_0)(Y^t \eta + Y \xi) \xi$. Using proposition 4, we can take the cross term $X(\xi, \eta) = (\alpha_1 - \alpha_0)(\xi \eta + \xi \eta) \eta$, thus:

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

By the tangent condition, $(\xi Y^t \eta + \eta Y^t \xi) = -(\xi \eta^t + \eta Y^t \xi - (\xi \eta + \eta \xi)Y)$, hence $K(\xi, \eta) = \frac{\alpha_1 - \alpha_0}{2} F$ with $F = Y(\xi \eta + \eta \xi) - 2(\xi \eta + \eta \xi)Y$, we see $Y^t F$ is symmetric so $\text{skew}_Y Y^t F = 0$,
therefore
\[
\Pi_g^{-1} F = \alpha_0^{-1}(I_n - YY^t)F = -2\alpha_0^{-1}(I_n - YY^t)(\xi^t\eta + \eta^t\xi)Y
\]
(5.3)
\[
\Pi_g^{-1}K(\xi, \eta) = \frac{\alpha_0 - \alpha_1}{\alpha_0}(I_n - YY^t)(\xi^t\eta + \eta^t\xi)Y
\]
Using \( g^{-1}J^t(J g^{-1}J)^{-1}(D\xi J)\eta = \frac{1}{2}Y(\xi^t\eta + \eta^t\xi) \) to evaluate eq. (2.13), the connection for two vector fields \( \xi, \eta \) is:
\[
\nabla_\xi\eta = D_\xi\eta + \frac{1}{2}Y(\xi^t\eta + \eta^t\xi) + \frac{\alpha_0 - \alpha_1}{\alpha_0}(I_n - YY^t)(\xi^t\eta + \eta^t\xi)Y
\]
(5.4)
With \( \Pi_0 = (I_n - YY^t) \) and let \( \hat{f}_{YY} \) be the ambient Hessian, from eq. (2.12):
\[
\text{rheat}_{\xi,\eta}^{11} = \Pi_g^{-1}xtrace(rheat_{\xi,\eta}^{02})\text{ by remark 1}
\]
(5.5)
\[
\text{rheat}_{\xi,\eta}^{11} = \Pi_{g^{-1}}Y(\hat{f}_{Y^t\xi} - \text{Tr}_{\hat{g}}(\hat{f}_{Y^t\xi}Y + \hat{f}_Y)) - \frac{\alpha_0 - \alpha_1}{\alpha_0}(I_n\hat{f}_{Y^t\xi} + Y\hat{f}_Y\Pi_0)\xi
\]
(5.6)
We note the term inside \( \Pi_{g^{-1}}Y \) is not unique as it can be modified by any expression sent to zero by \( \Pi_{g^{-1}} \). 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 \( YY^t(\xi^t\eta + \eta^t\xi) = \frac{1}{2}Y(\xi^t\eta + \eta^t\xi) \) 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 [19] we also provide computationally efficient closed-form geodesic formulas similar to those in [9].

6. Quotients of a Stiefel Manifold and Flag Manifolds

We will treat families of quotients of a Stiefel manifold, slightly more general than flag manifolds. Background materials for optimization on flag manifolds are in [22] [23], but the review below should be sufficient to understand the setup and the results. We generalize the formula for \( \text{rheat}_{\xi,\eta}^{11} \) in [29], Proposition 25, to the full family of metrics in [11] and provide a formula for \( \text{rheat}_{\xi,\eta}^{11} \).

Continuing with the setup in the previous section, consider a Stiefel manifold \( \text{St}_{m,n} \) (we will assume \( 0 < d < n \)). The metric induced by the operator \( g \) in eq. (5.1), \( \alpha_0 \text{Tr}_{\hat{g}}\omega_1^t\omega_2 + (\alpha_1 - \alpha_0)\text{Tr}_{\hat{g}}\omega_1^tYY^t\omega_2 \) with \( Y \in \text{St}_{m,n} \), \( \omega_1, \omega_2 \in \mathcal{E} \) is preserved if we replace \( Y, \omega_1, \omega_2 \) by \( YU, \omega_1U, \omega_2U, \) for \( U^tU = I_d \), or if we define the t-orthogonal group by \( U_{m,d} := \{ U \in \mathbb{K}^{d \times d} | U^tU = I_d \} \) then this is a group of isometries of \( g \). Therefore, any subgroup \( G \) of \( U_{m,d} \) acts on \( \text{St}_{m,n} \) by right-multiplication also preserves the metric, and if \( G \) is compact, we can consider the quotient manifold \( \text{St}_{m,n}/G \), identifying \( Y \in \text{St}_{m,n} \) with \( YU \) for \( U \in G \). The quotient manifold is useful for optimization problems with cost functions invariant under the action of the group \( G \). In that case, instead of optimization over \( \text{St}_{m,n} \), we can optimize over \( \text{St}_{m,n}/G \). The case of the Rayleigh quotient cost function \( \text{Tr}Y^tAY \) for \( Y \in \text{St}_{m,n} \), \( A \) is a positive-definite matrix is well-known, as the cost function is invariant under \( G = U_{m,d} \), we can optimize over the Grassmann manifolds \( \text{St}_{m,n}/G \). This section deals with the optimization of cost functions with a smaller group of symmetries \( G \) described below.
We will consider the case where $G$ is a group of block-diagonal matrices. Assume there is a sequence of positive integers $\mathbf{d} = \{d_1, \cdots, d_q\}$, $d_i > 0$ 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 $G = U_{d,d} = U_{d,d_1} \times U_{d,d_2} \cdots \times U_{d,d_q} \times \{I_{d_{q+1}}\}$ of $U_{d,d}$ of block-diagonal matrices $U$, with the $i$-th diagonal block from the top $U_{[ii]} \in U_{d,d_i}$, $1 \leq i \leq q$, and $U_{[q+1,q+1]} = I_{d_{q+1}}$. An element $U \in U_{d,d}$ has the form

