Preconditioned Riemannian Optimization on the Generalized Stiefel Manifold

Boris Shustin and Haim Avron
Tel Aviv University
June 23, 2021

Abstract

Optimization problems on the generalized Stiefel manifold (and products of it) are prevalent across science and engineering. For example, in computational science they arise in the symmetric (generalized) eigenvalue problem, in nonlinear eigenvalue problems, and in electronic structures computations, to name a few problems. In statistics and machine learning, they arise, for example, in various dimensionality reduction techniques such as canonical correlation analysis. In deep learning, regularization and improved stability can be obtained by constraining some layers to have parameter matrices that belong to the Stiefel manifold. Solving problems on the generalized Stiefel manifold can be approached via the tools of Riemannian optimization. However, using the standard geometric components for the generalized Stiefel manifold has two possible shortcomings: computing some of the geometric components can be too expensive and converge can be rather slow in certain cases. Both shortcomings can be addressed using a technique called Riemannian preconditioning, which amounts to using geometric components derived using a preconditioner that defines a Riemannian metric on the constraint manifold. In this paper we develop the geometric components required to perform Riemannian optimization on the generalized Stiefel manifold equipped with a non-standard metric, and illustrate theoretically and numerically the use of those components and the effect of Riemannian preconditioning for solving optimization problems on the generalized Stiefel manifold.

1 Introduction

In this paper we consider large-scale optimization problems on the generalized Stiefel manifold (and products of it), i.e. optimization with constraint spaces defined via generalized orthogonality constraints. One well known example of a problem with a generalized orthogonality constraint is the problem of finding the dominant generalized eigenspace of a symmetric positive-definite (SPD) matrix pencil. Indeed, given a pair of SPD matrices $A, B \in \mathbb{R}^{d \times d}$, minimizers of $-\text{Tr}(X^TAX)$ subject to $X^TBX = I_p$ (where $X \in \mathbb{R}^{d \times p}$) are bases for the subspace spanned by the $p$ generalized eigenvectors that correspond to the $p$ largest generalized eigenvalues of the pencil $(A, B)$ (this is a consequence of the Courant–Fisher characterization of generalized eigenvalues). More generally, problems with (generalized) orthogonality are prevalent across science and engineering. Examples include, the Trust-Region Subproblem, Canonical Correlation Analysis (CCA) [28], and Fisher Linear Discriminant Analysis [21].

Some optimization problems with generalized orthogonality constraints can be reformulated as (generalized) eigenvalue problems or (weighted) Singular Value Decomposition (SVD) problems. This is true for some of the cases mentioned in the previous paragraph. For example, CCA on a pair of matrices $(X, Y)$ amounts to computing the SVD of $P^TQ$ where $P$ and $Q$ are orthonormal matrices whose column space spans the column space of $X$ and $Y$ (respectively) [8]. This allows one to use direct methods, but that is unrealistic for large scale problems.
Using iterative method in lieu of direct methods is a common modus operandi for handling large scale problems. A natural framework for solving optimization problems with generalized orthogonality constraints is Riemannian optimization [19][11]. Indeed, when we have a single generalized orthogonality constraint of the form $X^TBX = I_p$, e.g. we want to minimize $f(X)$ s.t. $X^TBX = I_p$, one can impose the structure of a smooth manifold on the constraint set, thereby obtaining the generalized Stiefel manifold

$$\text{St}_B(p, d) := \{ X \in \mathbb{R}^{d \times p} : X^TBX = I_p \}. \quad (1.1)$$

(see [11] Propositions 3.3.3 and 3.3.4), and use Riemannian optimization to minimize $f(X)$ s.t. $X \in \text{St}_B(p, d)$. If we have $k > 1$ generalized orthogonality constraints, e.g. minimizing $f(X_1, \ldots, X_k)$ s.t. $X_i \in \text{St}_B(p_i, d_i) (i = 1, \ldots, k)$, as is the case in CCA (for $k = 2$), then each of the constraints constrain a disjoint set of variables, and the constraints are separable, so they define a product of generalized Stiefel manifolds, which is a smooth manifold as well, so Riemannian optimization can again be used.

In order to use Riemannian optimization for a constraint of the form $X \in \text{St}_B(p, d)$ we must further impose a Riemannian metric on the tangent bundle of $\text{St}_B(p, d)$. The standard Riemannian metric for $\text{St}_B(p, d)$ is the one naturally inherited by using the inner product $(U, V)_B = \text{Tr} (U^TBV)$ on $\mathbb{R}^{d \times p}$ (see [11] Section 3.6) for explanation on how a Riemannian metric is inherited from an ambient space in a natural way). Indeed, for the Stiefel manifold, i.e., when $B = I_d$, reference to the last metric as the standard metric appears in the seminal work of Edelman, Arias and Smith [19], and this is also the metric used in the implementation of the generalized Stiefel manifold in MANOPT [10]. Some of the geometric components for working with $\text{St}_B(p, d)$ equipped with the standard metric appears in [19] Section 4.5], while MANOPT implements all the geometric components, but without providing a reference.

This paper is motivated by the observation that using the standard metric in the context of Riemannian optimization with generalized orthogonality constraints has one severe shortcoming: the computation of some of the geometric components necessary for Riemannian optimization, e.g. the Riemannian gradient and Hessian, requires taking products with the inverse of $B$. Oftentimes, computing $B$ and its inverse is as expensive as the direct method. In such cases there is no reason to use Riemannian optimization as long as the standard metric is used. Another issue with using the standard metric is that in some cases it is suboptimal and using it will lead to slow convergence.

In this paper we propose to endow $\text{St}_B(p, d)$ with a metric inherited by the inner product $(U, V)_{M_X} = \text{Tr} (U^T M_X V)$ on $\mathbb{R}^{d \times p}$ for some smooth mapping $X \mapsto M_X$ that maps a $X \in \text{St}_B(p, d)$ to a SPD matrix $M_X$. Using such a mapping is an instance of so-called Riemannian preconditioning [37], so we call the mapping $X \mapsto M_X$ a preconditioning scheme. Indeed, using the metric defined by the mapping $X \mapsto M_X$ still requires computing $M_X$ in every iteration, and taking products with its inverse, however one is free to design the mapping so that $M_X$ can always be cheaply decomposed. On flip side, as we discuss later, one would like $M_X$ to well approximate $B$, or some other metric for which we can ensure well conditioning of the Riemannian Hessian at the optimum. Thus in designing the mapping $X \mapsto M_X$ we have the same tradeoffs as when designing a preconditioner for solving linear systems using a Krylov method.

In order to use Riemannian optimization with a preconditioning scheme, one needs to implement all the necessary geometric components for Riemannian optimization on $\text{St}_B(p, d)$ endowed with the metric defined by $X \mapsto M_X$. The majority of this paper is devoted to developing these geometric components. We complement these developments by considering the use of our approach on a couple of simple theoretical examples, and on the problem of finding the top canonical correlation between two datasets (which we explore both theoretically and numerically).

### 1.1 Related Work

**Riemannian Optimization.** Riemannian optimization is an approach for solving constrained optimization problems in which the constrains form a smooth manifold (e.g., nonlinear differentiable equality constraints). It is based on extending classical algorithms for unconstrained optimization on $\mathbb{R}^n$ (or any other vector space...
equipped with an inner product), by generalizing the main components needed to apply these algorithms to search spaces that form smooth manifolds. Some early works are \[35\] \[22\] \[42\]. A more recent and detailed introduction can be found in \[1\] and in \[11\].

**Riemannian Optimization on the (Generalized) Stiefel Manifold.** Optimization under orthogonality constraints are prevalent in many applications across science, naturally giving rise to Riemannian optimization on the (generalized) Stiefel manifold. Using Riemannian optimization to solve problems under orthogonality constraints was considered in the seminal work of Edelman et al. \[19\], and in particular the components of the Stiefel manifold were developed with the standard (and also the canonical) metric. Some recent works include \[43\] \[51\] \[34\], where the Cayley transform is used to define a retraction map which leads to more efficient algorithms. Another improved retraction computation is proposed in \[10\], where Sato and Aihara proposed a Cholesky QR-based retraction on the generalized Stiefel manifold. In \[40\], Kaneko et al. presented algorithms to compute inverses of several retractions on the Stiefel manifold in order to solve empirical arithmetic averaging problems over the Stiefel manifold. Also, several optimization algorithms for non-smooth optimization were developed on the Stiefel manifold such as a proximal gradient method and a fast iterative shrinkage-thresholding algorithm (FISTA \[5\]), see for example \[14\] \[29\]. Also in the context of this paper, a Riemannian optimization approach for adaptive CCA on a product manifold of two generalized Stiefel manifold was proposed in \[49\]. Unlike our work, all the aforementioned works use either the standard or the canonical metric when optimizing on the (generalized) Stiefel manifold.

**Riemannian Preconditioning.** In the context of Riemannian optimization, it is well-known that the condition number of the Riemannian Hessian at the optimum is highly indicative of the asymptotic convergence rate of Riemannian optimization (e.g., \[1\] Theorem 4.5.6, Theorem 7.4.11, Equation 7.5). If the objective function is convex (in the Riemannian sense \[46\] Chapter 3.2) then there also exist global convergence results depending on the condition number of the Riemannian Hessian at all the points on the manifold (e.g., \[46\] Chapter 7, Theorem 4.2), however these results are not applicable to optimization on the generalized Stiefel manifold, since every continuous and convex function (in the Riemannian sense) on the Stiefel manifold is constant.

The relation between convergence rate and condition number of the Riemannian Hessian at the optimum motivates adjusting the metric based on the cost or constraints, and this approach to preconditioning was presented in several works, see e.g., \[39\] \[36\] \[41\] \[50\]. Most of the aforementioned works attempt to lower the condition number of the Riemannian Hessian at the optimum by approximating the Euclidean Hessian of the cost function. However, it is possible for the Riemannian Hessian and the Euclidean Hessian to be very far from each other even for simple examples (see Section \[4\]). In \[37\], Mishra and Sepulchre showed that carefully selecting the metric based on both the cost and the constraint (inspired by the Lagrangian) used in Riemannian optimization affects convergence \[37\] of Riemannian steepest-descent (the iterations become a version of a Riemannian quasi-Newton close to the optimum). They demonstrated this technique on a quotient manifold (generalized Grassmann manifold) and on the fixed rank manifold. Unlike \[37\], we develop explicit components of Riemannian optimization on the generalized Stiefel manifold with non-standard metric and consider their costs with respect to the choice of metric (see Section \[3\]). This allows the use of various algorithms for smooth Riemannian optimization, e.g, conjugate-gradient, trust-region, etc. We also motivate the choice of metric by the condition number of the Riemannian Hessian at the optimum.

Another similar view of Riemannian preconditioning in the sense of Riemannian metric selection, which is specific for the Riemannian trust-region algorithm, is to precondition the solver used to solve the Trust-Region Subproblem \[37\]. The aforementioned preconditioning approach generalizes the preconditioning strategy for the unconstrained trust-region. Another example of using Riemannian preconditioning for the Trust-Region Subproblem can be found in \[38\].

A different approach for preconditioning of Riemannian methods can be found in \[32\] where linear systems with tensor product structure are considered. That paper proposed a Riemannian analogue to the
preconditioned Richardson method for Euclidean optimization based on the truncated Richardson iteration. Similarly to Euclidean Preconditioned Richardson, in each iteration the search direction is multiplied by an inverse of a SPD preconditioner (and then projected to the tangent space). Another method proposed in [32] is an approximate Riemannian Newton method where the search direction is determined by an equation evolving an approximation to the Riemannian Hessian (known as constrained Gauss–Newton, see e.g., [9]), and a preconditioning term replacing a component in that equation.

2 Preliminaries

2.1 Notation and Basic Definitions

We denote scalars using lower case Greek letters or using \(x, y, \ldots\). Vectors are denoted by \(\mathbf{x}, \mathbf{y}, \ldots\) and matrices by \(\mathbf{A}, \mathbf{B}, \ldots\) or upper case Greek letters. Tangent vectors (of a manifold) are denoted using lower case Greek letters with a subscript for the point on the manifold for which they correspond (e.g., \(\eta_{\mathbf{x}}\)). Normal vectors (of a manifold) are denoted using bold lower and upper case letters with a subscript for the point on the manifold for which they correspond (e.g., \(\mathbf{u}_{\mathbf{x}}\)). Vector fields on a manifold are denoted using lower case Greek letters with brackets indicating the point on the manifold they correspond (e.g., \(\eta(\mathbf{x})\)). Normal vector fields on a manifold are denoted using bold lower and upper case letters with brackets indicating the point on the manifold they correspond (e.g., \(\mathbf{u}(\mathbf{x})\)). We use the convention that vectors are column-vectors.

We denote by \((\cdot, \cdot)_{\mathbf{C}}\) the inner-product with respect to a matrix \(\mathbf{C}\): for vectors \(\mathbf{u}\) and \(\mathbf{v}\), \((\mathbf{u}, \mathbf{v})_{\mathbf{C}} := \mathbf{u}^T \mathbf{C} \mathbf{v}\), and for matrices \(\mathbf{U}\) and \(\mathbf{V}\), \((\mathbf{U}, \mathbf{V})_{\mathbf{C}} := \text{Tr}(\mathbf{U}^T \mathbf{C} \mathbf{V})\) where \(\text{Tr}(\cdot)\) denotes the trace operator. The \(s \times s\) identity matrix is denoted \(\mathbf{I}_s\). The \(s \times s\) zero matrix is denoted \(\mathbf{0}_s\). We denote by \(\mathcal{S}_{\text{sym}}(p)\) and \(\mathcal{S}_{\text{skew}}(p)\) the set of all symmetric and skew-symmetric matrices (respectively) in \(\mathbb{R}^{p \times p}\).

Given a \(d \times d\) matrix \(\mathbf{A}\) we denote by \(\text{sym}(\mathbf{A}) := (\mathbf{A} + \mathbf{A}^T)/2\) and by \(\text{skew}(\mathbf{A}) := (\mathbf{A} - \mathbf{A}^T)/2\) the symmetric and skew-symmetric (respectively) components of \(\mathbf{A}\). We describe a diagonal matrix using \(\text{diag}(\cdot)\) where the diagonal components appear in the parenthesis, and similarly block diagonal matrices are described using \(\text{blkdiag}(\cdot)\). For a SPD matrix \(\mathbf{B} \in \mathbb{R}^{d \times d}\), we denote by \(\mathbf{B}^{1/2}\) the unique SPD matrix such that \(\mathbf{B} = \mathbf{B}^{1/2} \mathbf{B}^{1/2}\). This matrix is obtained by keeping the same eigenvectors and taking the square root of the eigenvalues. We denote the inverse of \(\mathbf{B}^{1/2}\) by \(\mathbf{B}^{-1/2}\).

Let \(\mathbf{A}\) be a symmetric \(d \times d\) matrix. We use \(\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots \geq \lambda_d(\mathbf{A})\) to denote the eigenvalues of \(\mathbf{A}\), and use \(\kappa(\mathbf{A})\) to denote the condition number of \(\mathbf{A}\), which is the ratio between the largest and smallest eigenvalues in absolute value. Let \(\mathbf{B} \in \mathbb{R}^{d \times d}\) be another symmetric positive semidefinite matrix, and assume that \(\ker(\mathbf{B}) \subseteq \ker(\mathbf{A})\). If for \(\lambda \in \mathbb{R}\) and \(\mathbf{v} \notin \ker(\mathbf{B})\) it holds that \(\mathbf{A} \mathbf{v} = \lambda \mathbf{B} \mathbf{v}\) then \(\lambda\) is a generalized eigenvalue and \(\mathbf{v}\) is a generalized eigenvector of the matrix pencil \((\mathbf{A}, \mathbf{B})\). We use the notation \(\lambda_1(\mathbf{A}, \mathbf{B}) \geq \lambda_2(\mathbf{A}, \mathbf{B}) \geq \cdots \geq \lambda_{\text{rank}(\mathbf{B})}(\mathbf{A}, \mathbf{B})\) to denote the generalized eigenvalues of \((\mathbf{A}, \mathbf{B})\). The (generalized) condition number \(\kappa(\mathbf{A}, \mathbf{B})\) of the pencil \((\mathbf{A}, \mathbf{B})\) is the ratio between the largest and smallest generalized eigenvalues in absolute value. If \(\mathbf{B}\) is also non-singular, that is \(\mathbf{B}\) is a SPD matrix, then it holds that \(\kappa(\mathbf{A}, \mathbf{B}) = \kappa(\mathbf{B}^{-1/2} \mathbf{A} \mathbf{B}^{-1/2})\).

We denote by \(\text{St}_{\mathbf{B}}(p, d)\) the generalized Steifel manifold defined by Eq. (1.1). \(\text{St}_{\mathbf{B}}(p, d)\) is a submanifold of \(\mathbb{R}^{d \times p}\). Given a function or vector field defined on \(\text{St}_{\mathbf{B}}(p, d)\), we use a bar decorator to denote a smooth extension of that object to the entire \(\mathbb{R}^{d \times p}\), either by committing to a specific extension, or making sure that any statement made afterwards holds for any such smooth extension. For example, given a smooth objective function \(f : \text{St}_{\mathbf{B}}(p, d) \rightarrow \mathbb{R}\), we use \(\bar{f} : \mathbb{R}^{d \times p} \rightarrow \mathbb{R}\) to denote a smooth real-valued function defined on \(\mathbb{R}^{d \times p}\) whose restriction to \(\text{St}_{\mathbf{B}}(p, d)\) is \(f\).

