arXiv:submit/4600762 [math.NA] 15 Nov 2022

Riemannian optimization with a preconditioning scheme on the generalized Stiefel manifold

Boris Shustin and Haim Avron
Tel Aviv University

November 15, 2022

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 symmetric (generalized) eigenvalue problems, 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 convergence 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 by a precoditioner 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 constraints 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 $-\mathbf{Tr}\left(X^{\mathrm{T}}AX\right)$ subject to $X^{\mathrm{T}}BX = I_p$ (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 constraints are prevalent across science and engineering. Examples include, the Trust-Region Subproblem, Canonical Correlation Analysis (CCA) [1], and Fisher Linear Discriminant Analysis [2].

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 $P^{\mathrm{T}}Q$ where P and Q are orthonormal matrices whose column space spans the column space of X and Y (respectively) [3]. 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* [4, 5, 6]. Indeed, when we have a single generalized orthogonality constraint of the form $X^{\mathrm{T}}BX = I_p$, e.g., we want to minimize $f(X)$ s.t. $X^{\mathrm{T}}BX = I_p$, one can impose the structure of a smooth manifold on the constraint set, thereby obtaining the generalized Stiefel manifold

$$\mathbf{St}_B(p,d) \coloneqq \left\{X \in \mathbb{R}^{d\times p} :\ X^{\mathrm{T}}BX = I_p\right\} . \tag{1.1}$$

(see [5, Propositions 3.3.3 and 3.3.4]), and use Riemannian optimization to minimize $f(X)$ s.t. $X \in \mathbf{St}_B(p,d)$. If we have $k > 1$ generalized orthogonality constraints , e.g., minimizing $f(X_1,\dots,X_k)$ s.t. $X_i \in \mathbf{St}_B(p_i,d_i)\,(i = 1,\dots,k)$, 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 on the generalized Stiefel manifold $\mathbf{St}_B(p,d)$ we must further impose a Riemannian metric on the tangent bundle of $\mathbf{St}_B(p,d)$. We refer to the Riemannian metric naturally inherited by the scaled inner product $\langle U,V\rangle_B = \mathbf{Tr}\left(U^{\mathrm{T}}BV\right)$ on $\mathbb{R}^{d\times p}$ as the *standard metric* for (see [5, 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 [4], and this is also the metric used in the implementation of the generalized Stiefel manifold in Manopt [7]. Some of the geometric components for working with $\mathbf{St}_B(p,d)$ equipped with the standard metric appear in [4, 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 computations of some of the geometric components necessary for Riemaniann optimization on the generalized Stiefel manifold, e.g., the Riemannian gradient and Hessian, require 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, $\langle U,V\rangle_B = \mathbf{Tr}\left(U^{\mathrm{T}}BV\right)$, 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 $\mathbf{St}_B(p,d)$ with a metric inherited by the inner product $\langle U,V\rangle_{M_X} = \mathbf{Tr}\left(U^{\mathrm{T}}M_XV\right)$ on $\mathbb{R}^{d\times p}$ for some smooth mapping $X \mapsto M_X$ that maps a $X \in \mathbf{St}_B(p,d)$ to an SPD matrix M_X. Using such a mapping is an instance of so-called *Riemannian preconditioning* [8], 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 matrix 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 $\mathbf{St}_B(p,d)$ 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 constraints 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 [9, 10, 11]. A more recent and detailed introduction can be found in [5] and in [6].

Riemannian Optimization on the (Generalized) Stiefel Manifold. Optimization with 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 with orthogonality constraints was considered in the seminal work of Edelman et al. [4], and in particular the components of the Stiefel manifold were developed with the *standard* (and also the *canonical*) metric. Some recent works include [12, 13, 14], where the Cayley transform is used to define a retraction map which leads to more efficient algorithms. Another improved retraction computation is proposed in [15], where Sato and Aihara proposed a Cholesky QR-based retraction on the generalized Stiefel manifold. In [16], 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 [17]), see for example [18, 19, 20]. Also in the context of this paper, a Riemannian optimization approach for adaptive CCA on a product manifold of two generalized Stifel manifolds was proposed in [21]. In addition, components for the complex Stiefel manifold with the standard metric were developed in several works , e.g., [22, 23, 24, 25]. Unlike in our work, all the aforementioned works only use either the standard or the canonical [4] metrics 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 Riemannain Hessian at the optimum is highly indicative of the asymptotic convergence rate of Riemannian optimization (e.g., [5, Theorem 4.5.6, Theorem 7.4.11 and Eq. (7.50)]). If the objective function is convex (in the Riemannian sense [26, 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., [26, 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., [27, 28, 29, 30]. 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 [8], 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 [8] of Riemannian steepest-descent (the iterations become a version of 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 [8], we do not commit to a specific structure of the metric, as long as it is inherited from the ambient space. Our framework is suitable for the use of the metrics presented in [8], but also allows to use easier to compute metrics. Moreover, 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 [31]. The aforementioned preconditioning approach generalizes the preconditioning strategy for the unconstrained trust-region problem. Another example of using Riemannian preconditioning for the Trust-Region Subproblem can be found in [32].

A different approach for preconditioning of Riemannian methods can be found in [33] 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 an SPD preconditioner (and then projected to the tangent space). Another method proposed in [33] is an approximate Riemannian Newton method where the search direction is determined by an equation involving an approximation to the Riemannian Hessian (known as constrained Gauss–Newton, see e.g., [34]), 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 lower case English letters $x, y, \ldots$. Vectors are explicitly defined and also denoted by $x, y, \ldots$. Matrices are denoted by $A, B, \ldots$ 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 to which they correspond (e.g., η_x). Normal vectors (of a manifold) are denoted using lower and upper case English letters with a subscript for the point on the manifold to which they correspond (e.g., u_x). Vector fields on a manifold are denoted using lower case Greek letters with brackets indicating the point on the manifold to which they correspond (e.g., $\eta(x)$). Normal vector fields on a manifold are denoted using lower and upper case English letters with brackets indicating the point on the manifold to which they correspond (e.g., $u(x)$). We use the convention that vectors are column-vectors.

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

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

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

We denote by $\mathbf{St}_B(p, d)$ the generalized Stiefel manifold defined by (1.1). $\mathbf{St}_B(p, d)$ is a submanifold of $\mathbb{R}^{d \times p}$. Given a function or vector field defined on $\mathbf{St}_B(p, d)$, 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 $f : \mathbf{St}_B(p, d) \to \mathbb{R}$, we use $\bar{f} : \mathbb{R}^{d \times p} \to \mathbb{R}$ to denote a smooth real-valued function defined on $\mathbb{R}^{d \times p}$ whose restriction to $\mathbf{St}_B(p, d)$ is f.

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

$$\mathbb{S}^B \coloneqq \left\{ x \in \mathbb{R}^d \,:\, x^{\mathrm{T}} B x = 1 \right\} .$$

In the special case $B = I_d$, we denote by $\mathbf{St}(p, n)$ the Stiefel manifold defined by

$$\mathbf{St}(p, d) \coloneqq \left\{ X \in \mathbb{R}^{d \times p} :\ X^{\mathrm{T}} X = I_p \right\} .$$

Given an SPD matrix $B \in \mathbb{R}^{d \times d}$, we say that a decomposition $A = QR$ 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 $Q \in \mathbf{St}_B(p, d)$ and $R \in \mathbb{R}^{p \times p}$ is an upper triangular matrix. Note that the standard thin QR decomposition ([35, 36, 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 $Q \in \mathbf{St}(p, d)$ then $A = (B^{-1/2}Q)R$ and $B^{-1/2}Q \in \mathbf{St}_B(p, d)$. The thin QR decomposition is unique if A is full rank and we require R to have strictly positive diagonal elements ([35, 36, 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 $\mathbf{qf}_B(A)$ the unique Q factor of the thin B-QR decomposition. For the thin I_d-QR decomposition we abbreviate $\mathbf{qf}(A) \coloneqq \mathbf{qf}_{I_d}(A)$. Using this notation we have the following relation [15]:

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

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 $(\bar{\mathcal{M}}, \bar{g})$, if $\mathcal{M}$ is a submanifold of $\bar{\mathcal{M}}$ and it inherits the metric in a natural way: $g_x(\eta_x,) = \bar{g}_x(\eta_x,)$ for $\eta_x, \in T_x\mathcal{M}$ where in the right-side η_x and are viewed as elements in $T_x\bar{\mathcal{M}}$ (this is possible since $\mathcal{M}$ is a submanifold of $\bar{\mathcal{M}}$). 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 of the cost function, and then (possibly after several steps on the tangent space, e.g., [37, 38]) translates to the manifold using the retraction mapping $R_x : T_x\mathcal{M} \to \mathcal{M}$ [5, Section 4.1]. Manipulation of tangent vectors from different tangent spaces is performed via the vector transport $\tau_{\eta_x} \in T_{R_x(\eta_x)}\mathcal{M}$ [5, Section 8.1]. In particular, the exponential mapping [5, Section 5.4] and parallel translation [5, Section 5.4] are examples of retraction and vector transport based on geodesics. Note that computing them is costly, since it requires solving a system of differential equations which might be solvable only numerically.

The notions of Riemannian gradient and Riemannian Hessian [5, 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 $\mathbf{grad}f(x) \in T_x\mathcal{M}$ and $\mathbf{Hess}f(x) : T_x\mathcal{M} \to T_x\mathcal{M}$ respectively. Roughly speaking, the Levi-Civita (Riemannian connection) ∇ 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., [5, 6] for an extensive overview of smooth techniques, and [39, 40, 18, 19, 20] for some examples of non-smooth algorithms). For example, a variant of Riemannian gradient descent is

$$x_{k+1} = R_{x_k}(-\alpha_k \mathbf{grad}f(x_k))$$

where α_k is the step size (possibly chosen by the Armijo's backtracking procedure; see [5, 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 $\mathbf{St}_B(p,d)$ with a preconditioned Riemannian metric. In the following, $B \in \mathbb{R}^{d\times d}$ is an SPD matrix and we treat $\mathbf{St}_B(p,d)$ as an embedded submanifold of $\mathbb{R}^{d\times p}$. Unlike previous articles in the literature, we allow for a wider array of Riemannain metrics on $\mathbf{St}_B(p,d)$, 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 $\mathbf{St}_B(p,d)$. 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 local coordinates of the manifold $\mathbf{St}_B(p,d)$.

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., [21, 15, 13, 16, 14, 20]); we include these definitions and formulas, and their derivations (which appear in the appendix), for completeness. Additionally, most of the formulas in this section can be derived via the known components of the Stiefel manifold [5] via the change of variables $\hat{X} = B^{1/2}X$.

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 [4, 5, 13, 14]. The inverse retraction is used in several recent algorithms proposed in the literature: Riemannian CG with inverse retractions [41], Riemannian FISTA [20], and empirical arithmetic averaging over the Stiefel manifold [16].

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

$$T_X\mathbf{St}_B(p,d) = \left\{ Z \in \mathbb{R}^{d\times p} \,:\, Z^{\mathrm{T}}BX + X^{\mathrm{T}}BZ = 0_p \right\}. \tag{3.1}$$

To explain (3.1), note that $\mathrm{St}_B(p,d)$ is the kernel of $F(X) = X^{\mathrm{T}}BX - I_p$ which is a submersion [5, Proposition 3.3.3] (see further details in Appendix A.1). 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 $Z \in T_X\mathbf{St}_B(p,d)$ then the matrix $X^{\mathrm{T}}BZ$ is skew-symmetric. Thus, a different characterization of $T_X\mathbf{St}_B(p,d)$ is as a decomposition of every tangent vector into a sum of a product of a skew-symmetric matrix with X, and a term whose columns are B-orthogonal to the columns of X:

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

where Ω is a skew-symmetric matrix (i.e., $\Omega^{\mathrm{T}} = -\Omega$), K is arbitrary, and $X_{B\perp} \in \mathbb{R}^{d\times(d-p)}$ 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., $X_{B\perp}^{\mathrm{T}}BX_{B\perp} = I_{d-p}$, and $X_{B\perp}^{\mathrm{T}}BX = 0_{(d-p)\times p}$.

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 $Q \in \mathrm{St}_B(p,d)$ and P is an SPD matrix of the size $p \times p$; one such decomposition is $P = \left(A^{\mathrm{T}}BA\right)^{1/2}$ and $Q = A\left(A^{\mathrm{T}}BA\right)^{-1/2}$):

$$R_X^{\mathrm{polar}}(\xi_X) \coloneqq (X + \xi_X)(I_p + \xi_X^{\mathrm{T}}B\xi_X)^{-1/2}, \tag{3.3}$$

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

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

$$R_X^{\mathrm{polar}^{-1}}(Y) \coloneqq YZ - X, \tag{3.4}$$

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

$$2I_p = X^{\mathrm{T}}BYZ + ZY^{\mathrm{T}}BX. \tag{3.5}$$

Thus, once BX is computed we can compute $R_X^{\mathrm{polar}^{-1}}(Y)$ using $O(dp^2)$ operations. The expression for the inverse retraction in (3.4) is valid when the Lyapunov (3.5) has a unique SPD solution. If $Y = R_X^{\mathrm{polar}}(\xi_X)$ for some ξ_X, then (3.4) has an SPD solution $Z = (I_p + \xi_X^{\mathrm{T}}B\xi_X)^{-1/2}$ (see Appendix A.1 for more details). Let us now consider when (3.4) has a unique solution. It has a unique solution if and only if $X^{\mathrm{T}}BY$ and $-Y^{\mathrm{T}}BX$ do not share any eigenvalue [42, Theorem 2.4.4.1]. Both $X^{\mathrm{T}}BY$ and $-Y^{\mathrm{T}}BX$ are invertible since they are products of full rank matrices, thus all eigenvalues are not equal to zero. Next recall that $X^{\mathrm{T}}BX = I_p$, and that eigenvalues of a matrix are a continuous function of the matrix. Using the Bauer–Fike theorem [43] for a small enough perturbation of the matrix $X^{\mathrm{T}}BX$, i.e., $X^{\mathrm{T}}BX + \delta X^{\mathrm{T}}BX = X^{\mathrm{T}}BY$, the eigenvalues of $X^{\mathrm{T}}BY$ do not differ from the eigenvalues of $X^{\mathrm{T}}BX$ more than the norm of the perturbation. Thus, the real part of the eigenvalues of $X^{\mathrm{T}}BY$ remains strictly positive, leading to $X^{\mathrm{T}}BY$ and $-Y^{\mathrm{T}}BX$ not sharing any eigenvalue. The validity of (3.4) is the intersection of the image of $R_X^{\mathrm{polar}}(\cdot)$ and a neighborhood of X in which (3.5) has a unique solution.

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

$$R_X^{\mathrm{QR}}(\xi_X) \coloneqq \mathbf{qf}_B\left(X + \xi_X\right) = B^{-1/2}\mathbf{qf}\left(B^{1/2}\left(X + \xi_X\right)\right), \tag{3.6}$$

where $\xi_X \in T_X\mathrm{St}_B(p,d)$ [15]. One can show that if $R^{\mathrm{T}}R = (X+\xi_X)^{\mathrm{T}}B(X+\xi_X)$ is a Cholesky decomposition then $\mathbf{qf}_B\left(X + \xi_X\right) = (X + \xi_X)R^{-1}$, so once $B(X + \xi_X)$ has been computed we can compute $R_X^{\mathrm{QR}}(\xi_X)$ using $O(dp^2)$ operations [15].