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

When $q = 0$, we define $\mathbf{d} = \emptyset$ and $U_{d,d} = \{I_d\}$. We will consider the manifold $\text{St}_{d,d,n}/G$ for $G = U_{d,d}$.

Thus, when $\mathbf{d} = \emptyset$, this quotient is the Stiefel manifold $\text{St}_{d,d,n}$ itself, when $\mathbf{d} = \{d\}$, 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 $\text{Flag}(d_1, \cdots, d_q; n, \mathbb{K})$. Thus, 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 $\text{symf}$ acting on $\mathbb{K}^{d \times d}$, sending $A \in \mathbb{K}^{d \times d}$ to $A_{\text{symf}}$ such that $(A_{\text{symf}})_{ij} = \frac{1}{2}(A_{ij} + A_{ji})$ 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, $\text{symf}$ 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 $A = A_{ij} \in \mathbb{K}^{d \times d}$.

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

For the case $\mathbf{d} = \emptyset$ of the full Stiefel manifold, $\text{symf}$ is just $\text{sym}$, and for the case $\mathbf{d} = \{d\}$ of the Grassmann manifold, $\text{symf}$ is the identity map. We show these quotients, including Stiefel, Grassmann and flag manifolds, share similar Riemannian optimization settings.

**Theorem 2.** With the metric in eq. (5.1), the horizontal space $\mathcal{H}_Y$ at $Y \in \text{St}_{d,d,n}$ of the quotient $\text{St}_{d,d,n} \to \text{St}_{d,d,n}/U_{d,d}$ consists of matrices $\omega \in \mathcal{E} := \mathbb{K}^{n \times d}$ such that

$$
(Y^t \omega)_{\text{symf}} = 0
$$

or equivalently, $Y^t \omega$ is $t$-antisymmetric with first $q$ diagonal blocks $((Y^t \omega)_{\text{symf}})_{[ii]}$ 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

$$
\Pi_\mathcal{H}^{-1} \dot{f}_Y = \omega - Y(Y^t \omega)_{\text{symf}}
$$

$$
\Pi_\mathcal{H}^{-1} \dot{f}_Y = \omega - Y(Y^t \omega)_{\text{symf}}
$$
Let $\Pi_0 = I_n - YY^t$. For two vector fields $\xi, \eta$, the horizontal lift of the Levi-Civita connection and Riemannian Hessians are given by
\begin{equation}
\Pi_0 \nabla \xi \eta = D\xi \eta + Y(\xi^t \eta)^{\text{symf}} + \frac{\alpha_0 - \alpha_1}{\alpha_0} \Pi_0 (\xi^t \eta + \eta^t \xi) Y
\end{equation}
\begin{equation}
rhess^0_{f} (\xi, \eta) = \hat{f}_Y Y (\xi, \eta) - \text{Tr}_R (\hat{f}_Y^t Y (\xi^t \eta)^{\text{symf}} + \frac{\alpha_0 - \alpha_1}{\alpha_0} \Pi_0 (\xi^t \eta + \eta^t \xi) Y)
\end{equation}
\begin{equation}
rhess^1_{Y} \xi = \Pi_0 \Pi_0^{-1} (\hat{f}_Y Y \xi - \xi^t \hat{f}_Y Y)^{\text{symf}} - \frac{\alpha_0 - \alpha_1}{\alpha_0} \Pi_0 (\xi^t \hat{f}_Y Y^t + Y^t \hat{f}_Y Y \Pi_0) \xi
\end{equation}

Proof. First we note that $\text{symf}$ 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 $Y \in S^n_{t,k,d,n}$ under the action of $U(n, d)$ is $Y U(n, d)$, thus the vertical space consists of matrices of the form $Y D$ with $D$ is block-diagonal, $t$-skewsymmetric and $D_{[i,j]} = |d_{i,j}|$ if $i \neq j$ or $i = j = q + 1$. Since $g D = \alpha_1 Y D$, a horizontal vector $\omega$ satisfies $\text{skew}_t (Y^t \omega) = 0$ and $\text{Tr}_R (\omega Y D) = 0$. This shows $Y^t \omega$ has zero first $q$ diagonal blocks, hence $(Y^t \omega)^{\text{symf}} = 0$.

