Riemannian optimization using three different metrics for Hermitian PSD fixed-rank constraints: an extended version

Shixin Zheng
Department of Mathematics
Purdue University
West Lafayette, USA
zheng513@purdue.edu

Wen Huang
School of Mathematical Sciences
Xiamen University
Xiamen, China
wen.huang@xmu.edu.cn

Bart Vandereycken
Section of Mathematics
University of Geneva
Geneva, Switzerland
bart.vandereycken@unige.ch

Xiangxiong Zhang
Department of Mathematics
Purdue University
West Lafayette, USA
zhan1966@purdue.edu

Abstract

We consider smooth optimization problems with a Hermitian positive semi-definite fixed-rank constraint, where a quotient geometry with three Riemannian metrics $g^i(\cdot,\cdot)$ $(i = 1, 2, 3)$ is used to represent this constraint. By taking the nonlinear conjugate gradient method (CG) as an example, we show that CG on the quotient geometry with metric g^1 is equivalent to CG on the factor-based optimization framework, which is often called the Burer–Monteiro approach. We also show that CG on the quotient geometry with metric g^3 is equivalent to CG on the commonly-used embedded geometry. We call two CG methods equivalent if they produce an identical sequence of iterates $\{X_k\}$. In addition, we show that if the limit point of the sequence $\{X_k\}$ generated by an algorithm has lower rank, that is $X_k \in \mathbb{C}^{n\times n}, k = 1, 2, \dots$ has rank p and the limit point X_* has rank $r < p$, then the condition number of the Riemannian Hessian with metric g^1 can be unbounded, but those of the other two metrics stay bounded. Numerical experiments show that the Burer–Monteiro CG method has slower local convergence rate if the limit point has a reduced rank, compared to CG on the quotient geometry under the other two metrics. This slower convergence rate can thus be attributed to the large condition number of the Hessian near a minimizer.

Keywords Riemannian optimization · Hermitian fixed-rank positive semidefinite matrices · embedded manifold · quotient manifold · Burer–Monteiro · conjugate gradient · Riemannian Hessian

1 Introduction

1.1 The Hermitian PSD low-rank constraints

In this paper we are interested in algorithms for minimizing a real-valued function f with a Hermitian positive semi-definite (PSD) low-rank constraint

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & f(X) \\ \text{subject to} \quad & X \in \mathcal{H}_+^{n,p} \end{aligned}, \tag{1}$$

where $\mathcal{H}_+^{n,p}$ denotes the set of n-by-n Hermitian PSD matrices of fixed rank $p \ll n$. Even though $X \in \mathcal{H}_+^{n,p}$ is a nonconvex constraint, in practice (1) is often used for approximating solutions to a minimization with a convex PSD constraint:

$$\begin{array}{ll} \underset{X \in \mathbb{C}^{n\times n}}{\text{minimize}} & f(X) \\ \text{subject to} & X \succcurlyeq 0 \end{array}. \tag{2}$$

Among all kinds of matrix constraints, PSD matrices are abundant in applications and recent research. They arise in semidefinite programming serving as covariance matrices in statistics and kernels in machine learning, etc. See [1] and [2] for a reference of these applications. If the solution of (2) is of low rank and $\mathcal{O}(n^2)$ complexity is too large for storage or computation, it is preferable to consider a low-rank representation of PSD matrices. For example, real symmetric PSD fixed-rank matrices were used in [3, 4].

Since the elements in the constraint set $\mathcal{H}_+^{n,p}$ have a low-rank structure, they can be represented in a low-rank compact form on the order of $\mathcal{O}(np^2)$, which is smaller than the $\mathcal{O}(n^2)$ storage when directly using $X \in \mathbb{C}^{n\times n}$. In many applications, the cost function in (2) takes the form $f(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_F^2$ where $\mathcal{A}$ is a linear operator and the norm is the Frobebius norm, and $f(X)$ can be evaluated efficiently by $\mathcal{O}(pn \log n)$ flops for $X \in \mathcal{H}_+^{n,p}$; see, e.g., the PhaseLift problem [5, 6] and the interferometry recovery problem [7, 8]. For these kinds of problems, solving (1) with an iterative algorithm that works with low-rank representations for $X \in \mathcal{H}_+^{n,p}$ can lead to a good approximate solution to (2) with compact storage and computational cost.

1.2 The real inner product and Fréchet derivatives

Since $f(X)$ is real-valued and thus not holomorphic, $f(X)$ does not have a complex derivative with respect to $X \in \mathbb{C}^{n\times n}$. In this paper, all linear spaces of complex matrices will therefore be regarded as vector spaces over $\mathbb{R}$. For any real vector space $\mathcal{E}$, the inner product on $\mathcal{E}$ is denoted by $\langle ., . \rangle_{\mathcal{E}}$. For real matrices $A, B \in \mathbb{R}^{m\times n}$, the Hilbert–Schmidt inner product is $\langle A, B\rangle_{\mathbb{R}^{m\times n}} = tr(A^T B)$. Let $\Re(A)$ and $\Im(B)$ represent the real and imaginary parts of a complex matrix A. For $A, B \in \mathbb{C}^{m\times n}$, the real inner product for the real vector space $\mathbb{C}^{m\times n}$ then equals

$$\langle A, B\rangle_{\mathbb{C}^{m\times n}} := \Re(tr(A^* B)), \tag{3}$$

where * is the conjugate transpose. We emphasize that (3) is a real inner product, rather than the complex Hilbert—Schmidt inner product. It is straightforward to verify that (3) can be written as

$$\langle A, B\rangle_{\mathbb{C}^{m\times n}} = tr(\Re(A)^T \Re(B)) + tr(\Im(A)^T \Im(B)) = \langle \Re(A), \Re(B)\rangle_{\mathbb{R}^{m\times n}} + \langle \Im(A), \Im(B)\rangle_{\mathbb{R}^{m\times n}}.$$

With the real inner product (3) for the real vector space $\mathbb{C}^{m\times n}$, a Fréchet derivative for any real valued function $f(X)$ can be defined as

$$\nabla f(X) = \nabla f_{\Re(X)}(X) + \mathbb{i}\nabla f_{\Im(X)}(X) \in \mathbb{C}^{m\times n}, \tag{4}$$

where $\nabla f_{\Re(X)}(X), \nabla f_{\Im(X)}(X) \in \mathbb{R}^{m\times n}$ are the gradient of the cost function f with respect to the real and imaginary parts of X, respectively. In particular, for $f(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_F^2$ with a linear operator $\mathcal{A}$, the Fréchet derivative (4) becomes

$$\nabla f(X) = \mathcal{A}^*(\mathcal{A}(X) - b)$$

where $\mathcal{A}^*$ is the adjoint operator of $\mathcal{A}$. See Appendix A for details.

1.3 Three different methodologies

In this paper we will consider three straightforward ideas and methodologies for solving (1).

1.3.1 The Burer–Monteiro method

The first approach, often called the Burer–Monteiro method [9], is to solve the unconstrained problem

$$\min_{Y \in \mathbb{C}^{n\times p}} F(Y) := f(YY^*). \tag{5}$$

As proven in Appendix A, the chain rule of Fréchet derivatives gives

$$\nabla F(Y) = [\nabla f(YY^*) + \nabla f(YY^*)^*]\, Y \in \mathbb{C}^{n\times p}.$$

The gradient descent method simply takes the form of

$$Y_{n+1} = Y_n - \tau \nabla F(Y_n) = Y_n - \tau\, [\nabla f(Y_n Y_n^*) + \nabla f(Y_n Y_n^*)^*]\, Y_n, \tag{6}$$

which is one of the simplest low-rank algorithms. The nonlinear conjugate gradient and quasi-Newton type methods, like L-BFGS, can also be easily used for (5). On the other hand, $F(Y) = F(YO)$ for any unitary matrix $O \in \mathbb{O}^{p\times p}$, where

$$\mathcal{O}_p = \{O \in \mathbb{C}^{p\times p} : O^*O = OO^* = I\}.$$

Even though this ambiguity of unitary matrices is never explicitly addressed in the Burer–Monteiro method, in this paper we will prove that the gradient descent and nonlinear conjugate gradient methods for solving (5) are exactly equivalent to the Riemannian gradient descent and Riemannian conjugate gradient methods on a quotient manifold. Thus the convergence of the Burer–Monteiro method can be understood within the context of Riemannian optimizaiton on a quotient manifold.

1.3.2 Riemannian optimization with the embedded geometry of $\mathcal{H}_+^{n,p}$

Another natural approach is to regard $\mathcal{H}_+^{n,p}$ as an embedded manifold in $\mathbb{C}^{n\times n}$. For instance, Riemannian optimization algorithms on the embedded manifold of low-rank matrices and tensors are quite efficient and popular [10, 11]. Even though it is possible to study $\mathcal{H}_+^{n,p} \subset \mathbb{C}^{n\times n}$ as a complex manifold, we will regard $\mathbb{C}^{n\times n}$ as a $2n^2$-dimensional real vector space and $\mathcal{H}_+^{n,p} \subset \mathbb{C}^{n\times n}$ as a manifold over $\mathbb{R}$ since $f(X)$ is real-valued. In particular, the embedded geometry of $\mathcal{S}_+^{n,p}$, representing the set real symmetric PSD low-rank matrices, was studied in [12].

A *Riemannian metric* is a smoothly varying inner product defined on the tangent space. The Riemannian metric of the real embedded manifold $\mathcal{H}_+^{n,p}$ can simply be taken as the inner product (3) on $\mathbb{C}^{n\times n}$. The embedded geometry of the real manifold $\mathcal{H}_+^{n,p} \subset \mathbb{C}^{n\times n}$ will be discussed in Section 3.

Even though the tangent space and the Riemannian gradient in Section 3 for $\mathcal{H}_+^{n,p}$ look like a natural extension of those for $\mathcal{S}_+^{n,p}$, it is not obvious why this should be true. The subtlety lies in the fact that we have to regard $X \in \mathcal{H}_+^{n,p}$ as an element of a real vector space. For instance, for regarding $X \in \mathcal{H}_+^{n,p}$ as a real vector, one can either regard a complex matrix X as the pair of its real and imaginary part, or regard $X \in \mathbb{C}^{n\times n}$ with its realification, which is a $2n$-by-$2n$ real matrix generated by replacing each complex entry $a + \mathbb{i}b$ of X by a 2-by-2 block $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$. But neither way gives a straightforward generalization from the real case in [12] to the complex case in Section 3. Instead, with the real inner product (3) and the corresponding Fréchet derivative, it is possible to achieve the desired generalization.

1.3.3 Riemannian optimization by using quotient geometry

The third approach is to consider the quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$, which will be reviewed in Section 4. Here $\mathbb{C}_*^{n\times p}$ is the noncompact Stiefel manifold of full rank n-by-p matrices:

$$\mathbb{C}_*^{n\times p} = \{X \in \mathbb{C}^{n\times p} : \text{rank}(X) = p\}.$$

Define an equivalent class by

$$[Y] = \{Z \in \mathbb{C}_*^{n\times p} : Z = YO, O \in \mathcal{O}_p\}$$

and denote the natural projection as

$$\begin{aligned} \pi : \mathbb{C}_*^{n\times p} &\to \mathbb{C}_*^{n\times p}/\mathcal{O}_p \\ Y &\mapsto [Y] \end{aligned}.$$

Since there is a one-to-one correspondence between $X = YY^* \in \mathcal{H}_+^{n,p}$ and $\pi(Y) \in \mathbb{C}_*^{n\times p}/\mathcal{O}_p$, the optimization problem (1) is equivalent to

$$\begin{aligned} &\underset{\pi(Y)}{\text{minimize}} && h(\pi(Y)) \\ &\text{subject to} && \pi(Y) \in \mathbb{C}_*^{n\times p}/\mathcal{O}_p \end{aligned}, \tag{7}$$

where the cost function h is defined as $h(\pi(Y)) = F(Y) = f(YY^*)$.

For the quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$, one can first choose a metric for its total space $\mathbb{C}_*^{n\times p}$, which induces a Riemannian metric on the quotient manifold under suitable conditions. In particular, a special metric was used in [13] to construct efficient Riemannian optimization algorithms for the problem (5). The horizontal lift of the Riemannian gradient for $h(\pi(Y))$ under this particular metric satisfies

$$(\text{grad}\, h(\pi(Y)))_Y = \nabla F(Y)(Y^*Y)^{-1} = [\nabla f(YY^*) + \nabla f(YY^*)^*]\, Y(Y^*Y)^{-1}. \tag{8}$$

From the representation of the Riemannian gradient (8), we see that this approach generates different algorithms from the simpler Burer–Monteiro approach.

1.4 Main results: a unified representation and analysis of three methods using quotient geometry

A natural question arises: which of the three methods is the best? Even though the unconstrained Burer–Monteiro method is quite straightforward to use, it has an ambiguity up to a unitary matrix, and its performance is usually observed to be inferior to Riemannian optimization on embedded and quotient geometries. In order to compare these three methods, in this paper we will show that it is possible to equivalently rewrite both the Burer–Monteiro approach and embedded manifold approach as Riemannian optimization over the quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ with suitable metrics, retractions and vector transports.

For any $Y \in \mathbb{C}_*^{n\times p}$, we consider three different Riemannian metrics $g_Y^i(\cdot,\cdot)$ $(i = 1, 2, 3)$ for any A, B in the total space $\mathbb{C}_*^{n\times p}$:

$$\begin{aligned}
g_Y^1(A,B) &= \langle A, B\rangle_{\mathbb{C}^{n\times p}} = \Re(tr(A^*B))\\
g_Y^2(A,B) &= \langle AY^*, BY^*\rangle_{\mathbb{C}^{n\times n}} = \Re(tr((Y^*Y)A^*B))\\
g_Y^3(A,B) &= \langle YA^* + AY^*, YB^* + BY^*\rangle_{\mathbb{C}^{n\times n}} + \left\langle P_Y^{\mathcal{V}}(A)Y^*, P_Y^{\mathcal{V}}(B)Y^*\right\rangle_{\mathbb{C}^{n\times n}},
\end{aligned}$$

where $P_Y^{\mathcal{V}}$ is given by

$$P_Y^{\mathcal{V}}(A) = Y\left(\frac{(Y^*Y)^{-1}Y^*A - A^*Y(Y^*Y)^{-1}}{2}\right).$$

In particular, the Burer–Monteiro approach corresponds to the first metric g_Y^1 and the embedded manifold approach corresponds to the third metric g_Y^3. The second metric g_Y^2 is the one used in [13].

We will show that both the gradient descent and the conjugate gradient method for the unconstrained problem (5) are equivalent to a Riemannian gradient descent and a Riemannian conjugate gradient method on the quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ with the simplest metric g_Y^1 and a specific vector transport.

Furthermore, we will prove that the Riemannian gradient descent and the Riemannian conjugate gradient methods using the embedded geometry of $\mathcal{H}_+^{n,p}$ are equivalent to a Riemannian gradient descent and a Riemannian conjugate gradient algorithms on the quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ with the metric g_Y^3 and a specific vector transport.

Finally, we will analyze and compare the condition numbers of the Riemannian Hessian using these three different metrics by estimating their Rayleigh quotient. It is well known that the condition number of the Hessian of the cost function is closely related to the asymptotic performance of optimization methods. Under the assumption that the Fréchet Hessian $\nabla^2 f(X)$ is well conditioned, we will show that the the condition numbers of the Riemannian Hessian using the first metric can be significantly worse than the other two if the minimizer of (2) has a rank smaller than p.

1.5 Related work

The Burer–Monteiro approach for the PSD constraint has been popular in applications due to its simplicity. For instance, an L-BFGS method for (5) was used for solving convex recovery from interferometric measurements in [8]. It is straightforward to verify that (6) with $p = 1$ and a suitable step size τ for the PhaseLift problem [5] is precisely the Wirtinger flow algorithm [6]. In [14], it was shown that first-order and second-order optimality conditions of the nonconvex Burer–Monteiro approach are sufficient to find the global minimizer of the convex semi-definite program under certain assumptions.

Riemannian optimization on various matrix manifolds such as the Stiefel manifold, the Grassmann manifold and the set of fixed-rank matrices, have been used for applications in data science, machine learning, signal processing, bioscience, etc. The geometry of real symmetric PSD matrices of fixed rank $\mathcal{S}_+^{n,p}$ has also been studied intensively in the literature. Its embedded geometry was studied in [12]. The quotient geometry was studied in [15, 16, 1]. In [2], a new geometry of $\mathcal{S}_+^{n,p}$ as a homogeneous space of the general linear group of positive determinant GL_n^+ was discussed.

Riemannian optimization based on the embedded geometry has been well studied in [10] for real matrices of fixed rank, which can be easily extended to real symmetric PSD matrices of fixed rank [12]. As expected, Section 3 is its natural extensions to Hermitian PSD matrices of fixed rank. This is not surprising, but it is not a straightforward result either, because such a natural extension holds only when using the real inner product (3) and its associated Fréchet derivatives.

The quotient geometry of Hermitian PSD matrices of fixed-rank for the metric g_Y^2 has been studied in [17, 13]. The quotient geometry with metric g_Y^2 in this paper is exactly the same one as the one in [17, 13].

It is not uncommon to explore different metrics of a manifold for Riemannian optimization [18, 19]. In [2], a new embedded geometry and complete geodesics for real PSD fixed-rank matrices were for example obtained from a special quotient metric.

1.6 Contributions

In this paper, for simplicity, we only focus on the nonlinear conjugate gradient method.

First, we will prove that the nonlinear conjugate gradient method for the unconstrained Burer–Monteiro formulation (5) is equivalent to a Riemannian conjugate gradient method on the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^1)$ for solving (7). Thus the convergence of the simple Burer–Monteiro approach can be understood in the context of Riemannian optimization on the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^1)$. This is one major contribution of this paper.

Second, we will show that a Riemannian conjugate gradient method on the embedded manifold $\mathcal{H}_+^{n,p}$ for solving (1) is equivalent to a Riemannian conjugate gradient method on the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^3)$ for solving (7). For implementation, this is not necessary and there is no motivation to explicitly implement a Riemannian optimization algorithm on an embedded geometry as a Riemannian optimization algorithm on a quotient geometry. However, it is useful when comparing a Riemannian optimization algorithm on an embedded geometry with a Riemannian optimization algorithm on a quotient geometry.

Finally, for the sake of understanding the differences among the three methodologies, we will analyze the condition number of the Riemannian Hessian on the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^i)$ for the three different metrics g^i $(i = 1, 2, 3)$. One metric is equivalent to the simple Burer–Monteiro approach and another to Riemannian optimization on the embedded manifold $\mathcal{H}_+^{n,p}$. Since the three methods in Section 1.3 can all be regarded as Riemannian optimization algorithms on a quotient manifold with three different metrics, such a comparison is meaningful.

In certain problems, such as PhaseLift [5] and interferometry recovery [8], the rank r of the minimizer of (2) is known. However, it has been observed in practice that the basin of attraction is larger when solving the nonconvex problems (5) or (7) with rank $p > r$ instead of with rank $p = r$; see [8, 13]. We will also demonstrate this in the numerical tests in Section 8. Under suitable assumptions, we will show that the condition number of the Riemannian Hessian on the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^1)$ can be unbounded if $p > r$. On the other hand, the condition numbers of the Riemannian Hessians on the quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ with metrics g^1 and g^2 are still bounded. This is consistent with the numerical observation that the Burer–Monteiro approach has a much slower asymptotic convergence rate than the Riemannian optimization approach on the embedded manifold and the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^2)$ when $p > r$.