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

$$R_X^{\mathrm{QR}^{-1}}(Y) \coloneqq YR - X, \tag{3.7}$$

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:

$$2I_p = X^{\mathrm{T}}BYR + R^{\mathrm{T}}Y^{\mathrm{T}}BX. \tag{3.8}$$

Solving this equation takes $O(p^4)$ operations [16]. Thus, once BX is computed we can compute $R_X^{\mathrm{QR}^{-1}}(Y)$ 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 [16, Eq. (14) and Algorithm 1]. According to Kaneko et al., using the constraint that R is upper-triangular we can reformulate (3.8) as an equivalent set of linear equations which has a unique solution if and only if all the principal minors of $X^{\mathrm{T}}BY$ are non-singular (see Appendix (A.1)). Similarly to the argument for inverse of the polar inverse retraction, since $X^{\mathrm{T}}BX = I_p$, for a small enough perturbation, i.e., $X^{\mathrm{T}}BX + \delta X^{\mathrm{T}}BX = X^{\mathrm{T}}BY$, the real part of the eigenvalues of $X^{\mathrm{T}}BY$ remains strictly positive, leading to non-singularity of $X^{\mathrm{T}}BY$. 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 (3.7) is the intersection of the image of $R_X^{\mathrm{QR}}(\cdot)$ and a neighborhood of X in which (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 [16]):

$$R_X^{\mathrm{Cayley}}(\xi_X) \coloneqq \left(I_d - \frac{1}{2}W(\xi_X)\right)^{-1}\left(I_d + \frac{1}{2}W(\xi_X)\right)X, \tag{3.9}$$

where

$$W(\xi_X) \coloneqq (I_d - \frac{1}{2}XX^{\mathrm{T}}B)\xi_X X^{\mathrm{T}}B - X\xi_X^{\mathrm{T}}(I_d - \frac{1}{2}BXX^{\mathrm{T}})B.$$

Once the multiplications with B are computed, we need to compute the inverse of a $d \times d$ matrix in order to find $R_X^{\mathrm{Cayley}}(\xi_X)$. However, noticing that

$$W(\xi_X) = \begin{bmatrix} (I_d - \frac{1}{2}XX^{\mathrm{T}}B)\xi_X & X \end{bmatrix} \begin{bmatrix} X^{\mathrm{T}} \\ -\xi_X^{\mathrm{T}}(I_d - \frac{1}{2}XX^{\mathrm{T}}B)^{\mathrm{T}} \end{bmatrix} B,$$

(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 [16].

Similarly, there are several possible ways to compute a vector transport. It is possible to define a metric independent vector transport, using [5, Equation 8.6] by differentiating a retraction mapping

$$\tau_{\eta_X}^{(\mathrm{ind})}\xi_X \quad \coloneqq \quad \mathrm{D}R_X(\eta_X)[\xi_X].$$

In Appendix A.2, we derive concrete formulas based on the polar and QR retractions, (3.3) and (3.6). A vector transport based on the Cayley retraction is presented in [13]. The various vector transport have the same computational cost as computing the corresponding retractions. 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 $\hat{X} = B^{1/2}X$ unless a specific metric is used (corresponding to $M_X = B$ for all $X \in \mathbf{St}_B(p,d)$), since though this change of variables makes $\hat{X} \in \mathbf{St}(p,d)$, the induced metric on that manifold is not the standard metric. Note that if indeed $M_X = B$ for all $X \in \mathbf{St}_B(p,d)$, then via the change of variables $\hat{X} = B^{1/2}X$ we have that $\hat{X} \in \mathbf{St}(p,d)$ with the corresponding standard metric $M_{\hat{X}} = I$ for all $\hat{X} \in \mathbf{St}(p,d)$. Unfortunately, the change of variables $\hat{X} = B^{1/2}X$ forces us to form B explicitly, which is prohibited in problems where computing B is as expensive as solving them with a direct method.

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

$$\bar{g}_X(\bar{\xi}_X, \bar{\eta}_X) \coloneqq \left\langle \bar{\xi}_X, \bar{\eta}_X \right\rangle_{M_X} = \mathbf{Tr}\left(\bar{\xi}_X^{\mathrm{T}} M_X \bar{\eta}_X\right)$$

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 an 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 $X \in \mathrm{St}_B(p,d)$, $\xi_X, \eta_X \in T_X\mathbf{St}_B(p,d)$, given in ambient space coordinates, the Riemannian metric on $\mathrm{St}_B(p,d)$ is given by

$$g_X(\xi_X, \eta_X) \coloneqq \langle \xi_X, \eta_X \rangle_{M_X} = \mathbf{Tr}\left(\xi_X^{\mathrm{T}} M_X \eta_X\right). \tag{3.10}$$

The cost of computing $g_X(\xi_X, \eta_X)$ is $O\left(T_M p + dp\right)$ 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 3.5. Classically, the metric employed for the generalized Stiefel manifold corresponds to $M_X = B$ for all $X \in \mathbf{St}_B(p,d)$ [4]. 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 4.2). 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 $X \in \mathbf{St}_B(p,d)$. 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 $\mathrm{St}_B(p,d)$ 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 $T_X\mathrm{St}_B(p,d)$ by $\Pi_X(\cdot)$, and the orthogonal projection operator (with respect to the metric defined by $X \mapsto M_X$) on the normal space, $(T_X\mathrm{St}_B(p,d))^{\perp}$, by $\Pi_X^{\perp}(\cdot)$.

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

$$(T_X\mathbf{St}_B(p,d))^{\perp} = \left\{M_X^{-1}BXS \;:\; S \in \mathcal{S}_{\mathrm{sym}}(p)\right\}. \tag{3.11}$$

Indeed, recall that $\mathbf{Tr}\left(S^{\mathrm{T}}\Omega\right) = 0$ for any symmetric matrix S and anti-symmetric matrix Ω, thus by using the representation in (3.2) of tangent vectors we get that any vector of the form $M_X^{-1}BXS$ where $S \in \mathcal{S}_{\mathrm{sym}}(p)$ is orthogonal to the tangent space at $X \in \mathbf{St}_B(p,d)$. The dimension of the normal space should be $p(p+1)/2$, thus since the set $\left\{M_X^{-1}BXS \;:\; S \in \mathcal{S}_{\mathrm{sym}}(p)\right\}$ is $p(p+1)/2$ dimensional, it is indeed the normal space.

The following lemma gives a formula for the orthogonal projections to the tangent and normal spaces.

Lemma 3.1. *The orthogonal projections with respect to $g_X(\cdot,\cdot)$ on $(T_X\mathbf{St}_B(p,d))^{\perp}$ and on $T_X\mathbf{St}_B(p,d)$ (viewed as a subspace of $T_X\mathbb{R}^{d\times p} \simeq \mathbb{R}^{d\times p}$ and given in ambient coordinates) are:*

$$\Pi_X^{\perp}(\xi_X) = M_X^{-1}BXS_{\xi_X} \tag{3.12}$$

and

$$\Pi_X(\xi_X) = \left(id_{T_X\mathbb{R}^{d\times p}} - \Pi_X^{\perp}\right)(\xi_X) = \xi_X - M_X^{-1}BXS_{\xi_X} \tag{3.13}$$

where $\xi_X \in T_X\mathbb{R}^{d\times p}$, $id_{T_X\mathbb{R}^{d\times p}}$ denotes the identity mapping on $T_X\mathbb{R}^{d\times p}$, and $S_{\xi_X} \in \mathbb{R}^{p\times p}$ is the unique solution of the following Sylvester equation

$$\left(X^{\mathrm{T}}BM_X^{-1}BX\right)S_{\xi_X} + S_{\xi_X}\left(X^{\mathrm{T}}BM_X^{-1}BX\right) = X^{\mathrm{T}}B\xi_X + \left(X^{\mathrm{T}}B\xi_X\right)^{\mathrm{T}}.$$

The cost of computing (in ambient coordinates) $\Pi_X(\xi_X)$ for an arbitrary ξ_X is $O(T_Bp + T_{M^{-1}}p + dp^2)$, 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 $X \in \mathbf{St}_B(p,d)$.

Proof. Note that $T_X\mathbf{St}_B(p,d) \oplus (T_X\mathbf{St}_B(p,d))^\perp = T_X\mathbb{R}^{d\times p} \simeq \mathbb{R}^{d\times p}$. This implies that for any $\xi_X \in T_X\mathbb{R}^{d\times p} \simeq \mathbb{R}^{d\times p}$ there exists unique $\Omega_{\xi_X} \in \mathcal{S}_{\text{skew}}(p)$, $K_{\xi_X} \in \mathbb{R}^{(d-p)\times p}$ and $S_{\xi_X} \in \mathcal{S}_{\text{sym}}(p)$ such that ξ_X is decomposed to a unique component on the tangent space of $\mathbf{St}_B(p,d)$ and a unique component on the normal space of $\mathbf{St}_B(p,d)$:

$$\xi_X = \Pi_X(\xi_X) + \Pi_X^\perp(\xi_X) = (X\Omega_{\xi_X} + X_{B\perp}K_{\xi_X}) + M_X^{-1}BXS_{\xi_X}. \tag{3.14}$$

By left-multiplying (3.14) by $X^{\mathrm{T}}B$, we get

$$X^{\mathrm{T}}B\xi_X = \Omega_{\xi_X} + X^{\mathrm{T}}BM_X^{-1}BXS_{\xi_X}\ .$$

Summing $X^{\mathrm{T}}B\xi_X + \left(X^{\mathrm{T}}B\xi_X\right)^{\mathrm{T}}$, and using the fact that Ω_{ξ_X} is skew-symmetric so it vanishes in the sum, we get that S_{ξ_X} solves the following Sylvester equation ([42, Subsection 2.4.4]):

$$X^{\mathrm{T}}B\xi_X + \left(X^{\mathrm{T}}B\xi_X\right)^{\mathrm{T}} = \left(X^{\mathrm{T}}BM_X^{-1}BX\right)S_{\xi_X} + S_{\xi_X}\left(X^{\mathrm{T}}BM_X^{-1}BX\right)\ . \tag{3.15}$$

Indeed, according to [42, Theorem 2.4.4.1] there is a unique solution to (3.15) for any $X^{\mathrm{T}}B\xi_X + \left(X^{\mathrm{T}}B\xi_X\right)^{\mathrm{T}}$, since $\left(X^{\mathrm{T}}BM_X^{-1}BX\right)$ is positive definite ($X^{\mathrm{T}}BM_X^{-1}BX$ is a Gram matrix of $M_X^{-1/2}BX$, which consists of a product of three matrices, two invertible matrices $M_X^{-1/2}$ and B, and one full-column rank matrix $X \in \mathbf{St}_B(p,d)$) and $-\left(X^{\mathrm{T}}BM_X^{-1}BX\right)$ is negative definite, thus both matrices have no eigenvalues in common. Solving (3.15) costs $O(p^3)$ assuming we already computed $X^{\mathrm{T}}BM_X^{-1}BX$. Furthermore, as expected S_{ξ_X} is symmetric since $S_{\xi_X}^{\mathrm{T}}$ again satisfies (3.15), and the solution to the equation is unique.

After obtaining S_{ξ_X} by solving (3.15), analytical expressions for the orthogonal projections on the normal space and the tangent space are given by (3.12) and (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 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 $S_{\xi_{\Pi_X^\perp(\xi_X)}}$ and S_{ξ_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 M_X, i.e.,

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

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 ξ_X is $O(T_Bp + T_{M^{-1}}p + dp^2)$. Indeed, after obtaining S_{ξ_X} by solving a Sylvester equation which costs $O(p^3)$, we are left with taking product of B and 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 $M_X = B$ for all $X \in \mathbf{St}_B(p,d)$, $\mathbf{St}_B(p,d)$ is isometric to $\mathbf{St}(p,d)$ via the change of variables $\hat{X} = B^{1/2}X$. The orthogonal projections on the normal space (3.12) and on the tangent space (3.13) are reduced to a generalization of the orthogonal projection on the tangent space of the Stiefel manifold [5, Example 3.6.2]:

$$\Pi_X^\perp(\xi_X) = X\mathbf{sym}\left(X^{\mathrm{T}}B\xi_X\right) \tag{3.17}$$

and

$$\Pi_X(\xi_X) = \left(\mathrm{id}_{T_X\mathbb{R}^{d\times p}} - \Pi_X^\perp\right)(\xi_X) = \left(I_d - XX^{\mathrm{T}}B\right)\xi_X + X\mathbf{skew}\left(X^{\mathrm{T}}B\xi_X\right). \tag{3.18}$$

In such case, the cost of computing (in ambient coordinates) $\Pi_X(\xi_X)$ for an arbitrary ξ_X is $O(T_Bp + dp^2)$. 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 [5, Subsection 8.1.3]:

$$\tau_{\eta_X}^{(\mathrm{dep})}\xi_X \quad := \quad \Pi_{R_X(\eta_X)}\left(\xi_X\right), \tag{3.19}$$

where $R_X(\cdot)$ is a retraction mapping of our choice, e.g., (3.3), (3.6) or (3.9).

Let $f : \mathbf{St}_B(p,d) \to \mathbb{R}$ be a 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 [5, Eq. 3.37]: the Riemannian gradient can be computed by computing the Riemannian gradient in $\mathbb{R}^{d\times p}$ of $\bar{f}$, and orthogonally projecting it with respect to the Riemannnian metric to the tangent space of $\mathbf{St}_B(p,d)$ using the orthogonal projection on the tangent space, $\Pi_X(\cdot)$. In short, $\mathbf{grad}f(X) = \Pi_X\left(\mathbf{grad}\bar{f}(X)\right)$. First, we consider $\mathbf{grad}\bar{f}(X)$. Note that it is not the Euclidean gradient $\nabla\bar{f}(X)$, even though $\bar{f}$ is defined on $\mathbb{R}^{d\times p}$. The reason is that $\bar{f}$ is defined on a $\mathbb{R}^{d\times p}$ endowed with a non-standard inner product. According to [5, Eq. 3.31], we have

$$\mathbf{Tr}\left(\mathbf{grad}\bar{f}(X)^{\mathrm{T}}M_X\xi_X\right) = g_X(\mathbf{grad}\bar{f}(X),\xi_X) = \mathrm{D}\bar{f}(X)[\xi_X] = \mathbf{Tr}\left(\nabla\bar{f}(X)^{\mathrm{T}}\xi_X\right)$$

for every $\bar{\xi}_X \in T_X\mathbb{R}^{d\times p}$ (in the above, $\mathrm{D}f(X)$ denotes the (Frechet) differential of f at X), so $\mathbf{grad}\bar{f}(X) = M_X^{-1}\nabla\bar{f}(X)$. Thus, we have

$$\mathbf{grad}f(X) = \Pi_X\left(M_X^{-1}\nabla\bar{f}(X)\right). \tag{3.20}$$

The cost of computing the Riemannian gradient given the Euclidean gradient of $\bar{f}$ is the cost of computing the orthogonal projection on the tangent space, and taking the product of $\nabla\bar{f}(X)$ and 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 examining the condition number of the Hessian at the optimum. However, any expression for 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 $X \in \mathbf{St}_B(p,d)$). 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).