For the projection, we proceed like the Stiefel case, with the map $J \omega = (Y^t \omega)^{\text{symf}}$, mapping $E$ to $E_1 = \{ A \in \mathbb{K}^{t \times d} | A[i,j] = A^t[j,i], 1 \leq i \leq j \leq q + 1 \text{ or } i = j = q + 1 \}$. Since $\text{symf}$ is self-adjoint, $J^t A = Y A$ for $A \in E_1$. From here we get $(J g^{-1} J^t) A = \alpha_1^{-1} A$ and eq. (6.2) follows. Equation 6.3 is a substitution of $g^{-1} f_Y$ to eq. (6.2), noting $(Y^t (g^{-1} f_Y))^{\text{symf}} = \alpha_1^{-1} (Y^t f_Y)^{\text{symf}}$, using the eigen decomposition of $g$.

For the Levi-Civita connection, we use eq. (2.10). For two horizontal vector fields $\xi, \eta$, $(D_2 \Pi_0) \eta = -\xi (Y^t \eta)^{\text{symf}} - Y (\xi^t \eta)^{\text{symf}} = -Y (\xi^t \eta)^{\text{symf}}$ and $\Pi_0 = \Pi_0 \Pi_0$, where $\Pi$ is the Stiefel projection eq. (5.2), hence $\Pi_0 g^{-1} K (\xi, \eta) = \Pi_0 g^{-1} K (\xi, \eta) = Y (Y^t \Pi_0 g^{-1} K (\xi, \eta))^{\text{symf}}$. The last term vanishes from eq. (5.3), and we have eq. (6.4).

Equation (6.5) follows from eq. (2.12). We derive eq. (6.6) from remark 1 and self-adjointness of $\text{symf}$, expanding $\text{Tr}_R f_Y Y (\xi^t \eta)^{\text{symf}}$ to
\begin{equation}
\text{Tr}_R (Y^t f_Y) (\xi^t \eta)^{\text{symf}} = \text{Tr}_R (Y^t f_Y) (\xi^t \eta)^{\text{symf}} = \text{Tr}_R (Y^t f_Y) (\xi^t \eta)^{\text{symf}}
\end{equation}

7. Positive-definite matrices

Consider the manifold $S^+_{t,k,n}$ of $t$-symmetric positive-definite matrices in $\mathbb{K}^{n \times n}$. In our approach, we take $E = \mathbb{K}^{n \times n}$ with its Frobenius inner product. The metric is $(\xi, g \eta)_E = \text{Tr}(Y^t \eta Y^{-1})$, with the metric operator $g : \eta \mapsto Y^t \eta Y^{-1}$ for two vector fields $\xi, \eta$. The full tangent bundle $\mathcal{H} = T S^+_{t,k,n}$ is identified fiber-wise with the nullspace of the operator $J : \eta \mapsto J \eta = \eta - \eta^t$, with $\xi_J = \text{Skew}_{t,k,n}$. By item 5 in proposition 6, we have $J^t a = 2 a$ where $a$ is a $t$-antisymmetric matrix. From here $J g^{-1} J^t a = 4 Y a Y$ and (write $\Pi$ for $\Pi_0$):
\begin{equation}
\Pi \eta = \eta - g^{-1} J^t (J g^{-1} J^t)^{-1} J \eta = \eta - 2 Y (Y^{-1} (\eta - \eta^t) Y^{-1}) Y = \frac{1}{2} (\eta + \eta^t) = \text{symf} \eta
\end{equation}

Thus, the Riemannian gradient is $\Pi g^{-1} f_Y = \frac{1}{2} Y (f_Y + f_Y^t) Y$. Next, we compute $(D_2 g) \eta = D_{\xi \eta} \text{constant} (Y^{-1} \eta Y^{-1}) = -Y^{-1} \xi Y^{-1} \eta Y^{-1} - Y^{-1} \eta Y^{-1} \xi Y^{-1}$, where we keep $\eta$ constant in the derivative, as we evaluate $D_2 g$ as an operator-valued function. From here, $(D_0 g) \xi = (D_2 g) \eta$. We note for three vector fields $\xi, \eta, \xi_0$
\begin{equation}
\text{Tr}(Y^{-1} \xi_0 Y^{-1} \eta Y^{-1} + Y^{-1} \eta Y^{-1} \xi_0 Y^{-1})) = \text{Tr}(\xi_0 (Y^{-1} \eta Y^{-1} \xi Y^{-1} + Y^{-1} \xi Y^{-1} \eta Y^{-1}))
\end{equation}
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

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

(7.1)

$$\nabla_{\xi}\eta = D_\xi\eta - \frac{1}{2}(\xi Y^{-1}\eta + \eta Y^{-1}\xi)
$$

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

(7.2)

$$\tilde{f}_{YY}(\xi, \eta) + \frac{1}{2}\text{Tr}((\xi Y^{-1}\eta + \eta Y^{-1}\xi)\tilde{f}_Y) = \tilde{f}_{YY}(\xi, \eta) + \frac{1}{2}\text{Tr}((\tilde{f}_Y\xi Y^{-1} + Y^{-1}\xi \tilde{f}_Y)\eta)
$$