1.7 Organization of the paper

The outline of the paper is as follows. We summarize the notation in Section 2. Then we discuss the geometric operators such as the Riemannian gradient and vector transport in Section 3 for the embedded manifold $\mathcal{H}_+^{n,p}$ and in Section 4 for the quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$. In Section 5, we outline the Riemannian Conjugate Gradient (RCG) methods on different geometries and discuss equivalences among them. In particular, we show that RCG on the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^1)$ is exactly the Burer–Monteiro CG method, that is, CG directly on (5). We also show that Riemannian CG on the embedded manifold for solving (1) is equivalent to RCG on the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^3$ with a specific retraction and vector transport for solving (7). Implementation details are given in Section 6. In Section 7, we analyze and compare the condition numbers of the Riemannian Hessian operators, which can be used to understand the difference in the asymptotic convergence rates between using the simple Burer–Monteiro method and the more sophisticated Riemannian optimization using an embedded geometry or a quotient geometry with metric g^2. Numerical tests are given in Section 8.

2 Notation

Let $\mathbb{C}^{m\times n}$ denote all complex matrices of size $m \times n$. Let $p \le n$ and define

$$\begin{aligned}
\mathbb{C}_*^{n\times p} &= \{X \in \mathbb{C}^{n\times p} : \text{rank}(X) = p\},\\
\text{St}\,(p,n) &= \{X \in \mathbb{C}^{n\times p} : X^*X = I_p\},\\
\mathcal{H}_+^{n,p} &= \{X \in \mathbb{C}^{n\times n} : X^* = X, X \succcurlyeq 0, \text{rank}(X) = p\},\\
\mathcal{S}_+^{n,p} &= \{X \in \mathbb{R}^{n\times n} : X^T = X, X \succcurlyeq 0, \text{rank}(X) = p\},
\end{aligned}$$

$$\mathcal{O}_p \quad = \quad \{O \in \mathbb{C}^{p\times p} : O^*O = OO^* = I\},$$

where St (p, n) is also called the compact Stiefel manifold. For a matrix X, X^* denotes its conjugate transpose and $\overline{X}$ denotes its complex conjugate. If X is real, X^* becomes the matrix transpose and is denoted by X^T. We define

$$Herm(X) := \frac{X + X^*}{2}, \quad Skew(X) := \frac{X - X^*}{2}.$$

Let $\Re(X)$ and $\Im(X)$ denote the real and imaginary part of X respectively so that $X = \Re(X) + \mathbb{i}\Im(X)$. Let I_p be the identity matrix of size p-by-p. For any n-by-p matrix Z, $Z_\perp$ denotes the n-by-$(n-p)$ matrix such that $Z_\perp^* Z_\perp = I_{n-p}$ and $Z_\perp^* Z = \mathbf{0}$.

Let $\mathrm{Diag}(m, n)$ be the set of all m-by-n diagonal matrices. Let $\mathrm{diag}(M)$ be the n-by-1 vector that is the diagonal of the n-by-n matrix M. Given a vector v, $\mathrm{Diag}(v)$ is a square matrix with its ith diagonal entry equal to v_i. Given a matrix A, $tr(A)$ denotes the trace of A and A_{ij} denotes the (i, j)th entry of A.

For any $X \in \mathcal{H}_+^{n,p}$, its eigenvalues coincide with its singluar values. The compact singular value decomposition (SVD) of X is denoted by $X = U\Sigma U^*$, where $U \in \mathrm{St}\,(p, n)$ and $\Sigma = \mathrm{Diag}(\sigma)$ with $\sigma = (\sigma_1, \cdots, \sigma_p)^T$ and $\sigma_1 \geq \cdots \geq \sigma_p > 0$. In the rest of the paper, U and Σ are reserved for denoting the compact SVD of $X \in \mathcal{H}_+^{n,p}$.

In this paper, all manifolds of complex matrices are viewed as manifolds over $\mathbb{R}$ unless otherwise specified. Given a Euclidean space $\mathcal{E}$, the inner product on $\mathcal{E}$ is denoted by $\langle .,. \rangle_{\mathcal{E}}$. Specifically, $\langle A, B\rangle_{\mathbb{R}^{m\times n}} = tr(A^T B)$ for $A, B \in \mathbb{R}^{m\times n}$ and $\langle A, B\rangle_{\mathbb{C}^{m\times n}} = \Re(tr(A^* B))$ for $A, B \in \mathbb{C}^{m\times n}$ denotes the canonical inner product on $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$, respectively.

3 Embedded geometry of $\mathcal{H}_+^{n,p}$

The results in this section are natural extensions of results for $\mathcal{S}_+^{n,p}$ in [12]. Such an extension is not entirely obvious since $\mathcal{H}_+^{n,p}$ is treated as a real manifold and the real inner product (3) is not the complex Hilbert–Schmidt inner product. For completeness, we thus discuss these extensions in details.

3.1 Tangent space

We first need to show that $\mathcal{H}_+^{n,p}$ is a smooth embedded submanifold of $\mathbb{C}^{n\times n}$. See [20, Prop. 2.1] and [21, Chap. 5] for the case of $\mathcal{S}_+^{n,p}$.

Theorem 3.1. *Regard $\mathbb{C}^{n\times n}$ as a real vector space over $\mathbb{R}$ of dimension $2n^2$. Then $\mathcal{H}_+^{n,p}$ is a smooth embedded submanifold of $\mathbb{C}^{n\times n}$ of dimension $2np - p^2$.*

Proof. Let

$$E = \begin{bmatrix} I_{p\times p} & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & 0_{(n-p)\times(n-p)} \end{bmatrix} \tag{9}$$

and consider the smooth Lie group action

$$\begin{aligned} \Phi : \mathrm{GL}(n, \mathbb{C}) \times \mathbb{C}^{n\times n} &\to \quad \mathbb{C}^{n\times n} \\ (g, N) &\mapsto \quad gNg^* \end{aligned}$$

where

$$\begin{aligned} gNg^* \quad = \quad & (\Re(g)\Re(N) - \Im(g)\Im(N))\,\Re(g)^T + (\Im(g)\Re(N) + \Re(g)\Im(N))\Im(g)^T \\ & + \mathbb{i}\left((\Im(g)\Re(N) + \Re(g)\Im(N))\Re(g)^T - (\Re(g)\Re(N) - \Im(g)\Im(N))\Im(g)^T\right). \end{aligned}$$

It is easy to see that Φ is a rational mapping. Since $\mathrm{GL}(n, \mathbb{C})$ is a semialgebraic set by Lemma (B.1) in the Appendix, we have that $\mathrm{GL}(n, \mathbb{C}) \times \mathbb{C}^{n\times n}$ is also a semialgebraic set [22, section 2.1.1]. It follows from (B1) in [23] that Φ is a semialgebraic mapping. Observe that $\mathcal{H}_+^{n,p}$ is the orbit of E through Φ. It therefore follows from (B4) in [23] that $\mathcal{H}_+^{n,p}$ is a smooth submanifold of $\mathbb{C}^{n\times n}$.

Next, we compute the dimension of $\mathcal{H}_+^{n,p}$. Consider the smooth surjective mapping

$$\eta : \mathrm{GL}(n, \mathbb{C}) \to \mathcal{H}_+^{n,p} \quad \gamma \mapsto \gamma E \gamma^*. \tag{10}$$

The differential of η at $\gamma \in \mathrm{GL}(n,\mathbb{C})$ is the linear mapping $\mathrm{D}\eta(\gamma) : T_\gamma \mathrm{GL}(n,\mathbb{C}) = \mathbb{C}^{n\times n} \to T_X\mathcal{H}_+^{n,p}$, where $X = \eta(\gamma) = \gamma E\gamma^*$, by $\mathrm{D}\eta(\gamma)[\Delta] = \Delta E\gamma^* + \gamma E\Delta^*$. Observe that the differential at arbitrary γ is related to the differential at I_n by a full-rank linear transformation:

$$\mathrm{D}\eta(\gamma)[\Delta] = \gamma\mathrm{D}\eta(I_n)[\gamma^{-1}\Delta]\gamma^*. \tag{11}$$

Recall that the rank of a differentiable mapping f between two differentiable manifolds is the dimension of the image of the differential of f. So, from equation (11) we see that the rank of η is constant. It follows from Theorem 4.14 in [24] that η is a smooth submersion. As a consequence $\mathrm{D}\eta(\gamma)$ maps $T_\gamma\mathrm{GL}(n,\mathbb{C}) \cong \mathbb{C}^{n\times n}$ surjectively onto $T_X\mathcal{H}_+^{n,p}$ and we obtain

$$T_X\mathcal{H}_+^{n,p} = \left\{\Delta X + X\Delta^* : \Delta \in \mathbb{C}^{n\times n}\right\}. \tag{12}$$

Let $\Delta = \begin{bmatrix}\Delta_{11} & \Delta_{12}\\ \Delta_{21} & \Delta_{22}\end{bmatrix}$ be partitioned according to the partition of $E = \mathrm{diag}(I_{p\times p}) = \begin{bmatrix}I_{p\times p} & 0\\ 0 & 0\end{bmatrix}$. Then it can be easily verified that $\Delta \in \mathrm{Ker}\mathrm{D}\eta(I)$ if and only if

$$\Delta_{11} = -\Delta_{11}^*, \quad \Delta_{21} = 0.$$

This implies that Δ_{11} is a skew-Hermitian matrix, hence its diagonal entries are purely imaginary and its off diagonal entries satisfy $a_{ij} = -\overline{a_{ji}}$. This gives us $p+2\times(1+2+\cdots+(p-1))$ degrees of freedom. For Δ_{12} and Δ_{22} there are $2n(n-p)$ degrees of freedom. So, the dimension of $\mathrm{Ker}(\mathrm{D}\eta(I))$ is $2n(n-p)+p+2p(p-1)/2 = 2n^2-2np+p^2$ and by rank-nullity we get

$$\dim \mathrm{D}\,\eta(I) = 2n^2 - \dim\ker \mathrm{D}\,\eta(I) = 2np - p^2. \tag{13}$$

Since η is of constant rank, the dimension of $T_X\mathcal{H}_+^{n,p}$ is therefore $2np-p^2$. Remember that the dimension of the tangent space at every point of a connected manifold is the same as that of the manifold itself. Let $\mathrm{GL}^+(n,\mathbb{C})$ denote the connected subset of $\mathrm{GL}(n,\mathbb{C})$ with positive determinant, then $\mathcal{H}_+^{n,p}$ is the image of the connected set $\mathrm{GL}^+(n,\mathbb{C})$ under a continuous mapping η, so $\mathcal{H}_+^{n,p}$ is connected. We conclude that the dimension of $\mathcal{H}_+^{n,p}$ is $2np-p^2$. ☐

The next result characterizes the tangent space. See [10, Proposition 2.1] for the tangent space of $\mathcal{S}_+^{n,p}$.

Theorem 3.2. *Let $X = U\Sigma U^* \in \mathcal{H}_+^{n,p}$. Then the tangent space of $\mathcal{H}_+^{n,p}$ at X is given by*

$$T_X\mathcal{H}_+^{n,p} = \left\{[U \quad U_\perp]\begin{bmatrix}H & K^*\\ K & 0\end{bmatrix}\begin{bmatrix}U^*\\ U_\perp^*\end{bmatrix}\right\} \tag{14}$$

where $H = H^* \in \mathbb{C}^{p\times p}$, $K \in \mathbb{C}^{(n-p)\times p}$.

Remark 3.3. *Notice that there is no need to compute and store $U_\perp \in \mathbb{C}^{n\times(n-p)}$ and it suffices to store $U_\perp K \in \mathbb{C}^{n\times p}$. See Section 6 for the implementation details.*

Proof. Let $t \mapsto U(t)$ be any smooth curve in $\mathrm{St}\,(p,n)$ through U at $t=0$ such that $U(t) \in \mathbb{C}^{n\times p}, U(0) = U$ and $U(t)^*U(t) = I_p$ for all t. Let $t \mapsto \Sigma(t)$ be any smooth curve in $\mathrm{Diag}(p,p)$ through Σ at $t=0$. Then $X(t) := U(t)\Sigma(t)U(t)^*$ defines a smooth curve in $\mathcal{H}_+^{n,p}$ through X. It follows by differentiating $X(t) := U(t)\Sigma(t)U(t)^*$ that

$$X'(t) = U'(t)\Sigma(t)U(t)^* + U(t)\Sigma'(t)U(t)^* + U(t)\Sigma(t)U'(t)^*. \tag{15}$$

Without loss of generality, since $U'(t)$ is an element of $\mathbb{C}^{n\times p}$ and $U(t)$ has full rank, we can set

$$U'(t) = U(t)A(t) + U_\perp(t)B(t). \tag{16}$$

Hence, we have

$$X'(t) = [U(t) \quad U_\perp(t)]\begin{bmatrix}A(t)\Sigma(t)+\Sigma'(t)+\Sigma(t)A(t)^* & \Sigma(t)B(t)^*\\ B(t)\Sigma(t) & 0\end{bmatrix}\begin{bmatrix}U(t)^*\\ U_\perp(t)^*\end{bmatrix}. \tag{17}$$

Thus we consider the tangent vectors in the form of $[U \quad U_\perp]\begin{bmatrix}H & K^*\\ K & 0\end{bmatrix}\begin{bmatrix}U^*\\ U_\perp^*\end{bmatrix}$ with $H = H^*$. For any $H = H^* \in \mathbb{C}^{p\times p}$ and $K \in \mathbb{C}^{(n-p)\times p}$, taking $\Delta = (UH/2 + U_\perp K)\Sigma^{-1}(U^*U)^{-1}U^*$ in (12), we see that

$$\left\{[U \quad U_\perp]\begin{bmatrix}H & K^*\\ K & 0\end{bmatrix}\begin{bmatrix}U^*\\ U_\perp^*\end{bmatrix}\right\} \subseteq T_X\mathcal{H}_+^{n,p}. \tag{18}$$

Now counting the real dimension we see that H has $p + 2\times\frac{p(p-1)}{2} = p^2$ number of freedom and K has $2\times p(n-p)$ number of freedom. So the LHS of the inclusion (18) has freedom $2np-p^2$, which is equal to the dimension of $T_X\mathcal{H}_+^{n,p}$. Hence, the inclusion in (18) is an equality. ☐

3.2 Riemannian gradient

The *Riemannian metric* of the embedded manifold at $X \in \mathcal{H}_+^{n,p}$ is induced from the Euclidean inner product on $\mathbb{C}^{n\times n}$,

$$g_X(\zeta_1, \zeta_2) = \langle \zeta_1, \zeta_2 \rangle_{\mathbb{C}^{n\times n}} = \Re(tr(\zeta_1^* \zeta_2)), \quad \zeta_1, \zeta_2 \in T_X\mathcal{H}_+^{n,p}. \tag{19}$$

Let $f(X)$ be a smooth real-valued function for $X \in \mathbb{C}^{n\times n}$ and Fréchet gradient (4), denoted by $\nabla f(X)$. See Appendix A.1 for more details about Fréchet gradient.

The *Riemannian gradient* of f at $X \in \mathcal{H}_+^{n,p}$, denoted by $\operatorname{grad} f(X)$, is the projection of $\operatorname{grad} f(X)$ onto $T_X\mathcal{H}_+^{n,p}$ ([25, Sect. 3.6.1]):

$$\operatorname{grad} f(X) = P_X^t(\nabla f(X)), \tag{20}$$

where P_X^t denotes the orthogonal projection onto $T_X\mathcal{H}_+^{n,p}$. In order to get a closed-form expression of P_X^t, we should characterize the *normal space* to $\mathcal{H}_+^{n,p}$ at X, denoted by $(T_X\mathcal{H}_+^{n,p})^\perp$ or $N_X\mathcal{H}_+^{n,p}$,

$$N_X\mathcal{H}_+^{n,p} = \{\xi_X \in T_X\mathbb{C}^{n\times n} : \langle \xi_X, \eta_X \rangle_{\mathbb{C}^{n\times n}} = 0 \text{ for all } \eta_X \in T_X\mathbb{C}^{n\times n}\}, \tag{21}$$

which is the orthogonal complement of $T_X\mathcal{H}_+^{n,p}$ in $\mathbb{C}^{n\times n}$.

Lemma 3.4. *The normal space $N_X\mathcal{H}_+^{n,p}$ at $X = U\Sigma U^* \in \mathcal{H}_+^{n,p}$ is given by*

$$N_X\mathcal{H}_+^{n,p} = \left\{ \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} \Omega & -L^* \\ L & M \end{bmatrix} \begin{bmatrix} U^* \\ U_\perp^* \end{bmatrix} \right\}, \tag{22}$$

where $\Omega = -\Omega^* \in \mathbb{C}^{p\times p}$, $M \in \mathbb{C}^{(n-p)\times(n-p)}$ and $L \in \mathbb{C}^{(n-p)\times p}$.

Proof. First we show that every vector in (22) is orthogonal to $T_X\mathcal{H}_+^{n,p}$. Since U is orthonormal, we only need to show that $\left\langle \begin{bmatrix} H & K^* \\ K & 0 \end{bmatrix}, \begin{bmatrix} \Omega & -L^* \\ L & M \end{bmatrix} \right\rangle_{\mathbb{C}^{n\times n}} = 0$ for all H, K, Ω, L and M defined in Theorem 3.2 and Lemma 3.4. Indeed we have

$$\begin{aligned} \left\langle \begin{bmatrix} H & K^* \\ K & 0 \end{bmatrix}, \begin{bmatrix} \Omega & -L^* \\ L & M \end{bmatrix} \right\rangle_{\mathbb{C}^{n\times n}} &= \langle \Omega, H \rangle_{\mathbb{C}^{n\times n}} - \langle L^*, K^* \rangle_{\mathbb{C}^{n\times n}} + \langle L, K \rangle_{\mathbb{C}^{n\times n}} && (23) \\ &= \langle \Omega, H \rangle_{\mathbb{C}^{n\times n}} = 0. && (24) \end{aligned}$$

Next, we count the real dimension of $N_X\mathcal{H}_+^{n,p}$. Remember that a skew-Hermitian matrix has purely imaginary numbers on its diagonal entries, and $\omega_{ij} = -\overline{\omega}_{ji}$ on its off diagonal entries. So the number of degree of freedoms in Ω is $p + 2 \times \frac{p(p-1)}{2} = p^2$. The number of degree of freedoms in L is $2 \times p(n-p)$, and the number of degree of freedoms in M is $2 \times (n-p)^2$. So, the dimension of $N_X\mathcal{H}_+^{n,p}$ is $2n^2 + p^2 - 2np$. This gives us the desire dimension since the sum of the dimension of the tangent space and its normal space should be $2n^2$. □

The orthogonal projection from $\mathbb{C}^{n\times n}$ onto $T_X\mathcal{H}_+^{n,p}$ is given the following theorem.

Theorem 3.5. *Let $X = YY^* = U\Sigma U^*$ be the compact SVD for $X \in \mathcal{H}_+^{n,p}$ with $Y \in \mathbb{C}^{n\times p}$. Let $Z \in \mathbb{C}^{n\times n}$. Then the operator P_X^t defined below is the orthogonal projection onto $T_X\mathcal{H}_+^{n,p}$:*

$$\begin{aligned} P_X^t(Z) &= \frac{1}{2}\left(P_Y(Z+Z^*)P_Y + P_Y^\perp(Z+Z^*)P_Y + P_Y(Z+Z^*)P_Y^\perp\right) \\ &= \frac{1}{2}\left(P_U(Z+Z^*)P_U + P_U^\perp(Z+Z^*)P_U + P_U(Z+Z^*)P_U^\perp\right) \\ &= \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} U^*\frac{(Z+Z^*)}{2}U & U^*\frac{(Z+Z^*)}{2}U_\perp \\ U_\perp^*\frac{(Z+Z^*)}{2}U & 0 \end{bmatrix} \begin{bmatrix} U^* \\ U_\perp^* \end{bmatrix}, \end{aligned} \tag{25}$$

where $P_Y = Y(Y^*Y)^{-1}Y^*$, $P_Y^\perp = I - P_Y = P_{Y_\perp}$, $P_U = UU^*$ and $P_U^\perp = I - P_U = P_{U_\perp}$.