For \(p = 1\), we denote by \(S^\mathbf{B}\) the \(d - 1\) dimensional ellipsoid defined by

\[ S^\mathbf{B} := \{ \mathbf{x} \in \mathbb{R}^d : \mathbf{x}^T \mathbf{B} \mathbf{x} = 1 \} \, . \]

In the special case \(\mathbf{B} = \mathbf{I}_d\), we denote by \(\text{St}(p, n)\) the Steifel manifold defined by

\[ \text{St}(p, d) := \{ \mathbf{X} \in \mathbb{R}^{d \times p} : \mathbf{X}^T \mathbf{X} = \mathbf{I}_p \} \, . \]
Given a SPD matrix $B \in \mathbb{R}^{d \times d}$, we say that a decomposition $A = Q R$ of $A \in \mathbb{R}^{d \times p}$ where $Q \in \mathbb{R}^{d \times p}$ and $R \in \mathbb{R}^{p \times p}$ is a thin $B$-QR decomposition of $A$ if $Q \in \text{St}_B(p, d)$ and $R \in \mathbb{R}^{p \times p}$ is an upper triangular matrix. Note that the standard thin QR decomposition ([23, 16] Chapter 5 and Lecture 7]) is a thin $I_d$-QR decomposition. Moreover, the thin $B$-QR decomposition can be obtained using a standard thin QR decomposition of the matrix $B^{1/2} A$. Indeed, if $B^{1/2} A = QR$ with $Q \in \text{St}(p, d)$ then $A = (B^{-1/2} Q) R$ and $B^{-1/2} Q \in \text{St}_B(p, d)$. The thin QR decomposition is unique if $A$ is full rank and we require $R$ to have strictly positive diagonal elements ([23, 16] Theorem 5.2.2 and Theorem 7.2]). Consequently, we also have that the thin $B$-QR decomposition is unique if $A$ is full rank and we require $R$ to have strictly positive diagonal elements. In that case, we denote by $qf_B(A)$ the unique $Q$ factor of the thin $B$-QR decomposition. For the thin $I_d$-QR decomposition we abbreviate $qf(A) := qf_{I_d}(A)$. Using this notation we have the following relation [10]:

$$qf_B(A) = B^{-1/2} qf\left( B^{1/2} A \right).$$

### 2.2 Riemannian Optimization

In this section we recall some basic definitions of Riemannian optimization, and establish corresponding notations. A Riemannian manifold $\mathcal{M}$ is a real differentiable manifold $\mathcal{M}$ with a smoothly varying inner product $g_x$ on tangent spaces $T_x \mathcal{M}$ (where $x \in \mathcal{M}$). A Riemannian manifold $(\mathcal{M}, g)$ is a Riemannian submanifold of another Riemannian manifold $(\mathcal{M}, g_\mathcal{M})$, if $\mathcal{M}$ is a submanifold of $\mathcal{M}$ and it inherits the metric in a natural way: $g_x(\eta_x, \xi_x) = g_{\mathcal{M}}(\eta_x, \xi_x)$ for $\eta_x, \xi_x \in T_x \mathcal{M}$ where in the right-side $\eta_x$ and $\xi_x$ are viewed as elements in $T_x \mathcal{M}$ (this is possible since $\mathcal{M}$ is a submanifold of $\mathcal{M}$). The former notion is useful when the search space is embedded in a larger space and the objective function is given in the coordinates of the embedding space.

The fundamental idea in Riemannian optimization algorithms is to locally approximate the constraint manifold by its tangent space at every iteration. Each iterate on the tangent space minimizes some model embedding space.

The notions of Riemannian gradient and Riemannian Hessian [1] Section 3.6 and 5.5 generalize the corresponding concepts from the Euclidean setting. The Riemannian gradient is used for finding critical points, while the Riemannian Hessian classifies them. Moreover, (asymptotic) convergence of Riemannian methods is governed by the condition number of the Riemannian Hessian at the optimal point.

For a smooth (objective) function defined on the manifold, $f : \mathcal{M} \to \mathbb{R}$, denote the Riemannian gradient and Riemannian Hessian at $x \in \mathcal{M}$ by $\nabla f(x) \in T_x \mathcal{M}$ and $\nabla^2 f(x) : T_x \mathcal{M} \to T_x \mathcal{M}$ respectively. Roughly speaking, the Levi-Civita (Riemannian connection) $\nabla$ of $(\mathcal{M}, g)$ generalizes the notion of directional derivative of vector fields.

With these components, various optimization algorithms are naturally generalized from the Euclidean setting to the Riemannian setting (e.g., [1] for an extensive overview of smooth techniques, and [6] for some examples of non-smooth algorithms). For example, a variant of Riemannian gradient descent is

$$x_{k+1} = R_x(\tau_k \nabla f(x_k)) \quad (2.1)$$

where $\tau_k$ is the step size (possibly chosen by the Armijo’s backtracking procedure; see [1] Algorithm 1]).
3 Preconditioned Geometric Components for the Generalized Stiefel Manifold

In this section we describe the necessary geometric components required for Riemannian optimization on $\text{St}_B(p,d)$ with a preconditioned Riemannian metric. In the following, $B \in \mathbb{R}^{d \times d}$ is a SPD matrix and we treat $\text{St}_B(p,d)$ as an embedded submanifold of $\mathbb{R}^{d \times p}$. Unlike previous articles in the literature, we allow for a wider array of Riemannian metrics on $\text{St}_B(p,d)$, i.e., the metric is defined via a preconditioning scheme $X \mapsto M_X$. We refer to the components we develop as **preconditioned geometric components** for $\text{St}_B(p,d)$. In the following, we refer to $\mathbb{R}^{d \times p}$ as the ambient space. It is important to stress that all our formulas are given in ambient space coordinates, and not in some in some local coordinates of the manifold $\text{St}_B(p,d)$.

In terms of computational costs of the geometric components, we remark that an important feature of the components we develop is that they access $B$ only via matrix-matrix products. In particular, the formulas do not involve $B^{-1}$ but rather $M_X^{-1}$. In quite a few problems involving generalized orthogonality constraints, the matrix $B$ is given in a (semi-)implicit form, and it is desirable to avoid computing it. In many applications, just forming $B$ is as expensive as using a direct method. However, to use the preconditioned geometric components one can avoid computing $B$.

3.1 Metric Independent Notions

We first describe notions that are independent of the metric. This is not our main contribution, as most of the following definitions and formulas are well known (see e.g., [49, 40, 51, 30, 34]); we include these definitions and formulas, and their derivations (which appears in the appendix), for completeness. Additionally, most of the formulas in this section can be derived via the known components of the Stiefel manifold [1] via the change of variables $\tilde{X} = B^{1/2}X$.

We remark that the formulas for the inverses of various retractions do not appear in the previous literature. However, for the most part they too are simple generalizations of formulas for the Stiefel manifold, which are derived in [19, 1, 51, 34]. The inverse retraction is used in several recent algorithms proposed in the literature: Riemannian CG with inverse retractions [52], Riemannian FISTA [29], and empirical arithmetic averaging over the Stiefel manifold [30].

The tangent space of $\text{St}_B(p,d)$ at $X \in \text{St}_B(p,d)$, viewed as a subspace of $T_X \mathbb{R}^{d \times p} \simeq \mathbb{R}^{d \times p}$, is

$$T_X \text{St}_B(p,d) = \{ Z \in \mathbb{R}^{d \times p} : Z^T B X + X^T B Z = 0_{p} \}.$$  \hspace{1cm} (3.1)

To explain Eq. (3.1), note that $\text{St}_B(p,d)$ is the kernel of $F(X) = X^T B X - I_p$ which is a submersion. $F$ is a symmetric matrix valued function, so the dimension of the tangent space (and, as such, the manifold itself) is $dp - p(p+1)/2$.

Obviously, if $Z \in T_X \text{St}_B(p,d)$ then the matrix $X^T B Z$ is skew-symmetric. Thus, a different characterization of $T_X \text{St}_B(p,d)$ is as a decomposition of every tangent vector into a sum of a product of a skew-symmetric matrix with $X$, and term whose columns are $B$-orthogonal to the columns of $X$:

$$T_X \text{St}_B(p,d) = \left\{ Z = X \Omega + X_{B^\perp} K \in \mathbb{R}^{d \times p} : \Omega \in \mathcal{S}_{\text{skew}}(p), \ K \in \mathbb{R}^{(d-p) \times p} \right\},$$  \hspace{1cm} (3.2)

where $\Omega$ is a skew-symmetric matrix (i.e., $\Omega^T = -\Omega$), $K$ is arbitrary, and $X_{B^\perp} \in \mathbb{R}^{d \times (d-p)}$ satisfies that its columns are orthonormal for the orthogonal complement of the column space of $X$ with respect to the matrix $B$, i.e., $X_{B^\perp}^T B X_{B^\perp} = I_{d-p}$, and $X_{B^\perp}^T B X = 0_{(d-p) \times p}$.

There are several known retraction mappings suitable for the generalized Stiefel manifold. We mention three of them (not including the exponential map which is presented later and is also a retraction). The first retraction mapping is based on the polar decomposition of the matrix $X + \xi_X$ with respect to the inner
product defined by the matrix $B$ (i.e., decomposition of a matrix $A = QP$ where $Q \in \text{St}_B(p,d)$ and $P$ is a SPD matrix of the size $p \times p$; one such decomposition is $P = (A^TBA)^{1/2}$ and $Q = A (A^TBA)^{-1/2}$):

$$R^\text{polar}_X(\xi_X) := (X + \xi_X)(I_p + \xi X B\xi X)^{-1/2},$$

(3.3)

where $\xi_X \in T_X\text{St}_B(p,d)$. As for the arithmetic complexity, once $B\xi_X$ has been computed, we can compute $R^\text{polar}_X(\xi_X)$ in $O(dp^2)$ operations.

Given $Y \in \text{St}_B(p,d)$ close enough to $X$, the inverse of the polar retraction is

$$R^\text{polar}^{-1}_X(Y) := YZ - X,$$

(3.4)

where $Z$ is the unique SPD solution of the following Lyapunov equation

$$2I_p = X^TBYZ + ZY^TBX.$$

(3.5)

Thus, once $BX$ is computed we can compute $R^\text{polar}^{-1}_X(Y)$ using $O(dp^2)$ operations. The expression for the inverse retraction in Eq. (3.4) is valid when the Lyapunov Eq. (3.5) has a unique SPD solution. If $Y = R^\text{polar}_X(\xi_X)$ for some $\xi_X$, then Eq. (3.4) has a SPD solution $Z = (I_p + \xi X B\xi X)^{-1/2}$ (see Appendix A.1 for more details). Let us now consider when Eq. (3.4) has a unique solution. It has a unique solution if and only if $X^TBY$ and $-Y^T BX$ do not share any eigenvalue [27, Theorem 2.4.4.1]. Both $X^TBY$ and $-Y^T BX$ are invertible since they are products of full rank matrices, thus all eigenvalues are not equal to zero. Next recall that $X^T BX = I_p$, and that eigenvalues of a matrix are a continuous function of the matrix. Using the Bauer–Fike theorem [3] for a small enough perturbation of the matrix $X^T BX$, i.e., $X^T BX + \delta X^T BX = X^T BY$, the eigenvalues of $X^T BY$ do not differ from the eigenvalues of $X^T BX$ more than the norm of the perturbation. Thus, the real part of the eigenvalues of $X^T BY$ remains strictly positive, leading to $X^T BY$ and $-Y^T BX$ not sharing any eigenvalue. The validity of Eq. (3.4) is the intersection of the image of $R^\text{polar}_X(\cdot)$ and a neighborhood of $X$ in which Eq. (3.5) has a unique solution.

The second retraction mapping is based on the QR decomposition with respect to the matrix $B$:

$$R^\text{QR}_X(\xi_X) := qf_B(X + \xi X) = B^{-1/2}qf \left( B^{1/2}(X + \xi X) \right),$$

(3.6)

where $\xi_X \in T_X\text{St}_B(p,d)$. One can show that if $R^T R = (X + \xi X)^T B(X + \xi X)$ is a Cholesky decomposition then $qf_B(X + \xi X) = (X + \xi X)R^{-1}$, so once $B(X + \xi X)$ has been computed we can compute $R^\text{QR}_X(\xi_X)$ using $O(dp^2)$ operations.

Given $Y \in \text{St}_B(p,d)$ close enough to $X$ the inverse of the QR-based retraction is

$$R^\text{QR}^{-1}_X(Y) := YR - X,$$

(3.7)

where $R$ is the unique upper-triangular $p \times p$ matrix with strictly positive elements on its main diagonal which is a solution for the following Lyapunov-like equation:

$$2I_p = X^TBYR + R^T Y^T BX.$$

(3.8)

Solving this equation takes $O(p^4)$ operations. Thus, once $BX$ is computed we can compute $R^\text{QR}^{-1}_X(Y)$ using $O(p^4 + dp^2)$ operations. Note that this equation has a solution for $Y = X$, which is $R = I$. Then, by continuity arguments, if $Y$ is close enough to $X$, a solution exists. To show uniqueness of solution, we use [30, Eq. (14) and Algorithm 1]. According to Kaneko et al., using the constraint that $R$ is upper-triangular we can reformulate Eq. (3.8) as an equivalent set of linear equations which has a unique solution if and only if all the principal minors of $X^T BY$ are non-singular (see Appendix A.1). Similarly to the argument for inverse of the polar inverse retraction, since $X^T BX = I_p$, for a small enough perturbation, i.e.,
\( X^T B X + \delta X^T B X = X^T B Y \), the real part of the eigenvalues of \( X^T B Y \) remains strictly positive, leading to non-singularity of \( X^T B Y \). Note that in order to have consistency, it is also required that the diagonal elements of \( R \) are strictly positive. Again, using a similar continuity arguments we can achieve such a solution if \( Y \) is close enough to \( X \). The validity of Eq. (3.7) is the intersection of the image of \( R^Q X (\cdot) \) and a neighborhood of \( X \) in which Eq. (3.8) has a unique solution.

The third retraction mapping is based on the Cayley transform with respect to the matrix \( B \) (a generalization of the retraction presented in [30]):

\[
R^Cayley_X(\xi_X) := \left( I_d - \frac{1}{2} W(\xi_X) \right)^{-1} \left( I_d + \frac{1}{2} W(\xi_X) \right) X,
\]  

where

\[
W(\xi_X) := (I_d - \frac{1}{2} XX^T B)\xi_X X^T B - X\xi_X^T (I_d - \frac{1}{2} BXX^T) B.
\]

Once the multiplications with \( B \) are computed, we need to compute the inverse of a \( d \times d \) matrix in order to find \( R^Cayley_X(\xi_X) \). However, noticing that

\[
W(\xi_X) = \begin{bmatrix} (I_d - \frac{1}{2} XX^T B)\xi_X & X \\ \xi_X^T (I_d - \frac{1}{2} BXX^T) B & I_d \end{bmatrix}
\]

(i.e., a product of a \( d \times 2p \) matrix by a \( 2p \times d \) matrix), we can use the Sherman-Morrison-Woodbury formula to only invert a \( 2p \times 2p \) matrix. A closed form for the inverse of this retraction is only known when \( d \) is even [30].

Similarly, there are several possible ways to compute a vector transport. It is possible to define a metric independent vector transport, using [1, Equation 8.6] by differentiating a retraction mapping

\[
\tau_{\eta_X}(ind)^{\eta_X}(\xi_X) := D R_X(\eta_X)[\xi_X].
\]

In Appendix A.2 we derive concrete formulas based on the polar and QR retractions (Eq. 3.3 and 3.6). A vector transport based on the Cayley retraction is presented in [31]. The various vector transport have the same computational cost as computing the corresponding retraction. Note that it is also possible to define another vector transport that has this property by simply applying the projection on the tangent space. However, this vector transport is metric dependent, so we discuss it in the next subsection.

### 3.2 Metric Related Notions

This subsection is the main contribution of our paper. In this subsection we derive explicit formulas for the orthogonal projection with respect to the Riemannian metric, the Riemannian gradient and Hessian with respect to the non-standard metric which allow the use of various preconditioned Riemannian algorithms. Note that the formulas in this subsection, unlike the previous one, cannot be derived via a change of variables \( X = B^{\frac{1}{2}} X \) unless a specific metric is used (corresponding to \( M_X = B \) for all \( X \in St_B(p,d) \)), since though this change of variables makes \( X \in St(p,d) \), the induced metric on that manifold is not the standard metric.

Specifically, we define a Riemannian metric on the ambient space \( \mathbb{R}^{d \times p} \), and this uniquely defines a metric on \( St_B(p,d) \) that makes it a Riemannian submanifold. The metric we define on \( \mathbb{R}^{d \times p} \) is

\[
g_X(\xi_X, \eta_X) := (\xi_X, \eta_X)_{M_X} = \text{Tr} \left( \xi_X^T M_X \eta_X \right)
\]

where \( X \mapsto M_X \) is a smooth mapping on \( \mathbb{R}^{d \times p} \) (thus, the metric varies smoothly with \( X \) making it a Riemannian metric), and each \( M_X \) is assumed to be a SPD matrix so that we have a properly defined inner product on each tangent space, and a Riemannian metric for \( \mathbb{R}^{d \times p} \). Now, for any \( X \in St_B(p,d) \), \( \xi_X, \eta_X \in T_X St_B(p,d) \), given in ambient space coordinates, the Riemannian metric on \( St_B(p,d) \) is given by

\[
g_X(\xi_X, \eta_X) := (\xi_X, \eta_X)_{M_X} = \text{Tr} \left( \xi_X^T M_X \eta_X \right).
\]  

(3.10)
The cost of computing \( g_X(\xi_X, \eta_X) \) is \( O(T_M p + dp) \) where \( T_M \) is the maximal cost (possibly after preprocessing) of taking the product with \( M_X \) with a vector for all \( X \).