Using a symmetric version of remark [8], $\text{rhess}^{11} \xi = \Pi_2 g^{-1}\text{sym}_4(\tilde{f}_{YY}\xi + \frac{1}{2}(\tilde{f}_Y\xi Y^{-1} + Y^{-1}\xi \tilde{f}_Y))$. We get the following formula, as in [8]:

(7.3)

$$\text{rhess}^{11} \xi = Y\text{sym}_4(\tilde{f}_{YY}\xi)Y + \text{sym}_4(\xi\text{sym}_4(\tilde{f}_Y))Y$$

8. A FAMILY OF MACHINES FOR THE MANIFOLD OF POSITIVE-SEMDIFINITE MATRICES OF FIXED RANK

In [7], the authors defined a family of metrics on the manifold $S_{K,p,n}^+$ 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 $Y \in \text{St}_{K,p,n}(Y^tY = I_p)$ and $P$ is positive-definite of size $p \times p$, up to the equivalent relation $(Y, P) \sim (YU, UP^tU^t)$ for a matrix $U \in U_{K,p}$. (that means $U \in \mathbb{K}^{p \times p}$ and $UU^t = I_p$). So the manifold $S_{K,p,n}^+$ could be identified with the quotient space $(\text{St}_{K,p,n} \times S_{K,p}^+) / U_{K,p}$ of the product of the Stiefel manifold $\text{St}_{K,p,n}$ and the manifold of positive-definite matrices $S_{K,p}^+$ over the $t$-orthogonal group $U_{K,p}$. (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 $\mathbb{K}^{(n-p) \times p} \times \text{Sym}_{t,\mathbb{K},p}$ to $\mathcal{E}$, $N_1 : (B, D) \mapsto (Y_1 B, D)$, where the matrix $(Y_1 Y_1^t)$ is $t$-orthogonal. On the tangent space, [7] uses the metric $\text{Tr}(BB^t) + k \text{Tr}(DP^{-1}DP^{-1})$ for a positive number $k$. The action of the group gives us the vertical vectors $(Yq, Pq - qP)$ for a $t$-antisymmetric matrix $q$ such that $q + q^t = 0$. 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 $\Pi_H$. The horizontal lift of the Levi-Civita connection is $\Pi_H \nabla$, which we will denote by $\nabla^H$.

**Theorem 3.** Let $\mathcal{M} = \text{St}_{K,p,n} \times S_{K,p}^+ \subset \mathcal{E} := \mathbb{K}^{n \times p} \times \mathbb{K}^{p \times p}$ be an embedded ambient space of $\mathcal{M}$. Identifying the manifold $S_{K,p,n}^+$ of positive-semidefinite matrices with $B = (\text{St}_{K,p,n} \times S_{K,p}^+)/U_{K,p}$, where the pair $S = (Y, P)$ represents the matrix $YPY^t$ with $Y \in \text{St}_{K,p,n}$, $P \in S_{K,p}^+$, and the action of $U \in U_{K,p}$ sends $(Y, P)$ to $(YU, UP^tU^t)$. The self-adjoint metric operator

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

for $\omega = (\omega_Y, \omega_p) \in \mathcal{E} = \mathbb{K}^{n \times p} \times \mathbb{K}^{p \times p}$ defines the inner product $\text{Tr}(\alpha_0 \omega_Y + (\alpha_1 - \alpha_0)\omega_Y YY^t\omega_p + \beta \omega_p P^{-1}\omega_p P^{-1})$ on $\mathcal{E}$, which induces a metric on $\text{St}_{K,p,n} \times S_{K,p}^+$, invariant under the action of $U_{K,p}$ and induces a quotient metric on $S_{K,p,n}^+$. Its tangent bundle $TS_{K,p,n}^+$ lifts to the subbundle $\mathcal{H} \subset T(\text{St}_{K,p,n} \times S_{K,p}^+)$ 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:

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

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

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

where $Y_\perp$ is orthogonal complement matrix to $Y$, $(Y|Y_\perp) \in U_{\mathbb{K},n}$. 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

$$
\Pi_H(S)(\omega_Y, \omega_P) = (\beta Y(P^{-1} D - DP^{-1}) + \omega_Y - YY^\flat \omega_Y, \alpha_1 D)
$$

(8.4) with $D = D(P)\omega := L(P)^{-1} \text{sym}_t(\omega_P + Y^\flat \omega_Y P - P Y^\flat \omega_Y)$

where $L(P)X := (\alpha_1 - 2\beta)X + \beta (PX P^{-1} + P^{-1} XP)$

The operator $L(P)$ could be inverted by proposition 3. The Riemannian Hessian could be evaluated by eq. (8.2), and the lift of the Levi-Civita connection is given by

$$
(8.5) \quad \nabla^H \xi \eta := \Pi_H \nabla \xi \eta = D\xi \eta - (D\xi \Pi_H)\eta + \Pi_H g^{-1} K(\xi, \eta)
$$