Proof. First, observe that

$$P_X^t(Z) = \begin{bmatrix} P_Y & P_{Y_\perp} \end{bmatrix} \begin{bmatrix} \frac{Z+Z^*}{2} & \frac{Z+Z^*}{2} \\ \frac{Z+Z^*}{2} & 0 \end{bmatrix} \begin{bmatrix} P_Y \\ P_{Y_\perp} \end{bmatrix}$$

$$= \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} U^* \frac{(Z+Z^*)}{2} U & U^* \frac{(Z+Z^*)}{2} U_\perp \\ U_\perp^* \frac{(Z+Z^*)}{2} U & 0 \end{bmatrix} \begin{bmatrix} U^* \\ U_\perp^* \end{bmatrix}$$

is a tangent vector at X. So it suffices to show that $Z - P_X^t(Z)$ is a normal vector. Write Z as $Z = P_Y Z P_Y + P_Y Z P_{Y_\perp} + P_{Y_\perp} Z P_Y + P_{Y_\perp} Z P_{Y_\perp} = [P_Y \quad P_{Y_\perp}] \begin{bmatrix} Z & Z \\ Z & Z \end{bmatrix} \begin{bmatrix} P_Y \\ P_{Y_\perp} \end{bmatrix}$. Then we have

$$\begin{aligned} Z - P_X^t(Z) &= [P_Y \quad P_{Y_\perp}] \begin{bmatrix} \frac{Z-Z^*}{2} & \frac{Z-Z^*}{2} \\ \frac{Z-Z^*}{2} & Z \end{bmatrix} \begin{bmatrix} P_Y \\ P_{Y_\perp} \end{bmatrix} \\ &= [U \quad U_\perp] \begin{bmatrix} U^* \frac{(Z-Z^*)}{2} U & U^* \frac{(Z-Z^*)}{2} U_\perp \\ U_\perp^* \frac{(Z-Z^*)}{2} U & U_\perp^* Z U_\perp \end{bmatrix} \begin{bmatrix} U^* \\ U_\perp^* \end{bmatrix} \end{aligned}$$

Hence, $Z - P_X^t(Z)$ is a normal vector, which completes the proof. □

Remark 3.6. *We can write $P_X^t = P_X^s + P_X^p$ by introducing the two operators*

$$P_X^s : Z \mapsto P_U \frac{Z+Z^*}{2} P_U \tag{26}$$

$$P_X^p : Z \mapsto P_{U_\perp} \frac{Z+Z^*}{2} P_U + P_U \frac{Z+Z^*}{2} P_{U_\perp} \tag{27}$$

3.3 A retraction by projection to the embedded manifold

A retraction is essentially a first-order approximation to the exponential map; see [25, Def. 4.1.1]. Suppose $\mathcal{M}$ is an embedded submanifold of a Euclidean space $\mathcal{E}$, then by [26, Props. 3.2 and 3.3], the mapping R from the tangent bundle $T\mathcal{M}$ to the manifold $\mathcal{M}$ defined by

$$R : \begin{cases} T\mathcal{M} \to \mathcal{M} \\ (x, u) \mapsto P_{\mathcal{M}}(x+u) \end{cases} \tag{28}$$

is a retraction, where $P_{\mathcal{M}}$ is the orthogonal projection onto the manifold $\mathcal{M}$ with respect to the Euclidean distance, that is, the closest point. In our case $\mathcal{M} = \mathcal{H}_+^{n,p}$ and $\mathcal{E} = \mathbb{C}^{n\times n}$. Hence, a retraction on $\mathcal{H}_+^{n,p}$ is defined by the truncated SVD:

$$R_X(Z) := P_{\mathcal{H}_+^{n,p}}(X+Z) = \sum_{i=1}^{p} \sigma_i(X+Z) v_i v_i^*,$$

where v_i is the singular vector of $X+Z$ corresponding to the ith largest singular value $\sigma_i(X+Z)$.

Let $X = U\Sigma U^* \in \mathcal{H}_+^{n,p}$ be the compact SVD and let $Z = [U \quad U_\perp] \begin{bmatrix} H & K^* \\ K & 0 \end{bmatrix} \begin{bmatrix} U^* \\ U_\perp^* \end{bmatrix} \in T_X \mathcal{H}_+^{n,p}$. Then

$$X + Z = [U \quad U_\perp] \begin{bmatrix} H+\Sigma & K^* \\ K & 0 \end{bmatrix} \begin{bmatrix} U^* \\ U_\perp^* \end{bmatrix} = U(H+\Sigma)U^* + U_\perp K U^* + U K^* U_\perp^*. \tag{29}$$

Consider the compact QR factorization of $U_\perp K = Q_K R_K$ where Q_K is $n \times p$ and R_K is $p \times p$. Then (29) becomes

$$X + Z = U(H+\Sigma)U^* + Q_K R_K U^* + (Q_K R_K U^*)^* = [U \quad Q_K] \begin{bmatrix} H+\Sigma & R_K^* \\ R_K & 0 \end{bmatrix} \begin{bmatrix} U^* \\ Q_K^* \end{bmatrix}. \tag{30}$$

Now notice that $\begin{bmatrix} H+\Sigma & R_K^* \\ R_K & 0 \end{bmatrix}$ from the RHS of (30) is a small $2p \times 2p$ Hermitian matrix. We can therefore efficiently compute its SVD as

$$\begin{bmatrix} H+\Sigma & R_K^* \\ R_K & 0 \end{bmatrix} = [V_1 \quad V_2] \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} V_1^* \\ V_2^* \end{bmatrix}, \tag{31}$$

where S_1 and S_2 are $p \times p$ diagonal matrices that contain the singular values of $\begin{bmatrix} H+\Sigma & R_K^* \\ R_K & 0 \end{bmatrix}$ in descending order. The matrices V_1 and V_2 are $2p \times p$ contain the corresponding singular vectors. Combining (31) and (30), we can write $X+Z$ as

$$X + Z = [U \quad Q_K] [V_1 \quad V_2] \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} V_1^* \\ V_2^* \end{bmatrix} \begin{bmatrix} U^* \\ Q_K^* \end{bmatrix} \tag{32}$$

with $[U \quad Q_K]\,[V_1 \quad V_2]$ a unitary matrix. So (32) is the SVD of $X+Z$ with singular values in descending order. The orthogonal projection of $X+Z$ onto the manifold $\mathcal{H}_+^{n\times p}$ is therefore given by

$$P_{\mathcal{H}_+^{n,p}}(X+Z) = ([U \quad Q_K]\,V_1)S_1([U \quad Q_K]\,V_1)^*. \tag{33}$$

3.4 Vector transport

The vector transport is a mapping that transports a tangent vector from one tangent space to another tangent space.

Definition 3.1 ([25, definition 8.1.1])**.** *A vector transport on a manifold $\mathcal{M}$ is a smooth mapping*

$$T\mathcal{M} \oplus T\mathcal{M} \to T\mathcal{M} : (\eta_x, \xi_x) \mapsto \mathcal{T}_{\eta_x}(\xi_x) \in T\mathcal{M} \tag{34}$$

satisfying the following properties for all $x \in \mathcal{M}$:

1. *(Associated retraction) There exists a retraction R, called the retraction associated with $\mathcal{T}$, such that the following diagram commutes*

$$\begin{array}{ccc} (\eta_x, \xi_x) & \xrightarrow{\ \mathcal{T}\ } & \mathcal{T}_{\eta_x}(\xi_x) \\ \big\downarrow & & \big\downarrow \Pi \\ \eta_x & \xrightarrow{\ R\ } & \Pi(\mathcal{T}_{\eta_x}(\xi_x)) \end{array}$$

   where $\Pi(\mathcal{T}_{\eta_x}(\xi_x))$ denotes the foot of the tangent vector $\mathcal{T}_{\eta_x}(\xi_x)$.

2. *(Consistency) $\mathcal{T}_{0_x}\xi_x = \xi_x$ for all $\xi_x \in T_x\mathcal{M}$;*

3. *(Linearity) $\mathcal{T}_{\eta_x}(a\xi_x + b\zeta_x) = a\mathcal{T}_{\eta_x}(\xi_x) + b\mathcal{T}_{\eta_x}(\zeta_x)$.*

Let $\xi_X, \eta_X \in T_X\mathcal{H}_+^{n,p}$ and let R be a retraction on $\mathcal{H}_+^{n,p}$. By [25, section 8.1.3], the projection of one tangent vector onto another tangent space is a vector transport,

$$\mathcal{T}_{\eta_X}\xi_X := P^t_{R_X(\eta_X)}\xi_X, \tag{35}$$

where P^t_Z is the projection operator onto $T_Z\mathcal{H}_+^{n,p}$. Namely, we first apply retraction to $X+\eta_X$ to arrive at a new point on the manifold, then we project the old tangent vector ξ_X onto the tangent space at that new point.

Now, we derive the expression of the vector transport (35) in closed form. Given $X_1 = U_1\Sigma_1U_1^* \in \mathcal{H}_+^{n,p}$, the retracted point $X_2 = U_2\Sigma_2U_2^* \in \mathcal{H}_+^{n,p}$, and a tangent vector $\nu_1 = [U_1 \quad U_{1\perp}]\begin{bmatrix} H_1 & K_1^* \\ K_1 & 0 \end{bmatrix}\begin{bmatrix} U_1^* \\ U_{1\perp}^* \end{bmatrix} = U_1H_1U_1^* + U_{1\perp}K_1U_1^* + U_1K_1^*U_{1\perp}^* \in T_{X_1}\mathcal{H}_+^{n,p}$, we need to determine H_2 and K_2 of the transported tangent vector $\nu_2 = [U_2 \quad U_{2\perp}]\begin{bmatrix} H_2 & K_2^* \\ K_2 & 0 \end{bmatrix}\begin{bmatrix} U_2^* \\ U_{2\perp}^* \end{bmatrix} \in T_{X_2}\mathcal{H}_+^{n,p}$.

By the projection formula (25), we have

$$\nu_2 = P^t_{X_2}(\nu_1) = [U_2 \quad U_{2\perp}]\begin{bmatrix} U_2^*\nu_1U_2 & U_2^*\nu_1U_{2\perp} \\ U_{2\perp}^*\nu_1U_2 & 0 \end{bmatrix}\begin{bmatrix} U_2^* \\ U_{2\perp}^* \end{bmatrix}. \tag{36}$$

Hence, H_2 and K_2 are satisfy

$$\begin{aligned} H_2 &= U_2^*\nu_1U_2 = U_2^*U_1H_1U_1^*U_2 + U_2^*U_{1\perp}K_1U_1^*U_2 + U_2^*U_1K_1^*U_{1\perp}^*U_2, \\ K_2 &= U_{2\perp}^*\nu_1U_2 = U_{2\perp}^*U_1H_1U_1^*U_2 + U_{2\perp}^*U_{1\perp}K_1U_1^*U_2 + U_{2\perp}^*U_1K_1^*U_{1\perp}^*U_2. \end{aligned}$$

In implementation, we observe a better numerical performance if we only keep the first term in the above sum of H_2 and the second term of K_2. That is, we define H_2 and K_2 by

$$H_2 = U_2^*U_1H_1U_1^*U_2 \tag{37a}$$
$$K_2 = U_{2\perp}^*U_{1\perp}K_1U_1^*U_2. \tag{37b}$$

It is straightforward to verify that this choice of H_2 and K_2 also defines a vector transport:

Proposition 3.7. *The operation defined by* (37) *is a vector transport.*

One can verify that the vector transport in (37) is a vector transport by parallelization in [27].

In numerical tests we have observed that the nonlinear conjugate gradient method using this simpler version of vector transport is usually more efficient. So in all our numerical tests, we do not use the more complicated (35), instead we use the following simplified vector transport:

1. Given $X_1 = U_1\Sigma_1 U_1^* \in \mathcal{H}_+^{n,p}$, and $\eta_{X_1}, \xi_{X_1} \in T_{X_1}\mathcal{H}_+^{n,p}$, first compute
$$X_2 = R_{X_1}(\eta_{X_1}) := P_{\mathcal{H}_+^{n,p}}(X_1 + \eta_{X1}) = U_2\Sigma_2 U_2^* \in \mathcal{H}_+^{n,p}.$$
2. Let $\xi_{X_1} = [U_1 \quad U_{1\perp}] \begin{bmatrix} H_1 & K_1^* \\ K_1 & 0 \end{bmatrix} \begin{bmatrix} U_1^* \\ U_{1\perp}^* \end{bmatrix} \in T_{X_1}\mathcal{H}_+^{n,p}$, then compute
$$\mathcal{T}_{\eta_{X_1}}\xi_{X_1} = [U_2 \quad U_{2\perp}] \begin{bmatrix} H_2 & K_2^* \\ K_2 & 0 \end{bmatrix} \begin{bmatrix} U_2^* \\ U_{2\perp}^* \end{bmatrix} \in T_{X_2}\mathcal{H}_+^{n,p}. \tag{37c}$$

3.5 Riemannian Hessian operator

For a real-valued function $f(X)$ defined on the Euclidean space $\mathbb{C}^{n\times n}$, the Hessian $\nabla^2 f(X)$ is defined in the sense of the Fréchet derivative; see Appendix A.2 for the definition of the Fréchet Hessian.

The following proposition gives the Riemannian Hessian of f. The proof follows similar ideas as in [4, Prop. 5.10] and [28, Prop. 2.3] where a second-order retraction based on a simple power expansion is constructed. We will leave the outline of the proof to Appendix B.1.

Proposition 3.8. *Let $f(X)$ be the a real-valued function defined on $\mathcal{H}_+^{n,p}$. Let $X \in \mathcal{H}_+^{n,p}$ and $\xi_X \in T_X\mathcal{H}_+^{n,p}$. Then the Riemannian Hessian operator of f at X is given by*
$$\textit{Hess}\, f(X)[\xi_X] = P_X^t(\nabla^2 f(X)[\xi_X]) + P_X^p\left(\nabla f(X)(X^\dagger \xi_X^p)^* + (\xi_X^p X^\dagger)^* \nabla f(X)\right). \tag{38}$$

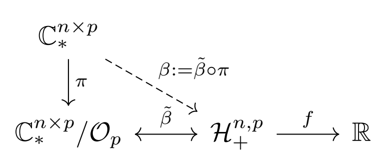
4 The quotient geometry of $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ using three Riemannian metrics

Besides being regarded as an embedded manifold in $\mathbb{C}^{n\times n}$, $\mathcal{H}_+^{n,p}$ can also be viewed as a quotient set $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ since any $X \in \mathcal{H}_+^{n,p}$ can be written as $X = YY^*$ with $Y \in \mathbb{C}_*^{n\times p}$. We define an equivalence relation on $\mathbb{C}_*^{n\times p}$ through the smooth Lie group action of $\mathcal{O}_p$ on the manifold $\mathbb{C}_*^{n\times p}$:
$$\begin{aligned} \mathbb{C}_*^{n\times p} \times \mathcal{O}_p &\to \quad \mathbb{C}_*^{n\times p} \\ (Y, O) &\mapsto \quad YO. \end{aligned} \tag{39}$$
This action defines an equivalence relation on $\mathbb{C}_*^{n\times p}$ by setting $Y_1 \sim Y_2$ if there exists an $O \in \mathcal{O}_p$ such that $Y_1 = Y_2 O$. Hence we have constructed a quotient space $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ that removes this ambiguity. The set $\mathbb{C}_*^{n\times p}$ is called the *total space* of $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$.

Denote the natural projection as
$$\pi : \mathbb{C}_*^{n\times p} \to \mathbb{C}_*^{n\times p}/\mathcal{O}_p.$$
For any $Y \in \mathbb{C}_*^{n\times p}$, $\pi(Y)$ is an element in $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$. We denote the equivalent class containing Y as
$$[Y] = \pi^{-1}(\pi(Y)) = \{YO | O \in \mathcal{O}_p\}.$$

Define
$$\begin{aligned} \beta : \mathbb{C}_*^{n\times p} &\to \mathcal{H}_+^{n,p} \\ Y &\mapsto YY^*. \end{aligned}$$

Then β is invariant under the equivalence relation $\sim$ and induces a unique function $\tilde{\beta}$ on $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$, called the projection of β, such that $\beta = \tilde{\beta} \circ \pi$ [25, section 3.4.2]. One can easily check that $\tilde{\beta}$ is a bijection. For any real-valued function $f(X)$ defined on $X = YY^* \in \mathcal{H}_+^{n,p}$, $F(Y) := f \circ \beta(Y) = f(YY^*)$ is a real-valued function defined on $\mathbb{C}_*^{n\times p}$ and F induces f. This is summarized in the diagram below:



(40)

The next theorem shows that $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ is a smooth manifold.

Theorem 4.1. *The quotient space $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ is a quotient manifold over $\mathbb{R}$ of dimension $2np - p^2$ and has a unique smooth structure such that the natural projection π is a smooth submersion.*

Proof. The proof follows from Corollary 21.6 and Theorem 21.10 of [24]. □

The next theorem shows that $\mathcal{H}_+^{n,p}$ and $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ are essentially the same in the sense that there is a diffeomorphism between them. The proof uses the same technique in [1, Prop. A.7]

Theorem 4.2. *The quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ is diffeomorphic to the embedded manifold $\mathcal{H}_+^{n,p}$ under $\tilde{\beta}$.*

Proof. Recall from Theorem 3.2, any tangent vector in $T_{\beta(Y)}\mathcal{H}_+^{n,p}$ can be written as

$$\zeta_{\beta(Y)} = YHY^* + Y_\perp KY^* + YK^*Y_\perp^*. \tag{41}$$

Let $V = YH/2 + Y_\perp K$, then $D\beta(Y)[V] = \zeta_{\beta(Y)}$. This implies that β is a submersion.

Now notice that $\pi = \tilde{\beta}^{-1} \circ \beta$ and $\beta = \tilde{\beta} \circ \pi$. By [29, Prop. 6.1.2], we conclude that $\tilde{\beta}^{-1}$ and $\tilde{\beta}$ are both differentiable. So $\tilde{\beta}$ is a diffeomorphism between $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ and $\mathcal{H}_+^{n,p}$. □

4.1 Vertical space, three Riemannian metrics and horizontal space

The equivalence class $[Y]$ is an embedded submanifold of $\mathbb{C}_*^{n\times p}$([25, Prop. 3.4.4]). The tangent space of $[Y]$ at Y is therefore a subspace of $T_Y\mathbb{C}_*^{n\times p}$ called the *vertical space* at Y and is denoted by $\mathcal{V}_Y$. The following proposition characterizes $\mathcal{V}_Y$.