The metric selection is how we propose to incorporate a preconditioner, and so the mapping \( X \mapsto M_X \) is termed a preconditioning scheme. It should be chosen so that the Riemannian Hessian at the optimum is well conditioned. We discuss this further in Subsection 3.3. Classically, the metric employed for the generalized Stiefel manifold corresponds to \( M_X = B \) for all \( X \in St_B(p, d) \). In quite a few applications this choice minimizes a-priori bounds on the condition number of the Riemannian Hessian at the optimum (see Subsections 3.5 and 12). However, as we shall see, various operations required for Riemannian optimization require products with \( M_X^{-1} \), and in many applications this results in algorithms that are too expensive when \( M_X = B \) for some \( X \in St_B(p, d) \). In such cases, there is a need to balance in the chosen \( X \mapsto M_X \) between minimizing the condition number, and efficient products with \( M_X^{-1} \). This is a typical trade-off for preconditioning.

After defining the Riemannian metric we can derive the metric related notions required for Riemannian optimization. Since \( St_B(p, d) \) is an embedded submanifold of \( \mathbb{R}^{d \times p} \), the orthogonal projection on the tangent space with respect to the Riemannian metric is a key component. We denote the orthogonal projection operator on \( T_X St_B(p, d) \) by \( \Pi_X(\cdot) \), and the orthogonal projection operator (with respect to the metric defined by \( X \mapsto M_X \)) on the normal space, \((T_X St_B(p, d))^\perp \), by \( \Pi_X^N(\cdot) \).

In order to find analytic formulas for these operators, we first note that the normal space is:

\[
(T_X St_B(p, d))^\perp = \{ M_X^{-1} B X : S \in S_{sym}(p) \} \tag{3.11}
\]

since \( \text{Tr}(S^T \Omega) = 0 \) for any symmetric matrix \( S \) and anti-symmetric matrix \( \Omega \) (see Appendix A.2). The dimension of the normal space is \( p(p + 1)/2 \). The following lemma gives a formula for the orthogonal projections to the tangent and normal spaces.

**Lemma 1.** The orthogonal projections with respect to \( g_X(\cdot, \cdot) \) on \((T_X St_B(p, d))^\perp \) and on \( T_X St_B(p, d) \) (viewed as a subspace of \( T_X \mathbb{R}^{d \times p} \approx \mathbb{R}^{d \times p} \) and given in ambient coordinates) are:

\[
\Pi_X^T(\xi_X) = M_X^{-1} B X \xi_X \tag{3.12}
\]

and

\[
\Pi_X(\xi_X) = \left( \text{id}_{T_X St_B(p, d)} - \Pi_X^N \right) (\xi_X) = \xi_X - M_X^{-1} B X \xi_X \tag{3.13}
\]

where \( \xi_X \in T_X \mathbb{R}^{d \times p} \), \( \text{id}_{T_X St_B(p, d)} \) denotes the identity mapping on \( T_X St_B(p, d) \), and \( S_{\xi_X} \in \mathbb{R}^{p \times p} \) is the unique solution of the following Sylvester equation:

\[
(X^T B M_X^{-1} B X) S_{\xi_X} + S_{\xi_X} (X^T B M_X^{-1} B X) = X^T B \xi_X + (X^T B \xi_X)^T.
\]

The cost of computing (in ambient coordinates) \( \Pi_X(\xi_X) \) for an arbitrary \( \xi_X \) is \( O(T_B p + T_{M^{-1}} p + dp^2) \), where \( T_B \) and \( T_{M^{-1}} \) are the cost of computing the product of \( B \) with a vector and the maximal cost of taking the product with \( M_X^{-1} \) with a vector for all \( X \in St_B(p, d) \).

**Proof.** Note that \( T_X St_B(p, d) \oplus (T_X St_B(p, d))^\perp = T_X \mathbb{R}^{d \times p} \approx \mathbb{R}^{d \times p} \). This implies that for any \( \xi_X \in T_X \mathbb{R}^{d \times p} \approx \mathbb{R}^{d \times p} \) there exists unique \( \Omega_{\xi_X} \in S_{skew}(p) \), \( K_{\xi_X} \in \mathbb{R}^{(d - p) \times p} \) and \( S_{\xi_X} \in S_{sym}(p) \) such that \( \xi_X \) is decomposed to a unique component on the tangent space of \( St_B(p, d) \) and a unique component on the normal space of \( St_B(p, d) \):

\[
\xi_X = \Pi_X(\xi_X) + \Pi_X^N(\xi_X) = \left( X \Omega_{\xi_X} + X_B K_{\xi_X} \right) + M_X^{-1} B X S_{\xi_X} \tag{3.14}
\]

By left-multiplying Eq. (3.14) by \( X^T B \), we get

\[
X^T B \xi_X = \Omega_{\xi_X} + X^T B M_X^{-1} B X S_{\xi_X}.
\]
Summing $X^TB\xi_X + (X^TB\xi_X)^T$, and using the fact that $\Omega_{\xi_X}$ is skew-symmetric so it cancels in the sum, we get that $S_{\xi_X}$ solves the following Sylvester equation (27 Subsection 2.4.4.1):

$$X^TB\xi_X + (X^TB\xi_X)^T = (X^TB\text{M}_X^{-1}BX)S_{\xi_X} + S_{\xi_X}(X^TB\text{M}_X^{-1}BX).$$ (3.15)

Indeed, according to [27] Theorem 2.4.4.1] there is a unique solution to Eq. (3.15) for any $X^TB\xi_X + (X^TB\xi_X)^T$, since $(X^TB\text{M}_X^{-1}BX)$ is positive definite $(X^TB\text{M}_X^{-1}BX)$ is a Gram matrix of $\text{M}_X^{-1/2}BX$, which consists of a product of three matrices, two invertible matrices $\text{M}_X^{-1/2}$ and $B$, and one full-column rank matrix $X \in \text{St}_B(p,d)$ and $-(X^TB\text{M}_X^{-1}BX)$ is negative definite, thus both matrices have no eigenvalues in common. Solving Eq. (3.15) costs $O(p^3)$ assuming we already computed $X^TB\text{M}_X^{-1}BX$. Furthermore, as expected $S_{\xi_X}$ is symmetric since $S_{\xi_X}$ again satisfies Eq. (3.15), and the solution to the equation is unique.

After obtaining $S_{\xi_X}$ by solving Eq. (3.15), analytical expressions for the orthogonal projections on the normal space and the tangent space are given by Eq. (3.12) and Eq. (3.13).

Note that the orthogonal projection on the normal space and the tangent space satisfy the definition of an orthogonal projection with respect to the inner product defined on $\mathbb{R}^{n \times p}$ with the matrix $\text{M}_X$. Indeed, both projections satisfy the projection property $\Pi_X^2(\cdot) = \Pi_X(\cdot)$ and $\left(\Pi_X^\perp\right)^2(\cdot) = \Pi_X^\perp(\cdot)$, since $S_{\text{st}_X(\xi_X)}$ and $S_{\xi_X}$ satisfy the same Sylvester equation. In addition, both projections are orthogonal with respect to the inner product defined on $\mathbb{R}^{n \times p}$ with the matrix $\text{M}_X$, i.e.,

$$g_X(\Pi_X(\xi_X),\eta_X) = g_X(\xi_X,\Pi_X(\eta_X)), g_X(\Pi_X^\perp(\xi_X),\eta_X) = g_X(\xi_X,\Pi_X^\perp(\xi_X))$$ (3.16)

for all $\xi_X, \eta_X \in \mathbb{R}^{n \times p}$, since by using the properties of the trace operator.

The cost of computing (in ambient coordinates) $\Pi_X(\xi_X)$ for an arbitrary $\xi_X$ is $O(\text{T}_Bp + \text{T}_M^{-1}p + dp^2)$. Indeed, after obtaining $S_{\xi_X}$ by solving a Sylvester equation which costs $O(p^3)$, we are left with taking product of $B$ and $\text{M}_X^{-1}$ with matrices, and products of matrices of the dimensions $p \times d$ by $d \times p$, $d \times p$ by $p \times p$ and $p \times p$ by $p \times p$.

In the special case where $\text{M}_X = B$ for all $X \in \text{St}_B(p,d)$, the orthogonal projections on the normal space Eq. (3.12) and on the tangent space Eq. (3.13), are reduced to a generalization of the orthogonal projection on the tangent space of the Stiefel manifold [1] Example 3.6.2):

$$\Pi_X^\perp(\xi_X) = X_{\text{sym}}(X^TB\xi_X)$$ (3.17)

and

$$\Pi_X(\xi_X) = (\text{id}_{XX^T}\text{St}_B(p,d) - \Pi_X^\perp)(\xi_X) = (\text{id} - XX^TB)\xi_X + X_{\text{skew}}(X^TB\xi_X).$$ (3.18)

In such case, the cost of computing (in ambient coordinates) $\Pi_X(\xi_X)$ for an arbitrary $\xi_X$ is $O(\text{T}_Bp + dp^2)$. The cost is evident from the formulas once we observe that none of the operations require forming $B$, but instead require taking product of $B$ with a matrix of $p$ columns.

Using the orthogonal projection we can also propose a simple metric dependent vector transport using the vector transport definition on Riemannian submanifolds [1] Subsection 8.1.3):

$$\tau_{X(\xi_X)}^{(\text{dep})} := \Pi_{R_X(\xi_X)}(\xi_X),$$ (3.19)

where $R_X(\cdot)$ is a retraction mapping of our choice (e.g., Eq. (3.3), Eq. (3.9) or Eq. (3.9)).

Let $f : \text{St}_B(p,d) \to \mathbb{R}$ be smooth function, and let $\bar{f}$ be a smooth extension of $f$ to $\mathbb{R}^{d \times p}$ (typically, $f$ is given in ambient coordinates, thereby making the extension $\bar{f}$ natural). We now develop first and second order Riemannian components for $f$. The Riemannian gradient is an element of the tangent space, and to derive an analytic formula for it we use [1] Eq. 3.37]: the Riemannian gradient can be computed by computing
the Riemannian gradient in $\mathbb{R}^{d \times p}$ of $\tilde{f}$, and orthogonally projecting it with respect to the Riemannian metric to the tangent space of $\text{St}_B(p, d)$ using the orthogonal projection on the tangent space, $\Pi_X(\cdot)$. In short, $\text{grad} f(X) = \Pi_X(\text{grad} f(X))$. First, we consider $\text{grad} f(X)$. Note that it is not the Euclidean gradient $\nabla f(X)$, even though $f$ is defined on $\mathbb{R}^{d \times p}$. The reason is that $\tilde{f}$ is defined on a $\mathbb{R}^{d \times p}$ endowed with a non-standard inner product. According to [1, Eq. 3.31], we have

$$\text{Tr} (\text{grad} \tilde{f}(X)^T M_X \xi_X) = g_X(\text{grad} \tilde{f}(X), \xi_X) = D\tilde{f}(X)[\xi_X] = \text{Tr} (\nabla \tilde{f}(X)^T \xi_X)$$

for every $\xi_X \in T_X \mathbb{R}^{d \times p}$ (in the above, $Df(X)$ denotes the (Frechet) differential of $f$ at $X$), so $\text{grad} \tilde{f}(X) = M_X^{-1} \nabla \tilde{f}(X)$. Thus, we have

$$\text{grad} f(X) = \Pi_X (M_X^{-1} \nabla \tilde{f}(X)). \quad (3.20)$$

The cost of computing the Riemannian gradient given the Euclidean gradient of $\tilde{f}$ is the cost of computing the orthogonal projection on the tangent space, and products with $M_X^{-1}$.

The components developed so far, allow the application of any first order Riemannian optimization algorithm, e.g., Riemannian gradient and Riemannian conjugate-gradient. In order to apply second-order methods, e.g., Riemannian Newton and Riemannian trust-region, the Riemannian Hessian must also be derived. An expression for the Riemannian Hessian is also useful for reasoning on the convergence rate by deriving an algorithm, e.g., Riemannian gradient and Riemannian conjugate-gradient. In order to apply second-order algorithms, e.g., Riemannian gradient and Riemannian conjugate-gradient, the Riemannian Hessian must depend on the specifics of the mapping of $X$ to $M_X$. Thus, we focus on the simpler case where $M_X = M$, i.e. $M_X$ is constant for all $X \in \text{St}_B(p, d)$. This is a reasonable choice for a preconditioning metric since it still allows the use of different cheap-to-invert constant approximations of $B$ (see Subsection 4.2 for an example).

We note that due to [1, Proposition 5.5.6], at a critical point $X^*$, i.e. $\text{grad} f(X^*) = 0$, the Riemannian Hessian equals to the Euclidean Hessian of a composition of the cost function with a retraction map (retraction map typically does not depend on the choice of the Riemannian metric). Thus, the formula we derive for the Riemannian Hessian in ambient coordinates is valid at a critical point $X^*$ when using any preconditioning scheme $X \mapsto M_X$ as well if we set $M = M_X^{\ast \ast}$. This property allows the analysis of the condition number of the Riemannian Hessian at the critical points with a preconditioning scheme $X \mapsto M_X$, giving indication for the asymptotic convergence Riemannian optimization (e.g., [1] Theorem 4.5.6, Theorem 7.4.11, Equation 7.5)).

The Riemannian Hessian of $f$ at a point on the manifold is a linear transformation from the tangent space to itself. When $M_X = M$ for all $X \in \text{St}_B(p, d)$, we can compute the result of applying the Riemannian Hessian to a tangent vector in ambient coordinates via the formula [2]:

$$\text{Hess} f(X)[\eta_X] = \Pi_X (M^{-1} \nabla^2 \tilde{f}(X) \eta_X) + W_X (\eta_X, \Pi_X^{-1} (M^{-1} \nabla \tilde{f}(X))) \quad (3.21)$$

where $\nabla^2 \tilde{f}(X)$ is the Euclidean Hessian of $\tilde{f}$ and $W_X$ is the Weingarten map on $\text{St}_B(p, d)$. The Weingarten map is an operator that takes as arguments a tangent vector $\eta_X \in T_X \text{St}_B(p, d)$ and a normal vector $U_X \in (T_X \text{St}_B(p, d))^\perp$ and returns a tangent vector. An analytic formula for Weingarten map on $\text{St}_B(p, d)$, in ambient coordinates, is

$$W_X (\eta_X, U_X) = -\Pi_X (M^{-1} B \eta_X (X^T M U_X)).$$

The derivation of Eq. (3.21) is based on the Riemannian connection on $\text{St}_B(p, d)$, which is the classical directional derivative on $\mathbb{R}^{d \times p}$ projected on the tangent space (when $M_X = M$ for all $X$). The complete derivation of the Riemannian connection, the Weingarten map, and the Riemannian Hessian appears in Appendix A.2. Based on these formula, we have the following formula for the Riemannian Hessian when $M_X = M$ for all $X$:

$$\text{Hess} f(X)[\eta_X] = \Pi_X (M^{-1} \nabla^2 \tilde{f}(X) \eta_X - M^{-1} B \eta_X (X^T \nabla \tilde{f}(X) - X^T M \text{grad} f(X))). \quad (3.22)$$

The cost of applying the Riemannian Hessian to a tangent vector given the Euclidean Hessian of $\tilde{f}$ is the cost of computing the orthogonal projection on the tangent space, and products with $B$, $M$ and $M^{-1}$. 

11
Exponential Map. An important metric related retraction map on a Riemannian manifold is the exponential mapping. According to [1, Proposition 5.4.1], the exponential map induced by the Riemannian connection defined on the manifold is a retraction map, termed the exponential retraction. In particular, the exponential map is based on moving on geodesic curves in the direction of a tangent vector. In the derivation of the exponential map we assume $\mathbf{M}_\mathbf{X} = \mathbf{M}$ for all $\mathbf{X} \in \mathbf{St}_B(p,d)$.

First, let us recall the definition of a geodesic curve. A geodesic $\gamma(t)$ on a manifold $\mathcal{M}$ endowed with an affine connection $\nabla$ is a curve with zero acceleration

$$\frac{D^2}{dt^2} \gamma(t) = 0,$$

for all $t$ in the domain of $\gamma(t)$, where $\frac{D^2}{dt^2} \gamma(t) = \frac{D}{dt} \gamma$ [1, Section 5.4].

On the generalized Stiefel manifold, the function $\xi_x \mapsto \frac{D}{dt} \xi_x$ from the set of all (smooth) vector fields on $\mathbf{St}_B(p,d)$ to itself is $\frac{D}{dt} (\cdot) := \Pi_x (t) \left( \frac{d}{dt} (\cdot) \right)$. For every $\xi_x \in T_x \mathcal{M}$, there exists an interval $I$ about 0 and a unique geodesic $\gamma(t; x, \xi): I \rightarrow \mathcal{M}$ such that $\gamma(0) = x$ and $\dot{\gamma}(0) = \xi_x$. Moreover, we have the homogeneity property $\gamma(t; x, a\xi_x) = \gamma(at; x, \xi_x)$. The mapping

$$\text{Exp}_x : T_x \mathcal{M} \rightarrow \mathcal{M} : \xi_x \mapsto \text{Exp}_x \xi_x = \gamma(1; x, \xi_x),$$

is called the exponential map at $x$ [1, Section 5.4].

To find the exponential map on the Stiefel manifold $\mathbf{St}_B(p,d)$, we need to find the geodesic given $\mathbf{X} = \gamma(0) \in \mathbf{St}_B(p,d)$ and $\dot{\xi}_x = \dot{\gamma}(0) \in T_x \mathbf{St}_B(p,d)$, i.e., we need to solve the differential equation

$$\frac{D^2}{dt^2} \gamma(t) = 0$$

$$\Pi_{\gamma(t)} \left( \frac{d}{dt} \left( \frac{d}{dt} (\gamma(t)) \right) \right) = 0$$

$$\Pi_{\gamma(t)} (\dot{\gamma}(t)) = 0$$

$$\ddot{\gamma}(t) = M^{-1}B \gamma(t) S_{\dot{\gamma}(t)}.$$

where the matrix $S_{\dot{\gamma}(t)}$ satisfies the following Sylvester equation

$$\gamma(t)^T B \ddot{\gamma}(t) + (\gamma(t)^T B \dot{\gamma}(t))^T = (\gamma(t)^T B M^{-1} B \gamma(t)) S_{\dot{\gamma}(t)} + S_{\dot{\gamma}(t)} (\gamma(t)^T B M^{-1} B \gamma(t)).$$

Note that we can replace $\gamma(t)^T B \ddot{\gamma}(t) + (\gamma(t)^T B \dot{\gamma}(t))^T$ by $-2 \dot{\gamma}(t)^T B \ddot{\gamma}(t)$ since $\gamma(t)^T B \gamma(t) = I_p$ when $\gamma(t) \in \mathbf{St}_B(p,d)$ and by differentiating two times with respect to $t$ we get the equality.