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

$$
(8.6) \quad K(\xi, \eta)_Y = \frac{\alpha_1 - \alpha_0}{2} (Y(\eta^t_Y \xi_Y + \xi^t_Y \eta_Y) - 2(\eta Y \xi_Y + \xi Y \eta_Y)Y)
$$

$$
K(\xi, \eta)_P = -\beta \frac{1}{2} (P^{-1} \eta_P P^{-1} \xi_P P^{-1} + P^{-1} \xi_P P^{-1} \eta_P P^{-1})
$$

Set $\hat{D} = (D\xi D)\omega = L(P)^{-1} \text{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 DP^{-1} + P^{-1} D \xi_P - P DP^{-1} \xi_P P^{-1} - P^{-1} \xi_P P^{-1} DP)$, we have

$$
(8.7) \quad (D\xi \Pi_H)\omega_Y = \beta \xi_Y (P^{-1} D - DP^{-1}) + \beta Y(P^{-1} \hat{D} - \hat{D} P^{-1} + DP^{-1} \xi_P P^{-1} - P^{-1} \xi_P P^{-1} D) - (\xi Y^t + Y \xi^t_Y)\omega_Y
$$

Thus, the Hessian could be evaluated at $O(np^2)$-complexity, by operator composition using eq. (8.1). Note that $(D\xi \Pi_H)\eta$ could be further simplified if $\eta = (\eta_Y, \eta_P)$ is a horizontal tangent vector, as $D = \frac{1}{\alpha_1} \eta_P$ in that case.

Proof. We have $\text{Tr}_{\mathbb{K}}(\alpha_1 \eta^t_Y Y q + \beta(\eta^t_P P^{-1} (Pq - qP) P^{-1})) = 0$ for a tangent vector $(\eta_Y, \eta_P)$ and a $t$-antisymmetric matrix $q$ from the horizontal condition. Using proposition 3 this means skew-$t(\alpha_1 \eta^t_Y Y + \beta P^{-1} \eta^t_P - \beta P^{-1} \eta^t_P P^{-1}) = 0$. Using the fact that $\eta^t_P = \eta_P$ and $\eta^t_P Y$ is $t$-antisymmetric, we have eq. (8.2).

It is clear that $N(B, D)$ satisfies this equation and is one-to-one: if $N(B, D) = 0$ then immediately $D = 0$, and $B = 0$ since $Y^t Y_\perp = I$. It is onto the tangent space by a dimension count. The adjoint $N^t = (N^t_B, N^t_D)$ has two components corresponding
positive-definite matrix $P$ has the following unique solution that could be computed at the Schur decomposition. Substitute on the left-hand and right-hand sides of that equation we get

$$N^t(\omega_Y, \omega_P)_B = Y_{\perp}^t\omega_Y$$

$$N^t(\omega_Y, \omega_P)_D = \text{sym}_i(\alpha_1\omega_P + \beta P^{-1}Y^t\omega_Y - \beta Y^t\omega_Y P^{-1})$$

$$N^t g(\omega_Y, \omega_P)_B = \alpha_0 Y_{\perp}^t\omega_Y$$

$$N^t g(\omega_Y, \omega_P)_D = \alpha_1 \text{sym}_i(P^{-1}\omega_P P^{-1} + P^{-1}Y^t\omega_Y - Y^t\omega_P P^{-1})$$

$$(N^t g N(B, D))_B = \alpha_0 B$$

$$(N^t g N(B, D))_D = \alpha_1 \text{sym}_i(P^{-1}(\alpha_1 D) P^{-1} + P^{-1}Y^t(\beta Y(P^{-1}D - DP^{-1}) + Y_{\perp} B) - Y^t(\beta Y(P^{-1}D - DP^{-1}) + Y_{\perp} B) P^{-1})$$

$$= \alpha_1 \text{sym}_i(\alpha_1 P^{-1}DP^{-1} + \beta P^{-2}D - \beta P^{-1}DP^{-1} - \beta P^{-1}DP^{-1} + \beta DP^{-2})$$

$$= \alpha_1 \beta P^{-1}((\alpha_1 - 2\beta)D + \beta PDP^{-1} + \beta P^{-1}DP^{-1})$$

$$= \alpha_1 \beta P^{-1}L(P)DP^{-1}$$

Hence $(N^t g N)^{-1}(B, D) = (\alpha_0^{-1}B, (\alpha_1^{-1}I(P)^{-1}(PDP))$ and $(N^t g N)^{-1}N^t g (\omega_Y, \omega_P) = (Y_{\perp}^t\omega_Y, D)$. The projection formula eq. (8.4) is just $N(N^t g N)^{-1}N^t g$. The formulas for $K$ follow from the corresponding Stiefel and positive-definite manifold formulas. For $D(\Pi_d)$, we take directional derivative of eq. (8.4) using standard matrix calculus, the only difficulty is $\hat{D} = (D_{\xi}D)\omega$. We evaluate it by evaluating $D_{\xi}(L(P)D)\omega$ as

$$(\alpha_1 - 2\beta)(\hat{D} + \beta(\hat{D}PDP^{-1} + P^{-1}\hat{D}P)) + \beta(\hat{D}PDP^{-1} + P^{-1}\hat{D}P) - \beta(PDP^{-1}\xi_P P^{-1} + P^{-1}\xi_P P^{-1}DP) = \text{sym}_i(\xi_P^t\omega_Y - P\xi_P^t\omega_Y + Y^t\omega_Y \xi_P - \xi_P Y^t\omega_Y)$$

by differentiating the equation for $D$. From here, we get the equation for $\hat{D}$. □ □