Recall that in [5, Proposition 5.5.6], it is shown that at a critical point $X^\star$, i.e. $\mathbf{grad}f(X^\star) = 0$, the Riemannian Hessian equals to the Riemannian Hessian of a composition of the cost function with a retraction map (known in the literature as the *pullback* function). The pullback function is a function from the tangent space which is a Euclidean space to $\mathbb{R}$, thus its Riemannian Hessian is the Euclidean Hessian. In addition, retraction maps typically do not depend on the choice of the Riemannian metric. Therefore, the Euclidean Hessian of the pullback function only depends on the Riemannian metric at a critical point through the directional derivative of Riemannian gradient of the pullback function on the tangent space at the critical point. Thus, the formula we derive for the Riemannian Hessian in ambient coordinates is valid at a critical point $X^\star$ when using any preconditioning scheme $X \mapsto M_X$ as well if we set $M = M_{X^\star}$. 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 of Riemannian optimization algorithms (e.g., [5, Theorem 4.5.6, Theorem 7.4.11 and Eq. (7.50)]).

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 $X \in \mathbf{St}_B(p,d)$, we can compute the result of applying the Riemannian Hessian to a tangent vector in ambient coordinates via the formula [44]:

$$\mathbf{Hess}f(X)[\eta_X] = \Pi_X(M^{-1}\nabla^2\bar{f}(X)\eta_X) + W_X(\eta_X, \Pi_X^{\perp}(M^{-1}\nabla\bar{f}(X))) \tag{3.21}$$

where $\nabla^2\bar{f}(X)$ is the Euclidean Hessian of $\bar{f}$ and W_X is the Weingarten map on $\mathbf{St}_B(p,d)$. The Weingarten map is an operator that takes as arguments a tangent vector $\eta_X \in T_X\mathbf{St}_B(p,d)$ and a normal vector $U_X \in$

$(T_X\mathbf{St}_B(p,d))^{\perp}$ and returns a tangent vector. An analytic formula for Weingarten map on $\mathbf{St}_B(p,d)$, in ambient coordinates, is

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

The derivation of (3.21) is based on Lemma A.2. The complete derivation of the Riemannian connection, the Weingarten map, and the Riemannian Hessian appears in Appendix A.2. Based on these formulas, we have the following formula for the Riemannian Hessian when $M_X = M$ for all X:

$$\mathbf{Hess}f(X)[\eta_X] = \Pi_X\left(M^{-1}\nabla^2\bar{f}(X)\eta_X - M^{-1}B\eta_X\left(X^{\mathrm{T}}\nabla\bar{f}(X) - X^{\mathrm{T}}M\mathbf{grad}f(X)\right)\right). \tag{3.22}$$

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

Exponential Map. An important metric related retraction map on a Riemannian manifold is the exponential mapping. According to [5, 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 $M_X = M$ for all $X \in \mathbf{St}_B(p,d)$.

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

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

for all t in the domain of $\gamma(t)$, where $\frac{\mathrm{D}^2}{\mathrm{dt}^2}\gamma(t) = \frac{\mathrm{D}}{\mathrm{dt}}\dot{\gamma}$ [5, Section 5.4] .

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

$$\mathrm{Exp}_x : T_x\mathcal{M} \to \mathcal{M} : \mapsto \mathrm{Exp}_x = \gamma(1;x,),$$

is called the exponential map at x [5, Section 5.4].

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

$$\begin{aligned} \frac{\mathrm{D}^2}{\mathrm{dt}^2}\gamma(t) &= 0 \\ \Pi_{\gamma(t)}\left(\frac{\mathrm{d}}{\mathrm{dt}}\left[\frac{\mathrm{d}}{\mathrm{dt}}(\gamma(t))\right]\right) &= 0 \\ \Pi_{\gamma(t)}(\ddot{\gamma}(t)) &= 0 \\ \ddot{\gamma}(t) &= M^{-1}B\gamma(t)S_{\ddot{\gamma}(t)}\ . \end{aligned} \tag{3.23}$$

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

$$\gamma(t)^{\mathrm{T}}B\ddot{\gamma}(t) + \left(\gamma(t)^{\mathrm{T}}B\ddot{\gamma}(t)\right)^{\mathrm{T}} = \left(\gamma(t)^{\mathrm{T}}BM^{-1}B\gamma(t)\right)S_{\ddot{\gamma}(t)} + S_{\ddot{\gamma}(t)}\left(\gamma(t)^{\mathrm{T}}BM^{-1}B\gamma(t)\right).$$

Note that we can replace $\gamma(t)^{\mathrm{T}}B\ddot{\gamma}(t) + \left(\gamma(t)^{\mathrm{T}}B\ddot{\gamma}(t)\right)^{\mathrm{T}}$ by $-2\dot{\gamma}(t)^{\mathbf{T}}B\dot{\gamma}(t)$ since $\gamma(t)^{\mathrm{T}}B\gamma(t) = I_p$ 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 (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 where $M_X = B$ for all $X \in \mathbf{St}_B(p,d)$ such that $\mathbf{St}_B(p,d)$ is isometric to $\mathbf{St}(p,d)$ via the change of variables $\hat{X} = B^{1/2}X$, the equation can be solved analytically in a manner similar to [5, Equation 5.26]. For $M_X = B$, the equation for the geodesic is reduced to

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

We perform a small modification of the calculations given in [4, Subsection 2.2.2] (also developed by Ross Lippert). Denote

$$C \coloneqq \gamma(t)^{\mathrm{T}} B \gamma(t),\ A \coloneqq \gamma(t)^{\mathrm{T}} B \dot{\gamma}(t),\ S \coloneqq \dot{\gamma}(t)^{\mathrm{T}} B \dot{\gamma}(t).$$

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

$$\begin{aligned} \dot{C} &= A + A^{\mathrm{T}} \ , \\ \dot{A} &= S + \gamma(t)^{\mathrm{T}} B \ddot{\gamma}(t) = S - CS \ , \\ \dot{S} &= \ddot{\gamma}(t)^{\mathrm{T}} B \dot{\gamma}(t) + \dot{\gamma}(t)^{\mathrm{T}} B \ddot{\gamma}(t) = -\left[SA + A^{\mathrm{T}} S\right] \ . \end{aligned}$$

Recall that since $\gamma(t) \in \mathbf{St}_B(p,d)$ we get that $C = I_p$. Thus, $\dot{C} = 0_p$ so $A = -A^{\mathrm{T}}$, i.e., A is skew-symmetric. Moreover, $\dot{A} = 0_p$ 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 [45, 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

$$\frac{\mathrm{d}}{\mathrm{dt}}\left[\gamma(t)e^{At},\ \dot{\gamma}(t)e^{At}\right] = \left[\gamma(t)e^{At},\ \dot{\gamma}(t)e^{At}\right] \begin{pmatrix} A & -S(0) \\ I_p & A \end{pmatrix} \ ,$$

to find a closed form for the geodesic curve

$$\gamma(t) = [X,\ \xi] \exp\left(t \begin{pmatrix} A & -S(0) \\ I_p & A \end{pmatrix}\right) \begin{bmatrix} I_p \\ 0_p \end{bmatrix} e^{-At} \ . \tag{3.24}$$

Substituting $t = 1$ into (3.24) gives us the exponential mapping $\mathrm{Exp}_X \xi_X$.

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 $X \in \mathbf{St}_B(p,d)$) 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 compared to the standard metric on $\mathbf{St}_B(p,d)$ (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.

Table 1: Summary of the cost of the Riemnnian components on the generalized Stiefel manifold

Operation	Cost
Retraction maps (Eqs. (3.3), (3.6), (3.9))	$O\left(T_B p + dp^2\right)$
Inverse of the polar-based retraction (Eq. (3.4))	$O\left(T_B p + dp^2\right)$
Inverse of the QR-based retraction (Eq. (3.7))	$O\left(T_B p + dp^2 + p^4\right)$
Vector Transport, associated with retractions (Eqs. (A.8), (A.10), [13, Eq. (16)])	$O\left(T_B p + dp^2\right)$
Inner product on the tangent space (Eq. (3.10))	$O\left(T_M p + dp\right)$
Orthogonal projections on the tangent/normal space, M_X metric (Eqs. (3.13),(3.12))	$O\left(T_B p + T_{M^{-1}} p + dp^2\right)$
Orthogonal projections on the tangent/normal space, B metric (Eqs. (3.18),(3.17))	$O\left(T_B p + dp^2\right)$
Vector Transport, based on the orthogonal projection (Eq. (3.19))	$O\left(T_B p + T_{M^{-1}} p + dp^2\right)$
Riemannian gradient computation (Eq. (3.20))	$O\left(T_B p + T_{M^{-1}} p + dp^2 + T_{\nabla \bar{f}}\right)$
Applying the Riemannian Hessian to a tangent vector (Eq. (3.22))	$O\left(T_B p + T_{M^{-1}} p + T_M p + dp^2 + T_{\nabla \bar{f}} + T_{\nabla^2 \bar{f}}\right)$

3.4 Product Manifold of Generalized Stiefel Manifolds

In some cases it is desirable to solve optimization problems with several sets of variables, in which each set of variables 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 [5, 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 $X_i \in \mathbf{St}_{B_i}(p, d_i)$ for $i = 1, \ldots, k$. The problem can be solved using Riemannian optimization on the product manifold $\mathbf{St}_{B_1}(p, d_1) \times \mathbf{St}_{B_2}(p, d_2) \times \cdots \times \mathbf{St}_{B_k}(p, d_k)$, i.e., $X \in \mathbf{St}_{B_1}(p, d_1) \times \mathbf{St}_{B_2}(p, d_2) \times \cdots \times \mathbf{St}_{B_k}(p, d_k)$. Indeed, for the product manifold, there is a natural way to define the differentiable structure so that manifold topology of $\mathbf{St}_{B_1}(p, d_1) \times \mathbf{St}_{B_2}(p, d_2) \times \cdots \times \mathbf{St}_{B_k}(p, d_k)$ 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 $\mathbf{St}_{B_k}(p, d_k)$ the metric is defined by a smooth mapping $X_i \mapsto M^{(i)}_{X_i}$ such that $M^{(i)}_{X_i}$ is an SPD matrix (i.e., the metric $g^{(i)}$ on $\mathbf{St}_{B_i}(p, d_i)$ is defined in ambient coordinates by $g^{(i)}_X(\eta_X, \xi_X) = \mathbf{Tr}\left(\eta_X^{\mathrm{T}} M^{(i)}_{X_i} \xi_X\right)$). The product manifold $\mathbf{St}_{B_1}(p, d_1) \times \mathbf{St}_{B_2}(p, d_2) \times \cdots \times \mathbf{St}_{B_k}(p, d_k)$ 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 $\mathbf{St}_{B_1}(p, d_1) \times \mathbf{St}_{B_2}(p, d_2) \times \cdots \times \mathbf{St}_{B_k}(p, d_k)$ can be viewed as a Riemannian embedded submanifold of $\mathbb{R}^{d \times p}$ endowed with the metric defined by the $d \times d$ matrix $M_X \coloneqq \mathbf{blkdiag}\left(M^{(1)}_{X_1}, M^{(2)}_{X_2}, ..., M^{(k)}_{X_k}\right)$, 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 $\mathbf{St}_{B_1}(p, d_1) \times \mathbf{St}_{B_2}(p, d_2) \times \cdots \times \mathbf{St}_{B_k}(p, d_k)$ 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 premultiplying by M_X^{-1}, i.e., $\mathbf{grad}f(X) = \Pi_X\left(M_X^{-1}\nabla\bar{f}(X)\right)$ for $X \in \mathbf{St}_{B_1}(p,d_1)\times\mathbf{St}_{B_2}(p,d_2)\times\cdots\times\mathbf{St}_{B_k}(p,d_k)$, 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. 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

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

where $W_{X_i}(\xi_{X_i}, U_{X_i})$ is the Weingarten map on $\mathbf{St}_{B_i}(p,d_i)$. 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, (3.21), 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 the manifolds in the product. One can also work on the product manifold $\mathbf{St}_{B_1}(p_1,d_1)\times\mathbf{St}_{B_2}(p_2,d_2)\times\cdots\times\mathbf{St}_{B_k}(p_k,d_k)$ 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 [5, Theorem 4.5.6, Theorem 7.4.11 and Eq. (7.50)]. 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., [46]). 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 linear convergence is observed (see experiments in Subsection 4.2) as guaranteed by [5, 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 [5, Propositin 5.5.3]). Thus, its condition number at the optimum, $x^\star$, can be found using the ratio between the maximal and minimal value of the Rayleigh quotient

$$q(\xi_{x^\star}) \coloneqq \frac{g_{x^\star}(\xi_{x^\star}, \mathbf{Hess}f(x^\star)[\xi_{x^\star}])}{g_{x^\star}(\xi_{x^\star}, \xi_{x^\star})}.$$

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

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

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

$$\begin{aligned} q(\xi_{x^\star}) &= \frac{\xi_{x^\star}^{\mathrm{T}} M_{x^\star} \Pi_{x^\star} \left(M_{x^\star}^{-1} \left[\nabla^2 \bar{f}(x^\star) - \left((x^\star)^{\mathrm{T}} \nabla \bar{f}(x^\star) \right) B \right] \xi_{x^\star} \right)}{\xi_{x^\star}^{\mathrm{T}} M_{x^\star} \xi_{x^\star}} \\ &= \frac{\left(\Pi_{x^\star}(\xi_{x^\star}) \right)^{\mathrm{T}} \left[\nabla^2 \bar{f}(x^\star) - \left((x^\star)^{\mathrm{T}} \nabla \bar{f}(x^\star) \right) B \right] \xi_{x^\star}}{\xi_{x^\star}^{\mathrm{T}} M_{x^\star} \xi_{x^\star}} \\ &= \frac{\xi_{x^\star}^{\mathrm{T}} \left[\nabla^2 \bar{f}(x^\star) - \left((x^\star)^{\mathrm{T}} \nabla \bar{f}(x^\star) \right) B \right] \xi_{x^\star}}{\xi_{x^\star}^{\mathrm{T}} M_{x^\star} \xi_{x^\star}} . \end{aligned}$$

This is the Rayleigh quotient of the matrix pencil

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

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

$$\left(\Pi_{x^\star} \left(\nabla^2 \bar{f}(x^\star) - \left((x^\star)^{\mathrm{T}} \nabla \bar{f}(x^\star) \right) B \right) \Pi_{x^\star}, \Pi_{x^\star} M_{x^\star} \Pi_{x^\star} \right) . \tag{3.25}$$