Thus, in order to compute the exponential map, we simply need to solve Eq. (3.23). Unfortunately, in the general case we are unaware of any analytical solution, and so the equation needs to be solved numerically. However, in the special case $M = B$, the equation can be solved analytically in a manner similar to [1, Equation 5.26]. For $M = B$, the equation for the geodesic is reduced to

$$\ddot{\gamma}(t) = -\gamma(t) (\dot{\gamma}(t)^T B \dot{\gamma}(t)) .$$

We perform a small modification of the calculation given in [19, Subsection 2.2.2] (also developed by Ross Lippert). Denote

$$C := \gamma(t)^T B \gamma(t), \quad A := \gamma(t)^T B \dot{\gamma}(t), \quad S := \dot{\gamma}(t)^T B \dot{\gamma}(t).$$

By differentiating $C, A, S$ we get the following equations:

$$\dot{C} = A + A^T ,$$

$$\dot{A} = S + \gamma(t)^T B \dot{\gamma}(t) = S - CS ,$$

$$\dot{S} = \dot{\gamma}(t)^T B \dot{\gamma}(t) + \dot{\gamma}(t)^T B \dot{\gamma}(t) = - [SA + A^T S] .$$
Recall that since $\gamma(t) \in \text{St}_B(p, d)$ we get that $C = I_p$. Thus, $\dot{C} = 0_p$, so $A = -A^T$, i.e., $A$ is skew-symmetric. Moreover, $A = 0_p$ so that $A(t) = A(0)$. In addition, the last equation can be rewritten as
\[
\dot{S} = AS - SA,
\]
and it has a closed form (see [7, Theorem 9.2] for a constant matrix $A$) solution of the form
\[
S(t) = e^{At}S(0)e^{-At}.
\]
Finally, we can use the following equation
\[
\frac{d}{dt} [\gamma(t)e^{At}, \dot{\gamma}(t)e^{At}] = [\gamma(t)e^{At}, \dot{\gamma}(t)e^{At}] \begin{pmatrix} A & -S(0) \\ I_p & A \end{pmatrix},
\]
to find a closed form for the geodesic curve
\[
\gamma(t) = [X, \xi] \exp \left( t \begin{pmatrix} A & -S(0) \\ I_p & A \end{pmatrix} \right) \begin{pmatrix} I_p \\ 0_p \end{pmatrix} e^{-At}.
\]
Substituting $t = 1$ into Eq. (3.25) gives us the exponential mapping $\text{Exp}_X\xi_X$.

### 3.3 Computational Costs

Table 1 summarizes the computational costs, measured in terms of arithmetic operations, of computing the Riemannian components on the generalized Stiefel manifold described in Subsections 3.1 and 3.2. Note that all the costs are for operations in ambient coordinates. In the table, we denote by $T_C$ the cost of computing the product of $C$ with a vector (potentially, after preprocessing $C$), for some matrix $C$. Specifically, we use $T_B, T_{B^{-1/2}}, T_{B^{1/2}}, T_M$ and $T_{M^{-1}}$. In particular, $T_M$ and $T_{M^{-1}}$ denote the maximal cost (over $X \in \text{St}_B(p, d)$) of taking the product of $M_X$ and $M_X^{-1}$ (respectively) with a vector. Also, we denote by $T_{\nabla\bar{f}}$ and by $T_{\nabla^2\bar{f}}$ the cost of computing the Euclidean gradient and the cost of applying the Euclidean Hessian to a tangent vector.

Note that when compared to the standard metric on $\text{St}_B(p, d)$ (i.e., $M_X = B$ for all $X$), we replace products with $B^{-1}$ by products with $M_X^{-1}$, and $B$ is accessed only through matrix-vector products.

| Operation | Cost |
|-----------|------|
| Retraction maps (Eq. 3.3, 3.6, 3.9) | $O(T_{BP} + dp^2)$ |
| Inverse of the polar-based retraction (Eq. 3.4) | $O(T_{BP} + dp^2)$ |
| Inverse of the QR-based retraction (Eq. 3.7) | $O(T_{BP} + dp^2 + p^2)$ |
| Vector Transport, associated with retraction (Eq. A.8, A.10) | $O(T_{BP} + dp^2)$ |
| Inner product on the tangent space (Eq. 3.10) | $O(T_{MP} + dp)$ |
| Orthogonal projections on the tangent/norma, $M_X$ metric (Eq. 3.13, 3.12) | $O(T_{BP} + T_{M^{-1}p} + dp^2)$ |
| Orthogonal projections on the tangent/norma, $B$ metric (Eq. 3.18, 3.17) | $O(T_{BP} + dp^2)$ |
| Vector Transport, based on the orthogonal projection (Eq. 3.19) | $O(T_{BP} + T_{M^{-1}p} + dp^2)$ |
| Riemannian gradient computation (Eq. 3.20) | $O(T_{BP} + T_{M^{-1}p} + T_{MP} + dp^2 + T_{\nabla\bar{f}} + T_{\nabla^2\bar{f}})$ |

Applying the Riemannian Hessian to a tangent vector (Eq. 3.21) | $O(T_{BP} + T_{M^{-1}p} + T_{MP} + dp^2 + T_{\nabla\bar{f}} + T_{\nabla^2\bar{f}})$ |
3.4 Product Manifold of Generalized Stiefel Manifolds

In some cases it is desirable to solve optimization problems with several groups of variables, in which each group is constrained to a different generalized Stiefel manifold. For example, the CCA problem is formulated as an optimization problem with two generalized orthogonality constraints. Such cases are easily addressed by using the notion of product manifold [1, Section 3.1.6]. Here, we briefly summarize how it applies to our settings.

The basic idea of the product manifold of generalized Stiefel manifolds is to simply consider the Cartesian product of separately computed Riemannian components on each of the manifolds in the product. In particular, when the number of columns is equal for all the generalized Stiefel manifolds in the product, then it is possible to simply stack the component matrices on top of each other, and performing the operations separably on each manifold.

Specifically, let \( B_1, \ldots, B_k \) be SPD matrices, where the dimension of \( B_i \) is \( d_i \times d_i \), and denote \( d = d_1 + \cdots + d_k \). Suppose that the goal is to minimize \( f(X_1, \ldots, X_k) = f(X) \) with the constraint \( X_i \in \text{St}_{B_i}(p, d_i) \) for \( i = 1, \ldots, k \). The problem can be solved using Riemannian optimization on the product manifold \( \text{St}_{B_1}(p, d_1) \times \text{St}_{B_2}(p, d_2) \times \cdots \times \text{St}_{B_k}(p, d_k) \), i.e., \( X \in \text{St}_{B_1}(p, d_1) \times \text{St}_{B_2}(p, d_2) \times \cdots \times \text{St}_{B_k}(p, d_k) \). Indeed, for the product manifold, there is a natural way to define the differentiable structure so that manifold topology of \( \text{St}_{B_1}(p, d_1) \times \text{St}_{B_2}(p, d_2) \times \cdots \times \text{St}_{B_k}(p, d_k) \) is the product topology. However, to employ Riemannian optimization it is also necessary to define a metric on the product manifold.

Suppose that on each \( \text{St}_{B_i}(p, d_k) \) the metric is defined by a smooth mapping \( X_i \mapsto M_{X_i}^{(i)} \) such that \( M_{X_i}^{(i)} \) is a SPD matrix (i.e., the metric \( g^{(i)} \) on \( \text{St}_{B_i}(p, d_i) \) is defined in ambient coordinates by \( g^{(i)}_X(\eta_X, \xi_X) = \text{Tr} \left( \eta_X M_{X_i}^{(i)} \xi_X \right) \)). The product manifold \( \text{St}_{B_1}(p, d_1) \times \text{St}_{B_2}(p, d_2) \times \cdots \times \text{St}_{B_k}(p, d_k) \) is a Riemannian submanifold of \( \mathbb{R}^{d_1 \times p} \times \mathbb{R}^{d_2 \times p} \times \cdots \times \mathbb{R}^{d_k \times p} \) endowed with the product metric (sum of the metric values on each product component). Since \( \mathbb{R}^{d_1 \times p} \times \mathbb{R}^{d_2 \times p} \times \cdots \times \mathbb{R}^{d_k \times p} \) is naturally isomorphic to \( \mathbb{R}^{d \times p} \) by stacking the matrices on top of each other, then \( \text{St}_{B_1}(p, d_1) \times \text{St}_{B_2}(p, d_2) \times \cdots \times \text{St}_{B_k}(p, d_k) \) can be viewed as a Riemannian embedded submanifold of \( \mathbb{R}^{d \times p} \) endowed by the metric defined by the \( d \times d \) matrix \( M_X := \text{blkdiag}(M_{X_1}^{(1)}, M_{X_2}^{(2)}, \ldots, M_{X_k}^{(k)}) \), and the mapping \( X \mapsto M_X \) is smooth.

The various notions introduced previously now extend to the product manifold in a straightforward way. Indeed, the tangent space of \( \text{St}_{B_1}(p, d_1) \times \text{St}_{B_2}(p, d_2) \times \cdots \times \text{St}_{B_k}(p, d_k) \) is the Cartesian product of tangent spaces of each of the generalized Stiefel manifolds. The retraction and vector transport, and orthogonal projection on the tangent space is stacking the operations performed separably on each manifold on top of each other. The Riemannian gradient is computed using the orthogonal projection to the tangent space after pre-multiplying by \( M_X^{-1} \), i.e., \( \text{grad}_{X}(f) = \Pi_X (M_X^{-1} \nabla f(X)) \) for \( X \in \text{St}_{B_1}(p, d_1) \times \text{St}_{B_2}(p, d_2) \times \cdots \times \text{St}_{B_k}(p, d_k) \), where \( \Pi_X (\cdot) \) is stacking the orthogonal projections on the tangent space of each of the manifolds on top of each other. The normal space is the product of the normal spaces of each of the manifolds manifolds. Similarly to Subsection 3.2, for the next components we assume \( M_X \) is constant. The Weingarten map is again obtained by stacking the Weingarten maps of each of the manifolds

\[
W_X (\xi_X, U_X) = \begin{bmatrix} W_{X_1}(\xi_{X_1}, U_{X_1}) \\ \vdots \\ W_{X_k}(\xi_{X_k}, U_{X_k}) \end{bmatrix}
\]  

(3.26)

where \( W_X (\xi_X, U_X) \) is the Weingarten map on \( \text{St}_{B_i}(p, d_i) \). The Riemannian connection on the product manifold is the classical directional derivative on \( \mathbb{R}^{d \times p} \) projected on the tangent space. Thus, the Riemannian Hessian can be computed using the same formula for the Riemannian Hessian on the generalized Stiefel manifold, Eq. [4.21], following similar reasoning as in Appendix A.2.

In the above, we assume the number of columns in each Stiefel component is the same in all manifold in the product. One can also work on the product manifold \( \text{St}_{B_1}(p_1, d_1) \times \text{St}_{B_2}(p_2, d_2) \times \cdots \times \text{St}_{B_k}(p_k, d_k) \) where the \( p_1, \ldots, p_k \) are not necessarily equal. In this case, we cannot simply stack the tangent vectors etc.,
but can still work with Cartesian product of the different components, and operators like $M_X$ and $B$ that operate on each component separately. Logically, this is the same as we do above for $p_1 = \cdots = p_k$, although the description is somewhat more complex, so we omit the details.

### 3.5 Metric Selection and Riemannian Hessian Conditioning

In this subsection we discuss the effects of metric selection with relation to the condition number of the Riemannian Hessian at the optimum. Similarly to the unconstrained case, the condition number of the Riemannian Hessian affects the asymptotic convergence of the various optimization algorithms – see [1, Theorem 4.5.6, Theorem 7.4.11 and Equation 7.5]. We remark that there are also (worst-case) global convergence results which guarantee sublinear convergence to first and second order (approximate) critical points (e.g., [12]). However, these guarantees require additional assumptions, e.g., Lipschitz gradient for first-order conditions and Lipschitz Hessian with second-order retraction for second-order conditions. Moreover, these guarantees do not depend on the condition number of the Riemannian Hessian. In practice, as the iterations progress and Lipschitz Hessian with second-order retraction for second-order conditions. Moreover, these guarantees require additional assumptions, e.g., Lipschitz gradient for first-order conditions and Lipschitz Hessian with second-order retraction for second-order conditions. Moreover, these guarantees do not depend on the condition number of the Riemannian Hessian. In practice, as the iterations progress linear convergence is observed (see experiments in Subsection 4.2) as guaranteed by [1, Theorem 4.5.6], and for smaller condition number the convergence is faster.

For simplicity of analysis, consider the case $p = 1$, i.e., the generalized Stiefel manifold in this case is an ellipsoid $\mathbb{S}^B$. We also assume that for all $x \in \mathbb{S}^B$ we have $M_x = M$ for some fixed SPD matrix $M$. In order to analyze the condition number of the Riemannian Hessian at the optimum, recall that the Riemannian Hessian is self-adjoint with respect to the Riemannian metric (see [1, Proposition 5.5.3]). Thus, its condition number at the optimum, $x^*$, can be found using the ratio between the maximal and minimal value of the Rayleigh quotient

$$q(\xi_{x^*}) := \frac{g_{x^*}(\xi_{x^*}, \text{Hess} f(x^*)[\xi_{x^*}])}{g_{x^*}(\xi_{x^*}, \xi_{x^*})}.$$  

Using Eq. (3.22), the Riemannian Hessian for $p = 1$ is reduced to

$$\text{Hess} f(x^*)[\eta_{x^*}] = \Pi_{x^*} \left( M_{x^*}^{-1} \left[ \nabla^2 \bar{f}(x^*) - \left( (x^*)^T \nabla \bar{f}(x^*) - g_{x^*}(x^*, \text{grad} f(x^*)) \right) B \right] \eta_{x^*} \right). \quad (3.27)$$

Recall that $\text{grad} f(x^*) = 0$, also the projection on the tangent space is self-adjoint with respect to the Riemannian metric (Eq. (3.16)), and for any $\xi_{x^*} \in T_{x^*}\mathbb{S}^B$ we have $\Pi_{x^*} (\xi_{x^*}) = \xi_{x^*}$, we get:

$$q(\xi_{x^*}) = \frac{\xi_{x^*}^T M_{x^*} \Pi_{x^*} \left( M_{x^*}^{-1} \left[ \nabla^2 \bar{f}(x^*) - \left( (x^*)^T \nabla \bar{f}(x^*) \right) B \right] \xi_{x^*} \right)}{\xi_{x^*}^T M_{x^*} \xi_{x^*}}$$

$$= \frac{(\Pi_{x^*} (\xi_{x^*}))^T \left[ \nabla^2 \bar{f}(x^*) - \left( (x^*)^T \nabla \bar{f}(x^*) \right) B \right] \xi_{x^*}}{\xi_{x^*}^T M_{x^*} \xi_{x^*}}$$

$$= \frac{\xi_{x^*}^T \left[ \nabla^2 \bar{f}(x^*) - \left( (x^*)^T \nabla \bar{f}(x^*) \right) B \right] \xi_{x^*}}{\xi_{x^*}^T M_{x^*} \xi_{x^*}}. \quad (3.28)$$

This is the Rayleigh quotient of the matrix pencil

$$\left( \nabla^2 \bar{f}(x^*) - \left( (x^*)^T \nabla \bar{f}(x^*) \right) B, M_{x^*} \right)$$

on $T_{x^*}\mathbb{S}^B$. So, if we want to bound the condition number of the Riemannian Hessian at the optimum we need to look at the pencil

$$\left( \Pi_{x^*} \left( \nabla^2 \bar{f}(x^*) - \left( (x^*)^T \nabla \bar{f}(x^*) \right) B \right) \Pi_{x^*}, \Pi_{x^*} M_{x^*} \Pi_{x^*} \right). \quad (3.29)$$
Therefore, choosing a preconditioning scheme \( x \mapsto M_x \) such that \( M_x \) is SPD for any \( x \in S^B \) and

\[
M_{x^*} \approx \nabla^2 f(x^*) - \left( (x^*)^T \nabla f(x^*) \right) B
\]

will precondition the Riemannian Hessian at the optimum. For the generalized Stiefel manifold with \( p > 1 \) such a choice is less obvious, and we leave it for future work.

Recall that the standard choice for metric selection on the generalized Stiefel manifold with \( p = 1 \) is \( M_x = B \) for all \( x \in S^B \). If \( \nabla^2 f(x^*) \) is well conditioned, it is often the case that the pencil Eq. (3.29) is well conditioned under certain assumptions. We demonstrate this in Section 4 for the problem of finding the leading correlation in CCA. In such cases, if we use a preconditioning scheme \( x \mapsto M_x \) such that \( M_{x^*} \approx B \), the condition number grows by at most \( \kappa(B, M_{x^*}) \), so if that quantity is small (i.e., \( M_{x^*} \) well approximates \( B \)) we can expect fast convergence.

### 4 Theoretical and Numerical Illustrations

#### 4.1 Simple Theoretical Examples

Our proposed preconditioning strategy for orthogonality constrained problems is based on using a preconditioning scheme to define the Riemannian metric. In this section we illustrate this point using a couple of simple examples. All examples correspond to the case \( p = 1 \), i.e., the ellipsoid.