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

**Proposition 7.** Let $P \in \text{Sym}_{4,4}$ be an $t$-symmetric matrix having the eigenvalue decomposition $P = U\Lambda U^t$ with eigenvalues $(\Lambda_i)^p_{i=1}$ and $UU^t = I$. Let the set of coefficients $(c_{st})_{-k \leq s, t \leq k}$ 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 $(\Lambda_i, \Lambda_j)$ of eigenvalues of $P$, then the equation

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

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

$$X = U((U^tBU)/M)U^t$$

with $M = (M_{ij})_{i=1, j=1}^{i=p, j=p}$ and $\slash$ denotes the by-entry division. In particular, for a positive-definite matrix $P$ and positive scalars $(\alpha, \beta)$, the equation $(\alpha - 2\beta)X + \beta(P^{-1}XP + PXP^{-1}) = B$ 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 $P = U\Lambda U^t$ to eq. (8.8) and multiply $U^t$ and $U$ on the left-hand and right-hand sides of that equation we get $\sum_{s=-k, t=-k}^{s=k, t=k} c_{st}\Lambda^sU^tXUA^t = U^tBU$, which is equivalent to $(U^tXU)_{ij}M_{ij} = (U^tBU)_{ij}$ or $U^tXU = (U^tBU)/M$, $\square$
and we have eq. (8.9). If \( \alpha \) and \( \beta \) are positive then \( \alpha_1 + \beta(\Lambda_j^{-1}A_j + \Lambda_A^{-1} - 2) > 0 \) by the AGM-inequality. The eigenvalue decomposition has \( O(p^3) \) cost. □ □

Horizontal lifts of geodesics on \( S^+_{K,p,n} \) are geodesics on \( S_{K,p,n} \times S^+_{K,p} \). It is clear \( S^+_{K,p,n} \) is a complete Riemannian manifold under this quotient metric, as both factors above are. If \( (\eta_Y, \eta_P) \) is a horizontal tangent vector at \( S = (Y, P) \), a horizontal geodesic \( \gamma \) 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 section 5 described in 19.

9. Implementation and numerical experiments

We developed a Python package based on Pymanopt and Manopt [27, 8] implementing the manifolds with metrics considered in this paper, for both the real and complex cases in the package [20]. The basic optimization algorithms are from [27, 8], and the methods described in this paper provide the translation from ambient gradient and Hessian to the Riemannian ones. Besides the numerical implementation, we also include notebooks showing symbolic calculus results for each manifold, and scripts with numerical tests for the manifolds, including geodesics in most cases. We implemented the main formulas in theorem 1 and proposition 3 as template methods in the base class NullRangeManifold. To use the method described here for a new manifold, it is convenient to derive from NullRangeManifold. Then, the user needs to implement the constraint operator \( \mathbf{J} \), its transpose and derivative, the metric operator \( g \), its inverse, derivative, and the third Christoffel term \( \mathbf{X} \), as well as a method to solve \( \mathbf{J} g^{-1} \mathbf{J}^t \) (defaulted to use a conjugate-gradient solver otherwise). A retraction is also required. The template automatically provides the projection, Riemannian gradient, and Hessian derived in this paper.

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 based on the manifold. Using the method in this paper but with details appearing elsewhere, we also derive and implement optimization framework for fixed-rank matrix manifolds. For flag manifolds, we initially constructed a large family of metrics (including the metrics in section 6 in [20]). We also implemented the metrics in section 6 separately in [21], which we use below. For testing, we numerically verify the projection satisfying the nullspace condition. We also tested metric compatibility and torsion-freeness of the covariance derivative and confirmed the relation between bilinear and operator Hessians. As the manifolds considered here are constructed from Stiefel or positive-definite matrix manifolds, both have effective retractions, we use these retractions in our implementation. As we would like to 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, which handles them comfortably.

