Journal of Machine Learning Research 25 (2024) 1-59 Submitted 8/21; Revised 9/24; Published 9/24

Faster Randomized Methods for Orthogonality Constrained Problems

Boris Shustin BORISSHY@GMAIL.COM
Haim Avron HAIMAV@TAUEX.TAU.AC.IL
Department of Applied Mathematics
Tel Aviv University
Tel Aviv, 69978, Israel

Editor: Michael Mahoney

Abstract

Recent literature has advocated the use of randomized methods for accelerating the solution of various matrix problems arising in machine learning and data science. One popular strategy for leveraging randomization in numerical linear algebra is to use it as a way to reduce problem size. However, methods based on this strategy lack sufficient accuracy for some applications. Randomized preconditioning is another approach for leveraging randomization in numerical linear algebra, which provides higher accuracy. The main challenge in using randomized preconditioning is the need for an underlying iterative method, thus randomized preconditioning so far has been applied almost exclusively to solving regression problems and linear systems. In this article, we show how to expand the application of randomized preconditioning to another important set of problems prevalent in machine learning: optimization problems with (generalized) orthogonality constraints. We demonstrate our approach, which is based on the framework of Riemannian optimization and Riemannian preconditioning, on the problem of computing the dominant canonical correlations and on the Fisher linear discriminant analysis problem. More broadly, our method is designed for problems with input matrices featuring one dimension much larger than the other (e.g., the number of samples much larger than the number of features). For both problems, we evaluate the effect of preconditioning on the computational costs and asymptotic convergence and demonstrate empirically the utility of our approach.

Keywords: Riemannian optimization, randomized preconditioning, generalized orthogonality constraints, generalized Stiefel manifold; matrix sketching, machine learning

1. Introduction

Matrix sketching has recently emerged as a powerful tool for accelerating the solution of many important matrix computations, with widespread use throughout machine learning and data science. Important examples of matrix computations that have been accelerated using matrix sketching include linear regression, low rank approximation, and principal component analysis (see recent surveys Woodruff, 2014; Yang et al., 2016). Matrix sketching is one of the main techniques used in so-called *Randomized Numerical Linear Algebra*.

Roughly speaking, matrix sketching provides a transformation that embeds a high dimensional space in a lower dimensional space, while preserving some desired properties of the high dimensional space (Woodruff, 2014). There are several ways in which such an

embedding can be used. The most popular approach is *sketch-and-solve*, in which matrix sketching is used to form a smaller problem. That is, sketch-and-solve based algorithms attempt to find a "good" approximate solution by sketching the input problem so that with high probability the exact solution of the sketched problem is a good approximate solution to the original problem. For example, the sketch-and-solve approach for solving unconstrained overdetermined linear regression problem, i.e., $\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2$, where $\mathbf{X} \in \mathbb{R}^{n\times d}$ $(n \geq d)$ is assumed to be full rank matrix, is to randomly generate a matrix $\mathbf{S} \in \mathbb{R}^{s\times n}$ and solve the reduced size problem $\min_{\mathbf{w}} \|\mathbf{S}\mathbf{X}\mathbf{w} - \mathbf{S}\mathbf{y}\|_2$ (Drineas et al., 2011). If $\mathbf{S}$ is chosen appropriately, then with high probability the solution of the sketched problem, $\hat{\mathbf{w}}$, is close to the exact solution, $\mathbf{w}^\star$, in the sense that the inequality $\|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|_2 \leq (1+\varepsilon)\|\mathbf{X}\mathbf{w}^\star - \mathbf{y}\|_2$ holds, and that $\mathbf{SX}$ and $\mathbf{Sy}$ can be computed quickly.

Sketch-and-solve algorithms have been proposed for a wide spectrum of linear algebra problems relevant for machine learning applications: linear regression (Drineas et al., 2011), principal component analysis (Kannan et al., 2014), canonical correlation analysis (CCA) (Avron et al., 2014a), kernelized methods (Avron et al., 2014b), low-rank approximations (Clarkson and Woodruff, 2017), structured decompositions (Boutsidis and Woodruff, 2017), etc. However, there are two main drawbacks to the sketch-and-solve approach. First, it is unable to deliver highly accurate results (typically, for sketch-and-solve algorithms, the running time dependence on the accuracy parameter ϵ is $\Theta(\epsilon^{-2})$). The second drawback is that sketch-and-solve algorithms typically have only Monte-Carlo type guarantees, i.e., they return a solution within the prescribed accuracy threshold only with high probability (on the positive side, the running time is deterministic).

These drawbacks have prompted researchers to develop a second approach, typically termed *sketch preconditioning* or *randomized preconditioning.* The main idea in randomized preconditioning is to use an iterative method which, in turn, uses a preconditioner that is formed using a sketched matrix. For example, consider again the unconstrained overdetermined linear regression problem problem. It is possible to accelerate the solution of $\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2$ by first sketching the matrix $\mathbf{X}$ to form $\mathbf{SX}$, and then using a factorization of $\mathbf{SX} = \mathbf{QR}$ to form a preconditioner, $\mathbf{R}$, for an iterative Krylov method (e.g., LSQR). By choosing $\mathbf{S}$ properly, with high probability the preconditioner $\mathbf{R}$ is such that the condition number governing the convergence of the Krylov method, $\kappa(\mathbf{X}\mathbf{R}^{-1})$, is bounded by a small constant (Rokhlin and Tygert, 2008; Avron et al., 2010; Meng et al., 2014; Clarkson and Woodruff, 2017; Gonen et al., 2016). Thus, when using a Krylov method to solve $\min_{\mathbf{w}} \|\mathbf{X}\mathbf{R}^{-1}\mathbf{w} - \mathbf{y}\|_2$ only a small number of iterations are necessary for convergence.

More generally, by using an iterative method, it is typically possible to reduce the running time dependence on the accuracy parameter to be logarithmic instead of polynomial. Furthermore, since we can control the stopping criteria of the iterative methods, sketch preconditioning algorithms typically entertain Las-Vegas type guarantees, i.e., they return a solution within the accuracy threshold, albeit at the cost of probabilistic running time.

The sketch-and-solve approach is more prevalent in the literature than sketch preconditioning. Indeed, in one way or the other, almost all sketch preconditioning methods have essentially been designed for linear regression or solving linear systems. The main reason is that sketch preconditioning requires an iterative method that can be preconditioned, and such a method is not always known for the various problems addressed by sketching. In-

deed, linear regression and solving linear systems are cases where the use of preconditioning is straightforward.

The goal of this paper is to go beyond linear regression, and design sketch preconditioning algorithms for another important class of problems: optimization problems under generalized orthogonality constraints. We would like to emphasize that in the context of optimization, preconditioning can be performed on some algorithms, e.g., algorithms based on conjugate gradients (Vandereycken and Vandewalle, 2010). However, our approach aims to be more general; rather than preconditioning each optimization method on its own, we aim to precondition the problem itself through randomized preconditioning. In general, we are interested in solving problems of the form

$$\min f(\mathbf{X}_1, \ldots, \mathbf{X}_k) \text{ s.t. } \mathbf{X}_i^{\mathrm{T}} \mathbf{B}_i \mathbf{X}_i = \mathbf{I}_p \quad (i = 1, \ldots, k) \tag{1}$$

where $f(\mathbf{X}_1, \ldots, \mathbf{X}_k) : \mathbb{R}^{d_1 \times p} \times \cdots \times \mathbb{R}^{d_k \times p} \to \mathbb{R}$ is a smooth function, and $\mathbf{B}_i \in \mathbb{R}^{d_i \times d_i}$ are fixed symmetric positive definite (SPD) matrices. An important example of such problems is the problem of finding the dominant subspace: given a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, if $f(\mathbf{X}_1) = -\mathbf{Tr}\left(\mathbf{X}_1^{\mathrm{T}} \mathbf{A} \mathbf{X}_1\right)$ and $\mathbf{B}_1 = \mathbf{I}_n$, then Problem (1) corresponds to finding a basis for the dominant eigenspace.

Problems of the form of Eq. (1) frequently appear in machine learning. In this paper, we focus on a specific important subset of these problems; the constraint matrices $\mathbf{B}_i$ are not given explicitly, however, as is typically happening in various machine learning problems, we assume that they can be written as a Gram matrix of a tall-and-skinny data matrix $\mathbf{Z}_i$, i.e., $\mathbf{B}_i = \mathbf{Z}_i^{\mathrm{T}} \mathbf{Z}_i$ for $\mathbf{Z}_i \in \mathbb{R}^{n_i \times d_i}$. More generally, we allow the following form $\mathbf{B}_i = \mathbf{Z}_i^{\mathrm{T}} \mathbf{Z}_i + \lambda_i \mathbf{I}_d$ where $\lambda_i \geq 0$ is some regularization parameter. Moreover, we assume that $n_i \gg d_i$. Indeed, that is often the case that $\mathbf{Z}_i$ represent data matrices of samples stacked in rows, such that the number of samples, n_i, is much larger than the dimension of each data point, d_i. We aim to address the class of problems where $\mathbf{Z}_1, \ldots, \mathbf{Z}_k$ are given as inputs and not $\mathbf{B}_1, \ldots, \mathbf{B}_k$, and $n_i \gg d_i$. Both requirements are fundamental to our paper. Thus, our goal is to avoid explicitly forming $\mathbf{B}_i$ or factorizing it (e.g., Cholesky decomposition) computations that require $O(nd^2)$ operations. We aim to design algorithms that use $o(nd^2)$ operations to find the preconditioner, and use $O(nd)$ operations per iteration, where $n = \max_i n_i$ and $d = \max_i d_i$.

Two important unsupervised machine learning methods that reduce to Problem (1) are canonical correlation analysis (CCA) and Fisher linear discriminant analysis (FDA). We illustrate our approach, demonstrating its effectiveness both theoretically and empirically, on both of these problems. In particular, we improve on the $\Theta(nd^2)$ running time possible for both problems using direct methods. Moreover, $\Theta(nd^2)$ is also the cost of forming $\mathbf{B}_i$, which is one of the reasons for prohibiting the explicit formation of $\mathbf{B}_i$ in our setting. For CCA and FDA, the proposed approach can be viewed as a novel way of solving an old problem.

The underlying iterative methods we precondition are based on the framework of *Riemannian optimization* (Absil et al., 2008; Boumal, 2022). Riemannian optimization is well suited for problems with manifold constraints, e.g., under generalized orthogonality constraints (Edelman et al., 1998). It makes use of the Riemannian geometry components associated with the constraining manifolds, which in the case of Eq. (1) are products of generalized Stiefel manifolds. *Riemannian Preconditioning* (Mishra and Sepulchre, 2016), which

is a technique for preconditioning Riemannian optimization algorithms based on carefully choosing the Riemannian metric is another component of our proposed method. By combining randomized preconditioning with Riemannian preconditioning we obtain randomized preconditioners and faster methods for solving Eq. (1). Specifically, we propose a constant randomized preconditioning scheme $\mathbf{X} \mapsto \mathbf{M}_{\mathbf{X}} \coloneqq \mathbf{M}$ that is an SPD matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$ which defines the Riemannian metric on $\mathbf{St}_{\mathbf{B}}(p,d)$ (see Section 3).

Remark 1 *We note that our proposed preconditioner is based on the search space, as we assume that the main computational burden lies in forming* $\mathbf{B}_i$*. Nevertheless, our method can be combined with the approach presented in (Mishra and Sepulchre, 2016), i.e., also include the a cost function dependent component, which would make the preconditioner non-constant. We leave this for future work, and in particular one example of this kind of preconditioner can be found in (Mor and Avron, 2020).*

1.1 Contributions

In this subsection, we emphasize our main contributions. In this paper, we propose to expand the application of randomized preconditioning to optimization problems with (generalized) orthogonality constraints. In order to achieve that goal we use the framework of Riemannian optimization (Absil et al., 2008; Boumal, 2022), and perform Riemannian preconditioning via Riemannian metric selection (Mishra and Sepulchre, 2016). Thus, we utilize the Riemannian components of the generalized Stiefel manifold developed in (Shustin and Avron, 2023).

- Our main contribution, presented in Section 3, is designing a constant randomized preconditioning scheme which is incorporated via the Riemannian metric for the generalized Stiefel manifold (and products of it) with $\mathbf{B}_i = \mathbf{Z}_i^{\mathrm{T}}\mathbf{Z}_i + \lambda\mathbf{I}_d$ where $\mathbf{Z}_i$ are given. Moreover, we exemplify a specific sketching transformation, CountSketch (Charikar et al., 2004), list the corresponding computational costs of the geometric components required for sketching based Riemannian preconditioning in Table 2 (we also provide costs for the *Subsampled Randomized Hadamard Transform* (SRHT) (Tropp, 2011; Boutsidis and Gittens, 2013; Clarkson and Woodruff, 2017)), and present Lemma 2 (and its analogue for SRHT, Lemma 22), which would later aid in assessing the effectiveness of our proposed preconditioning.

- In sections 4 and 5, we exemplify our approach by developing preconditioned iterative methods for CCA and FDA correspondingly. We identify the corresponding critical points, and analyze their stability. We theoretically analyze the effect of a generic preconditioner on the asymptotic convergence rate, and identify the optimal preconditioner. Then, we present end-to-end sketching based algorithms for CCA and FDA, analyze their asymptotic convergence and evaluate the computational costs. Finally, we demonstrate numerically our randomized preconditioning approach in Section 6.

1.2 Related Work

We begin with the related work on **matrix sketching**. The literature on sketching has so far mostly focused on the sketch-and-solve approach. In particular, in the context of solving problems with generalized orthogonality constraints, a sketch-and-solve based approach for

CCA was developed by Avron et al. (2014a). Sketch preconditioning was predominantly applied to linear least-squares regression problems (e.g., Avron et al., 2010; Meng et al., 2014), and also to non least-squares variants (e.g., ℓ_1-regression by Meng and Mahoney, 2013). For a regularized version of FDA, a randomized iterative method was developed by Chowdhury et al. (2019). Even though the method of Chowdhury et al. (2019) does not use sketch preconditioning per se, it can be viewed as a preconditioned Richardson iteration. A randomized iterative method for solving LP problems was proposed by Chowdhury et al. (2020), where sketch preconditioning is used to find the Newton search direction as part of an interior point method. To the best of our knowledge, sketch preconditioning has neither been applied to CCA before nor has it been used in the context of Riemannian optimization. We remark, that one approach to solving CCA is via alternating least squares (ALS) presented by Golub and Zha (1995) (see Algorithm 5.2 in their paper). Thus, one natural idea is to use randomized preconditioning for least squares for each of the sub-problems in ALS approach. While a series of least squares solved exactly is equivalent to a specific choice of Riemannian algorithm, and a prescribed step size, our preconditioning approach allows for a wide variety of algorithms (e.g., first, and second-order methods). In addition, solving a series of least squares requires to solve each of the sub-problems, whereas in our approach one does not need to solve sub-problems. Finally, on a conceptual level our approach aims to precondition the problem directly, rather than performing a two-level solution of replacing the problem with a series of sub-problems, and then preconditioning these.

Next, we recall the related work on **Riemannian optimization**. Recent works introducing Riemannian optimization are (Absil et al., 2008; Boumal, 2022). Earlier works are (Luenberger, 1972; Gabay, 1982; Smith, 1994), and specifically, using Riemannian optimization to solve problems under orthogonality constraints is presented in the seminal work of Edelman et al. (1998).

In particular, we mention related work on **Riemannian preconditioning**. Preconditioning of Riemannian optimization methods based on the cost function alone is presented in several works (see e.g., Ngo and Saad, 2012; Mishra and Sepulchre, 2014; Shi et al., 2016; Zhou et al., 2017). Most of the aforementioned works attempt to perform preconditioning by approximating the Euclidean Hessian of the cost function. However, the Riemannian Hessian and the Euclidean Hessian are quite different even for simple examples (see Shustin and Avron, 2023, Section 3.5). The Riemannian Hessian is related to the Hessian of the Lagrangian, and indeed in (Mishra and Sepulchre, 2016) it is shown that selecting the Riemannian metric inspired by the Hessian of the Lagrangian affects convergence of Riemannian steepest-descent, coining the term *Riemannian preonditioning* for judiciously choosing the Riemannian metric in order to accelerate convergence. Unlike in (Mishra and Sepulchre, 2016), we present a randomized preconditioning strategy, and analyze the condition number of the Riemannian Hessian at the optimum for specific examples (CCA and FDA), which allows us to quantify the quality of the proposed preconditioner. Recently, a new paper, (Gao et al., 2023), that appeared after our paper was sent for initial review, presented a preconditioning technique for product manifolds by defining the Riemannian metric via the diagonal blocks of the Riemannian Hessian at the optimum. In addition, Gao et. al analyzed the effect of preconditioning via the condition number of the Riemannian Hessian at the optimum, in a similar manner to the analysis in our paper. They demonstrated their technique on CCA, truncated singular value decomposition, and matrix and tensor com-

pletion problems. Unlike in (Gao et al., 2023), we present a randomized preconditioning strategy via matrix sketching.

Additional works regarding preconditioning of Riemannian methods are a work by Kressner et al. (2016) where they proposed a Riemannian analogue to the preconditioned Richardson method, and the works (Baker, 2008, Algorithm 5) and (Vandereycken and Vandewalle, 2010) where a Riemannian version of preconditioning the trust-region sub-problem was proposed. To some extent that work can be also viewed as Riemannian metric selection. The aforementioned preconditioning approaches are essentially two-stage approaches, in which a specific choice of a Riemannian algorithm is made, and then a preconditioning is applied specific to this method. In contrast, our approach aims to allow a wider choice of algorithms (zero, first, and second-order methods), without the need to solve sub-problems. Moreover, from a conceptual approach, we aim to precondition the problem itself, regardless of the algorithm at hand, and avoid two-stage procedures.

Another work on Riemannian preconditioning for the trust-region sub-problem is (Mor and Avron, 2020). We remark that apart from the last work mentioned, no previous work proposed to use randomized preconditioners in the context of Riemannian preconditioning.

Finally, we also recall related work on **iterative methods for CCA and FDA**. Several iterative methods for solving CCA have been proposed in the literature. Golub and Zha (1995) presented an iterative method for CCA based on ALS. Each iteration requires the solution of two least squares problems. The authors suggest using LSQR for that task. Wang et al. (2016) proposed to replace LSQR with either accelerated gradient descent, stochastic variance reduce gradient (SVRG) or accelerated SVRG. They also proposed a different approach based on shift-and-invert preconditioning. Ma et al. (2015) developed an algorithm for CCA based on augmented approximate gradients. Ge et al. (2016b) provided an iterative algorithm for the generalized eigenvalue problem, and used a standard reduction of CCA to generalized eigenvalue problems to derive an algorithm for CCA. They assume a fast black box access to an approximate linear system solver.

Convergence bounds of all the aforementioned algorithms depend on the condition number of the input matrices, which might be large. In contrast, the condition number bounds for our proposed sketching based algorithms are independent of the conditioning of the input matrices. As aside, (Yger et al., 2012) proposed a Riemannian method for adaptive CCA.

Recently, Chapman et al. (2024) proposed a stochastic method to solve large-scale CCA problems. Their method is based on unifying multiview CCA and self-supervised learning. To make the method feasible for large-scale problems a family of fast stochastic algorithms was proposed for a novel unconstrained objective function for generalized eigenvalue problems (which CCA generalizes).

In the context of FDA, a recent work by Chowdhury et al. (2019) proposed an iterative, sketching-based algorithm for regularized FDA.

Note that both CCA and FDA are closely related to the singular value decomposition (SVD) and generalized eigenvalue problems correspondingly. In the context of optimization on manifolds Helmke and Moore (1992) and Brockett (1991) did an analysis of SVD and generalized eigenvalue problems correspondingly, as dynamical systems, i.e., gradient flows on manifolds (of unitary matrices). In particular, the stability of the equilibrium point was established, which we use in Subsections 4.3 and 5.3.

2. Preliminaries

In this section, we present notation and recall preliminaries on Riemannian optimization and preconditioning.

2.1 Notation and Basic Definitions

Scalars are denoted by lower case letters $\alpha, \beta, ..., x, y, \ldots$, while vectors are denoted by bold letters $\boldsymbol{\alpha}, \boldsymbol{\beta}, ..., \boldsymbol{x}, \boldsymbol{y}, \ldots$, and matrices are denoted by bold uppercase English letters $\mathbf{A}, \mathbf{B} \ldots$ or upper case Greek letters. We use the convention that vectors are column-vectors. We describe a diagonal matrix using $\mathbf{diag}(\cdot)$, and a block diagonal matrix using $\mathbf{blkdiag}(\cdot)$.

We denote by $(\cdot,\cdot)_{\mathbf{M}}$ the inner-product with respect to a matrix $\mathbf{M}$: for $\mathbf{U}$ and $\mathbf{V}$, $(\mathbf{U},\mathbf{V})_{\mathbf{M}} \coloneqq \mathbf{Tr}\left(\mathbf{U}^{\mathrm{T}}\mathbf{M}\mathbf{V}\right)$ where $\mathbf{Tr}(\cdot)$ denotes the trace operator. The $s \times s$ identity matrix is denoted $\mathbf{I}_s$. The $s \times s$ zero matrix is denoted $\mathbf{0}_s$. We denote by $\mathcal{S}_{\mathrm{sym}}(p)$ and $\mathcal{S}_{\mathrm{skew}}(p)$ the set of all symmetric and skew-symmetric matrices (respectively) in $\mathbb{R}^{p\times p}$. The symmetric and skew-symmetric components of $\mathbf{A}$ are denoted by $\mathbf{sym}(\mathbf{A}) \coloneqq \left(\mathbf{A}+\mathbf{A}^{\mathrm{T}}\right)/2$ and $\mathbf{skew}(\mathbf{A}) \coloneqq \left(\mathbf{A}-\mathbf{A}^{\mathrm{T}}\right)/2$ respectively.

For a SPD matrix $\mathbf{B} \in \mathbb{R}^{d\times d}$, we denote by $\mathbf{B}^{1/2}$ the unique SPD matrix such that $\mathbf{B} = \mathbf{B}^{1/2}\mathbf{B}^{1/2}$, obtained by keeping the same eigenvectors and taking the square root of the eigenvalues. We denote the inverse of $\mathbf{B}^{1/2}$ by $\mathbf{B}^{-1/2}$.

The eigenvalues of a symmetric $d \times d$ matrix $\mathbf{A}$ are denoted by $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots \geq \lambda_d(\mathbf{A})$. The condition number of $\mathbf{A}$, i.e. the ratio between the largest and smallest eigenvalues in absolute value, is denoted by $\kappa(\mathbf{A})$. The generalized condition number of a matrix pencil $(\mathbf{A},\mathbf{B})$, which is the ratio between the largest and the smallest generalized eigenvalues, is denoted by $\kappa(\mathbf{A},\mathbf{B})$, and it holds that if $\mathbf{B}$ is a SPD matrix, then $\kappa(\mathbf{A},\mathbf{B}) = \kappa(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2})$.

For differential geometry objects we use the following notations: tangent vectors (of a manifold) are denoted using lower case Greek letters with a subscript indicating the point of the manifold for which they correspond (e.g., $\eta_{\mathbf{x}}$), while normal vectors (of a manifold) are denoted using bold lower and upper case letters with a subscript (e.g., $\mathbf{u}_{\mathbf{x}}$). We denote by $\mathbf{St}_{\mathbf{B}}(p,d)$ the *generalized Stiefel manifold*, which is a submanifold of $\mathbb{R}^{d\times p}$ defined by

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

Given a function defined on $\mathbf{St}_{\mathbf{B}}(p,d)$, a smooth extension of that function to the entire $\mathbb{R}^{d\times p}$ is denoted by a bar decorator. For example, given a smooth function $f : \mathbf{St}_{\mathbf{B}}(p,d) \to \mathbb{R}$, we use $\bar{f} : \mathbb{R}^{d\times p} \to \mathbb{R}$ to denote a smooth function on $\mathbb{R}^{d\times p}$ such that on $\mathbf{St}_{\mathbf{B}}(p,d)$ that function agrees with f.

2.2 Riemannian Optimization and Preconditioning for the Generalized Stiefel Manifold

In this subsection we detail several necessary ingredients required for preconditioning strategy: Riemannian optimization, Riemannian preconditioning, and the geometric optimization components of the generalized Stiefel manifold with a non-standard metric. In following sections we make use of these components to present our randomized Riemannian preconditioning strategy.

Riemannian optimization (Absil et al., 2008; Boumal, 2022) is a framework of designing algorithms for solving constrained optimization problems, where the constraints form a smooth manifold $\mathcal{M}$. The general idea of these algorithms is to make use of the differential geometry components of $\mathcal{M}$ in order to generalize iterative methods for unconstrained optimization. Iterative algorithms for smooth problems such as gradient methods, trust-region, and conjugate gradient (CG), are adapted to the Riemannian setting using the following components: a *Riemannian metric* which is a smoothly varying inner product $\mathbf{x} \mapsto g_{\mathbf{x}}$ on the tangent bundle $T\mathcal{M}$ such that $(\mathcal{M}, g)$ becomes a Riemannian manifold, a *retraction mapping* $R_{\mathbf{x}} : T_{\mathbf{x}}\mathcal{M} \to \mathcal{M}$ which allows to take a step at point $\mathbf{x} \in \mathcal{M}$ in a direction $\xi_{\mathbf{x}} \in T_{\mathbf{x}}\mathcal{M}$, a *vector transport* $\mathcal{T}_{\eta_{\mathbf{x}}} : T_{\mathbf{x}}\mathcal{M} \to T_{R_{\mathbf{x}}(\eta_{\mathbf{x}})}\mathcal{M}$ which allows operations between tangent vectors from two different tangent spaces, a *Riemannian gradient* $\mathbf{grad} f(\mathbf{x}) \in T_{\mathbf{x}}\mathcal{M}$, and a *Riemannian Hessian* $\mathbf{Hess} f(\mathbf{x}) : T_{\mathbf{x}}\mathcal{M} \to T_{\mathbf{x}}\mathcal{M}$. Another important notion in the context of this paper is the notion of Riemannian submanifold, which allows to easily compute the aforementioned geometric components if they are known for the ambient manifold. Usually, a Riemannian optimization algorithm is built from iterations on the tangent space $T_{\mathbf{x}}\mathcal{M}$ which are then retracted to the manifold. For example, Riemannian gradient methods are given by the formula $\mathbf{x}_{k+1} = R_{\mathbf{x}_k}(\tau_k \mathbf{grad} f(\mathbf{x}_k))$ where τ_k is the step size. Many of the Riemannian algorithms and common manifolds are implemented in MANOPT, which is a MATLAB library (Boumal et al., 2014). There is also a PYTHON parallel for MANOPT called PYMANOPT (Townsend et al., 2016). The experiments reported in Section 6 use the MANOPT library.

Preconditioning of iterative methods is often challenging since it is not initially clear how to actually precondition the problem for an iterative method. Riemannian preconditioning (Mishra and Sepulchre, 2016) generalizes preconditioning for Riemannian optimization methods. In Riemannian preconditioning, preconditioning is performed via a Riemannian metric selection, i.e., the preconditioner is incorporated via the Riemannian metric.

A motivation for metric selection is the observation that the condition number of the Riemannian Hessian at the optimum affects the asymptotic convergence rate of Riemannian optimization, e.g., (Absil et al., 2008, Theorem 4.5.6, Theorem 7.4.11 and Equation 7.50), and (Boumal, 2022, Section 4.6). In the case of convex objective function in the Riemannian sense (Udriste, 2013, Chapter 3.2) there are also global convergence results (e.g., Udriste, 2013, Chapter 7, Theorem 4.2). Thus, selecting a metric such that this condition number is lowered should improve convergence.

In this paper, we propose to utilize randomized preconditioning to accelerate the solution of orthogonality constrained problems by using Riemannian preconditioning. Optimization problems under generalized orthogonality constraints can be written as Riemannian optimization problems on the generalized Stiefel manifold. The standard Riemannian metric for the generalized Stiefel manifold (e.g., Edelman et al., 1998; Boumal et al., 2014) is

$$g_{\mathbf{X}}(\xi_{\mathbf{X}}, \eta_{\mathbf{X}}) \coloneqq (\xi_{\mathbf{X}}, \eta_{\mathbf{X}})_{\mathbf{B}} = \mathbf{Tr}\left(\xi_{\mathbf{X}}^{\mathrm{T}} \mathbf{B} \eta_{\mathbf{X}}\right).$$

The use of this metric has two possible shortcomings. First, it might not be the optimal metric with respect to convergence of Riemannian optimization methods. Second, and more relevant for this paper, the computation of most Riemannian components requires taking products of $\mathbf{B}^{-1}$ with a vector. Computing $\mathbf{B}$ and/or factorizing is in many cases as costly

Table 1: Summary of the Preconditioned various Riemannian components of the generalized Stiefel manifold (Shustin and Avron, 2023), i.e., a generalized Stiefel manifold with a Riemannian metric defined via an SPD matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$. Note that, the matrix $\mathbf{S}_{\xi_\mathbf{X}}$ denotes a solution for a Sylvester equation, arising from the orthogonal projection of $\xi_\mathbf{X}$ on the tangent space.

Component	**Formula**
Tangent space $T_\mathbf{X}\mathbf{St}_\mathbf{B}(p,d)$	$\left\{\mathbf{Z} \in \mathbb{R}^{d\times p} \,:\, \mathbf{Z}^\mathrm{T}\mathbf{B}\mathbf{X} + \mathbf{X}^\mathrm{T}\mathbf{B}\mathbf{Z} = \mathbf{0}_p\right\}$
Tangent space $T_\mathbf{X}\mathbf{St}_\mathbf{B}(p,d)$ alternative form	$\left\{\mathbf{Z} = \mathbf{X}\boldsymbol{\Omega} + \mathbf{X}_{\mathbf{B}\perp}\mathbf{K} \in \mathbb{R}^{d\times p} \,:\, \boldsymbol{\Omega} = -\boldsymbol{\Omega}^\mathrm{T},\ \mathbf{K} \in \mathbb{R}^{(d-p)\times p}\right\}$ such that $\mathbf{X}_{\mathbf{B}\perp}^\mathrm{T}\mathbf{B}\mathbf{X}_{\mathbf{B}\perp} = \mathbf{I}_{d-p},\ \mathbf{X}_{\mathbf{B}\perp}^\mathrm{T}\mathbf{B}\mathbf{X} = \mathbf{0}_{(d-p)\times p}$
The Riemannian metric	$g_\mathbf{X}(\xi_\mathbf{X}, \eta_\mathbf{X}) = (\xi_\mathbf{X}, \eta_\mathbf{X})_\mathbf{M} = \mathbf{Tr}\left(\xi_\mathbf{X}^\mathrm{T}\mathbf{M}\eta_\mathbf{X}\right)$
Normal space with respect to the Riemannian metric	$(T_\mathbf{X}\mathrm{St}_\mathbf{B}(p,d))^\perp = \left\{\mathbf{M}^{-1}\mathbf{B}\mathbf{X}\mathbf{S} \,:\, \mathbf{S} = \mathbf{S}^\mathrm{T}\right\}$
Polar-based retraction	$R_\mathbf{X}^\mathrm{polar}(\xi_\mathbf{X}) = (\mathbf{X} + \xi_\mathbf{X})(\mathbf{I}_p + \xi_\mathbf{X}^\mathrm{T}\mathbf{B}\xi_\mathbf{X})^{-1/2}$
QR-based retraction	$R_\mathbf{X}^\mathrm{QR}(\xi_\mathbf{X}) = \mathbf{B}^{-1/2}\mathbf{qf}\left(\mathbf{B}^{1/2}\left(\mathbf{X} + \xi_\mathbf{X}\right)\right)$
Orthogonal projection on the normal space	$\boldsymbol{\Pi}_\mathbf{X}^\perp\left(\xi_\mathbf{X}\right) = \mathbf{M}^{-1}\mathbf{B}\mathbf{X}\mathbf{S}_{\xi_\mathbf{X}}$
Orthogonal projection on the tangent space	$\boldsymbol{\Pi}_\mathbf{X}\left(\xi_\mathbf{X}\right) = \left(\mathbf{id}_{T_\mathbf{X}\mathrm{St}_\mathbf{B}(p,d)} - \boldsymbol{\Pi}_\mathbf{X}^\perp\right)\left(\xi_\mathbf{X}\right) = \xi_\mathbf{X} - \mathbf{M}^{-1}\mathbf{B}\mathbf{X}\mathbf{S}_{\xi_\mathbf{X}}$
Sylvester equation for $\mathbf{S}_{\xi_\mathbf{X}}$ (orthogonal projection)	$\left(\mathbf{X}^\mathrm{T}\mathbf{B}\mathbf{M}^{-1}\mathbf{B}\mathbf{X}\right)\mathbf{S}_{\xi_\mathbf{X}} + \mathbf{S}_{\xi_\mathbf{X}}\left(\mathbf{X}^\mathrm{T}\mathbf{B}\mathbf{M}^{-1}\mathbf{B}\mathbf{X}\right) = \mathbf{X}^\mathrm{T}\mathbf{B}\xi_\mathbf{X} + \left(\mathbf{X}^\mathrm{T}\mathbf{B}\xi_\mathbf{X}\right)^\mathrm{T}$
Vector transport	$\tau_{\eta_\mathbf{X}}\xi_\mathbf{X} = \boldsymbol{\Pi}_{R_\mathbf{X}(\eta_\mathbf{X})}\left(\xi_\mathbf{X}\right)$
Riemannian gradient	$\mathbf{grad}f(\mathbf{X}) = \boldsymbol{\Pi}_\mathbf{X}\left(\mathbf{M}^{-1}\nabla\bar{f}(\mathbf{X})\right)$
Riemannian Hessian applied on a tangent vector	$\mathbf{Hess}f(\mathbf{X})[\eta_\mathbf{X}] =$ $\boldsymbol{\Pi}_\mathbf{X}(\mathbf{M}^{-1}(\nabla^2\bar{f}(\mathbf{X})\eta_\mathbf{X} - \mathbf{B}\eta_\mathbf{X}(\mathbf{X}^\mathrm{T}\nabla\bar{f}(\mathbf{X}) - \mathbf{X}^\mathrm{T}\mathbf{M}\mathbf{grad}f(\mathbf{X}))))$

as solving the problem directly or even solving it via standard iterative methods, e.g., for CCA (see Section 4).