Proposition 4.3. *The vertical space at $Y \in [Y] = \{YO|O \in \mathcal{O}_p\}$, which is the tangent space of $[Y]$ at Y is*

$$\mathcal{V}_Y = \left\{Y\Omega|\Omega^* = -\Omega, \Omega \in \mathbb{C}^{p\times p}\right\}. \tag{42}$$

Proof. The tangent space of $\mathcal{O}_p$ at I_p is $T_{I_p}\mathcal{O}_p = \{\Omega : \Omega^* = -\Omega, \Omega \in \mathbb{C}^{p\times p}\}$, which is also the set $\{\gamma'(0) : \gamma \text{ is a curve in } \mathcal{O}_p, \gamma(0) = I_p\}$. Hence $T_Y\{YO|O \in \mathcal{O}_p\} = \{Y\gamma'(0) : \gamma \text{ is a curve in } \mathcal{O}_p, \gamma(0) = I_p\} = \{Y\Omega|\Omega^* = -\Omega, \Omega \in \mathbb{C}^{p\times p}\}$. □

A *Riemannian metric* g is a smoothly varying inner product defined on the tangent space. That is, $g_Y(\cdot,\cdot)$ is an inner product on $T_Y\mathbb{C}_*^{n\times p}$. Once we choose a Riemannian metric g for $\mathbb{C}_*^{n\times p}$, we can obtain the orthogonal complement in $T_Y\mathbb{C}_*^{n\times p}$ of $\mathcal{V}_Y$ with respect to the metric. In other words, we choose the *horizontal distribution* as orthogonal complement w.r.t. Riemannian metric, see [25, Section 3.5.8]. This orthogonal complement to $\mathcal{V}_Y$ is called *horizontal space* at Y and is denoted by $\mathcal{H}_Y$. We thus have

$$T_Y\mathbb{C}_*^{n\times p} = \mathcal{H}_Y \oplus \mathcal{V}_Y. \tag{43}$$

Once we have the horizontal space, there exists a unique vector $\bar{\xi}_Y \in \mathcal{H}_Y$ that satisfies $\mathrm{D}\,\pi(Y)[\bar{\xi}_Y] = \xi_{\pi(Y)}$ for each $\xi_{\pi(Y)} \in T_{\pi(Y)}\mathbb{C}_*^{n\times p}/\mathcal{O}_p$. This $\bar{\xi}_Y$ is called the *horizontal lift* of $\xi_{\pi(Y)}$ at Y. The next lemma shows the relationship between the horizontal lifts of the quotient tangent vector $\xi_{\pi(Y)}$ lifted at different representatives in $[Y]$.

Lemma 4.4. *Let η be a vector field on $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$, and let $\bar{\eta}$ be the horizontal lift of η. Then for each $Y \in \mathbb{C}_*^{n\times p}$, we have*

$$\bar{\eta}_{YO} = \bar{\eta}_Y O$$

for all $O \in \mathcal{O}_p$.

Proof. See [1, Prop. A.8] and [13, Lemma 5.1]. □

There exist more than one choice of Riemannian metric on $\mathbb{C}_*^{n\times p}$. Different Riemanian metrics do not affect the vertical space, but generally result in different horizontal spaces. In this paper, we discuss three Riemannian metrics on $\mathbb{C}_*^{n\times p}$ and study how each metric affects the convergence of Riemannian optimization algorithms.

The most straightforward choice of a Riemannian metric on $\mathbb{C}_*^{n\times p}$ is the canonical Euclidean inner product on $\mathbb{C}^{n\times p}$ defined by

$$g_Y^1(A, B) := \langle A, B\rangle_{\mathbb{C}^{n\times p}} = \Re(tr(A^*B)), \quad \forall A, B \in T_Y\mathbb{C}_*^{n\times p} = \mathbb{C}^{n\times p}. \tag{44}$$

Proposition 4.5. *Under metric g^1, the horizontal space at Y satisfies*

$$\begin{aligned}\mathcal{H}_Y^1 &= \left\{Z \in \mathbb{C}^{n\times p} : Y^*Z = Z^*Y\right\}\\ &= \left\{Y(Y^*Y)^{-1}S + Y_\perp K | S^* = S, S \in \mathbb{C}^{p\times p}, K \in \mathbb{C}^{(n-p)\times p}\right\}.\end{aligned}$$

Another Riemannian metric used in [17, 13] is defined by

$$g_Y^2(A,B) := \langle AY^*, BY^*\rangle_{\mathbb{C}^{n\times n}} = \Re(tr((Y^*Y)A^*B)), \quad \forall A,B \in T_Y\mathbb{C}_*^{n\times p} = \mathbb{C}^{n\times p}. \tag{45}$$

Proposition 4.6. *Under metric g^2, the horizontal space at Y satisfies*

$$\begin{aligned}\mathcal{H}_Y^2 &= \left\{Z \in \mathbb{C}^{n\times p} : (Y^*Y)^{-1}Y^*Z = Z^*Y(Y^*Y)^{-1}\right\}\\ &= \left\{YS + Y_\perp K | S^* = S, S \in \mathbb{C}^{p\times p}, K \in \mathbb{C}^{(n-p)\times p}\right\}.\end{aligned}$$

The third Riemannian metric for $\mathbb{C}_*^{n\times p}$ is motivated by the Riemannian metric of $\mathcal{H}_+^{n,p}$ and the diffeomorphism between $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ and $\mathcal{H}_+^{n,p}$. We know that β is a submersion. Every tangent vector of $\mathcal{H}_+^{n,p}$ therefore corresponds to a tangent vector of $\mathbb{C}_*^{n\times p}$. We can use the Riemannian metric of $\mathcal{H}_+^{n,p}$ and the correspondence of tangent vectors between $\mathcal{H}_+^{n,p}$ and $\mathbb{C}_*^{n\times p}$ to define a Riemannian metric for $\mathbb{C}_*^{n\times p}$. A natural first attempt would be to use

$$g_Y(A,B) := \langle \mathrm{D}\,\beta(Y)[A], \mathrm{D}\,\beta(Y)[B]\rangle_{\mathbb{C}^{n\times n}} = \langle YA^* + AY^*, YB^* + BY^*\rangle_{\mathbb{C}^{n\times n}}, \tag{46}$$

which is however not a Riemannian metric because it is not positive-definite. To see this, notice that $\ker(\mathrm{D}\,\beta(Y)[\cdot]) = \mathcal{V}_Y$. Consider $C \neq 0 \in \mathcal{V}_Y$, then $g_Y^3(C,C) = 0$. To modify this definition for g^3, we can use the Riemannian metric g^2 and the decomposition $T_Y\mathbb{C}_*^{n\times p} = \mathcal{H}_Y^2 \oplus \mathcal{V}_Y$, by which $A \in T_Y\mathbb{C}_*^{n\times p}$ can be uniquely decomposed as

$$A = A^{\mathcal{V}} + A^{\mathcal{H}^2}, \tag{47}$$

where $A^{\mathcal{V}} \in \mathcal{V}_Y$ and $A^{\mathcal{H}^2} \in \mathcal{H}_Y^2$. Now define g^3 as

$$\begin{aligned}g_Y^3(A,B) : &= \left\langle \mathrm{D}\,\beta(Y)[A^{\mathcal{H}^2}], \mathrm{D}\,\beta(Y)[B^{\mathcal{H}^2}]\right\rangle + g_Y^2\left(A^{\mathcal{V}}, B^{\mathcal{V}}\right)\\ &= \langle \mathrm{D}\,\beta(Y)[A], \mathrm{D}\,\beta(Y)[B]\rangle_{\mathbb{C}^{n\times n}} + \left\langle P_Y^{\mathcal{V}}(A)Y^*, P_Y^{\mathcal{V}}(B)Y^*\right\rangle_{\mathbb{C}^{n\times n}},\\ &= \langle YA^* + AY^*, YB^* + BY^*\rangle_{\mathbb{C}^{n\times n}} + \left\langle P_Y^{\mathcal{V}}(A)Y^*, P_Y^{\mathcal{V}}(B)Y^*\right\rangle_{\mathbb{C}^{n\times n}}\end{aligned}$$

where $P_Y^{\mathcal{V}}$ is the projection of any tangent vector of $\mathbb{C}_*^{n\times p}$ to the vertical space $\mathcal{V}_Y$. It is straightforward to verify that g^3 defined above is now a Riemannian metric.

Proposition 4.7. *Under metric g^3, the horizontal space at Y is the same set as $\mathcal{H}_Y^2$. That is,*

$$\begin{aligned}\mathcal{H}_Y^3 &= \left\{Z \in \mathbb{C}^{n\times p} : (Y^*Y)^{-1}Y^*Z = Z^*Y(Y^*Y)^{-1}\right\}\\ &= \left\{YS + Y_\perp K | S^* = S, S \in \mathbb{C}^{p\times p}, K \in \mathbb{C}^{(n-p)\times p}\right\}.\end{aligned}$$

4.2 $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ as Riemannian quotient manifold

If the expression $g_Y(\bar{\xi}_Y, \bar{\zeta}_Y)$ does not depend on the choice of $Y \in [Y]$ for every $\pi(Y) \in \mathbb{C}_*^{n\times p}/\mathcal{O}_p$ and every $\xi_{\pi(Y)}, \zeta_{\pi(Y)} \in T_{\pi(Y)}\mathbb{C}_*^{n\times p}/\mathcal{O}_p$, then

$$g_{\pi(Y)}(\xi_{\pi(Y)}, \zeta_{\pi(Y)}) := g_Y(\bar{\xi}_Y, \bar{\zeta}_Y) \tag{48}$$

defines a Riemannian metric on the quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$, see [25, Section 3.6.2]. By Lemma 4.4, it is straightforward to verify that each Riemannian metric g^i $\mathbb{C}_*^{n\times p}$ induces a Riemannian metric on $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$. The quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ endowed with a Riemannian metric defined in (48) is called a *Riemannian quotient manifold*. By abuse of notation, we use g^i for denoting Riemannian metrics on both total space $\mathbb{C}_*^{n\times p}$ and quotient space $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$.

4.3 Projections onto vertical space and horizontal space

Due to the direct sum property (43), for our choices of $\mathcal{H}_Y^i$, there exist projection operators for any $A \in T_Y\mathbb{C}_*^{n\times p}$ to $\mathcal{H}_Y^i$ as

$$A = P_Y^{\mathcal{V}}(A) + P_Y^{\mathcal{H}^i}(A).$$

It is straightforward to verify the following formulae for projection operators $P_Y^{\mathcal{V}}$ and $P_Y^{\mathcal{H}^i}$.

Proposition 4.8. *If we use g^1 as our Riemannian metric on $\mathbb{C}_*^{n\times p}$, then the orthogonal projections of any $A \in \mathbb{C}^{n\times p}$ to $\mathcal{V}_Y$ and $\mathcal{H}_Y^1$ are*

$$P_Y^{\mathcal{V}}(A) = Y\Omega, \quad P_Y^{\mathcal{H}^1}(A) = A - Y\Omega,$$

where Ω is the skew-symmetric matrix that solves the Lyapunov equation

$$\Omega Y^*Y + Y^*Y\Omega = Y^*A - A^*Y. \tag{49}$$

Remark 4.9. *The solution X to the Lyapunov equation $XE + EX = Z$ for a Hermitian E is unique if E is Hermitian positive-definite [1, Section 2.2]. Let $E = U\Lambda U^*$ be the SVD, then the Lyapunov equation $XE + EX = Z$ becomes*

$$(U^*XU)\Lambda + \Lambda(U^*XU) = U^*ZU,$$

which gives the solution

$$(U^*XU)_{i,j} = (U^*ZU)_{i,j}/(\Lambda_{i,i} + \Lambda_{j,j}).$$

Proposition 4.10. *If we use g^2 as our Riemannian metric on $\mathbb{C}_*^{n\times p}$, then the orthogonal projection of any $A \in \mathbb{C}^{n\times p}$ to vertical space $\mathcal{V}_Y$ satisfies*

$$P_Y^{\mathcal{V}}(A) = Y\left(\frac{(Y^*Y)^{-1}Y^*A - A^*Y(Y^*Y)^{-1}}{2}\right) = Y\,Skew\left((Y^*Y)^{-1}Y^*A\right),$$

and the orthogonal projection of any $A \in \mathbb{C}^{n\times p}$ to the horizontal space $\mathcal{H}_Y^2$ is

$$\begin{aligned} P_Y^{\mathcal{H}^2}(A) &= A - P_Y^{\mathcal{V}}(A) \\ &= Y\left(\frac{(Y^*Y)^{-1}Y^*A + A^*Y(Y^*Y)^{-1}}{2}\right) + Y_\perp Y_\perp^* A \\ &= Y\,Herm\left((Y^*Y)^{-1}Y^*A\right) + Y_\perp Y_\perp^* A. \end{aligned}$$

Proposition 4.11. *If we use g^3 as our Riemannian metric on $\mathbb{C}_*^{n\times p}$, then the orthogonal projection of any $A \in \mathbb{C}^{n\times p}$ to vertical space $\mathcal{V}_Y$ satifies*

$$P_Y^{\mathcal{V}}(A) = Y\left(\frac{(Y^*Y)^{-1}Y^*A - A^*Y(Y^*Y)^{-1}}{2}\right) = Y\,skew((Y^*Y)^{-1}Y^*A),$$

and the orthogonal projection of any $A \in \mathbb{C}^{n\times p}$ to the horizontal space $\mathcal{H}_Y^3$ is

$$\begin{aligned} P_Y^{\mathcal{H}^3}(A) &= A - P_Y^{\mathcal{V}}(A) \\ &= Y\left(\frac{(Y^*Y)^{-1}Y^*A + A^*Y(Y^*Y)^{-1}}{2}\right) + Y_\perp Y_\perp^* A \\ &= Y\,Herm\left((Y^*Y)^{-1}Y^*A\right) + Y_\perp Y_\perp^* A. \end{aligned}$$

4.4 Riemannian gradient

Recall that $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ is diffeomorphic to $\mathcal{H}_+^{n,p}$ under $\tilde{\beta}$. Given a smooth real-valued function $f(X)$ on $X \in \mathcal{H}_+^{n,p}$, the corresponding cost function on $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ satisfies

$$\begin{aligned} &h : \mathbb{C}_*^{n\times p}/\mathcal{O}_p \to \mathbb{R} \\ &\qquad \pi(Y) \mapsto f(\tilde{\beta}(\pi(Y))) = f(\beta(Y)) = f(YY^*). \end{aligned} \tag{50}$$

Observe that the function $F(Y) := f(YY^*)$ satisfies $F(Y) = h \circ \pi(Y) = f \circ \beta(Y)$.

The Riemannian gradient of h at $\pi(Y)$ is a tangent vector in $T_{\pi(Y)}\mathbb{C}_*^{n\times p}/\mathcal{O}_p$. The next theorem shows that the horizontal lift of grad $h(\pi(Y))$ can be obtained from the gradient of F defined on $\mathbb{C}_*^{n\times p}$.

Theorem 4.12. *The horizontal lift of the gradient of h at $\pi(Y)$ is the Riemannian gradient of F at Y. That is,*

$$\overline{grad\, h(\pi(Y))}_Y = grad\, F(Y) \tag{51}$$

Proof. See [25, Section 3.6.2]. □

The next proposition summarizes the expression of grad $F(Y)$ under different metrics.

Proposition 4.13. *Let f be a smooth real-valued function defined on $\mathcal{H}_+^{n,p}$ and let $F : \mathbb{C}_*^{n\times p} \to \mathbb{R} : Y \mapsto f(YY^*)$. Assume $YY^* = X$. Then the Riemannian gradient of F is given by*

$$grad\, F(Y) = \begin{cases} (\nabla f(YY^*) + \nabla f(YY^*)^*)Y, & \text{if using metric } g^1 \\ (\nabla f(YY^*) + \nabla f(YY^*)^*)Y(Y^*Y)^{-1}, & \text{if using metric } g^2 \\ \left(I - \frac{1}{2}P_Y\right) \frac{\nabla f(YY^*) + \nabla f(YY^*)^*}{2} Y(Y^*Y)^{-1} & \text{if using metric } g^3 \end{cases} \tag{52}$$

where ∇f denotes Fréchet gradient (4) *and $P_Y = Y(Y^*Y)^{-1}Y^*$.*

Proof. Let $A \in T_Y\mathbb{C}_*^{n\times p}$. By chain rule, we have

$$\mathrm{D}\, F(Y)[A] = \mathrm{D}\, f(YY^*)[YA^* + AY^*]. \tag{53}$$

This yields to

$$g_Y^i(\mathrm{grad}\, F(Y), A) = g_X\left(\mathrm{grad}\, f(YY^*), YA^* + AY^*\right), \tag{54}$$

where g_X is the metric (19). Since $YA^* + AY^* \in T_{YY^*}\mathcal{H}_+^{n,p}$, we have

$$g_X\left(\mathrm{grad}\, f(YY^*), YA^* + AY^*\right) = \left\langle P_{YY^*}^t(\nabla f(YY^*)), YA^* + YA^*\right\rangle_{\mathbb{C}^{n\times n}} = \left\langle \nabla f(YY^*), YA^* + AY^*\right\rangle_{\mathbb{C}^{n\times n}}.$$

It is straightforward to verify that

$$\begin{aligned} \langle \nabla f(YY^*), YA^* + YA^* \rangle_{\mathbb{C}^{n\times n}} &= g_Y^1\left((\nabla f(YY^*) + \nabla f(YY^*)^*)Y, A\right) \\ &= g_Y^2\left((\nabla f(YY^*) + \nabla f(YY^*)^*)Y(Y^*Y)^{-1}, A\right), \end{aligned}$$

which yields the expression of grad $F(Y)$ under g^1 and g^2.

The Riemannian gradient for g^3 is due to

$$\begin{aligned} \left\langle P_{YY^*}^t(\nabla f(YY^*)), YA^* + YA^*\right\rangle_{\mathbb{C}^{n\times n}} &= g_Y^3\left(\left(I - \frac{1}{2}P_Y\right) P_X^t(f')Y(Y^*Y)^{-1}, A\right) \\ &= g_Y^3\left(\left(I - \frac{1}{2}P_Y\right) \frac{f' + f'^*}{2} Y(Y^*Y)^{-1}, A\right). \end{aligned}$$

□

4.5 Retraction

The retraction on the quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ can be defined using the retraction on the total space $\mathbb{C}_*^{n\times p}$. For any $A \in T_Y\mathbb{C}_*^{n\times p}$ and a step size $\tau > 0$,

$$\overline{R}_Y(\tau A) := Y + \tau A, \tag{55}$$

is a retraction on $\mathbb{C}_*^{n\times p}$ if $Y + \tau A$ remains full rank, which is ensured for small enough τ. Then Lemma 4.4 indicates that $\overline{R}$ satisfies the conditions of [25, Prop. 4.1.3], which implies that

$$R_{\pi(Y)}(\tau\eta_{\pi(Y)}) := \pi(\overline{R}_Y(\tau\overline{\eta}_Y)) = \pi(Y + \tau\overline{\eta}_Y) \tag{56}$$

defines a retraction on the quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ for a small enough step size $\tau > 0$. With Lemma 4.4, it is straightforward to verify that the retraction (56) does not depend on the choice of Y for the same equivalent class.

4.6 Vector transport

A vector transport on $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ introduced in [25, Section 8.1.4] is projection to horizontal space.

$$\overline{\left(\mathcal{T}_{\eta_{\pi(Y)}}\xi_{\pi(Y)}\right)}_{Y+\overline{\eta}_Y} := P^{\mathcal{H}}_{Y+\overline{\eta}_Y}(\overline{\xi}_Y). \tag{57}$$

It can be shown that this vector transport is actually the differential of the retraction R defined in (56) (see [25, Section 8.1.2]) since

$$\begin{aligned}
\mathrm{D}\,R_{\pi(Y)}(\eta_{\pi(Y)})[\xi_{\pi(Y)}] &= \mathrm{D}\,\pi\left(\overline{R}_Y(\overline{\eta}_Y)\right)\left[\mathrm{D}\,\overline{R}_Y(\overline{\eta}_Y)[\overline{\xi}_Y]\right] \\
&= \mathrm{D}\,\pi(Y+\overline{\eta}_Y)\left[\left.\frac{d}{dt}\right|_{t=0}(Y+\overline{\eta}_Y+t\overline{\xi}_Y)\right] \\
&= \mathrm{D}\,\pi(Y+\overline{\eta}_Y)\left[\overline{\xi}_Y\right] \\
&= \mathrm{D}\,\pi(Y+\overline{\eta}_Y)\left[P^{\mathcal{H}}_{Y+\overline{\eta}_Y}(\overline{\xi}_Y)\right].
\end{aligned}$$

Based on the projection formulae in Section 4.3, we can obtain formulae of vector transports using different Riemannian metrics. Denote $Y_2 = Y_1 + \overline{\eta}_{Y_1}$. If we use metric g^1, then

$$\overline{\left(\mathcal{T}_{\eta_{\pi(Y_1)}}\xi_{\pi(Y_1)}\right)}_{Y_1+\overline{\eta}_{Y_1}} = \overline{\xi}_{Y_1} - Y_2\Omega, \tag{58}$$

where Ω solves the Lyapunov equation

$$Y_2^*Y_2\Omega + \Omega Y_2^*Y_2 = Y_2^*\overline{\xi}_{Y_1} - \overline{\xi}_{Y_1}^*Y_2.$$

See Remark 4.9 for the expression of Ω.