**Example. Linear Objective.** Consider the following problem

\[
\max_{x \in \mathbb{R}^d} b^T x \quad \text{s.t.} \quad x^T B x = 1
\]

for some vector \( 0 \neq b \in \mathbb{R}^d \). It is easy to show that the solution is \( x^* = B^{-1} b / \| B^{-1} b \|_B \). It is well known that solving a linear system is equivalent to an unconstrained minimization of a quadratic objective. Here we can see that solving a linear system is also equivalent to maximizing a linear objective subject to a quadratic constraint. Note that this problem is constrained on the ellipsoid manifold \( S^B \). Let the inner product on each tangent space (the Riemannian metric) be endowed from the ambient space \( \mathbb{R}^d \). Using \( S^B \) with a metric selection \( g_x(\xi, \eta) = \xi^T M_x \eta \) (in ambient coordinates), where \( x \mapsto M_x \in \mathbb{R}^{d \times d} \) is a smooth mapping that maps \( x \in S^B \) to an SPD matrix \( M_x \), the Riemannian gradient is

\[
\nabla f(x) = (I_n - (x^T B M_x^{-1} B x)^{-1} M_x^{-1} B x x^T B) M_x^{-1} b
\]

since the Euclidean gradient is simply \( b \), independent of \( x \). Thus, using Riemannian gradient ascent on \( S^B \) with QR based retraction (Eq. (3.6)) we get the iteration

\[
y_{k+1} = x_k + \tau_k \left( M_{x_k}^{-1} b - \frac{x_k^T B M_{x_k}^{-1} b}{x_k^T B M_{x_k}^{-1} B x_k} M_{x_k}^{-1} B x_k \right)
\]

\[
x_{k+1} = \frac{y_{k+1}}{\| y_{k+1} \|_B}.
\]

We see, as expected, that the iterations depend on the choice of the Riemannian metric defined by the matrix \( M_x \). If we impose the metric \( M_x = B \) for all \( x \in S^B \), and take step size \( \tau_0 = 1/x_k^T b \), then the iteration reduces to \( x_1 = B^{-1} b / \| B^{-1} b \|_B \), and the problem is solved in a single iteration.

As expected, with \( M_x = B \) for all \( x \), the Riemannian Hessian at \( x^* \) is well conditioned. Indeed, we have

\[
\text{Hess} f(x^*) = -\Pi_{x^*} \left( (x^*^T b) I_d \right),
\]
and its corresponding Rayleigh quotient is
\[ q(\xi_x^*) = \frac{\xi_x^T B [\Pi_{x^*} (x^{*T} B) B_d] \xi_{x^*} - (x^{*T} B) = -\|B^{-1} b\|_B, \]
which is constant so the condition number equals 1.

**Example. Inverse Power Iteration.** Consider the following problem
\[
\max_{x \in \mathbb{R}^d} x^T x \quad \text{s.t.} \quad x^T B x = 1
\]
where B is an SPD matrix. The solution is equal the eigenvector corresponding the smallest eigenvalue of B, \( \lambda_d(B) \), (which is also the eigenvector corresponding to the maximum eigenvalue of \( B^{-1} \)), since this problem is equivalent to maximizing the Rayleigh quotient \( x^T x / x^T B x \). Note that this problem is constrained on the ellipsoid manifold \( S^B \). Using \( S^B \) with metric selection \( g_x(\xi_x, \eta_x) = \xi_x^T M_x \eta_x \) (in ambient coordinates), where \( M_x \in \mathbb{R}^{d \times d} \) is a SPD matrix for any \( x \in S^B \), the Riemannian gradient is
\[
\text{grad}(x) = 2(I_n - (x^T B M_x^{-1} Bx)^{-1} M_x^{-1} Bxx^T B)M_x^{-1} x
\]
since the Euclidean gradient is \( 2x \). Thus, using Riemannian gradient ascent on \( S^B \) with QR based retraction (Eq. (3.6)) we get the iteration
\[
y_{k+1} = x_k + 2\tau_k \left( M^{-1}_x x_k - \frac{x_k^T B M_x^{-1} x_k}{x_k^T B M_x^{-1} B x_k} M_x^{-1} B x_k \right)
\]
\[
x_{k+1} = \frac{y_{k+1}}{\|y_{k+1}\|_B}.
\]
If we impose the metric \( M_x = B \) for all \( x \in S^B \), and take step sizes \( \tau_k = (x_k^T x_k)^{-1} \), then the iterations reduce to \( x_{k+1} = B^{-1} x_k / \|B^{-1} x_k\|_2 \), i.e., the inverse power method, which is well known for its good convergence properties for eigenvalues near zero.

Let us examine the Riemannian Hessian at the optimal point \( x^* \) (i.e \( (x^*)^T x^* = 1/\lambda_{\text{min}}(B) = 1/\lambda_d(B) \)):
\[
\text{Hess}_f(x^*) = \Pi_{x^*} \left( B^{-1} \left[ 2I_d - 2 \left( (x^*)^T x^* \right) B \right] \right) = \Pi_{x^*} \left( B^{-1} \left[ 2I_d - 2/\lambda_d(B) \right] B \right)
\]
The corresponding Rayleigh quotient is the following from using similar reasoning as in Subsection 3.5
\[
q(\xi_{x^*}) = \frac{\xi_{x^*}^T B [\Pi_{x^*} (B^{-1} [2I_d - 2/\lambda_d(B)] B)] \xi_{x^*} - (x^{*T} B) = \xi_{x^*}^T B \xi_{x^*} \cdot \xi_{x^*}^T B \xi_{x^*}.
\]
Thus, the eigenvalues of the Riemannian Hessian at \( x^* \) correspond to the generalized eigenvalues of the matrix pencil \( (2I_d - 2/\lambda_d(B)) B \) on \( T_{x^*} S^B \), i.e., the eigenvalues of \( 2B^{-1} \) deflated by \( -2/\lambda_d(B) \) on \( T_{x^*} S^B \), thus constrained not to correspond to \( \lambda_d(B) \). The condition number is bounded by
\[
\frac{1/\lambda_{d-1}(B) - 1/\lambda_d(B)}{1/\lambda_1(B) - 1/\lambda_d(B)}
\]
which for \( \lambda_d(B) \) close to 0 and \( \lambda_{d-1}(B) \gg 0 \) is close to 1.

### 4.2 Canonical Correlation Analysis: Theory and Experiment

In this subsection we illustrate our approach on the problem of finding the top correlation between two datasets. This problem can be written as optimization problem whose constraint set is the product of two ellipsoids.
CCA, originally introduced by [18], is a well-established method in statistical learning with numerous applications (e.g., [11, 13, 17, 18, 19, 21]). In CCA the relation between a pair of datasets in matrix form is analyzed, where the goal is to find the directions of maximal correlation between a pair of observed variables. In the language of linear algebra, CCA measures the similarities between two subspaces spanned by the columns of the two matrices. Here, we consider a regularized version of CCA defined below:

**Definition 2.** Let \( X \in \mathbb{R}^{n \times d_x} \) and \( Y \in \mathbb{R}^{n \times d_y} \) be two data matrices, and \( \lambda_x, \lambda_y \geq 0 \) be two regularization parameters. Let \( q = \max \{ \text{rank} (X^T X + \lambda_x I_{d_x}), \text{rank} (Y^T Y + \lambda_y I_{d_y}) \} \).

The \( (\lambda_x, \lambda_y) \) canonical correlations \( \sigma_1 \geq \cdots \geq \sigma_q \) and the \( (\lambda_x, \lambda_y) \) canonical weights \( u_1, \ldots, u_q, v_1, \ldots, v_q \) are the ones that maximize

\[
\text{Tr} \left( U^T X^T Y V \right)
\]

subject to

\[
U^T (X^T X + \lambda_x I_{d_x}) U = I_{d_x}, \quad V^T (Y^T Y + \lambda_y I_{d_y}) V = I_{d_y}
\]

where \( U^T X^T Y V = \text{diag} (\sigma_1, \ldots, \sigma_q) \), \( U = [u_1 \ldots u_q] \) and \( V = [v_1 \ldots v_q] \).

In his paper, we focus on finding the top correlation, i.e., finding \( \sigma_1, u_1 \) and \( v_1 \). It is useful to introduce the following notations:

\[
\Sigma_{xx} = X^T X + \lambda_x I_{d_x}, \Sigma_{yy} = Y^T Y + \lambda_y I_{d_y}, \Sigma_{xy} = X^T Y.
\]

Restricting to finding the top correlation, the optimization problem becomes:

\[
\max u^T \Sigma_{xy} v \quad \text{s.t.} \quad u \in \mathbb{S}^{\Sigma_{xx}}, v \in \mathbb{S}^{\Sigma_{yy}} \tag{4.3}
\]

It is well known ([8]) that the optimal solution of Problem (4.3) is (up to the sign of the vectors)

\[
u_1 := \Sigma_{xx}^{-1/2} \phi \quad v_1 := \Sigma_{yy}^{-1/2} \psi \tag{4.4}
\]

where \( \phi \) and \( \psi \) are the left and right unit-length singular vector corresponding to the largest singular value \( \sigma_1 \) of the matrix

\[
R := \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2}. \tag{4.5}
\]

In order to conveniently use the Riemannian optimization framework, we also denote \( d = d_x + d_y \), and \( z = [u^T, v^T]^T \in \mathbb{R}^d \) where \( u \in \mathbb{R}^{d_x} \) and \( v \in \mathbb{R}^{d_y} \). Then the constraint set is a product manifold of two ellipsoids \( z \in \mathbb{S}_{xy} := \mathbb{S}^{\Sigma_{xx}} \times \mathbb{S}^{\Sigma_{yy}} \). The objective function to be minimized is then

\[
f(z) = -\frac{1}{2} z^T \begin{bmatrix} 0 & \Sigma_{xy} \\ \Sigma_{xy} & 0 \end{bmatrix} z. \tag{4.6}
\]

We endow the manifold \( \mathbb{S}^{\Sigma_{xx}} \) and \( \mathbb{S}^{\Sigma_{yy}} \) with a metric defined by two preconditioning schemes \( u \mapsto M^{(xx)}_u \) and \( v \mapsto M^{(yy)}_v \). The metric on the product manifold \( \mathbb{S}_{xy} \) is defined by \( z \mapsto M_{z} = \text{blkdiag} \left( M^{(xx)}_u, M^{(yy)}_v \right) \) as explained in Section 3.4. Using the formulas in Section 3.2, we find that the Riemannian gradient and the Riemannian Hessian (at the critical points or if \( M_{z} := M = \text{blkdiag} \left( M^{(xx)}, M^{(yy)} \right) \)) are given by:

\[
\text{grad} f(z) = \Pi_{z} \left( M_{z}^{-1} \nabla f(z) \right) = -\left[ \Pi_{u} \left( \left( M^{(xx)}_u \right)^{-1} \Sigma_{xy} v \right) \right] \quad \Pi_{v} \left( \left( M^{(yy)}_v \right)^{-1} \Sigma_{xy} u \right), \tag{4.7}
\]
\[ \text{Hess} f(z)[\eta] = \Pi_z \left( M_z^{-1} \begin{pmatrix} \begin{bmatrix} (u^T M^{(xx)} \Pi_u^{-1} (M^{(xx)})^{-1} \Sigma_{xy} v) \cdot \Sigma_{xx} \\ -\Sigma_{xy}^T \\ -\Sigma_{xy}^T \\ -\Sigma_{xy} \end{bmatrix} & \Sigma_{xy} \\ \Sigma_{xy} & -\frac{1}{\lambda} \end{bmatrix} \right) \eta_z \right) \]

Along with formulas for the retraction and vector transport (see Subsection 3.1), various Riemannian optimization algorithms can be applied to solve Problem (4.3).

As expected, at the optimal solution \( z^* = [u^T, v^T]^T \) (see Eq. (4.4)) the Riemannian gradient vanishes: \( \text{grad} f(z^*) = 0 \). Moreover, the Riemannian Hessian at the optimum becomes

\[ \text{Hess} f(z^*) = \Pi_{z^*} \left( M_{z^*}^{-1} \begin{pmatrix} 0 & \Sigma_{xy} \\ -\Sigma_{xy} & 0 \end{pmatrix} \right). \] (4.9)

Next, we demonstrate the effect of preconditioning on the condition number of the Riemannian Hessian at \( z^* \). We show that if the leading correlation is strictly larger than the second largest one, and we select a smooth preconditioning scheme \( z \mapsto M_z \) such that \( M_{z^*} = \Sigma := \text{blkdiag}(\Sigma_{xx}, \Sigma_{yy}) \), the condition number of the Riemannian Hessian at the optimum is equal to \( (\sigma_1+\sigma_2)/(\sigma_1-\sigma_2) \). Thus, if the leading correlation gap \( \sigma_1-\sigma_2 \) is \( O(\sigma_1) \) then the condition number at the optimum is \( O(1) \), and we can expect fast convergence (dependence on the gap between the correlations is expected). Furthermore, if we select a smooth preconditioning scheme \( z \mapsto M_z \) such that \( M_{z^*} \approx \Sigma \) (see for example Figure 4.1) the condition number bound grows by at most a small factor: \( \kappa(B, M_{z^*}) \).

**Lemma 3.** Assuming \( \sigma_1 - \sigma_2 > 0 \) and that \( \Sigma \) is a SPD matrix, if \( S_{xy} \) is equipped with a metric defined by a smooth preconditioning scheme \( z \mapsto M_z \) such that \( M_{z^*} = \Sigma \), then the condition number of Riemannian Hessian on \( S_{xy} \) of Eq. (4.6) at \( z^* \) is equal to \( \frac{\sigma_1+\sigma_2}{\sigma_1-\sigma_2} \). Additionally, if \( M_{z^*} \approx \Sigma \) then the condition number is bounded by \( \frac{\sigma_1+\sigma_2}{\sigma_1-\sigma_2} \cdot \kappa(B, M_{z^*}). \)

**Proof.** In order to bound the condition number of Riemannian Hessian on \( S_{xy} \) of Eq. (4.6) at \( z^* \) we use the Courant-Fischer Theorem for the compact self-adjoint linear operator \( \text{Hess} f(z^*)[\cdot] : T_{z^*} S_{xy} \to T_{z^*} S_{xy} \) over the finite dimensional vector space \( T_{z^*} S_{xy} \):

\[ \lambda_k(\text{Hess} f(z^*)) = \min_{U, \dim(U) = k-1} \max_{0 \neq \xi \in U} q(\xi), \] (4.10)

\[ \lambda_k(\text{Hess} f(z^*)) = \max_{U, \dim(U) = k} \min_{0 \neq \xi \in U} q(\xi), \] (4.11)

where

\[ q(\xi) := \frac{g_{z^*}(\xi, \text{Hess} f(z^*)(\xi))}{g_{z^*}(\xi, \xi)}, \]

is the Rayleigh quotient. In the above, \( \lambda_k(\text{Hess} f(z^*)) \) is the \( k \)-th largest eigenvalue (i.e., eigenvalues are ordered in a descending order) of \( \text{Hess} f(z^*) \), and \( U \) is a linear subspace of \( T_{z^*} S_{xy} \). In particular, the maximal and minimal eigenvalues are given by the formulas

\[ \lambda_{\max}(\text{Hess} f(z^*)) = \max_{0 \neq \xi \in T_{z^*} S_{xy}} q(\xi), \] (4.12)

\[ \lambda_{\min}(\text{Hess} f(z^*)) = \min_{0 \neq \xi \in T_{z^*} S_{xy}} q(\xi), \] (4.13)

and the condition number of the Riemannian Hessian at \( z^* \) is the ratio of these two eigenvalues.

\[ \kappa(\text{Hess} f(z^*)) = \frac{\lambda_{\max}(\text{Hess} f(z^*))}{\lambda_{\min}(\text{Hess} f(z^*))}. \]
We begin by simplifying the quotient $q(\xi^*)$. At the optimum, $z^*$, we have $f(z^*) = -u_1^T \Sigma_{xy} v_1 = -v_1^T \Sigma_{xy} u_1 = -\sigma_1$. The formula for the Riemannian Hessian, $\text{Hess} f(z^*)$, is given by Eq. (4.9). Using the following notation for the Euclidean Hessian of

$$\nabla^2 \tilde{f}(z^*) := \begin{bmatrix} 0 & -\Sigma_{xy} \\ -\Sigma_{xy}^T & 0 \end{bmatrix},$$

(4.14)

and $\Sigma$ we can compactly write Eq. (4.9):

$$\text{Hess} f(z^*) = \Pi_{z^*} (M_{z^*} \left( \nabla^2 \tilde{f}(z^*) + \sigma_1 \Sigma \right)).$$

Next, as in Subsection 3.5, recall that $\Pi_{z^*}$ is self-adjoint with respect to the Riemannian metric (Eq. (3.16)), and that for any $\xi^* \in T_{z^*} S_{xy}$ we have $\Pi_{z^*} (\xi^*) = \delta_{\xi^*}$, we get:

$$q(\delta_{z^*}) = \frac{\xi_1^T (\nabla^2 \tilde{f}(z^*) + \sigma_1 \Sigma) \xi_1}{\xi_1^T \Sigma \xi_1},$$

where we use the fact that $\Sigma$ is not singular. Note that the quotient

$$\frac{\xi_1^T (\nabla^2 \tilde{f}(z^*) + \sigma_1 \Sigma) \xi_1}{\xi_1^T \Sigma \xi_1},$$

corresponds to the Rayleigh quotient of the Riemannian Hessian at $z^*$ if $M_{z^*} = \Sigma$.