For flag manifolds, we optimize the function \( f(Y) = \text{Tr}_R((Y \Lambda Y^t A)^2) \) over matrices \( Y \in \text{St}_{d,n} \). Here, \( \Lambda \) is a positive-definite matrix, \( d < n \) are two positive integers, \( \hat{d} = (d_1, \cdots, d_q) \) is a partition of \( d \) with \( \sum_{i=1}^q d_i = d, \Lambda = \text{diag}(\lambda_1 I_{d_1}, \cdots, \lambda_q I_{d_q}) \) for positive numbers \( \lambda_1, \cdots, \lambda_q \). This cost function is invariant under the action of of \( U_{d,n} \); this could be considered as a function on the flag manifold \( \text{St}_{d,n}/U_{d,n} \). The Euclidean gradient is \( 4(AYA)Y^t(AYA) \), and the Euclidean Hessian is computed by routine matrix calculus. For testing, we consider \( d = 60, \hat{d} = (30, 20, 10), n = 1000 \).
and use a trust-region solver. In this case, there is no noticeable variation in time when different values of $\alpha$ are used, typically a few seconds on a free colab engine in the notebook [21]. 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 non-linear weighted PCA (principal component analysis) problem, which could be solved by optimizing over the positive-semidefinite matrix manifold. Given a symmetric matrix $A \in \text{Sym}_{T,\mathbb{R},n}$ and a weight vector $W \in \mathbb{R}^n$ we want to minimize the cost function

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

of a positive-semidefinite matrix $S = YPY^T \in S^+_{\mathbb{R},p,n}$, with $Y \in \text{St}_{\mathbb{R},p,n}, P \in S^+_{\mathbb{R},p}$. Here, $W_d$ denotes the diagonal matrix diag($W$). When $W$ has identical weight $\lambda$, $W_d = \lambda I_n$, expanding the cost function, we need to minimize $\text{Tr}(P^2 - 2\text{Tr}(Y^TA)PYP)$ in $Y$ and $P$, which implies $P = Y^TY$. Thus, the problem is optimizing $-\text{Tr}(Y^TA)^2$ 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 non identical weights, it is difficult to reduce the problem to a Stiefel manifold, hence we optimize over the positive-semidefinite manifold, with a trust-region solver.

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

$$(D_\xi \hat{f})(Y, P) = \text{Tr} W_d(-A\xi_YPY^T - AY\xi_PY^T - AYP\xi_Y^T - \xi_YPY^TA - Y\xi_PY^TA - YP\xi_Y^TA + \xi_YP^2Y^T + Y(\xi_PP + P\xi_P)Y^T + YP^2\xi_Y^T)$$

We have $\text{Tr} W_d(-A\xi_YPY^T - AY\xi_PY^T - AYP\xi_Y^T - \xi_YPY^TA - YP\xi_Y^TA) = -2\text{Tr}(AW_d + W_dA)YP\xi_Y^T$, $\text{Tr} W_d(YP^2Y^T + YP^2\xi_Y^T) = 2\text{Tr} W_dYP^2\xi_Y^T$ and similar equalities for $\xi_P$ give us

$$\text{grad} \hat{f} = (-4\text{sym} \tau(AW_d)YP + 2W_dYP^2, -2\text{sym} \tau(Y^TW_d(AY - YP)))$$
Figure 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.

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

$$
\text{hess} \hat{f}(\xi) = -4\text{sym}_T(AW_d)(\xi_Y P + YY_P) + 2W_d\xi_Y P^2 + 2W_d Y(\xi P + P Y) + 2W_d\xi_Y Y_P + 2W_d Y(\xi P + P Y);
$$

$$
-2\text{sym}_T(\xi_Y W_d(AY - Y P)) - 2\text{sym}_T(Y^T W_d(A\xi Y - \xi Y P - Y P)).
$$

The Riemannian gradient is computed as $\Pi_{H,g}^{-1} \text{grad} \hat{f}$, with $\Pi_{H,g}$ is given by eq. (8.4), the Riemannian Hessian is computed from eq. (2.11), with all components given in theorem 3. We use the built-in trust region solver in [27].

In our experiment (implemented in the notebook colab/WeightedPCA.ipynb in [20]), we take $n = 1000, p = 50$, with $A$ and $W$ generated randomly. To find the optimum $S = YPY^T$, we optimize with $\alpha_0 = \alpha_1 = 1$, with $\beta$ is 0.1 for the first 20 iterations, $\beta = 10$ for the next 20 and $\beta = 30$ for the remaining iterations. This choice of $\beta$ comes from our limited experiments, we find varying $\beta$ has a strong effect on the speed of convergence, and updating $\beta$ as such gives better convergence rates than a static $\beta$. Philosophically, the small starting $\beta$ 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 and statistical learning. The optimization platform for positive-semidefinite matrices should help to learn sparse plus low-rank probability densities in statistical problems. We hope the research community will find the method useful in future works.