In order to overcome these shortcomings, we leverage our recently developed geometric components for the generalized Stiefel manifold with a Riemannian metric which is based on a preconditioning scheme $\mathbf{X} \mapsto \mathbf{M}_\mathbf{X}$ that maps a $\mathbf{X} \in \mathbf{St}_\mathbf{B}(p,d)$ to an SPD matrix $\mathbf{M}_\mathbf{X} \in \mathbb{R}^{d\times d}$, instead using the standard metric defined by $\mathbf{B}$ (Shustin and Avron, 2023). In this paper, we apply a preconditioning scheme independent of $\mathbf{X}$, i.e., $\mathbf{M}_\mathbf{X} \coloneqq \mathbf{M}$ for all $\mathbf{X} \in \mathbf{St}_\mathbf{B}(p,d)$ (see Section 3). As a reference we summarize (without proofs) in Table 1 the various geometric components developed in (Shustin and Avron, 2023).

In addition, we also address optimization problems that are constrained on a product of generalized Stiefel manifolds such as CCA. In this case, the Riemannian components of the generalized Stiefel manifold can be generalized to be the Riemannian components of a product of generalized Stiefel manifolds in a straightforward way (Shustin and Avron, 2023).

3. Randomized Riemannian Preconditioning

In this section we present our main contribution: randomized Riemannian preconditioners for optimization problems which feature generalized orthogonality constraints defined by Gram matrices. That is, our goal is to address constraints of the form $\mathbf{X}_i^{\mathrm{T}}\mathbf{B}_i\mathbf{X}_i = \mathbf{I}_p \quad (i = 1, \ldots, k)$, where $\mathbf{B}_i$ are not given explicitly but can be written as a regularized Gram matrix of a given tall-and-skinny matrix $\mathbf{Z}_i \in \mathbb{R}^{n_i \times d_i}$, $n_i \gg d_i$, with a regularization parameter $\lambda \geq 0$, i.e., $\mathbf{B}_i = \mathbf{Z}_i^{\mathrm{T}}\mathbf{Z}_i + \lambda\mathbf{I}_d$. As alluded to earlier, our proposed solution incorporates a preconditioner by selecting a non-standard metric for $\mathbf{St}_{\mathbf{B}_1} \times \cdots \times \mathbf{St}_{\mathbf{B}_k}$.

For simplicity, let us focus our exposition on the case of $k = 1$ and drop the subscript from $\mathbf{B}_1, \mathbf{Z}_1$ etc. Generalization to an arbitrary k is straightforward. We define the metric on $\mathbf{St}_{\mathbf{B}}(p, d)$ using a constant preconditioning scheme $\mathbf{M}_{\mathbf{X}} \coloneqq \mathbf{M}$ for all $\mathbf{X} \in \mathbf{St}_{\mathbf{B}}(p, d)$ formed from sketching $\mathbf{Z}$ prior to computing the Gram matrix. Our proposed construction of $\mathbf{M}$ requires $o(nd^2)$ operations, which is cheaper than computing $\mathbf{B}$ as required when the standard metric is used.

Our randomized construction of $\mathbf{M}$ is based on the following observation: in many cases, if $\mathbf{M}$ approximates $\mathbf{B}$ in the sense that the condition number $\kappa(\mathbf{B}, \mathbf{M})$ is small, then convergence will be fast. Conceptually, the last observation is synonymous with the observation that usually the standard metric is a good choice iteration complexity-wise, albeit a computationally expensive choice, and thus we should aim at cheaply approximating it. Mathematically, the underlying reason is that in many cases the condition number of the Riemannian Hessian at the optimum is bounded by the product of $\kappa(\mathbf{B}, \mathbf{M})$ and a problem/input dependent quantity which is typically small. It is known in the Riemannian optimization literature that the local convergence rate can be analyzed by inspecting the condition number of the Riemnnian Hessian at the optimum, see (Absil et al., 2008, Section 4.1 and Theorem 4.5.6) and (Boumal, 2022, Section 4.6).

We demonstrate the observation that when $\mathbf{M}$ approximates $\mathbf{B}$ well, then the Riemannian Hessian at the optimum is well conditioned on two important use-cases: CCA and FDA (see Theorems 9 and 18). Indeed, we show that under certain assumptions, both for CCA and FDA we can bound the condition number at the optimum by the product of some baseline condition number that depends on eigengaps, and the condition number of $\kappa(\mathbf{B}, \mathbf{M})$.

We propose to construct $\mathbf{M}$ using the technique of sketching (Woodruff, 2014). Let $\mathbf{S} \in \mathbb{R}^{s \times n}$ be some *sketching matrix* (a certain distribution on matrices; we discuss a concrete choice in the next paragraph), where $s < n$. We then use $\mathbf{M} = \mathbf{Z}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Z} + \lambda\mathbf{I}_d$. However, we do not propose to actually compute $\mathbf{M}$. Using the metric defined by $\mathbf{M}$ requires only taking products with $\mathbf{M}$ and $\mathbf{M}^{-1}$ (see Table 1). Suppose we have already computed $\mathbf{SZ}$. The product of $\mathbf{M}$ with a vector can be computed in $O(\mathbf{nnz}(\mathbf{SZ}))$ operations. As for $\mathbf{M}^{-1}$, by computing a QR factorization of $\begin{bmatrix} \mathbf{SZ} \\ \sqrt{\lambda}\mathbf{I}_d \end{bmatrix}$ ($O(sd^2)$ operations), we can obtain a Cholesky factorization of $\mathbf{M}$, and then taking the product of $\mathbf{M}^{-1}$ with a vector requires $O(d^2)$. So our goal is to design sketching matrices $\mathbf{S}$ such that $\mathbf{SZ}$ is cheap to compute, and for which $\kappa(\mathbf{B}, \mathbf{M})$ is bounded by a constant.

There are quite a few sketching distributions proposed in the literature, and most of them are good choices for $\mathbf{S}$. For concreteness, we describe a specific choice: the CountSketch[1] transformation (Charikar et al., 2004). CountSketch is specified by a random hash function $h : \{1, \ldots, d\} \to \{1, \ldots, s\}$ and random sign function $g : \{1, \ldots, d\} \to \{-1, +1\}$. Applying $\mathbf{S}$ to a vector $\mathbf{x} \in \mathbb{R}^n$ is given by the formula

$$(\mathbf{S}\mathbf{x})_i = \sum_{j|h(j)=i} g(j)x_j \,.$$

It is easy to see that $\mathbf{S}$ is a random matrix where the jth column contains a single nonzero entry $g(j)$ in the $h(j)$th row. Clearly, $\mathbf{SZ}$ can be computed using $O(\mathbf{nnz}\,(\mathbf{Z})) = O(nd)$ arithmetic operations. Thus, it only remains to bound the condition number. The following lemma shows that if the sketch size is large enough, then with high probability the condition number is bounded by a constant.

Lemma 2 *Assume that $\lambda > 0$ or that $\mathbf{Z} \in \mathbb{R}^{n\times d}$ has full column rank. Let $s_\lambda(\mathbf{Z}) \coloneqq \mathbf{Tr}\left((\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \lambda\mathbf{I})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)$. Suppose that $\mathbf{S} \in \mathbb{R}^{s\times n}$ is a CountSketch matrix with $s \geq 20s_\lambda(\mathbf{Z})^2/\delta$ for some $\delta \in (0, 1)$. Then with probability of at least $1 - \delta$ we have that all the generalized eigenvalues of the pencil $(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}+\lambda\mathbf{I}, \mathbf{Z}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Z}+\lambda\mathbf{I})$ are contained in the interval $[1/2, 3/2]$ and $\kappa(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \lambda\mathbf{I}, \mathbf{Z}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Z} + \lambda\mathbf{I}) \leq 3$.*

Proof The argument is rather standard and appeared in similar forms in the literature. To show that the generalized eigenvalues of the pencil $(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}+\lambda\mathbf{I}, \mathbf{Z}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Z}+\lambda\mathbf{I})$ are contained in the interval $[1/2, 3/2]$ and that $\kappa(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \lambda\mathbf{I}, \mathbf{Z}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Z} + \lambda\mathbf{I}) \leq 3$ to hold, it is enough to show that

$$\frac{1}{2}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \lambda\mathbf{I}) \preceq \mathbf{Z}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Z} + \lambda\mathbf{I} \preceq \frac{3}{2}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \lambda\mathbf{I})\,.$$

Let $\mathbf{Z} = \mathbf{Q}\mathbf{R}$ be a λ-QR factorization of $\mathbf{Z}$, i.e., $\mathbf{Q}$ is a full-rank matrix and $\mathbf{R}$ is upper triangular such that $\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \lambda\mathbf{I}$ (Avron et al., 2017, Definition 28). Note that such a factorization always exists (Avron et al., 2017, remark following Definition 28). Left-multiplying by $\mathbf{R}^{-\mathrm{T}}$ and right-multiplying by $\mathbf{R}^{-1}$ on both sizes, we find it suffices to show that with probability of at least $1 - \delta$ we have

$$\frac{1}{2}\mathbf{I}_d \preceq \mathbf{Q}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Q} + \lambda\mathbf{R}^{-\mathrm{T}}\mathbf{R}^{-1} \preceq \frac{3}{2}\mathbf{I}_d$$

or, equivalently,

$$\|\mathbf{Q}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Q} + \lambda\mathbf{R}^{-\mathrm{T}}\mathbf{R}^{-1} - \mathbf{I}_d\|_2 \leq \frac{1}{2}\,.$$

Since $\mathbf{Q}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Q} + \lambda\mathbf{R}^{-\mathrm{T}}\mathbf{R}^{-1} - \mathbf{I}_d = \mathbf{Q}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Q} - \mathbf{Q}^{\mathrm{T}}\mathbf{Q}$ (Avron et al., 2017, Fact 29) and the spectral norm is dominated by the Frobenius norm, it is enough to show that

$$\|\mathbf{Q}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Q} - \mathbf{Q}^{\mathrm{T}}\mathbf{Q}\|_F \leq \frac{1}{2}\,.$$

1. Note that SRHT is also a good choice, in particular for dense data sets. In Appendix A, we provide a similar analysis for SRHT.

It is known (Avron et al., 2014b, Lemma 2) that for any two fixed matrices $\mathbf{A}$ and $\mathbf{B}$, and a CountSketch matrix $\mathbf{S}_0$ with $m \geq 5/(\epsilon^2\delta)$ rows, we have that with probability of at least $1-\delta$,

$$\|\mathbf{A}^{\mathrm{T}}\mathbf{S}_0^{\mathrm{T}}\mathbf{S}_0\mathbf{B} - \mathbf{A}^{\mathrm{T}}\mathbf{B}\|_F \leq \epsilon \cdot \|\mathbf{A}\|_F \cdot \|\mathbf{B}\|_F .$$

Since $\|\mathbf{Q}\|_F^2 = s_\lambda(\mathbf{Z})$ (Avron et al., 2017, Fact 30), then with $s \geq 20 s_\lambda(\mathbf{Z})^2/\delta$ we have

$$\|\mathbf{Q}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Q} - \mathbf{Q}^{\mathrm{T}}\mathbf{Q}\|_F \leq \frac{1}{2}$$

with probability of at least $1-\delta$. ■

The last lemma justifies the use of CountSketch to form the preconditioner $\mathbf{M}$. In practice, additional heuristics that improve running time and robustness can be inserted into the construction of randomized preconditioners, and these can improve running time considerably; see (Avron et al., 2010).

Furthermore, our sketching-based preconditioner construction naturally allows for a warm-start. While this is not captured by our theory, heuristically (and empirically) the Riemannian optimization part of our proposed algorithm converges faster if the starting vectors are close to the optimum. Our sketching approach lets us quickly compute good starting vectors (these are the sketch-and-solve approximations). We demonstrate warm-start numerically in Section 6 for CCA and FDA.

In Table 2 we detail the computational cost, measured in terms of arithmetic operations, of computing the Riemannian components of Table 1, for our construction of $\mathbf{M}$ as a preconditioner. Table 2 is based on (Shustin and Avron, 2023, Table 1). Note that all the costs are for operations in ambient coordinates. In the table, 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. Instead of committing to a specific sketch size, we use s for sketch size (number of rows), and consider two possible sketching distributions: CountSketch and SRHT.

To put Table 2 in the contexts of the total computation complexities, we can use the results presented in (Boumal et al., 2019). For example, Riemannian gradient descent with a fixed step-size (which requires a Riemannian gradient and a retraction computation per iteration) takes at most $O(\varepsilon^{-2})$ iterations to reach to an ε-approximate first-order KKT point (Boumal et al., 2019, Theorem 2.5). Using our proposed preconditioned components, preprocessing takes $O(\mathbf{nnz}\,(\mathbf{Z}) + sd^2)$ operations for CountSketch, and $O(ndlog(s) + sd^2)$ operations for SRHT. On the other hand, using the standard metric would require $O(nd^2)$ operations in preprocessing. In addition, computing a Riemannian gradient and a retraction is unaffected of the choice of the Riemannian metric after preprocessing is done. Thus, in the worst-case scenario, using our preconditioned components for a fixed-step Riemannian gradient descent would result in a total of $O(\mathbf{nnz}\,(\mathbf{Z}) + sd^2 + \varepsilon^{-2}(ndp + d^2p + dp^2 + T_{\nabla \bar{f}}))$ operations for CountSketch, and $O(nd\log(s) + sd^2 + \varepsilon^{-2}(ndp + d^2p + dp^2 + T_{\nabla \bar{f}}))$ operations for SRHT, while using the standard metric would result in a total of $O(nd^2 + \varepsilon^{-2}(ndp + d^2p + dp^2 + T_{\nabla \bar{f}}))$ operations. As we demonstrate in Section 6, we have $s \ll n$, thus our preconditioned algorithm is expected to be faster, at least in the worst-case scenario. As for the asymptotic convergence, we demonstrate for CCA and FDA how our preconditioner affects the asymptotic convergence via analysis of the condition number of the Riemannian Hessian at the optimum in Subsection 4.5 and Subsection 5.5.

Table 2: Summary of the cost of various components for using sketching based Riemannian preconditioning for generalized orthogonality constraints.

Operation	**Cost using** CountSketch	**Cost using** SRHT
Preprocessing: computing $\mathbf{SZ}$	$O(\mathbf{nnz}(\mathbf{Z})) = O(nd)$	$O(nd\log(s))$
Preprocessing: given $\mathbf{SZ}$ forming the inverse of $\mathbf{M} = \mathbf{Z}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SZ} + \lambda\mathbf{I}_d$	$O(sd^2)$	$O(sd^2)$
Applying $\mathbf{M}^{-1}$ on a vector	$O(d^2)$	$O(d^2)$
Retraction	$O(ndp + dp^2)$	$O(ndp + dp^2)$
Inner product on the tangent space (Riemannian metric)	$O(\mathbf{nnz}(\mathbf{SZ})\,p + dp)$	$O(\mathbf{nnz}(\mathbf{SZ})\,p + dp)$
Orthogonal projections on the tangent/normal space	$O(ndp + d^2p + dp^2)$	$O(ndp + d^2p + dp^2)$
Vector Transport	$O(ndp + d^2p + dp^2)$	$O(ndp + d^2p + dp^2)$
Riemannian gradient	$O(ndp + d^2p + dp^2 + T_{\nabla\bar{f}})$	$O(ndp + d^2p + dp^2 + T_{\nabla\bar{f}})$
Applying the Riemannian Hessian to a tangent vector	$O(ndp + d^2p + \mathbf{nnz}(\mathbf{SZ})\,p + dp^2 + T_{\nabla\bar{f}} + T_{\nabla^2\bar{f}})$	$O(ndp + d^2p + \mathbf{nnz}(\mathbf{SZ})\,p + dp^2 + T_{\nabla\bar{f}} + T_{\nabla^2\bar{f}})$

Remark 3 *In Table 2 we assumed, for simplicity, that $s \geq d$. However, if λ is sufficiently large, it is possible for the prescribed values of s ($s \geq 20s_\lambda(\mathbf{Z})^2/\delta$) to be smaller than d. In such cases, we can reduce the $O(sd^2)$ term in the complexity to $O(sd\min(s,d))$ by employing the Woodbury formula. We omit the details.*

4. Sketched Iterative CCA

CCA, originally introduced by Hotelling in 1936 (Hotelling, 1936), is a well-established method in statistical learning with numerous applications (e.g., Sun et al., 2010; Kim et al., 2007; Su et al., 2012; Dhillon et al., 2012, 2011; Chaudhuri et al., 2009). In CCA, the relation between a pair of data sets in matrix form is analyzed, where the goal is to find the directions of maximal correlation between a pair of observed variables. In the language of linear algebra, CCA measures the similarities between two subspaces spanned by the columns of the two matrices, whereas in the geometric point of view, CCA computes the cosine of the principal angles between the two subspaces. We consider a regularized version of CCA defined below[2]:

Definition 4 *Let $\mathbf{X} \in \mathbb{R}^{n\times d_\mathbf{x}}$ and $\mathbf{Y} \in \mathbb{R}^{n\times d_\mathbf{y}}$ be two data matrices, and $\lambda_\mathbf{x}, \lambda_\mathbf{y} \geq 0$ be two regularization parameter. Let $q = \max\left(\mathbf{rank}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_\mathbf{x}\mathbf{I}_{d_\mathbf{x}}\right), \mathbf{rank}\left(\mathbf{Y}^{\mathrm{T}}\mathbf{Y} + \lambda_\mathbf{y}\mathbf{I}_{d_\mathbf{y}}\right)\right)$. The $(\lambda_\mathbf{x}, \lambda_\mathbf{y})$-canonical correlations $\sigma_1 \geq \cdots \geq \sigma_q$ and the $(\lambda_\mathbf{x}, \lambda_\mathbf{y})$-canonical weights $\mathbf{u}_1, \ldots, \mathbf{u}_q$, $\mathbf{v}_1, \ldots, \mathbf{v}_q$ are the arguments that maximize $\mathbf{Tr}\left(\mathbf{U}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{YV}\right)$ subject to*

$$\mathbf{U}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_\mathbf{x}\mathbf{I}_{d_\mathbf{x}})\mathbf{U} = \mathbf{I}_q,\ \mathbf{V}^{\mathrm{T}}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y} + \lambda_\mathbf{y}\mathbf{I}_{d_\mathbf{y}})\mathbf{V} = \mathbf{I}_q,\ \mathbf{U}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{YV} = \mathbf{diag}(\sigma_1, \ldots, \sigma_q),$$

2. The definition is formulated as a linear algebra problem. While the problem can be motivated, and described, in the language of statistics, the linear algebraic formulation is more convenient for our purposes.

where $\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_q \end{bmatrix} \in \mathbb{R}^{d_\mathbf{x} \times q}$ *and* $\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \dots & \mathbf{v}_q \end{bmatrix} \in \mathbb{R}^{d_\mathbf{y} \times q}$.

4.1 CCA as an Optimization Problem on a Product of Generalized Stiefel Manifolds

We focus on finding $\sigma_1, ..., \sigma_p, \mathbf{u}_1, ..., \mathbf{u}_p$ and $\mathbf{v}_1, ..., \mathbf{v}_p$, where $p \leq q$ is a parameter, i.e., on finding the top p-canonical correlations and the corresponding left and right vectors. For convenience, we use the following notations:

$$\Sigma_{\mathbf{xx}} \coloneqq \mathbf{X}^\mathrm{T}\mathbf{X} + \lambda_\mathbf{x}\mathbf{I}_{d_\mathbf{x}}, \quad \Sigma_{\mathbf{yy}} \coloneqq \mathbf{Y}^\mathrm{T}\mathbf{Y} + \lambda_\mathbf{y}\mathbf{I}_{d_\mathbf{y}}, \quad \Sigma_{\mathbf{xy}} \coloneqq \mathbf{X}^\mathrm{T}\mathbf{Y}\,.$$

With these notations, we can reformulate the problem of finding the top p-canonical correlations succinctly in the following way:

$$\max \mathbf{Tr}\left(\mathbf{U}^\mathrm{T}\Sigma_{\mathbf{xy}}\mathbf{V}\right), \quad \text{s.t.} \quad \mathbf{U}^\mathrm{T}\Sigma_{\mathbf{xx}}\mathbf{U} = \mathbf{I}_p,\ \mathbf{V}^\mathrm{T}\Sigma_{\mathbf{yy}}\mathbf{V} = \mathbf{I}_p, \\ \mathbf{U}^\mathrm{T}\Sigma_{\mathbf{xy}}\mathbf{V} \text{ is diagonal with non-increasing diagonal} \quad (2)$$

Notice that *without* the last constraint, Problem (2) is a maximization over the product of two generalized Stiefel manifolds. However, without this constraint the solution is not unique. The reason for that is that the trace operator and the constraint set are invariant to multiplication by orthonormal matrices, and so there are optimal values that are non-diagonal (and so additional steps are needed to extract the canonical correlations from such values). In order to circumvent this issue, we use a well known method to modify such problems so to make the solution unique. The modification is based on the *von Neumann cost function* (Von Neumann, 1962). That is, we replace the objective function $\mathbf{Tr}\left(\mathbf{U}^\mathrm{T}\Sigma_{\mathbf{xy}}\mathbf{V}\right)$ with $\mathbf{Tr}\left(\mathbf{U}^\mathrm{T}\Sigma_{\mathbf{xy}}\mathbf{V}\mathbf{N}\right)$ where $\mathbf{N} = \mathbf{diag}\left(\mu_1, ..., \mu_p\right)$ and we take arbitrary $\mu_1, ..., \mu_p$ such that $\mu_1 > ... > \mu_p > 0$. In other words, the problem we wish solve is

$$\max \mathbf{Tr}\left(\mathbf{U}^\mathrm{T}\Sigma_{\mathbf{xy}}\mathbf{V}\mathbf{N}\right), \ \text{s.t.}\ \mathbf{U}^\mathrm{T}\Sigma_{\mathbf{xx}}\mathbf{U} = \mathbf{I}_p,\ \mathbf{V}^\mathrm{T}\Sigma_{\mathbf{yy}}\mathbf{V} = \mathbf{I}_p.$$

In the next subsection, we detail the Riemannian components which allow to solve the CCA problem. We show that critical points of the corresponding objective function consist of coordinated left and right canonical correlation vectors not necessarily on the same phase. In particular, the optimal solutions are critical points consisting of coordinated left and right top p-canonical correlation vectors on the same phase.

4.2 Preconditioned Riemannian Components for CCA

In this subsection we derive the Riemannian components associated with the CCA problem. The CCA problem is a constraint maximization on the product of two generalized Stiefel manifolds: $\mathbf{St}_{\Sigma_{\mathbf{xx}}}(p, d_\mathbf{x})$ and $\mathbf{St}_{\Sigma_{\mathbf{yy}}}(p, d_\mathbf{y})$. We denote the search space by $\mathbb{S}_{\mathbf{xy}} \coloneqq \mathbf{St}_{\Sigma_{\mathbf{xx}}}(p, d_\mathbf{x}) \times \mathbf{St}_{\Sigma_{\mathbf{yy}}}(p, d_\mathbf{y})$. We consider the use of Riemannian optimization for solving the CCA problem, while exploiting the geometry of the preconditioned generalized Stiefel manifold and use the notion of product manifold. To make the calculations easier, we denote $d = d_\mathbf{x} + d_\mathbf{y}$ and $\mathbf{Z} \coloneqq \begin{bmatrix} \mathbf{U}^\mathrm{T} & \mathbf{V}^\mathrm{T} \end{bmatrix}^\mathrm{T} \in \mathbb{R}^{d \times p}$ where $\mathbf{U} \in \mathbf{St}_{\Sigma_{\mathbf{xx}}}(p, d_\mathbf{x})$ and $\mathbf{V} \in \mathbf{St}_{\Sigma_{\mathbf{yy}}}(p, d_\mathbf{y})$. Henceforth, we abuse notation and view $\mathbb{S}_{\mathbf{xy}}$ as a subset of $\mathbb{R}^{d \times p}$ given

by this coordinate split, and also write $\mathbf{Z} = (\mathbf{U}, \mathbf{V})$ as a shorthand for $\mathbf{Z} = \begin{bmatrix} \mathbf{U}^{\mathrm{T}} & \mathbf{V}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$. With these conventions, the optimization problem can be rewritten in the following way:

$$\min_{\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}} f_{\mathbf{CCA}}(\mathbf{Z}) \ , \ f_{\mathbf{CCA}}(\mathbf{Z}) \coloneqq -\frac{1}{2}\mathbf{Tr}\left(\mathbf{Z}^{\mathrm{T}} \begin{bmatrix} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{xy}}^{\mathrm{T}} & \end{bmatrix} \mathbf{Z}\mathbf{N}\right) \tag{3}$$

For a product of disjoint Generalized Stiefel manifolds, if the number of columns of the matrices that belong to each of the manifolds is the same, the various Riemannian components can be computed separately on each of the manifolds and then stacked on top of each other. We use two preconditioning schemes $\mathbf{U} \mapsto \mathbf{M}_{\mathbf{U}}^{(\mathbf{xx})}$ and $\mathbf{V} \mapsto \mathbf{M}_{\mathbf{V}}^{(\mathbf{yy})}$ to define Riemannian metrics on $\mathbf{St}_{\Sigma_{\mathbf{xx}}}(p, d_{\mathbf{x}})$ and $\mathbf{St}_{\Sigma_{\mathbf{yy}}}(p, d_{\mathbf{y}})$ respectively, and these, in turn, define a Riemannian metric on $\mathbb{S}_{\mathbf{xy}}$ via $\mathbf{Z} \mapsto \mathbf{M}_{\mathbf{Z}} \coloneqq \mathbf{blkdiag}\left(\mathbf{M}_{\mathbf{U}}^{(\mathbf{xx})}, \mathbf{M}_{\mathbf{V}}^{(\mathbf{yy})}\right)$. For $\mathbf{U} \in \mathbf{St}_{\Sigma_{\mathbf{xx}}}(p, d_{\mathbf{x}})$, let $\mathbf{\Pi}_{\mathbf{U}}(\cdot)$ denote the projection on $T_{\mathbf{U}}\mathbf{St}_{\Sigma_{\mathbf{xx}}}(p, d_{\mathbf{x}})$, and similarly for $\mathbf{\Pi}_{\mathbf{V}}(\cdot)$ where $\mathbf{V} \in \mathbf{St}_{\Sigma_{\mathbf{yy}}}(p, d_{\mathbf{y}})$. Similarly, let $\Pi_{\mathbf{U}}^{\perp}(\cdot)$ and $\Pi_{\mathbf{V}}^{\perp}(\cdot)$ be the projections on the corresponding normal spaces. Given $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$ the orthogonal projection on the tangent space $T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}}$ is $\mathbf{\Pi}_{\mathbf{Z}}(\xi_{\mathbf{Z}}) = (\mathbf{\Pi}_{\mathbf{U}}(\xi_{\mathbf{U}}), \mathbf{\Pi}_{\mathbf{V}}(\xi_{\mathbf{V}}))$ where $\xi_{\mathbf{Z}} \coloneqq (\xi_{\mathbf{U}}, \xi_{\mathbf{V}}) \in \mathbb{R}^{d \times p}$.

Let $\bar{f}_{\mathbf{CCA}}$ be $f_{\mathbf{CCA}}$ extended smoothly to be defined on $\mathbb{R}^{d \times p}$ where $\bar{f}_{\mathbf{CCA}}$ is defined by Eq. (3) as well. The following are analytical expressions for the Riemannian gradient and the Riemannian Hessian in ambient coordinates:

$$\mathbf{grad} f_{\mathbf{CCA}}(\mathbf{Z}) = \mathbf{\Pi}_{\mathbf{Z}}\left(\mathbf{M}_{\mathbf{Z}}^{-1}\nabla \bar{f}_{\mathbf{CCA}}(\mathbf{Z})\right) = -\begin{bmatrix} \mathbf{\Pi}_{\mathbf{U}}\left(\left(\mathbf{M}_{\mathbf{U}}^{(\mathbf{xx})}\right)^{-1}\Sigma_{\mathbf{xy}}\mathbf{V}\mathbf{N}\right) \\ \mathbf{\Pi}_{\mathbf{V}}\left(\left(\mathbf{M}_{\mathbf{V}}^{(\mathbf{yy})}\right)^{-1}\Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{U}\mathbf{N}\right) \end{bmatrix}, \tag{4}$$

$$\begin{aligned} \mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z}) = \mathbf{\Pi}_{\mathbf{Z}}\Bigg(\mathbf{M}_{\mathbf{Z}}^{-1}\Bigg[-\Sigma_{\nabla^2 f_{\mathbf{CCA}}}\xi_{\mathbf{Z}}\mathbf{N} + \Sigma\begin{bmatrix} \xi_{\mathbf{U}}\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{V}\mathbf{N} \\ \xi_{\mathbf{V}}\mathbf{V}^{\mathrm{T}}\Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{U}\mathbf{N} \end{bmatrix} + \\ + \Sigma\begin{bmatrix} \xi_{\mathbf{U}}\mathbf{U}^{\mathrm{T}} \\ \xi_{\mathbf{V}}\mathbf{V}^{\mathrm{T}} \end{bmatrix}\mathbf{M}_{\mathbf{Z}}\mathbf{grad} f_{\mathbf{CCA}}(\mathbf{Z})\Bigg]\Bigg), \end{aligned} \tag{5}$$

$$\Sigma_{\nabla^2 f_{\mathbf{CCA}}} \coloneqq \begin{bmatrix} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{xy}}^{\mathrm{T}} & \end{bmatrix}, \ \Sigma \coloneqq \mathbf{blkdiag}\left(\Sigma_{\mathbf{xx}}, \Sigma_{\mathbf{yy}}\right),$$

where Eq. (5) is valid for critical points or if $\mathbf{M}_{\mathbf{Z}} \coloneqq \mathbf{M} = \mathbf{blkdiag}\left(\mathbf{M}^{(\mathbf{xx})}, \mathbf{M}^{(\mathbf{yy})}\right)$.

Along with formulas for the retraction and vector transport, various Riemannian optimization algorithms can be applied to solve Problem (3). In the next theorem, we summarize the critical points. Note that an important outcome of this theorem is that we can obtain σ_i from $\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{V}$ for critical $\mathbf{U}$ and $\mathbf{V}$. In the theorem statement, a pair of left and right canonical correlation vectors $\mathbf{u}$ and $\mathbf{v}$ are *on the same phase* if $\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{v} \geq 0$. The proof is delegated to Appendix B.1.

Theorem 5 *A point $\mathbf{Z} = (\mathbf{U}, \mathbf{V}) \in \mathbb{S}_{\mathbf{xy}}$ is a critical point of $f_{\mathbf{CCA}}(\mathbf{Z})$ on $\mathbb{S}_{\mathbf{xy}}$ if and only if the columns of $\mathbf{U}$ and $\mathbf{V}$ are left and right coordinated canonical correlation vectors not necessarily on the same phase.*

The optimal solutions of minimizing $f_{\mathbf{CCA}}(\mathbf{Z})$ on $\mathbb{S}_{\mathbf{xy}}$ are critical points $\mathbf{Z}$ such that the columns of $\mathbf{U}$ and $\mathbf{V}$ are coordinated left and right top p-canonical correlation vectors on the

same phase. Moreover, the optimal solution is unique up to sign of the columns of $\mathbf{U}$ *and* $\mathbf{V}$ *if* $\sigma_1 > \sigma_2 > ... > \sigma_{p+1} \geq 0$.

4.3 Stability of the Critical Points of $f_{\mathbf{CCA}}$

Naturally, we want our proposed optimization algorithm to converge to an optimal point. In Theorem 5 we characterized all the critical points of $f_{\mathbf{CCA}}(\mathbf{Z})$ on $\mathbb{S}_{\mathbf{xy}}$. In general, a typical guarantee for Riemannian optimization algorithms is that for a sequence of iterates, all the accumulation points of the sequence are critical points (e.g., Absil et al., 2008, Theorem 4.3.1). Unfortunately, this guarantee does not specify to which of the critical points the convergence is to. However, we can utilize the fact that in practice, for a sufficiently close initial guess, Riemannian optimization methods converge to the stable critical points and do not converge to unstable critical points (Absil et al., 2008, Section 4.4). We analyze the stability of the various critical points of $f_{\mathbf{CCA}}(\mathbf{Z})$ on $\mathbb{S}_{\mathbf{xy}}$. Note that solving CCA on the generalized Stiefel manifold (and its products) can be reduced to an SVD problem and that the discussion of the stability of critical points is independent of the choice of the Riemannian metric. In (Helmke and Moore, 1992, Theorem 3.4), the stability of the global minimum of SVD problem is established assuming distinct singular values. In the context of Riemannian optimization, the SVD problem was discussed by Sato and Iwai (2013). However, apart from (Absil et al., 2008, Theorem 4.6.3) where the stability of the critical points of finding the extreme eigenvalue, i.e., $p = 1$, none of the aforementioned references provides an analysis of the stability of the critical points in the context of Riemannian optimization. Thus, for completeness of the presentation, we present our analysis in this paper and delegate the proofs to the appendix. In Theorem 6, we show that under reasonable assumptions the critical points which solve Problem (3) are asymptotically stable. Moreover, we show that critical points which are saddle points or local maxima of Problem (3) are unstable. The proof of Theorem 6 is in Appendix B.2.

To do so, we use (Absil et al., 2008, Proposition 4.4.1) and (Absil et al., 2008, Proposition 4.4.2). First, recall the definition of a descent mapping from (Absil et al., 2008, Chapter 4.4): we say that F is a *descent mapping* for a cost function f on $\mathcal{M}$ if $f(F(\mathbf{x})) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{M}$. Now, (Absil et al., 2008, Proposition 4.4.1) shows that if we use any Riemannian algorithm that induces a descent mapping, and for which for every starting point the series of points generated by the algorithm has only accumulation points that are critical points, then any critical point which is not a local minimum with a compact neighborhood where the cost function achieves the same value for all other critical points is *unstable*. Additionally, (Absil et al., 2008, Proposition 4.4.2) shows that if the same conditions hold, and the distance on the manifold between iterations goes to zero as the algorithm approaches a local minimum, then if this minimum is an isolated critical point, it is an *asymptotically stable* critical point.

Theorem 6 *Consider using Riemannian optimization to minimize* $f_{\mathbf{CCA}}(\mathbf{Z})$ *subject to* $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$*, and assume that the mapping defined by the algorithm is a descent mapping. Assume that* $\sigma_1 > \sigma_2 > ... > \sigma_{p+1} \geq 0$*, then* $\mathbf{Z}$ *that minimize* $f_{\mathbf{CCA}}(\mathbf{Z})$ *on* $\mathbb{S}_{\mathbf{xy}}$ *are asymptotically stable. Furthermore, critical points which are not a local minimum of Problem (3) are unstable.*

Remark 7 *Consider the case that one or more of the top p-canonical correlations is not simple, i.e., some $\sigma_i = \sigma_j$ for $1 \leq i, j \leq p$. Then there is no longer an isolated minimum of the form $\mathbf{Z} = (\mathbf{U}, \mathbf{V}) \in \mathbb{S}_{\mathbf{xy}}$ such that the columns of $\mathbf{U}$ and $\mathbf{V}$ are the left and right top p-canonical vectors, because permutations and linear combinations of the columns of $\mathbf{U}$ and $\mathbf{V}$ associated with σ_i which maintain the phase do not change the optimal value. In such case, we can still guarantee that there exists a neighborhood of the space of the left and right top p-canonical correlation spaces for which all the starting points converge. However, note that linear convergence of Riemannian gradient descent is no longer guaranteed, as the Riemannian Hessian is only positive semi-definite, but theorems such as (Absil et al., 2008, e.g., Theorem 4.5.6), require a positive definite Riemannian Hessian.*
Indeed, any neighborhood of these spaces contains a sublevel set of $f_{\mathbf{CCA}}$ where the only critical points of $f_{\mathbf{CCA}}$ in this sublevel set belongs to the space of the left and right top p-canonical correlation spaces. Thus, if a Riemannian optimization algorithm which induces a descent mapping is started with an initial point within such a sublevel set, and assuming all accumulation points are critical points of $f_{\mathbf{CCA}}$, then it converges to the space of the left and right top p-canonical correlation spaces.