Let us first find the eigenvalues of the Riemannian Hessian for the case $M_{z^*} = \Sigma$. We perform the following invertible change of variables $\delta_{z^*} := \Sigma^{1/2} \delta_{z^*}$, to find that

$$q(\delta_{z^*}) = \frac{\eta_1^T (\Sigma^{-1/2} \nabla^2 \tilde{f}(z^*) \Sigma^{-1/2} + \sigma_1 \cdot I_d) \eta_1}{\eta_1^T \eta_1},$$

(4.10)

Denote the space of vectors $\eta_{z^*}$ such that $\Sigma^{-1/2} \eta_{z^*} \in T_{z^*} S_{xy}$ by $\Sigma^{1/2} T_{z^*} S_{xy}$, and the orthogonal space to it by $(\Sigma^{1/2} T_{z^*} S_{xy})^\perp$. The above expression, $\tilde{q}(\eta_{z^*})$, is the Rayleigh quotient for the symmetric matrix $\Sigma^{-1/2} \nabla^2 \tilde{f}(z^*) \Sigma^{-1/2} + \sigma_1 \cdot I_d$. Thus, applying the Courant-Fischer theorem for $\tilde{q}(\eta_{z^*})$, where $\xi_{z^*} \in \Sigma^{1/2} T_{z^*} S_{xy}$, the minimal and the maximal values of $R(\xi_{z^*})$, where $\xi_{z^*} \in T_{z^*} S_{xy}$, are the minimal and the maximal eigenvalues of the matrix $\Sigma^{-1/2} \nabla^2 \tilde{f}(z^*) \Sigma^{-1/2} + \sigma_1 \cdot I_d$ in the space $\Sigma^{1/2} T_{z^*} S_{xy}$.

To find the eigenvalues of the matrix $\Sigma^{-1/2} \nabla^2 \tilde{f}(z^*) \Sigma^{-1/2} + \sigma_1 \cdot I_d$ in the space $\Sigma^{1/2} T_{z^*} S_{xy}$, we first note that all the eigenvalues of $\Sigma^{-1/2} \nabla^2 \tilde{f}(z^*) \Sigma^{-1/2}$ are $-\sigma_1 < -\sigma_2 < \ldots < -\sigma_q \leq 0 < \ldots < 0 < \sigma_q \leq \ldots \leq \sigma_2 < \sigma_1$ (see [24]). So, all the eigenvalue of $\Sigma^{-1/2} \nabla^2 \tilde{f}(z^*) \Sigma^{-1/2} + \sigma_1 \cdot I_d$ are $0 < \sigma_1 - \sigma_2 < \ldots < \sigma_1 - \sigma_q \leq \ldots \leq \sigma_2 - \sigma_1 \leq \ldots \leq 0$. Next, note that the eigenspaces of $\Sigma^{-1/2} \nabla^2 \tilde{f}(z^*) \Sigma^{-1/2} + \sigma_1 \cdot I_d$ corresponding to the eigenvalues 0 and 2$\sigma_1$ is exactly the two dimensional space $(\Sigma^{1/2} T_{z^*} S_{xy})^\perp$. Indeed, according to Eq. (3.11) and Subsection 3.4

$$(\Sigma^{1/2} T_{z^*} S_{xy})^\perp = \text{span} \left\{ \Sigma^{1/2} \begin{bmatrix} u_1 \\ v_1 \end{bmatrix}, \Sigma^{1/2} \begin{bmatrix} u_1 \\ -v_1 \end{bmatrix} \right\},$$

where using Eq. (4.4)

$$\Sigma^{1/2} \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} = \begin{bmatrix} \phi \\ \psi \end{bmatrix} \quad \text{and} \quad \Sigma^{1/2} \begin{bmatrix} u_1 \\ -v_1 \end{bmatrix} = \begin{bmatrix} \phi \\ -\psi \end{bmatrix}. $$

Then, using Eq. (4.5) and Eq. (4.14) we have

$$(\Sigma^{-1/2} \nabla^2 \tilde{f}(z^*) \Sigma^{-1/2} + \sigma_1 \cdot I_d) \Sigma^{1/2} \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} = \left( \begin{bmatrix} -R^T & -R \\ -R & +\sigma_1 I_d \end{bmatrix} \begin{bmatrix} \phi \\ \psi \end{bmatrix} = 0 ,$$

20
Finally, we get $A \in \Lambda$ (we use leading $k$). Preconditioning and, Variance Reduced Gradient via an approximation of the empirical correlation matrix. In our experiments we \[ \Sigma \] via the preconditioning strategy described by Gonen et al. [25], which we term as Dominant Subspace Preconditioning was originally designed for ridge regression to speed up Stochastic Variance Reduced Gradient via an approximation of the empirical correlation matrix. In our experiments we\[ \text{use MATLAB's svds}. \]

\[
\begin{align*}
\lambda_{\max}(\text{Hess} f(z^*)) &= \max_{0 \neq \eta_* \in T_{z^*} S_{xy}} \frac{\eta_*^T (\nabla^2 f(z^*) + \sigma_1 \cdot \Sigma) \eta_*}{\eta_*^T \Sigma \eta_*}, \\
\lambda_{\min}(\text{Hess} f(z^*)) &= \min_{0 \neq \eta_* \in T_{z^*} S_{xy}} \frac{\eta_*^T \Sigma \eta_*}{\eta_*^T \Sigma \eta_*},
\end{align*}
\]

The condition number for the case $M_{z^*} = \Sigma$ is obtained by dividing the last two quantities.

Finally, we get
\[
\kappa(\text{Hess} f(z^*)) = \frac{\lambda_{\max}(\text{Hess} f(z^*))}{\lambda_{\min}(\text{Hess} f(z^*))} \leq \frac{\sigma_1 + \sigma_2}{\sigma_1 - \sigma_2} \cdot \kappa(B, M_{z^*}).
\]

We now illustrate the effect of the preconditioning scheme $z \mapsto M_z$ numerically. In our experiments, we use six metric choices with constant matrices, i.e., $M_z := M$ independent of $z \in S_{xy}$: the trivial choice of a unit matrix $M = I_d$, the standard optimal but expensive choice $M = \Sigma$, and four approximations of $\Sigma$ via the preconditioning strategy described by Gonen et al. [25], which we term as Dominant Subspace Preconditioning.

Dominant Subspace Preconditioning was originally designed for ridge regression to speed up Stochastic Variance Reduced Gradient via an approximation of the empirical correlation matrix. In our experiments we use this preconditioning strategy to approximate $X$ and $Y$. The approximation is done as follows: suppose $A \in R^{d \times d}$ be some SPD matrix, and let $A = U \Lambda U^T$ be an eigendecomposition, with the diagonal entries in $\Lambda$ sorted in descending order. Given $k$, let us denote by $U_k$ the first $k$ columns of $U$, $\Lambda_k$ denote the leading $k \times k$ minor of $\Lambda$, and $\lambda_k$ the $k$-th largest eigenvalue of $A$. The $k$-dominant subspace preconditioner of $A + \lambda I_d$ is $U(\Lambda - \lambda_k I)U^T + (\lambda_k + \lambda)I_d$. The dominant subspace can be found using a sparse SVD solver (we use MATLAB's svds).
Figure 4.1: Results for CCA with Riemannian conjugate-gradient (left) and Riemannain Trust-Region (right and bottom) with various choices of metrics.

The experiments are performed with the MEDIANILIL dataset where the dimensions are $n = 43907$, $d_x = 120$, and $d_y = 101$. The implementation is via MANOPT which is a MATLAB library that implements some Riemannian optimization algorithms [10]. In Figure 4.1, the left graph presents suboptimality vs. iteration count for Riemannian CG, and the right graph presents suboptimality vs. products with the data matrices for Riemannian Trust-Region. Note that in Riemannian Trust-Region, different iterations do a variable amount of passes over the data, thus, this is the dominant cost of the Trust-Region method. The figures demonstrate that the choice $M = \Sigma$ leads to the lowest iteration count. This observation is also supported by the condition number of the Riemannian Hessian at the optimum. We evaluated it using Manopt, and indeed, the lowest condition number, 4.03, is achieved when $M = \Sigma$, and the highest, 60.2, for $M = I_d$.

5 Conclusions

In this paper, we developed the preconditioned geometric components for optimization on the generalized Stiefel manifold. The main mechanism for introducing a preconditioner is via the Riemannian metric. The technique can be used to precondition any underlying Riemannian optimization method. Our method can also be applied to constraints which are described by the product of two or more generalized Stiefel manifolds. We demonstrated our method both theoretically and numerically on the problem of computing the dominant canonical correlation. As part of developing the related geometrical components of the generalized Stiefel manifold equipped with a non standard Riemannian metric, we evaluate the costs of computing these components and relate the preconditioner to asymptotic convergence via the condition number of the Riemannian Hessian at the optimum.

In a sense, this paper presents only part of the picture. While it presents a methodology for building preconditioned algorithms for optimization with generalized orthogonality constraints, it does not explains how to build effective preconditioners to be used in conjunction with those algorithms, and we leave it for future work. Additional research directions include addressing other constraints using similar ideas, e.g., fixed-rank matrices, products of different types of manifolds, quotient manifolds, etc.

1 Datasets were downloaded for libsvm’s website: https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
References

[1] P. A. Absil, R. Mahony, and R. Sepulchre, Optimization algorithms on matrix manifolds, Princeton University Press, 2009.

[2] P. A. Absil, R. Mahony, and J. Trumpf, An Extrinsic Look at the Riemannian Hessian, Geometric Science of Information, Springer, 2013, pp. 361–368.

[3] R. H. Bartels and G. W. Stewart, Solution of the Matrix Equation AX + XB = C, Commun. ACM 15 (1972), no. 9, 820–826.

[4] Friedrich L Bauer and Charles T Fike, Norms and exclusion theorems, Numerische Mathematik 2 (1960), no. 1, 137–141.

[5] Amir Beck and Marc Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM journal on imaging sciences 2 (2009), no. 1, 183–202.

[6] Gladyston C Bento, Orizon P Ferreira, and Jefferson G Melo, Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds, Journal of Optimization Theory and Applications 175 (2017), no. 2, 548–562.

[7] Rajendra Bhatia and Peter Rosenthal, How and why to solve the operator equation ax − xb = y, Bulletin of the London Mathematical Society 29 (1997), no. 1, 1–21.

[8] Ake Bjorck and Gene H Golub, Numerical methods for computing angles between linear subspaces, Mathematics of Computation 27 (1973), no. 123, 579–594.

[9] Hans Georg Bock, Randwertproblemmethoden zur parameteridentifizierung in systemen nichtlinerar differentialgleichungen, no. 183, Der Math.-Naturwiss. Fakultät der Universität Bonn, 1987.

[10] 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.

[11] Nicolas Boumal, An introduction to optimization on smooth manifolds, Available online, May 2020.

[12] Nicolas Boumal, Pierre-Antoine Absil, and Coralia Cartis, Global rates of convergence for nonconvex optimization on manifolds, IMA Journal of Numerical Analysis 39 (2019), no. 1, 1–33.

[13] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan, Multi-view clustering via canonical correlation analysis, Proceedings of the 26th Annual International Conference on Machine Learning (ICML), ACM, 2009, pp. 129–136.

[14] Shixiang Chen, Shiqian Ma, Anthony Man-Cho So, and Tong Zhang, Proximal gradient method for nonsmooth optimization over the Stiefel manifold, SIAM Journal on Optimization 30 (2020), no. 1, 210–239.

[15] Shixiang Chen, Shiqian Ma, Lingzhou Xue, and Hui Zou, An alternating manifold proximal gradient method for sparse principal component analysis and sparse canonical correlation analysis, INFORMS Journal on Optimization 2 (2020), no. 3, 192–208.

[16] Chris Criscitiello and Nicolas Boumal, An accelerated first-order method for non-convex optimization on manifolds, arXiv preprint arXiv:2008.02252 (2020).

[17] Paramveer Dhillon, Dean P Foster, and Lyle H Ungar, Multi-view learning of word embeddings via CCA, Advances in Neural Information Processing Systems, 2011, pp. 199–207.
[18] Paramveer S. Dhillon, Jordan Rodu, Dean P. Foster, and Lyle H. Ungar, Two step CCA: A new spectral method for estimating vector models of words, Proceedings of the 29th International Conference on International Conference on Machine Learning (USA), ICML’12, Omnipress, 2012, pp. 67–74.

[19] A. Edelman, T. Arias, and S. Smith, The geometry of algorithms with orthogonality constraints, SIAM Journal on Matrix Analysis and Applications 20 (1998), no. 2, 303–353.

[20] OP Ferreira and PR Oliveira, Proximal point algorithm on Riemannian manifolds, Optimization 51 (2002), no. 2, 257–270.

[21] Ronald A Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (1936), no. 2, 179–188.

[22] Daniel Gabay, Minimizing a differentiable function over a differential manifold, Journal of Optimization Theory and Applications 37 (1982), no. 2, 177–219.

[23] Gene H. Golub and Charles F. Van Loan, Matrix computations (4rd ed.), Johns Hopkins University Press, Baltimore, MD, USA, 2013.

[24] Gene H Golub and Hongyuan Zha, The canonical correlations of matrix pairs and their numerical computation, Linear Algebra for Signal Processing, Springer, 1995, pp. 27–49.

[25] Alon Gonen, Francesco Orabona, and Shai Shalev-Shwartz, Solving ridge regression using sketched pre-conditioned SVRG, Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML’16, JMLR.org, 2016, pp. 1397–1405.

[26] Alfred Gray, Elsa Abbena, and Simon Salamon, Modern differential geometry of curves and surfaces with mathematica, third edition (studies in advanced mathematics), Chapman & Hall/CRC, 2006.

[27] Roger A Horn, Roger A Horn, and Charles R Johnson, Matrix analysis, 2nd ed., Cambridge University Press, 2012.

[28] Harold Hotelling, Relations between two sets of variates, Biometrika 28 (1936), no. 3/4, 321–377.

[29] Wen Huang and Ke Wei, An extension of FISTA to Riemannian optimization for sparse PCA, arXiv preprint arXiv:1909.05485 (2019).

[30] Tetsuya Kaneko, Simone Fiori, and Toshihisa Tanaka, Empirical arithmetic averaging over the compact Stiefel manifold, IEEE Transactions on Signal Processing 61 (2012), no. 4, 883–894.

[31] Tae-Kyun Kim, Josef Kittler, and Roberto Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007), no. 6, 1005–1018.

[32] Daniel Kressner, Michael Steinlechner, and Bart Vandereycken, Preconditioned low-rank Riemannian optimization for linear systems with tensor product structure, SIAM Journal on Scientific Computing 38 (2016), no. 4, A2018–A2044.

[33] Mario Lezcano Casado, Trivializations for gradient-based optimization on manifolds, Advances in Neural Information Processing Systems (H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[34] Jun Li, Fuxin Li, and Sinisa Todorovic, Efficient Riemannian Optimization on the Stiefel Manifold via the Cayley Transform, International Conference on Learning Representations, 2020.
[35] David G Luenberger, *The gradient projection method along geodesics*, Management Science 18 (1972), no. 11, 620–631.

[36] Bamdev Mishra and Rodolphe Sepulchre, *R3MC: A Riemannian three-factor algorithm for low-rank matrix completion*, Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on, IEEE, 2014, pp. 1137–1142.

[37] Bamdev Mishra and Rodolphe Sepulchre, *Riemannian preconditioning*, SIAM Journal on Optimization 26 (2016), no. 1, 635–660.

[38] Uria Mor and Haim Avron, *Solving trust region subproblems using Riemannian optimization*, arXiv preprint arXiv:2010.07547 (2020).

[39] Thanh Ngo and Yousef Saad, *Scaled gradients on Grassmann manifolds for matrix completion*, Advances in Neural Information Processing Systems, 2012, pp. 1412–1420.

[40] Hiroyuki Sato and Kensuke Aihara, *Cholesky QR-based retraction on the generalized Stiefel manifold*, Computational Optimization and Applications 72 (2019), no. 2, 293–308.

[41] Yuanming Shi, Jun Zhang, and Khaled B Letaief, *Low-rank matrix completion for topological interference management by Riemannian pursuit*, arXiv preprint arXiv:1603.01729 (2016).

[42] Steven T Smith, *Optimization techniques on Riemannian manifolds*, (1994).

[43] Ya Su, Yun Fu, Xinbo Gao, and Qi Tian, *Discriminant learning through multiple principal angles for visual recognition*, IEEE Transactions on Image Processing 21 (2012), no. 3, 1381–1390.

[44] Liang Sun, Betul Ceran, and Jieping Ye, *A scalable two-stage approach for a class of dimensionality reduction techniques*, Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2010, pp. 313–322.

[45] Lloyd N Trefethen and David Bau III, *Numerical linear algebra*, vol. 50, SIAM, 1997.

[46] Constantin Udriste, *Convex functions and optimization methods on Riemannian manifolds*, vol. 297, Springer Science & Business Media, 2013.

[47] Bart Vandereycken and Stefan Vandewalle, *A Riemannian optimization approach for computing low-rank solutions of Lyapunov equations*, SIAM Journal on Matrix Analysis and Applications 31 (2010), no. 5, 2553–2579.

[48] Zaiwen Wen and Wotao Yin, *A feasible method for optimization with orthogonality constraints*, Mathematical Programming 142 (2013), no. 1, 397–434.

[49] Florian Yger, Maxime Berar, Gilles Gasso, and Alain Rakotomamonjy, *Adaptive Canonical Correlation Analysis based on matrix manifolds*, Proceedings of the 29th International Conference on Machine Learning (USA), ICML’12, Omnipress, 2012, pp. 299–306.

[50] Tengfei Zhou, Hui Qian, Zebang Shen, Chao Zhang, and Congfu Xu, *Tensor completion with side information: A Riemannian manifold approach*, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 3539–3545.

[51] Xiaojing Zhu, *A Riemannian conjugate gradient method for optimization on the Stiefel manifold*, Computational optimization and Applications 67 (2017), no. 1, 73–110.

[52] Xiaojing Zhu and Hiroyuki Sato, *Riemannian conjugate gradient methods with inverse retraction*, Computational Optimization and Applications 77 (2020), no. 3, 779–810.
A Further Details on the Preconditioned Geometric Components

In this section we elaborate on the derivations of the Riemannian components that appear in Section 3. Our main contribution is the metric dependent components in Subsection A.2. The metric independent components are included for completeness.