If we use metric g^2 or g^3, then

$$\begin{aligned}
\overline{\left(\mathcal{T}_{\eta_{\pi(Y_1)}}\xi_{\pi(Y_1)}\right)}_{Y_1+\overline{\eta}_{Y_1}} &= \overline{\xi}_{Y_1} - P^{\mathcal{V}}_{Y_2}(\overline{\xi}_{Y_1}) \\
&= \overline{\xi}_{Y_1} - Skew\left((Y_2^*Y_2)^{-1}Y_2^*\overline{\xi}_{Y_1}\right) \\
&= Y_2\left(\frac{(Y_2^*Y_2)^{-1}Y_2^*\overline{\xi}_{Y_1} + \overline{\xi}_{Y_1}^*Y_2(Y_2^*Y_2)^{-1}}{2}\right) + Y_{2\perp}Y_{2\perp}^*\overline{\xi}_{Y_1}.
\end{aligned}$$

4.7 Riemannian Hessian operator

Recall that the cost function h on $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ is defined in (50). In this section, we summarize the Riemannian Hessian of h under the three different metrics g^i. The proofs are tedious calculations and given in Appendix C.1.

Proposition 4.14. *Using g^1, the Riemannian Hession of h is given by*

$$\overline{(\mathit{Hess}\,h(\pi(Y))[\xi_{\pi(Y)}])}_Y = P_Y^{\mathcal{H}^1}\left(2Herm\{\nabla^2 f(YY^*)[Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*]\}Y + 2Herm(\nabla f(YY^*))\overline{\xi}_Y\right). \tag{59}$$

Proposition 4.15. *Using g^2, the Riemannian Hession of h is given by*

$$\begin{aligned}
\overline{(\mathit{Hess}\,h(\pi(Y))[\xi_{\pi(Y)}])}_Y = \; & P_Y^{\mathcal{H}^2}\Big\{2Herm\{\nabla^2 f(YY^*)[Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*]\}Y(Y^*Y)^{-1} \\
&+Herm(\nabla f(YY^*))P_Y^{\perp}\overline{\xi}_Y(Y^*Y)^{-1} + P_Y^{\perp}Herm(\nabla f(YY^*))\overline{\xi}_Y(Y^*Y)^{-1} \\
&+2skew(\overline{\xi}_Y Y^*)Herm(\nabla f(YY^*))Y(Y^*Y)^{-2} \\
&+ 2skew\{\overline{\xi}_Y(Y^*Y)^{-1}Y^*Herm(\nabla f(YY^*))\}Y(Y^*Y)^{-1}\Big\}.
\end{aligned}$$

Proposition 4.16. *Using g^3, the Riemannian Hession of h is given by*

$$\begin{aligned}
\overline{(\mathit{Hess}\,h(\pi(Y))[\xi_{\pi(Y)}])}_Y &= P_Y^{\mathcal{H}^3}(\mathit{Hess}\,F(Y)[\overline{\xi}_Y]) \\
&= \left(I - \frac{1}{2}P_Y\right)Herm\{\nabla^2 f(YY^*)[Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*]\}Y(Y^*Y)^{-1} \\
&\quad +(I-P_Y)Herm(\nabla f(YY^*))(I-P_Y)\overline{\xi}_Y(Y^*Y)^{-1}.
\end{aligned}$$

5 The Riemannian conjugate gradient method

For simplicity, in this paper we only consider the Riemannian conjugate gradient (RCG) method described as Algorithm 1 in [10] with the geometric variant of Polak–Ribiére (PR+) for computing the conjugate direction. It is possible to explore other methods such as the limited-memory version of the Riemannian BFGS method (LRBFGS) as in [30]. However, RCG performs very well on a wide variety of problems.

In this section, we focus on establishing two equivalences. First, we show that the Burer–Monteiro CG method, which is simply applying CG method for the unconstrainted problem (5), is equivalent to RCG on the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^1)$ with our retraction and vector transport. Second, we show that RCG on the embedded manifold $\mathcal{H}_+^{n,p}$ is equivalent to RCG on the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^3)$ with a specific retraction and vector transport.

For convenience, let $\mathcal{T}_{X_{k-1}\to X_k}$ denote a vector transport that maps tangent vectors from $T_{X_{k-1}}\mathcal{H}_+^{n,p}$ to $T_{X_k}\mathcal{H}_+^{n,p}$, defined as

$$\mathcal{T}_{X_{k-1}\to X_k} : T_{X_{k-1}}\mathcal{H}_+^{n,p} \to T_{X_k}\mathcal{H}_+^{n,p}, \quad \zeta_{X_{k-1}} \mapsto \mathcal{T}_{R^{-1}_{X_{k-1}}(X_k)}(\zeta_{X_{k-1}}), \tag{60}$$

where R_X^{-1} exists locally for every $X \in \mathcal{H}_+^{n,p}$ by the inverse function theorem. Hence $\mathcal{T}_{X_{k-1}\to X_k}$ should be understood locally in the sense that X_k is sufficiently close to X_k. See [28, Section 2.4].

Similarly, Let $\mathcal{T}_{Y_{k-1}\to Y_k}$ denote a vector transport that maps tangent vectors from $\mathcal{H}_{Y_{k-1}}$ to $\mathcal{H}_{Y_k}$ as

$$\mathcal{T}_{Y_{k-1}\to Y_k} : \mathcal{H}_{Y_{k-1}} \to \mathcal{H}_{Y_k}, \quad \overline{\xi}_{Y_{k-1}} \mapsto \overline{\left(\mathcal{T}_{R^{-1}_{\pi(Y_{k-1})}\xi_{\pi(Y_k)}}\right)_{Y_k}}, \tag{61}$$

where $R^{-1}_{\pi(Y)}$ also exists locally for every $\pi(Y) \in \mathbb{C}_*^{n\times p}/\mathcal{O}_p$. $\mathcal{T}_{Y_{k-1}\to Y_k}$ and should again be understood locally in the sense that $\pi(Y_{k-1})$ is sufficiently close to $\pi(Y_k)$.

We first summarize two Riemannian CG algorithm in Algorithm 1 and Algorithm 2 below. Algorithm 1 is the RCG on the embedded manifold for solving 1 and Algorithm 2 is the RCG on the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^i)$ for solving (7). We remark that the explicit constants 0.0001 and 0.5 in the Armijo backtracking are chosen for convenience.

Algorithm 1 Riemannian Conjugate Gradient on the embedded manifold $\mathcal{H}_+^{n,p}$

Require: initial iterate $X_1 \in \mathcal{H}_+^{n,p}$, tolerance $\varepsilon > 0$, tangent vector $\eta_0 = 0$

1: **for** $k = 1, 2, \dots$ **do**
2: Compute gradient
$\xi_k := \operatorname{grad} f(X_k)$ ▷ See Algorithm 3
3: Check convergence
if $\|\xi_k\| := \sqrt{g_{X_k}(\xi_k, \xi_k)} < \varepsilon$, then break
4: Compute a conjugate direction by PR_+ and vector transport
$\eta_k == -\xi_k + \beta_k \mathcal{T}_{X_{k-1}\to X_k}(\eta_{k-1})$ ▷ See Algorithm **??**

$$\beta_k = \frac{g_{X_k}\left(\xi_k, \xi_k - \mathcal{T}_{X_{k-1}\to X_k}(\xi_{k-1})\right)}{g_{X_{k-1}}\left(\xi_{k-1}, \xi_{k-1}\right)}.$$

5: Compute an initial step t_k. For special cost functions, it is possible to compute:
$t_k = \arg\min_t f(X_k + t\eta_k)$ ▷ See Algorithm 8
6: Perform Armijo backtracking to find the smallest integer $m \geq 0$ such that

$$f(X_k) - f(R_{X_k}(0.5^m t_k \eta_k)) \geq -0.0001 \times 0.5^m t_k g_{X_k}(\xi_k, \eta_k) \tag{62}$$

7: Obtain the new iterate by retraction
$X_{k+1} = R_{X_k}(0.5^m t_k \eta_k)$ ▷ See Algorithm 5
8: **end for**

5.1 Equivalence between Burer–Monteiro CG and RCG on the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^1)$

Theorem 5.1. *Using retraction (56), vector transport (57) and metric g^1, Algorithm 2 is equivalent to the conjugate gradient method solving* (5) *in the sense that they produce exactly the same iterates if started from the same initial point.*

Algorithm 2 Riemannian Conjugate Gradient on the quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ with metric g^i

Require: initial iterate $Y_1 \in \pi^{-1}(\pi(Y_1))$, tolerance $\varepsilon > 0$, tangent vector $\eta_0 = 0$

for $k = 1, 2, \dots$ **do**

  Compute the horizontal lift of gradient
    $\xi_k := \overline{(\operatorname{grad} h(\pi(Y_k)))}_{Y_k} = \operatorname{grad} F(Y_k)$ ▷ See Algorithm 6

  Check convergence
    if $\|\xi_k\| := \sqrt{g^i_{Y_k}(\xi_k, \xi_k)} < \varepsilon$, then break

  Compute a conjugate direction by PR$_+$ and vector transport
    $\eta_k = -\xi_k + \beta_k \mathcal{T}_{Y_{k-1}\to Y_k}(\eta_{k-1})$ ▷ See Algorithm 7

$$\beta_k = \frac{g^i_{Y_k}\left(\operatorname{grad} F(Y_k), \operatorname{grad} F(Y_k) - \mathcal{T}_{Y_{k-1}\to Y_k}(\xi_{k-1})\right)}{g^i_{Y_{k-1}}\left(\operatorname{grad} F(Y_{k-1}), \operatorname{grad} F(Y_{k-1})\right)}.$$

  Compute an initial step t_k. For special cost functions, it is possible to compute:
    $t_k = \arg\min_t F(Y_k + t\eta_k)$ ▷ See Algorithm 9

  Perform Armijo backtracking to find the smallest integer $m \geq 0$ such that

$$F(Y_k) - F(\overline{R}_{Y_k}(0.5^m t_k \eta_k)) \geq -0.0001 \times 0.5^m t_k g^i_{Y_k}(\xi_k, \eta_k)$$

  Obtain the new iterate by the simple retraction
    $Y_{k+1} = \overline{R}_{Y_k}(0.5^m t_k \eta_k) = Y_k + 0.5^m t_k \eta_k$

end for

Proof. First of all, for g^1, $\operatorname{grad} F(Y) = (\nabla f(YY^*) + \nabla f(YY^*)^*)\, Y$, is equal to the Fréchet gradient of $F(Y) := f(YY^*)$ at Y. Since vector transport is the orthogonal projection to the horizontal space, the PR$_+$ β_k used in Riemannian CG becomes

$$\beta_k = \frac{g^1_{Y_k}\left(\operatorname{grad} F(Y_k), \operatorname{grad} F(Y_k) - P^{\mathcal{H}^1}_{Y_k}(\operatorname{grad} F(Y_{k-1}))\right)}{g^1_{Y_{k-1}}\left(\operatorname{grad} F(Y_{k-1}), \operatorname{grad} F(Y_{k-1})\right)}. \tag{63}$$

Now observe that

$$P^{\mathcal{H}^1}_{Y_k}(\operatorname{grad} F(Y_{k-1})) = \operatorname{grad} F(Y_{k-1}) - P^{\mathcal{V}}_{Y_k}(\operatorname{grad} F(Y_{k-1})) \tag{64}$$

and g^1 is equivalent to the classical inner product for $\mathbb{C}^{n\times p}$. Hence β_k computed by (63) is equal to PR$_+$ β_k in conjugate gradient for (5).

The first conjugate direction is $\eta_1 = -\operatorname{grad} F(Y_1) = -\nabla F(Y_1)$, so Burer–Monteiro CG coincides with Riemannian CG for the first iteration. It remains to show that η_k generated in Riemannian CG by

$$\eta_k = -\xi_k + \beta_k P^{\mathcal{H}^1}_{Y_k}(\eta_{k-1})$$

is equal to η_k generated in Burer–Monteiro CG for each $k \geq 2$. It suffices to show that

$$P^{\mathcal{H}^1}_{Y_k}(\eta_{k-1}) = \eta_{k-1}, \quad \forall k \geq 2.$$

Equivalently we need to show that for all $k \geq 2$, the Lyapunov equation

$$(Y_k^* Y_k)\Omega + \Omega(Y_k^* Y_k) = Y_k^* \eta_{k-1} - \eta_{k-1}^* Y_k \tag{65}$$

only has trivial solution $\Omega = 0$. By invertibility of the equation, this means that we only need to show the right hand side is zero. We prove it by induction.

For $k = 2$, $\eta_{k-1} = \eta_1 = -\xi_1 = -\operatorname{grad} F(Y_1)$. The following computations show that the RHS of (65) satisfies

$$\begin{aligned} Y_2^*\eta_1 - \eta_1^* Y_2 &= -Y_2^*\xi_1 + \xi_1^* Y_2 \\ &= -(Y_1 - c\xi_1)^*\xi_1 + \xi_1^*(Y_1 - c\xi_1) \\ &= \xi_1^* Y_1 - Y_1^*\xi_1 \\ &= Y_1^*[\nabla f(Y_1Y_1^*) + \nabla f(Y_1Y_1^*)^*]Y_1 - Y_1^*[\nabla f(Y_1Y_1^*) + \nabla f(Y_1Y_1^*)^*]Y_1 \\ &= 0. \end{aligned}$$

Hence $\Omega = 0$ and $P^{\mathcal{H}^1}_{Y_k}(\eta_{k-1}) = \eta_{k-1}$ for $k = 2$.

Now suppose for $k \geq 2$, the RHS of (65) is 0 and hence $P^{\mathcal{H}^1}_{Y_k}(\eta_{k-1}) = \eta_{k-1}$ holds. Then the RHS of the Lyapunov equation of step $k+1$ is

$$\begin{aligned} Y_{k+1}^*\eta_k - \eta_k^* Y_{k+1} &= (Y_k + c\eta_k)^*\eta_k - \eta_k^*(Y_k + c\eta_k) \\ &= Y_k^*\eta_k - \eta_k^* Y_k \\ &= Y_k^*\left(-\xi_k + \beta_k P^{\mathcal{H}^1}_{Y_k}(\eta_{k-1})\right) - \left(-\xi_k + \beta_k P^{\mathcal{H}^1}_{Y_k}(\eta_{k-1})\right)^* Y_k \\ &= Y_k^*(-\xi_k + \beta_k\eta_{k-1}) - (-\xi_k + \beta_k\eta_{k-1})^* Y_k \\ &= -Y_k^*\xi_k + \xi_k^* Y_k \\ &= -Y_k^*[\nabla f(Y_kY_k^*) + \nabla f(Y_kY_k^*)^*]Y_k + Y_k^*[\nabla f(Y_kY_k^*) + \nabla f(Y_kY_k^*)^*]Y_k \\ &= 0. \end{aligned}$$

Hence $P^{\mathcal{H}^1}_{Y_{k+1}}(\eta_k) = \eta_k$ also holds. We have thus proven that Riemannian CG is equivalent to Burer–Monteiro CG. ☐

Since the gradient descent corresponds to $\beta_k \equiv 0$, the same discussion also implies the following

Theorem 5.2. *Using retraction (56) and metric g^1, the Riemannian gradient descent on the quotient manifold is equivalent to the Burer–Monteiro gradient descent method with suitable step size* (6) *in the sense that they produce exactly the same iterates.*

5.2 Equivalence between RCG on embedded manifold and RCG on the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^3)$

In this subsection we show that Algorithm 1 is equivalent to Algorithm 2 with Riemannian metric g^3, a specific initial line-search in step 5, a specific retraction and a specific vector transport. The idea is to take the advantage of the diffeomorphism $\tilde{\beta}$ between $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ and $\mathcal{H}_+^{n,p}$, as well as the fact that the metric g^3 of $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ is induced from the metric of $\mathcal{H}_+^{n,p}$.

The Lemma below shows that there is a one-to-one correspondence between grad f and grad h.

Lemma 5.3. *If we use g^3 as the Riemannian metric for $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$, then the Riemannian gradient of f and h is related by the diffeomorphism $\tilde{\beta}$ in the following way:*

$$D\,\tilde{\beta}(\pi(Y))[\mathit{grad}\, h(\pi(Y))] = \mathit{grad}\, f(YY^*). \tag{66}$$

Proof. Recall that $\beta = \tilde{\beta}\circ\pi$ and we have Theorem 4.12. By chain rule and the definition of horizontal lift we have

$$\begin{aligned} LHS = \mathrm{D}\,\tilde{\beta}(\pi(Y))[\mathrm{grad}\, h(\pi(Y))] &= \mathrm{D}\,\tilde{\beta}(\pi(Y))\left[\mathrm{D}\,\pi(Y)\left[\overline{\mathrm{grad}\, h(\pi(Y))}_Y\right]\right] \\ &= \mathrm{D}\,\beta(Y)\left[\overline{\mathrm{grad}\, h(\pi(Y))}_Y\right] \\ &= \mathrm{D}\,\beta(Y)\left[\mathrm{grad}\, F(Y)\right]. \end{aligned}$$

Now recall that $F = f\circ\beta$. Let $A \in \mathbb{C}^{n\times p}$ then

$$\mathrm{D}\, F(Y)[A] = \mathrm{D}\, f(YY^*)[YA^* + YA^*].$$

Let $X = YY^*$. Then we have

$$g_Y^3(\mathrm{grad}\, F(Y), A) = g_X(\mathrm{grad}\, f(YY^*), YA^* + AY^*).$$

Since $\mathrm{grad}\, F(Y) \in \mathcal{H}_Y^3$, we have

$$g_X\left(\mathrm{D}\,\beta(Y)[\mathrm{grad}\, F(Y)], YA^* + AY^*\right) = g_X\left(\mathrm{grad}\, f(YY^*), YA^* + AY^*\right),$$

or

$$g_X\left(LHS, YA^* + AY^*\right) = g_X\left(RHS, YA^* + AY^*\right).$$

Now notice that A is arbitrary and $YA^* + AY^*$ can be any tangent vector in $T_X\mathcal{H}_+^{n,p}$. Hence we must have $LHS = RHS$ ☐

Remark 5.4. *Since $\tilde{\beta}$ is a diffeomorphism bewteen $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ and $\mathcal{H}_+^{n,p}$, $D\tilde{\beta}(\pi(Y))[\cdot]$ defines an isomorphism between the two tangent space $T_{\pi(Y)}\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ and $T_{YY^*}\mathcal{H}_+^{n,p}$. We denote this isomorphism by $L_{\pi(Y)}$. When the tangent space is clear from the context, $\pi(Y)$ is omitted and we only use the notation L for simplicity. The previous lemma then simply reads*

$$L_{\pi(Y)}(\mathit{grad}\, h(\pi(Y))) = \mathit{grad}\, f(\tilde{\beta}(\pi(Y))). \tag{67}$$

In Algorithm 1, we have a retraction R^E and a vector transport $\mathcal{T}^E$ on the embedded manifold $\mathcal{H}_+^{n,p}$, with the superscript E for *Embedded*, such that R^E is the retraction associated with $\mathcal{T}^E$. Then we claim that there are a retraction R^Q and a vector transport $\mathcal{T}^Q$, with the superscript Q denoting *Quotient*, on the quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ with metric g^3, such that Algorithm 2 is equivalent to Algorithm (1). The idea is again to use the diffeomorphism $\tilde{\beta}$ and the isomorphism $L_{\pi(Y)}$. We give the desired construction of R^Q and $\mathcal{T}^Q$ as follows.

$$R^Q_{\pi(Y)}(\xi_{\pi(Y)}) := \tilde{\beta}^{-1}\left(R^E_{\tilde{\beta}(\pi(Y))}\left(L(\xi_{\pi(Y)})\right)\right). \tag{68}$$

$$\mathcal{T}^Q_{\eta_{\pi(Y)}}(\xi_{\pi(Y)}) := L^{-1}_{\pi(Y_2)}\left(\mathcal{T}^E_{L(\eta_{\pi(Y)})}\left(L(\xi_{\pi(Y)})\right)\right), \tag{69}$$

where $\tilde{\beta}(\pi(Y_2))$ denotes the foot of the tangent vector $\mathcal{T}^E_{L(\eta_{\pi(Y)})}\left(L(\xi_{\pi(Y)})\right)$.