In the next theorem, we show that under certain assumptions, critical points which do not solve Problem (3) are saddle points or local maximizers. Furthermore, the algorithm is likely to converge to the desired global minimizer since under some assumptions it is the only local minimizer (up to the signs of the columns) among the critical points, thus making it the only asymptotically stable critical point. The proof relies on the proof of Theorem 9, Thus, its proof appears in Appendix B.4, after the proof of Theorem 9 (Appendix B.3).

Theorem 8 *Consider using Riemannian optimization to minimize $f_{\mathbf{CCA}}(\mathbf{Z})$ subject to $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$, where we use the Riemannian metric defined by $\mathbf{M}_{\mathbf{Z}} = \mathbf{blkdiag}\left(\mathbf{M}_{\mathbf{U}}^{(\mathbf{xx})}, \mathbf{M}_{\mathbf{V}}^{(\mathbf{yy})}\right)$ where $\mathbf{M}_{\mathbf{U}}^{(\mathbf{xx})} \in \mathbb{R}^{d_{\mathbf{x}} \times d_{\mathbf{x}}}$ and $\mathbf{M}_{\mathbf{V}}^{(\mathbf{yy})} \in \mathbb{R}^{d_{\mathbf{y}} \times d_{\mathbf{y}}}$ are given preconditioning schemes. Assume that for all $i = 1, ..., q$ the values σ_i are distinct, and that Σ is a SPD matrix. Then the global minimizer of $f_{\mathbf{CCA}}(\mathbf{Z})$ subject to $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$, denoted by $\mathbf{Z}^\star = (\mathbf{U}^\star, \mathbf{V}^\star)$, is the only local minimizer of $f_{\mathbf{CCA}}(\mathbf{Z})$ on $\mathbb{S}_{\mathbf{xy}}$, up to the signs of the columns, and it is also strict, and all other critical points are either saddle points or global maximizers. Thus, $\mathbf{Z}^\star$ is the only asymptotically stable critical point, and all other critical points are unstable.*

4.4 The Effect of Preconditioning on the Convergence

The following theorem allows us to reason about the quality of a preconditioner for the CCA problem. In this theorem, we bound the condition number of the Riemannian Hessian at the optimum based on how well the preconditioner approximates a specific matrix (Σ). The proof appears in Appendix B.3.

Theorem 9 *Consider using Riemannian optimization to minimize $f_{\mathbf{CCA}}(\mathbf{Z})$ subject to $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$, where we use the Riemannian metric defined by $\mathbf{M}_{\mathbf{Z}} = \mathbf{blkdiag}\left(\mathbf{M}_{\mathbf{U}}^{(\mathbf{xx})}, \mathbf{M}_{\mathbf{V}}^{(\mathbf{yy})}\right)$ where $\mathbf{M}_{\mathbf{U}}^{(\mathbf{xx})} \in \mathbb{R}^{d_{\mathbf{x}} \times d_{\mathbf{x}}}$ and $\mathbf{M}_{\mathbf{V}}^{(\mathbf{yy})} \in \mathbb{R}^{d_{\mathbf{y}} \times d_{\mathbf{y}}}$ are given preconditioning schemes. Also assume that $\sigma_1 > \sigma_2 > ... > \sigma_{p+1} \geq 0$ and that Σ is a SPD matrix. Then at the global minimizer of $f_{\mathbf{CCA}}(\mathbf{Z})$ subject to $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$, denoted by $\mathbf{Z}^\star = (\mathbf{U}^\star, \mathbf{V}^\star)$, the following bound on the*

condition number of the Riemannian Hessian at $\mathbf{Z}^\star$ *holds*

$$\kappa(\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z}^\star)) \leq \kappa^\star_{\mathbf{CCA}} \cdot \kappa\left(\Sigma, \mathbf{M}_{\mathbf{Z}^\star}\right)$$

where

$$\kappa^\star_{\mathbf{CCA}} \coloneqq \frac{\max\left\{\mu_1(\sigma_1+\sigma_{p+1}), \frac{1}{2}(\mu_1+\mu_2)(\sigma_1+\sigma_2)\right\}}{\min\left\{\mu_p(\sigma_p-\sigma_{p+1}), \min_{1\leq j<p} \frac{1}{2}\left(\mu_j-\mu_{j+1}\right)\left(\sigma_j-\sigma_{j+1}\right)\right\}}$$

and $\mu_1 > \cdots > \mu_p > 0$. *If* $\mathbf{M}_{\mathbf{Z}^\star} = \Sigma$ *then* $\kappa(\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z}^\star)) = \kappa^\star_{\mathbf{CCA}}$.

Note that from Theorem 9, we have a cumbersome connection between $\mathbf{N}$ and the condition number. Nevertheless, from the expression for $\kappa^\star_{\mathbf{CCA}}$, one can conclude that to balance the numerator and denominator the differences between the values on the diagonal of $\mathbf{N}$ should not be too small nor too large.

The condition number for the case $p = 1$ does not involve the values in $\mathbf{N}$, and thus is simple and illuminating:

Corollary 10 *For* $p = 1$, *the condition number of the Riemannian Hessian at the optimum is at most* $\frac{\sigma_1+\sigma_2}{\sigma_1-\sigma_2} \cdot \kappa(\Sigma, \mathbf{M}_{\mathbf{z}^\star})$. *If* $\mathbf{M}_{\mathbf{z}^\star} = \Sigma$ *then the condition number of the Riemannian Hessian at the optimum is* $\frac{\sigma_1+\sigma_2}{\sigma_1-\sigma_2}$.

The condition number bound from Corollary 10 decomposes into two components: the first is the relative eigengap ($(\sigma_1+\sigma_2)/(\sigma_1-\sigma_2)$), which forms a natural condition number for the problem (if the first and second correlations are very close, it is very hard to distinguish between them) that almost always appear in problems of this form, and a second component which measures how close the preconditioner-defined metric approximates the natural metric for the constraints. The optimal preconditioner, according to the bound, is $\mathbf{M} = \Sigma$. However, using this preconditioner requires explicitly computing it in $O(nd^2)$ time. This is too expensive since the exact correlations can be computed analytically in $O(nd^2)$ time as well Björck and Golub (1973).

4.5 Randomized Preconditioning for CCA

The condition number bound in Theorem 9 separates two factors: $\kappa^\star_{\mathbf{CCA}}$ and $\kappa\left(\Sigma, \mathbf{M}\right)$. The first, $\kappa^\star_{\mathbf{CCA}}$, depends on the gap between the $p+1$ largest canonical correlations and on the differences between the values in $\mathbf{N}$, which are parameters for the CCA problem as a Riemannian optimization problem. The dependence on the gap almost always appears in problems of this form, since the more the singular values are distinct it is easier to distinguish between them. The second component, $\kappa\left(\Sigma, \mathbf{M}_{\mathbf{Z}^\star}\right)$, measures how close the preconditioner, which defines the Riemannian metric, approximates Σ. A preconditioner that minimizes the bound in Theorem 9 is such that $\mathbf{M}_{\mathbf{Z}^\star} = \Sigma$. However, using that preconditioner, requires explicitly computing a factorization of Σ which classically requires $\Omega(nd^2)$ arithmetic operations which is non-beneficial in light of direct solution methods.

Thus, Theorem 9 provides an argument in favor of our proposed randomized preconditioner, i.e., easy to factorize $\mathbf{M}_{\mathbf{Z}}$ such that $\kappa\left(\Sigma, \mathbf{M}_{\mathbf{Z}^\star}\right)$ is bounded. In order to achieve this goal, the preconditioning schemes $\mathbf{U} \mapsto \mathbf{M}^{(\mathbf{xx})}_{\mathbf{U}}$ and $\mathbf{V} \mapsto \mathbf{M}^{(\mathbf{yy})}_{\mathbf{V}}$ should approximate $\Sigma_{\mathbf{xx}} = \mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_{\mathbf{x}}\mathbf{I}_{d_{\mathbf{x}}}$ and $\Sigma_{\mathbf{yy}} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} + \lambda_{\mathbf{y}}\mathbf{I}_{d_{\mathbf{y}}}$ respectively (at least at the optimum). Thus, as described in Section 3 we propose constant randomized preconditioning scheme

Algorithm 1 Sketched Riemannian Iterative CCA with warm-start.

1: **Input:** $\mathbf{X} \in \mathbb{R}^{n \times d_{\mathbf{x}}}$, $\mathbf{Y} \in \mathbb{R}^{n \times d_{\mathbf{y}}}$, $s \geq \max(d_{\mathbf{x}}, d_{\mathbf{y}})$, $\lambda_{\mathbf{x}}, \lambda_{\mathbf{y}} \geq 0$.
2: **Generate random** $h : \{1, \ldots, d\} \to \{1, \ldots, s\}$ and $g : \{1, \ldots, d\} \to \{-1, +1\}$. Let $\mathbf{S}$ denote the corresponding CountSketch matrix.
3: $\mathbf{X_S} \leftarrow \mathbf{SX}$, $\mathbf{Y_S} \leftarrow \mathbf{SY}$.
4: $\tilde{\mathbf{U}}, \tilde{\mathbf{V}} \leftarrow \text{exact} - \text{cca}(\mathbf{X_S}, \mathbf{Y_S})$.
5: $\mathbf{M}^{(\mathbf{xx})} \leftarrow \mathbf{X_S}^{\mathrm{T}}\mathbf{X_S} + \lambda_{\mathbf{x}}\mathbf{I}_{d_{\mathbf{x}}}$, $\mathbf{M}^{(\mathbf{yy})} \leftarrow \mathbf{Y_S}^{\mathrm{T}}\mathbf{Y_S} + \lambda_{\mathbf{y}}\mathbf{I}_{d_{\mathbf{y}}}$.
6: **Notation:** $\Sigma_{\mathbf{xx}} = \mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_{\mathbf{x}}\mathbf{I}$, $\Sigma_{\mathbf{yy}} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} + \lambda_{\mathbf{y}}\mathbf{I}$. Do not compute these matrices (algorithms only require taking products with them).
7: **Choose:** any $\mathbf{N} = \mathbf{diag}\,(\mu_1, ..., \mu_p)$ s.t. $\mu_1 > ... > \mu_p > 0$.
8: Using Riemannian CG, solve $\max \mathbf{Tr}\left(\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{VN}\right)$ s.t. $\mathbf{U} \in \mathbf{St}_{\Sigma_{\mathbf{xx}}}(p, d_{\mathbf{x}})$, $\mathbf{V} \in \mathbf{St}_{\Sigma_{\mathbf{yy}}}(p, d_{\mathbf{y}})$. Use $\mathbf{M}^{(\mathbf{xx})}$ and $\mathbf{M}^{(\mathbf{yy})}$ for the metric. Start the iteration from $\mathbf{qf}_{\Sigma_{\mathbf{xx}}}\left(\tilde{\mathbf{U}}\right)$ and $\mathbf{qf}_{\Sigma_{\mathbf{yy}}}\left(\tilde{\mathbf{V}}\right)$.

$\mathbf{M} \coloneqq \mathbf{blkdiag}\left(\mathbf{M}^{(\mathbf{xx})}, \mathbf{M}^{(\mathbf{yy})}\right)$, i.e., using sketching for $\mathbf{X}$ and $\mathbf{Y}$ correspondingly to define $\mathbf{M}_{\mathbf{U}}^{(\mathbf{xx})} \coloneqq \mathbf{M}^{(\mathbf{xx})}$ and $\mathbf{M}_{\mathbf{V}}^{(\mathbf{yy})} \coloneqq \mathbf{M}^{(\mathbf{yy})}$. A pseudocode description of an end-to-end randomized preconditioned CCA algorithm with warm-start appears in Algorithm 1. The following corollary summarizes our theoretical results regarding the proposed algorithm. We remark that CountSketch can possibly be replaced with other sketching transforms (such as SRHT), and Riemannian CG can be replaced with other Riemannian optimization methods, although the bounds in the corollary might change. We prove Corollary 11 in Appendix B.5.

Corollary 11 *Consider Algorithm 1. Let $\delta \in (0, 1)$ and denote $s_\lambda = \max(s_{\lambda_{\mathbf{x}}}(\mathbf{X}), s_{\lambda_{\mathbf{y}}}(\mathbf{Y}))$. If $s = \max(\lceil 40 s_\lambda^2/\delta \rceil, d)$, then with probability of at least $1 - \delta$, the condition number of the Riemannian Hessian at the optimum is bounded by $3\kappa^\star_{\mathbf{CCA}}$, regardless of the condition number of $\Sigma_{\mathbf{xx}}$ and $\Sigma_{\mathbf{yy}}$. Furthermore, assuming we use Riemannian CG, $n \geq d \geq p$, and all computations are done in ambient $\mathbb{R}^{d \times p}$ coordinates, then the preprocessing steps take $O(\mathbf{nnz}\,(\mathbf{X}) + \mathbf{nnz}\,(\mathbf{Y})) = O(nd)$ and $O(sd^2)$. Assuming a bounded number of line-search steps in each iteration then each iteration takes $O\left(p\,(\mathbf{nnz}\,(\mathbf{X}) + \mathbf{nnz}\,(\mathbf{Y})) + dp^2 + d^2p\right)$ operations.*

5. Sketched Iterative FDA

Fisher's linear discriminant analysis (FDA), introduced by Fisher (1936), is a well-known method for classification (Fisher, 1936; Mika et al., 1999), and more commonly for dimensionality reduction before classification (Chen et al., 2012). The latter is achieved by finding an embedding such that simultaneously the between-class scatter is maximized and the within-class scatter is minimized. In this paper, we consider a regularized version of FDA as defined below:

Definition 12 *(Duda et al., 2001, Section 4.11) Let $\mathbf{x}_1^{(1)}, ..., \mathbf{x}_{n_1}^{(1)}, \mathbf{x}_1^{(2)}, ..., \mathbf{x}_{n_2}^{(2)} ..., \mathbf{x}_1^{(2)}, ..., \mathbf{x}_{n_l}^{(l)} \in \mathbb{R}^d$ be samples from $l \leq d$ different classes, and denote by $\mathbf{x}_1, ..., \mathbf{x}_n$ the union of the different classes (the entire data set in a sequential index). For $i = 1, \ldots, n$, let y_i denote the*

label corresponding to $\mathbf{x}_i$*, i.e.,* $y_i = k$ *if* $\mathbf{x}_i = \mathbf{x}_j^{(k)}$ *for some* j*. Let* $\mathbf{m}_k$*, for* $k = 1, \ldots, l$*, denote the sample mean of class* k *(i.e.,* $\mathbf{m}_k \coloneqq n_k^{-1} \sum_{i=1}^{n_k} \mathbf{x}_i^{(k)}$*), and* $\mathbf{m} \coloneqq n^{-1} \sum_{i=1}^{n} \mathbf{x}_i = n^{-1} \sum_{k=1}^{l} n_k \mathbf{m}_k$ *denote the data set sample mean of the entire data set. Let* $\mathbf{S_B}$ *and* $\mathbf{S_w}$ *be the* between-class[3] *and* within-class *scatter matrices (respectively):*

$$\mathbf{S_B} \coloneqq \sum_{k=1}^{l} n_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^{\mathrm{T}} \quad \mathbf{S_w} \coloneqq \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m}_{y_i})(\mathbf{x}_i - \mathbf{m}_{y_i})^{\mathrm{T}}$$

Let $\lambda \geq 0$ *be a regularization parameter. The* $l-1$ *FDA weight vectors* $\mathbf{w}_1, ..., \mathbf{w}_{l-1}$ *are the columns of* $\mathbf{W} \in \mathbb{R}^{d \times (l-1)}$ *such that* $\mathbf{W}$ *is the maximizer of the following cost function:*

$$J(\mathbf{W}) = \frac{\det(\mathbf{W}^{\mathrm{T}} \mathbf{S_B} \mathbf{W})}{\det(\mathbf{W}^{\mathrm{T}} (\mathbf{S_w} + \lambda \mathbf{I}_d) \mathbf{W})}. \tag{6}$$

5.1 FDA as an Optimization Problem on a Generalized Stiefel Manifold

It is well known that the solution of maximizing Eq. (6) (i.e., finding the FDA weight vectors) is equivalent to finding a matrix $\mathbf{W}$ such that its columns are the leading $l-1$ generalized eigenvectors of the matrix pencil $(\mathbf{S_B}, \mathbf{S_w} + \lambda \mathbf{I}_d)$ (Duda et al., 2001, Section 4.11). Note that this generalized eigenproblem has at most $l-1$ nonzero generalized eigenvalues since the matrix $\mathbf{S_w} + \lambda \mathbf{I}_d$ is a SPD matrix, and $\mathbf{S_B}$ is the sum of l matrices of rank one or less, where only $l-1$ of these are independent, thus, $\mathbf{S_B}$ is of rank $l-1$ or less. We denote the eigenvalues of the matrix pencil $(\mathbf{S_B}, \mathbf{S_w} + \lambda \mathbf{I}_d)$ with correspondence to the FDA weight vectors by $\rho_1 \geq \rho_2 \geq ... \geq \rho_d \geq 0$.

We focus on finding the p leading FDA weight vectors, i.e., $\mathbf{w}_1, ..., \mathbf{w}_p$ corresponding to $\rho_1 \geq \rho_2 \geq ... \geq \rho_p$ where $p \leq l-1$. For the purpose of describing and analyzing our algorithm, it is useful to write,

$$\mathbf{S_w} = \hat{\mathbf{X}}^{\mathrm{T}} \hat{\mathbf{X}}, \quad \hat{\mathbf{X}} \coloneqq \mathbf{X} - \mathbf{Y}, \quad \mathbf{S_B} = \hat{\mathbf{Y}}^{\mathrm{T}} \hat{\mathbf{Y}} \ ,$$

where $\hat{\mathbf{X}} \in \mathbb{R}^{n \times d}$ is a matrix such that each i-th row of $\hat{\mathbf{X}}$ is $(\mathbf{x}_i - \mathbf{m}_{y_i})^{\mathrm{T}}$, $\mathbf{X} \in \mathbb{R}^{n \times d}$ is a matrix such that each i-th row of $\mathbf{X}$ is $\mathbf{x}_i^{\mathrm{T}}$, $\mathbf{Y} \in \mathbb{R}^{n \times d}$ is a matrix such that each i-th row is of the form $\mathbf{m}_{y_i}$ (thus, there are at most l different rows in $\mathbf{Y}$), and $\hat{\mathbf{Y}} \in \mathbb{R}^{l \times d}$ is a matrix such that each kth row of $\hat{\mathbf{Y}}$ is $\sqrt{n_k}(\mathbf{m}_k - \mathbf{m})^{\mathrm{T}}$. With these notations, we can reformulate the problem of finding the p leading FDA weight vectors as a Riemannian optimization problem on the generalized Stiefel manifold (see Fukunaga, 2013, Section 10.2, Eq. (10.5)), i.e., finding the generalized eigenvalues of the matrix pencil $(\mathbf{S_B}, \mathbf{S_w} + \lambda \mathbf{I}_d)$. We use the *Brockett cost function* (Brockett, 1991) and obtain the following optimization problem

$$\max_{\mathbf{W} \in \mathbb{R}^{d \times p}} \mathbf{Tr}\left(\mathbf{W}^{\mathrm{T}} \hat{\mathbf{Y}}^{\mathrm{T}} \hat{\mathbf{Y}} \mathbf{W} \mathbf{N}\right), \ \text{s.t.} \ \mathbf{W}^{\mathrm{T}} (\hat{\mathbf{X}}^{\mathrm{T}} \hat{\mathbf{X}} + \lambda \mathbf{I}_d) \mathbf{W} = \mathbf{I}_p, \tag{7}$$

where $\mathbf{N} = \mathbf{diag}\,(\mu_1, ..., \mu_p)$ where we take arbitrary $\mu_1, ..., \mu_p$ such that $\mu_1 > ... > \mu_p > 0$.

3. For $l = 2$ the matrix $\mathbf{S_B}$ is defined by $\mathbf{S_B} \coloneqq (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^{\mathrm{T}}$. The definitions coincide after multiplying $\mathbf{S_B} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^{\mathrm{T}}$ by $2(n_1 n_2)/n$.

Remark 13 *Problem (7) is actually a relaxation of the* trace-ratio *problem (Wang et al., 2007; Ngo et al., 2012), which provides superior results in various tasks (e.g., classification and clustering Wang et al., 2007). We focus on Problem (7) since it is amenable to our preconditioning strategy.*

In the next subsection, we detail the Riemannian components which allow to solve the FDA problem. We show that critical points of the corresponding objective function are matrices $\mathbf{W}$ such that the columns are some p FDA weight vectors. In particular, the optimal solutions are critical points consisting of the p leading FDA weight vectors. We further show that if $\rho_1, \ldots, \rho_{p+1}$ are distinct, the optimal solution is unique up to sign of the columns of $\mathbf{W}$.

5.2 Preconditioned Riemannian Components for FDA

We first transform Problem (7) into a minimization problem:

$$\min_{\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)} f_{\mathbf{FDA}}(\mathbf{W}) \ , \ f_{\mathbf{FDA}}(\mathbf{W}) \coloneqq -\frac{1}{2}\mathbf{Tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{S_B}\mathbf{W}\mathbf{N}\right). \tag{8}$$

We use the components of the generalized Stiefel manifold with a Riemannian metric defined by a preconditioning scheme $\mathbf{M_W} \in \mathbb{R}^{d\times d}$ to apply Riemannian optimization to solve Problem (8).

Let $\bar{f}_{\mathbf{FDA}}$ be $f_{\mathbf{FDA}}$ extended smoothly to be defined on $\mathbb{R}^{d\times p}$ where $\bar{f}_{\mathbf{FDA}}$ is defined by Eq. (8) as well.

For $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$, let $\mathbf{\Pi_W}(\cdot)$ denote the projection on $T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$. Similarly, let $\mathbf{\Pi}^{\perp}_{\mathbf{W}}(\cdot)$ be the projection on the corresponding normal space. The following are analytical expressions for the Riemannian gradient and Hessian in ambient coordinates:

$$\mathbf{grad} f_{\mathbf{FDA}}(\mathbf{W}) = \mathbf{\Pi_W}\left(\mathbf{M}_{\mathbf{W}}^{-1}\nabla\bar{f}_{\mathbf{FDA}}(\mathbf{W})\right) - \mathbf{\Pi_W}\left(\mathbf{M}_{\mathbf{W}}^{-1}\mathbf{S_B}\mathbf{W}\mathbf{N}\right), \tag{9}$$

$$\begin{aligned}\mathbf{Hess} f_{\mathbf{FDA}}(\mathbf{W})[\xi_{\mathbf{W}}] = \mathbf{\Pi_W}(\mathbf{M}_{\mathbf{W}}^{-1}[-\mathbf{S_B}\xi_{\mathbf{W}}\mathbf{N}+ \\ (\mathbf{S_w}+\lambda\mathbf{I}_d)\xi_{\mathbf{W}}(\mathbf{W}^{\mathrm{T}}\mathbf{S_B}\mathbf{W}\mathbf{N}+\mathbf{grad} f(\mathbf{W}))]),\end{aligned} \tag{10}$$

where Eq. (10) is valid for critical points or if $\mathbf{M_W} \coloneqq \mathbf{M}$.

Along with formulas for the retraction and vector transport, various Riemannian optimization algorithms can be applied to solve Problem (8). In the next theorem, we summarize the critical points. Note that an important outcome of this theorem is that we can obtain ρ_i from $\mathbf{W}^{\mathrm{T}}\mathbf{S_B}\mathbf{W}$ with a critical $\mathbf{W}$. The proof is almost the same as the proof of Theorem 5, and presented in Appendix C.1.

Theorem 14 *A point $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ is a critical point of $f_{\mathbf{FDA}}(\mathbf{W})$ on the manifold $\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ if and only if the columns of $\mathbf{W}$ are some p FDA weight vectors.*

The optimal solutions of minimizing $f_{\mathbf{FDA}}(\mathbf{W})$ on $\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ are critical points $\mathbf{W}$ such that the columns are the p leading FDA weight vectors. Moreover, the optimal solution is unique up to sign of the columns if $\rho_1 > \rho_2 > \ldots > \rho_{p+1} \geq 0$.

5.3 Stability of the Critical Points of $f_{\mathbf{FDA}}$

In the FDA problem, we want any optimization algorithm to converge to an optimal point. In Theorem 14 we prove that the critical points of $f_{\mathbf{FDA}}(\mathbf{W})$ on $\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ are matrices $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ such that the columns are some p FDA weight vectors. Similarly to the CCA problem, we analyze the stability of the critical points using (Absil et al., 2008, Proposition 4.4.1) and (Absil et al., 2008, Proposition 4.4.2). Note that similarly to the CCA problem, solving FDA on the generalized Stiefel manifold can be reduced to a generalized eigenvalues problem, and that the discussion of the stability of critical points is independent of the choice of Riemannian metric. In (Brockett, 1991, Theorem 4), the stability of the global minimum of the generalized eigenvalue problem is established assuming distinct eigenvalues. In the context of Riemannian optimization, the eigenvalue problem was discussed in various references, e.g., (Absil et al., 2008, Sections 4.6 and 4.8). However, apart from (Absil et al., 2008, Theorem 4.6.3) where the stability of the critical points of finding the extreme eigenvalue, i.e., $p = 1$, none of the aforementioned references provides a complete analysis of the stability of the critical points in the context of Riemannian optimization. Thus, for completeness of the presentation, we present our analysis in this paper and relegate the proofs to the appendix.

The following theorem shows that under reasonable assumptions the critical points which solve Problem (8) are asymptotically stable and critical points which are saddle points of Problem (8) are unstable. The proof is analogous to the proof of Theorem 6, and presented in Appendix C.2.

Theorem 15 *Consider using Riemannian optimization to minimize $f_{\mathbf{FDA}}(\mathbf{W})$ subject to $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$, and assume that the mapping defined by the algorithm is a descent mapping. Assume that $\rho_1 > \rho_2 > ... > \rho_{p+1} \geq 0$, then $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ such that the columns are the p leading FDA weight vectors are asymptotically stable. Furthermore, critical points which are not a local minimum of Problem (8) are unstable.*

Remark 16 *Consider the case that one or more of the p-dominant generalized eigenvalues of the matrix pencil $(\mathbf{S_B}, \mathbf{S_w} + \lambda\mathbf{I}_d)$ is not simple, i.e., some $\rho_i = \rho_j$ for $1 \leq i, j \leq p$. Then there is no longer an isolated minimum of the form $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ such that the columns are the p leading FDA weight vectors, because permutations and linear combinations of the columns of $\mathbf{W}$ associated with ρ_i do not change the optimal value. In such case, we can still guarantee that there exists a neighborhood of the space of the p-dominant generalized eigenspaces for which all the starting points converge to the space of the p-dominant generalized eigenspaces. However, note that linear convergence of Riemannian gradient descent is no longer guaranteed, as the Riemannian Hessian is only positive semi-definite, but theorems such as (Absil et al., 2008, e.g., Theorem 4.5.6), require a positive definite Riemannian Hessian.*
Indeed, any neighborhood of these generalized eigenspaces contains a sublevel set of $f_{\mathbf{FDA}}$ where the only critical points of $f_{\mathbf{FDA}}$ in this sublevel set belong to the space of the p-dominant generalized eigenspaces. Thus, if a Riemannian optimization algorithm which induces a descent mapping is started with an initial point within such a sublevel set, and assuming all accumulation points are critical points of $f_{\mathbf{FDA}}$, then it converges to the space of the p-dominant generalized eigenspaces.

In the next theorem, we show that under certain assumptions, critical points which do not solve Problem (7) are saddle points or local maximizers. Furthermore, the algorithm is likely to converge to the desired global minimizer since under some assumptions it is the only local minimizer (up to the signs of the columns) among the critical points, thus making it the only asymptotically stable critical point. The proof relies on the proof of Theorem 18, and is presented in Appendix C.4.

Theorem 17 *Consider using Riemannian optimization to minimize $f_{\mathbf{FDA}}(\mathbf{W})$ subject to $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$, where we use the Riemannian metric defined by $\mathbf{M}_{\mathbf{W}} \in \mathbb{R}^{d\times d}$, which is a given preconditioning scheme. Assume that for all $i = 1, ..., d$ the values ρ_i are distinct and that $\mathbf{S_w} + \lambda\mathbf{I}_d$ is a SPD matrix. Let $\mathbf{W}^\star$ denote the global minimizer of $f_{\mathbf{FDA}}(\mathbf{W})$ subject to $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$. Then $\mathbf{W}^\star$ is the only local minimizer of $f_{\mathbf{FDA}}(\mathbf{W})$ on $\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$, up to the signs of the columns, and it is also strict, and all other critical points are either saddle points or strict local maximizers. Thus, $\mathbf{W}^\star$ is the only asymptotically stable critical point, and all other critical points are unstable.*

5.4 The Effect of Preconditioning on the Convergence

The following theorem provides a bound on the condition number of the Riemannian Hessian at the optimum based on how well the preconditioner approximates a specific matrix $(\mathbf{S_w} + \lambda\mathbf{I}_d)$. This theorem provides a general guideline in designing a Riemannian preconditioner via the Riemannian metric, and in particular motivates our proposed approach detailed in Subsection 5.5. The proof uses the same arguments as the proof of Theorem 9, though they are simplified considerably since only a single generalized Stiefel manifold is considered. The proof appears in Appendix C.3.

Theorem 18 *Consider using Riemannian optimization to minimize $f_{\mathbf{FDA}}(\mathbf{W})$ subject to $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$, where we use the Riemannian metric defined by $\mathbf{M}_{\mathbf{W}} \in \mathbb{R}^{d\times d}$, which is a given preconditioning scheme. Assume that $\rho_1 > \rho_2 > ... > \rho_{p+1} \geq 0$ and that $\mathbf{S_w} + \lambda\mathbf{I}_d$ is an SPD matrix. Let $\mathbf{W}^\star$ denote the global minimizer of $f_{\mathbf{FDA}}(\mathbf{W})$ subject to $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$. Then,*

$$\kappa(\mathbf{Hess} f_{\mathbf{FDA}}(\mathbf{W}^\star)) \leq \kappa^\star_{\mathbf{FDA}} \cdot \kappa\left(\mathbf{S_w} + \lambda\mathbf{I}_d, \mathbf{M}_{\mathbf{W}^\star}\right)$$

where

$$\kappa^\star_{\mathbf{FDA}} := \frac{\mu_1\left(\rho_1 - \rho_d\right)}{\min\left\{\mu_p\left(\rho_p - \rho_{p+1}\right), \min_{1\leq j<p} \frac{1}{2}\left(\mu_j - \mu_{j+1}\right)\left(\rho_j - \rho_{j+1}\right)\right\}}$$

and $\mu_1 > \cdots > \mu_p > 0$. If $\mathbf{M}_{\mathbf{W}^\star} = \mathbf{S_w} + \lambda\mathbf{I}_d$ then $\kappa(\mathbf{Hess} f_{\mathbf{FDA}}(\mathbf{W}^\star)) = \kappa^\star_{\mathbf{FDA}}$.

Note that from Theorem 18, we have a cumbersome connection between $\mathbf{N}$ and the condition number. Nevertheless, from the expression for $\kappa^\star_{\mathbf{FDA}}$, one can conclude that to balance the numerator and denominator the differences between the values on the diagonal of $\mathbf{N}$ should not be too small nor too large.

The condition number for the case $p = 1$ does not involve the values in $\mathbf{N}$, and thus is simple and illuminating:

Corollary 19 *For $p = 1$, the condition number of the Riemannian Hessian at the optimum is at most $\frac{\rho_1-\rho_d}{\rho_1-\rho_2} \cdot \kappa\left(\mathbf{S_w} + \lambda\mathbf{I}_d, \mathbf{M}_{\mathbf{W}^\star}\right)$. If $\mathbf{M}_{\mathbf{W}^\star} = \mathbf{S_w} + \lambda\mathbf{I}_d$ then the condition number of the Riemannian Hessian at the optimum is $\frac{\rho_1-\rho_d}{\rho_1-\rho_2}$.*

The condition number bound from Corollary 19 decomposes to two components: the first is the relative eigengap $(\rho_1/(\rho_1-\rho_2))$ (when ρ_d is very small or zero), which form a natural condition number for the problem (if the first and second generalized eigenvalues are very close, it is very hard to distinguish between invariant subspaces corresponding to them) that almost always appears in problems of this form, and a second component which measures how close the preconditioner-defined metric approximates the natural metric for the constraints. The optimal preconditioner according to the bound, is $\mathbf{M} = \mathbf{S_w} + \lambda \mathbf{I}_d$. However, using this preconditioner requires explicitly computing it in $O(nd^2)$ time. This is too expensive, since the exact discriminant variables can be computed analytically in $O(nd^2)$ time as well (exact solution requires finding the inverse of $\mathbf{S_w} + \lambda \mathbf{I}_d$ and then finding the first eigenvalue and corresponding eigenvector of $(\mathbf{S_w} + \lambda \mathbf{I}_d)^{-1}\mathbf{S_B}$).