A.1 Metric Independent Notions

We begin with the metric independent notions that appear in Subsection 3.1. Recall that the tangent space has two common characterizations. The first characterization

\[ T_X \text{St}_B(p, d) = \left\{ Z \in \mathbb{R}^{d \times p} : Z^T B X + X^T B Z = 0_p \right\} , \tag{A.1} \]

is based on the Submersion Theorem \[1\], Proposition 3.3.3. \( \text{St}_B(p, d) \) is the kernel of the mapping \( F(X) = X^T B X - I_p \), i.e., \( \text{St}_B(p, d) = F^{-1}(0_p) \). This mapping is a submersion since the rank of \( F \) is \( p(p + 1)/2 \) (i.e., \( F \) is full rank); indeed, the rank of \( F \) is determined by the range of \( DF(X)[\cdot] : \mathbb{R}^{d \times p} \to \text{Sym}(p) \). For every \( Z \in \text{Sym}(p) \), the matrix \( Z = \frac{1}{2} X Z \in \mathbb{R}^{d \times p} \) satisfies \( DF(X)[Z] = Z \). According to \[1\] Proposition 3.3.3 then \( \text{St}_B(p, d) \) is an embedded submanifold of \( \mathbb{R}^{d \times p} \), and its dimension is \( dp - \frac{p(p+1)}{2} \).

The second characterization is

\[ T_X \text{St}_B(p, d) = \left\{ Z = X \Omega + X_{B \perp} K \in \mathbb{R}^{d \times p} : \Omega \in \mathcal{S}_{\text{skew}}(p), K \in \mathbb{R}^{(d-p) \times p} \right\} , \tag{A.2} \]

where \( \Omega \) is a skew-symmetric matrix (i.e., \( \Omega^T = -\Omega \)), \( K \) is arbitrary, and \( X_{B \perp} \subseteq \mathbb{R}^{d \times (d-p)} \) satisfies that its columns are an orthonormal basis for the orthogonal complement of the column space of \( X \) with respect to the matrix \( B \), i.e., \( X_{B \perp} B X_{B \perp} = I_{d-p} \), and \( X_{B \perp}^T B X = 0_{(d-p) \times p} \). The dimension of the space defined in Eq. (A.2) is \( p(p - 1)/2 + p(d - p) = dp - p(p + 1)/2 \). Both characterizations of \( T_X \text{St}_B(p, d) \), Eq. (A.1) and Eq. (A.2), are equal. Indeed, every \( Z \in \mathbb{R}^{d \times p} \) can be represented by \( X \Omega + X_{B \perp} K \) for arbitrary \( \Omega \in \mathbb{R}^{p \times p} \) and \( K \in \mathbb{R}^{(d-p) \times p} \) (\( dp \) degrees of freedom), where the columns of \( X \) and \( X_{B \perp} \) are linearly independent, thus each of the columns of \( Z \) can be any vector in \( \mathbb{R}^d \), and \( Z \) any matrix in \( \mathbb{R}^{d \times p} \). Suppose \( Z \) satisfies Eq. (A.1), then \( \Omega^T = -\Omega \), so that \( Z \) belongs to the set defined in Eq. (A.2). Thus, the set defined in Eq. (A.1) is a subset (subspace) of the set defined in Eq. (A.2). Finally, since both the sets defined in Eq. (A.1) and Eq. (A.2) are subspaces of \( T_X \mathbb{R}^{d \times p} \simeq \mathbb{R}^{d \times p} \), and both are with the same dimension we get that Eq. (A.1) and Eq. (A.2) are equal.

In this article, we consider the use of three retractions mappings:

\[ R_X^{\text{polar}}(\xi_X) := (X + \xi_X)(I_p + \xi_X B \xi_X)^{-1/2} \tag{A.3} \]

\[ R_X^{\text{QR}}(\xi_X) := qf(B(X + \xi_X) = B^{-1/2}qf(B^{1/2}(X + \xi_X)) \tag{A.4} \]

\[ R_X^{\text{Cayley}}(\xi_X) := (I_d - \frac{1}{2} W(\xi_X))^{-1}(I_d + \frac{1}{2} W(\xi_X))X \tag{A.5} \]

where

\[ W(\xi_X) := (I_d - \frac{1}{2}XX^TB)\xi_XXX^TB - X^TX\xi_X(I_d - \frac{1}{2}BXX^T)B. \]

The cost of computing the polar-based retraction (Eq. (A.3)) is \( O(T_B p + dp^2) \) where \( T_B \) is the cost of computing the product of \( B \) with a vector. This is evident from the formulas since none of the operations require forming \( B \), but instead require taking product of \( B \) with matrices, finding the inverse of a square root of a \( p \times p \) matrix, multiplying a \( d \times p \) matrix by a \( p \times p \) matrix, and multiplying a \( p \times d \) matrix by a \( d \times p \) matrix. This is also mentioned in \[10\] Section 3.2. The cost of computing the QR-based retraction (Eq. (A.4)) is also \( O(T_B p + dp^2) \). This is shown in \[10\] Section 3.2. Though, in \[10\], it is claimed that for large
For the second condition, we have

\[ Q \in [1, \text{Definition } 4.1.1]. \]

For the first condition we have

\[ \text{Stiefel manifold. In order to generalize it to the generalized Stiefel manifold, we show it meets the conditions } \]

\[ \lambda \text{ in } [1, \text{Definition } 4.1.1]. \]

The first condition of [1, Definition 4.1.1] is that \( R \) is a retraction. Therefore, we show this by showing that it meets the conditions in [1, Definition 4.1.1]. The first condition of [1, Definition 4.1.1] is that for every vector \( \xi \in T_x \text{St}_B(p,d) \) we have \( \frac{d}{dt} R_X(t\xi)|_{t=0} = \xi \). Denote by \( \lambda_1, \ldots, \lambda_p \geq 0 \) the eigenvalues of \( \xi^T B \xi \), then

\[
(I_p + t^2 \xi^T B \xi)^{-1/2} = Q \begin{pmatrix} 1/\sqrt{1+t^2\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 1/\sqrt{1+t^2\lambda_p} & \cdots & 1/\sqrt{1+t^2\lambda_p} \end{pmatrix} Q^T,
\]

where \( Q \) is an orthogonal matrix that diagonalizes \( \xi^T B \xi \). Then,

\[
\frac{d}{dt} R_X^\text{polar}(t\xi)|_{t=0} = \frac{d}{dt} \left[ (X + t\xi)(I_p + t^2 \xi^T B \xi)^{1/2} \right]|_{t=0} =
\]

\[
= \frac{d}{dt} \left[ (X + t\xi)Q \begin{pmatrix} 1/\sqrt{1+t^2\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 1/\sqrt{1+t^2\lambda_p} & \cdots & 1/\sqrt{1+t^2\lambda_p} \end{pmatrix} Q^T \right]|_{t=0} =
\]

\[
= \xi X \begin{pmatrix} 1/\sqrt{1+t^2\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 1/\sqrt{1+t^2\lambda_p} & \cdots & 1/\sqrt{1+t^2\lambda_p} \end{pmatrix} Q^T - (X + t\xi)Q \begin{pmatrix} \frac{t\lambda_1}{(1+t^2\lambda_1)^{1/2}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ \frac{t\lambda_p}{(1+t^2\lambda_p)^{1/2}} & \cdots & 1/\sqrt{1+t^2\lambda_p} \end{pmatrix} Q^T = \xi X.
\]

Similarly, the retraction in Eq. (A.5) is also proven to be a retraction mapping in [34, Eq. (4)] for the Stiefel manifold. In order to generalize it to the generalized Stiefel manifold, we show it meets the conditions in [1, Definition 4.1.1]. For the first condition we have \( W(0) = 0_d, \) thus

\[
R_X^\text{Cayley}(0) = (I_d - 0_d)^{-1}(I_d + 0_d)X = X.
\]

For the second condition, we have

\[
\frac{d}{dt} R_X^\text{Cayley}(t\xi)|_{t=0} = W(\xi)X = \xi X,
\]

where we used \( X^T B X = I_p \) and the definition of tangent vectors on \( \text{St}_B(p,d) \) (Eq. (A.1)), i.e., \( \xi^T B X + X^T B \xi = 0_p. \)
Let us consider the inverse of the polar retraction. Suppose that $Y = R^\text{polar}_X(\xi_X)$. Using the definition of the polar retraction, and reordering the equation we find that
\[
\xi_X = Y(I_p + \xi_X^T B \xi_X)^{1/2} - X.
\] (A.6)
Left multiply by $X^T B$, and recall that $X^T B X = I_p$, to find that
\[
X^T B \xi_X = X^T B Y(I_p + \xi_X^T B \xi_X)^{1/2} - I_p.
\]
Now using the fact that $X^T B \xi_X + \xi_B^T B X = 0_p$ (since $\xi_X$ is a tangent vector), we find that
\[
X^T B Y(I_p + \xi_X^T B \xi_X)^{1/2} + (I_p + \xi_B^T B X)^{1/2} Y^T B X - 2I_p = 0_p.
\]
Thus $Z = (I_p + \xi_X^T B \xi_X)^{1/2}$ is SPD solution to Eq. (3.5). If we can uniquely recover $(I_p + \xi_X^T B \xi_X)^{1/2}$ by solving Eq. (3.5) (something we can do in a small neighborhood of $X$ that intersects with the image of the polar retraction), we can use Eq. (A.6) to invert the polar retraction.

The derivation of the inverse of the QR retraction is similar. Suppose that $Y = R^\text{QR}_X(\xi_X)$. Using the definition of the QR-based retraction, and reordering the equation we find that
\[
\xi_X = Y R - X,
\] (A.7)
where $R$ is an upper-triangular matrix with strictly positive elements on its main diagonal such that
\[
qf \left( B^{1/2} (X + \xi_X) \right) R = B^{1/2} (X + \xi_X).
\]
To find $R$, left multiply by $X^T B$, and recall that $X^T B X = I_p$ to find that
\[
X^T B \xi_X = X^T B Y R - I_p.
\]
Now using the fact that $X^T B \xi_X + \xi_B^T B X = 0_p$ (since $\xi_X$ is a tangent vector), we find that
\[
X^T B Y R + R^T Y^T B X - 2I_p = 0_p.
\]
Thus, $R$ is an upper-triangular matrix with strictly positive elements on its main diagonal solving to Eq. (3.8). If we can uniquely recover $R$ by solving Eq. (3.8) (something we can do in a small neighborhood of $X$ that intersects with the image of the QR-based retraction), we can use Eq. (A.7) to invert the QR-based retraction.

We remind here the conditions for a unique solution for Eq. (3.8). According to [30], Eq. (14) and Algorithm 1], Eq. (3.8) is equivalent to the set of the following $p$ linear equations
\[
\bar{M}_i \bar{r}_i = b_i, \ i = 1, \ldots, p,
\]
where $\bar{M}_i$ is the $i$-th principal minor extracted from the matrix $X^T B Y$, $\bar{r}_i$ is the column-vector formed by the first $i$ elements of the $i$-th column of the matrix $R$, and $b_i$ is the column-vector whose first $i - 1$ elements are the product
\[- [m_{i1}, \ldots, m_{ij}]^T \bar{r}_j,
\]
where $j = 1, \ldots, i - 1$, $m_{ik}$ are elements of the $i$-th row of $\bar{M}_i$, and the $i$-th element of $b_i$ equals 1. Thus, this set of linear equations has a unique solution if and only if all the principal minors of $X^T B Y$ are non-singular. In addition, we also demand that the diagonal elements of $R$ are strictly positive. Note that, since for $Y$ close enough to $X$ the eigenvalues of $X^T B Y$ are strictly positive, thus $\det(\bar{M}_i) > 0$. Moreover, using Cramer’s rule for $r_{ii}$ the denominator is positive and the nominator is also positive for $Y$ close enough to $X$ which satisfies the second constraint on $R$. 

28
We also show the derivation of retraction based vector transports using equations (A.3) and (A.4) similarly to [21]. For Eq. (A.3) denote \( A(t) := I_p + (\eta X + t \xi X)^T B (\eta X + t \xi X) \), then,

\[
\tau^{(\text{polar})}_{\eta X \xi X} := D R^{\text{polar}}_X (\eta X)[\xi X] = \left. \frac{d}{dt} R^{\text{polar}}_X (\eta X + t \xi X) \right|_{t=0} = \left. \frac{d}{dt} (X + \eta X + t \xi X)(A(t))^{-1/2} \right|_{t=0} = \xi X (A(0))^{-1/2} + (X + \eta X) \left. \frac{d}{dt} (A(t))^{-1/2} \right|_{t=0} = \xi X (A(0))^{-1/2} - (X + \eta X) (A(0))^{-1/2} \left. \frac{d}{dt} (A(t))^{1/2} \right|_{t=0} (A(0))^{-1/2},
\]

where the last equality is due to the differentiation of the following two identities

\[
I = (A(t))^{-1/2} (A(t))^{1/2}, \quad A(t) = (A(t))^{1/2} (A(t))^{1/2},
\]

which leads to

\[
0 = \left. \frac{d}{dt} (A(t))^{-1/2} (A(t))^{1/2} + (A(t))^{-1/2} \frac{d}{dt} (A(t))^{1/2} \right|_{t=0},
\]

and

\[
\frac{d}{dt} A(t) = \left. \frac{d}{dt} (A(t))^{1/2} (A(t))^{1/2} + (A(t))^{1/2} \frac{d}{dt} (A(t))^{1/2} \right|_{t=0},
\]

\[
\xi^T X B \eta X + \eta^T X B \xi X + 2 \xi^T X B \xi X = \left. \frac{d}{dt} (A(t))^{1/2} (A(t))^{1/2} + (A(t))^{1/2} \frac{d}{dt} (A(t))^{1/2} \right|_{t=0}.
\]

Thus, \( \left. \frac{d}{dt} (A(t))^{1/2} \right|_{t=0} \) is a \( p \times p \) matrix which is the solution of the following Sylvester equation:

\[
\left. \frac{d}{dt} (A(t))^{1/2} \right|_{t=0} (A(0))^{1/2} + (A(0))^{1/2} \left. \frac{d}{dt} (A(t))^{1/2} \right|_{t=0} = \xi^T X B \eta X + \eta^T X B \xi X.
\]

According to [27] Theorem 2.4.4.1, there is a unique solution to Eq. (A.9) for any \( \xi^T X B \eta X + \eta^T X B \xi X \), since \( (A(t))^{-1/2} = (I_p + \eta X B \eta X)^{-1/2} \) is positive definite (\( \eta X B \eta X \) is a symmetric positive semi-definite matrix) and \( -(I_p + \eta X B \eta X)^{-1/2} \) is negative definite, thus they have no eigenvalues in common. Solving Eq. (A.9) costs \( O(p^3) \) (e.g., using the Bartels–Stewart algorithm [3]). In addition, to compute this vector transport we need to find the square root of a \( p \times p \) matrix and its inverse which also costs \( O(p^3) \), compute the product of \( B \) with matrices, compute the matrix multiplication of \( d \times p \) matrices by \( p \times p \) matrices, of \( p \times d \) matrices by \( d \times p \) matrices and of \( p \times p \) matrices by \( p \times p \) matrices. Thus, the total computational cost of using the vector transport given in Eq. (A.8) is \( O(T_B p + dp^2) \).
For Eq. (3.6):

\[
\tau^{(QR)}_{\eta X}(\eta X) : = \text{DR}^{QR}_{X}(\eta X)[\xi_X] = \text{Dqf}_B(X + \eta X)[\xi_X] \\
= B^{-1/2}\text{Dqf}\left(B^{1/2}(X + \xi_X)\right)[B^{1/2}\xi_X] \\
= B^{-1/2}\left[\text{qf}\left(B^{1/2}(X + \xi_X)\right)\rho_{\text{skew}}\left(\text{qf}\left(B^{1/2}(X + \xi_X)\right)^T B^{1/2}(X + \xi_X)\right)^{-1}\right] + \\
\quad \left(I_n - \text{qf}\left(B^{1/2}(X + \xi_X)\right)\text{qf}\left(B^{1/2}(X + \xi_X)\right)^T\right)B^{1/2}\xi_X \left(\text{qf}\left(B^{1/2}(X + \xi_X)\right)^T B^{1/2}(X + \xi_X)\right)^{-1},
\]

where the last equality is due to [1, Example 8.1.5]:

\[
\text{Dqf}(Y)[U] = \text{qf}(Y)\rho_{\text{skew}}\left(\text{qf}(Y)^T U \text{qf}(Y)^T Y\right)^{-1} + \left(I_n - \text{qf}(Y)\text{qf}(Y)^T\right) U \left(\text{qf}(Y)^T Y\right)^{-1},
\]

and \(\rho_{\text{skew}}(\cdot)\) is the skew-symmetric term of the decomposition of a square matrix \(A\) into the sum of a skew-symmetric term and an upper triangular term, i.e,

\[
(\rho_{\text{skew}}(A))_{i,j} = \begin{cases} 
A_{i,j} & i > j \\
0 & i = j \\
-A_{j,i} & i < j
\end{cases}
\]

Computing (A.10) can be done in the following way. First, computing \(B^{-1/2}\text{qf}\left(B^{1/2}(X + \xi_X)\right)\) costs \(O(T_B p + dp^2)\) (see computational cost of Eq. (3.6)). Also, computing \(\text{qf}\left(B^{1/2}(X + \xi_X)\right)^TB^{1/2}\) has the same cost since it is equivalent to computing \(R^{-1}(X + \xi_X)^TB\), where \(R\) is the \(R\) matrix of the thin I-QR decomposition of \(B^{1/2}(X + \xi_X)\), and it can be found using the Cholesky decomposition of \((X + \xi_X)^TB(X + \xi_X)\). Applying \(\rho_{\text{skew}}(\cdot)\) takes \(O(1)\). Finally, all other computations evolve products of matrices which cost at most \(O(dp^2)\) and computing the inverse of a \(p \times p\) matrix. Thus, the total computational cost of Eq. (A.10) is \(O(T_B p + dp^2)\).