Next we need to show that R^Q defined in (68) and $\mathcal{T}^Q$ defined in (69) are indeed a retraction and a vector transport, respectively.

Theorem 5.5. *R^Q defined in (68) is a retraction.*

Proof. First it is easy to see that $R^Q_{\pi(Y)}(0_{\pi(Y)}) = \pi(Y)$. Then we also have for all $v_{\pi(Y)} \in T_{\pi(Y)}\mathbb{C}_*^{n\times p}/\mathcal{O}_p$

$$\begin{aligned}
\mathrm{D}\, R^Q_{\pi(Y)}(0_{\pi(Y)})[v_{\pi(Y)}] &= \mathrm{D}\,\tilde{\beta}^{-1}(\tilde{\beta}(\pi(Y))\left[\mathrm{D}\, R^E_{\tilde{\beta}(\pi(Y))}(0)\left[\mathrm{D}\, L(0)\left[v_{\pi(Y)}\right]\right]\right] \\
&= \mathrm{D}\,\tilde{\beta}^{-1}(\tilde{\beta}(\pi(Y))\left[\mathrm{D}\, R^E_{\tilde{\beta}(\pi(Y))}(0)\left[L(v_{\pi(Y)})\right]\right] \\
&= \mathrm{D}\,\tilde{\beta}^{-1}(\tilde{\beta}(\pi(Y))\left[L(v_{\pi(Y)})\right] \\
&= \left(\mathrm{D}\,\tilde{\beta}(\pi(Y))\right)^{-1}\left[L(v_{\pi(Y)})\right] \\
&= L^{-1}(L(v_{\pi(Y)})) \\
&= v_{\pi(Y)}
\end{aligned}$$

Hence $\mathrm{D}\, R^Q_{\pi(Y)}(0_{\pi(Y)})[\cdot]$ is an identity map. □

Theorem 5.6. *$\mathcal{T}^E$ defined in (69) is a vector transport and R^Q is the retraction associated with $\mathcal{T}^E$.*

Proof. Consistency and linearity are straightforward. It thus suffices to verify that the foot of $\mathcal{T}^Q_{\eta_{\pi(Y)}}(\xi_{\pi(Y)})$ is equal to $R^Q_{\pi(Y)}(\eta_{\pi(Y)})$. Since R^E is the associated retraction with $\mathcal{T}^E$, the foot of $\mathcal{T}^E_{L(\eta_{\pi(Y)})}(L(\xi_{\pi(Y)}))$ is equal to $R^E_{\tilde{\beta}(\pi(Y))}\left(L(\eta_{\pi(Y)})\right)$, which we denote by $\tilde{\beta}(\pi(Y_2))$ for some $\pi(Y_2)$. Hence $R^Q_{\pi(Y)}(\eta_{\pi(Y)}) = \tilde{\beta}^{-1}\left(R^E_{\tilde{\beta}(\pi(Y))}\left(L(\eta_{\pi(Y)})\right)\right) = \pi(Y_2)$.

Furthermore, we have that $\mathcal{T}^Q_{\eta_{\pi(Y)}}(\xi_{\pi(Y)}) = L^{-1}_{\pi(Y_2)}\left(\mathcal{T}^E_{L(\eta_{\pi(Y)})}\left(L(\xi_{\pi(Y)})\right)\right)$ is a tangent vector in $T_{\pi(Y_2)}\mathbb{C}_*^{n\times p}/\mathcal{O}_p$. Hence, the foot of $\mathcal{T}^Q_{\eta_{\pi(Y)}}(\xi_{\pi(Y)})$ is also $\pi(Y_2)$. □

Finally, in order to reach an equivalence, we also need the initial step size to match the one in step 5 of Algorithm 2. We simply replace the original initial step size t_k by

$$t_k = \arg\min_t f(Y_k Y_k^* + t(Y_k\eta_k^* + \eta_k Y_k^*)). \tag{70}$$

This value of t_k now is equivalent to the initial step size in step 5 of Algorithm 1. This gives us the following result:

Theorem 5.7. *With the newly constructed initial step size, retraction and vector transport in this subsection, Algorithm 2 for solving* (7) *is equivalent to Algorithm 1 solving* (1) *in the sense that they produce exactly the same iterates.*

6 Implementation details

The algorithms in this paper can be applied for minimizing any smooth function $f(X)$ in (1). For problems with large n, however, it is advisable to avoid constructing and storing the Fréchet derivative $\nabla f(X) \in \mathbb{C}^{n\times n}$ explicitly. Instead, one directly computes the matrix-vector multiplications $\nabla f(X)U$ and $(\nabla f(X))^*U$. In the PhaseLift problem [5], for example, these matrix-vector multiplications can be implemented via the FFT at a cost of $\mathcal{O}(pn\log n)$ when $U \in \mathbb{C}^{n\times p}$; see [13].

Below, we detail the calculations needed in Algorithms 1 and 2. When giving flop counts, we assume that $\nabla f(X)U$ and $(\nabla f(X))^*U \in \mathbb{C}^{n\times p}$ can be computed in $spn\log n$ flops with s small. For g^2 and g^3 in Algorithms 6 and 7, we use forwardslash "/" and backslash "\" in Matlab command to compute the inverse of Y^*Y.

6.1 Embedded manifold

Algorithm 3 Calculate the Riemannian gradient grad $f(X)$

Require: $X = U\Sigma U^* \in \mathcal{H}_+^{n,p}$
Ensure: grad $f(X) = UHU^* + U_pU^* + UU_p^* \in T_X\mathcal{H}_+^{n,p}$

$T \leftarrow \frac{1}{2}(\nabla f(X) + \nabla f(X)^*)U$	▷ # $2spn\log n$ flops
$H \leftarrow U^*T$	▷ # $2np^2$ flops
$U_p \leftarrow T - UH$	▷ # $2np^2 + np$ flops

Algorithm 4 Calculate the embedded vector transport $P_{X_2}^t(\nu)$

Require: $X_1 = U_1\Sigma_1U_1^*$, $X_2 = U_2\Sigma_2U_2^*$ and tangent vector $\nu = U_1H_1U_1^* + U_{p_1}U_1^* + U_1U_{p_1}^* \in T_{X_1}\mathcal{H}_+^{n,p}$.
Ensure: $P_{X_2}^t(\nu) = U_2H_2U_2^* + U_{p_2}U_2^* + U_2U_{p_2}^*$

$A \leftarrow U_1^*U_2$	▷ # $2np^2$ flops
$H_2^{(1)} \leftarrow A^*H_1A, \quad U_p^{(1)} \leftarrow U_1(H_1A)$	▷ # $6p^3 + 2np^2$ flops
$H_2^{(2)} \leftarrow U_2^*U_{p_1}A, \quad U_p^{(2)} \leftarrow U_{p_1}A$	▷ # $4np^2 + 2p^3$ flops
$H_2^{(3)} \leftarrow {H_2^{(2)}}^*, \quad U_p^{(3)} \leftarrow U_1({U_1}_p^*U_2)$	▷ # $4np^2$ flops
$H_2 \leftarrow H_2^{(1)} + H_2^{(2)} + H_2^{(3)}$	▷ # $4np^2$ flops
$U_{p_2} \leftarrow U_p^{(1)} + U_p^{(2)} + U_p^{(3)}, \quad U_{p_2} \leftarrow U_{p_2} - U_2(U_2^*U_{p_2})$	▷ # $4np^2$ flops

Algorithm 5 Calculate the retraction $P_{\mathcal{H}_+^{n,p}}(X+Z)$

Require: $X = U\Sigma U^* \in \mathcal{H}_+^{n,p}$, tangent vector $Z = UHU^* + U_pU^* + UU_p^*$.
Ensure: $P_{\mathcal{H}_+^{n,p}}(X+Z) = U_+\Sigma_+U_+^*$.

$(Q,R) \leftarrow \mathrm{qr}(U_p,0) \quad M \leftarrow \begin{bmatrix}\Sigma + H & R^* \\ R & 0\end{bmatrix}$	▷ # $10np^2$ flops
$[V,S] \leftarrow \mathrm{eig}(M)$	▷ $O(p^3)$ flops
$\Sigma+ \leftarrow S(1:p,1:p), \quad U_+ \leftarrow [U \quad Q]\,V(:,1:p)$	▷ # $4np^2$ flops

6.2 Quotient manifold

Algorithm 6 Calculate the Riemannian gradient grad $F(Y)$

Require: $Y \in \mathbb{C}_*^{n\times p}$
Ensure: $T = \operatorname{grad} F(Y)$
1: **if** metric is g^1 **then**
  $T \leftarrow (\nabla f(YY^*) + \nabla f(YY^*)^*)Y.$ ▷ # $2spn\log n$ flops
2: **else if** metric is g^2 **then**
  $Z \leftarrow Y(Y^*Y)^{-1}$ ▷ # $4np^2 + O(p^3)$ flops
  $T \leftarrow (\nabla f(YY^*) + \nabla f(YY^*)^*)Z$ ▷ # $2spn\log n$ flops
3: **else if** metric is g^3 **then**
  $Z \leftarrow Y(Y^*Y)^{-1}$ ▷ # $4np^2 + O(p^3)$ flops
  $T \leftarrow (\nabla f(YY^*) + \nabla f(YY^*)^*)Z$ ▷ # $2spn\log n$ flops
  $M \leftarrow Y^*T, \quad T \leftarrow T - \frac{1}{2}ZM$ ▷ # $4np^2 + np + p^3$ flops
4: **end if**

Algorithm 7 Calculate the quotient vector transport $P_{Y_2}^{\mathcal{H}}(h_1)$

Require: $Y_1 \in \mathbb{C}_*^{n\times p}$, $Y_2 \in \mathbb{C}_*^{n\times p}$ and horizontal vector $h_1 \in \mathcal{H}_{Y_1}$.
Ensure: $h_2 = P_{Y_2}^{\mathcal{H}}(h_1) \in \mathcal{H}_{Y_2}$.
1: **if** metric is g^1 **then**
  $E \leftarrow Y_2^*Y_2$ ▷ # $2np^2$ flops
  $(Q, S) \leftarrow \operatorname{eig}(E), \quad d \leftarrow \operatorname{diag}(S)$ ▷ # $O(p^3)$ flops
  $\lambda \leftarrow d\,[1, 1, \cdots, 1] + [1, 1, \cdots, 1]^T\, d^T$ ▷ # $2p^2$ flops
  $A \leftarrow Q^*(Y_2^*h_1 - h_1^*Y_2)Q$ ▷ # $4np^2 + 4p^3$ flops
  $\Omega \leftarrow Q(A./\lambda)Q^*$ ▷ # $p^2 + 4p^3$ flops
  $h_2 \leftarrow h_1 - Y_2\Omega$ ▷ # $np + 2np^2$ flops
2: **else if** metric is g^2 or g^3 **then**
  $\tilde{\Omega} \leftarrow (Y^*Y)^{-1}(Y_2^*h_1)$ ▷ # $4np^2 + O(p^3)$ flops
  $\Omega \leftarrow \frac{1}{2}(\tilde{\Omega} - \tilde{\Omega}^*)$ ▷ # $2p^2$ flops
  $h_2 \leftarrow h_1 - Y_2\Omega$ ▷ # $np + 2np^2$ flops
3: **end if**

6.3 Initial guess for the line search

The initial guess for the line search generally depends on the expression of the cost function $f(X)$. For the important case of $f(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_F^2$ where $\mathcal{A}$ is a linear operator and b is a matrix, the initial guess for embedded CG requires solving a linear equation and for quotient CG it requires solving a cubic equation. Below this calculation is detailed for b of size mn for some m and assuming that $\mathcal{A}(X)$ can be evaluated in $sp^\alpha n\log n$ flops for $X \in \mathcal{H}_+^{n,p}$, $\mathcal{A}(T)$ for $T \in T_X\mathcal{H}_+^{n,p}$ and $\mathcal{A}(Y\eta^*)$ for $Y, \eta \in \mathbb{C}_*^{n\times p}$.

Algorithm 8 Calculate the initial guess $t_* = \arg\min_t f(X + tT)$

Require: $X \in \mathcal{H}_+^{n,p}$ and a descend direction $T \in T_X\mathcal{H}_+^{n,p}$
Ensure: $t_* = \arg\min_t f(X + tT) = \arg\min_t \frac{1}{2}\|\mathcal{A}(X + tT) - b\|_F^2$
  $R \leftarrow \mathcal{A}(X) - b$ ▷ # $sp^\alpha n\log n + mn$ flops
  $S \leftarrow \mathcal{A}(T)$ ▷ # $sp^\alpha n\log n$ flops
  $t_* \leftarrow -\frac{\langle R,S\rangle}{\langle S,S\rangle}$ ▷ # $2mn + 1$ flops

---

Algorithm 9 Calculate the initial guess $t_* = \arg\min_t F(Y + t\eta)$

Require: $Y \in \mathbb{C}_*^{n\times p}$, a descend direction $\eta \in \mathcal{H}_Y$,
Ensure: $t_* = \arg\min_t F(Y + t\eta) = \arg\min_t \frac{1}{2}\|\mathcal{A}((Y+t\eta)(Y+t\eta)^*) - b\|_F^2$
$c_0 \leftarrow \mathcal{A}(YY^*) - b$ ▷ # $sp^\alpha n \log n + mn$ flops
$c_1^{(1)} \leftarrow \mathcal{A}(Y\eta^*), \quad c_1^{(2)} \leftarrow \mathcal{A}(\eta Y^*), \quad c_1 \leftarrow c_1^{(1)} + c_1^{(2)}$ ▷ # $2sp^\alpha n \log n + mn$ flops
$c_2 \leftarrow \mathcal{A}(\eta\eta^*)$ ▷ # $sp^\alpha n \log n$ flops
$d_4 \leftarrow \langle c_2, c_2\rangle, \quad d_3 \leftarrow 2\langle c_2, c_1\rangle$ ▷ # $2mn$ flops
$d_2 \leftarrow 2\langle c_2, c_0\rangle + \langle c_1, c_1\rangle, \quad d_1 \leftarrow 2\langle c_1, c_0\rangle$ ▷ # $3mn$ flops
$C \leftarrow [4d_4 \quad 3d_3 \quad 2d_2 \quad d_1]$
$S \leftarrow roots(C), \quad t_* \leftarrow$ the smallest real positive root in S

7 Estimates of Rayleigh quotient for Riemannian Hessians

In many applications, (1) or (7) is often used for solving (2). In [14], it was proven that first-order and second-order optimality conditions for the nonconvex Burer–Monteiro approach are sufficient to find the global minimizer of certain convex semi-definite programs under certain assumptions. In practice, even if the minimizer $\hat{X}$ of (2) has a known rank r, one might consider solving (1) or (7) for Hermitian PSD matrices with fixed rank $p > r$. For instance, in PhaseLift [5] and interferometry recovery [8], the minimizer to (2) is has rank one, but in practice optimization over the set of PSD Hermitian matrices of rank p with $p \geq 2$ is often used because of a larger basin of attraction [8, 13].

If $p > r$, then an algorithm that solves (1) or (7) can generate a sequence that goes to the boundary of the manifold. Numerically, the smallest $p - r$ singular values of the iterates X_k will become very small as $k \to \infty$. In this section, we analyze the eigenvalues of the Riemannian Hessian. In particular we will obtain upper and lower bounds of the Rayleigh quotient at the point $X = YY^*$ (or $\pi(Y)$) that is close to the global minimum $\hat{X} = \hat{Y}\hat{Y}^*$ (or $\pi(\hat{Y})$). We assume that the Fréchet Hessian $\nabla^2 f$ is well conditioned when restricted to the tangent space. Formally, our bounds will be stated in terms of the constants A, B defined in the following assumption:

Assumption 7.1. *For a fixed $\epsilon > 0$, there exists constants $A > 0$ and B such that for all X with $\left\|X - \hat{X}\right\|_F < \epsilon$, the following inequality holds for all $\zeta_X \in T_X\mathcal{H}_+^{n,p}$.*

$$A\|\zeta_X\|_F^2 \leq \left\langle \nabla^2 f(X)[\zeta_X], \zeta_X\right\rangle_{\mathbb{C}^{n\times n}} \leq B\|\zeta_X\|_F^2. \tag{71}$$

An important case for which this assumption holds is $f(X) = \frac{1}{2}\|X - H\|_F^2$ with H a given Hermitian PSD matrix. In this case, $\nabla^2 f(X)$ is the identity operator and thus $A = B = 1$.

We summarize the main result in the following theorem. Its proof is outlined the subsections below.

Theorem 7.1. *Let $\hat{X} = \hat{Y}\hat{Y}^*$ be a minimizer of (2) with rank $r < p$. For $X = YY^*$ near $\hat{X}$ where $Y \in \mathbb{C}_*^{n\times p}$, let $\xi_{\pi(Y)} \in T_{\pi(Y)}\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ be any quotient tangent vector at $\pi(Y)$ and let $\overline{\xi}_Y \in \mathcal{H}_Y^i$ be its horizontal lift at Y w.r.t. the metric g^i. Define the Rayleigh quotient of the Riemannian Hessian of $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^i)$ as*

$$\rho^i(Y) = \frac{g^i_{\pi(Y)}(\text{Hess}\, h(\pi(Y))[\xi_{\pi(Y)}], \xi_{\pi(Y)})}{g^i_{\pi(Y)}(\xi_{\pi(Y)}, \xi_{\pi(Y)})}.$$

Then, under Assumption 7.1, it holds

$$0 \leq \lim_{\|Y-\hat{Y}\|_F \to 0} \rho^1(Y) \leq 4B\|\hat{X}\|, \quad 2A \leq \lim_{\|Y-\hat{Y}\|_F \to 0} \rho^2(Y) \leq 4B, \quad A \leq \lim_{\|Y-\hat{Y}\|_F \to 0} \rho^3(Y) \leq B,$$

where $\|\hat{X}\|$ is the spectral norm, that is, the largest eigenvalue of $\hat{X}$.