5.5 Randomized Preconditioning for FDA

The bound on the condition number in Theorem 18 decomposes two components: the first, $\kappa^\star_{\mathbf{FDA}}$, depends only on the FDA problem (and its formulation as a Riemannian optimization problem). $\kappa^\star_{\mathbf{FDA}}$ depends on both the gap between the $p+1$ largest generalized eigenvalues, and on the diagonal elements of $\mathbf{N}$, which are parameters of the optimization problem. The dependence on the gap almost always appears in problems of this form, since the more the generalized eigenvalues associated with the FDA weight vectors we search for are distinct it is easier to distinguish between them. The second component, $\kappa\left(\mathbf{S_w} + \lambda \mathbf{I}_d, \mathbf{M_{W^\star}}\right)$, measures how close the preconditioner, which defines the Riemannian metric, approximates $\mathbf{S_w} + \lambda \mathbf{I}_d$. A preconditioner that minimizes the bound in Theorem 18 is such that $\mathbf{M_{W^\star}} = \mathbf{S_w} + \lambda \mathbf{I}_d$. However, using that preconditioner requires explicitly computing $\mathbf{S_w} + \lambda \mathbf{I}_d$ which takes $O(nd^2)$ arithmetic operations. As for the CCA problem, direct methods for solving FDA require $\Theta(nd^2)$ arithmetic operations as well (exact solution requires finding the inverse of $\mathbf{S_w}+\lambda \mathbf{I}_d$ and then finding the eigenvalues and corresponding eigenvectors of $(\mathbf{S_w}+\lambda \mathbf{I}_d)^{-1}\mathbf{S_B}$). Thus, we want $\mathbf{M_W}$ to approximate $\mathbf{S_w} + \lambda \mathbf{I}_d$ at the optimum, while allowing a cheap factorization which is satisfied by our proposed randomized preconditioning approach.

We propose to design the preconditioner in the following way: $\mathbf{M_W} \coloneqq \mathbf{M}$ approximates $\mathbf{S_w} + \lambda \mathbf{I}_d = \hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{X}} + \lambda \mathbf{I}_d$ via a matrix sketching procedure for $\hat{\mathbf{X}}$ as described in Section 3. A full description of a randomized preconditioned algorithm for FDA with warm-start appears in Algorithm 2. The following corollary summarize our theoretical results regarding the proposed algorithm. Note that CountSketch can possibly be replaced with other sketching transforms (such as SRHT), and Riemannian CG can be replaced with any other Riemannian optimization methods, although the bound in the corollary might change. The proof is in Appendix C.5.

Corollary 20 *Consider Algorithm 2. Let $\delta \in (0,1)$. If $s = \max(\left\lceil 20 s_\lambda(\hat{\mathbf{X}})^2/\delta \right\rceil, d)$, then with probability of at least $1-\delta$, the condition number of the Riemannian Hessian at the optimum is bounded by $3\kappa^\star_{\mathbf{FDA}}$, regardless of the condition number of $\mathbf{S_w}+\lambda\mathbf{I}$. Furthermore, assuming we use Riemannian CG, $n \geq d \geq p$, and all computations are done in ambient $\mathbb{R}^{d\times p}$ coordinates, then the preprocessing steps take $O(\mathbf{nnz}\left(\hat{\mathbf{X}}\right)) = O(nd)$ and $O(sd^2)$. Assuming a bounded number of line-search steps in each iteration then each iteration takes $O(p\left(\mathbf{nnz}\left(\mathbf{X}\right) + ld\right) + \mathbf{nnz}\left(\hat{\mathbf{X}}\right)p + d^2p + dp^2)$ operations.*

Algorithm 2 Sketched Riemannian Iterative FDA with warm-start.

1: **Input:** $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \mathbb{N}^n$, $s \geq d$, $\lambda \geq 0$.
2: Compute matrices $\mathbf{S_B}$ and $\hat{\mathbf{X}}$.
3: **Generate random** $h : \{1, \ldots, d\} \to \{1, \ldots, s\}$ and $g : \{1, \ldots, d\} \to \{-1, +1\}$. Let $\mathbf{S}$ denote the corresponding CountSketch matrix.
4: $\mathbf{X_S} \leftarrow \mathbf{S}\hat{\mathbf{X}}$.
5: $\tilde{\mathbf{W}} \leftarrow \text{exact} - \text{fda}(\mathbf{X_S})$.
6: $\mathbf{M} \leftarrow \mathbf{X_S}^{\mathrm{T}}\mathbf{X_S} + \lambda \mathbf{I}_d$.
7: **Notation:** $\mathbf{S_w} = \hat{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{X}}$. Do not compute this matrix (algorithms only require taking products with it).
8: **Choose:** any $\mathbf{N} = \mathbf{diag}\,(\mu_1, ..., \mu_p)$ s.t. $\mu_1 > ... > \mu_p > 0$.
9: Using Riemannian CG, solve $\max \mathbf{Tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{S_B}\mathbf{W}\mathbf{N}\right)$ s.t. $\mathbf{W} \in \mathbf{St}_{\mathbf{S_w}+\lambda\mathbf{I}_d}(p, d)$. Use $\mathbf{M}$ for the metric. Start the iteration from $\mathbf{qf}_{\mathbf{S_w}+\lambda\mathbf{I}}\left(\tilde{\mathbf{W}}\right)$.

6. Numerical Experiments

In the following section, we present our numerical experiments illustrating our randomized preconditioning approach.

6.1 Experiments with Real-World Data Sets

We report experiments with our proposed preconditioned Riemannian optimization algorithms. The experiments are not designed to be exhaustive; we use a prototype MATLAB implementation. In particular we present experiments with the preconditioned CCA and FDA algorithms presented in Sections 4 and 5. Our aim is to assess the effectiveness of our randomized preconditioning approach.

In addition to Algorithms 1 and 2, we experiment with an additional preconditioning strategy based on work by Gonen et al. (2016), which we term as *Dominant Subspace Preconditioning*. This preconditioner was designed via an approximation of the empirical correlation matrix to speed up SVRG when solving ridge regression problems. In our experiments, we use this method to approximate $\Sigma_{\mathbf{xx}}$ and $\Sigma_{\mathbf{yy}}$ for CCA, and $\mathbf{S_w} + \lambda\mathbf{I}_d$ for FDA. Specifically, the matrices $\mathbf{X}, \mathbf{Y}$ and $\hat{\mathbf{X}}$ are approximated in the following way: suppose $\mathbf{A} = \mathbf{X}^{\mathrm{T}}\mathbf{X} \in \mathbb{R}^{d\times d}$, and let $\mathbf{X} = \mathbf{U}\Lambda^{1/2}\mathbf{V}^{\mathrm{T}}$ be an SVD decomposition of $\mathbf{X}$ such that $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}}$ is an eigendecomposition, with the diagonal entries in Λ sorted in a descending order. Given k, let us denote by $\mathbf{U}_k$ the first k columns of $\mathbf{U}$, Λ_k denote the leading $k \times k$ minor of Λ, and λ_k the k-th largest eigenvalue of $\mathbf{A}$. Then, the k-dominant subspace preconditioner of $\mathbf{A} + \lambda\mathbf{I}_d$ is $\mathbf{U}_k(\Lambda_k - \lambda_k\mathbf{I})\mathbf{U}_k^{\mathrm{T}} + (\lambda_k + \lambda)\mathbf{I}_d$. 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

$$\mathbf{U}_k(\Lambda_k + \lambda\mathbf{I})^{-1}\mathbf{U}_k^{\mathrm{T}} + \frac{1}{\lambda_k + \lambda}(\mathbf{I}_d - \mathbf{U}_k\mathbf{U}_k^{\mathrm{T}}).$$

We also experiment with variants of sketched iterative CCA (Algorithm 1) and sketched iterative FDA (Algorithm 2) in which Riemannian CG is replaced with Riemannian trust-region Method.

We use MATLAB for our implementations, relying on the MANOPT library (Boumal et al., 2014) for Riemannian optimization. In the MANOPT library we implemented the generalized Stiefel manifold with a non-standard metric[4]. The experiments we present here are with $p = 3$ and $\mathbf{N} = \mathbf{diag}\,(3, 2.75, 2)$. Recall that the only condition on the matrix $\mathbf{N}$ is that it is diagonal, with strictly descending and positive diagonal elements. In the literature, typically, descending integers are chosen. We remark that this choice of $\mathbf{N}$ is an arbitrary (not optimized) choice, and we achieved similar results for other choices of $\mathbf{N}$ as a diagonal matrix with strictly decreasing values on the diagonal, and p as well (which we exclude from this text). In the plots (Fig. 1, 2, 3, and 4), the role of s is the number of rows in the sketched data matrices, after applying a CountSketch transformation on the data matrices. Similarly, k is the number of singular values and vectors we use for the Dominant Subspace Preconditioning.

We did not optimize the implementation, so wall clock time is not an appropriate metric for performance, instead, we use an alternative metrics. In particular, the direct methods are about 10 times faster than the iterative methods measured in wall clock time in the experiments in this subsection, however, as n becomes larger, iterative methods become faster than the direct methods (see Subsection 6.2). In Riemannian trust-region Method, different iterations do a variable amount of passes over the data, so we measure passes directly, i.e., products with the data matrices, as this is the dominant cost of our algorithm. As for Riemannian CG, the MANOPT solver restricts the number of line-search steps in each iteration to 25, otherwise the step is rejected. In practice, in our experiments the number of products in different iterations is between 6 to 26 for CCA and between 3 to 13 for FDA, so we report the number of iterations. For every iteration or pass over the data, t, we plot the suboptimality of the current iterate: $|\sum_{i=1}^{p} \sigma_i\mu_i + f_{\mathbf{CCA}}(\mathbf{Z}_t)|/\sum_{i=1}^{p} \sigma_i\mu_i$ for CCA where $\mathbf{Z}_t$ represents the parameters at iteration t, and $|\sum_{i=1}^{p} \rho_i\mu_i + f_{\mathbf{FDA}}(\mathbf{W}_t)|/\sum_{i=1}^{p} \rho_i\mu_i$ for FDA where $\mathbf{W}_t$ represents the parameters at iteration t. We use MANOPT's default stopping criteria: the optimization process terminates if the norm of the Riemannian gradient drops below 10^{-6}. We cap the number of iterations by 1000.

We use in our experiments three popular data sets: MNIST (Figures 1 and 2), MEDIANILL (Figure 3) and COVTYPE (Figure 4)[5]. MNIST is used for testing CCA and FDA, where for CCA we try to correlate the left side of the image to the right side of the image. MEDIANILL ($43,907$ examples) is a multilabel data set, so we use it to test CCA. COVTYPE is a large ($581,012$ examples) labeled data set, and we use it to test FDA.

Consider Figure 1 (CCA on MNIST). For Riemannian CG the number of products per iteration is never bigger than 20. As a reference, for CCA the number of iterations required for CG and Trust-Region with $\mathbf{M} = \Sigma$ is 218 and 15 correspondingly, whereas without a preconditioner ($\mathbf{M} = \mathbf{I}_d$) the CG does not converge even after 1000 iterations and the Trust-Region required 21 iterations to converge. We clearly see the direct correspondence between sketch quality (as measured by the sketch size s) and number of iterations. Furthermore, the number of iterations is close to optimal after sketching to only $s = 2000$ rows (there are

4. MANOPT has an implementation of the generalized Stiefel manifold, but only with the standard metric $\mathbf{M} = \mathbf{B}$.
5. data sets were downloaded for libsvm's website: https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.

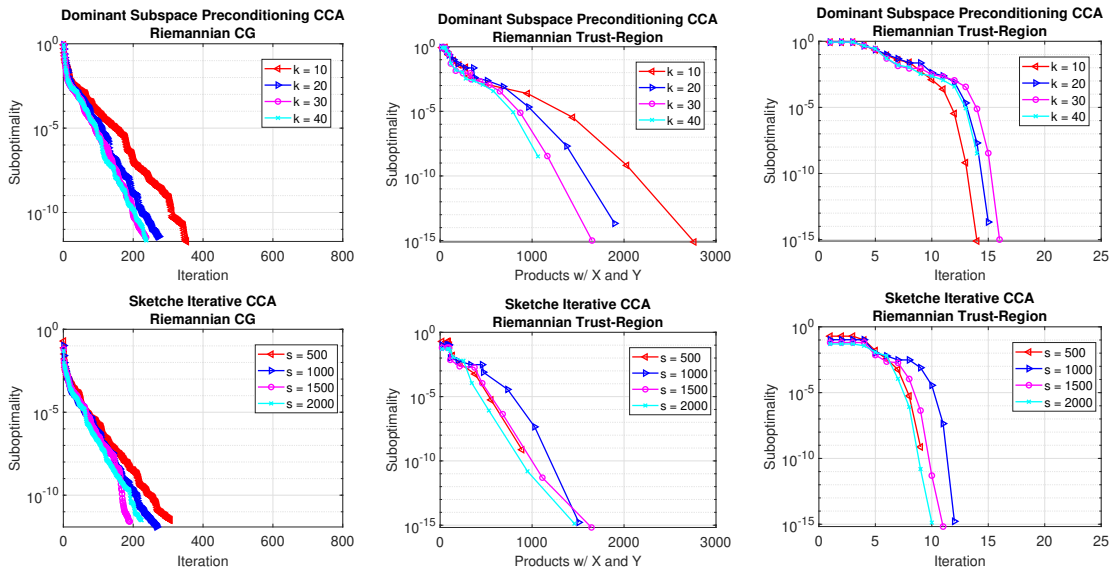

Figure 1: Results for CCA on MNIST. For CountSketch the number of rows is s, and k is the number of singular vectors for the dominant subspace preconditioner.

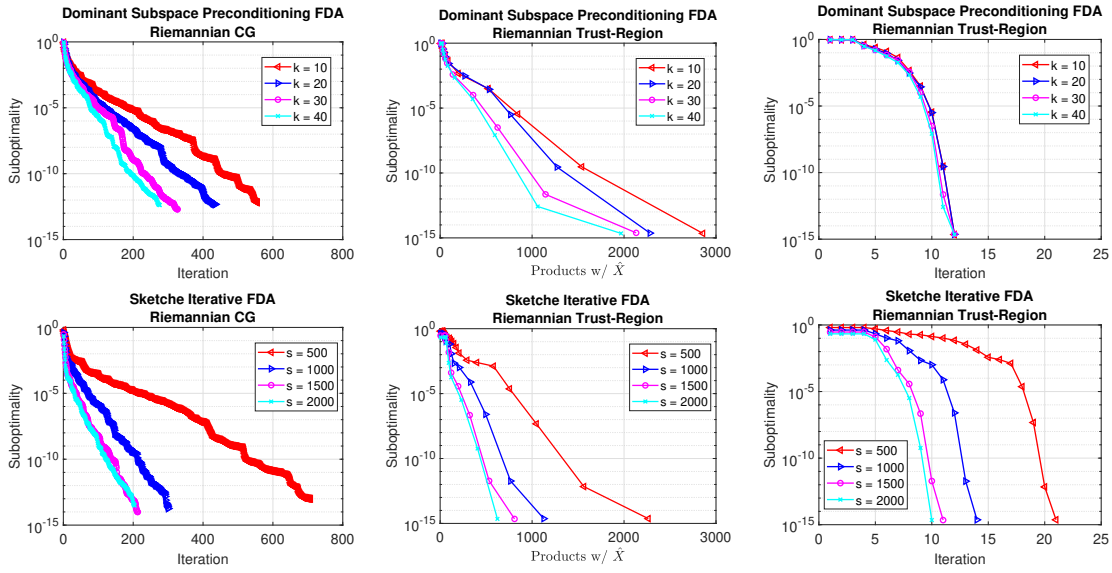

Figure 2: Results for FDA on MNIST. For CountSketch the number of rows is s, and k is the number of singular vectors for the dominant subspace preconditioner.

60,000 examples in the original data set) or using only 40 singular vectors for the dominant subspace preconditioner (there are 784 features in the data set)[6].

6. Interestingly, with $s \geq 500$ the subspsace embedding preconditioner uses less iterations than the optimal preconditioner. This is because of the use of sketching based warm-start.

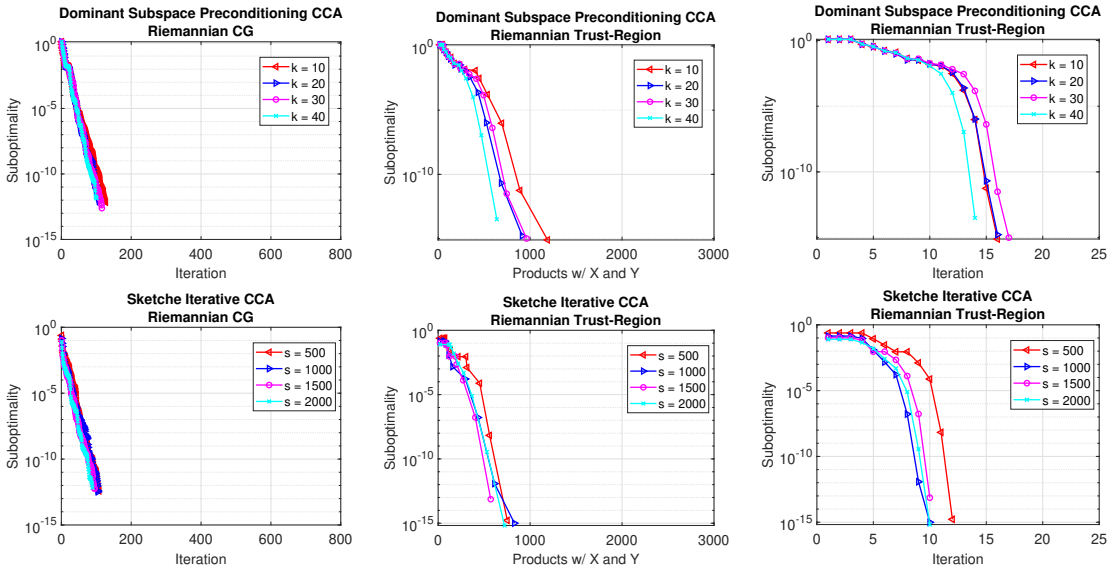

Figure 3: Results for CCA on MEDIAMILL. For CountSketch the number of rows is s, and k is the number of singular vectors for the dominant subspace preconditioner.

Consider Figure 2 (FDA on MNIST). For Riemannian CG the number of products per iteration is never bigger than 9. As a reference, for FDA the number of iterations required for CG and Trust-Region with $\mathbf{M} = \mathbf{S_w} + \lambda \mathbf{I}_d$ is 95 and 11 correspondingly, whereas without a preconditioner ($\mathbf{M} = \mathbf{I}_d$) the CG does not converge even after 1000 iterations and the Trust-Region required 20 iterations to converge, so again we see that the sketching reduces the number of iterations.

Consider Figure 3 (CCA on MEDIANILL). For Riemannian CG the number of products per iteration is never bigger than 20. As a reference, the number of iterations required for CG and Trust-Region with $\mathbf{M} = \Sigma$ is 107 and 15 correspondingly, whereas without a preconditioner ($\mathbf{M} = \mathbf{I}_d$) the number of iterations for CG is 550 and for Trust-Region is 23. The data set has 30,993 examples and 221 features, so again we see that we can sketch to relatively small size ($s = 2000$ or $k = 40$) and get an effective preconditioner.

Consider Figure 4 (FDA on COVTYPE). For Riemannian CG the number of products per iteration is never bigger than 10. As a reference, the number of iterations required for CG and Trust-Region with $\mathbf{M} = \mathbf{S_w} + \lambda \mathbf{I}_d$ is 71 and 19 correspondingly, whereas without a preconditioner ($\mathbf{M} = \mathbf{I}_d$) the CG does not converge even after 1000 iterations and the Trust-Region required 44 iterations to converge. Considering that the data set has over half a million examples, subspace embedding preconditioning is highly effective, as it sketches the data to a comparatively very small size. The data set has only 50 features, so dominant subspace preconditioning is less effective for this data set.

6.2 Synthetic Experiments and Comparison with ALS

In addition, we report a set of synthetic experiments for CCA with comparisons to other methods. To perform comparisons to existing methods, we conducted a set of synthetic

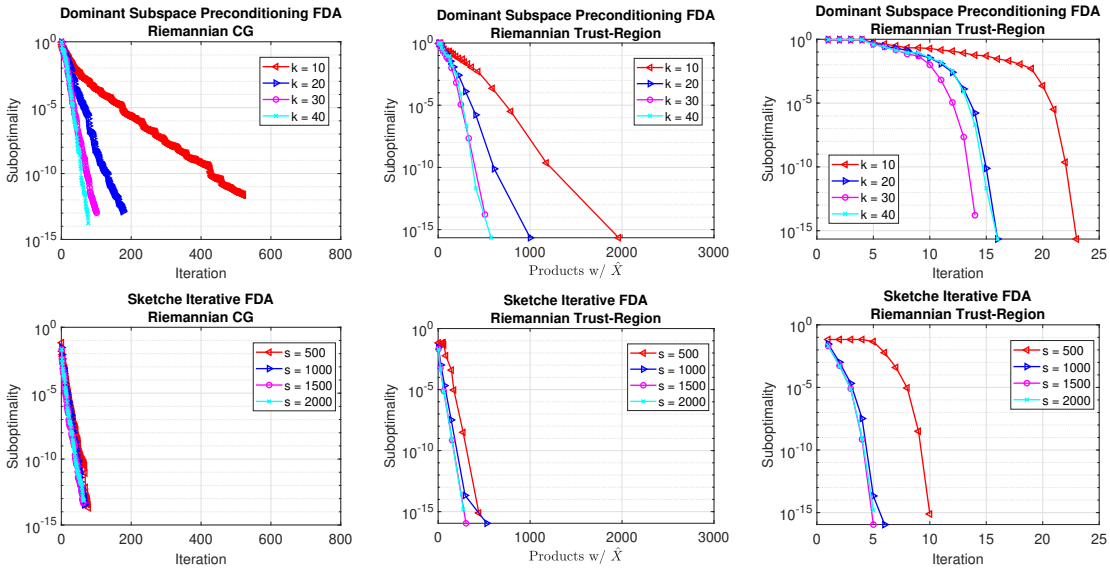

Figure 4: Results for FDA on COVTYPE. For CountSketch the number of rows is s, and k is the number of singular vectors for the dominant subspace preconditioner.

experiments in which we compared the running times of our algorithms for CCA to ALS algorithm for CCA (Ge et al., 2016a, Algorithm 1), that is (Golub and Zha, 1995, Algorithm 5.2). Note that the aforementioned algorithm is equivalent to a Riemannian gradient descent with a standard metric and a step-size of $\alpha_t = (\mathbf{u}_t^{\mathrm{T}} \Sigma_{\mathbf{xy}} \mathbf{v}_t)^{-1}$ at each iteration t (this is valid whenever $\alpha_0 = (\mathbf{u}_0^{\mathrm{T}} \Sigma_{\mathbf{xy}} \mathbf{v}_0)^{-1} > 0$), for $p = 1$. Thus, we add in Fig. 5 experiments for $p = 1$ of Riemannian gradient descent with a step-size of $\alpha_t = (\mathbf{u}_t^{\mathrm{T}} \Sigma_{\mathbf{xy}} \mathbf{v}_t)^{-1}$. We compare using the standard metric with a prescribed step-size (Golub and Zha, 1995, Algorithm 5.2) vs. line search and our sketched metrics. In addition, we add the running times for direct methods. For FDA, one can perform similar experiments, comparing our algorithms to the power method for $(\mathbf{S_w} + \lambda \mathbf{I}_d)^{-1/2} \mathbf{S_B} (\mathbf{S_w} + \lambda \mathbf{I}_d)^{-1/2}$, as it is equivalent to Riemannian gradient descent with the standard metric and a step-size of $\alpha_t = (\mathbf{w}_t^{\mathrm{T}} \mathbf{S_B} \mathbf{w}_t)^{-1}$ for $p = 1$ (this is valid whenever $\alpha_0 = (\mathbf{w}_0^{\mathrm{T}} \mathbf{S_B} \mathbf{w}_0)^{-1} > 0$). We do not include these experiments in the paper as the results are similar to the experiments for CCA. Moreover, we also added experiments with Riemannian trust-region in Fig. 6, to illustrate an advantage Riemannian optimization framework provides, e.g., second-order methods.

Our method is most beneficial in the regime $n \gg d$ and ill-conditioned data matrices. Thus, the experiments presented in Fig. 5 and Fig. 6 are performed as follows. A synthetic data set is generated with two ill-conditioned random data matrices $\mathbf{X} \in \mathbb{R}^{10^6 \times 500}$ and $\mathbf{Y} \in \mathbb{R}^{10^6 \times 500}$ with a condition number of the order of 10^{14} correspondingly. We use a small regularization of 10^{-6} multiplied by the average eigenvalue of the Gram matrices of $\mathbf{X}$ and $\mathbf{Y}$ correspondingly. We measure suboptimality vs. time in seconds (including pre-processing time of creating the sketched matrices) and iteration count.

In Fig. 5, we present results for ALS algorithm and gradient descent with line search with $\mathbf{M} = \mathbf{I}$, i.e., no preconditioning, $\mathbf{M} = \Sigma$, i.e., optimal preconditioning, and four sketched matrices with CountSketch of sizes $s = 10^3, 10^4, 2 * 10^4, 3 * 10^4$.

ALS is the fastest (109.7 seconds). Solving with $\mathbf{M} = \Sigma$ takes 295.72 seconds including pre-processing, for $s = 10^4$ (0.1% of the data) it takes 245.99 seconds including pre-processing, for $s = 2*10^4$ (0.2% of the data) it takes 231.62 seconds including pre-processing, and for $s = 3 * 10^4$ (0.3% of the data) it takes 202.64 seconds including pre-processing. In particular, excluding pre-processing time, solving with $\mathbf{M} = \Sigma$ takes 289.5 seconds, whereas, for $s = 10^4$ it takes 241.77 seconds, for $s = 2 * 10^4$ it takes 227.74 seconds, for $s = 3 * 10^4$ it takes 197.74 seconds. Note that the pre-processing time is almost identical and faster for the sketched matrices since storing and computing the full data matrices takes considerable time. For comparison, solving directly using (Björck and Golub, 1973, Section 5) takes 216.33 seconds. Note that for $s = 3 * 10^4$ the iterative method is faster, due to the larger n. Note that the condition number of the Riemannian Hessian at the optimum changes according to the preconditioning, as the ill-conditioning of the problem arises from the data matrices. For $\mathbf{M} = \mathbf{I}$ the condition number computed via Manopt is $2.66 * 10^5$, for $s = 10^3$ it is 34.1, however for $s = 10^4$ it is reduced to 2.48, for $s = 2 * 10^4$ it becomes 1.89, for $s = 3*10^4$ it becomes 1.68, and for $\mathbf{M} = \Sigma$ it becomes 1.09. Indeed, we can observe that our randomized preconditioning scheme behaves as theory predicts, having better conditioning of the Riemannian Hessian at the optimum, as our sketched matrices approximate better and better the standard metric. Moreover, as seen in these results, for very large matrices, sketching indeed accelerates the algorithms, and results in faster methods.

In Fig. 6, we present Riemannian trust-region compared to ALS, with the metric $\mathbf{M} = \Sigma$, and the metrics achieved by sketching with $s = 10^3$ and $s = 10^4$ (results for $\mathbf{M} = \mathbf{I}$ were very similar to $s = 10^3$). All the iterative methods were faster than solving directly using (Björck and Golub, 1973, Section 5) (201.78 seconds). While $s = 10^3$ achieved the slowest convergence (113.32 seconds including preprocessing) among the iterative methods, due to its ill-conditioning, both $\mathbf{M} = \Sigma$ and $s = 10^4$ were faster than the ALS, and $s = 10^4$ was the fastest (45.18 seconds including preprocessing, and similar runtime without preprocessing of the algorithm to $\mathbf{M} = \Sigma$).

7. Conclusions

In this paper we propose faster randomized methods for orthogonality constrained problems. Our method is specifically designed for typical structure of orthogonality constraints in machine learning, i.e., the constraints are defined by a Gram matrix of a data matrix where one dimension is much larger than the other. We use the framework of Riemannian optimization as the underlying iterative methods which we precondition by incorporating a randomized preconditioner via the Riemannian metric (Riemannian preconditioning). The aforementioned technique can be used to precondition any core Riemannian 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 and proposed a preconditioning strategy for two problems: CCA and FDA. For both of these examples we evaluate the computational costs, and bound

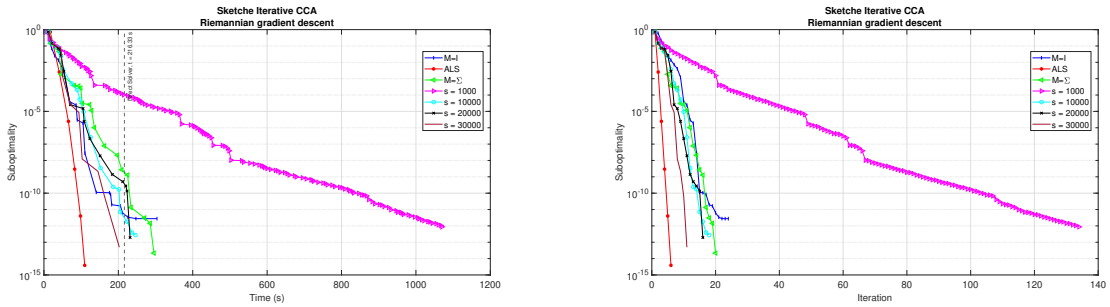

Figure 5: Results for CCA on a synthetic experiment. Left - suboptimality vs. time in seconds. Right - suboptimality vs. iteration count. we use CountSketch as the sketching transform, where s is the number of rows after sketching is applied to the data matrices.

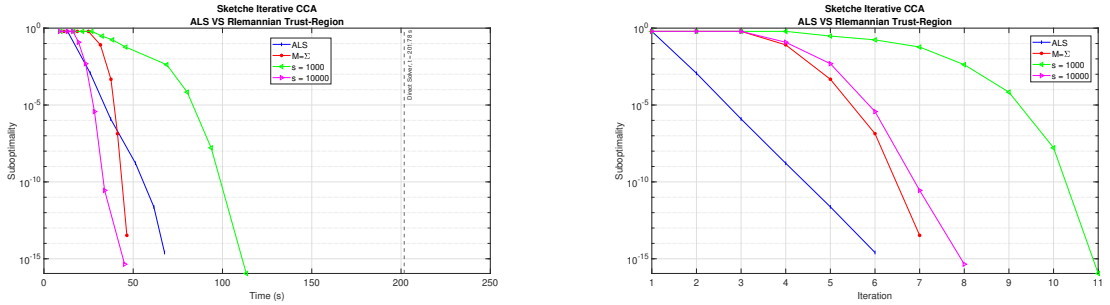

Figure 6: Results for CCA with Riemannian trust-region on a synthetic experiment. Left - suboptimality vs. time in seconds. Right - suboptimality vs. iteration count. we use CountSketch as the sketching transform, where s is the number of rows after sketching is applied to the data matrices.

the condition number of the Riemanninan Hessian at the optimum. This in turn, allows us to reason about the effect of the proposed randomized preconditioner.

As a future research direction we believe our method can be extended beyond orthogonality constrained problems for other equality constraints by identifying similar problem structure.

Acknowledgments

The authors thank Bart Vandereycken for useful discussions. This research was supported by the Israel Science Foundation (grant no. 1272/17).

Appendix A. Analogue of Lemma 2 for SRHT

In this section, we formulate an analogue for Lemma 2, which was formulated for CountSketch sketching transform to SRHT. First, we recall the definition of SRHT from (Tropp, 2011, Section 1.1) and (Boutsidis and Gittens, 2013, Definition 1.2).

Definition 21 *Fix integers s and $n = 2^m$ where $s < n$, and $m = 1, 2, 3, \ldots$. An SRHT matrix is an $s \times n$ matrix of the form*

$$\mathbf{S} = \sqrt{\frac{s}{n}}\mathbf{RHD},$$

where $\mathbf{D} \in \mathbb{R}^{n\times n}$ is a random diagonal matrix whose entries are independent random signs, i.e., random variables uniformly distributed on ± 1, $\mathbf{H} \in \mathbb{R}^{n\times n}$ is a normalized Walsh-Hadamard matrix, and $\mathbf{R} \in \mathbb{R}^{s\times n}$ is a subset of s rows from the $n \times n$ identity matrix, where the rows are chosen uniformly at random and without replacement.

Next, recall that according to (Ailon and Liberty, 2009, Theorem 2.1) and (Boutsidis and Gittens, 2013, Lemma 1.3), constructing $\mathbf{S}$ and computing $\mathbf{S}x$ given $\mathbf{x} \in \mathbb{R}^n$, $s < n$, can be done in $O(n \log(s))$ operations. Thus, computing $\mathbf{SZ}$ for $\mathbf{Z} \in \mathbb{R}^{n\times d}$ takes $O(nd \log(s))$ operations (see Table 2).

Finally, we prove an analogue of Lemma 2 for SRHT using the approximate matrix multiplication property for SRHT (Cohen et al., 2015, Theorem 9).

Lemma 22 *Assume that $\lambda > 0$ or that $\mathbf{Z} \in \mathbb{R}^{n\times d}$ has full column rank. Suppose that $\mathbf{S} \in \mathbb{R}^{s\times n}$ is an SRHT matrix with*

$$s = \Omega(4s_\lambda(\mathbf{Z})^2(s_\lambda(\mathbf{Z})/\|\mathbf{Q}\|_2^2 + \log(2s_\lambda(\mathbf{Z})/\delta)\log(s_\lambda(\mathbf{Z})/(\|\mathbf{Q}\|_2^2\delta)))),$$

where $\mathbf{Q}$ is the matrix $\mathbf{Q}$ in the λ-QR factorization of $\mathbf{Z}$ (Avron et al., 2017, Definition 28), for some $\delta \in (0, 1)$. Then with probability of at least $1 - \delta$ we have that all the generalized eigenvalues of the pencil $(\mathbf{Z}^\mathrm{T}\mathbf{Z} + \lambda\mathbf{I}, \mathbf{Z}^\mathrm{T}\mathbf{S}^\mathrm{T}\mathbf{SZ} + \lambda\mathbf{I})$ are contained in the interval $[1/2, 3/2]$ and $\kappa(\mathbf{Z}^\mathrm{T}\mathbf{Z} + \lambda\mathbf{I}, \mathbf{Z}^\mathrm{T}\mathbf{S}^\mathrm{T}\mathbf{SZ} + \lambda\mathbf{I}) \le 3$.