Therefore, choosing a preconditioning scheme $x \mapsto M_x$ such that M_x is SPD for any $x \in \mathbb{S}^B$ and

$$M_{x^\star} \approx \nabla^2 \bar{f}(x^\star) - \left((x^\star)^{\mathrm{T}} \nabla \bar{f}(x^\star) \right) B \tag{3.26}$$

will precondition the Riemannian Hessian at the optimum. One such example can be found in [32]. In addition, the preconditioners proposed in [8], which are inspired by the Lagrangian, can be viewed in such manner, thus, approximating the Riemannian Newton method. 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 $x \in \mathbb{S}^B$. If $\nabla^2 \bar{f}(x^\star)$ is well conditioned, it is often the case that the pencil (3.25) 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 $M_{x^\star} \approx B$, the condition number grows by at most $\kappa(B, M_{x^\star})$, so if that quantity is small (i.e., $M_{x^\star}$ 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 4.1. Linear Objective. Consider the following problem

$$\max_{x \in \mathbb{R}^d} b^{\mathrm{T}} x \ \text{ s.t. } \ x^{\mathrm{T}} B x = 1$$

for some vector $0 \neq b \in \mathbb{R}^d$, where $B \in \mathbb{R}^{d \times d}$. It is easy to show that the solution is $x^\star = B^{-1}b / \|B^{-1}b\|_B$. 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 $\mathbb{S}^B$. Let the inner product on each tangent space (the Riemannian metric) be endowed from the ambient space $\mathbb{R}^d$. Using $\mathbb{S}^B$ with a metric selection $g_x(,\eta_x) = M_x\eta_x$ (in ambient coordinates), where $x \mapsto M_x \in \mathbb{R}^{d\times d}$ is a smooth mapping that maps $x \in \mathbb{S}^B$ to an SPD matrix M_x, the Riemannian gradient is

$$\mathbf{grad}f(x) = (I_n - (x^{\mathrm{T}}BM_x^{-1}Bx)^{-1}M_x^{-1}Bxx^{\mathrm{T}}B)M_x^{-1}b$$

since the Euclidean gradient is simply b, independent of x. Thus, using Riemannian gradient ascent on $\mathbb{S}^B$ with the polar based retraction, (3.3), we get the iteration

$$\begin{aligned} y_{k+1} &= x_k + \alpha_k\left(M_{x_k}^{-1}b - \frac{x_k^{\mathrm{T}}BM_{x_k}^{-1}b}{x_k^{\mathrm{T}}BM_{x_k}^{-1}Bx_k}M_{x_k}^{-1}Bx_k\right) \\ x_{k+1} &= \frac{y_{k+1}}{\|y_{k+1}\|_B}. \end{aligned}$$

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 $x \in \mathbb{S}^B$, and take step size $\alpha_0 = 1/x_k^{\mathrm{T}}b$, then the iterations reduce 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 $x^\star$ is well conditioned. Indeed, we have

$$\mathbf{Hess}f(x^\star) = -\Pi_{x^\star}\left(\left(x^{\star\mathrm{T}}b\right)I_d\right),$$

and its corresponding Rayleigh quotient is

$$q(\xi_{x^\star}) = \frac{\xi_{x^\star}^{\mathrm{T}}B\left[-\Pi_{x^\star}\left(\left(x^{\star\mathrm{T}}b\right)I_d\right)\right]\xi_{x^\star}}{\xi_{x^\star}^{\mathrm{T}}B\xi_{x^\star}} = -\left(x^{\star\mathrm{T}}b\right) = -\|B^{-1}b\|_B\ ,$$

which is constant so the condition number equals 1. Note that the metric selection $M_x = B$ also satisfies (3.26).

Example 4.2. Inverse Power Iteration. Consider the following problem

$$\max_{x\in\mathbb{R}^d}\frac{1}{2}x^{\mathrm{T}}x \text{ s.t. } x^{\mathrm{T}}Bx = 1$$

where $B \in \mathbb{R}^{d\times d}$ is an SPD matrix. The solution is an eigenvector corresponding the smallest eigenvalue of B, $\lambda_d(B)$, (which is also the eigenvector corresponding to the maximal eigenvalue of B^{-1}), since this problem is equivalent to maximizing the Rayleigh quotient $x^{\mathrm{T}}x/x^{\mathrm{T}}Bx$. Note that this problem is constrained on the ellipsoid manifold $\mathbb{S}^B$. Using $\mathbb{S}^B$ with metric selection $g_x(,\eta_x) = M_x\eta_x$ (in ambient coordinates), where $M_x \in \mathbb{R}^{d\times d}$ is an SPD matrix for any $x \in \mathbb{S}^B$, the Riemannian gradient is

$$\mathbf{grad}f(x) = (I_d - (x^{\mathrm{T}}BM_x^{-1}Bx)^{-1}M_x^{-1}Bxx^{\mathrm{T}}B)M_x^{-1}x$$

since the Euclidean gradient is x. Thus, using Riemannian gradient ascent on $\mathbb{S}^B$ with the polar based retraction, (3.3), we get the iteration

$$\begin{aligned} y_{k+1} &= x_k + \alpha_k\left(M_x^{-1}x_k - \frac{x_k^{\mathrm{T}}BM_x^{-1}x_k}{x_k^{\mathrm{T}}BM_x^{-1}Bx_k}M_x^{-1}Bx_k\right) \\ x_{k+1} &= \frac{y_{k+1}}{\|y_{k+1}\|_B}. \end{aligned}$$

If we impose the metric $M_x = B$ for all $x \in \mathbb{S}^B$, and take step sizes $\alpha_k = (x_k^{\mathrm{T}}x_k)^{-1}$, 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 $x^\star$ (i.e $(x^\star)^{\mathrm{T}} x^\star = 1/\lambda_{\min}(B) = 1/\lambda_d(B)$):

$$\mathbf{Hess} f(x^\star) = \Pi_{x^\star}\left(B^{-1}\left[I_d - \left((x^\star)^{\mathrm{T}} x^\star\right) B\right]\right) = \Pi_{x^\star}\left(B^{-1}\left[I_d - (1/\lambda_d(B))\, B\right]\right).$$

The corresponding Rayleigh quotient is reduced to the following form using similar reasoning as in Subsection 3.5:

$$q(\xi_{x^\star}) = \frac{\xi_{x^\star}^{\mathrm{T}} B\left[\Pi_{x^\star}\left(B^{-1}\left[I_d - (1/\lambda_d(B)))\, B\right]\right)\right]\xi_{x^\star}}{\xi_{x^\star}^{\mathrm{T}} B \xi_{x^\star}} = \frac{\xi_{x^\star}^{\mathrm{T}}\left[I_d - (1/\lambda_d(B))\, B\right]\xi_{x^\star}}{\xi_{x^\star}^{\mathrm{T}} B \xi_{x^\star}}.$$

Thus, the eigenvalues of the Riemannian Hessian at $x^\star$ correspond to the generalized eigenvalues of the matrix pencil $(I_d - (1/\lambda_d(B))\, B, B)$ on $T_{x^\star}\mathbb{S}^B$, i.e., the eigenvalues of B^{-1} deflated by $-1/\lambda_d(B)$ on $T_{x^\star}\mathbb{S}^B$. Moreover, since $\xi_{x^\star} \in T_{x^\star}\mathbb{S}^B$, we have $\xi_{x^\star}^{\mathrm{T}} B x^\star = 0$, thus $\xi_{x^\star}$ is constrained not to correspond to $1/\lambda_d(B)$. Assume that $\lambda_{d-1}(B) > \lambda_d(B)$, then the condition number is bounded by

$$\frac{1/\lambda_d(B) - 1/\lambda_1(B)}{1/\lambda_d(B) - 1/\lambda_{d-1}(B)},$$

which for $\lambda_d(B)$ that is close to 0, and $\lambda_{d-1}(B) \gg 0$ is close to 1.

Note that if we try to impose the metric $M_x = -(I_d - (1/\lambda_d(B))\, B)$ for all $x \in \mathbb{S}^B$ (following (3.26)), we have that M_x is singular since it has a zero eigenvalue (corresponding to the eigenvector $x^\star$), thus it cannot be a Riemannian metric inherited from the ambient space $\mathbb{R}^d$.

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 [1], is a well-established method in statistical learning with numerous applications (e.g., [47, 48, 49, 50, 51, 52]). 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 of the two matrices. Here, we consider a regularized version of CCA defined below:

Definition 4.3. 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 parameter. Let

$$q = \max\left(\mathbf{rank}\left(X^{\mathrm{T}}X + \lambda_x I_{d_x}\right), \mathbf{rank}\left(Y^{\mathrm{T}}Y + \lambda_y I_{d_y}\right)\right).$$

The (λ_x, λ_y) canonical correlations $\sigma_1 \geq \cdots \geq \sigma_q$ and the (λ_x, λ_y) canonical weights $u_1, \ldots, u_q \in \mathbb{R}^{d_x}$, $v_1, \ldots, v_q \in \mathbb{R}^{d_y}$, are the ones that maximize

$$\mathbf{Tr}\left(U^{\mathrm{T}} X^{\mathrm{T}} Y V\right)$$

subject to

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

where $U^{\mathrm{T}} X^{\mathrm{T}} Y V = \mathbf{diag}(\sigma_1, \ldots, \sigma_q)$, $U = \begin{bmatrix} u_1 & \ldots & u_q \end{bmatrix} \in \mathbb{R}^{d_x \times q}$ and $V = \begin{bmatrix} v_1 & \ldots & v_q \end{bmatrix} \in \mathbb{R}^{d_y \times q}$.

In this paper, we focus on finding the top correlation, i.e., finding σ_1, u_1 and v_1. It is useful to introduce the following notations:

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

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

$$\max u^{\mathrm{T}} \Sigma_{xy} v \ \text{ s.t. } \ u \in \mathbb{S}^{\Sigma_{xx}}, v \in \mathbb{S}^{\Sigma_{yy}} \tag{4.1}$$

It is well known ([3]) that the optimal solution of Problem (4.1) is (up to the sign of the vectors)

$$u_1 \coloneqq \Sigma_{xx}^{-1/2}\phi \quad v_1 \coloneqq \Sigma_{yy}^{-1/2}\psi \tag{4.2}$$

where $\phi \in \mathbb{R}^{d_x}$ and $\psi \in \mathbb{R}^{d_y}$ are the left and right unit-length singular vector corresponding to the largest singular value σ_1 of the matrix

$$R \coloneqq \Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}\,. \tag{4.3}$$

In order to conveniently use the Riemannian optimization framework, we also denote $d = d_x + d_y$, and $z = [u^{\mathrm{T}}, v^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^d$ 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 $z \in \mathbb{S}_{xy} \coloneqq \mathbb{S}^{\Sigma_{xx}} \times \mathbb{S}^{\Sigma_{yy}}$. The objective function to be minimized is then

$$f(z) = -\frac{1}{2} z^{\mathrm{T}} \begin{bmatrix} 0 & \Sigma_{xy} \\ \Sigma_{xy}^{\mathrm{T}} & 0 \end{bmatrix} z\,. \tag{4.4}$$

We endow the manifold $\mathbb{S}^{\Sigma_{xx}}$ and $\mathbb{S}^{\Sigma_{yy}}$ with a metric defined by two preconditioning schemes $u \mapsto M_u^{(xx)}$ and $v \mapsto M_v^{(yy)}$. The metric on the product manifold $\mathbb{S}_{xy}$ is defined by $z \mapsto M_z = \mathbf{blkdiag}\left(M_u^{(xx)}, M_v^{(yy)}\right)$ 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 $M_z \coloneqq M = \mathbf{blkdiag}\left(M^{(xx)}, M^{(yy)}\right)$) are given by:

$$\mathbf{grad} f(z) = \Pi_z\left(M_z^{-1}\nabla \bar{f}(z)\right) = -\begin{bmatrix} \Pi_u\left(\left(M_u^{(xx)}\right)^{-1}\Sigma_{xy}v\right) \\ \Pi_v\left(\left(M_v^{(yy)}\right)^{-1}\Sigma_{xy}^{\mathrm{T}}u\right) \end{bmatrix},$$

$$\mathbf{Hess} f(z)[\eta_z] = \Pi_z\left(M_z^{-1}\begin{bmatrix} \left(u^{\mathrm{T}}M^{(xx)}\Pi_u^{\perp}\left(\left(M^{(xx)}\right)^{-1}\Sigma_{xy}v\right)\right)\cdot\Sigma_{xx} & -\Sigma_{xy} \\ -\Sigma_{xy}^{\mathrm{T}} & \left(v^{\mathrm{T}}M^{(yy)}\Pi_v^{\perp}\left(\left(M^{(yy)}\right)^{-1}\Sigma_{xy}^{\mathrm{T}}u\right)\right)\cdot\Sigma_{yy} \end{bmatrix}\eta_z\right).$$

Along with formulas for the retraction and vector transport (see Subsection 3.1), various Riemannian optimization algorithms can be applied to solve Problem (4.1).

As expected, at the optimal solution $z^\star = [u_1^{\mathrm{T}}, v_1^{\mathrm{T}}]^{\mathrm{T}}$ (see (4.2)) the Riemannian gradient vanishes: $\mathbf{grad} f(z^\star) = 0$. Moreover, the Riemannian Hessian at the optimum becomes

$$\mathbf{Hess} f(z^\star) = \Pi_{z^\star}\left(M_{z^\star}^{-1}\begin{bmatrix} \sigma_1\cdot\Sigma_{xx} & -\Sigma_{xy} \\ -\Sigma_{xy}^{\mathrm{T}} & \sigma_1\cdot\Sigma_{yy} \end{bmatrix}\right)\,. \tag{4.5}$$

Next, we demonstrate the effect of preconditioning on the condition number of the Riemannian Hessian at $z^\star$. 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 $M_{z^\star} = \Sigma \coloneqq \mathbf{blkdiag}(\Sigma_{xx}, \Sigma_{yy})$, 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 $M_{z^\star} \approx \Sigma$ (see for example Fig. 1) the condition number bound grows by at most a small factor: $\kappa(B, M_{z^\star})$.

Lemma 4.4. *Assuming $\sigma_1 - \sigma_2 > 0$ and that Σ is an SPD matrix, if $\mathbb{S}_{xy}$ is equipped with a metric defined by a smooth preconditioning scheme $z \mapsto M_z$ such that $M_{z^\star} = \Sigma$, then the condition number of Riemannian Hessian on $\mathbb{S}_{xy}$ of (4.4) at $z^\star$ is equal to $\frac{\sigma_1+\sigma_2}{\sigma_1-\sigma_2}$. Additionally, if $M_{z^\star} \approx \Sigma$ then the condition number is bounded by $\frac{\sigma_1+\sigma_2}{\sigma_1-\sigma_2}\cdot\kappa(B, M_{z^\star})$.*

Proof. In order to bound the condition number of Riemannian Hessian on $\mathbb{S}_{xy}$ of (4.4) at $z^\star$ we use the Courant-Fischer Theorem for the compact self-adjoint linear operator $\mathbf{Hess}f(z^\star)[\cdot] : T_{z^\star}\mathbb{S}_{xy} \to T_{z^\star}\mathbb{S}_{xy}$ over the finite dimensional vector space $T_Z\mathbb{S}_{xy}$:

$$\lambda_k(\mathbf{Hess}f(z^\star)) = \min_{U,\dim(U)=k-1} \max_{0\neq\xi_{z^\star}\in U^\perp} q(\xi_{z^\star}),$$

$$\lambda_k(\mathbf{Hess}f(z^\star)) = \max_{U,\dim(U)=k} \min_{0\neq\xi_{z^\star}\in U} q(\xi_{z^\star}),$$

where

$$q(\xi_{z^\star}) := \frac{g_{z^\star}(\xi_{z^\star}, \mathbf{Hess}f(z^\star)[\xi_{z^\star}])}{g_{z^\star}(\xi_{z^\star}, \xi_{z^\star})},$$

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

$$\lambda_{\max}(\mathbf{Hess}f(z^\star)) = \max_{0\neq\xi_{z^\star}\in T_{z^\star}\mathbb{S}_{xy}} q(\xi_{z^\star}),$$

$$\lambda_{\min}(\mathbf{Hess}f(z^\star)) = \min_{0\neq\xi_{z^\star}\in T_{z^\star}\mathbb{S}_{xy}} q(\xi_{z^\star}),$$

and the condition number of the Riemannian Hessian at $z^\star$ is the ratio of these two eigenvalues.