7.1 Quotient manifold

Let $\pi(Y) \in \mathbb{C}_*^{n\times p}/\mathcal{O}_p$ such that each $Y \in [Y]$ gives YY^* close enough to $\hat{X}$. Let $\xi_{\pi(Y)} \in T_{\pi(Y)}\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ be any quotient tangent vector at $\pi(Y)$ and let $\overline{\xi}_Y \in \mathcal{H}_Y^i$ be its horizontal lift at Y w.r.t. the metric g^i. We calculate the Rayleigh quotients ρ^i defined in Theorem 7.1 for the three metrics g^i, $i = 1, 2, 3$. Observe first that by definition of h, we have for each g^i that

$$\rho^i(Y) = \frac{g^i_Y(\text{Hess}\, F(Y)[\overline{\xi}_Y], \overline{\xi}_Y)}{g^i_Y(\overline{\xi}_Y, \overline{\xi}_Y)}.$$

Simple calculations for g^1 then gives

$$\rho^1(Y) = \frac{g^1_Y(2Herm\{\nabla^2 f(YY^*)[Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*]\}Y, \overline{\xi}_Y)}{g^1_Y(\overline{\xi}_Y, \overline{\xi}_Y)} + \frac{g^1_Y(2Herm(\nabla f(YY^*))\overline{\xi}_Y, \overline{\xi}_Y)}{g^1_Y(\overline{\xi}_Y, \overline{\xi}_Y)}.$$

Likewise, for g^2 we obtain

$$\begin{aligned}\rho^2(Y) = {} & \frac{g^2_Y(2Herm\{\nabla^2 f(YY^*)[Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*]\}Y(Y^*Y)^{-1}, \overline{\xi}_Y)}{g^2_Y(\overline{\xi}_Y, \overline{\xi}_Y)} \\ & + \frac{g^2_Y(Herm(\nabla f(YY^*))P_Y^{\perp}\overline{\xi}_Y(Y^*Y)^{-1}, \overline{\xi}_Y)}{g^2_Y(\overline{\xi}_Y, \overline{\xi}_Y)} + \frac{g^2_Y(P_Y^{\perp} Herm(\nabla f(YY^*))\overline{\xi}_Y(Y^*Y)^{-1}, \overline{\xi}_Y)}{g^2_Y(\overline{\xi}_Y, \overline{\xi}_Y)} \\ & + \frac{g^2_Y(2skew(\overline{\xi}_Y Y^*)Herm(\nabla f(YY^*))Y(Y^*Y)^{-2}, \overline{\xi}_Y)}{g^2_Y(\overline{\xi}_Y, \overline{\xi}_Y)} \\ & + \frac{2skew\{\overline{\xi}_Y(Y^*Y)^{-1}Y^* Herm(\nabla f(YY^*))\}Y(Y^*Y)^{-1}, \overline{\xi}_Y)}{g^2_Y(\overline{\xi}_Y, \overline{\xi}_Y)}.\end{aligned}$$

Finally, for g^3, we get

$$\rho^3(Y) = \frac{g^3_Y\left(\left(I - \frac{1}{2}P_Y\right) Herm\{\nabla^2 f(Y^*Y)[Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*]\}Y(Y^*Y)^{-1}, \overline{\xi}_Y\right)}{g^3_Y(\overline{\xi}_Y, \overline{\xi}_Y)} \tag{72}$$

$$+ \frac{g^3_Y((I - P_Y)Herm(\nabla f(YY^*))(I - P_Y)\overline{\xi}_Y(Y^*Y)^{-1}, \overline{\xi}_Y)}{g^3_Y(\overline{\xi}_Y, \overline{\xi}_Y)}. \tag{73}$$

Observe that when we use metric g^i for any i, the leading term in the Rayleigh quotient takes the same form

$$\frac{\left\langle \nabla^2 f(YY^*)[Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*], Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^* \right\rangle_{\mathbb{C}^{n\times n}}}{g^i_Y(\overline{\xi}_Y, \overline{\xi}_Y)}. \tag{74}$$

By Theorem 4.3 in [13], all other terms become arbitrarily small as $Y \to \hat{Y}$. Therefore, it suffices to analyze the fraction (74) near the true solution $\pi(\hat{Y})$. Observe that $Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^* \in T_{YY^*}\mathcal{H}_+^{n,p}$. Hence Assumption 7.1 also applies and becomes the following:

Lemma 7.2. *For a fixed $\epsilon > 0$, there exists constants A and B such that for all $Y \in \mathbb{C}_*^{n\times p}$ with $\left\|YY^* - \hat{X}\right\|_F < \epsilon$, the following inequality holds for all $\overline{\xi}_Y \in \mathcal{H}_Y^i$:*

$$A\left\|Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*\right\|_F^2 \leq \left\langle \nabla^2 f(YY^*)[Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*], Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^* \right\rangle_{\mathbb{C}^{n\times n}} \leq B\left\|Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*\right\|_F^2, \tag{75}$$

where $Y \in [Y]$ is a representation of $[Y]$.

Equation (74) therefore satisfies

$$A\frac{\left\|Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*\right\|_F^2}{g^i_Y(\overline{\xi}_Y, \overline{\xi}_Y)} \leq \frac{\left\langle \nabla^2 f(YY^*)[Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*], Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^* \right\rangle_{\mathbb{C}^{n\times n}}}{g^i_Y(\overline{\xi}_Y, \overline{\xi}_Y)} \leq B\frac{\left\|Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*\right\|_F^2}{g^i_Y\left(\overline{\xi}_Y, \overline{\xi}_Y\right)} \tag{76}$$

and it suffices to estimate the fraction

$$\frac{\left\|Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*\right\|_F^2}{g^i_Y\left(\overline{\xi}_Y, \overline{\xi}_Y\right)} \tag{77}$$

for different g^i. Below, we will bound this fraction as ξ_Y varies in $\mathcal{H}_Y^i$ for each g^i. The cases of g^2 and g^3 are simple while the case of g^1 involves more analysis.

7.2 Riemannian metric 1

For g^1, write $\overline{\xi}_Y = Y(Y^*Y)^{-1}S + Y_\perp K$ for some $S = S^* \in \mathbb{C}^{p\times p}$ and $K \in \mathbb{C}^{n\times p}$. Then (77) becomes

$$\frac{\left\|Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*\right\|_F^2}{g_Y^i(\overline{\xi}_Y, \overline{\xi}_Y)} = \frac{\left\|Y((Y^*Y)^{-1}S + S(Y^*Y)^{-1})Y^*\right\|_F^2 + 2\left\|KY^*\right\|_F^2}{\left\|Y(Y^*Y)^{-1}S\right\|_F^2 + \left\|K\right\|_F^2}. \tag{78}$$

Notice that the Rayleigh quotient is independent of the representative for $[Y]$. Hence, we can choose $Y = U\sqrt{\Sigma}$ where $YY^* = U\Sigma U^*$ is the SVD of YY^*.

Let K_i denote the ith column of K. Let σ_i denote the ith diagonal entry of Σ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq \epsilon > 0$. Similarly for $\hat{X}$, let $\hat{U}\hat{\Sigma}\hat{U}^*$ be its SVD and let $\hat{\sigma}_1 \geq \cdots, \geq \hat{\sigma}_r > 0$ be the singular values of $\hat{X}$. Define $\hat{\sigma}_{\max} = \max\{\hat{\sigma}_1, \cdots, \hat{\sigma}_r\}$ and $\hat{\sigma}_{\min} = \min\{\hat{\sigma}_1, \cdots, \hat{\sigma}_r\}$. Then (78) can be simplied to

$$\frac{\left\|Y((Y^*Y)^{-1}S + S(Y^*Y)^{-1})Y^*\right\|_F^2 + 2\left\|KY^*\right\|_F^2}{\left\|Y(Y^*Y)^{-1}S\right\|_F^2 + \left\|K\right\|_F^2} = \frac{\left\|\Sigma^{-\frac{1}{2}}S\Sigma^{\frac{1}{2}} + \Sigma^{\frac{1}{2}}S\Sigma^{-\frac{1}{2}}\right\|_F^2 + 2\left\|K\Sigma^{\frac{1}{2}}\right\|_F^2}{\left\|\Sigma^{-\frac{1}{2}}S\right\|_F^2 + \left\|K\right\|_F^2} \tag{79}$$

$$= \frac{\sum_{i,j=1}^p \left(\frac{\sqrt{\sigma_i}}{\sqrt{\sigma_j}} + \frac{\sqrt{\sigma_j}}{\sqrt{\sigma_i}}\right)^2 |S_{ij}|^2 + 2\sum_{i=1}^p \sigma_i \left\|K_i\right\|_F^2}{\sum_{i,j=1}^p \frac{|S_{ij}|^2}{\sigma_i} + \sum_{i=1}^p \left\|K_i\right\|_F^2} \tag{80}$$

$$= \frac{\sum_{i,j=1}^p \left(\frac{\sigma_i}{\sigma_j} + \frac{\sigma_j}{\sigma_i} + 2\right) |S_{ij}|^2 + 2\sum_{i=1}^p \sigma_i \left\|K_i\right\|_F^2}{\sum_{i,j=1}^p \frac{|S_{ij}|^2}{\sigma_i} + \sum_{i=1}^p \left\|K_i\right\|_F^2} \tag{81}$$

$$= \frac{2\sum_{i,j=1}^p \frac{\sigma_j}{\sigma_i} |S_{ij}|^2 + 2\sum_{i,j=1}^p |S_{ij}|^2 + 2\sum_{i=1}^p \sigma_i \left\|K_i\right\|_F^2}{\sum_{i,j=1}^p \frac{|S_{ij}|^2}{\sigma_i} + \sum_{i=1}^p \left\|K_i\right\|_F^2}. \tag{82}$$

Recall that $\hat{X}$ has rank r. By Weyl's theorem for the singular values of perturbed matrices [31], as $\pi(Y)$ approaches the true solution $\pi(\hat{Y})$ in the sense that $\left\|YY^* - \hat{X}\right\|_F$ goes to 0, the last $p - r$ eigenvalues in Σ will tend to 0, and the first r eigenvalues in Σ will approach the eigenvalues of $\hat{X}$. That is, $\sigma_i \to \hat{\sigma}_i$ for $1 \leq i \leq r$ and $\sigma_j \to 0$ for $r + 1 \leq j \leq p$. To simplify the formula from (82), introduce the two terms

$$M = 2\sum_{i=1}^r \frac{\sum_{j=1}^p \sigma_j |S_{i,j}|^2}{\sigma_i} + 2\sum_{i,j=1}^p |S_{i,j}|^2 + 2\sum_{i=1}^p \sigma_i \left\|K_i\right\|_F^2 \tag{83}$$

and

$$N = \sum_{i=1}^r \frac{\sum_{j=1}^p |S_{i,j}|^2}{\hat{\sigma}_i} + \sum_{i=1}^p \left\|K_i\right\|_F^2. \tag{84}$$

If $p \geq r + 1$, we can rewrite (82) as

$$\frac{2\sum_{i,j=1}^p \frac{\sigma_j}{\sigma_i} |S_{ij}|^2 + 2\sum_{i,j=1}^p |S_{ij}|^2 + 2\sum_{i=1}^p \sigma_i \left\|K_i\right\|_F^2}{\sum_{i,j=1}^p \frac{|S_{ij}|^2}{\sigma_i} + \sum_{i=1}^p \left\|K_i\right\|_F^2} \tag{85}$$

$$= \frac{2\left(\sum_{i=1}^r + \sum_{i=r+1}^p\right) \frac{\sum_{j=1}^p \sigma_j |S_{i,j}|^2}{\sigma_i} + 2\sum_{i,j=1}^p |S_{ij}|^2 + 2\sum_{i=1}^p \sigma_i \left\|K_i\right\|_F^2}{\sum_{i=1}^r \frac{\sum_{j=1}^p |S_{i,j}|^2}{\sigma_i} + \sum_{i=r+1}^p \frac{\sum_{j=1}^p |S_{i,j}|^2}{\sigma_i} + \sum_{i=1}^p \left\|K_i\right\|_F^2} \tag{86}$$

$$= \frac{2\sum_{i=r+1}^p \frac{\sum_{j=1}^p \sigma_j |S_{i,j}|^2}{\sigma_i} + M}{\sum_{i=r+1}^p \frac{\sum_{j=1}^p |S_{i,j}|^2}{\sigma_i} + N}. \tag{87}$$

If $p = r + 1$, then the limit of (87) as $\left\|YY^* - \hat{X}\right\|_F \to 0$ is

$$\lim_{\substack{\sigma_i \to \hat{\sigma}_i \\ \sigma_p \to 0}} \frac{2\frac{\sum_{j=1}^p \sigma_j |S_{p,j}|^2}{\sigma_p} + M}{\frac{\sum_{j=1}^p |S_{p,j}|^2}{\sigma_p} + N} = \frac{2\sum_{j=1}^r \hat{\sigma}_j |S_{p,j}|^2}{\sum_{j=1}^p |S_{p,j}|^2}. \tag{88}$$

If $S_{p,j} = 0$ for all $1 \leq j \leq r$ and $S_{p,p} \neq 0$, then the above limit equals zero. Otherwise it will be a nonzero constant.

If $p \geq r+2$, the limit of (87) as $\left\| YY^* - \hat{X} \right\|_F \to 0$ does not exist in general since the following ordered limit is dependent by its order that $\sigma_{r+1}, \cdots, \sigma_p$ goes to 0.

$$\lim_{\sigma_1 \to \hat{\sigma}_1} \cdots \lim_{\sigma_r \to \hat{\sigma}_r} \lim_{\sigma_{r+1} \to 0} \cdots \lim_{\sigma_p \to 0} \frac{2 \sum_{i=r+1}^{p} \frac{\sum_{j=1}^{p} \sigma_j |S_{i,j}|^2}{\sigma_i} + M}{\sum_{i=r+1}^{p} \frac{\sum_{j=1}^{p} |S_{i,j}|^2}{\sigma_i} + N} = \lim_{\sigma_1 \to \hat{\sigma}_1} \cdots \lim_{\sigma_r \to \hat{\sigma}_r} \lim_{\sigma_{r+1} \to 0} \cdots \lim_{\sigma_{p-1} \to 0} \frac{2 \sum_{j=1}^{p-1} \sigma_j \left|S_{p,j}\right|^2}{\sum_{j=1}^{p} \left|S_{p,j}\right|^2}$$
$$= \frac{2 \sum_{j=1}^{r} \hat{\sigma}_j \left|S_{p,j}\right|^2}{\sum_{j=1}^{p} \left|S_{p,j}\right|^2}.$$

If $S_{p,j} = 0$ for all $1 \leq j \leq r$ and $S_{p,j} \neq 0$ for some $r+1 \leq j \leq p$, then the above ordered limit is zero. Otherwise it will be a nonzero constant. Therefore, when $p = r$, the fraction (77) can be bounded between $2\sigma_{\min}$ and $4\sigma_{\max}$. When $p > r$, the range of the fraction (77) can be unbounded. This implies that the condition number of the Riemannian Hessian can be large when $p > r$ for metric g^1.

7.3 Riemannian metric 2

For g^2, write $\overline{\xi}_Y = YS + Y_\perp K$ for some $S = S^* \in \mathbb{C}^{p \times p}$ and $K \in \mathbb{C}^{n \times p}$. Then (77) becomes

$$\frac{\left\| Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^* \right\|_F^2}{g_Y^2(\overline{\xi}_Y, \overline{\xi}_Y)} = 2 + \frac{2 \left\| YSY^* \right\|_F^2}{\left\| YSY^* \right\|_F^2 + \left\| KY^* \right\|_F^2}. \tag{89}$$

When K is zero, (77) is upper bounded by 4. When S is a zero and K is nonzero, it is lower bounded by 2:

$$2A \leq \frac{\left\langle \nabla^2 f(YY^*)[Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*], Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^* \right\rangle_{\mathbb{C}^{n \times n}}}{g_Y^2(\overline{\xi}_Y, \overline{\xi}_Y)} \leq 4B. \tag{90}$$

Hence, the condition number of Hess $h(\pi(Y))$ is at most $\frac{2B}{A}$ when $\pi(Y) \to \pi(\hat{Y})$.

7.4 Riemannian metric 3

For g^3, (77) reduces to

$$\frac{\left\| Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^* \right\|_F}{g_Y^3(\overline{\xi}_Y, \overline{\xi}_Y)} = 1. \tag{91}$$

Hence we get directly

$$A \leq \frac{\left\langle \nabla^2 f(YY^*)[Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*], Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^* \right\rangle_{\mathbb{C}^{n \times n}}}{g_Y^3(\overline{\xi}_Y, \overline{\xi}_Y)} \leq B.$$

The condition number of Hess $h(\pi(Y))$ therefore is bounded by $\frac{B}{A}$ when $\pi(Y) \to \pi(\hat{Y})$.

8 Numerical experiments

In this section, we report on the numerical performance of the the conjugate gradient methods on three kinds of cost functions of $f(X)$: eigenvalue problem, matrix completion, phase-retrieval, and interferometry. In particular, we implement and compare the following four algorithms:

1. Riemannian CG on the quotient manifold $(\mathbb{C}_*^{n \times p}/\mathcal{O}_p, g^1)$, i.e., Algorithm 2 with metric g^1. This algorithm is equivalent to Burer–Monteiro CG, that is, CG applied directly to (5).
2. Riemannian CG on the quotient manifold $(\mathbb{C}_*^{n \times p}/\mathcal{O}_p, g^2)$, i.e., Algorithm 2 with metric g^2. The same metric g^2 was used in [13].

3. Riemannian CG on the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^3)$, i.e., Algorithm 2 with metric g^3, and also a specific retraction, vector transport and initial step as described in Section 5.2. This special implementation is equivalent to Riemannian CG on embedded manifold, i.e., Algorithm 1.

4. Burer–Monteiro L-BFGS method, that is, using the L-BFGS method directly applied to (5). This method was used in [8].

8.1 Eigenvalue problem

For any n-by-n Hermitian PSD matrix A, its top p eigenvalues and associated eigenvectors can be found by solving the following minimization problem:

$$\begin{array}{ll} \underset{X}{\text{minimize}} & f(X) := \frac{1}{2}\|X - A\|_F^2 \\ \text{subject to} & X \in \mathcal{H}_+^{n,p} \end{array},$$

or equivalently

$$\begin{array}{ll} \underset{\pi(Y)}{\text{minimize}} & h(\pi(Y)) := \frac{1}{2}\|YY^* - A\|_F^2 \\ \text{subject to} & \pi(Y) \in \mathbb{C}_*^{n\times p}/\mathcal{O}_p \end{array}.$$

It is easy to verify that

$$\nabla f(X) = X - A, \quad \nabla^2 f(X)[\zeta_X] = \zeta_X, \quad \zeta_X \in \mathbb{C}^{n\times n}.$$

We consider a numerical test for a random Hermitian PSD matrix A of size 50 000-by-50 000 with rank 10. We solve the minimization problem above with $p = 15$. Obivously, the minimizer is rank-10 thus rank deficient for $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ with $p = 15$. This corresponds to a scenario of finding the eigenvalue decomposition of a low rank Hermitian PSD matrix A with estimated rank at most 15. The results are shown in Figure 1. The initial guess is the same random initial matrix for all four algorithms. We see that the simpler Burer–Monteiro approach, including the L-BFGS method and the CG method with metric g^1, is significantly slower.

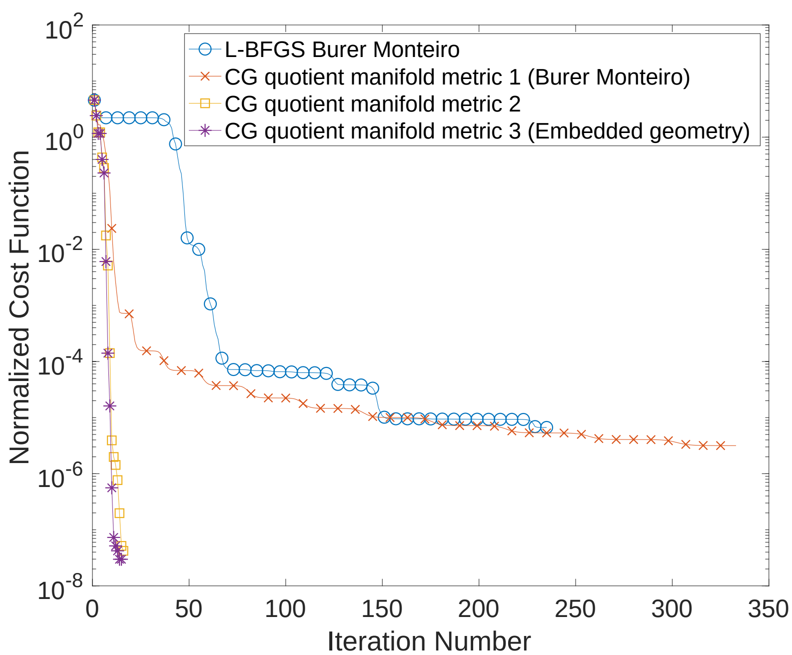


Figure 1: Eigenvalue problem of a random 50 000-by-50 000 PSD matrix of rank 10 on the rank 15 manifold.