Proof The proof is similar to the proof of Lemma 2. To show that the generalized eigenvalues of the pencil $(\mathbf{Z}^\mathrm{T}\mathbf{Z} + \lambda\mathbf{I}, \mathbf{Z}^\mathrm{T}\mathbf{S}^\mathrm{T}\mathbf{SZ} + \lambda\mathbf{I})$ are contained in the interval $[1/2, 3/2]$ and that $\kappa(\mathbf{Z}^\mathrm{T}\mathbf{Z} + \lambda\mathbf{I}, \mathbf{Z}^\mathrm{T}\mathbf{S}^\mathrm{T}\mathbf{SZ} + \lambda\mathbf{I}) \le 3$ to hold, it is enough to show that

$$\frac{1}{2}(\mathbf{Z}^\mathrm{T}\mathbf{Z} + \lambda\mathbf{I}) \preceq \mathbf{Z}^\mathrm{T}\mathbf{S}^\mathrm{T}\mathbf{SZ} + \lambda\mathbf{I} \preceq \frac{3}{2}(\mathbf{Z}^\mathrm{T}\mathbf{Z} + \lambda\mathbf{I}).$$

Let $\mathbf{Z} = \mathbf{QR}$ be a λ-QR factorization of $\mathbf{Z}$ (Avron et al., 2017, Definition 28). Left-multiplying by $\mathbf{R}^{-\mathrm{T}}$ and right-multiplying by $\mathbf{R}^{-1}$ on both sizes, we find it suffices to show that with probability of at least $1 - \delta$ we have

$$\frac{1}{2}\mathbf{I}_d \preceq \mathbf{Q}^\mathrm{T}\mathbf{S}^\mathrm{T}\mathbf{SQ} + \lambda\mathbf{R}^{-\mathrm{T}}\mathbf{R}^{-1} \preceq \frac{3}{2}\mathbf{I}_d$$

or, equivalently,

$$\|\mathbf{Q}^\mathrm{T}\mathbf{S}^\mathrm{T}\mathbf{SQ} + \lambda\mathbf{R}^{-\mathrm{T}}\mathbf{R}^{-1} - \mathbf{I}_d\|_2 \le \frac{1}{2}.$$

Since $\mathbf{Q}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Q} + \lambda\mathbf{R}^{-\mathrm{T}}\mathbf{R}^{-1} - \mathbf{I}_d = \mathbf{Q}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Q} - \mathbf{Q}^{\mathrm{T}}\mathbf{Q}$ (Avron et al., 2017, Fact 29), it is enough to show that

$$\|\mathbf{Q}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Q} - \mathbf{Q}^{\mathrm{T}}\mathbf{Q}\|_2 \leq \frac{1}{2}\,.$$

From (Cohen et al., 2015, Theorem 9), we have that for any fixed matrix $\mathbf{A}$, and an SRHT matrix $\mathbf{S}_0$ with $m = \Omega((\|\mathbf{A}\|_F^2/\|\mathbf{A}\|_2^2 + \log(1/(\varepsilon\delta))\log(\|\mathbf{A}\|_F^2/(\|\mathbf{A}\|_2^2\delta)))/\varepsilon^2)$ rows, taking $k \coloneqq \|\mathbf{A}\|_F^2/\|\mathbf{A}\|_2^2$ in (Cohen et al., 2015, Theorem 9), we have that with probability of at least $1-\delta$,

$$\|\mathbf{A}^{\mathrm{T}}\mathbf{S}_0^{\mathrm{T}}\mathbf{S}_0\mathbf{A} - \mathbf{A}^{\mathrm{T}}\mathbf{A}\|_2 \leq \epsilon \cdot \|\mathbf{A}\|_2^2 \leq \epsilon \cdot \|\mathbf{A}\|_F^2\,.$$

Since $\|\mathbf{Q}\|_F^2 = s_\lambda(\mathbf{Z})$ (Avron et al., 2017, Fact 30), then with

$$s = \Omega(4s_\lambda(\mathbf{Z})^2(s_\lambda(\mathbf{Z})/\|\mathbf{Q}\|_2^2 + \log(2s_\lambda(\mathbf{Z})/\delta)\log(s_\lambda(\mathbf{Z})/(\|\mathbf{Q}\|_2^2\delta)))),$$

we have

$$\|\mathbf{Q}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Q} - \mathbf{Q}^{\mathrm{T}}\mathbf{Q}\|_2 \leq \frac{1}{2}$$

with probability of at least $1-\delta$. ■

Appendix B. Omitted Proofs From Section 4

In this section we give omitted proofs from Section 4.

B.1 Proof of Theorem 5

Proof Recall that the critical points are defined to be the points where the Riemannian gradient is zero, and as such, whether a point is a critical point or not does not depend on the choice of Riemannian metric (see Absil et al., 2008, Eq. 3.31). Thus, for the sake of identifying the critical points, we can use $\mathbf{M}_\mathbf{Z} \coloneqq \Sigma$ for all $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$ and get a simplified form for the Riemannian gradient (Eq. (4)):

$$\begin{aligned}
\mathbf{grad} f_{\mathbf{CCA}}(\mathbf{Z}) &= -\begin{bmatrix} \Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\mathbf{VN} - \mathbf{Usym}\left(\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xx}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\mathbf{VN}\right) \\ \Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{UN} - \mathbf{Vsym}\left(\mathbf{V}^{\mathrm{T}}\Sigma_{\mathbf{yy}}\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{UN}\right) \end{bmatrix} \\
&= -\begin{bmatrix} \left(\mathbf{I}_{d_\mathbf{x}} - \mathbf{U}\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xx}}\right)\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\mathbf{VN} + \mathbf{Uskew}(\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{VN}) \\ \left(\mathbf{I}_{d_\mathbf{y}} - \mathbf{V}\mathbf{V}^{\mathrm{T}}\Sigma_{\mathbf{yy}}\right)\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{UN} + \mathbf{Vskew}(\mathbf{V}^{\mathrm{T}}\Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{UN}) \end{bmatrix}.
\end{aligned}$$

There is a connection between the canonical correlations and the singular value decomposition of $\mathbf{R} \coloneqq \Sigma_{\mathbf{xx}}^{-1/2}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1/2}$ (Wang et al., 2016): a pair of vectors $\mathbf{u}$ and $\mathbf{v}$ are canonical correlation vectors corresponding to the same canonical correlation if and only if $\tilde{\mathbf{u}} = \Sigma_{\mathbf{xx}}^{1/2}\mathbf{u}$ and $\tilde{\mathbf{v}} = \Sigma_{\mathbf{yy}}^{1/2}\mathbf{v}$ are left and right singular vectors of $\mathbf{R}$ corresponding to the same singular value.

canonical correlation vectors are the columns of $\mathbf{U}$ and $\mathbf{V}$ it and only if the columns of $(\tilde{\mathbf{U}}, \tilde{\mathbf{V}}) = (\Sigma_{\mathbf{xx}}^{1/2}\mathbf{U}, \Sigma_{\mathbf{yy}}^{1/2}\mathbf{V})$ are p left and right coordinated singular vectors of the matrix $\mathbf{R}$. Thus, we use this relation to prove this theorem.

To show that all $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$ such that the columns of $\mathbf{U}$ and $\mathbf{V}$ are left and right coordinated canonical correlation vectors not necessarily on the same phase are critical

points, let $\alpha_1, \ldots, \alpha_p$ be some singular values of $\mathbf{R}$, and let $\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_p, \tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_p$ be the corresponding left and right singular vectors not necessarily on the same phase. Writing $\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_p$ as the columns of $\tilde{\mathbf{U}}$ and and $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_p$ as the columns of $\tilde{\mathbf{V}}$, and defining $\mathbf{U} = \Sigma_{\mathbf{xx}}^{-1/2}\tilde{\mathbf{U}}, \mathbf{V} = \Sigma_{\mathbf{xx}}^{-1/2}\tilde{\mathbf{V}}$, the following two equations hold:

$$\Sigma_{\mathbf{xy}}\mathbf{V} = \Sigma_{\mathbf{xx}}\mathbf{U}\mathbf{A}, \quad \Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{U} = \Sigma_{\mathbf{yy}}\mathbf{V}\mathbf{A}$$

where $\mathbf{A} \coloneqq \mathbf{diag}\left(\beta_1, \ldots, \beta_p\right)$ such that $|\beta_i| = \alpha_i$ for $i = 1, ..., p$. Letting $\mathbf{Z} = (\mathbf{U}, \mathbf{V}) \in \mathbb{S}_{\mathbf{xy}}$, we have

$$\mathbf{grad} f_{\mathbf{CCA}}(\mathbf{Z}) = -\begin{bmatrix} \mathbf{U}\mathbf{A}\mathbf{N} - \mathbf{U}\mathbf{sym}\left(\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xx}}\mathbf{U}\mathbf{A}\mathbf{N}\right) \\ \mathbf{V}\mathbf{A}\mathbf{N} - \mathbf{V}\mathbf{sym}\left(\mathbf{V}^{\mathrm{T}}\Sigma_{\mathbf{yy}}\mathbf{V}\mathbf{A}\mathbf{N}\right) \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{d_{\mathbf{x}} \times p} \\ \mathbf{0}_{d_{\mathbf{y}} \times p} \end{bmatrix}$$

where the last line is true since $\mathbf{AN}$ is diagonal.

To show the other side, note that if the Riemannian gradient nullifies, then

$$\begin{bmatrix} \left(\mathbf{I}_{d_{\mathbf{x}}} - \mathbf{U}\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xx}}\right)\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\mathbf{V}\mathbf{N} + \mathbf{U}\mathbf{skew}(\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{V}\mathbf{N}) \\ \left(\mathbf{I}_{d_{\mathbf{y}}} - \mathbf{V}\mathbf{V}^{\mathrm{T}}\Sigma_{\mathbf{yy}}\right)\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{U}\mathbf{N} + \mathbf{V}\mathbf{skew}(\mathbf{V}^{\mathrm{T}}\Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{U}\mathbf{N}) \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{d_{\mathbf{x}} \times p} \\ \mathbf{0}_{d_{\mathbf{y}} \times p} \end{bmatrix}.$$

By using similar reasoning as in (Absil et al., 2008, Subsection 4.8.2),

$$(\mathbf{I}_{d_{\mathbf{x}}} - \mathbf{U}\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xx}})\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\mathbf{V}\mathbf{N},$$

belongs to the orthogonal compliment of the column space of $\mathbf{U}$ (with respect to the matrix $\Sigma_{\mathbf{xx}}$), and $\mathbf{U}\mathbf{skew}(\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{V}\mathbf{N})$ belongs to the column space of $\mathbf{U}$. The same is true for $(\mathbf{I}_{d_{\mathbf{y}}} - \mathbf{V}\mathbf{V}^{\mathrm{T}}\Sigma_{\mathbf{yy}})\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{U}\mathbf{N}$ and $\mathbf{V}\mathbf{skew}(\mathbf{V}^{\mathrm{T}}\Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{U}\mathbf{N})$ with respect to $\mathbf{V}$ and $\Sigma_{\mathbf{yy}}$. Thus, we get that the gradient vanishes if and only if the aforementioned four factors vanish.

From $\left(\mathbf{I}_{d_{\mathbf{x}}} - \mathbf{U}\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xx}}\right)\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\mathbf{V}\mathbf{N} = \mathbf{0}_{d_{\mathbf{x}} \times p}$ we get $\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\mathbf{V} = \mathbf{U}\left(\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{V}\right)$, since $\mathbf{N}$ is an invertible matrix. Also, since $\mathbf{U} \in \mathbf{St}_{\Sigma_{\mathbf{xx}}}(p, d_{\mathbf{x}})$ it is a full (column) rank matrix then $\mathbf{U}\mathbf{skew}(\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{V}\mathbf{N})$ vanishes if and only if $\mathbf{skew}(\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{V}\mathbf{N})$ vanishes, which leads to $\left(\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{V}\right)\mathbf{N} = \mathbf{N}\left(\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{V}\right)$. This implies that $\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{V}$ is diagonal because any rectangular matrix that commutes with a diagonal matrix with distinct entries is diagonal. Thus, we have

$$\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\mathbf{V} = \mathbf{U}\mathbf{D}\ , \tag{11}$$

where $\mathbf{D} = \mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{V}$ is diagonal. Similarly, we have

$$\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{U} = \mathbf{V}\mathbf{D}\ , \tag{12}$$

where $\mathbf{D} = \mathbf{V}^{\mathrm{T}}\Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{U} = \mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{V}$ is diagonal. From Eq. (11) and Eq. (12) we get

$$\mathbf{R}\Sigma_{\mathbf{yy}}^{1/2}\mathbf{V} = \Sigma_{\mathbf{xx}}^{1/2}\mathbf{U}\mathbf{D},\ \mathbf{R}^{\mathrm{T}}\Sigma_{\mathbf{xx}}^{1/2}\mathbf{U} = \Sigma_{\mathbf{yy}}^{1/2}\mathbf{V}\mathbf{D}.$$

This implies that the columns of $(\tilde{\mathbf{U}}, \tilde{\mathbf{V}}) = (\Sigma_{\mathbf{xx}}^{1/2}\mathbf{U}, \Sigma_{\mathbf{yy}}^{1/2}\mathbf{V})$ are some p left and right singular vectors of $\mathbf{R}$ not necessarily on the same phase, but corresponding to the same singular values.

Finally, to identify the optimal solutions, note that at the critical points the objective function is the sum of the canonical correlations multiplied by a diagonal element of $\mathbf{N}$, and

a sign corresponding to the canonical correlation vectors in the columns of $\mathbf{U}$ and $\mathbf{V}$ and the correspondence of their phase. Thus, the optimal solutions that minimize $f_{\mathbf{CCA}}(\mathbf{Z})$ on $\mathbb{S}_{\mathbf{xy}}$ are $\mathbf{Z} = (\mathbf{U}, \mathbf{V}) \in \mathbb{S}_{\mathbf{xy}}$ such that the columns of $\mathbf{U}$ and $\mathbf{V}$ correspond to the top p-canonical correlations on the *same phase*. Otherwise, we can increase the value of the objective function either by flipping the sign of one of the vectors or by replacing a canonical vector with another that corresponds to a smaller canonical correlation. Moreover, if we assume that $\sigma_1 > \sigma_2 > ... > \sigma_{p+1} \geq 0$, then for the aforementioned $\mathbf{Z} = (\mathbf{U}, \mathbf{V}) \in \mathbb{S}_{\mathbf{xy}}$, the columns of $(\tilde{\mathbf{U}}, \tilde{\mathbf{V}})$ belong each to a one dimensional singular left and right space and so do the columns of the corresponding $(\mathbf{U}, \mathbf{V})$, i.e., unique solution up the the signs of the columns of $(\mathbf{U}, \mathbf{V})$. In the case where some $\sigma_i = \sigma_j$ for $1 \leq i, j \leq p$, permutations of the columns of $\mathbf{U}$ and $\mathbf{V}$ associated with σ_i keep the solution optimal making it non-unique. ■

B.2 Proof of Theorem 6

Proof To prove the asymptotic stability of $\mathbf{Z} = (\mathbf{U}, \mathbf{V})$ that minimize $f_{\mathbf{CCA}}(\mathbf{Z})$ on $\mathbb{S}_{\mathbf{xy}}$, we use (Absil et al., 2008, Proposition 4.4.2). Recall from Theorem 5 that $\mathbf{Z}$ that solve Problem (3) are unique up the the signs of the columns of $\mathbf{U}$ and $\mathbf{V}$, making these points isolated global (and consequently local) minimizers of $f_{\mathbf{CCA}}(\mathbf{Z})$ on $\mathbb{S}_{\mathbf{xy}}$. According to (Absil et al., 2008, Proposition 4.4.2), such points $\mathbf{Z}$ are asymptotically stable.

Suppose $\mathbf{Z}$ is a critical point of $f_{\mathbf{CCA}}(\mathbf{Z})$ on $\mathbb{S}_{\mathbf{xy}}$ which is not a local minimum. Then, there exists compact neighborhoods with either no other critical points, if there are no multiplicities of the canonical correlations, or where all other critical point achieve the same value for the cost function, if there are multiplicities. Thus, according to (Absil et al., 2008, Proposition 4.4.1), such $\mathbf{Z}$ are unstable. ■

B.3 Proof of Theorem 9

Proof In order to bound the condition number of the Riemannian Hessian at $\mathbf{Z}^\star \in \mathbb{S}_{\mathbf{xy}}$, we need to bound its maximal and minimal eigenvalues. Thus, to prove the theorem we analyze the eigenvalues of the Riemannian Hessian at some critical point $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$ (in particular at $\mathbf{Z}^\star \in \mathbb{S}_{\mathbf{xy}}$) using the Courant-Fischer Theorem (also called the minimax principle, see (Kato, 2013, Chapter 1, Section 6.10)) for the compact self-adjoint linear operator $\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z})[\cdot] : T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}} \to T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}}$ over the finite dimensional vector space $T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}}$:

$$\lambda_k(\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z})) = \min_{U, \dim(U)=k-1} \max_{\mathbf{0} \neq \xi_{\mathbf{Z}} \in U^\perp} R(\xi_{\mathbf{Z}}), \tag{13}$$

$$\lambda_k(\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z})) = \max_{U, \dim(U)=k} \min_{\mathbf{0} \neq \xi_{\mathbf{Z}} \in U} R(\xi_{\mathbf{Z}}), \tag{14}$$

where

$$R(\xi_{\mathbf{Z}}) \coloneqq \frac{g_{\mathbf{Z}}(\xi_{\mathbf{Z}}, \mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z})[\xi_{\mathbf{Z}}])}{g_{\mathbf{Z}}(\xi_{\mathbf{Z}}, \xi_{\mathbf{Z}})},$$

In the above, $\lambda_k(\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z}))$ is the kth largest eigenvalue (i.e., eigenvalues are ordered in a descending order) of $\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z})$, and U is a linear subspace of $T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}}$. In particular,

the maximal and minimal eigenvalues are given by the formulas

$$\lambda_{\max}(\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z})) = \max_{\mathbf{0}\neq\xi_{\mathbf{Z}}\in T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}}} R(\xi_{\mathbf{Z}}), \tag{15}$$

$$\lambda_{\min}(\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z})) = \min_{\mathbf{0}\neq\xi_{\mathbf{Z}}\in T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}}} R(\xi_{\mathbf{Z}}). \tag{16}$$

We begin by simplifying the quotient $R(\xi_{\mathbf{Z}})$. Recall that any critical point of $f_{\mathbf{CCA}}$ is a matrix $\mathbf{Z} = (\mathbf{U}, \mathbf{V}) \in \mathbb{S}_{\mathbf{xy}}$ such that the columns of $(\tilde{\mathbf{U}}, \tilde{\mathbf{V}}) = (\Sigma_{\mathbf{xx}}^{1/2}\mathbf{U}, \Sigma_{\mathbf{yy}}^{1/2}\mathbf{V})$ are p left and right singular vectors (not necessarily on the same phase) of the matrix $\mathbf{R} \coloneqq \Sigma_{\mathbf{xx}}^{-1/2}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1/2}$ (see Theorem 5). Let $\alpha_1, \ldots, \alpha_p$ be some singular values of the matrix $\mathbf{R}$, and let $\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_p, \tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_p$ be the corresponding left and right singular vectors (not necessarily on the same phase). Writing $\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_p$ as the columns of $\tilde{\mathbf{U}}$ and and $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_p$ as the columns of $\tilde{\mathbf{V}}$, and defining $\mathbf{U} = \Sigma_{\mathbf{xx}}^{-\frac{1}{2}}\tilde{\mathbf{U}}, \mathbf{V} = \Sigma_{\mathbf{xx}}^{-\frac{1}{2}}\tilde{\mathbf{V}}$, the following two equations hold:

$$\Sigma_{\mathbf{xy}}\mathbf{V} = \Sigma_{\mathbf{xx}}\mathbf{U}\mathbf{A} \tag{17}$$

$$\Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{U} = \Sigma_{\mathbf{yy}}\mathbf{V}\mathbf{A} \tag{18}$$

where $\mathbf{A} \coloneqq \mathbf{diag}\,(\beta_1, \ldots, \beta_p)$ such that $|\beta_i| = \alpha_i$ for $i = 1, ..., p$. Letting $\mathbf{Z} = (\mathbf{U}, \mathbf{V}) \in \mathbb{S}_{\mathbf{xy}}$, plugging in the ambient coordinates formula for the Riemannian Hessian (Eq. (5)), the Riemannian gradient nullifies (see Theorem 5) and we have

$$\begin{aligned}\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z})[\xi_{\mathbf{Z}}] = \mathbf{\Pi}_{\mathbf{Z}}(\mathbf{M}_{\mathbf{Z}}^{-1}[-\Sigma_{\nabla^2 f_{\mathbf{CCA}}}\xi_{\mathbf{Z}}\mathbf{N} + \Sigma \begin{bmatrix} \xi_{\mathbf{U}}\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{V}\mathbf{N} \\ \xi_{\mathbf{V}}\mathbf{V}^{\mathrm{T}}\Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{U}\mathbf{N} \end{bmatrix} \\ + \Sigma \begin{bmatrix} \xi_{\mathbf{U}}\mathbf{U}^{\mathrm{T}} & \\ & \xi_{\mathbf{V}}\mathbf{V}^{\mathrm{T}} \end{bmatrix} \mathbf{M}_{\mathbf{Z}}\mathbf{grad} f_{\mathbf{CCA}}(\mathbf{Z})]) = \mathbf{\Pi}_{\mathbf{Z}}\left(\mathbf{M}_{\mathbf{Z}}^{-1}\left[-\Sigma_{\nabla^2 f_{\mathbf{CCA}}}\xi_{\mathbf{Z}} + \Sigma\xi_{\mathbf{Z}}\mathbf{A}\right]\mathbf{N}\right).\end{aligned} \tag{19}$$

Plugging in the formula for the Riemannian Hessian at a critical point (Eq. (19)), the quotient $R(\xi_{\mathbf{Z}})$ is reduced to to

$$R(\xi_{\mathbf{Z}}) = \frac{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\mathbf{M}_{\mathbf{Z}}\mathbf{\Pi}_{\mathbf{Z}}\left(\mathbf{M}_{\mathbf{Z}}^{-1}\left[-\Sigma_{\nabla^2 f_{\mathbf{CCA}}}\xi_{\mathbf{Z}} + \Sigma\xi_{\mathbf{Z}}\mathbf{A}\right]\mathbf{N}\right)\right)}{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\mathbf{M}_{\mathbf{Z}}\xi_{\mathbf{Z}}\right)}.$$

Now, using the fact that the projection to the tangent space is self-adjoint with respect to the Riemannian metric and that for any $\xi_{\mathbf{Z}} \in T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}}$ we have $\mathbf{\Pi}_{\mathbf{Z}}(\xi_{\mathbf{Z}}) = \xi_{\mathbf{Z}}$, we further see that

$$\frac{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\mathbf{M}_{\mathbf{Z}}\mathbf{\Pi}_{\mathbf{Z}}\left(\mathbf{M}_{\mathbf{Z}}^{-1}\left[-\Sigma_{\nabla^2 f_{\mathbf{CCA}}}\xi_{\mathbf{Z}} + \Sigma\xi_{\mathbf{Z}}\mathbf{A}\right]\mathbf{N}\right)\right)}{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\mathbf{M}_{\mathbf{Z}}\xi_{\mathbf{Z}}\right)} = \frac{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\left(-\Sigma_{\nabla^2 f_{\mathbf{CCA}}}\xi_{\mathbf{Z}} + \Sigma\xi_{\mathbf{Z}}\mathbf{A}\right)\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\mathbf{M}_{\mathbf{Z}}\xi_{\mathbf{Z}}\right)}.$$

Obviously, we can also write

$$\begin{aligned}&\frac{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\left(-\Sigma_{\nabla^2 f_{\mathbf{CCA}}}\xi_{\mathbf{Z}} + \Sigma\xi_{\mathbf{Z}}\mathbf{A}\right)\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\mathbf{M}_{\mathbf{Z}}\xi_{\mathbf{Z}}\right)} \\ &\qquad = \frac{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\left(-\Sigma_{\nabla^2 f_{\mathbf{CCA}}}\xi_{\mathbf{Z}} + \Sigma\xi_{\mathbf{Z}}\mathbf{A}\right)\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}}\right)} \cdot \frac{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\mathbf{M}_{\mathbf{Z}}\xi_{\mathbf{Z}}\right)}.\end{aligned} \tag{20}$$

Using Eq. (20), a simplified form of the quotient $R(\xi_{\mathbf{Z}})$, we can estimate upper and lower bounds on $R(\xi_{\mathbf{Z}})$ where $\mathbf{0} \neq \xi_{\mathbf{Z}} \in T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}}$ in order to bound the condition number of the Riemannian Hessian at $\mathbf{Z}^\star \in \mathbb{S}_{\mathbf{xy}}$. Since for $\xi_{\mathbf{Z}} \neq \mathbf{0}$ the term $\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}}\right)/\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\mathbf{M}_{\mathbf{Z}}\xi_{\mathbf{Z}}\right) > 0$, then the upper and lower bounds on

$$\frac{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\left(-\Sigma_{\nabla^2 f_{\mathbf{CCA}}}\xi_{\mathbf{Z}} + \Sigma\xi_{\mathbf{Z}}\mathbf{A}\right)\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}}\right)}, \tag{21}$$

together with the upper and lower bounds of $\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}}\right)/\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\mathbf{M}_{\mathbf{Z}}\xi_{\mathbf{Z}}\right) > 0$ bound the condition number of the Riemannian Hessian at $\mathbf{Z}^\star \in \mathbb{S}_{\mathbf{xy}}$.

We begin by estimating the term $\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}}\right)/\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\mathbf{M}_{\mathbf{Z}}\xi_{\mathbf{Z}}\right) > 0$. We use the vectorization operator, and the Kronecker product to rewrite it in the following form

$$\frac{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\mathbf{M}_{\mathbf{Z}}\xi_{\mathbf{Z}}\right)} = \frac{\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)^{\mathrm{T}}\left(\mathbf{I}_p \otimes \Sigma\right)\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)}{\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)^{\mathrm{T}}\left(\mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}}\right)\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)}. \tag{22}$$

Eq. (22) is the generalized Rayleigh quotient for the matrix pencil $(\mathbf{I}_p \otimes \Sigma, \mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}})$. Note that $\mathbf{I}_p \otimes \Sigma$ and $\mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}}$ are both SPD matrices, thus the generalized eigenvalues of the matrix pencil $(\mathbf{I}_p \otimes \Sigma, \mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}})$ are equivalent to the eigenvalues of the matrix $(\mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}})^{-1}(\mathbf{I}_p \otimes \Sigma) = \mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}}^{-1}\Sigma$. According to (Minka, 2000, Section 2) the eigenvalues $\mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}}^{-1}\Sigma$ are p copies of each of the eigenvalues of $\mathbf{M}_{\mathbf{Z}}^{-1}\Sigma$. Thus, the maximal and minimal eigenvalues of the matrix pencil $(\mathbf{I}_p \otimes \Sigma, \mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}})$ denoted by $\tilde{\lambda}_{\max}$ and $\tilde{\lambda}_{\min}$ are equivalent to the maximal and minimal generalized eigenvalues of the matrix pencil $(\Sigma, \mathbf{M}_{\mathbf{Z}})$, and so is the corresponding condition number

$$\kappa\left(\mathbf{I}_p \otimes \Sigma, \mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}}\right) = \frac{\tilde{\lambda}_{\max}}{\tilde{\lambda}_{\min}} = \kappa\left(\Sigma, \mathbf{M}_{\mathbf{Z}}\right).$$

Recall the definition of the generalized eigenvalues. The generalized eigenvalues of the matrix pencil $(\mathbf{A}, \mathbf{B})$, where $\mathbf{A} \in \mathbb{R}^{d\times d}$ and $\mathbf{B} \in \mathbb{R}^{d\times d}$ is a symmetric positive semi-definite matrix such that $\ker(\mathbf{B}) \subseteq \ker(\mathbf{A})$, are defined as follows: if for $\lambda \in \mathbb{R}$ and $\mathbf{v} \notin \ker(\mathbf{B})$ it holds that $\mathbf{A}\mathbf{v} = \lambda\mathbf{B}\mathbf{v}$ then λ is a generalized eigenvalue and $\mathbf{v}$ is a generalized eigenvector of the matrix pencil $(\mathbf{A}, \mathbf{B})$. The generalized eigenvalues are denoted by $\lambda_1(\mathbf{A}, \mathbf{B}) \geq \lambda_2(\mathbf{A}, \mathbf{B}) \geq \cdots \geq \lambda_{\mathbf{rank(B)}}(\mathbf{A}, \mathbf{B})$. Therefore, using the Courant-Fischer Theorem for the matrix pencil $(\mathbf{I}_p \otimes \Sigma, \mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}})$ we have

$$\begin{aligned}
\tilde{\lambda}_{\max} &:= \lambda_{\max}(\mathbf{I}_p \otimes \Sigma, \mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}}) \\
&= \max_{0\neq\xi_{\mathbf{Z}}\in\mathbb{R}^{d\times p}} \frac{\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)^{\mathrm{T}}\left(\mathbf{I}_p \otimes \Sigma\right)\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)}{\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)^{\mathrm{T}}\left(\mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}}\right)\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)} \\
&\geq \max_{0\neq\xi_{\mathbf{Z}}\in T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}}} \frac{\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)^{\mathrm{T}}\left(\mathbf{I}_p \otimes \Sigma\right)\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)}{\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)^{\mathrm{T}}\left(\mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}}\right)\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)} \\
&= \max_{0\neq\xi_{\mathbf{Z}}\in T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}}} \frac{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\mathbf{M}_{\mathbf{Z}}\xi_{\mathbf{Z}}\right)}
\end{aligned}$$

and

$$\begin{aligned}\tilde{\lambda}_{\min} &\coloneqq \lambda_{\min}(\mathbf{I}_p \otimes \Sigma, \mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}}) \\ &= \min_{0 \neq \xi_{\mathbf{Z}} \in \mathbb{R}^{d \times p}} \frac{\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)^{\mathrm{T}} \left(\mathbf{I}_p \otimes \Sigma\right) \mathbf{vec}\left(\xi_{\mathbf{Z}}\right)}{\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)^{\mathrm{T}} \left(\mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}}\right) \mathbf{vec}\left(\xi_{\mathbf{Z}}\right)} \\ &\leq \min_{0 \neq \xi_{\mathbf{Z}} \in T_{\mathbf{Z}} \mathbb{S}_{\mathbf{xy}}} \frac{\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)^{\mathrm{T}} \left(\mathbf{I}_p \otimes \Sigma\right) \mathbf{vec}\left(\xi_{\mathbf{Z}}\right)}{\mathbf{vec}\left(\xi_{\mathbf{Z}}\right)^{\mathrm{T}} \left(\mathbf{I}_p \otimes \mathbf{M}_{\mathbf{Z}}\right) \mathbf{vec}\left(\xi_{\mathbf{Z}}\right)} \\ &= \min_{0 \neq \xi_{\mathbf{Z}} \in T_{\mathbf{Z}} \mathbb{S}_{\mathbf{xy}}} \frac{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}} \Sigma \xi_{\mathbf{Z}}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}} \mathbf{M}_{\mathbf{Z}} \xi_{\mathbf{Z}}\right)}\end{aligned}$$

Next, we analyze Eq. (21). Recall that $\xi_{\mathbf{Z}} = (\xi_{\mathbf{U}}, \xi_{\mathbf{V}}) \in T_{\mathbf{Z}} \mathbb{S}_{\mathbf{xy}}$ where we have $\xi_{\mathbf{U}} \in T_{\mathbf{U}} \mathbf{St}_{\Sigma_{\mathbf{xx}}}(p, d_{\mathbf{x}})$ and $\xi_{\mathbf{V}} \in T_{\mathbf{V}} \mathbf{St}_{\Sigma_{\mathbf{yy}}}(p, d_{\mathbf{y}})$, thus we can rewrite the tangent vectors in the following form:

$$\xi_{\mathbf{U}} = \mathbf{U}\mathbf{\Omega}_{\mathbf{U}} + \mathbf{U}_{\Sigma_{\mathbf{xx}}\perp}\mathbf{K}_{\mathbf{U}},\ \xi_{\mathbf{V}} = \mathbf{V}\mathbf{\Omega}_{\mathbf{V}} + \mathbf{V}_{\Sigma_{\mathbf{yy}}\perp}\mathbf{K}_{\mathbf{V}},$$

where $\mathbf{U}_{\Sigma_{\mathbf{xx}}\perp}$ is $\Sigma_{\mathbf{xx}}$-orthogonal to $\mathbf{U}$ so that the union of the columns of $\mathbf{U}$ and $\mathbf{U}_{\Sigma_{\mathbf{xx}}\perp}$ is a basis to $\mathbb{R}^{d_{\mathbf{x}}}$, and similarly $\mathbf{V}_{\Sigma_{\mathbf{yy}}\perp}$ is $\Sigma_{\mathbf{yy}}$-orthogonal to $\mathbf{V}$ so that the union of the columns of $\mathbf{V}$ and $\mathbf{V}_{\Sigma_{\mathbf{yy}}\perp}$ is a basis to $\mathbb{R}^{d_{\mathbf{y}}}$, $\mathbf{\Omega}_{\mathbf{U}} = -\mathbf{\Omega}_{\mathbf{U}}^{\mathrm{T}} \in \mathbb{R}^{p \times p}$, $\mathbf{\Omega}_{\mathbf{V}} = -\mathbf{\Omega}_{\mathbf{V}}^{\mathrm{T}} \in \mathbb{R}^{p \times p}$, $\mathbf{K}_{\mathbf{U}} \in \mathbb{R}^{(d_{\mathbf{x}}-p) \times p}$ and $\mathbf{K}_{\mathbf{V}} \in \mathbb{R}^{(d_{\mathbf{y}}-p) \times p}$. Note that we can always make the choice of the columns of $\mathbf{U}_{\Sigma_{\mathbf{xx}}\perp}$ and $\mathbf{V}_{\Sigma_{\mathbf{yy}}\perp}$ to be such that $(\Sigma_{\mathbf{xx}}^{1/2}\mathbf{U}_{\Sigma_{\mathbf{xx}}\perp}, \Sigma_{\mathbf{yy}}^{1/2}\mathbf{V}_{\Sigma_{\mathbf{yy}}\perp})$ are some $\min\{d_{\mathbf{x}} - p, d_{\mathbf{y}} - p\}$ left and right singular vectors not necessarily on the same phase of the matrix $\mathbf{R}$ belonging to the same singular values. Without loss of generality suppose $d_{\mathbf{x}} \geq d_{\mathbf{y}}$. With this choice we have

$$\Sigma_{\mathbf{xy}}\mathbf{V}_{\Sigma_{\mathbf{yy}}\perp} = \Sigma_{\mathbf{xx}}\mathbf{U}_{\Sigma_{\mathbf{xx}}\perp}\tilde{\mathbf{A}},\ \Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{U}_{\Sigma_{\mathbf{xx}}\perp} = \Sigma_{\mathbf{yy}}\mathbf{V}_{\Sigma_{\mathbf{yy}}\perp}\tilde{\mathbf{A}}^{\mathrm{T}},$$

and

$$\mathbf{U}_{\Sigma_{\mathbf{xx}}\perp}^{\mathrm{T}}\Sigma_{\mathbf{xy}}\mathbf{V}_{\Sigma_{\mathbf{yy}}\perp} = \tilde{\mathbf{A}} \in \mathbb{R}^{(d_{\mathbf{x}}-p)\times(d_{\mathbf{y}}-p)},\ \mathbf{V}_{\Sigma_{\mathbf{yy}}\perp}^{\mathrm{T}}\Sigma_{\mathbf{xy}}^{\mathrm{T}}\mathbf{U}_{\Sigma_{\mathbf{xx}}\perp} = \tilde{\mathbf{A}}^{\mathrm{T}} \in \mathbb{R}^{(d_{\mathbf{y}}-p)\times(d_{\mathbf{x}}-p)},$$

where $\tilde{\mathbf{A}}$ is a diagonal matrix (but not necessarily a square matrix), with the corresponding values on the diagonal $\beta_{p+1}, ..., \beta_{d_{\mathbf{y}}}$, which satisfy $|\beta_i| = \alpha_i$ for $i = p+1, ..., d_{\mathbf{y}}$.