$$\kappa(\mathbf{Hess}f(z^\star)) = \frac{\lambda_{\max}(\mathbf{Hess}f(z^\star))}{\lambda_{\min}(\mathbf{Hess}f(z^\star))} .$$

We begin by simplifying the quotient $q(\xi_{z^\star})$. At the optimum, $z^\star$, we have $f(z^\star) = -u_1^{\mathrm{T}}\Sigma_{xy}v_1 = -v_1^{\mathrm{T}}\Sigma_{xy}^{\mathrm{T}}u_1 = -\sigma_1$. The formula for the Riemannian Hessian, $\mathbf{Hess}f(z^\star)$, is given by (4.5). Using the following notation for the Euclidean Hessian of

$$\nabla^2\bar{f}(z^\star) := \begin{bmatrix} 0 & -\Sigma_{xy} \\ -\Sigma_{xy}^{\mathrm{T}} & 0 \end{bmatrix}, \tag{4.6}$$

and Σ we can compactly write (4.5):

$$\mathbf{Hess}f(z^\star) = \Pi_{z^\star}\left(M_{z^\star}^{-1}\left(\nabla^2\bar{f}(z^\star) + \sigma_1\Sigma\right)\right).$$

Next, as in Subsection 3.5, recall that $\Pi_{z^\star}$ is self-adjoint with respect to the Riemannian metric, (3.16), and that for any $\xi_{z^\star} \in T_{z^\star}\mathbb{S}_{xy}$ we have $\Pi_{z^\star}(\xi_{z^\star}) = \xi_{z^\star}$, we get:

$$q(\xi_{z^\star}) = \frac{\xi_{z^\star}^{\mathrm{T}}\left(\nabla^2\bar{f}(z^\star) + \sigma_1\cdot\Sigma\right)\xi_{z^\star}}{\xi_{z^\star}^{\mathrm{T}}M_{z^\star}\xi_{z^\star}} = \frac{\xi_{z^\star}^{\mathrm{T}}\left(\nabla^2\bar{f}(z^\star) + \sigma_1\cdot\Sigma\right)\xi_{z^\star}}{\xi_{z^\star}^{\mathrm{T}}\Sigma\xi_{z^\star}} \cdot \frac{\xi_{z^\star}^{\mathrm{T}}\Sigma\xi_{z^\star}}{\xi_{z^\star}^{\mathrm{T}}M_{z^\star}\xi_{z^\star}} ,$$

where we use the fact that Σ is not singular. Note that the quotient

$$\frac{\xi_{z^\star}^{\mathrm{T}}\left(\nabla^2\bar{f}(z^\star) + \sigma_1\cdot\Sigma\right)\xi_{z^\star}}{\xi_{z^\star}^{\mathrm{T}}\Sigma\xi_{z^\star}},$$

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

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

$$q(\xi_{z^\star}) = \frac{\tilde{\xi}_{z^\star}^{\mathrm{T}}\left(\Sigma^{-1/2}\nabla^2\bar{f}(z^\star)\Sigma^{-1/2} + \sigma_1\cdot I_d\right)\tilde{\xi}_{z^\star}}{\tilde{\xi}_{z^\star}^{\mathrm{T}}\tilde{\xi}_{z^\star}} := \tilde{q}(\tilde{\xi}_{z^\star}) .$$

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

To find the eigenvalues of the matrix $\Sigma^{-1/2}\nabla^2\bar{f}(z^\star)\Sigma^{-1/2}+\sigma_1\cdot I_d$ in the space $\Sigma^{1/2}T_{z^\star}\mathbb{S}_{xy}$, we first note that all the eigenvalues of $\Sigma^{-1/2}\nabla^2\bar{f}(z^\star)\Sigma^{-1/2}$ are $-\sigma_1 < -\sigma_2 \le ... \le -\sigma_q \le 0 \le ... \le 0 \le \sigma_q \le ... \le \sigma_2 < \sigma_1$ (see [53]). So, all the eigenvalue of $\Sigma^{-1/2}\nabla^2\bar{f}(z^\star)\Sigma^{-1/2}+\sigma_1\cdot I_d$ are $0 < \sigma_1-\sigma_2 \le \cdots \le \sigma_1-\sigma_q \le \sigma_1 \le ... \le \sigma_q+\sigma_1 \le ... \le \sigma_2+\sigma_1 < 2\sigma_1$. Next, note that the eigenspaces of $\Sigma^{-1/2}\nabla^2\bar{f}(z^\star)\Sigma^{-1/2}+\sigma_1\cdot I_d$ corresponding to the eigenvalues 0 and $2\sigma_1$ is exactly the two dimensional space $(\Sigma^{\frac{1}{2}}T_{z^\star}\mathbb{S}_{xy})^\perp$. Indeed, according to (3.11) and Subsection 3.4:

$$(\Sigma^{1/2}T_{z^\star}\mathbb{S}_{xy})^\perp = \mathbf{span}\left\{\Sigma^{1/2}\begin{bmatrix} u_1 \\ v_1 \end{bmatrix}, \Sigma^{1/2}\begin{bmatrix} u_1 \\ -v_1 \end{bmatrix}\right\} ,$$

where using (4.2)

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

Recall that the normal space $(T_{z^\star}\mathbb{S}_{xy})^\perp$ is the Cartesian product of the normal spaces $(T_{u_1}\mathbb{S}^{\Sigma_{xx}})^\perp$ and $(T_{v_1}\mathbb{S}^{\Sigma_{yy}})^\perp$ which are spanned by u_1 and v_1 correspondingly when $M_{z^\star} = \Sigma$. Thus, the Cartesian product $(T_{z^\star}\mathbb{S}_{xy})^\perp$ can be spanned by $[u_1^{\mathrm{T}}, v_1^{\mathrm{T}}]^{\mathrm{T}}$ and $[u_1^{\mathrm{T}}, -v_1^{\mathrm{T}}]^{\mathrm{T}}$.

Then, using (4.3) and (4.6) we have

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

where the last equality follows from the fact that $\begin{bmatrix} & R \\ R^{\mathrm{T}} & \end{bmatrix}$ is the augmented matrix associated with R, so $\begin{bmatrix} \phi \\ \psi \end{bmatrix}$, which has the dominant left and right singular vectors stacked, is the eigenvalue corresponding to the largest eigenvalue σ_1 of the augmented matrix. Similarly, since the vector $\begin{bmatrix} \phi \\ -\psi \end{bmatrix}$ is the eigenvector corresponding to the smallest eigenvalue $-\sigma_1$ of the augmented matrix, then

$$(\Sigma^{-1/2}\nabla^2\bar{f}(z^\star)\Sigma^{-1/2}+\sigma_1\cdot I_d)\Sigma^{1/2}\begin{bmatrix} u_1 \\ -v_1 \end{bmatrix} = \left(\begin{bmatrix} & -R \\ -R^{\mathrm{T}} & \end{bmatrix} + \sigma_1 I_d\right)\begin{bmatrix} \phi \\ -\psi \end{bmatrix} = 2\sigma_1\begin{bmatrix} \phi \\ -\psi \end{bmatrix} .$$

Finally, the minimal and the maximal eigenvalues of the matrix $\Sigma^{-1/2}\nabla^2\bar{f}(z^\star)\Sigma^{-1/2}+\sigma_1\cdot I_d$ in the space of vectors ξ such that $\Sigma^{-1/2}\xi \in T_{z^\star}\mathbb{S}_{xy}$ are $\sigma_1-\sigma_2$ and $\sigma_1+\sigma_2$ correspondingly. Thus,

$$\lambda_{\max}(\mathbf{Hess}f(z^\star)) = \max_{0\neq\xi_{z^\star}\in T_{z^\star}\mathbb{S}_{xy}} q(\xi_{z^\star}) = \sigma_1+\sigma_2 > 0 ,$$

and,

$$\lambda_{\min}(\mathbf{Hess}f(z^\star)) = \min_{0\neq\xi_{z^\star}\in T_{z^\star}\mathbb{S}_{xy}} q(\xi_{z^\star}) = \sigma_1-\sigma_2 > 0 ,$$

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

If $M_{z^\star} \approx \Sigma$, we can bound the smallest and largest eigenvalues of the Riemannian Hessian at $z^\star$ by

$$\begin{aligned} \lambda_{\min}(\mathbf{Hess}f(z^\star)) &\ge \min_{0\neq\eta_{z^\star}\in T_{z^\star}\mathbb{S}_{xy}} \frac{\eta_{z^\star}^{\mathrm{T}}\left(\nabla^2\bar{f}(z^\star)+\sigma_1\cdot\Sigma\right)\eta_{z^\star}}{\eta_{z^\star}^{\mathrm{T}}\Sigma\eta_{z^\star}} \cdot \min_{\eta_{z^\star}\neq 0} \frac{\eta_{z^\star}^{\mathrm{T}}\Sigma\eta_{z^\star}}{\eta_{z^\star}^{\mathrm{T}}M_{z^\star}\eta_{z^\star}} \\ &= \lambda_{\min}(\Sigma, M_{z^\star})\cdot(\sigma_1-\sigma_2) , \end{aligned}$$

and

$$\begin{aligned}\lambda_{\max}(\mathbf{Hess}f(z^\star)) &\leq \max_{0\neq\eta_{z^\star}\in T_{z^\star}\mathbb{S}_{xy}} \frac{\eta_{z^\star}^{\mathrm{T}}\left(\nabla^2\bar{f}(z^\star)+\sigma_1\cdot\Sigma\right)\eta_{z^\star}}{\eta_{z^\star}^{\mathrm{T}}\Sigma\eta_{z^\star}}\cdot\max_{\eta_{z^\star}\neq 0}\frac{\eta_{z^\star}^{\mathrm{T}}\Sigma\eta_{z^\star}}{\eta_{z^\star}^{\mathrm{T}}M_{z^\star}\eta_{z^\star}} \\ &= \lambda_{\max}(\Sigma, M_{z^\star})\cdot(\sigma_1+\sigma_2)\ .\end{aligned}$$

Finally, we get

$$\kappa(\mathbf{Hess}f(z^\star)) = \frac{\lambda_{\max}(\mathbf{Hess}f(z^\star))}{\lambda_{\min}(\mathbf{Hess}f(z^\star))} \leq \frac{\sigma_1+\sigma_2}{\sigma_1-\sigma_2}\cdot\kappa\left(B, M_{z^\star}\right)\ .$$

□

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 $z \in \mathbb{S}_{xy}$: the trivial choice of a unit matrix $M = I_d$, the standard but expensive choice $M = \Sigma$ which achieves the optimal bound according to Lemma 4.4, and four approximations of Σ via the (exact) sketched preconditioning strategy described by Gonen et al. [54], 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 Σ_{xx} and Σ_{yy}. The approximation is done as follows: suppose $A = \hat{X}\hat{X}^{\mathrm{T}} \in \mathbb{R}^{d\times d}$ be some positive semi-definite matrix, and let $\hat{X} = U\Lambda^{1/2}V^{\mathrm{T}}$ be an SVD decomposition of $\hat{X}$ such that $A = U\Lambda U^{\mathrm{T}}$ is an eigendecomposition, with the diagonal entries in Λ sorted in descending order. Given k, let us denote by U_k the first k columns of U, Λ_k denote the leading $k \times k$ minor of Λ, and λ_k the k-th largest eigenvalue of A. The k-dominant subspace preconditioner of $A + \lambda I_d$ is $U_k(\Lambda_k - \lambda_k I)U_k^{\mathrm{T}} + (\lambda_k + \lambda)I_d$. The dominant subspace can be found using a sparse SVD solver (we use MATLAB's svds). Moreover, its inverse can be easily computed using the formula

$$U_k(\Lambda_k + \lambda I)^{-1}U_k^{\mathrm{T}} + \frac{1}{\lambda_k+\lambda}(I_d - U_kU_k^{\mathrm{T}}).$$

The experiments are performed with the MEDIANILL[1] dataset where the dimensions are $n = 43907$, $d_x = 120$, and $d_y = 101$. The implementation uses Manopt which is a MATLAB library that implements some Riemannian optimization algorithms [7]. In Fig. 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 graphs in Fig. 1 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

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

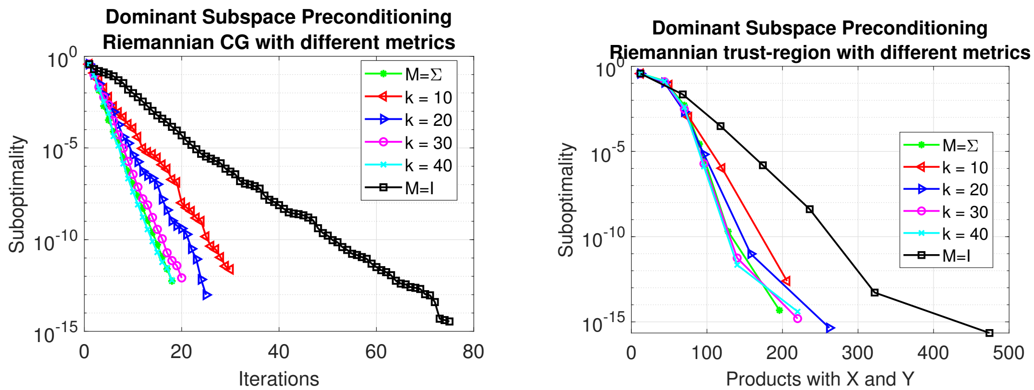

Figure 1: Results for CCA with Riemannian conjugate-gradient (left - suboptimality vs. #iterations) and Riemannian trust-region (right - suboptimality vs. products with the data matrices) with various choices of metrics for $p = 1$. The number of leading eigenvalues used to form the Dominant Subspace Preconditioner is denoted by k.

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.

Acknowledgement. The authors thank Bart Vandereycken for useful discussions. The authors would also like to thank the referees for their valuable comments. This research was supported by the Israel Science Foundation (grant no. 1272/17).

References

[1] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3/4) (1936) 321–377.

[2] R. A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.

[3] A. Björck, G. H. Golub, Numerical methods for computing angles between linear subspaces, Math. Comput. 27 (123) (1973) 579–594.

[4] A. Edelman, T. Arias, S. Smith, The Geometry of Algorithms with Orthogonality Constraints, SIAM J. Matrix Anal. Appl. 20 (2) (1998) 303–353. arXiv:https://doi.org/10.1137/S0895479895290954, doi:10.1137/S0895479895290954.
URL https://doi.org/10.1137/S0895479895290954

[5] P. A. Absil, R. Mahony, R. Sepulchre, Optimization Algorithms on Matrix Manifolds, Princeton University Press, 2009.

[6] N. Boumal, An introduction to optimization on smooth manifolds, Available online (May 2020).
URL http://www.nicolasboumal.net/book

[7] N. Boumal, B. Mishra, P. A. Absil, R. Sepulchre, Manopt, a Matlab Toolbox for Optimization on Manifolds, J Mach Learn Res. 15 (2014) 1455–1459.
URL `http://www.manopt.org`

[8] B. Mishra, R. Sepulchre, Riemannian Preconditioning, SIAM J. Optim. 26 (1) (2016) 635–660. `arXiv:http://dx.doi.org/10.1137/140970860`, `doi:10.1137/140970860`.
URL `http://dx.doi.org/10.1137/140970860`

[9] D. G. Luenberger, The gradient projection method along geodesics, Manage Sci. 18 (11) (1972) 620–631.