References

1. P.-A. Absil, R. Mahony, and R. Sepulchre, Optimisation algorithms on matrix manifolds, Princeton University Press, Princeton, NJ, USA, 2007.
2. P. A. Absil, Robert Mahony, and Jochen Trumpf, An extrinsic look at the Riemannian Hessian, Geometric Science of Information (Frank Nielsen and Frédéric Barbaresco, eds.), Springer Berlin Heidelberg, 2013, pp. 361–368.
3. Roy L. Adler, Jean-Pierre Dedieu, Joseph Y. Margulies, Marco Martens, and Mike Shub, Newton’s method on Riemannian manifolds and a geometric model for the human spine, IMA Journal of Numerical Analysis 22 (2002), no. 3, 359–390.
4. A. C. Aitken, On Least Squares and Linear Combination of Observations, Proceedings of the Royal Society of Edinburgh 55 (1936), 42–48.
5. R. H. Bartels and G. W. Stewart, Solution of the matrix equation $ax + xb = c$, Commun. ACM 15 (1972), no. 9, 820–826.
6. Rajendra Bhatia and Peter Rosenthal, How and why to solve the operator equation $AX - XB = Y$, Bulletin of the London Mathematical Society 29 (1997), 1–21.
7. Silvère Bonnabel and Rodolphe Sepulchre, Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank, SIAM Journal on Matrix Analysis and Applications 31 (2010), no. 3, 1055–1070.
8. N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, Manopt, a Matlab toolbox for optimization on manifolds, Journal of Machine Learning Research 15 (2014), 1455–1459.
9. Alan Edelman, Tomás A. Arias, and Steven T. Smith, The geometry of algorithms with orthogonality constraints, SIAM J. Matrix Anal. Appl. 20 (1999), no. 2, 303–353.
10. D. Gabay, Minimizing a differentiable function over a differential manifold, Journal of Optimization Theory and Applications 37 (1982), no. 2, 177–219.
11. Knut Hüper, Irina Markina, and Fátima Silva Leite, A Lagrangian approach to extremal curves on Stiefel manifolds, Journal of Geometric Mechanics 13 (2021), no. 1, 55–72.
12. Ben Jeuris, Raf Vandebril, and Bart Vandereycken, A survey and comparison of contemporary algorithms for computing the matrix geometric mean, Electron. Numer. Anal. 39 (2012), 379–402.
13. M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre, Low-rank optimization on the cone of positive semidefinite matrices, SIAM Journal on Optimization 20 (2010), no. 5, 2327–2351.
14. H. Kasai and B. Mishra, Low-rank tensor completion: a Riemannian manifold preconditioning approach, Proceedings of Machine Learning Research, vol. 48, 20–22 Jun 2016, pp. 1012–1021.
15. J. H. Manton, Optimization algorithms exploiting unitary constraints, IEEE Transactions on Signal Processing 50 (2002), no. 3, 635–650.
16. Nina Miolane, Alice Le Brigant, Johan Mathe, Benjamin Hou, Nicolas Guigui, Yann Thanverdas, Stefan Heyder, Olivier Peltre, Niklas Koep, Hadi Zaatiti, Hatem Hajri, Yann Cabanes, Thomas Gerald, Paul Chauchat, Christian Shewmake, Bernhard Kainz, Claire Donnat, Susan Holmes, and Xavier Pennec, Geomstats: A Python Package for Riemannian Geometry in Machine Learning, 2020.
17. Bamdev Mishra, Gilles Meyer, Silvère Bonnabel, and Rodolphe Sepulchre, Fixed-rank matrix factorizations and Riemannian low-rank optimization, Computational Statistics 29 (2014), 591, 621.
18. Bamdev. Mishra and Rodolphe. Sepulchre, Riemannian preconditioning, SIAM Journal on Optimization 26 (2016), no. 1, 635–660.
19. D. Nguyen, Closed-form geodesics and trust-region method to calculate Riemannian logarithms on Stiefel and its quotient manifolds, 2021.
20. Du Nguyen, Project ManNullRange, https://github.com/dnguyend/ManNullRange, 2020.
21. ______, Project SimpleFlag, https://github.com/dnguyend/SimpleFlag, 2021.
22. Yasunori Nishimori, Shotaro Akaho, and Mark D. Plumbley, Riemannian optimization method on the flag manifold for independent subspace analysis, Independent Component Analysis and Blind Signal Separation (Justinian Rosca, Deniz Erdogmus, José C. Príncipe, and Simon Haykin, eds.), Springer Berlin Heidelberg, 2006, pp. 295–302.
23. B. O’Neill, Semi-riemannian geometry with applications to relativity, Pure and Applied Mathematics, vol. 103, Academic Press, Inc, New York, NY, 1983.
24. X. Pennec, P. Fillard, and N. Ayache, A Riemannian Framework for Tensor Computing, Int J Comput Vision (2006), 41–66.
25. S. T. Smith, Covariance, subspace, and intrinsic Cramér—Rao bounds, IEEE Transactions on Signal Processing 53 (2005), no. 5, 1610–1630.
26. Suvrit Sra and Reshad Hosseini, Conic geometric optimization on the manifold of positive definite matrices, SIAM Journal on Optimization 25 (2015), no. 1, 713–739.
27. James Townsend, Niklas Koep, and Sebastian Weichwald, *Pymanopt: A Python toolbox for optimization on manifolds using automatic differentiation*, J. Mach. Learn. Res. **17** (2016), no. 137, 1–5.

28. Bart Vandereycken, P.-A. Absil, and Stefan Vandewalle, *A Riemannian geometry with complete geodesics for the set of positive semidefinite matrices of fixed rank*, IMA Journal of Numerical Analysis **33** (2012), no. 2, 481–514.

29. Ke Ye, Ken Sze-Wai Wong, and Lek-Heng Lim, *Optimization on flag manifolds*, Mathematical Programming, Series A (2021).

*Email address:* nguyendu@post.harvard.edu