Now,

$$\xi_{\mathbf{Z}}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}} = \begin{bmatrix} \mathbf{\Omega}_{\mathbf{U}}^{\mathrm{T}} & \mathbf{\Omega}_{\mathbf{V}}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{\Omega}_{\mathbf{U}} \\ \mathbf{\Omega}_{\mathbf{V}} \end{bmatrix} + \begin{bmatrix} \mathbf{K}_{\mathbf{U}}^{\mathrm{T}} & \mathbf{K}_{\mathbf{V}}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{K}_{\mathbf{U}} \\ \mathbf{K}_{\mathbf{V}} \end{bmatrix}$$

and

$$\xi_{\mathbf{Z}}^{\mathrm{T}}\Sigma_{\nabla^2 f_{\mathbf{CCA}}}\xi_{\mathbf{Z}} = \begin{bmatrix} \mathbf{\Omega}_{\mathbf{U}}^{\mathrm{T}} & \mathbf{\Omega}_{\mathbf{V}}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} & \mathbf{A} \\ \mathbf{A} & \end{bmatrix} \begin{bmatrix} \mathbf{\Omega}_{\mathbf{U}} \\ \mathbf{\Omega}_{\mathbf{V}} \end{bmatrix} + \begin{bmatrix} \mathbf{K}_{\mathbf{U}}^{\mathrm{T}} & \mathbf{K}_{\mathbf{V}}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} & \tilde{\mathbf{A}} \\ \tilde{\mathbf{A}}^{\mathrm{T}} & \end{bmatrix} \begin{bmatrix} \mathbf{K}_{\mathbf{U}} \\ \mathbf{K}_{\mathbf{V}} \end{bmatrix}.$$

Let

$$\mathbf{m}_{\mathbf{Z}} \coloneqq \mathbf{vec}\left(\begin{bmatrix} \mathbf{\Omega}_{\mathbf{U}} \\ \mathbf{\Omega}_{\mathbf{V}} \end{bmatrix}\right), \quad \mathbf{k}_{\mathbf{Z}} \coloneqq \mathbf{vec}\left(\begin{bmatrix} \mathbf{K}_{\mathbf{U}} \\ \mathbf{K}_{\mathbf{V}} \end{bmatrix}\right)$$

Then,

$$\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}}\right) = \mathbf{m}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{m}_{\mathbf{Z}} + \mathbf{k}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{k}_{\mathbf{Z}},$$

$$\mathbf{Tr}\left(-\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{U}}^{\mathrm{T}} & \boldsymbol{\Omega}_{\mathbf{V}}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\ & \mathbf{A}\\ \mathbf{A} & \end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{U}}\\ \boldsymbol{\Omega}_{\mathbf{V}}\end{bmatrix}\mathbf{N}+\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{U}}^{\mathrm{T}} & \boldsymbol{\Omega}_{\mathbf{V}}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{U}}\\ \boldsymbol{\Omega}_{\mathbf{V}}\end{bmatrix}\mathbf{AN}\right)=$$
$$\mathbf{m}_{\mathbf{Z}}^{\mathrm{T}}\left(\mathbf{AN}\otimes\mathbf{I}_{2p}-\mathbf{N}\otimes\begin{bmatrix}\ & \mathbf{A}\\ \mathbf{A} & \end{bmatrix}\right)\mathbf{m}_{\mathbf{Z}}$$

$$\mathbf{Tr}\left(-\begin{bmatrix}\mathbf{K}_{\mathbf{U}}^{\mathrm{T}} & \mathbf{K}_{\mathbf{V}}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\ & \tilde{\mathbf{A}}\\ \tilde{\mathbf{A}}^{\mathrm{T}} & \end{bmatrix}\begin{bmatrix}\mathbf{K}_{\mathbf{U}}\\ \mathbf{K}_{\mathbf{V}}\end{bmatrix}\mathbf{N}+\begin{bmatrix}\mathbf{K}_{\mathbf{U}}^{\mathrm{T}} & \mathbf{K}_{\mathbf{V}}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{K}_{\mathbf{U}}\\ \mathbf{K}_{\mathbf{V}}\end{bmatrix}\mathbf{AN}\right)=$$
$$\mathbf{k}_{\mathbf{Z}}^{\mathrm{T}}\left(\mathbf{AN}\otimes\mathbf{I}_{2p}-\mathbf{N}\otimes\begin{bmatrix}\ & \tilde{\mathbf{A}}\\ \tilde{\mathbf{A}}^{\mathrm{T}} & \end{bmatrix}\right)\mathbf{k}_{\mathbf{Z}}$$

thus,

$$\frac{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\left(-\Sigma_{\nabla^2 f_{\mathbf{CCA}}}\xi_{\mathbf{Z}}+\Sigma\xi_{\mathbf{Z}}\mathbf{A}\right)\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}}\right)}$$
$$=\frac{\mathbf{m}_{\mathbf{Z}}^{\mathrm{T}}\left(\mathbf{AN}\otimes\mathbf{I}_{2p}-\mathbf{N}\otimes\begin{bmatrix}\ & \mathbf{A}\\ \mathbf{A} & \end{bmatrix}\right)\mathbf{m}_{\mathbf{Z}}+\mathbf{k}_{\mathbf{Z}}^{\mathrm{T}}\left(\mathbf{AN}\otimes\mathbf{I}_{2p}-\mathbf{N}\otimes\begin{bmatrix}\ & \tilde{\mathbf{A}}\\ \tilde{\mathbf{A}}^{\mathrm{T}} & \end{bmatrix}\right)\mathbf{k}_{\mathbf{Z}}}{\mathbf{m}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{m}_{\mathbf{Z}}+\mathbf{k}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{k}_{\mathbf{Z}}}.$$

Recalling that $\boldsymbol{\Omega}_{\mathbf{U}}=-\boldsymbol{\Omega}_{\mathbf{U}}^{\mathrm{T}}$ and $\boldsymbol{\Omega}_{\mathbf{V}}=-\boldsymbol{\Omega}_{\mathbf{V}}^{\mathrm{T}}$, and both are real matrices (so the elements of the main diagonals are 0), we have

$$\mathbf{m}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{m}_{\mathbf{Z}}=2\left(\sum_{1\leq j<i\leq p}\left(\boldsymbol{\Omega}_{\mathbf{U}}\right)_{ij}^{2}+\sum_{1\leq j<i\leq p}\left(\boldsymbol{\Omega}_{\mathbf{V}}\right)_{ij}^{2}\right),\tag{23}$$

and

$$\mathbf{m}_{\mathbf{Z}}^{\mathrm{T}}\left(\mathbf{AN}\otimes\mathbf{I}_{2p}-\mathbf{N}\otimes\begin{bmatrix}\ & \mathbf{A}\\ \mathbf{A} & \end{bmatrix}\right)\mathbf{m}_{\mathbf{Z}}$$
$$=\sum_{1\leq j<i\leq p}\left[\left(\mu_i\beta_i+\mu_j\beta_j\right)\left(\left(\boldsymbol{\Omega}_{\mathbf{U}}\right)_{ij}^{2}+\left(\boldsymbol{\Omega}_{\mathbf{V}}\right)_{ij}^{2}\right)-2\left(\beta_i\mu_j+\beta_j\mu_i\right)\left(\boldsymbol{\Omega}_{\mathbf{U}}\right)_{ij}\left(\boldsymbol{\Omega}_{\mathbf{V}}\right)_{ij}\right].\tag{24}$$

Thus, only the $p(p-1)/2$ entries below the diagonal of $\boldsymbol{\Omega}_{\mathbf{U}}$ and the $p(p-1)/2$ entries below the diagonal of $\boldsymbol{\Omega}_{\mathbf{V}}$ determine the values of Eq. (23) and Eq. (24). Let us now denote by $\tilde{\mathbf{m}}_{\mathbf{Z}}$ the column stack of $(\boldsymbol{\Omega}_{\mathbf{U}},\boldsymbol{\Omega}_{\mathbf{V}})$, but only with the the subdiagonal entries of $\boldsymbol{\Omega}_{\mathbf{U}}$ and of $\boldsymbol{\Omega}_{\mathbf{V}}$ (i.e., $\mathbf{m}_{\mathbf{Z}}$ "purged" of the superdiagonal elements). We then have,

$$\mathbf{m}_{\mathbf{Z}}^{\mathrm{T}}\mathbf{m}_{\mathbf{Z}}=2\tilde{\mathbf{m}}_{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{m}}_{\mathbf{Z}}$$

and

$$\mathbf{m}_{\mathbf{Z}}^{\mathrm{T}}\left(\mathbf{AN}\otimes\mathbf{I}_{2p}-\mathbf{N}\otimes\begin{bmatrix}\ & \mathbf{A}\\ \mathbf{A} & \end{bmatrix}\right)\mathbf{m}_{\mathbf{Z}}=\tilde{\mathbf{m}}_{\mathbf{Z}}^{\mathrm{T}}\Psi\tilde{\mathbf{m}}_{\mathbf{Z}}$$

where $\boldsymbol{\Psi}\in\mathbb{R}^{p(p-1)\times p(p-1)}$ is a matrix defined as follows: Ψ is a block diagonal matrix, where the blocks are of descending order: $2\left(p-1\right),2\left(p-2\right),...,2=2$. The jth $(1\leq j\leq p-1)$ block, denoted by $\boldsymbol{\Psi}_j$, is of the order $2\left(p-j\right)$ and has following form:

$$\boldsymbol{\Psi}_j\coloneqq\begin{bmatrix}\mathbf{D}_j & -\mathbf{T}_j\\ -\mathbf{T}_j & \mathbf{D}_j\end{bmatrix},$$

where

$$\mathbf{D}_j = \mathbf{diag}\left(\mu_{j+1}\beta_{j+1} + \mu_j\beta_j, \mu_{j+2}\beta_{j+2} + \mu_j\beta_j, ..., \mu_p\beta_p + \mu_j\beta_j\right),$$

and

$$\mathbf{T}_j = \mathbf{diag}\left(\beta_{j+1}\mu_j + \beta_j\mu_{j+1}, \beta_{j+2}\mu_j + \beta_j\mu_{j+2}, ..., \beta_p\mu_j + \beta_j\mu_p\right).$$

We make the following change of variables: $\tilde{\mathbf{d}}_{\mathbf{Z}} \coloneqq \sqrt{2}\tilde{\mathbf{m}}_{\mathbf{Z}}$. Finally, Eq. (21) is rewritten in the following way

$$\frac{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\left(-\Sigma_{\nabla^2 f_{\mathbf{CCA}}}\xi_{\mathbf{Z}} + \Sigma\xi_{\mathbf{Z}}\mathbf{A}\right)\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}}\right)} = \frac{\begin{bmatrix} \tilde{\mathbf{d}}_{\mathbf{Z}}^{\mathrm{T}} & \mathbf{k}_{\mathbf{Z}}^{\mathrm{T}} \end{bmatrix} \mathbf{blkdiag}\left(\frac{1}{2}\mathbf{\Psi}, \mathbf{AN}\otimes\mathbf{I}_{d-2p} - \mathbf{N}\otimes\begin{bmatrix} & \tilde{\mathbf{A}} \\ \tilde{\mathbf{A}}^{\mathrm{T}} & \end{bmatrix}\right)\begin{bmatrix} \tilde{\mathbf{d}}_{\mathbf{Z}} \\ \mathbf{k}_{\mathbf{Z}} \end{bmatrix}}{\begin{bmatrix} \tilde{\mathbf{d}}_{\mathbf{Z}}^{\mathrm{T}} & \mathbf{k}_{\mathbf{Z}}^{\mathrm{T}} \end{bmatrix}\begin{bmatrix} \tilde{\mathbf{d}}_{\mathbf{Z}} \\ \mathbf{k}_{\mathbf{Z}} \end{bmatrix}}. \quad (25)$$

Note that the mapping $\varphi(\cdot) : T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}} \to \mathbb{R}^{pd-p(p+1)}$ defined by

$$\varphi(\xi_{\mathbf{Z}}) \coloneqq \begin{bmatrix} \tilde{\mathbf{d}}_{\mathbf{Z}} \\ \mathbf{k}_{\mathbf{Z}} \end{bmatrix},$$

is a coordinate chart of the elements of $T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}}$, since $\varphi(\cdot)$ is a bijection (one-to-one correspondence) of the elements of $T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}}$ onto $\mathbb{R}^{pd-p(p+1)}$. Indeed, $\mathbf{k}_{\mathbf{Z}}$ is a column stack of $(\mathbf{K}_{\mathbf{U}}, \mathbf{K}_{\mathbf{V}})$, thus we can retract the matrices $\mathbf{K}_{\mathbf{U}}$ and $\mathbf{K}_{\mathbf{V}}$. Similarly $\tilde{\mathbf{d}}_{\mathbf{Z}}$ is proportional to $\tilde{\mathbf{m}}_{\mathbf{Z}}$ which is a column stack of $(\mathbf{\Omega}_{\mathbf{U}}, \mathbf{\Omega}_{\mathbf{V}})$, but only with the the subdiagonal entries of $\mathbf{\Omega}_{\mathbf{U}}$ and of $\mathbf{\Omega}_{\mathbf{V}}$. Since $\mathbf{\Omega}_{\mathbf{U}}$ and $\mathbf{\Omega}_{\mathbf{V}}$ are skew-symmetric matrices, we can retract $\mathbf{\Omega}_{\mathbf{U}}$ and $\mathbf{\Omega}_{\mathbf{V}}$. With the matrices $\mathbf{K}_{\mathbf{U}}, \mathbf{K}_{\mathbf{V}}, \mathbf{\Omega}_{\mathbf{U}}$ and $\mathbf{\Omega}_{\mathbf{V}}$ at hand, we can fully retract $\xi_{\mathbf{Z}} = (\xi_{\mathbf{U}}, \xi_{\mathbf{V}})$.

The eigenvalues and corresponding eigenvectors of any linear operator over a finite dimensional vector space do not depend on the choice of coordinate chart and basis, thus the eigenvalues and eigenvectors of $\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z})[\cdot] : T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}} \to T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}}$ which are computed using the Courant-Fischer Theorem for compact self-adjoint linear operators (Eq. (13) and Eq. (14)), can be also computed by the Courant-Fischer Theorem for symmetric matrices (Horn and Johnson, 2012, Theorem 4.2.6) after applying $\varphi(\cdot)$. In particular, Eq. (25) determines the signs of the eigenvalues of the Riemannian Hessian at any $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$ (in the special case $\mathbf{M}_{\mathbf{Z}} \coloneqq \Sigma$, the eigenvalues of Eq. (26) are the eigenvalues of the Riemannian Hessian at a critical point $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$), and the bounds of Eq. (25) together with the bounds of the term $\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}}\right)/\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\mathbf{M}_{\mathbf{Z}}\xi_{\mathbf{Z}}\right)$ bound the condition number of the Riemannian Hessian at $\mathbf{Z}^\star \in \mathbb{S}_{\mathbf{xy}}$.

To that end, we perform the following computation. The righthand side of Eq. (25) is a Rayleigh quotient, so according to the Courant-Fischer Theorem for symmetric matrices the eigenvalues of the $pd - p(p+1) \times pd - p(p+1)$ symmetric matrix

$$\mathbf{blkdiag}\left(\frac{1}{2}\mathbf{\Psi}, \mathbf{AN}\otimes\mathbf{I}_{d-2p} - \mathbf{N}\otimes\begin{bmatrix} & \tilde{\mathbf{A}} \\ \tilde{\mathbf{A}}^{\mathrm{T}} & \end{bmatrix}\right), \quad (26)$$

are determined by critical values of Eq. (25). The set of eigenvalues of the matrix in Eq. (26) is equal to the union of the set of eigenvalues of $\frac{1}{2}\mathbf{\Psi}$ and

$$\Phi \coloneqq \mathbf{AN} \otimes \mathbf{I}_{d-2p} - \mathbf{N} \otimes \begin{bmatrix} & \tilde{\mathbf{A}} \\ \tilde{\mathbf{A}}^{\mathrm{T}} & \end{bmatrix}.$$

The matrix Φ is a $p(d-2p) \times p(d-2p)$ block diagonal matrix, where all the blocks are $p \times p$, and the jth block is

$$\mu_j \beta_j \mathbf{I}_{d-2p} - \mu_j \begin{bmatrix} & \tilde{\mathbf{A}} \\ \tilde{\mathbf{A}}^{\mathrm{T}} & \end{bmatrix}.$$

Thus, the eigenvalues of Φ are $\mu_j(\beta_j \pm \beta_i)$ and $\mu_j\beta_j$ for $j = 1, ..., p$ and $i = p+1, ..., d_{\mathbf{y}}$. In summary, we have $p(d_{\mathbf{y}} - p)$ eigenvalues of the form $\mu_j(\beta_j + \beta_i)$, similarly $p(d_{\mathbf{y}} - p)$ eigenvalues of the form $\mu_j(\beta_j - \beta_i)$, and $p(d_{\mathbf{x}} - d_{\mathbf{y}})$ eigenvalues of the form $\mu_j\beta_j$. From the definition of Ψ, we see that the eigenvalues of $\frac{1}{2}\mathbf{\Psi}$ are: $\frac{1}{2}[(\mu_i\beta_i + \mu_j\beta_j) \pm (\beta_i\mu_j + \beta_j\mu_i)]$ for $1 \leq j < i \leq p$. These eigenvalues can also be rewritten as: $\frac{1}{2}(\mu_j + \mu_i)(\beta_j + \beta_i)$ and $\frac{1}{2}(\mu_j - \mu_i)(\beta_j - \beta_i)$.

Now, we have all the eigenvalues of the matrix in Eq. (26): $p(d_{\mathbf{y}} - p)$ eigenvalues of the form $\mu_j(\beta_j + \beta_i)$ where $j = 1, ..., p$ and $i = p+1, ..., d_{\mathbf{y}}$, $p(d_{\mathbf{y}} - p)$ eigenvalues of the form $\mu_j(\beta_j - \beta_i)$ where $j = 1, ..., p$ and $i = p+1, ..., d_{\mathbf{y}}$, $p(d_{\mathbf{x}} - d_{\mathbf{y}})$ eigenvalues of the form $\mu_j\beta_j$ where $j = 1, ..., p$, $p(p-1)/2$ eigenvalues of the form $\frac{1}{2}(\mu_j + \mu_i)(\beta_j + \beta_i)$ where $1 \leq j < i \leq p$, and $p(p-1)/2$ eigenvalues of the form $\frac{1}{2}(\mu_j - \mu_i)(\beta_j - \beta_i)$ where $1 \leq j < i \leq p$.

Finally, we bound the condition number of the Riemannian Hessian at $\mathbf{Z}^\star = (\mathbf{U}^\star, \mathbf{V}^\star) \in \mathbb{S}_{\mathbf{xy}}$. In such case, $\beta_1 = \sigma_1 > ... > \beta_p = \sigma_p$. Without loss of generality, we can always choose $\mathbf{U}_{\Sigma_{\mathbf{xx}}\perp}$ and $\mathbf{V}_{\Sigma_{\mathbf{yy}}\perp}$ such that $\beta_{p+1} = \sigma_{p+1} \geq ... \geq \beta_{d_{\mathbf{y}}} = \sigma_{d_{\mathbf{y}}}$. Then, we have that Eq. (21) is bounded by the minimal and maximal eigenvalues of Eq. (26). Thus,

$$0 < \max_{\mathbf{0} \neq \xi_{\mathbf{Z}^\star} \in T_{\mathbf{Z}^\star}\mathbb{S}_{\mathbf{xy}}} \frac{\mathbf{Tr}\left(\xi_{\mathbf{Z}^\star}^{\mathrm{T}}\left(-\Sigma_{\nabla^2 f_{\mathbf{CCA}}}\xi_{\mathbf{Z}^\star} + \Sigma\xi_{\mathbf{Z}^\star}\mathbf{A}\right)\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{Z}^\star}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}^\star}\right)} = \max\left\{\mu_1(\sigma_1 + \sigma_{p+1}), \frac{1}{2}(\mu_1 + \mu_2)(\sigma_1 + \sigma_2)\right\},$$

and

$$\min_{\mathbf{0} \neq \xi_{\mathbf{Z}^\star} \in T_{\mathbf{Z}^\star}\mathbb{S}_{\mathbf{xy}}} \frac{\mathbf{Tr}\left(\xi_{\mathbf{Z}^\star}^{\mathrm{T}}\left(-\Sigma_{\nabla^2 f_{\mathbf{CCA}}}\xi_{\mathbf{Z}^\star} + \Sigma\xi_{\mathbf{Z}^\star}\mathbf{A}\right)\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{Z}^\star}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}^\star}\right)} = \min\left\{\mu_p(\sigma_p - \sigma_{p+1}), \min_{1 \leq j < i \leq p} \frac{1}{2}(\mu_j - \mu_i)(\sigma_j - \sigma_i)\right\} > 0.$$

We use Eq. (15) and Eq. (16) to bound the condition number of the Riemannian Hessian at $\mathbf{Z}^\star = (\mathbf{U}^\star, \mathbf{V}^\star) \in \mathbb{S}_{\mathbf{xy}}$:

$$\begin{aligned} \lambda_{\max}(\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z}^\star)) &= \max_{\mathbf{0} \neq \xi_{\mathbf{Z}^\star} \in T_{\mathbf{Z}^\star}\mathbb{S}_{\mathbf{xy}}} \frac{g_{\mathbf{Z}^\star}(\xi_{\mathbf{Z}^\star}, \mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z}^\star)[\xi_{\mathbf{Z}^\star}])}{g_{\mathbf{Z}^\star}(\xi_{\mathbf{Z}^\star}, \xi_{\mathbf{Z}^\star})} \\ &\leq \max\left\{\mu_1(\sigma_1 + \sigma_{p+1}), \frac{1}{2}(\mu_1 + \mu_2)(\sigma_1 + \sigma_2)\right\} \cdot \tilde{\lambda}_{\max}\ , \end{aligned}$$

and

$$\begin{aligned}\lambda_{\min}(\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z}^\star)) &= \min_{\mathbf{0}\neq\xi_{\mathbf{Z}^\star}\in T_{\mathbf{Z}^\star}\mathbb{S}_{\mathbf{xy}}} \frac{g_{\mathbf{Z}^\star}(\xi_{\mathbf{Z}^\star}, \mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z}^\star)[\xi_{\mathbf{Z}^\star}])}{g_{\mathbf{Z}^\star}(\xi_{\mathbf{Z}^\star}, \xi_{\mathbf{Z}^\star})} \\ &\geq \min\left\{\mu_p(\sigma_p - \sigma_{p+1}), \min_{1\leq j<p} \frac{1}{2}\left(\mu_j - \mu_{j+1}\right)\left(\sigma_j - \sigma_{j+1}\right)\right\} \cdot \tilde{\lambda}_{\min}\ .\end{aligned}$$

Finally,

$$\kappa(\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z}^\star)) = \frac{\lambda_{\max}(\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z}^\star))}{\lambda_{\min}(\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z}^\star))} \leq \kappa^\star_{\mathbf{CCA}} \cdot \kappa\left(\Sigma, \mathbf{M}_{\mathbf{Z}}\right),$$

where

$$\kappa^\star_{\mathbf{CCA}} = \frac{\max\left\{\mu_1(\sigma_1 + \sigma_{p+1}), \frac{1}{2}(\mu_1 + \mu_2)(\sigma_1 + \sigma_2)\right\}}{\min\left\{\mu_p(\sigma_p - \sigma_{p+1}), \min_{1\leq j<p} \frac{1}{2}\left(\mu_j - \mu_{j+1}\right)\left(\sigma_j - \sigma_{j+1}\right)\right\}}.$$

In the special case $\mathbf{M}_{\mathbf{Z}^\star} = \Sigma$, the bound on the condition number of the Riemannian Hessian at $\mathbf{Z}^\star = (\mathbf{U}^\star, \mathbf{V}^\star) \in \mathbb{S}_{\mathbf{xy}}$ is reduced to an equality

$$\kappa(\mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z}^\star)) = \kappa^\star_{\mathbf{CCA}}.$$

■

B.4 Proof of Theorem 8

Proof We prove that the global minimizer of $f_{\mathbf{CCA}}(\mathbf{Z})$ subject to $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$, denoted by $\mathbf{Z}^\star$, is the only local minimum of $f_{\mathbf{CCA}}(\mathbf{Z})$ (and it is also strict) and all other critical points are either saddle points or strict local maximizers.

Eq. (20) helps to determine the signs of the eigenvalues of the Riemannian Hessian at any critical point $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$ and in particular at $\mathbf{Z}^\star$: the matrices Σ and $\mathbf{M}_{\mathbf{Z}}$ are both SPD matrices, therefore for $\xi_{\mathbf{Z}} \neq \mathbf{0}$ the term $\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\Sigma\xi_{\mathbf{Z}}\right)/\mathbf{Tr}\left(\xi_{\mathbf{Z}}^{\mathrm{T}}\mathbf{M}_{\mathbf{Z}}\xi_{\mathbf{Z}}\right) > 0$, thus only Eq. (21), where $\mathbf{0} \neq \xi_{\mathbf{Z}} \in T_{\mathbf{Z}}\mathbb{S}_{\mathbf{xy}}$ determines the signs. In addition, at a critical point $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$, Eq. (21) equals to the quotient $R(\xi_{\mathbf{Z}})$ for the choice $\mathbf{M}_{\mathbf{Z}} \coloneqq \Sigma$, since in (Absil et al., 2008, Proposition 5.5.6 and Eq. 5.35) it is shown that at a critical point the term $g_{\mathbf{Z}}(\xi_{\mathbf{Z}}, \mathbf{Hess} f_{\mathbf{CCA}}(\mathbf{Z})[\xi_{\mathbf{Z}}])$, which is the numerator of $R(\xi_{\mathbf{Z}})$, does not depend on the choice of Riemannian metric. Thus, the optimal values of Eq. (21) satisfying Eq. (13) or Eq. (14) are the eigenvalues of the Riemannian Hessian at $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$ with the choice $\mathbf{M}_{\mathbf{Z}} \coloneqq \Sigma$. Obviously, classification of the critical points does not depend on the Riemannian metric. Therefore, we can classify the critical points using the signs of the eigenvalues of the Riemannian Hessian at any critical point $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$ with the choice $\mathbf{M}_{\mathbf{Z}} \coloneqq \Sigma$.

Recall from the proof of Theorem 9 that in the special case $\mathbf{M}_{\mathbf{Z}} \coloneqq= \Sigma$, the eigenvalues of Eq. (26) are also the eigenvalues of the Riemannian Hessian at a critical point $\mathbf{Z} \in \mathbb{S}_{\mathbf{xy}}$. The eigenvalues are: $p(d_{\mathbf{y}} - p)$ eigenvalues of the form $\mu_j\left(\beta_j + \beta_i\right)$ where $j = 1, ..., p$ and $i = p+1, ..., d_{\mathbf{y}}$, $p(d_{\mathbf{y}} - p)$ eigenvalues of the form $\mu_j\left(\beta_j - \beta_i\right)$ where $j = 1, ..., p$ and $i = p+1, ..., d_{\mathbf{y}}$, $p(d_{\mathbf{x}} - d_{\mathbf{y}})$ eigenvalues of the form $\mu_j\beta_j$ where $j = 1, ..., p$, $p(p-1)/2$ eigenvalues of the form $\frac{1}{2}\left(\mu_j + \mu_i\right)\left(\beta_j + \beta_i\right)$ where $1 \leq j < i \leq p$, and $p(p-1)/2$ eigenvalues of the form $\frac{1}{2}\left(\mu_j - \mu_i\right)\left(\beta_j - \beta_i\right)$ where $1 \leq j < i \leq p$. $\mu_i > 0$ for $i = 1, ..., p$, and $\mu_j - \mu_i > 0$ for $j < i$. Also $|\beta_i| = \alpha_i$ for $i = 1, ..., d_{\mathbf{y}}$. Thus, the signs of the eigenvalues of the matrix in

Eq. (26) are only determined by β_j, $\beta_j + \beta_i$ and $\beta_j - \beta_i$, where β_j are up to their sign the p singular values of the matrix $\mathbf{R} \coloneqq \Sigma_{\mathbf{xx}}^{-1/2}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1/2}$ corresponding to left and right singular vectors which are the columns of $(\tilde{\mathbf{U}}, \tilde{\mathbf{V}}) = (\Sigma_{\mathbf{xx}}^{1/2}\mathbf{U}, \Sigma_{\mathbf{yy}}^{1/2}\mathbf{V})$. In particular, for an optimal $\mathbf{Z}^\star = (\mathbf{U}^\star, \mathbf{V}^\star) \in \mathbb{S}_{\mathbf{xy}}$ such that the columns of $(\tilde{\mathbf{U}}, \tilde{\mathbf{V}}) = (\Sigma_{\mathbf{xx}}^{1/2}\mathbf{U}^\star, \Sigma_{\mathbf{yy}}^{1/2}\mathbf{V}^\star)$ are ordered left and right p-dominant singular vectors of the matrix $\mathbf{R}$ on the same phase, then we have that $\beta_1 = \sigma_1 > ... > \beta_p = \sigma_p$. We conclude that β_i such that $i = p+1, ..., d_{\mathbf{y}}$ satisfies $|\beta_i| = \sigma_i$, which leads to $\beta_j > 0$, $\beta_j + \beta_i > 0$ and $\beta_j - \beta_i > 0$ where $j = 1, ..., p$ and $i = p+1, ..., d_{\mathbf{y}}$ or $1 \leq j < i \leq p$. Therefore, in this case all the eigenvalues of the matrix Eq. (26) are strictly positive, the matrix Eq. (26) is SPD, and consequently the eigenvalues of the Riemannian Hessian at $\mathbf{Z}^\star \in \mathbb{S}_{\mathbf{xy}}$ are all strictly positive. This proves that $\mathbf{Z}^\star$ is a strict local minimum of $f_{\mathbf{CCA}}(\mathbf{Z})$ on $\mathbb{S}_{\mathbf{xy}}$, see (Boumal, 2022, Proposition 6.5.).

If we prove that $\mathbf{Z}^\star$ is the only local minimum (up to the signs of the columns of $\mathbf{U}^\star$ and $\mathbf{V}^\star$), then $\mathbf{Z}^\star$ is the only asymptotically stable critical point following Theorem 6. In order to prove it we further assume that for all $i = 1, ..., q$ the values σ_i are distinct, then we can conclude the following. Suppose $\mathbf{Z} = (\mathbf{U}, \mathbf{V})$ is any other critical point differs from $\mathbf{Z}^\star$ at the optimal value, i.e., such that the columns of $(\tilde{\mathbf{U}}, \tilde{\mathbf{V}}) = (\Sigma_{\mathbf{xx}}^{1/2}\mathbf{U}, \Sigma_{\mathbf{yy}}^{1/2}\mathbf{V})$ are left and right singular vectors corresponding to some p singular values of the matrix $\mathbf{R}$ not necessarily on the same phase, so that Eq. (17) and Eq. (18) hold, and there exists at least one $1 \leq j \leq p$ for which $\beta_j \neq \sigma_j$. We consider the different cases:

1. Suppose $\beta_1, \ldots, \beta_p$ are not ordered in any particular order (possible for $p \geq 3$), then there exists j such that β_j is larger than some β_k and smaller than β_m where $j < k, m \leq p$, then $\beta_j - \beta_k > 0$ and $\beta_j - \beta_m < 0$. In this case there are both strictly positive and strictly negative eigenvalues of the Riemannian Hessian at $\mathbf{Z}$ for the choice $\mathbf{M}_{\mathbf{Z}} \coloneqq \Sigma$, therefore, $\mathbf{Z}$ is a saddle point.

2. Suppose $\beta_1, \ldots, \beta_p$ are ordered in a descending order. Since $\mathbf{Z}$ is not an optimal solution of Problem (3), then there exists at least one $1 \leq j \leq p$ for which $\beta_j \neq \sigma_j$. Thus, on the one hand $\beta_j - \beta_i > 0$ where $1 \leq j < i \leq p$, but on the other hand there exists at least one pair of indexes $j = 1, ..., p$ and $i = p+1, ..., d_{\mathbf{y}}$ such that $\beta_j + \beta_i < 0$ or $\beta_j - \beta_i < 0$, otherwise, it contradicts the assumption that there exists at least one $1 \leq j \leq p$ for which $\beta_j \neq \sigma_j$. In this case there are both strictly positive and strictly negative eigenvalues of the Riemannian Hessian at $\mathbf{Z}$ for the choice $\mathbf{M}_{\mathbf{Z}} \coloneqq \Sigma$, therefore, $\mathbf{Z}$ is a saddle point.

3. Suppose $\beta_1, \ldots, \beta_p$ are ordered in an ascending order. Then, $\beta_j - \beta_i < 0$ where $1 \leq j < i \leq p$. Now, we consider two sub-cases:

   (a) There exists at least one $1 \leq j \leq p$ for which $\beta_j \neq -\sigma_j$. Then, there exists at least one pair of indexes $j = 1, ..., p$ and $i = p+1, ..., d_{\mathbf{y}}$ such that $\beta_j + \beta_i > 0$ or $\beta_j - \beta_i > 0$, otherwise it contradicts the assumption that there exists at least one $1 \leq j \leq p$ for which $\beta_j \neq -\sigma_j$. In this case there are both strictly positive and strictly negative eigenvalues of the Riemannian Hessian at $\mathbf{Z}$ for the choice $\mathbf{M}_{\mathbf{Z}} :== \Sigma$, therefore, $\mathbf{Z}$ is a saddle point.

   (b) Consider the case $\beta_1 = -\sigma_1 < ... < \beta_p = -\sigma_p$. Then, $\beta_j < 0$, $\beta_j + \beta_i < 0$ and $\beta_j - \beta_i < 0$ where $j = 1, ..., p$ and $i = p+1, ..., d_{\mathbf{y}}$ or $1 \leq j < i \leq p$. Therefore, in

this case all the eigenvalues of the Riemannian Hessian at $\mathbf{Z}$ are strictly negative for the choice $\mathbf{M}_{\mathbf{Z}} := \Sigma$, therefore, $\mathbf{Z}$ is a strict local maximizer. Since it is the only local maximizer up to the signs of the columns of $\mathbf{U}$ and $\mathbf{V}$, and it is also a global maximizer, it is the unique maximizer.