[10] D. Gabay, Minimizing a differentiable function over a differential manifold, J Optim Theory Appl. 37 (2) (1982) 177–219.

[11] S. T. Smith, Optimization Techniques on Riemannian Manifolds, Fields institute communications 3 (3) (1994) 113–135.

[12] Z. Wen, W. Yin, A feasible method for optimization with orthogonality constraints, Math Program 142 (1) (2013) 397–434.

[13] X. Zhu, A Riemannian conjugate gradient method for optimization on the Stiefel manifold, Comput Optim Appl. 67 (1) (2017) 73–110.

[14] J. Li, F. Li, S. Todorovic, Efficient Riemannian Optimization on the Stiefel Manifold via the Cayley Transform, in: International Conference on Learning Representations, 2020.
URL `https://openreview.net/forum?id=HJxV-ANKDH`

[15] H. Sato, K. Aihara, Cholesky QR-based retraction on the generalized Stiefel manifold, Comput Optim Appl. 72 (2) (2019) 293–308.

[16] T. Kaneko, S. Fiori, T. Tanaka, Empirical arithmetic averaging over the compact Stiefel manifold, IEEE Trans. Signal 61 (4) (2012) 883–894.

[17] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (1) (2009) 183–202.

[18] S. Chen, S. Ma, A. Man-Cho So, T. Zhang, Proximal Gradient Method for Nonsmooth Optimization over the Stiefel Manifold, SIAM J. Optim. 30 (1) (2020) 210–239. `arXiv:https://doi.org/10.1137/18M122457X`, `doi:10.1137/18M122457X`.
URL `https://doi.org/10.1137/18M122457X`

[19] S. Chen, S. Ma, L. Xue, H. Zou, An Alternating Manifold Proximal Gradient Method for Sparse Principal Component Analysis and Sparse Canonical Correlation Analysis, INFORMS J Optim. 2 (3) (2020) 192–208. `arXiv:https://doi.org/10.1287/ijoo.2019.0032`, `doi:10.1287/ijoo.2019.0032`.
URL `https://doi.org/10.1287/ijoo.2019.0032`

[20] W. Huang, K. Wei, An Extension of FISTA to Riemannian Optimization for Sparse PCA, arXiv preprint arXiv:1909.05485 (2019).

[21] F. Yger, M. Berar, G. Gasso, A. Rakotomamonjy, Adaptive Canonical Correlation Analysis Based on Matrix Manifolds, in: Proceedings of the 29th International Coference on International Conference on Machine Learning, ICML'12, Omnipress, USA, 2012, pp. 299–306.
URL `http://dl.acm.org/citation.cfm?id=3042573.3042615`

[22] A. Pechen, D. Prokhorenko, R. Wu, H. Rabitz, Control landscapes for two-level open quantum systems, J. Phys. A Math. 41 (4) (2008) 045205.

[23] A. Oza, A. Pechen, J. Dominy, V. Beltrani, K. Moore, H. Rabitz, Optimization search effort over the control landscapes for open quantum systems with Kraus-map evolution, J. Phys. A Math. 42 (20) (2009) 205305.

[24] H. Sato, T. Iwai, A complex singular value decomposition algorithm based on the Riemannian Newton method, in: Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on, IEEE, 2013, pp. 2972–2978.

[25] H. Sato, Riemannian conjugate gradient method for complex singular value decomposition problem, in: Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on, IEEE, 2014, pp. 5849–5854.

[26] C. Udriste, Convex functions and optimization methods on Riemannian manifolds, Vol. 297, Springer Science & Business Media, 2013.

[27] T. Ngo, Y. Saad, Scaled gradients on Grassmann manifolds for matrix completion, in: Adv Neural Inf Process Syst., 2012, pp. 1412–1420.

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

[29] Y. Shi, J. Zhang, K. B. Letaief, Low-rank matrix completion for topological interference management by Riemannian pursuit, IEEE Trans. Wirel. 15 (7) (2016) 4703–4717.

[30] T. Zhou, H. Qian, Z. Shen, C. Zhang, C. Xu, Tensor Completion with Side Information: A Riemannian Manifold Approach, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 3539–3545. `doi:10.24963/ijcai.2017/495`.
URL `https://doi.org/10.24963/ijcai.2017/495`

[31] B. Vandereycken, S. Vandewalle, A Riemannian optimization approach for computing low-rank solutions of Lyapunov equations, SIAM J. Matrix Anal. Appl. 31 (5) (2010) 2553–2579.

[32] U. Mor, H. Avron, Solving Trust Region Subproblems Using Riemannian Optimization, arXiv preprint arXiv:2010.07547 (2020).

[33] D. Kressner, M. Steinlechner, B. Vandereycken, Preconditioned low-rank Riemannian optimization for linear systems with tensor product structure, SIAM J Sci Comput. 38 (4) (2016) A2018–A2044.

[34] H.-G. Bock, Randwertproblemmethoden zur Parameteridentifizierung in systemen nichtlinearer Differentialgleichungen, Bonner mathematische Schriften, Rheinische Friedrich-Wilhelm Universität, Academic Dissertation 16795956, 1987 (in German).

[35] G. H. Golub, C. F. Van Loan, Matrix Computations (4rd Ed.), Johns Hopkins University Press, Baltimore, MD, USA, 2013.

[36] L. N. Trefethen, D. Bau III, Numerical linear algebra, Vol. 50, SIAM, 1997.

[37] M. Lezcano Casado, Trivializations for gradient-based optimization on manifolds, Adv Neural Inf Process Syst. 32 (2019).

[38] C. Criscitiello, N. Boumal, An accelerated first-order method for non-convex optimization on manifolds, arXiv preprint arXiv:2008.02252 (2020).

[39] G. C. Bento, O. P. Ferreira, J. G. Melo, Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds, J Optim Theory Appl. 173 (2) (2017) 548–562.

[40] O. P. Ferreira, P. R. Oliveira, Proximal point algorithm on Riemannian manifolds, Optim. 51 (2) (2002) 257–270.

[41] X. Zhu, H. Sato, Riemannian conjugate gradient methods with inverse retraction, Comput Optim Appl. 77 (3) (2020) 779–810.

[42] R. A. Horn, C. R. Johnson, Matrix Analysis, 2nd Ed., Cambridge University Press, 2012.

[43] F. L. Bauer, C. T. Fike, Norms and exclusion theorems, Numer Math (Heidelb) 2 (1) (1960) 137–141.

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

[45] R. Bhatia, P. Rosenthal, How and why to solve the operator equation $AX - XB = Y$, Bull. London Math. Soc. 29 (1) (1997) 1–21.

[46] N. Boumal, P. A. Absil, C. Cartis, Global rates of convergence for nonconvex optimization on manifolds, IMA J. Numer. Anal. 39 (1) (2019) 1–33.

[47] L. Sun, B. Ceran, J. Ye, A scalable two-stage approach for a class of dimensionality reduction techniques, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2010, pp. 313–322.

[48] K. Chaudhuri, S. M. Kakade, K. Livescu, K. Sridharan, Multi-view clustering via canonical correlation analysis, in: Proceedings of the 26th Annual International Conference on Machine Learning (ICML), ACM, 2009, pp. 129–136.

[49] P. Dhillon, D. P. Foster, L. H. Ungar, Multi-view learning of word embeddings via CCA, in: Adv Neural Inf Process Syst., 2011, pp. 199–207.

[50] P. S. Dhillon, J. Rodu, D. P. Foster, L. H. Ungar, Two Step CCA: A New Spectral Method for Estimating Vector Models of Words, in: Proceedings of the 29th International Conference on International Conference on Machine Learning, ICML'12, Omnipress, USA, 2012, pp. 67–74.
URL `http://dl.acm.org/citation.cfm?id=3042573.3042586`

[51] Y. Su, Y. Fu, X. Gao, Q. Tian, Discriminant learning through multiple principal angles for visual recognition, IEEE Trans Image Process. 21 (3) (2012) 1381–1390.

[52] T.-K. Kim, J. Kittler, R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, IEEE Trans Pattern Anal Mach Intell. 29 (6) (2007) 1005–1018.

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

[54] A. Gonen, F. Orabona, S. Shalev-Shwartz, Solving Ridge Regression Using Sketched Preconditioned SVRG, in: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16, JMLR.org, 2016, pp. 1397–1405.
URL `http://dl.acm.org/citation.cfm?id=3045390.3045538`

[55] R. H. Bartels, G. W. Stewart, Solution of the Matrix Equation $AX + XB = C$, Commun. ACM 15 (9) (1972) 820–826. `doi:10.1145/361573.361582`.
URL `https://doi.org/10.1145/361573.361582`

[56] P. A. Absil, J. Trumpf, R. Mahony, B. Andrews, All roads lead to Newton: Feasible second-order methods for equality-constrained optimization, Technical Report UCL-INMA-2009.024 (2009).

[57] J. Von Neumann, Some matrix-inequalities and metrization of matric-space. Tomsk Univ. Rev. 1 (1937) 286–300.

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

$$T_X\mathrm{St}_B(p,d)=\left\{Z\in\mathbb{R}^{d\times p}\;:\;Z^{\mathrm{T}}BX+X^{\mathrm{T}}BZ=0_p\right\},\tag{A.1}$$

is based on the Submersion Theorem [5, Proposition 3.3.3]. $\mathrm{St}_B(p,d)$ is the kernel of the mapping $F(X)=X^{\mathrm{T}}BX-I_p$, i.e., $\mathrm{St}_B(p,d)=F^{-1}(0_p)$. 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 $DF(X)[\cdot]:\mathbb{R}^{d\times p}\to\mathcal{S}_{\mathrm{sym}}(p)$. For every $\hat{Z}\in\mathcal{S}_{\mathrm{sym}}(p)$, the matrix $Z=\frac{1}{2}X\hat{Z}\in\mathbb{R}^{d\times p}$ satisfies $DF(X)[Z]=\hat{Z}$. According to [5, Proposition 3.3.3] then $\mathrm{St}_B(p,d)$ is an embedded submanifold of $\mathbb{R}^{d\times p}$, and its dimension is $dp-\frac{p(p+1)}{2}$.

The second characterization is:

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

where Ω is a skew-symmetric matrix (i.e., $\Omega^{\mathrm{T}}=-\Omega$), K is arbitrary, and $X_{B\perp}\in\mathbb{R}^{d\times(d-p)}$ 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., $X_{B\perp}^{\mathrm{T}}BX_{B\perp}=I_{d-p}$, and $X_{B\perp}^{\mathrm{T}}BX=0_{(d-p)\times p}$. The dimension of the space defined in (A.2) is $p(p-1)/2+p(d-p)=dp-p(p+1)/2$. Both characterizations of $T_X\mathrm{St}_B(p,d)$, (A.1) and (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 (A.1), then $\Omega^{\mathrm{T}}=-\Omega$, so that Z belongs to the set defined in (A.2). Thus, the set defined in (A.1) is a subset (subspace) of the set defined in (A.2). Finally, since both the sets defined in (A.1) and (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 (A.1) and (A.2) are equal.

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

$$R_X^{\mathrm{polar}}(\xi_X)\coloneqq(X+\xi_X)(I_p+\xi_X^{\mathrm{T}}B\xi_X)^{-1/2}\tag{A.3}$$

$$R_X^{\mathrm{QR}}(\xi_X)\coloneqq\mathbf{qf}_B\left(X+\xi_X\right)=B^{-1/2}\mathbf{qf}\left(B^{1/2}\left(X+\xi_X\right)\right)\tag{A.4}$$

$$R_X^{\mathrm{Cayley}}(\xi_X)\coloneqq(I_d-\frac{1}{2}W(\xi_X))^{-1}(I_d+\frac{1}{2}W(\xi_X))X\tag{A.5}$$

where

$$W(\xi_X)\coloneqq(I_d-\frac{1}{2}XX^{\mathrm{T}}B)\xi_XX^{\mathrm{T}}B-X\xi_X^{\mathrm{T}}(I_d-\frac{1}{2}BXX^{\mathrm{T}})B.$$

The cost of computing the polar-based retraction, (A.3), is $O\left(T_Bp+dp^2\right)$ 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 [15, Section 3.2]. The cost of computing the QR-based retraction, (A.4), is also $O\left(T_Bp+dp^2\right)$. This is shown in [15, Section 3.2]. Though, in [15], it is claimed that for large p ($p\leq d$) the QR-based retraction has an advantage in computational costs compared to the polar-based retraction, since the eigenvalue decomposition of $(X+\xi_X)^{\mathrm{T}}B(X+\xi_X)$ (or SVD decomposition of $X+\xi_X$) can be replaced with a Cholesky decomposition of the same matrix. The cost of computing the Cayley transform based retraction, (A.5), is $O\left(T_B p + dp^2\right)$ which follows using the Sherman-Morrison-Woodbury formula as described in Subsection 3.1. Another approach suggested in [14] by Li et al. is to use a fixed point method to approximate the retraction.

The retraction in (A.4) is proven to be indeed a retraction mapping in [15, Theorem 3.1]. For the retraction in (A.3), though we found the equation in the literature, we could not find a formal argument that it is a retraction. Therefore, we show this by showing that it meets the conditions in [5, Definition 4.1.1]. The first condition of [5, Definition 4.1.1] is that $R_X(0_X)=X$, and it indeed holds since $R_X^{\mathrm{polar}}(0_X) = (X+0_X)(I_p+0_X^{\mathrm{T}}B0_X)^{-1/2}=X$. The second condition of [5, Definition 4.1.1] is that $\mathrm{D}R_x(0_x)=\mathrm{id}_{T_X\mathrm{St}_B(p,d)}$, where $\mathrm{id}_{T_X\mathrm{St}_B(p,d)}$ denotes the identity mapping on $T_X\mathrm{St}_B(p,d)$. This condition is equivalent to the condition that for every vector $\xi_X\in T_X\mathrm{St}_B(p,d)$ we have $\frac{\mathrm{d}}{\mathrm{dt}}R_X(t\xi_X)\big|_{t=0}=\xi_X$. Denote by $\lambda_1,...,\lambda_p\geq 0$ the eigenvalues of $\xi_X^{\mathrm{T}}B\xi_X$, then

$$(I_p+t^2\xi_X^{\mathrm{T}}B\xi_X)^{-1/2}=Q\begin{pmatrix}\frac{1}{\sqrt{1+t^2\lambda_1}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{1+t^2\lambda_p}}\end{pmatrix}Q^{\mathrm{T}},$$

where Q is an orthogonal matrix that diagonalizes $\xi_X^{\mathrm{T}}B\xi_X$. Then,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}R_X^{\mathrm{polar}}(t\xi_X)\bigg|_{t=0} &= \frac{\mathrm{d}}{\mathrm{dt}}\left[(X+t\xi_X)(I_p+t^2\xi_X^{\mathrm{T}}B\xi_X)^{1/2}\right]\bigg|_{t=0} = \\ &= \frac{\mathrm{d}}{\mathrm{dt}}\left[(X+t\xi_X)Q\begin{pmatrix}\frac{1}{\sqrt{1+t^2\lambda_1}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{1+t^2\lambda_p}}\end{pmatrix}Q^{\mathrm{T}}\right]\Bigg|_{t=0}\end{aligned}$$

$$= \xi_X Q\begin{pmatrix}\frac{1}{\sqrt{1+t^2\lambda_1}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{1+t^2\lambda_p}}\end{pmatrix}Q^{\mathrm{T}} - (X+t\xi_X)Q\begin{pmatrix}\frac{t\lambda_1}{(1+t^2\lambda_1)^{1.5}} & & \\ & \ddots & \\ & & \frac{t\lambda_p}{(1+t^2\lambda_p)^{1.5}}\end{pmatrix}Q^{\mathrm{T}}\Bigg|_{t=0} = \xi_X.$$

Similarly, the retraction in (A.5) is also proven to be a retraction mapping in [14, Eq. (4)] for the Stiefel manifold. In order to generalize it to the generalized Stiefel manifold, we show it meets the conditions in [5, Definition 4.1.1]. For the first condition we have $W(0_X)=0_d$, thus

$$R_X^{\mathrm{Cayley}}(0_X)=(I_d-0_d)^{-1}(I_d+0_d)X=X\ .$$

For the second condition, we have

$$\frac{\mathrm{d}}{\mathrm{dt}}R_X^{\mathrm{Cayley}}(t\xi_X)\bigg|_{t=0}=W(\xi_X)X=\xi_X\ ,$$

where we used $X^{\mathrm{T}}BX=I_p$ and the definition of tangent vectors on $\mathrm{St}_B(p,d)$, (A.1), i.e., $\xi_X^{\mathrm{T}}BX+X^{\mathrm{T}}B\xi_X = 0_p$.