Both forms of vector transport (A.8) and (A.10) satisfy [1, Definition 8.1.1]. The vector transport based on the Cayley transform is derived in [11, Eq. (16)]. It features the same computational complexity as computing the retraction (Eq. (A.5)).

A.2 Metric Related Notions

We detail the derivation of the Riemannian Hessian that led to Eq. (3.21) stated in Subsection 3.2. For the derivation of the Riemannian Hessian we assume that the preconditioning scheme defining the Riemannian metric is constant, i.e., \(M_X := M\) for all \(X \in \text{St}_B(p,d)\). We remark again that Eq. (3.21) holds also with a non-constant \(M_X\) at the critical points.

We use [1, Definition 5.5.1] of the Riemannian Hessian: For a real-valued function \(f\) on \(\text{St}_B(p,d)\), at a point \(X \in \text{St}_B(p,d)\) the Riemannian Hessian \(\text{Hess}f(X)\) is a linear mapping of \(T_X \text{St}_B(p,d)\) into itself such that

\[
\text{Hess}f(X)[\eta X] = \nabla_{\eta X} \text{grad}f(X),
\]

for all \(\eta X \in T_X \text{St}_B(p,d)\). In the previous equation, \(\nabla\) is the Riemannian connection, which should not be confused with the Euclidean gradient.

First, we find the Riemannian connection on \(\text{St}_B(p,d)\) and show that it is the classical directional derivative of vector fields projected on the tangent space. We can find the Riemannian connection in a similar
manner to the gradient computation performed in Section 3.2 by using [1 Proposition 5.3.2]: composing the connection in the ambient space with the projection on the tangent space. Let $\nabla$ be the Levi-Civita connection on $\mathbb{R}^{d \times p}$ endowed with the metric $\bar{g}$. Let $(e_1, \ldots, e_{dp}) = (E_{11}, E_{21}, \ldots, E_{d1}, E_{12}, \ldots, E_{d2}, \ldots, E_{dp})$ be the canonical basis of $\mathbb{R}^{d \times p}$, that is matrices $E_{ij} \in \mathbb{R}^{d \times p}$ such that their only non-zero element is in the $ij$-th position and its value is 1. The matrices are ordered by columns, i.e., for $i = kd + r$ where $k, r \in \mathbb{N} \cup \{0\}$ and $0 \leq k \leq p$, $0 \leq r < d$ we have that

$$e_i = \begin{cases} E_{r,(k+1)} & r \neq 0 \\ E_{d,k} & r = 0 \end{cases}$$

(first only the matrices with 1 in their first column appear, then in the second column, as so on). Then we have

$$\nabla_{\eta} \xi = \sum_{i,j} (\eta^j \xi^i - \eta^i \xi^j) \nabla_{e_i} e_j + \sum_{i,j} \eta^j \partial_i \xi^j,$$

where $\eta, \xi, e_i, \nabla_{\eta} \xi, \nabla_{e_i} \xi$ are all vector fields on $\mathbb{R}^{d \times p}$ (i.e., given a point $X \in \mathbb{R}^{d \times p}$ the vector field assigns a tangent vector in $T_X \mathbb{R}^{d \times p} \cong \mathbb{R}^{d \times p}$, e.g., $\eta(X) = \eta_X$). In particular, $\eta$ and $\xi$ are smooth local extensions of the vector fields $\eta$ and $\xi$ on $\mathbb{R}^{d \times p}$. Given a point $X \in \mathbb{R}^{d \times p}$, in the sense that for $X \in \mathbb{R}^{d \times p}$ the vector fields $\eta$ and $\xi$ assign the same tangent vectors as $\eta$ and $\xi$. Note that the vector field $\nabla_{\eta} \xi$ at $X$ depends on $\eta(X) = \eta_X$ (see [1 Proposition 5.18]). Thus, we can write $\nabla_{\eta} \xi$ at $X$ in the following way $\nabla_{\eta} \xi(X)$. In addition, given $\eta \in T_X \mathbb{R}^{d \times p}$ and a vector field $\xi$ on $\mathbb{R}^{d \times p}$, the connection $\nabla_{\eta} \xi$ is defined by $\nabla_{\eta} \xi$ according to [1 Equation 5.13] and it does not depend on the local extension of $\xi$. Recall that $(\nabla_{e_i} e_j)_{i,j} = \Gamma_{ij}^k$ ($k$-th coordinate of $\nabla_{e_i} e_j$) are the Christoffel symbols. These symbols determine the connection $\nabla$ uniquely, using the Fundamental Theorem of Riemannian Geometry for the Levi-Civita connection. The Christoffel symbols can be calculated using

$$\Gamma_{ij}^k = \frac{1}{2} \sum_{l=1}^{dp} g^{kl} (\partial_i g_{lj} + \partial_j g_{li} - \partial_l g_{ij}),$$

where $g^{kl}$ is the $(k,l)$th entry of the inverse of the matrix $dp \times dp$ matrix $G$ which is defined by

$$(G)_{kl} := g_{kl} = g_X(e_k, e_l) = g_X(E_{ij}, E_{hm}) = Tr(E_{ij}^T M E_{hm}) = \begin{cases} 0 & j \neq m \\ M_{ih} & j = m \end{cases}.$$ 

Since the components of the matrix $M$ do not depend on $X$ and on $(e_1, \ldots, e_{dp})$ (it is a constant matrix) we have $\forall i,j,k : \Gamma_{ij}^k = 0$. Therefore, $\nabla$ is reduced to the classical directional derivative in $\mathbb{R}^{d \times p}$

$$\nabla_{\eta} \xi(X) = \sum_{j=1}^{dp} \left( \eta_X \partial_i \xi^j(X) e_j \right) = J_{\xi} \eta_X,$$

where $J_{\xi} \eta_X$ denotes the Jacobian matrix of $\xi$ at $X$ in the direction $\eta_X$. Now that we have the connection on the ambient space $\mathbb{R}^{d \times p}$, which is a Riemannian manifold, we can compute the connection on the submanifold $\mathbb{R}^{d \times p}$. Given $\eta_X \in T_X \mathbb{R}^{d \times p}$ and a vector field $\xi$ on $\mathbb{R}^{d \times p}$, the Riemannian connection is (written, as usual, in terms of ambient coordinates):

$$\nabla_{\eta} \xi(X) = \Pi_X (\nabla_{\eta_X} \xi(X)) = \Pi_X \left( J_{\xi} \eta_X \right)$$

where $\eta_X = \eta_X$ and $\xi$ is any smooth local extension of $\xi$ in a neighborhood of $X \in \mathbb{R}^{d \times p}$.
Next, we can find the Riemannian Hessian using Eq. \[ \text{Hess}_f (X)[\eta_X] = \nabla_{\eta_X} \grad f (X) \]
\[ = \Pi_X (J_{h(X)} \eta_X) \]
\[ = \Pi_X \left[ P_X M^{-1} \nabla^2 \tilde{f} (X) \eta_X + (D \Pi_X) [\eta_X] M^{-1} \nabla \tilde{f} (X) \right] \]
\[ = \Pi_X \left( M^{-1} \nabla^2 \tilde{f} (X) \eta_X \right) + \Pi_X \left( (D \Pi_X)(X)[\eta_X] M^{-1} \nabla \tilde{f} (X) \right) \quad \text{(A.12)} \]
where \( \nabla \tilde{f} (X) \) and \( \nabla^2 \tilde{f} (X) \) are the Euclidean gradient and Hessian (respectively) of \( \tilde{f} \) and
\[
h : \mathbb{R}^{d \times p} \rightarrow \mathbb{R}^{d \times p}, \quad h(X) = \Pi_X \left( M^{-1} \nabla \tilde{f} (X) \right).
\]
Note that for \( X \in \text{St}_B(p, d) \) we have \( h(X) = \grad f (X) \) so \( h \) is a smooth local extension of the vector field \( \grad f \) to \( \mathbb{R}^{d \times p} \), and its Jacobian is calculated as follows
\[
J_{h(X)} \eta_X = (D \Pi_X)[\eta_X] M^{-1} \nabla \tilde{f} (X) + \Pi_X \left( M^{-1} \nabla^2 \tilde{f} (X) \eta_X \right),
\]
where \( (D \Pi_X)[\eta_X] \) (here and in Eq. \[ \text{(A.12)} \]) is the derivative at \( X \) along \( \eta_X \) of the function that maps \( X \) to \( \Pi_X \).

The main challenge in computing the Riemannian Hessian from Eq. \[ \text{(A.12)} \] is in computing \( (D \Pi_X)[\eta_X] \). In order to circumvent this issue, we use a simple modification of a result found in [2] to the case in which the Riemannian metric induced from \( \mathbb{R}^{d \times p} \) on any Riemannian submanifold of \( \mathbb{R}^{d \times p} \) is of the form \( g_X(\xi, \eta_X) = \text{Tr} (\xi^\top M \eta_X) \) where \( M \in \mathbb{R}^{d \times d} \) is any constant, SPD matrix. In order to so, first we introduce the notion of the Weingarten map (also known as the shape operator).

**Definition 4.** (Expansion of [20] Definition 13.1 and [2] Definition 1 beyond \( \mathbb{R}^d \)) Given a Riemannian manifold \( \mathcal{M} \), a point \( x \in \mathcal{M} \) on the manifold, a tangent vector \( \eta_X \in T_x \mathcal{M} \) at \( x \), and a normal vector \( u_X \in (T_x \mathcal{M})^\perp \), we define the Weingarten map by
\[
W_X(\eta_X, u_X) := -\nabla_{\eta_X} u(x) \quad \text{(A.13)}
\]
where \( u(\cdot) \) is a smooth normal vector field on \( \mathcal{M} \) which satisfies \( u(x) = u_X \). Naturally, the Weingarten map is linear in \( \eta_X \).

For the manifold \( \text{St}_B(p, d) \), viewed as an embedded submanifold of \( \mathbb{R}^{d \times p} \), Eq. \[ \text{(A.13)} \] reduces to
\[
W_X(\eta_X, U(X)) = -\nabla_{\eta_X} U(X) = -\Pi_X (J_{U(X)} \eta_X), \quad \text{(A.14)}
\]
where \( U(\cdot) \) is any smooth local extension of the normal vector field \( U(\cdot) \) such that \( U(X) = U_X \) on \( \text{St}_B(p, d) \). Now, that at a point \( X \in \text{St}_B(p, d) \) any normal vector is of the form \( U_X = M^{-1} B X \) for some \( S_X \in S_{\text{sym}}(p) \). Left multiplying by \( X^\top M \) we get \( X^\top M U_X = X^\top M M^{-1} B X S_X = S_X \). Now we can define a normal field on \( \text{St}_B(p, d) \) by the formula \( U(X) = M^{-1} B X S_X \) such that \( U(X) = U_X \) with \( S_X = X^\top M U_X \) such that \( S_X \in S_{\text{sym}}(p) \). The vector field can be extended to \( \mathbb{R}^{d \times p} \) by the same formula such that \( U(\cdot) \) and \( U(\cdot) \) coincide on \( \text{St}_B(p, d) \). Next, we calculate the Jacobian of \( U(X) \) at the direction \( \eta_X \):
\[
J_{U(X)} \eta_X = M^{-1} B \eta_X S_X;
\]
Therefore the Weingarten map for \( \text{St}_B(p, d) \) is
\[
W_X(\eta_X, U_X) = -\Pi_X (M^{-1} B \eta_X S_X) \quad \text{(A.15)}
\]
\[
=-\Pi_X (M^{-1} B \eta_X (X^\top M U_X)) \quad \text{(A.16)}
\]

The following lemma is a simple modification of [2] Theorem 1. Although the proof is almost identical, we include it here for completeness.
Lemma 5. For the Riemannian submanifold $\mathbf{St}_B(p, d)$ of $\mathbb{R}^{d \times p}$ endowed with $\bar{g}_X(\bar{\xi}_X, \bar{\eta}_X) = \text{Tr} \left( \bar{\xi}_X^T M \bar{\eta}_X \right)$ we have
\[
W_X \left( \eta_X, \Pi^\perp_X (M^{-1}U) \right) = \Pi_X \left( (D\Pi_X)[\eta_X] (M^{-1}U) \right) = \Pi_X \left( (D\Pi_X)[\eta_X] \left( \Pi^\perp_X (M^{-1}U) \right) \right),
\]
for all $X \in \mathbf{St}_B(p, d)$, $\eta_X \in T_X \mathbf{St}_B(p, d)$ and $U \in \mathbb{R}^{d \times p}$.

Proof. First, we show that
\[
\Pi_X \left( (D\Pi_X)[\eta_X] \right) = \Pi_X \left( (D\Pi_X)[\eta_X] \left( \Pi^\perp_X (M^{-1}U) \right) \right)
\]
holds. Then applying both sided on $M^{-1}U$ gives us the equality
\[
\Pi_X \left( (D\Pi_X)[\eta_X]M^{-1}U \right) = \Pi_X \left( (D\Pi_X)[\eta_X] \left( \Pi^\perp_X (M^{-1}U) \right) \right).
\]
To show this, we take the directional derivative of the equality $\Pi_X \left( \Pi^\perp_X (\cdot) \right) = 0$ in the direction $\eta_X$, and we use $\Pi_X(\cdot) = (\text{id}_{T_X \mathbf{St}_B(p, d)} - \Pi_X)(\cdot)$ to get
\[
0 = (D\Pi_X)[\eta_X] \Pi^\perp_X (\cdot) + \Pi_X \left( (D\Pi_X)[\eta_X] \left( \Pi^\perp_X (\cdot) \right) \right)
\]
Substituting any tangent vector in both sides of the equation nullifies the term $(D\Pi_X)[\eta_X] \Pi^\perp_X (\cdot)$. Thus, we substitute $\Pi_X$ and use $\Pi_X(\Pi_X(\cdot)) = 0$
\[
0 = \Pi_X \left( (D\Pi_X)[\eta_X] \left( \Pi_X (\cdot) \right) \right).
\]
Finally, to get (A.17), we use $\text{id}_{T_X \mathbf{St}_B(p, d)}(\cdot) = (\Pi_X + \Pi^\perp_X)(\cdot)$ and get
\[
\Pi_X \left( (D\Pi_X)[\eta_X] \right) = \Pi_X \left( (D\Pi_X)[\eta_X] \left( \left( \Pi_X + \Pi^\perp_X \right) (\cdot) \right) \right) = \Pi_X \left( (D\Pi_X)[\eta_X] \left( \Pi^\perp_X (\cdot) \right) \right).
\]
To conclude the proof we show that $W_X \left( \eta_X, \Pi^\perp_X (M^{-1}U) \right) = \Pi_X \left( (D\Pi_X)[\eta_X] (M^{-1}U) \right)$. Note that for embedded submanifolds of $\mathbb{R}^{d \times p}$ with a metric derived from $M$, the Weingarten map reduces to $W_X (\eta_X, U_X) = -\Pi_X \left( J_{U_X} \eta_X \right)$. Using Definition 4 along this observation, we have
\[
W_X \left( \eta_X, \Pi^\perp_X (M^{-1}U) \right) = -\Pi_X \left( J_{\Pi^\perp_X (M^{-1}U_X)} \eta_X \right) = -\Pi_X \left( (D\Pi_X)[\eta_X] (M^{-1}U_X) \right) - \Pi_X \left( \Pi^\perp_X (J_{M^{-1}U_X} \eta_X) \right) = \Pi_X \left( (D\Pi_X)[\eta_X] (M^{-1}U_X) \right),
\]
where in the last equality we used $\Pi_X \left( \Pi^\perp_X (\cdot) \right) = 0$ and $\Pi_X(\cdot) = (\text{id}_{T_X \mathbf{St}_B(p, d)} - \Pi_X)(\cdot)$.

As a consequence of Lemma 5, we can replace $\Pi_X \left( (D\Pi_X)[\eta_X] (M^{-1} \nabla \bar{f}(X)) \right)$ with $W_X \left( \eta_X, \Pi^\perp_X (M^{-1} \nabla \bar{f}(X)) \right)$ in Eq. (A.12). Therefore the expression for the Riemannian Hessian becomes
\[
\text{Hess}_f(X)[\eta_X] = \Pi_X \left( (M^{-1} \nabla^2 \bar{f}(X)) \eta_X \right) + W_X \left( \eta_X, \Pi^\perp_X (M^{-1} \nabla \bar{f}(X)) \right).
\]
In particular, the Riemannian Hessian on $\text{St}_B(p,d)$ is

$$\text{Hess}_f(X)[\eta_X] = \Pi_X \left( M^{-1} \nabla^2 \bar{f}(X) \eta_X \right) - \Pi_X \left( M^{-1} B \eta_X \left( X^T M \left( \Pi_X (M^{-1} \nabla \bar{f}(X)) \right) \right) \right) \quad (A.19)$$

Note that some simplification of these expressions can be made by using $\Pi_X = \text{id}_{T_X \text{St}_B(p,d)} - \Pi_X$:

$$\text{Hess}_f(X)[\eta_X] = \Pi_X \left( M^{-1} \nabla^2 \bar{f}(X) \eta_X \right) - \Pi_X \left( M^{-1} B \eta_X \left( X^T \nabla \bar{f}(X) - X^T M \left( \Pi_X (M^{-1} \nabla \bar{f}(X)) \right) \right) \right)$$

$$= \Pi_X \left( M^{-1} \nabla^2 \bar{f}(X) \eta_X \right) - \Pi_X \left( M^{-1} B \eta_X \left( X^T \nabla \bar{f}(X) - X^T M \text{grad} f(X) \right) \right).$$