In all cases, $\mathbf{Z}$ is not a local minimizer. Thus, $\mathbf{Z}^\star$ is the only local minimum (up to the signs of the columns of $\mathbf{U}^\star$ and $\mathbf{V}^\star$). According to Theorem 6 all the other critical points are unstable. ■

B.5 Proof of Corollary 11

Proof The condition number bound follows from Lemma 2 and Theorem 9 with one additional argument. From Lemma 2 we know that with probability of at least $1-\delta$ we have that all the generalized eigenvalues of the pencil $(\Sigma_{\mathbf{xx}}, \mathbf{M}^{(\mathbf{xx})})$ are contained in the interval $[1/2, 3/2]$, and the same is true for the pencil $(\Sigma_{\mathbf{yy}}, \mathbf{M}^{(\mathbf{yy})})$. Recall that all the generalized eigenvalues of the pencil $(\Sigma_{\mathbf{xx}}, \mathbf{M}^{(\mathbf{xx})})$ are also generalized eigenvalues of the pencil $(\Sigma, \mathbf{M})$, and the same is true for the generalized eigenvalues of $(\Sigma_{\mathbf{yy}}, \mathbf{M}^{(\mathbf{yy})})$. Indeed, after an appropriate padding with zeros each generalized eigenvector of $(\Sigma_{\mathbf{xx}}, \mathbf{M}^{(\mathbf{xx})})$ or $(\Sigma_{\mathbf{yy}}, \mathbf{M}^{(\mathbf{yy})})$ is also a generalized eigenvector of $(\Sigma, \mathbf{M})$. Thus, since $(\Sigma_{\mathbf{xx}}, \mathbf{M}^{(\mathbf{xx})})$ and $(\Sigma_{\mathbf{yy}}, \mathbf{M}^{(\mathbf{yy})})$ have $d_{\mathbf{x}}$ and $d_{\mathbf{y}}$ generalized eigenvalues and corresponding eigenvectors, they characterize all the generalized eigenvalues and corresponding eigenvectors of $(\Sigma, \mathbf{M})$. Thus, also the eigenvalues of the pencil $(\Sigma, \mathbf{M})$ are contained in the interval $[1/2, 3/2]$, and subsequently $\kappa(\Sigma, \mathbf{M}) \leq 3$. Using Theorem 9 we get the require bound for the condition number.

The costs are evident from Table 2, once we observe that none of the operations require forming $\Sigma_{\mathbf{xx}}, \Sigma_{\mathbf{yy}}$ or $\Sigma_{\mathbf{xy}}$, but instead require taking product of these matrices with vectors. These products can be computed in cost proportional to the number of non-zeros in $\mathbf{X}$ and/or $\mathbf{Y}$ by iterated products. In addition, we use the fact that $\mathbf{SX}$ and $\mathbf{SY}$ can be computed in $O(\mathbf{nnz}(\mathbf{X})) = O(nd_{\mathbf{x}})$ and $O(\mathbf{nnz}(\mathbf{Y})) = O(nd_{\mathbf{y}})$ operations. The required preprocessing is to factorize $\mathbf{M}^{(\mathbf{xx})}$ and $\mathbf{M}^{(\mathbf{yy})}$, so we can efficiently take products with $\left(\mathbf{M}^{(\mathbf{xx})}\right)^{-1}$ and $\left(\mathbf{M}^{(\mathbf{yy})}\right)^{-1}$, which can be done in $O(sd^2 + d^2)$ as explained in Section 3. Assuming a bounded number of line-search steps in each iteration of Riemannian CG, each iteration requires a bounded number of computations of each of the following: objective function evaluation costs $O\left(p\left(\mathbf{nnz}(\mathbf{X}) + \mathbf{nnz}(\mathbf{Y})\right)\right) = O(pnd)$, retraction costs $O\left(p\left(\mathbf{nnz}(\mathbf{X}) + \mathbf{nnz}(\mathbf{Y})\right) + dp^2\right)$, vector transport and Riemannian gradient computation take $O\left(p\left(\mathbf{nnz}(\mathbf{X}) + \mathbf{nnz}(\mathbf{Y})\right) + dp^2 + d^2p\right)$. ■

Appendix C. Omitted Proofs From Section 5

In this section we give omitted proofs from Section 5.

C.1 Proof of Theorem 14

Proof Recall that the critical points are defined to be the points where the Riemannian gradient is zero, and as such, whether a point is a critical point or not does not depend on the choice of Riemannian metric (see Absil et al., 2008, Eq. 3.31). Thus, for the sake of identifying the critical points, we can assume that $\mathbf{M_W} \coloneqq \mathbf{S_w} + \lambda \mathbf{I}_d$ and use a simplified form for the Riemannian gradient:

$$\begin{aligned}
\mathbf{grad} f_{\mathbf{FDA}}(\mathbf{W}) &= \mathbf{\Pi_W} \left((\mathbf{S_w} + \lambda \mathbf{I}_d)^{-1} \nabla \bar{f}_{\mathbf{FDA}}(\mathbf{W}) \right) \\
&= -\mathbf{\Pi_W} \left((\mathbf{S_w} + \lambda \mathbf{I}_d)^{-1} \mathbf{S_B W N} \right) \\
&= -\left[(\mathbf{S_w} + \lambda \mathbf{I}_d)^{-1} \mathbf{S_B W N} - \mathbf{Wsym} \left(\mathbf{W}^{\mathrm{T}} \mathbf{S_B W N} \right) \right] \\
&= -\left[\left(\mathbf{I}_d - \mathbf{W W}^{\mathrm{T}} (\mathbf{S_w} + \lambda \mathbf{I}_d) \right) (\mathbf{S_w} + \lambda \mathbf{I}_d)^{-1} \mathbf{S_B W N} \right. \\
&\quad \left. + \mathbf{Wskew} \left(\mathbf{W}^{\mathrm{T}} \mathbf{S_B W N} \right) \right] .
\end{aligned}$$

Let $\alpha_1, \ldots, \alpha_p$ be some generalized eigenvalues of the matrix pencil $(\mathbf{S_B}, \mathbf{S_w} + \lambda \mathbf{I}_d)$, and let $\mathbf{w}_1, ..., \mathbf{w}_p$ be the corresponding generalized eigenvectors (making them some p FDA weight vectors). Writing $\mathbf{w}_1, ..., \mathbf{w}_p$ as the columns of $\mathbf{W}$, the following equation holds:

$$\mathbf{S_B W} = (\mathbf{S_w} + \lambda \mathbf{I}_d) \mathbf{W A}$$

where $\mathbf{A} = \mathbf{diag} (\alpha_1, \ldots, \alpha_p)$. Letting $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w} + \lambda \mathbf{I}_d)}(p, d)$, we have

$$\begin{aligned}
\mathbf{grad} f_{\mathbf{FDA}}(\mathbf{W}) &= -\left[(\mathbf{S_w} + \lambda \mathbf{I}_d)^{-1} \mathbf{S_B W N} - \mathbf{Wsym} \left(\mathbf{W}^{\mathrm{T}} \mathbf{S_B W N} \right) \right] \\
&= -\left[\mathbf{WAN} - \mathbf{Wsym} (\mathbf{AN}) \right] = \mathbf{0}_{d \times p}.
\end{aligned}$$

To show the other side, note that if the Riemannian gradient nullifies, then

$$\left[\left(\mathbf{I}_d - \mathbf{W W}^{\mathrm{T}} (\mathbf{S_w} + \lambda \mathbf{I}_d) \right) (\mathbf{S_w} + \lambda \mathbf{I}_d)^{-1} \mathbf{S_B W N} + \mathbf{Wskew} \left(\mathbf{W}^{\mathrm{T}} \mathbf{S_B W N} \right) \right] = \mathbf{0}_{d \times p} .$$

By using similar reasoning as in (Absil et al., 2008, Subsection 4.8.2),

$$\left(\mathbf{I}_d - \mathbf{W W}^{\mathrm{T}} (\mathbf{S_w} + \lambda \mathbf{I}_d) \right) (\mathbf{S_w} + \lambda \mathbf{I}_d)^{-1} \mathbf{S_B W N},$$

belongs to the orthogonal compliment of the column space of $\mathbf{W}$ (with respect to the matrix $(\mathbf{S_w} + \lambda \mathbf{I}_d)$), and $\mathbf{Wskew} \left(\mathbf{W}^{\mathrm{T}} \mathbf{S_B W N} \right)$ belongs to the column space of $\mathbf{W}$. Thus, we get that the gradient vanishes if and only if the following two formulas hold:

$$\left(\mathbf{I}_d - \mathbf{W W}^{\mathrm{T}} (\mathbf{S_w} + \lambda \mathbf{I}_d) \right) (\mathbf{S_w} + \lambda \mathbf{I}_d)^{-1} \mathbf{S_B W N} = \mathbf{0}_{d \times p} , \tag{27}$$

and

$$\mathbf{Wskew} \left(\mathbf{W}^{\mathrm{T}} \mathbf{S_B W N} \right) = \mathbf{0}_{d \times p} . \tag{28}$$

From Eq. (27) we get

$$(\mathbf{S_w} + \lambda \mathbf{I}_d)^{-1} \mathbf{S_B W} = \mathbf{W} \left(\mathbf{W}^{\mathrm{T}} \mathbf{S_B W} \right),$$

since $\mathbf{N}$ is an invertible matrix. Also, since $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w} + \lambda \mathbf{I}_d)}(p, d)$ it is a full (column) rank matrix then Eq. (28) vanishes if and only if

$$\mathbf{skew} \left(\mathbf{W}^{\mathrm{T}} \mathbf{S_B W N} \right) = \mathbf{0}_p ,$$

which leads to $\left(\mathbf{W}^{\mathrm{T}}\mathbf{S_B}\mathbf{W}\right)\mathbf{N} = \mathbf{N}\left(\mathbf{W}^{\mathrm{T}}\mathbf{S_B}\mathbf{W}\right)$. This implies that $\left(\mathbf{W}^{\mathrm{T}}\mathbf{S_B}\mathbf{W}\right)$ is diagonal because any rectangle matrix that commutes with a diagonal matrix with distinct entries is diagonal. Finally we get

$$(\mathbf{S_w} + \lambda\mathbf{I}_d)^{-1}\mathbf{S_B}\mathbf{W} = \mathbf{W}\mathbf{D} \ ,$$

where $\mathbf{D} = \mathbf{W}^{\mathrm{T}}\mathbf{S_B}\mathbf{W}$ is a diagonal matrix. This implies that the columns of $\mathbf{W}$ correspond to some p generalized eigenvectors the matrix pencil $(\mathbf{S_B}, \mathbf{S_w} + \lambda\mathbf{I}_d)$, thus, making the columns some p FDA weight vectors.

Finally, to identify the optimal solutions, note that at the critical points the objective function is a sum of the generalized eigenvalues corresponding to the columns of $\mathbf{W}$ multiplied by a diagonal element of $\mathbf{N}$. Thus, the optimal solutions that minimize $f_{\mathbf{FDA}}(\mathbf{W})$ on $\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ are $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ such that the columns correspond to the p leading FDA weight vectors. Otherwise, we can increase the value of the objective function by replacing a weight vector with another that corresponds to a smaller generalized eigenvalue. Moreover, if we assume that $\rho_1 > \rho_2 > ... > \rho_{p+1} \geq 0$, then for the aforementioned $\mathbf{W}$, the columns belong each to a one dimensional generalized eigenspace, i.e., unique solution up the the signs of the columns. In the case where some $\rho_i = \rho_j$ for $1 \leq i, j \leq p$, permutations of the columns of $\mathbf{W}$ associated with ρ_i keep the solution optimal making it non-unique. ■

C.2 Proof of Theorem 15

Proof To prove the asymptotic stability of a $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ such that the columns are the p leading FDA weight vectors we use (Absil et al., 2008, Proposition 4.4.2). Recall from Theorem 14 that $\mathbf{W}$ that solve Problem (8) are unique up the the signs of the columns of $\mathbf{W}$, making these points isolated global (and consequently local) minimizers of $f_{\mathbf{FDA}}$ on $\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$. According to (Absil et al., 2008, Proposition 4.4.2), such points $\mathbf{W}$ are asymptotically stable.

Suppose $\mathbf{W}$ is a critical point of $f_{\mathbf{FDA}}(\mathbf{W})$ on $\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ which is not a local minimum. Then, there exists compact neighborhoods with either no other critical points, if there are no multiplicities of the generalized eigenspaces, or where all other critical point achieve the same value for the cost function, if there are multiplicities of the generalized eigenspaces. Thus, according to (Absil et al., 2008, Proposition 4.4.1), such $\mathbf{W}$ are unstable. ■

C.3 Proof of Theorem 18

Proof The proof is similar to the proof of Theorem 9. In order to bound the condition number of the Riemannian Hessian at $\mathbf{W}^\star$, we need to bound its maximal and minimal eigenvalues. Thus, to prove the theorem we analyze the eigenvalues of the Riemannian Hessian at some critical point $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ (in particular at $\mathbf{W}^\star$) using the Courant-Fischer Theorem for the compact self-adjoint linear operator

$$\mathbf{Hess} f_{\mathbf{FDA}}(\mathbf{W})[\cdot] : T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d) \to T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d),$$

over the finite-dimensional vector space $T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$, see (Kato, 2013, Chapter 1, Section 6.10):

$$\lambda_k(\mathbf{Hess} f_{\mathbf{FDA}}(\mathbf{W})) = \min_{U,\dim(U)=k-1} \max_{\mathbf{0}\neq\xi_{\mathbf{W}}\in U^{\perp}} R(\xi_{\mathbf{W}}), \tag{29}$$

$$\lambda_k(\mathbf{Hess} f_{\mathbf{FDA}}(\mathbf{W})) = \max_{U,\dim(U)=k} \min_{\mathbf{0}\neq\xi_{\mathbf{W}}\in U} R(\xi_{\mathbf{W}}), \tag{30}$$

where

$$R(\xi_{\mathbf{W}}) \coloneqq \frac{g_{\mathbf{W}}(\xi_{\mathbf{W}}, \mathbf{Hess} f_{\mathbf{FDA}}(\mathbf{W})[\xi_{\mathbf{W}}])}{g_{\mathbf{W}}(\xi_{\mathbf{W}}, \xi_{\mathbf{W}})},$$

$\lambda_k(\mathbf{Hess} f_{\mathbf{FDA}}(\mathbf{W})) = \rho_k$ is the kth largest eigenvalue (descending order) of $\mathbf{Hess} f_{\mathbf{FDA}}(\mathbf{W})$, and U is a linear subspace of $T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$. In particular, the maximal and minimal eigenvalues are given by the formulas:

$$\lambda_{\max}(\mathbf{Hess} f_{\mathbf{FDA}}(\mathbf{W})) = \max_{\mathbf{0}\neq\xi_{\mathbf{W}}\in T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)} R(\xi_{\mathbf{W}}), \tag{31}$$

$$\lambda_{\min}(\mathbf{Hess} f_{\mathbf{FDA}}(\mathbf{W})) = \min_{\mathbf{0}\neq\xi_{\mathbf{W}}\in T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)} R(\xi_{\mathbf{W}}). \tag{32}$$

We begin by simplifying the quotient $R(\xi_{\mathbf{W}})$. Recall that any critical point of $f_{\mathbf{FDA}}(\cdot)$ is a matrix $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ such that the columns are some p generalized eigenvectors of the matrix pencil $(\mathbf{S_B}, \mathbf{S_w} + \lambda\mathbf{I}_d)$ (see Theorem 14). Let $\alpha_1, \ldots, \alpha_p$ be some generalized eigenvalues of the matrix pencil $(\mathbf{S_B}, \mathbf{S_w} + \lambda\mathbf{I}_d)$, and let $\mathbf{w}_1, ..., \mathbf{w}_p$ be the corresponding generalized eigenvectors. Writing $\mathbf{w}_1, ..., \mathbf{w}_p$ as the columns of $\mathbf{W}$, the following equation holds:

$$\mathbf{S_B W} = (\mathbf{S_w} + \lambda\mathbf{I}_d)\mathbf{W A}$$

where $\mathbf{A} = \mathbf{diag}(\alpha_1, \ldots, \alpha_p)$. Letting $\mathbf{W} \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$, plugging in the ambient coordinates formula for the Riemannian Hessian (Eq. (10)), the Riemannian gradient nullifies (see Theorem 14) and we have

$$\begin{aligned}
\mathrm{Hess} f_{\mathbf{FDA}}(\mathbf{W})[\xi_{\mathbf{W}}] &= \mathbf{\Pi_W}\left(\mathbf{M_W^{-1}}\left[-\mathbf{S_B}\xi_{\mathbf{W}}\mathbf{N} + (\mathbf{S_w} + \lambda\mathbf{I}_d)\xi_{\mathbf{W}}\left(\mathbf{W}^{\mathrm{T}}\mathbf{S_B W N} + \mathbf{grad} f(\mathbf{W})\right)\right]\right) \\
&= \mathbf{\Pi_W}\left(\mathbf{M_W^{-1}}\left[-\mathbf{S_B}\xi_{\mathbf{W}}\mathbf{N} + (\mathbf{S_w} + \lambda\mathbf{I}_d)\xi_{\mathbf{W}}\left(\mathbf{W}^{\mathrm{T}}(\mathbf{S_w} + \lambda\mathbf{I}_d)\mathbf{W A N}\right)\right]\right) \\
&= \mathbf{\Pi_W}\left(\mathbf{M_W^{-1}}\left[-\mathbf{S_B}\xi_{\mathbf{W}} + (\mathbf{S_w} + \lambda\mathbf{I}_d)\xi_{\mathbf{W}}\mathbf{A}\right]\mathbf{N}\right).
\end{aligned} \tag{33}$$

Plugging in the formula for the Riemannian Hessian at a critical point (Eq. (33)), the quotient $R(\xi_{\mathbf{W}})$ reduces to

$$R(\xi_{\mathbf{W}}) = \frac{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\mathbf{M_W}\mathbf{\Pi_W}\left(\mathbf{M_W^{-1}}\left[-\mathbf{S_B}\xi_{\mathbf{W}} + (\mathbf{S_w} + \lambda\mathbf{I}_d)\xi_{\mathbf{W}}\mathbf{A}\right]\mathbf{N}\right)\right)}{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\mathbf{M_W}\xi_{\mathbf{W}}\right)}.$$

Now using the fact that the projection to the tangent space is self-adjoint with respect to the Riemannian metric, and that for any $\xi_{\mathbf{W}} \in T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ we have $\mathbf{\Pi_W}(\xi_{\mathbf{W}}) = \xi_{\mathbf{W}}$, we further see that

$$\frac{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\mathbf{M_W}\mathbf{\Pi_W}\left(\mathbf{M_W^{-1}}\left[-\mathbf{S_B}\xi_{\mathbf{W}} + (\mathbf{S_w} + \lambda\mathbf{I}_d)\xi_{\mathbf{W}}\mathbf{A}\right]\mathbf{N}\right)\right)}{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\mathbf{M_W}\xi_{\mathbf{W}}\right)} =$$

$$\frac{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\left[-\mathbf{S_B}\xi_{\mathbf{W}} + (\mathbf{S_w} + \lambda\mathbf{I}_d)\xi_{\mathbf{W}}\mathbf{A}\right]\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\mathbf{M_W}\xi_{\mathbf{W}}\right)}$$

Obviously, we can also write

$$\frac{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\left[-\mathbf{S_B}\xi_{\mathbf{W}}+(\mathbf{S_w}+\lambda\mathbf{I}_d)\xi_{\mathbf{W}}\mathbf{A}\right]\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\mathbf{M_W}\xi_{\mathbf{W}}\right)} = \frac{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\left[-\mathbf{S_B}\xi_{\mathbf{W}}+(\mathbf{S_w}+\lambda\mathbf{I}_d)\xi_{\mathbf{W}}\mathbf{A}\right]\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}(\mathbf{S_w}+\lambda\mathbf{I}_d)\xi_{\mathbf{W}}\right)}\cdot\frac{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}(\mathbf{S_w}+\lambda\mathbf{I}_d)\xi_{\mathbf{W}}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\mathbf{M_W}\xi_{\mathbf{W}}\right)}. \quad (34)$$

Using Eq. (34), a simplified form of the quotient $R(\xi_{\mathbf{W}})$, we can estimate upper and lower bounds on $R(\xi_{\mathbf{W}})$ where $\mathbf{0}\neq\xi_{\mathbf{W}}\in T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ in order to bound the condition number of the Riemannian Hessian at $\mathbf{W}^\star$. Since for $\xi_{\mathbf{W}}\neq\mathbf{0}$ the term

$$\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}(\mathbf{S_w}+\lambda\mathbf{I}_d)\xi_{\mathbf{W}}\right)/\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\mathbf{M_W}\xi_{\mathbf{W}}\right)>0,$$

the upper and lower bounds on

$$\frac{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\left[-\mathbf{S_B}\xi_{\mathbf{W}}+(\mathbf{S_w}+\lambda\mathbf{I}_d)\xi_{\mathbf{W}}\mathbf{A}\right]\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}(\mathbf{S_w}+\lambda\mathbf{I}_d)\xi_{\mathbf{W}}\right)}, \quad (35)$$

together with the upper and lower bounds of $\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}(\mathbf{S_w}+\lambda\mathbf{I}_d)\xi_{\mathbf{W}}\right)/\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\mathbf{M_W}\xi_{\mathbf{W}}\right)$ bound the condition number of the Riemannian Hessian at $\mathbf{W}^\star$.

We begin by estimating the term $\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}(\mathbf{S_w}+\lambda\mathbf{I}_d)\xi_{\mathbf{W}}\right)/\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\mathbf{M_W}\xi_{\mathbf{W}}\right)$. We use the vectorization operator and the Kronecker Product to rewrite it in the following form

$$\frac{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}(\mathbf{S_w}+\lambda\mathbf{I}_d)\xi_{\mathbf{W}}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\mathbf{M_W}\xi_{\mathbf{W}}\right)}=\frac{\mathbf{vec}\left(\xi_{\mathbf{W}}\right)^{\mathrm{T}}\left(\mathbf{I}_p\otimes(\mathbf{S_w}+\lambda\mathbf{I}_d)\right)\mathbf{vec}\left(\xi_{\mathbf{W}}\right)}{\mathbf{vec}\left(\xi_{\mathbf{W}}\right)^{\mathrm{T}}\left(\mathbf{I}_p\otimes\mathbf{M_W}\right)\mathbf{vec}\left(\xi_{\mathbf{W}}\right)}. \quad (36)$$

The righthand side of Eq. (36) is the generalized Rayleigh quotient for the matrix pencil

$$\left(\mathbf{I}_p\otimes(\mathbf{S_w}+\lambda\mathbf{I}_d),\mathbf{I}_p\otimes\mathbf{M_W}\right).$$

Note that $\mathbf{I}_p\otimes(\mathbf{S_w}+\lambda\mathbf{I}_d)$ and $\mathbf{I}_p\otimes\mathbf{M_W}$ are both SPD matrices, thus the eigenvalues of the matrix pencil

$$\left(\mathbf{I}_p\otimes(\mathbf{S_w}+\lambda\mathbf{I}_d),\mathbf{I}_p\otimes\mathbf{M_W}\right)$$

are equivalent to the eigenvalues of the matrix $\left(\mathbf{I}_p\otimes\mathbf{M_W}\right)^{-1}\left(\mathbf{I}_p\otimes(\mathbf{S_w}+\lambda\mathbf{I}_d)\right)=\mathbf{I}_p\otimes\mathbf{M_W}^{-1}(\mathbf{S_w}+\lambda\mathbf{I}_d)$ and all positive. According to (Minka, 2000, Section 2) the eigenvalues $\mathbf{I}_p\otimes\mathbf{M_W}^{-1}(\mathbf{S_w}+\lambda\mathbf{I}_d)$ are p copies of each of the eigenvalues of $\mathbf{M_W}^{-1}(\mathbf{S_w}+\lambda\mathbf{I}_d)$. Thus, the maximal and minimal eigenvalues of the matrix pencil $(\mathbf{I}_p\otimes(\mathbf{S_w}+\lambda\mathbf{I}_d),\mathbf{I}_p\otimes\mathbf{M_W})$ denoted by $\tilde{\lambda}_{\max}$ and $\tilde{\lambda}_{\min}$ are equivalent to the maximal and minimal eigenvalues of the matrix pencil $(\mathbf{S_w}+\lambda\mathbf{I}_d,\mathbf{M_W})$, and so is the corresponding condition number

$$\kappa\left(\mathbf{I}_p\otimes(\mathbf{S_w}+\lambda\mathbf{I}_d),\mathbf{I}_p\otimes\mathbf{M_W}\right)=\frac{\tilde{\lambda}_{\max}}{\tilde{\lambda}_{\min}}=\kappa\left((\mathbf{S_w}+\lambda\mathbf{I}_d),\mathbf{M_W}\right).$$

Therefore, using the Courant-Fischer Theorem

$$\begin{aligned}\tilde{\lambda}_{\max} &=: \lambda_{\max}(\mathbf{I}_p\otimes(\mathbf{S_w}+\lambda\mathbf{I}_d),\mathbf{I}_p\otimes\mathbf{M_W})\\ &= \max_{\mathbf{0}\neq\xi_{\mathbf{W}}\in\mathbb{R}^{d\times p}}\frac{\mathbf{vec}\left(\xi_{\mathbf{W}}\right)^{\mathrm{T}}\left(\mathbf{I}_p\otimes(\mathbf{S_w}+\lambda\mathbf{I}_d)\right)\mathbf{vec}\left(\xi_{\mathbf{W}}\right)}{\mathbf{vec}\left(\xi_{\mathbf{W}}\right)^{\mathrm{T}}\left(\mathbf{I}_p\otimes\mathbf{M_W}\right)\mathbf{vec}\left(\xi_{\mathbf{W}}\right)}\\ &\geq \max_{\mathbf{0}\neq\xi_{\mathbf{W}}\in T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)}\frac{\mathbf{vec}\left(\xi_{\mathbf{W}}\right)^{\mathrm{T}}\left(\mathbf{I}_p\otimes(\mathbf{S_w}+\lambda\mathbf{I}_d)\right)\mathbf{vec}\left(\xi_{\mathbf{W}}\right)}{\mathbf{vec}\left(\xi_{\mathbf{W}}\right)^{\mathrm{T}}\left(\mathbf{I}_p\otimes\mathbf{M_W}\right)\mathbf{vec}\left(\xi_{\mathbf{W}}\right)}\end{aligned}$$

and

$$\begin{aligned}\tilde{\lambda}_{\min} &=: \lambda_{\min}(\mathbf{I}_p \otimes (\mathbf{S_w} + \lambda \mathbf{I}_d), \mathbf{I}_p \otimes \mathbf{M_W}) \\ &= \min_{\mathbf{0} \neq \xi_\mathbf{W} \in \mathbb{R}^{d \times p}} \frac{\mathbf{vec}\,(\xi_\mathbf{W})^{\mathrm{T}} (\mathbf{I}_p \otimes (\mathbf{S_w} + \lambda \mathbf{I}_d))\,\mathbf{vec}\,(\xi_\mathbf{W})}{\mathbf{vec}\,(\xi_\mathbf{W})^{\mathrm{T}} (\mathbf{I}_p \otimes \mathbf{M_W})\,\mathbf{vec}\,(\xi_\mathbf{W})} \\ &\leq \min_{\mathbf{0} \neq \xi_\mathbf{W} \in T_\mathbf{W}\mathbf{St}_{(\mathbf{S_w}+\lambda \mathbf{I}_d)}(p,d)} \frac{\mathbf{vec}\,(\xi_\mathbf{W})^{\mathrm{T}} (\mathbf{I}_p \otimes (\mathbf{S_w} + \lambda \mathbf{I}_d))\,\mathbf{vec}\,(\xi_\mathbf{W})}{\mathbf{vec}\,(\xi_\mathbf{W})^{\mathrm{T}} (\mathbf{I}_p \otimes \mathbf{M_W})\,\mathbf{vec}\,(\xi_\mathbf{W})}\end{aligned}$$

Next, we consider Eq. (35). Recall that we can rewrite any tangent vector, $\xi_\mathbf{W} \in T_\mathbf{W}\mathbf{St}_{(\mathbf{S_w}+\lambda \mathbf{I}_d)}(p,d)$, in the following form:

$$\xi_\mathbf{W} = \mathbf{W}\mathbf{\Omega_W} + \mathbf{W}_{(\mathbf{S_w}+\lambda \mathbf{I}_d)\perp}\mathbf{K_W},$$

where $\mathbf{W}_{(\mathbf{S_w}+\lambda \mathbf{I}_d)\perp}$ is $(\mathbf{S_w} + \lambda \mathbf{I}_d)$-orthogonal to $\mathbf{W}$ so that the union of the columns of $\mathbf{W}$ and $\mathbf{W}_{(\mathbf{S_w}+\lambda \mathbf{I}_d)\perp}$ is a basis to $\mathbb{R}^d$, and $\mathbf{\Omega_W} = -\mathbf{\Omega}_\mathbf{W}^{\mathrm{T}} \in \mathbb{R}^{p \times p}$. Note that we can always make the choice of the columns of $\mathbf{W}_{(\mathbf{S_w}+\lambda \mathbf{I}_d)\perp}$ to be some $d - p$ generalized eigenvalues of the matrix pencil $(\mathbf{S_B}, \mathbf{S_w} + \lambda \mathbf{I}_d)$. With this choice we have

$$\mathbf{S_B}\mathbf{W}_{(\mathbf{S_w}+\lambda \mathbf{I}_d)\perp} = (\mathbf{S_w} + \lambda \mathbf{I}_d)\mathbf{W}_{(\mathbf{S_w}+\lambda \mathbf{I}_d)\perp}\tilde{\mathbf{A}},\ \mathbf{W}_{(\mathbf{S_w}+\lambda \mathbf{I}_d)\perp}^{\mathrm{T}}\mathbf{S_B}\mathbf{W}_{(\mathbf{S_w}+\lambda \mathbf{I}_d)\perp} = \tilde{\mathbf{A}},$$

where $\tilde{\mathbf{A}} \in \mathbb{R}^{(d-p)\times(d-p)}$ is a diagonal matrix with the corresponding generalized eigenvalues on the diagonal $\alpha_{p+1}, ..., \alpha_d$.

Now, we have

$$\xi_\mathbf{W}^{\mathrm{T}}(\mathbf{S_w} + \lambda \mathbf{I}_d)\xi_\mathbf{W} = \mathbf{\Omega}_\mathbf{W}^{\mathrm{T}}\mathbf{\Omega_W} + \mathbf{K}_\mathbf{W}^{\mathrm{T}}\mathbf{K_W}$$

and

$$\xi_\mathbf{W}^{\mathrm{T}}\mathbf{S_B}\xi_\mathbf{W} = \mathbf{\Omega}_\mathbf{W}^{\mathrm{T}}\mathbf{A}\mathbf{\Omega_W} + \mathbf{K}_\mathbf{W}^{\mathrm{T}}\tilde{\mathbf{A}}\mathbf{K_W}\ .$$

Substitute these computations into Eq. (35)

$$\begin{aligned}\frac{\mathbf{Tr}\left(\xi_\mathbf{W}^{\mathrm{T}}\left[-\mathbf{S_B}\xi_\mathbf{W} + (\mathbf{S_w} + \lambda \mathbf{I}_d)\xi_\mathbf{W}\mathbf{A}\right]\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_\mathbf{W}^{\mathrm{T}}(\mathbf{S_w} + \lambda \mathbf{I}_d)\xi_\mathbf{W}\right)} &= \frac{\mathbf{Tr}\left(-\left(\mathbf{\Omega}_\mathbf{W}^{\mathrm{T}}\mathbf{A}\mathbf{\Omega_W} + \mathbf{K}_\mathbf{W}^{\mathrm{T}}\tilde{\mathbf{A}}\mathbf{K_W}\right)\mathbf{N} + \left(\mathbf{\Omega}_\mathbf{W}^{\mathrm{T}}\mathbf{\Omega_W} + \mathbf{K}_\mathbf{W}^{\mathrm{T}}\mathbf{K_W}\right)\mathbf{AN}\right)}{\mathbf{Tr}\left(\mathbf{\Omega}_\mathbf{W}^{\mathrm{T}}\mathbf{\Omega_W} + \mathbf{K}_\mathbf{W}^{\mathrm{T}}\mathbf{K_W}\right)} \\ &= \frac{\mathbf{vec}\,(\mathbf{\Omega_W})^{\mathrm{T}}\,\mathbf{D}_1\mathbf{vec}\,(\mathbf{\Omega_W}) + \mathbf{vec}\,(\mathbf{K_W})^{\mathrm{T}}\,\mathbf{D}_2\mathbf{vec}\,(\mathbf{K_W})}{\mathbf{vec}\,(\mathbf{\Omega_W})^{\mathrm{T}}\,\mathbf{vec}\,(\mathbf{\Omega_W}) + \mathbf{vec}\,(\mathbf{K_W})^{\mathrm{T}}\,\mathbf{vec}\,(\mathbf{K_W})},\end{aligned}$$

where

$$\mathbf{D}_1 \coloneqq \mathbf{AN} \otimes \mathbf{I}_p - \mathbf{N} \otimes \mathbf{A},\ \mathbf{D}_2 \coloneqq \mathbf{AN} \otimes \mathbf{I}_{d-p} - \mathbf{N} \otimes \tilde{\mathbf{A}}.$$

RRecall that $\mathbf{\Omega_W} = -\mathbf{\Omega}_\mathbf{W}^{\mathrm{T}} \in \mathbb{R}^{p \times p}$, and both are real matrices (so the elements of the main diagonals are 0), we have

$$\mathbf{vec}\,(\mathbf{\Omega_W})^{\mathrm{T}}\,\mathbf{vec}\,(\mathbf{\Omega_W}) = 2\sum_{p \geq i > j} (\mathbf{\Omega_W})_{ij}^2 \tag{37}$$

and

$$\mathbf{vec}\,(\mathbf{\Omega_W})^{\mathrm{T}}\,(\mathbf{AN} \otimes \mathbf{I}_p - \mathbf{N} \otimes \mathbf{A})\,\mathbf{vec}\,(\mathbf{\Omega_W}) = \sum_{1 \leq j < i \leq p} (\mathbf{\Omega_W})_{ij}^2\,(\mu_j - \mu_i)\,(\alpha_j - \alpha_i)\,. \tag{38}$$

Thus, only the $p(p-1)/2$ entries below the diagonal of $\mathbf{\Omega_W}$ determine the values of Eq. (37) and Eq. (38). Let us now denote by $\mathbf{m_W}$ the column stack of $\mathbf{\Omega_W}$, but only with the the subdiagonal entries of $\mathbf{\Omega_W}$ (i.e., $\mathbf{vec}\left(\mathbf{\Omega_W}\right)$ purged of the superdiagonal elements). We then have

$$\mathbf{vec}\left(\mathbf{\Omega_W}\right)^{\mathrm{T}}\mathbf{vec}\left(\mathbf{\Omega_W}\right)=2\mathbf{m_W^{\mathrm{T}}m_W}$$

and

$$\mathbf{vec}\left(\mathbf{\Omega_W}\right)^{\mathrm{T}}\left(\mathbf{AN}\otimes\mathbf{I}_p-\mathbf{N}\otimes\mathbf{A}\right)\mathbf{vec}\left(\mathbf{\Omega_W}\right)=\mathbf{m_W^{\mathrm{T}}\Psi m_W}$$

where $\mathbf{\Psi}\in\mathbb{R}^{p(p-1)/2\times p(p-1)/2}$ is a block diagonal matrix, where the blocks are of descending order from $p-1,p-2,...,1$, and each block is a diagonal matrix as well. The jth ($1\leq j\leq p-1$) block, denoted by $\mathbf{\Psi}_j$, is of the order $p-j$ and has the following form

$$\mathbf{\Psi}_j=\mathbf{diag}\left(\left(\mu_j-\mu_{j+1}\right)\left(\alpha_j-\alpha_{j+1}\right),\left(\mu_j-\mu_{j+2}\right)\left(\alpha_j-\alpha_{j+2}\right),...,\left(\mu_j-\mu_p\right)\left(\alpha_j-\alpha_p\right)\right)$$

Now, we make the following change of variables: $\mathbf{d_W}\coloneqq\sqrt{2}\mathbf{m_W}$, $\mathbf{k_W}\coloneqq\mathbf{vec}\left(\mathbf{K_W}\right)$. Finally, Eq. (35) is rewritten in the following way

$$\frac{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\left[-\mathbf{S_B}\xi_{\mathbf{W}}+\left(\mathbf{S_w}+\lambda\mathbf{I}_d\right)\xi_{\mathbf{W}}\mathbf{A}\right]\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\left(\mathbf{S_w}+\lambda\mathbf{I}_d\right)\xi_{\mathbf{W}}\right)}$$

$$=\frac{\begin{bmatrix}\mathbf{d_W^{\mathrm{T}}} & \mathbf{k_W^{\mathrm{T}}}\end{bmatrix}\mathbf{blkdiag}\left(\frac{1}{2}\mathbf{\Psi},\mathbf{AN}\otimes\mathbf{I}_{d-p}-\mathbf{N}\otimes\tilde{\mathbf{A}}\right)\begin{bmatrix}\mathbf{d_W}\\ \mathbf{k_W}\end{bmatrix}}{\begin{bmatrix}\mathbf{d_W^{\mathrm{T}}} & \mathbf{k_W^{\mathrm{T}}}\end{bmatrix}\begin{bmatrix}\mathbf{d_W}\\ \mathbf{k_W}\end{bmatrix}}. \quad (39)$$

Note that the mapping $\varphi(\cdot):T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)\to\mathbb{R}^{pd-p(p+1)/2}$ defined by

$$\varphi(\xi_{\mathbf{Z}})\coloneqq\begin{bmatrix}\mathbf{d_W}\\ \mathbf{k_W}\end{bmatrix},$$

is a coordinate chart of the elements of $T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$, since $\varphi(\cdot)$ is a bijection (one-to-one correspondence) of the elements of $T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ onto $\mathbb{R}^{pd-p(p+1)/2}$. Indeed, $\mathbf{k_W}$ is a column stack of $\mathbf{K_W}$, thus we can retract the matrices $\mathbf{K_W}$. Similarly $\mathbf{d_W}$ is proportional to $\mathbf{m_W}$ which is a column stack of $\mathbf{\Omega_W}$, but only with the the subdiagonal entries of $\mathbf{\Omega_W}$. Since $\mathbf{\Omega_W}$ is a skew-symmetric matrix, we can retract $\mathbf{\Omega_W}$. With the matrices $\mathbf{K_W}$ and $\mathbf{\Omega_W}$ at hand, we can fully retract $\xi_{\mathbf{W}}$.