Let us consider the inverse of the polar retraction. Suppose that $Y=R_X^{\mathrm{polar}}(\xi_X)$. Using the definition of the polar retraction, and reordering the equation we find that

$$\xi_X = Y(I_p+\xi_X^{\mathrm{T}}B\xi_X)^{1/2}-X\ . \tag{A.6}$$

Left multiply by $X^{\mathrm{T}}B$, and recall that $X^{\mathrm{T}}BX = I_p$, to find that

$$X^{\mathrm{T}}B\xi_X = X^{\mathrm{T}}BY(I_p + \xi_X^{\mathrm{T}}B\xi_X)^{1/2} - I_p \ .$$

Now using the fact that $X^{\mathrm{T}}B\xi_X + \xi_B^{\mathrm{T}}BX = 0_p$ (since ξ_X is a tangent vector), we find that

$$X^{\mathrm{T}}BY(I_p + \xi_X^{\mathrm{T}}B\xi_X)^{1/2} + (I_p + \xi_X^{\mathrm{T}}B\xi_X)^{1/2}Y^{\mathrm{T}}BX - 2I_p = 0_p \,.$$

Thus $Z = (I_p + \xi_X^{\mathrm{T}}B\xi_X)^{1/2}$ is SPD solution to (3.5). If we can uniquely recover $(I_p + \xi_X^{\mathrm{T}}B\xi_X)^{1/2}$ by solving (3.5) (something we can do in a small neighborhood of X that intersects with the image of the polar retraction), we can use (A.6) to invert the polar retraction.

The derivation of the inverse of the QR retraction is similar. Suppose that $Y = R_X^{\mathrm{QR}}(\xi_X)$. Using the definition of the QR-based retraction, and reordering the equation we find that

$$\xi_X = YR - X, \tag{A.7}$$

where R is an upper-triangular matrix with strictly positive elements on its main diagonal such that

$$\mathbf{qf}\left(B^{1/2}\left(X + \xi_X\right)\right) R = B^{1/2}\left(X + \xi_X\right).$$

To find R, left multiply by $X^{\mathrm{T}}B$, and recall that $X^{\mathrm{T}}BX = I_p$ to find that

$$X^{\mathrm{T}}B\xi_X = X^{\mathrm{T}}BYR - I_p \ .$$

Now using the fact that $X^{\mathrm{T}}B\xi_X + \xi_B^{\mathrm{T}}BX = 0_p$ (since ξ_X is a tangent vector), we find that

$$X^{\mathrm{T}}BYR + R^{\mathrm{T}}Y^{\mathrm{T}}BX - 2I_p = 0_p \,.$$

Thus, R is an upper-triangular matrix with strictly positive elements on its main diagonal solving to (3.8). If we can uniquely recover R by solving (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 (A.7) to invert the QR-based retraction.

We remind here the conditions for a unique solution for (3.8). According to [16, Eq. (14) and Algorithm 1], (3.8) is equivalent to the set of the following p linear equations

$$\tilde{M}_i\tilde{r}_i = b_i, \ i = 1, ..., p,$$

where $\tilde{M}_i$ is the i-th principal minor extracted from the matrix $X^{\mathrm{T}}BY$, $\tilde{r}_i$ 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

$$-[m_{i1}, ...m_{ij}]\tilde{r}_j \ ,$$

where $j = 1, ..., i-1$, m_{ik} are elements of the i-th row of $\tilde{M}_i$, 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 $X^{\mathrm{T}}BY$ 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 $X^{\mathrm{T}}BY$ are strictly positive, thus $\det(\tilde{M}_i) > 0$. 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.

We also show the derivation of retraction based vector transports using equations (A.3) and (A.4) similarly to [13]. For (A.3) denote

$$A(t) \coloneqq I_p + \left(\eta_X + t\xi_X\right)^{\mathrm{T}} B\left(\eta_X + t\xi_X\right),$$

then,

$$\begin{aligned}
\tau_{\eta_X}^{(\text{polar})} &\coloneqq \mathrm{D}R_X^{\text{polar}}(\eta_X)[\xi_X] \qquad\qquad (A.8)\\
&= \frac{\mathrm{d}}{\mathrm{dt}} R_X^{\text{polar}}(\eta_X + t\xi_X)\bigg|_{t=0}\\
&= \frac{\mathrm{d}}{\mathrm{dt}} \left[(X+\eta_X+t\xi_X)\left(A(t)\right)^{-1/2}\right]\bigg|_{t=0}\\
&= \xi_X \left(A(0)\right)^{-1/2} + (X+\eta_X)\,\frac{\mathrm{d}}{\mathrm{dt}}\left(A(t)\right)^{-1/2}\bigg|_{t=0}\\
&= \xi_X \left(A(0)\right)^{-1/2} - (X+\eta_X)\left(A(0)\right)^{-1/2}\frac{\mathrm{d}}{\mathrm{dt}}\left(A(t)\right)^{1/2}\bigg|_{t=0}\left(A(0)\right)^{-1/2},
\end{aligned}$$

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

$$I = \left(A(t)\right)^{-1/2}\left(A(t)\right)^{1/2},$$
$$A(t) = \left(A(t)\right)^{1/2}\left(A(t)\right)^{1/2},$$

which leads to

$$0 = \frac{\mathrm{d}}{\mathrm{dt}}\left(A(t)\right)^{-1/2}\left(A(t)\right)^{1/2} + \left(A(t)\right)^{-1/2}\frac{\mathrm{d}}{\mathrm{dt}}\left(A(t)\right)^{1/2},$$
$$\frac{\mathrm{d}}{\mathrm{dt}}\left(A(t)\right)^{-1/2} = -\left(A(t)\right)^{-1/2}\frac{\mathrm{d}}{\mathrm{dt}}\left(A(t)\right)^{1/2}\left(A(t)\right)^{-1/2},$$

and

$$\frac{\mathrm{d}}{\mathrm{dt}}A(t) = \frac{\mathrm{d}}{\mathrm{dt}}\left(A(t)\right)^{1/2}\left(A(t)\right)^{1/2} + \left(A(t)\right)^{1/2}\frac{\mathrm{d}}{\mathrm{dt}}\left(A(t)\right)^{1/2},$$
$$\xi_X^{\mathrm{T}}B\eta_X + \eta_X^{\mathrm{T}}B\xi_X + 2t\xi_X^{\mathrm{T}}B\xi_X = \frac{\mathrm{d}}{\mathrm{dt}}\left(A(t)\right)^{1/2}\left(A(t)\right)^{1/2} + \left(A(t)\right)^{1/2}\frac{\mathrm{d}}{\mathrm{dt}}\left(A(t)\right)^{1/2}.$$

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

$$\frac{\mathrm{d}}{\mathrm{dt}}\left(A(t)\right)^{1/2}\bigg|_{t=0}\left(A(0)\right)^{1/2} + \left(A(0)\right)^{1/2}\frac{\mathrm{d}}{\mathrm{dt}}\left(A(t)\right)^{1/2}\bigg|_{t=0} = \xi_X^{\mathrm{T}}B\eta_X + \eta_X^{\mathrm{T}}B\xi_X. \qquad (A.9)$$

According to [42, Theorem 2.4.4.1], there is a unique solution to (A.9) for any $\xi_X^{\mathrm{T}}B\eta_X + \eta_X^{\mathrm{T}}B\xi_X$, since $\left(A(0)\right)^{1/2} = \left(I_p + \eta_X^{\mathrm{T}}B\eta_X\right)^{1/2}$ is positive definite ($\eta_X^{\mathrm{T}}B\eta_X$ is a symmetric positive semi-definite matrix) and $-\left(I_p + \eta_X^{\mathrm{T}}B\eta_X\right)^{1/2}$ is negative definite, thus they have no eigenvalues in common. Solving (A.9) costs $O(p^3)$ (e.g., using the Bartels–Stewart algorithm [55]). 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 (A.8) is $O(T_B p + dp^2)$.

For (3.6):

$$\begin{aligned}
\tau_{\eta_X}^{(\text{QR})} &\coloneqq \mathrm{D}R_X^{\text{QR}}(\eta_X)[\xi_X] = \mathrm{Dqf}_B(X+\eta_X)[\xi_X] \qquad\qquad (A.10)\\
&= B^{-1/2}\mathrm{Dqf}\left(B^{1/2}\left(X+\xi_X\right)\right)[B^{1/2}\xi_X]\\
&= B^{-1/2}\Bigg[\mathrm{qf}\left(B^{1/2}\left(X+\xi_X\right)\right)\rho_{\text{skew}}\left(\mathrm{qf}\left(B^{1/2}\left(X+\xi_X\right)\right)^{\mathrm{T}}B^{1/2}\xi_X\left(\mathrm{qf}\left(B^{1/2}\left(X+\xi_X\right)\right)^{\mathrm{T}}B^{1/2}\left(X+\xi_X\right)\right)^{-1}\right)+\\
&\quad + \left(I_n - \mathrm{qf}\left(B^{1/2}\left(X+\xi_X\right)\right)\mathrm{qf}\left(B^{1/2}\left(X+\xi_X\right)\right)^{\mathrm{T}}\right)B^{1/2}\xi_X\left(\mathrm{qf}\left(B^{1/2}\left(X+\xi_X\right)\right)^{\mathrm{T}}B^{1/2}\left(X+\xi_X\right)\right)^{-1}\Bigg],
\end{aligned}$$

where the last equality is due to [5, Example 8.1.5]:

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

and $\rho_{\mathrm{skew}}(\cdot)$ 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,

$$\left(\rho_{\mathrm{skew}}(A)\right)_{i,j} \;=\; \begin{cases} A_{i,j} & i>j \\ 0 & i=j \\ -A_{j,i} & i<j \end{cases}.$$

Computing (A.10) can be done in the following way. First, computing $B^{-1/2}\mathbf{qf}\left(B^{1/2}\left(X+\xi_X\right)\right)$ costs $O\left(T_Bp+dp^2\right)$ (see computational cost of (3.6)). Also, computing $\mathrm{qf}\left(B^{1/2}\left(X+\xi_X\right)\right)^{\mathrm{T}}B^{1/2}$ has the same cost since it is equivalent to computing $R^{-\mathrm{T}}\left(X+\xi_X\right)^{\mathrm{T}}B$, where R is the R matrix of the thin I-QR decomposition of $B^{1/2}\left(X+\xi_X\right)$, and it can be found using the Cholesky decomposition of $(X+\xi_X)^{\mathrm{T}}B(X+\xi_X)$. Applying $\rho_{\mathrm{skew}}(\cdot)$ 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 (A.10) is $O(T_Bp+dp^2)$.

Both forms of vector transport (A.8) and (A.10) satisfy [5, Definition 8.1.1]. The vector transport based on the Cayley transform is derived in [13, Eq. (16)]. It features the same computational complexity as computing the retraction (A.5).

A.2 Metric Related Notions

We detail the derivation of the Riemannian Hessian that led to (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 $X\in\mathbf{St}_B(p,d)$. We remark again that (3.21) holds also with a non-constant M_X at the critical points.

We use [5, Definition 5.5.1] of the Riemannian Hessian: For a real-valued function f on $\mathrm{St}_B(p,d)$, at a point $X\in\mathrm{St}_B(p,d)$ the Riemannian Hessian $\mathbf{Hess}f(X)$ is a linear mapping of $T_X\mathbf{St}_B(p,d)$ into itself such that

$$\mathbf{Hess}f(X)[\eta_X]=\nabla_{\eta_X}\mathbf{grad}f(X),$$

for all $\eta_X\in T_X\mathbf{St}_B(p,d)$. In the previous equation, ∇ is the Riemannian connection, which should not be confused with the Euclidean gradient.

First, we find the Riemannian connection on $\mathbf{St}_B(p,d)$ 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 [5, Proposition 5.3.2]: composing the connection in the ambient space with the projection on the tangent space. Let $\bar{\nabla}$ be the Levi-Civita connection on $\mathbb{R}^{d\times p}$ endowed with the metric $\bar{g}$. Let $(e_1,...,e_{dp})=(E_{11},E_{21},...,E_{d1},E_{12},...,E_{d2},...,E_{dp})$ 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 $0\le k\le p,\ 0\le r<d$ 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

$$\overline{\nabla}_{\overline{\eta}(\cdot)}\overline{\xi}(\cdot)=\sum_{i,j}\left(\overline{\eta^i}(\cdot)\,\overline{\xi^j}(\cdot)\overline{\nabla}_{e_i(\cdot)}e_j(\cdot)+\overline{\eta^i}(\cdot)\partial_i\overline{\xi^j}(\cdot)e_j(\cdot)\right),$$

where $\overline{\eta}(\cdot), \overline{\xi}(\cdot), e_i(\cdot), \overline{\nabla}_{\overline{\eta}(\cdot)}\overline{\xi}(\cdot), \overline{\nabla}_{e_i(\cdot)}e_j(\cdot)$ 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., $\overline{\eta}(X) = \overline{\eta}_X$). In particular, $\overline{\eta}(\cdot)$ and $\overline{\xi}(\cdot)$ are smooth local extensions of the vector fields $\eta(\cdot)$ and $\xi(\cdot)$ on $\mathbf{St}_B(p,d)$ in a neighborhood of $X \in \mathbf{St}_B(p,d)$ in $\mathbb{R}^{d\times p}$, in the sense that for $X \in \mathbb{R}^{d\times p}$ the vector fields $\overline{\eta}(\cdot)$ and $\overline{\xi}(\cdot)$ assign the same tangent vectors as $\eta(\cdot)$ and $\xi(\cdot)$. Note that the vector field $\overline{\nabla}_{\overline{\eta}(\cdot)}\overline{\xi}(\cdot)$ at X depends on $\overline{\eta}(X) = \overline{\eta}_X$ (see [6, Proposition 5.18.]). Thus, we can write $\overline{\nabla}_{\overline{\eta}(\cdot)}\overline{\xi}(\cdot)$ at X in the following way $\overline{\nabla}_{\overline{\eta}_X}\overline{\xi}(X)$. In addition, given $\eta_X \in T_X\mathbf{St}_B(p,d)$ and a vector field $\xi(X)$ on $\mathrm{St}_B(p,d)$, the connection $\overline{\nabla}_{\eta_X}\xi(X)$ is defined by $\overline{\nabla}_{\overline{\eta}_X}\overline{\xi}(X)$ according to [5, Equation 5.13] and it does not depend on the local extension of $\xi(X)$. Recall that $\left(\overline{\nabla}_{e_i(\cdot)}e_j(\cdot)\right)_k = \Gamma^k_{i,j}$ (k-th coordinate of $\overline{\nabla}_{e_i}e_j$) are the Christoffel symbols. These symbols determine the connection $\overline{\nabla}$ uniquely, using the Fundamental Theorem of Riemannian Geometry for the Levi-Civita connection. The Christoffel symbols can be calculated using

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

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

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

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

$$\overline{\nabla}_{\overline{\eta}_X}\overline{\xi}(X) = \sum_{j=1}^{dp}\sum_{i=1}^{dp}\left(\overline{\eta^i_X}\partial_i\overline{\xi^j}(X)e_j\right) = J_{\overline{\xi}(X)}\overline{\eta}_X,$$

where $J_{\overline{\xi}(X)}\overline{\eta}_X$ denotes the Jacobian matrix of $\overline{\xi}(X)$ at X in the direction $\overline{\eta}_X$. 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 $\mathbf{St}_B(p,d)$. Given $\eta_X \in T_X\mathbf{St}_B(p,d)$ and a vector field $\xi(X)$ on $\mathbf{St}_B(p,d)$, the Riemannian connection is (written, as usual, in terms of ambient coordinates):

$$\nabla_{\eta_X}\xi(X) = \Pi_X\left(\overline{\nabla}_{\overline{\eta}_X}\overline{\xi}(X)\right) = \Pi_X\left(J_{\overline{\xi}(X)}\eta_X\right) \tag{A.11}$$

where $\eta_X = \overline{\eta}_X$ and $\overline{\xi}(\cdot)$ is any smooth local extension of $\xi(\cdot)$ in a neighborhood of $X \in \mathbf{St}_B(p,d)$ in $\mathbb{R}^{d\times p}$.

Next, we can find the Riemannian Hessian using (A.11), the product rule for derivation and according to [44]:

$$\begin{aligned}
\mathbf{Hess}f(X)[\eta_X] &= \nabla_{\eta_X}\mathbf{grad}f(X) \\
&= \Pi_X\left(J_{h(X)}\eta_X\right) \\
&= \Pi_X\left[P_X M^{-1}\nabla^2\bar{f}(X)\eta_X + (\mathrm{D}\Pi_X)[\eta_X]M^{-1}\nabla\bar{f}(X)\right] \\
&= \Pi_X\left(M^{-1}\nabla^2\bar{f}(X)\eta_X\right) + \Pi_X\left((\mathrm{D}\Pi_X)(X)[\eta_X]M^{-1}\nabla\bar{f}(X)\right)
\end{aligned} \tag{A.12}$$

where $\nabla\bar{f}(X)$ and $\nabla^2\bar{f}(X)$ are the Euclidean gradient and Hessian (respectively) of $\bar{f}$ and

$$h: \mathbb{R}^{d\times p} \to \mathbb{R}^{d\times p}, \ h(X) = \Pi_X\left(M^{-1}\nabla\bar{f}(X)\right).$$

Note that for $X \in \mathbf{St}_B(p,d)$ we have $h(X) = \mathbf{grad}f(X)$ so h is a smooth local extension of the vector field $\mathbf{grad}f$ to $\mathbb{R}^{d\times p}$, and its Jacobian is calculated as follows

$$J_{h(X)}\eta_X = (\mathrm{D}\Pi_X)[\eta_X]M^{-1}\nabla\bar{f}(X) + \Pi_X\left(M^{-1}\nabla^2\bar{f}(X)\eta_X\right),$$

where $(\mathrm{D}\Pi_X)[\eta_X]$ (here and in (A.12)) is the derivative at X along η_X of the function that maps X to Π_X.

The main challenge in computing the Riemannian Hessian from (A.12) is in computing $(\mathrm{D}\Pi_X)[\eta_X]$. In order to circumvent this issue, we use a simple modification of a result found in [44] 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 $g_X(\xi_X,\eta_X) = \mathbf{Tr}\left(\xi_X^{\mathrm{T}} M \eta_X\right)$ 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.*

Definition A.1. ([56, Section 6.1], [44, Definition 1]) 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

$$W_x\left(\eta_x, u_x\right) := -\Pi_x(\mathrm{D}u(x)[\eta_x]) \tag{A.13}$$

where $u(\cdot)$ is a smooth normal vector field on $\mathcal{M}$ which satisfies $u(x) = u_x$.

For the manifold $\mathbf{St}_B(p,d)$, viewed as an embedded submanifold of $\mathbb{R}^{d\times p}$, (A.13) reduces to

$$W_X\left(\eta_X, U(X)\right) \;=\; -\Pi_X\left(J_{\bar{U}(X)}\eta_X\right),$$

where $\bar{U}(\cdot)$ is any smooth local extension of the normal vector field $U(\cdot)$ such that $U(X) = U_X$ on $\mathbf{St}_B(p,d)$. Now, that at a point $X \in \mathbf{St}_B(p,d)$ any normal vector is of the form $U_X = M^{-1}BXS_X$ for some $S_X \in \mathcal{S}_{\mathrm{sym}}(p)$. Left multiplying by $X^{\mathrm{T}}M$ we get $X^{\mathrm{T}}MU_X = X^{\mathrm{T}}MM^{-1}BXS_X = S_X$. Now we can define a normal field on $\mathbf{St}_B(p,d)$ by the formula $U(X) = M^{-1}BXS_X$ such that $U(X) = U_X$ with $S_X = X^{\mathrm{T}}MU_X$ such that $S_X \in \mathcal{S}_{\mathrm{sym}}(p)$. The vector field can be extended to $\mathbb{R}^{d\times p}$ by the same formula such that $\bar{U}(\cdot)$ and $U(\cdot)$ coincide on $\mathbf{St}_B(p,d)$. Next, we calculate the Jacobian of $\bar{U}(X)$ at the direction η_X:

$$J_{\bar{U}(X)}\eta_X = M^{-1}B\eta_X S_X\ ,$$

Therefore the Weingarten map for $\mathbf{St}_B(p,d)$ is

$$\begin{aligned} W_X\left(\eta_X, U_X\right) &= -\Pi_X(M^{-1}B\eta_X S_X) \\ &= -\Pi_X\left(M^{-1}B\eta_X\left(X^{\mathrm{T}}MU_X\right)\right)\ . \end{aligned}$$

The following lemma is a simple modification of [44, Theorem 1]. Although the proof is almost identical, we include it here for completeness.

Lemma A.2. *For the Riemannian submanifold $\mathbf{St}_B(p,d)$ of $\mathbb{R}^{d\times p}$ endowed with $\bar{g}_X(\bar{\xi}_X,\bar{\eta}_X) = \mathbf{Tr}\left(\bar{\xi}_X^{\mathrm{T}} M \bar{\eta}_X\right)$ we have*

$$\begin{aligned} W_X\left(\eta_X, \Pi_X^\perp\left(M^{-1}U\right)\right) &= \Pi_X\left((D\Pi_X)[\eta_X]\left(M^{-1}U\right)\right) \\ &= \Pi_X\left((D\Pi_X)[\eta_X]\left(\Pi_X^\perp\left(M^{-1}U\right)\right)\right)\ , \end{aligned}$$

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

$$\Pi_X\left((\mathrm{D}\Pi_X)[\eta_X]\,(\cdot)\right) = \Pi_X\left((\mathrm{D}\Pi_X)[\eta_X]\left(\Pi_X^\perp(\cdot)\right)\right) \tag{A.14}$$

holds. Then applying both sided on $M^{-1}U$ gives us the equality

$$\Pi_X\left((\mathrm{D}\Pi_X)[\eta_X]M^{-1}U\right) = \Pi_X\left((\mathrm{D}\Pi_X)[\eta_X]\Pi_X^\perp\left(M^{-1}U\right)\right).$$

To show this, we take the directional derivative of the equality $\Pi_X\left(\Pi_X^\perp(\cdot)\right) = 0$ in the direction η_X, and we use $\Pi_X^\perp(\cdot) = \left(\mathrm{id}_{T_X\mathbb{R}^{d\times p}} - \Pi_X\right)(\cdot)$ to get

$$\begin{aligned} 0 &= (\mathrm{D}\Pi_X)[\eta_X]\Pi_X^\perp(\cdot) + \Pi_X\left((\mathrm{D}\Pi_X^\perp)[\eta_X]\,(\cdot)\right) \\ &= (\mathrm{D}\Pi_X)[\eta_X]\Pi_X^\perp(\cdot) - \Pi_X\left((\mathrm{D}\Pi_X)[\eta_X]\,(\cdot)\right). \end{aligned}$$

Substituting any tangent vector in both sides of the equation nullifies the term $(\mathrm{D}\Pi_X)[\eta_X]\Pi_X^{\perp}(\cdot)$. Thus, we substitute Π_X and use $\Pi_X^{\perp}(\Pi_X(\cdot)) = 0$

$$0 = \Pi_X((\mathrm{D}\Pi_X)[\eta_X](\Pi_X(\cdot))).$$

Finally, to get (A.14), we use $\mathrm{id}_{T_X\mathbb{R}^{d\times p}}(\cdot) = \left(\Pi_X + \Pi_X^{\perp}\right)(\cdot)$ and get

$$\begin{aligned} \Pi_X((\mathrm{D}\Pi_X)[\eta_X](\cdot)) &= \Pi_X\left((\mathrm{D}\Pi_X)[\eta_X]\left(\left(\Pi_X + \Pi_X^{\perp}\right)(\cdot)\right)\right) \\ &= \Pi_X\left((\mathrm{D}\Pi_X)[\eta_X]\left(\Pi_X^{\perp}(\cdot)\right)\right). \end{aligned}$$

To conclude the proof we show that $W_X\left(\eta_X, \Pi_X^{\perp}\left(M^{-1}U\right)\right) = \Pi_X\left((\mathrm{D}\Pi_X)[\eta_X]\left(M^{-1}U\right)\right)$. 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 A.1 along this observation, we have

$$\begin{aligned} W_X\left(\eta_X, \Pi_X^{\perp}\left(M^{-1}U\right)\right) &= -\Pi_X\left(J_{\Pi_X^{\perp}(M^{-1}U(X))}\eta_X\right) \\ &= -\Pi_X\left((\mathrm{D}\Pi_X^{\perp})[\eta_X]\left(M^{-1}U_X\right)\right) - \Pi_X\left(\Pi_X^{\perp}\left(J_{M^{-1}U(X)}\eta_X\right)\right) \\ &= \Pi_X\left((\mathrm{D}\Pi_X)[\eta_X]\left(M^{-1}U_X\right)\right), \end{aligned}$$

where in the last equality we used $\Pi_X\left(\Pi_X^{\perp}(\cdot)\right) = 0$ and $\Pi_X^{\perp}(\cdot) = (\mathrm{id}_{T_X\mathbb{R}^{d\times p}} - \Pi_X)(\cdot)$. □

As a consequence of Lemma A.2, we can replace $\Pi_X\left((\mathrm{D}\Pi_X)[\eta_X]\left(M^{-1}\nabla\bar{f}(X)\right)\right)$ by $W_X\left(\eta_X, \Pi_X^{\perp}\left(M^{-1}\nabla\bar{f}(X)\right)\right)$ in (A.12). Therefore the expression for the Riemannian Hessian becomes

$$\mathbf{Hess}f(X)[\eta_X] = \Pi_X\left(M^{-1}\nabla^2\bar{f}(X)\eta_X\right) + W_X\left(\eta_X, \Pi_X^{\perp}\left(M^{-1}\nabla\bar{f}(X)\right)\right).$$

In particular, the Riemannian Hessian on $\mathbf{St}_B(p,d)$ is

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

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

$$\begin{aligned} \mathbf{Hess}f(X)[\eta_X] &= \Pi_X\left(M^{-1}\nabla^2\bar{f}(X)\eta_X\right) - \Pi_X\left(M^{-1}B\eta_X\left(X^T\nabla\bar{f}(X) - X^{\mathrm{T}}M\left(\Pi_X\left(M^{-1}\nabla\bar{f}(X)\right)\right)\right)\right) \\ &= \Pi_X\left(M^{-1}\nabla^2\bar{f}(X)\eta_X\right) - \Pi_X\left(M^{-1}B\eta_X\left(X^T\nabla\bar{f}(X) - X^{\mathrm{T}}M\mathbf{grad}f(X)\right)\right). \end{aligned}$$

B Experiments With $p = 2$

Similarly to the experiments in Subsection 4.2, we perform experiments with the MEDIANILL dataset to demonstrate CCA for $p = 2$. We use the same choices for Riemannian metric: the trivial choice of a unit matrix $M = I_d$, the standard but expensive choice $M = \Sigma$, and four approximations of Σ via the (exact) sketched preconditioning strategy. Finding the top two correlations requires the *von Neumann cost function* [57] formulation:

$$\begin{gathered} \max \mathbf{Tr}\left(U^{\mathrm{T}}\Sigma_{xy}VN\right) \\ \text{subject to} \\ U^{\mathrm{T}}\Sigma_{xx}U = I_p \\ V^{\mathrm{T}}\Sigma_{yy}V = I_p \end{gathered}$$

where $N = \mathbf{diag}(\mu_1, \mu_2)$ and any $\mu_1 > \mu_2 > 0$ (here we take $\mu_1 = 5$ and $\mu_2 = 1$). The corresponding Riemannian components are constructed in a similar manner to Subsection 4.2.

The graphs in Fig. 2 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, which is evaluated using Manopt: the lowest condition number, 115.68, is achieved when $M = \Sigma$, and the highest, 805.2, for $M = I_d$.

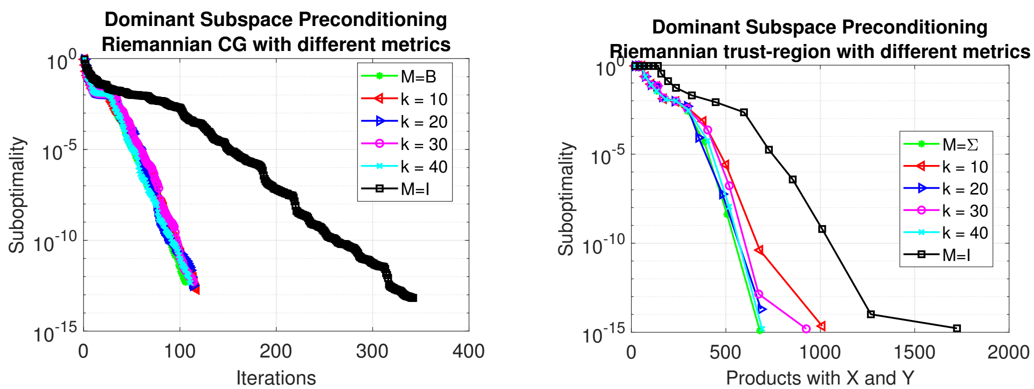

Figure 2: Results for CCA with Riemannian conjugate-gradient (left - suboptimality vs. #iterations) and Riemannain trust-region (right - suboptimality vs. products with the data matrices) with various choices of metrics for $p = 2$. The number of leading eigenvalues used to form the Dominant Subspace Preconditioner is denoted by k.