The eigenvalues and corresponding eigenvectors of any linear operator over a finite dimensional vector space do not depend on the choice of coordinate chart and basis, thus the eigenvalues and eigenvectors of $\mathbf{Hess}f_{\mathbf{FDA}}(\mathbf{W})[\cdot]:T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)\to T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ which are computed using the Courant Fischer Theorem for compact self-adjoint linear operators (Eq. (29) and Eq. (30)), can be also computed by the Courant Fischer Theorem for symmetric matrices (Horn and Johnson, 2012, Theorem 4.2.6) after applying $\varphi(\cdot)$. In particular, Eq. (35) determines the signs of the eigenvalues of the Riemannian Hessian at any $\mathbf{W}\in\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$ (in the special case $\mathbf{M_W}\coloneqq\mathbf{S_w}+\lambda\mathbf{I}_d$, the eigenvalues of Eq. (40) are the eigenvalues of the Riemannian Hessian at $\mathbf{W}\in\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$), and the bounds of Eq. (35) together with the bounds of the term $\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\left(\mathbf{S_w}+\lambda\mathbf{I}_d\right)\xi_{\mathbf{W}}\right)/\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}\mathbf{M_W}\xi_{\mathbf{W}}\right)$ bound the condition number of the Riemannian Hessian at $\mathbf{W}^{\star}\in\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$.

To that end, we perform the following computation. The righthand side of Eq. (39) is a Rayleigh quotient, so according to the Courant-Fischer Theorem for symmetric matrices the

eigenvalues (in particular, the maximal and the minimal) of the $pd-p(p+1)/2 \times pd-p(p+1)/2$ symmetric matrix

$$\mathbf{blkdiag}\left(\frac{1}{2}\mathbf{\Psi}, \mathbf{AN}\otimes\mathbf{I}_{d-p} - \mathbf{N}\otimes\tilde{\mathbf{A}}\right), \tag{40}$$

are determined by optimal (in particular the maximal and the minimal) values of Eq. (39). The set of eigenvalues of Eq. (40) is equal to the union of the set of eigenvalues of $\frac{1}{2}\mathbf{\Psi}$ and $\mathbf{AN}\otimes\mathbf{I}_{d-p} - \mathbf{N}\otimes\tilde{\mathbf{A}}$. Since Eq. (40) is a diagonal matrix, its eigenvalues are simply the diagonal elements:

$$\frac{1}{2}\left(\mu_j - \mu_i\right)\left(\alpha_j - \alpha_i\right),\ 1 \leq j < i \leq p,$$

and

$$\mu_j\left(\alpha_j - \alpha_i\right),\ j = 1, ..., p,\ i = p+1, ..., d.$$

Finally, we bound the condition number of the Riemannian Hessian at an optimal point, $\mathbf{W}^\star \in \mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$. In such case, $\alpha_1 = \rho_1 > ... > \alpha_p = \rho_p$. Without loss of generality, we can always choose $\mathbf{W}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)\perp}$ such that the corresponding eigenvalues to its columns are in a descending order, thus, $\alpha_{p+1} = \rho_{p+1} \geq ... \geq \alpha_d = \rho_d \geq 0$. Then, we have that Eq. (35) is bounded by the minimal and maximal eigenvalues of (40). Thus,

$$0 < \max_{\mathbf{0}\neq\xi_{\mathbf{W}^\star}\in T_{\mathbf{W}^\star}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)} \frac{\mathbf{Tr}\left(\xi_{\mathbf{W}^\star}^{\mathrm{T}}\left[-\mathbf{S_B}\xi_{\mathbf{W}^\star} + (\mathbf{S_w}+\lambda\mathbf{I}_d)\xi_{\mathbf{W}^\star}\mathbf{A}\right]\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{W}^\star}^{\mathrm{T}}(\mathbf{S_w}+\lambda\mathbf{I}_d)\xi_{\mathbf{W}^\star}\right)} \leq \mu_1\left(\rho_1 - \rho_d\right),$$

and

$$\begin{aligned}\min_{\mathbf{0}\neq\xi_{\mathbf{W}^\star}\in T_{\mathbf{W}^\star}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)} \frac{\mathbf{Tr}\left(\xi_{\mathbf{W}^\star}^{\mathrm{T}}\left[-\mathbf{S_B}\xi_{\mathbf{W}^\star} + (\mathbf{S_w}+\lambda\mathbf{I}_d)\xi_{\mathbf{W}^\star}\mathbf{A}\right]\mathbf{N}\right)}{\mathbf{Tr}\left(\xi_{\mathbf{W}^\star}^{\mathrm{T}}(\mathbf{S_w}+\lambda\mathbf{I}_d)\xi_{\mathbf{W}^\star}\right)} &\geq \\ \min\left\{\mu_p\left(\rho_p - \rho_{p+1}\right), \min_{1\leq j<p}\frac{1}{2}\left(\mu_j - \mu_{j+1}\right)\left(\rho_j - \rho_{j+1}\right)\right\} &> 0.\end{aligned}$$

We use Eq. (31) and Eq. (32) to bound the condition number:

$$\begin{aligned}\lambda_{\max}(\mathbf{Hess}f_{\mathbf{FDA}}(\mathbf{W}^\star)) &= \max_{\mathbf{0}\neq\xi_{\mathbf{W}^\star}\in T_{\mathbf{W}^\star}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)} \frac{g_{\mathbf{W}^\star}(\xi_{\mathbf{W}^\star}, \mathbf{Hess}f_{\mathbf{FDA}}(\mathbf{W}^\star)[\xi_{\mathbf{W}^\star}])}{g_{\mathbf{W}^\star}(\xi_{\mathbf{W}^\star}, \xi_{\mathbf{W}^\star})} \\ &\leq \mu_1\left(\rho_1 - \rho_d\right)\tilde{\lambda}_{\max}\ ,\end{aligned}$$

and

$$\begin{aligned}\lambda_{\min}(\mathbf{Hess}f_{\mathbf{FDA}}(\mathbf{W}^\star)) &= \min_{\mathbf{0}\neq\xi_{\mathbf{W}^\star}\in T_{\mathbf{W}^\star}\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)} \frac{g_{\mathbf{W}^\star}(\xi_{\mathbf{W}^\star}, \mathbf{Hess}f_{\mathbf{FDA}}(\mathbf{W}^\star)[\xi_{\mathbf{W}^\star}])}{g_{\mathbf{W}^\star}(\xi_{\mathbf{W}^\star}, \xi_{\mathbf{W}^\star})} \\ &\geq \min\left\{\mu_p\left(\rho_p - \rho_{p+1}\right), \min_{1\leq j<p}\frac{1}{2}\left(\mu_j - \mu_{j+1}\right)\left(\rho_j - \rho_{j+1}\right)\right\}\tilde{\lambda}_{\min}\ .\end{aligned}$$

Finally,

$$\kappa(\mathbf{Hess}f_{\mathbf{FDA}}(\mathbf{W}^\star)) = \frac{\lambda_{\max}(\mathbf{Hess}f_{\mathbf{FDA}}(\mathbf{W}^\star))}{\lambda_{\min}(\mathbf{Hess}f_{\mathbf{FDA}}(\mathbf{W}^\star))} \leq \kappa^\star_{\mathbf{FDA}}\cdot\kappa\left(\mathbf{S_w}+\lambda\mathbf{I}_d, \mathbf{M_W}\right),$$

where

$$\kappa^{\star}_{\mathbf{FDA}} = \frac{\mu_1 \left(\rho_1 - \rho_d\right)}{\min\left\{\mu_p \left(\rho_p - \rho_{p+1}\right), \min_{1 \le j < p} \frac{1}{2} \left(\mu_j - \mu_{j+1}\right)\left(\rho_j - \rho_{j+1}\right)\right\}}.$$

In the special case $\mathbf{M}_{\mathbf{W}^\star} = \mathbf{S}_{\mathbf{w}} + \lambda \mathbf{I}_d$, the bound on the condition number of the Riemannian Hessian at $\mathbf{W}^\star \in \mathbf{St}_{(\mathbf{S}_{\mathbf{w}}+\lambda \mathbf{I}_d)}(p,d)$ is reduced to

$$\kappa(\mathbf{Hess} f_{\mathbf{FDA}}(\mathbf{W}^\star)) = \kappa^{\star}_{\mathbf{FDA}}.$$

■

C.4 Proof of Theorem 17

Proof To show that $\mathbf{W}^\star$ is a strict local minimum of $f_{\mathbf{FDA}}(\mathbf{W})$ we show that the eigenvalues of the Riemannian Hessian at $\mathbf{W}^\star$ are strictly positive (see (Boumal, 2022, Proposition 6.5.)). Moreover, under the assumption that for all $i = 1, ..., d$ the values ρ_i are distinct, we prove that $\mathbf{W}^\star$ is the only local minimum of $f_{\mathbf{FDA}}(\mathbf{W})$ (and it is also strict) and all other critical points are either saddle points or strict local maximizers.

Eq. (34) helps to determine the signs of the eigenvalues of the Riemannian Hessian at any critical point $\mathbf{W} \in \mathbf{St}_{(\mathbf{S}_{\mathbf{w}}+\lambda \mathbf{I}_d)}(p,d)$ and in particular at $\mathbf{W}^\star$: the matrices $\mathbf{S}_{\mathbf{w}} + \lambda \mathbf{I}_d$ and $\mathbf{M}_{\mathbf{W}}$ are both SPD matrices, therefore for $\xi_{\mathbf{W}} \neq \mathbf{0}$ the term

$$\mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}}(\mathbf{S}_{\mathbf{w}} + \lambda \mathbf{I}_d)\xi_{\mathbf{W}}\right) / \mathbf{Tr}\left(\xi_{\mathbf{W}}^{\mathrm{T}} \mathbf{M}_{\mathbf{W}} \xi_{\mathbf{W}}\right) > 0,$$

thus only Eq. (35), where $\mathbf{0} \neq \xi_{\mathbf{W}} \in T_{\mathbf{W}}\mathbf{St}_{(\mathbf{S}_{\mathbf{w}}+\lambda \mathbf{I}_d)}(p,d)$ determines the signs. In addition, at a critical point $\mathbf{W} \in \mathbf{St}_{(\mathbf{S}_{\mathbf{w}}+\lambda \mathbf{I}_d)}(p,d)$ Eq. (35) equals to the quotient $R(\xi_{\mathbf{W}})$ for the choice $\mathbf{M}_{\mathbf{W}} \coloneqq \mathbf{S}_{\mathbf{w}} + \lambda \mathbf{I}_d$, since (Absil et al., 2008, Proposition 5.5.6 and Eq. (5.25)) show that at a critical point the term $g_{\mathbf{W}}(\xi_{\mathbf{W}}, \mathbf{Hess} f_{\mathbf{FDA}}(\mathbf{W})[\xi_{\mathbf{W}}])$ which is the numerator of $R(\xi_{\mathbf{W}})$ do not depend on the choice of Riemannian metric. Thus, the optimal values of Eq. (35) satisfying Eq. (29) or Eq. (30) are the eigenvalues of the Riemannian Hessian at $\mathbf{W} \in \mathbf{St}_{(\mathbf{S}_{\mathbf{w}}+\lambda \mathbf{I}_d)}(p,d)$ with the choice $\mathbf{M}_{\mathbf{W}} \coloneqq \mathbf{S}_{\mathbf{w}} + \lambda \mathbf{I}_d$. Obviously, classification of the critical points does not depend on the Riemannian metric. Therefore, we can classify the critical points using the signs of the eigenvalues of the Riemannian Hessian at $\mathbf{W} \in \mathbf{St}_{(\mathbf{S}_{\mathbf{w}}+\lambda \mathbf{I}_d)}(p,d)$ with the choice $\mathbf{M}_{\mathbf{W}} \coloneqq \mathbf{S}_{\mathbf{w}} + \lambda \mathbf{I}_d$.

Recall from the proof of Theorem 18 that in the special case $\mathbf{M}_{\mathbf{W}} \coloneqq \mathbf{S}_{\mathbf{w}} + \lambda \mathbf{I}_d$, the eigenvalues of Eq. (40 are also the eigenvalues of the Riemannian Hessian at a critical point $\mathbf{W} \in \mathbf{St}_{(\mathbf{S}_{\mathbf{w}}+\lambda \mathbf{I}_d)}(p,d)$. The eigenvalues are:

$$\frac{1}{2}\left(\mu_j - \mu_i\right)\left(\alpha_j - \alpha_i\right), \ 1 \le j < i \le p,$$

and

$$\mu_j \left(\alpha_j - \alpha_i\right), \ j = 1, ..., p, \ i = p+1, ..., d.$$

Now, we can conclude the signs of the eigenvalues of the Riemannian Hessian at any critical point $\mathbf{W} \in \mathbf{St}_{(\mathbf{S}_{\mathbf{w}}+\lambda \mathbf{I}_d)}(p,d)$ for the choice $\mathbf{M}_{\mathbf{W}} \coloneqq \mathbf{S}_{\mathbf{w}} + \lambda \mathbf{I}_d$, and classify these critical points. Recall that $\mu_i > 0$ for $i = 1, ..., p$, and $\mu_j - \mu_i > 0$ for $j < i$. Also $\alpha_i \ge 0$. Thus, the signs of the eigenvalues of Eq. (40) are only determined by the differences between

α_j, $j = 1, ..., p$ which are some generalized eigenvalues of the matrix pencil $(\mathbf{S_B}, \mathbf{S_w} + \lambda \mathbf{I}_d)$, corresponding to the generalized eigenvectors on ordered $1 \leq j < i \leq p$, but on the other hand there exists at least one pair of indexes $j = 1, ..., p$ and $i = p + 1, ..., d$ such that $\alpha_j - \alpha_i < 0$, otherwise it contradicts the assumption on the columns of $\mathbf{W}$, and the $d - j$ trailing α_i's, i.e., $\alpha_j - \alpha_i$ where $1 \leq j < i \leq p$ or $j = 1, ..., p$ and $i = p + 1, ..., d$.

In particular, for $\mathbf{W}^\star$ such that the columns are the p-dominant generalized eigenvectors of the matrix pencil $(\mathbf{S_B}, \mathbf{S_w} + \lambda \mathbf{I}_d)$, then by the assumption that $\alpha_1 = \rho_1 > ... > \alpha_p = \rho_p$, we have $\alpha_j - \alpha_i > 0$ where $1 \leq j < i \leq p$ or $j = 1, ..., p$ and $i = p + 1, ..., d$. Therefore, in this case all the eigenvalues of Eq. (40) are strictly positive, thus the matrix in Eq. (40) is SPD, and consequently the eigenvalues of the Riemannian Hessian at $\mathbf{W}^\star$ are all strictly positive. Thus, $\mathbf{W}^\star$ is a strict local minimum of $f_{\mathbf{FDA}}(\mathbf{W})$ on $\mathbf{St}_{(\mathbf{S_w} + \lambda \mathbf{I}_d)}(p, d)$.

If we prove that $\mathbf{W}^\star$ it is the only local minimum (up to the signs of the columns), then it is the only asymptotically stable critical point according to Theorem 15. In order to prove it we further assume that for all $i = 1, ..., d$ the values ρ_i are distinct, then we can conclude the following. Suppose $\mathbf{W}$ is any other critical point differs from $\mathbf{W}^\star$ at the optimal value, i.e., such that the columns of $\mathbf{W}$ are ordered generalized eigenvectors corresponding to some p singular values of the matrix pencil $(\mathbf{S_B}, \mathbf{S_w} + \lambda \mathbf{I}_d)$, $\alpha_1, ..., \alpha_p$, which are not the leading p values. We consider the different cases:

1. Suppose $\alpha_1, ..., \alpha_p$ are not ordered in any particular order (possible only for $p \geq 3$), then there exists j such that α_j is larger than some α_k and smaller than α_m where $j < k, m \leq p$, then $\alpha_j - \alpha_k > 0$ and $\alpha_j - \alpha_m < 0$. In this case there are both strictly positive and strictly negative eigenvalues of the Riemannian Hessian at $\mathbf{W}$ for the choice $\mathbf{M_W} \coloneqq \mathbf{S_w} + \lambda \mathbf{I}_d$, thus $\mathbf{W}$ is a saddle point.

2. Suppose $\alpha_1, ..., \alpha_p$ are ordered in a descending order. Since $\mathbf{Z}$ is not an optimal solution of Problem (8), then there exists at least one $1 \leq j \leq p$ for which $\alpha_j \neq \rho_j$. Thus, on the one hand $\alpha_j - \alpha_i > 0$ where $1 \leq j < i \leq p$, but on the other hand there exists at least one pair of indexes $j = 1, ..., p$ and $i = p + 1, ..., d$ such that $\alpha_j - \alpha_i < 0$, otherwise it contradicts the assumption that there exists at least one $1 \leq j \leq p$ for which $\alpha_j \neq \rho_j$. In this case there are both strictly positive and strictly negative eigenvalues of the Riemannian Hessian at $\mathbf{W}$ for the choice $\mathbf{M_W} \coloneqq \mathbf{S_w} + \lambda \mathbf{I}_d$, therefore, $\mathbf{W}$ is a saddle point.

3. Suppose $\alpha_1, ..., \alpha_p$ are ordered in an ascending order. Then, $\alpha_j - \alpha_i < 0$ where $1 \leq j < i \leq p$. Now, we consider two sub-cases:

   (a) There exists at least one $1 \leq j \leq p$ for which $\alpha_j \neq \rho_{d-j+1}$. Then, there exists at least one pair of indexes $j = 1, ..., p$ and $i = p + 1, ..., d$ such that $\alpha_j - \alpha_i > 0$, otherwise it contradicts the assumption that there exists at least one $1 \leq j \leq p$ for which $\alpha_j \neq \rho_{d-j+1}$. In this case there are both strictly positive and strictly negative eigenvalues of the Riemannian Hessian at $\mathbf{W}$ for the choice $\mathbf{M_W} \coloneqq \mathbf{S_w} + \lambda \mathbf{I}_d$, therefore, $\mathbf{W}$ is a saddle point.

   (b) Consider the case $\alpha_1 = \rho_d < \alpha_2 = \rho_{d-1} < ... < \alpha_p = \rho_{d-p+1}$. Thus, $\alpha_j - \alpha_i < 0$ where $1 \leq j < i \leq p$ or $j = 1, ..., p$ and $i = p + 1, ..., d$. In this case all the eigenvalues of the Riemannian Hessian at $\mathbf{W}$ are strictly negative for the choice

$\mathbf{M_W} \coloneqq \mathbf{S_w} + \lambda \mathbf{I}_d$, thus, $\mathbf{W}$ is a local maximizer. Since it is the only strict local maximizer up to the signs of the columns of $\mathbf{W}$, and it is also the global maximizer.

In all cases, $\mathbf{W}$ is not a local minimizer of $f_{\mathbf{FDA}}(\mathbf{W})$ on $\mathbf{St}_{(\mathbf{S_w}+\lambda\mathbf{I}_d)}(p,d)$. Thus, $\mathbf{W}^\star$ is the only local minimum (up to the signs of the columns. According to Theorem 15 all these critical points are unstable. ■

C.5 Proof of Corollary 20

Proof The condition number bound follows immediately from Lemma 2 and Theorem 18. For the costs, the same arguments as in the proof of Corollary 11 hold: none of the operations require forming $\mathbf{S_B}$ or $\mathbf{S_w}$, but instead require taking product of these matrices with vectors. These products can be computed in cost proportional to the iterated products of the matrices $\hat{\mathbf{X}}$ and/or $\hat{\mathbf{Y}}$ with vectors. Computing a product of the matrix $\hat{\mathbf{X}} = \mathbf{X} - \mathbf{Y}$ with a vector is equivalent to computing the product of the same vector with the matrices $\mathbf{X}$ and $\mathbf{Y}$ and subtracting the result. Computing the product of $\mathbf{X}$ with a vector is proportional number of non-zeros in $\mathbf{X}$, and the cost of the product of $\mathbf{Y}$ with a vector is $O(ld)$ since $\mathbf{Y}$ has exactly l distinct rows. Computing a product of the matrix $\hat{\mathbf{Y}}$ with a vector costs $O(ld)$ since $\hat{\mathbf{Y}}$ is a $l \times d$ matrix. The preprocessing steps follow from Table 2. Assuming a bounded number of line-search steps in each iteration of Riemannian CG, the costs follows from Table 2, as each iteration requires a bounded number of computations of each of the following: objective function evaluation $O\left(p\left(\mathbf{nnz}\left(\mathbf{X}\right) + ld\right)\right)$, retraction $O\left(\mathbf{nnz}\left(\hat{\mathbf{X}}\right)p + dp^2\right)$, vector transport and Riemannian gradient $O\left(p\left(\mathbf{nnz}\left(\mathbf{X}\right) + ld\right) + \mathbf{nnz}\left(\hat{\mathbf{X}}\right)p + dp^2 + d^2p\right)$. ■

References

P. A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008. ISBN 978-0-6911-3298-3.

N. Ailon and E. Liberty. Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete & Computational Geometry*, 42(4):615–630, 2009.

H. Avron, P. Maymounkov, and S. Toledo. Blendenpik: Supercharging LAPACK's least-squares solver. *SIAM J Sci Comput.*, 32(3):1217–1236, 2010.

H. Avron, C. Boutsidis, S. Toledo, and A. Zouzias. Efficient Dimensionality Reduction for Canonical Correlation Analysis. *SIAM J Sci Comput .*, 36(5):S111–S131, 2014a. doi: 10.1137/130919222. URL `http://dx.doi.org/10.1137/130919222`.

H. Avron, H. Nguyen, and D. Woodruff. Subspace Embeddings for the Polynomial Kernel. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Adv Neural Inf Process Syst. 27*, pages 2258–

2266. Curran Associates, Inc., 2014b. URL http://papers.nips.cc/paper/5240-subspace-embeddings-for-the-polynomial-kernel.pdf.

H. Avron, K. L. Clarkson, and D. P. Woodruff. Sharper Bounds for Regularized Data Fitting. In Klaus Jansen, José D. P. Rolim, David Williamson, and Santosh S. Vempala, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017)*, volume 81 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 27:1–27:22, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-044-6. doi: 10.4230/LIPIcs.APPROX-RANDOM.2017.27. URL http://drops.dagstuhl.de/opus/volltexte/2017/7576.

C. G. Baker. *Riemannian manifold trust-region methods with applications to eigenproblems.* The Florida State University, 2008.

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

N. Boumal. An introduction to optimization on smooth manifolds. To appear with Cambridge University Press, Jan 2022. URL http://www.nicolasboumal.net/book.

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

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

C. Boutsidis and A. Gittens. Improved matrix algorithms via the subsampled randomized Hadamard transform. *SIAM Journal on Matrix Analysis and Applications*, 34(3):1301–1340, 2013.

C. Boutsidis and D. P. Woodruff. Optimal CUR matrix decompositions. *SIAM Journal on Computing*, 46(2):543–589, 2017.

R. W. Brockett. Dynamical systems that sort lists, diagonalize matrices, and solve linear programming problems. *Linear Algebra Appl.*, 146:79–91, 1991.

J. Chapman, L. Wells, and A. L. Aguila. Unconstrained Stochastic CCA: Unifying Multiview and Self-Supervised Learning. In *The Twelfth International Conference on Learning Representations*, 2024.

M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. *Theor Comput Sci.*, 312(1):3 – 15, 2004. ISSN 0304-3975. doi: https://doi.org/10.1016/S0304-3975(03)00400-6. URL http://www.sciencedirect.com/science/article/pii/S0304397503004006. Automata, Languages and Programming.

K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 129–136. ACM, 2009.

M. Chen, W. Carson, M. Rodrigues, R. Calderbank, and L. Carin. Communications Inspired Linear Discriminant Analysis. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 1507–1514, USA, 2012. Omnipress. ISBN 978-1-4503-1285-1. URL http://dl.acm.org/citation.cfm?id=3042573.3042766.

A. Chowdhury, J. Yang, and P. Drineas. Randomized Iterative Algorithms for Fisher Discriminant Analysis. In *Proceedings of the 35th Conference in Uncertainty in Artificial Intelligence (UAI 2019)*, 2019.

A. Chowdhury, P. London, H. Avron, and P. Drineas. Faster Randomized Infeasible Interior Point Methods for Tall/Wide Linear Programs. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8704–8715. Curran Associates, Inc., 2020.

K. L. Clarkson and D. P. Woodruff. Low-Rank Approximation and Regression in Input Sparsity Time. *J. ACM*, 63(6):54:1–54:45, January 2017. ISSN 0004-5411. doi: 10.1145/3019134. URL http://doi.acm.org/10.1145/3019134.

M. B. Cohen, J. Nelson, and D. P. Woodruff. Optimal approximate matrix product in terms of stable rank, 2015. URL https://arxiv.org/abs/1507.02268.

P. Dhillon, D. P. Foster, and L. H. Ungar. Multi-view learning of word embeddings via CCA. In *Adv Neural Inf Process Syst.*, pages 199–207, 2011.

P. S. Dhillon, J. Rodu, D. P. Foster, and 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, pages 67–74, USA, 2012. Omnipress. ISBN 978-1-4503-1285-1. URL http://dl.acm.org/citation.cfm?id=3042573.3042586.

P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numer Math (Heidelb)*, 117(2):219–249, 2011.

R. O. Duda, P. E. Hart, and D. G. Stork. Pattern classification 2nd ed. *John Willey & Sons Inc*, 2001.

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

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

K. Fukunaga. *Introduction to statistical pattern recognition.* Elsevier, 2013.

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

B. Gao, R. Peng, and Y. Yuan. Optimization on product manifolds under a preconditioned metric, 2023.

R. Ge, C. Jin, S. Kakade, P. Netrapalli, and A. Sidford. Efficient Algorithms for Large-scale Generalized Eigenvector Computation and Canonical Correlation Analysis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 2741–2750. JMLR.org, 2016a. URL `http://dl.acm.org/citation.cfm?id=3045390.3045679`.

R. Ge, C. Jin, S. M. Kakade, P. Netrapalli, and A. Sidford. Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis. In *International Conference on Machine Learning (ICML)*, pages 2741–2750, 2016b.

G. H. Golub and H. Zha. The canonical correlations of matrix pairs and their numerical computation. In *Linear Algebra for Signal Processing*, pages 27–49. Springer, 1995.

A. Gonen, F. Orabona, and 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, pages 1397–1405. JMLR.org, 2016. URL `http://dl.acm.org/citation.cfm?id=3045390.3045538`.

U. Helmke and J. B. Moore. Singular-value decomposition via gradient and self-equivalent flows. *Linear algebra and its applications*, 169:223–248, 1992.

R. A. Horn and C. R. Johnson. *Matrix Analysis, 2nd Ed.* Cambridge University Press, 2012.

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

R. Kannan, S. Vempala, and D. Woodruff. Principal component analysis and higher correlations for distributed data. In *Conference on Learning Theory*, pages 1040–1057. PMLR, 2014.

T. Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.

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

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

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

Z. Ma, Y. Lu, and D. Foster. Finding Linear Structure in Large Datasets with Scalable Canonical Correlation Analysis. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 169–178. JMLR.org, 2015. URL `http://dl.acm.org/citation.cfm?id=3045118.3045138`.

X. Meng and M. W. Mahoney. Low-distortion Subspace Embeddings in Input-sparsity Time and Applications to Robust Linear Regression. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 91–100, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2029-0. doi: 10.1145/2488608.2488621. URL http://doi.acm.org/10.1145/2488608.2488621.

X. Meng, M. A. Saunders, and M. W. Mahoney. LSRN: A parallel iterative solver for strongly over-or underdetermined systems. *SIAM J Sci Comput.*, 36(2):C95–C118, 2014.

S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K-R. Mullers. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.*, pages 41–48. IEEE, 1999.

T. P. Minka. Old and new matrix algebra useful for statistics. *https://tminka.github.io/papers/matrix/*, 2000.

B. Mishra and R. Sepulchre. R3MC: A Riemannian three-factor algorithm for low-rank matrix completion. In *53rd IEEE Conference on Decision and Control*, pages 1137–1142. IEEE, 2014.

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

U. Mor and H. Avron. Solving Trust Region Subproblems Using Riemannian Optimization. *arXiv preprint arXiv:2010.07547*, 2020.

T. Ngo and Y. Saad. Scaled gradients on Grassmann manifolds for matrix completion. In *Adv Neural Inf Process Syst.*, pages 1412–1420, 2012.

T. T. Ngo, M. Bellalij, and Y. Saad. The trace ratio optimization problem. *SIAM review*, 54(3):545–569, 2012.

V. Rokhlin and M. Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, 105(36):13212–13217, 2008. ISSN 0027-8424. doi: 10.1073/pnas.0804869105. URL https://www.pnas.org/content/105/36/13212.

H. Sato and T. Iwai. A Riemannian optimization approach to the matrix singular value decomposition. *SIAM Journal on Optimization*, 23(1):188–212, 2013.

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

B. Shustin and H. Avron. Riemannian optimization with a preconditioning scheme on the generalized Stiefel manifold. *Journal of Computational and Applied Mathematics*, 423: 114953, 2023. ISSN 0377-0427. doi: https://doi.org/10.1016/j.cam.2022.114953. URL https://www.sciencedirect.com/science/article/pii/S0377042722005519.

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

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

L. Sun, B. Ceran, and 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*, pages 313–322. ACM, 2010.

J. Townsend, N. Koep, and S. Weichwald. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *J Mach Learn Res.*, 17(1):4755–4759, 2016.

J. A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis*, 3(01n02):115–126, 2011.

C. Udriste. *Convex functions and optimization methods on Riemannian manifolds*, volume 297. Springer Science & Business Media, 2013.

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

J. Von Neumann. Some matrix-inequalities and metrization of matric-space. Tomsk Univ. Rev., 1: 286–300, 1937. Reprinted in Collected Works, 1962.

H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang. Trace ratio vs. ratio trace for dimensionality reduction. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

W. Wang, J. Wang, D. Garber, and N. Srebro. Efficient Globally Convergent Stochastic Optimization for Canonical Correlation Analysis. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 766–774. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper_files/paper/2016/file/42998cf32d552343bc8e460416382dca-Paper.pdf`.

D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1–2):1–157, 2014.

J. Yang, X. Meng, and M. W. Mahoney. Implementing randomized matrix algorithms in parallel and distributed environments. *Proc IEEE Inst Electr Electron Eng.*, 104(1):58–92, 2016.

F. Yger, M. Berar, G. Gasso, and 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, pages 299–306, USA, 2012. Omnipress. ISBN 978-1-4503-1285-1. URL `http://dl.acm.org/citation.cfm?id=3042573.3042615`.

T. Zhou, H. Qian, Z. Shen, C. Zhang, and 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*, pages 3539–3545, 2017. doi: 10.24963/ijcai.2017/495. URL `https://doi.org/10.24963/ijcai.2017/495`.