8.2 Matrix completion

Let Ω be a subset of of the complete set $\{1, \cdots, n\} \times \{1, \cdots, n\}$. Then the projection operator onto Ω is a sampling operator defined as

$$P_\Omega : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n} : X_{i,j} \mapsto \begin{cases} X_{i,j} & \text{if } (i,j) \in \Omega, \\ 0 & \text{if } (i,j) \notin \Omega. \end{cases} \tag{92}$$

The original matrix completion problem has no symmetry or Hermitian constraint. Here, we just consider an artificial Hermitian matrix completion problem for a given $A \in \mathcal{H}_+^{n,p}$:

$$\begin{array}{ll} \underset{X}{\text{minimize}} & f(X) := \frac{1}{2} \|P_\Omega(X - A)\|_F^2 \\ \text{subject to} & X \in \mathcal{H}_+^{n,p} \end{array}, \tag{93}$$

or equivalently

$$\begin{array}{ll} \underset{\pi(Y)}{\text{minimize}} & h(\pi(Y)) := \frac{1}{2} \|P_\Omega(YY^* - A)\|_F^2 \\ \text{subject to} & \pi(Y) \in \mathbb{C}_*^{n\times p}/\mathcal{O}_p \end{array}. \tag{94}$$

Straightforward calculation shows

$$\nabla f(X) = P_\Omega(X - A), \quad \nabla^2 f(X)[\zeta_X] = P_\Omega(\zeta_X), \quad \zeta_X \in \mathbb{C}^{n\times n}.$$

We consider a Hermitian PSD matrix $A \in \mathbb{C}^{n\times n}$ with $n = 10\,000$ and P_Ω a random 90% sampling operator. In the first test of Figure 2, the minimizer has rank $r = 25$, and the fixed rank for the manifold is set to $p = 30$. In the second test of Figure 3, the minimizer has rank $r = 25$, and the fixed rank for the manifold is set to $p = 25$. The initial guess is the same random matrix for all four algorithms. For both cases, we see that the simpler Burer–Monteiro approach, including the L-BFGS method and the CG method with metric g^1, is significantly slower.

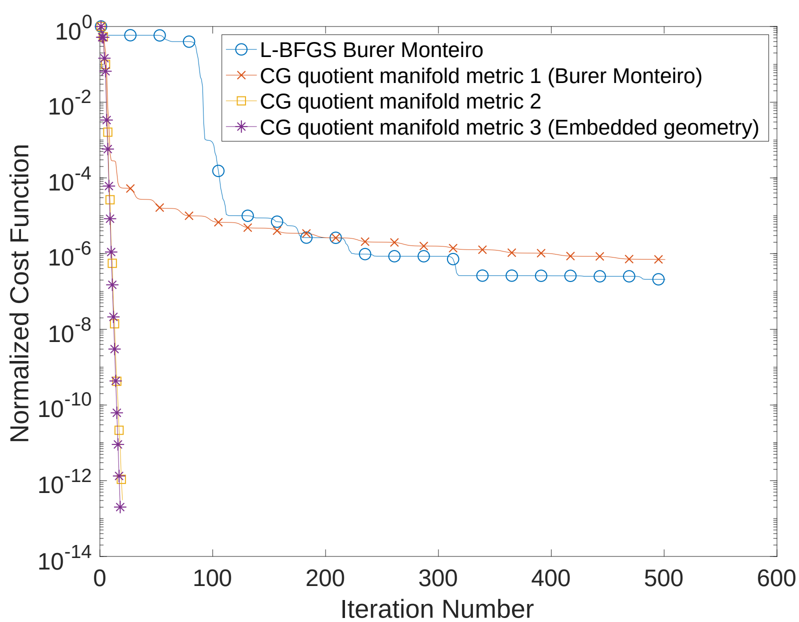


Figure 2: Matrix completion of a random 10 000-by-10 000 PSD matrix of rank 25 observed at random 90% entries. The algorithms are solved on the rank 30 manifold.

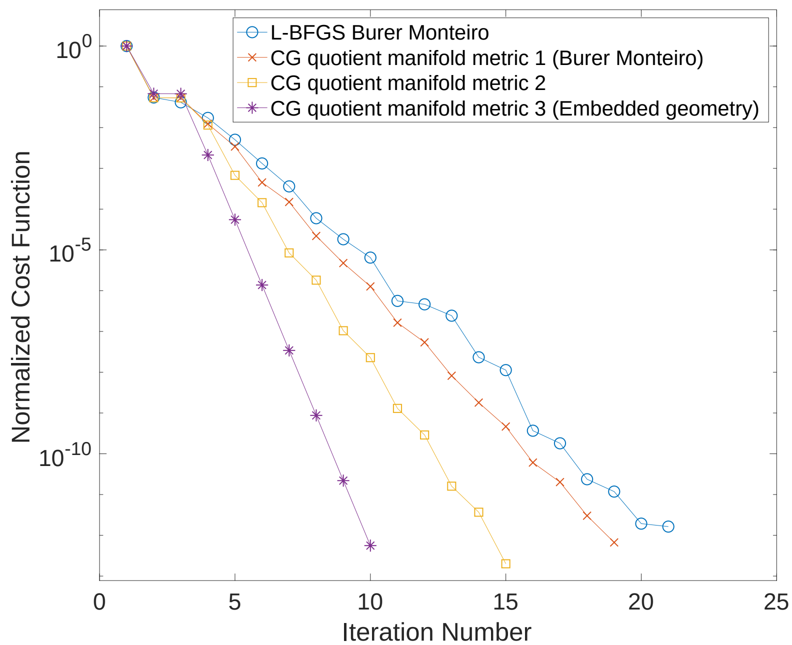


Figure 3: Same setting as in Figure 2 but minimized on the rank 25 manifold.

8.3 The PhaseLift problem

We now solve the phase retrieval problem as described in [5]: Take an image $x \in \mathbb{C}^{N^2\times 1}$ and a collection of masks denoted by $\{M_i\}_{i=1}^m$ where $N^2 = n$ is the size of the flattened image. Each M_i is of the same size as x and the elements in each M_i are real or complex numbers with both real and imaginary parts between 0 and 1. We can choose M_i to be random numbers or i.i.d. Gaussian. We have m number of observations for each mask $i = 1, \cdots, m$:

$$d^i = \mathcal{N}(x) := |(\text{DFT}(\text{Diag}(M_i) * x)|^2, \tag{95}$$

where $\mathcal{N}$ denotes the nonlinear operator. The squared power is taken element-wisely. $\text{Diag}(M_i)$ denotes the diagonal matrix whose diagonal is M_i. DFT denotes the $n \times n$ discrete fourier transform matrix. Therefore, d^i is vector of size $n \times 1$.

Now we lift x so that $\mathcal{N}$ can be treated as a linear operator. Let d^i_j denote the jth component of d^i. Let ${z^i}^*$ denote $\text{DFT} \cdot \text{Diag}(M_i)$ and ${z^i_j}^*$ denote the jth row of $\text{DFT} \cdot \text{Diag}(M_i)$. Then equation (95) can be written as

$$|\langle z^i_j, x\rangle|^2 = {z^i_j}^* x x^* z^i_j = d^i_j, \quad j = 1, \dots n, \quad i = 1, \dots, m. \tag{96}$$

Denoting $X := xx^*$, the nonlinear operator $\mathcal{N}$ can be rewritten as the linear operator

$$\mathcal{A} : \mathbb{C}^{n\times n} \to \mathbb{R}^{mn\times 1}, \quad X \mapsto [tr(z^1_1 {z^1_1}^* X), \cdots, tr(z^1_n {z^1_n}^* X), \cdots, tr(z^m_1 {z^m_1}^* X), \cdots, tr(z^m_n {z^m_n}^* X)]^T. \tag{97}$$

Let $Z^i := \text{DFT} \cdot \text{Diag}(M_i) = \begin{bmatrix} -{z^i_1}^* - \\ \cdots \\ -{z^i_n}^* - \end{bmatrix}$, then we have alternatively

$$\mathcal{A} : \mathbb{C}^{n\times n} \to \mathbb{R}^{mn\times 1}, \quad X \mapsto [\text{diag}(Z^1 X {Z^1}^*), \cdots, \text{diag}(Z^m X Z^{m*})]^T. \tag{98}$$

Denote $b = [d^1, \cdots, d^m]^T$. Then the cost function can be written as

$$f(X) = \frac{1}{2} \|\mathcal{A}(X) - b\|^2$$

The conjugate of operator $\mathcal{A}$, detoted by $\mathcal{A}^*$ is given by

$$A^*(b) = \begin{cases} \sum_{i=1}^{m}\sum_{j=1}^{n} b^i_j z^i_j {z^i_j}^* = \sum_{i=1}^{m} {Z^i}^* \operatorname{Diag}(b^i) Z^i, & \text{if domain of } \mathcal{A} \text{ is } \mathbb{C}^{n\times n} \\ \Re\left(\sum_{i=1}^{m}\sum_{j=1}^{n} b^i_j z^i_j {z^i_j}^*\right) = \Re\left(\sum_{i=1}^{m} {Z^i}^* \operatorname{Diag}(b^i) Z^i\right), & \text{if domain of } \mathcal{A} \text{ is } \mathbb{R}^{n\times n}. \end{cases}$$

Straightforward calculation shows

$$\nabla f(X) = \mathcal{A}^*(\mathcal{A}(X) - b), \quad \nabla^2 f(X)[\zeta_X] = \mathcal{A}^*(\mathcal{A}(\zeta_X)) \quad \text{for all } \zeta_X \in \mathbb{C}^{n\times n}.$$

For the numerical experiments, we take the phase retrieval problem for a complex gold ball image of size 256×256 as in [13]. Thus $n = 256^2 = 65,536$ in (2) or (1). We consider two different kinds of operator $\mathcal{A}$: the first corresponds to 6 Gaussian random masks and the second one to 8 Gaussian random masks. Hence, the size of b is $8n = 524,288$. Remark that problem is easier to solve with more masks.

We first test the algorithms on the rank 1 manifold, and then on the rank 3 manifolds. The results are visible in Figures 4–7. The initial guess is randomly generated. First, we observe that solving the PhaseLift problem on the rank p manifold with $p > 1$ can accelerate the convergence, compared to solving it on the rank 1 manifold. Second, when $p = r = 1$, the asymptotic convergence rates of all algorithms are essentially the same, though the algorithms differ in the length of their convergence "plateaus". When $p = 3 > r = 1$, we can see that the Burer–Monteiro approach has slower asymptotic convergence rates.

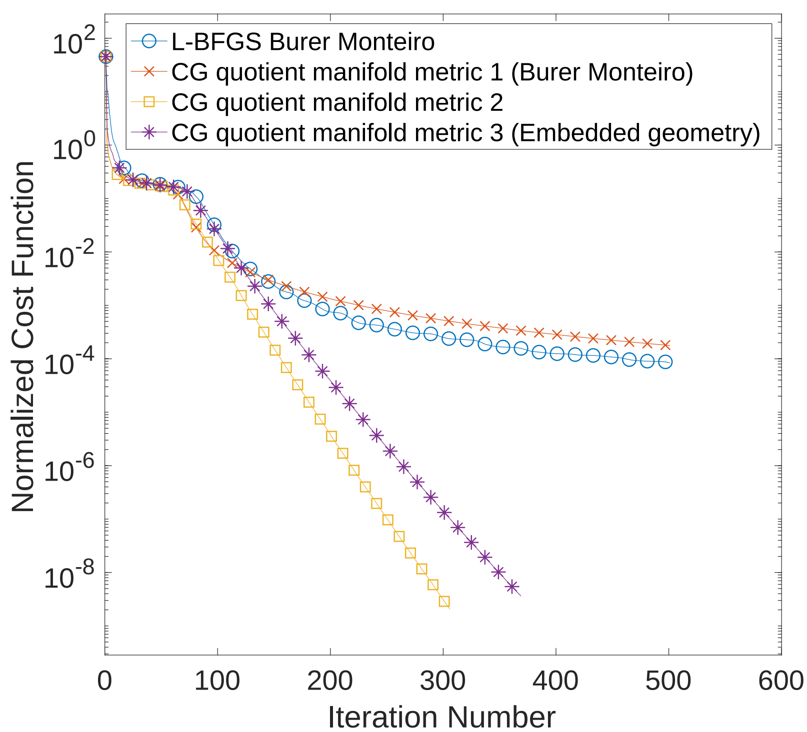


Figure 4: Phase retrieval of a 256-by-256 image with 6 Gaussian masks. The algorithms are solved on the rank 3 manifold.

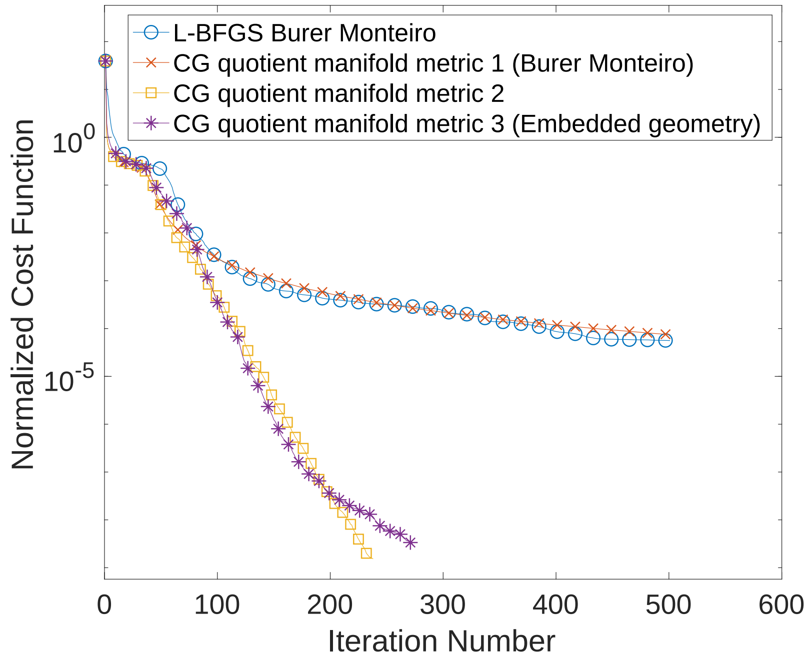


Figure 5: Phase retrieval of a 256-by-256 image with 8 Gaussian masks. The algorithms are solved on the rank 3 manifold.

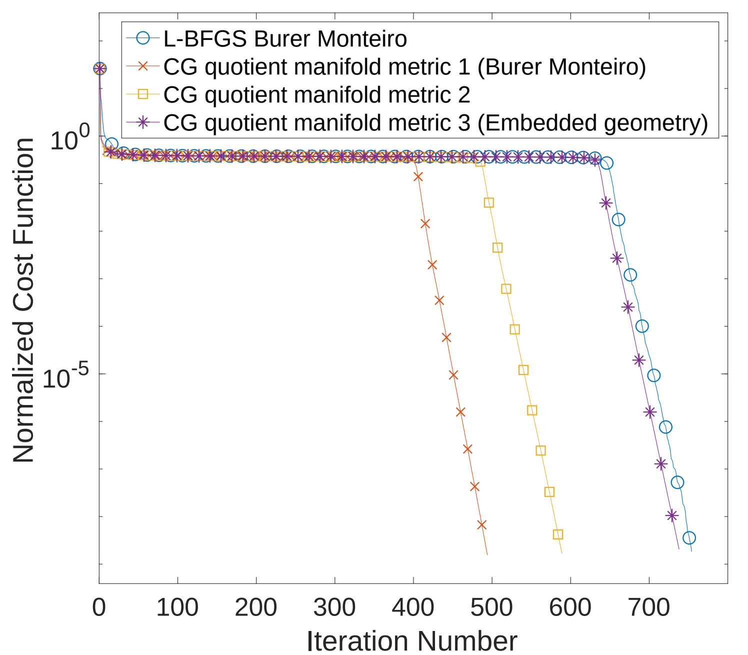


Figure 6: Same phase retrieval problem setting as in Figure 4 but The algorithms are solved on the rank 1 manifold.

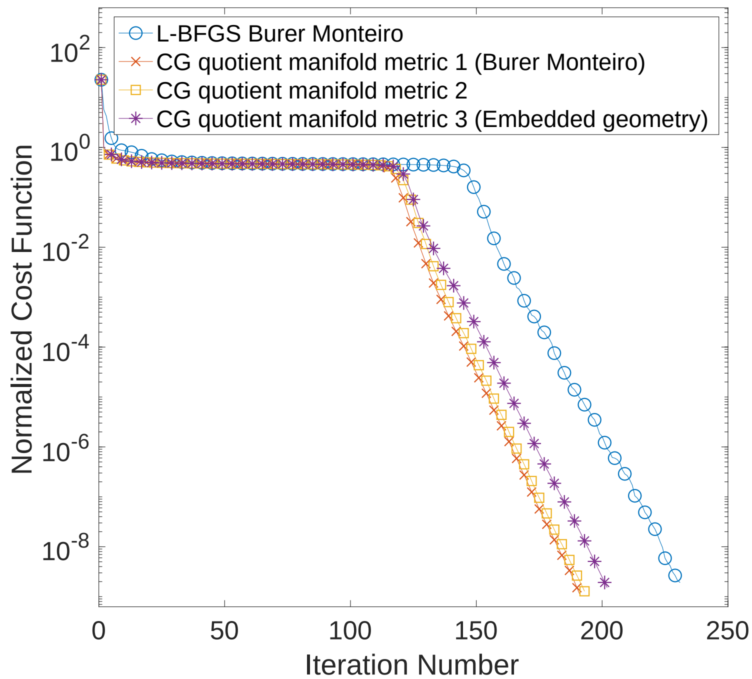


Figure 7: Same phase retrieval problem setting as in Figure 5 but The algorithms are solved on the rank 1 manifold.

8.4 Interferometry recovery problem

As last example, we consider solving the interferometry recovery problem described in [8]. Consider solving the linear system $Fx = d$ where $F \in \mathbb{C}_*^{m\times n}$ with $m > n$ and $x \in \mathbb{C}^{n\times 1}$. For the sake of robustness, the interferometry recovery [8] requires solving the lifted problem

$$\begin{array}{ll} \underset{X}{\text{minimize}} & f(X) = \frac{1}{2}\left\|P_\Omega(FXF^* - dd^*)\right\|_F^2 \\ \text{subject to} & X \in \mathcal{H}_+^{n,p} \end{array}, \tag{99}$$

where Ω is a sparse and symmetric sampling index that includes all of the diagonal.

Straightforward calculation again shows

$$\nabla f(X) = F^* P_\Omega(FXF^* - dd^*)F, \quad \nabla^2 f(X)[\zeta_X] = F^* P_\Omega(F\zeta_X F^*)F \quad \text{for all } \zeta_X \in \mathbb{C}^{n\times n}.$$

We solve an interferometry problem with a randomly generated $F \in \mathbb{C}^{10\,000\times 1000}$. Hence $n = 1000$ in (2) or (1). The sampling operator Ω is also randomly generated, with 70% density. In Figure8, $p = 3$ and $r = 1$ and we can see that the Burer–Monteiro approach has slower asymptotic convergence rates. In Figure9, $p = r = 1$ and we can see now that all algorithms have more or less the same asymptotic convergence rates.

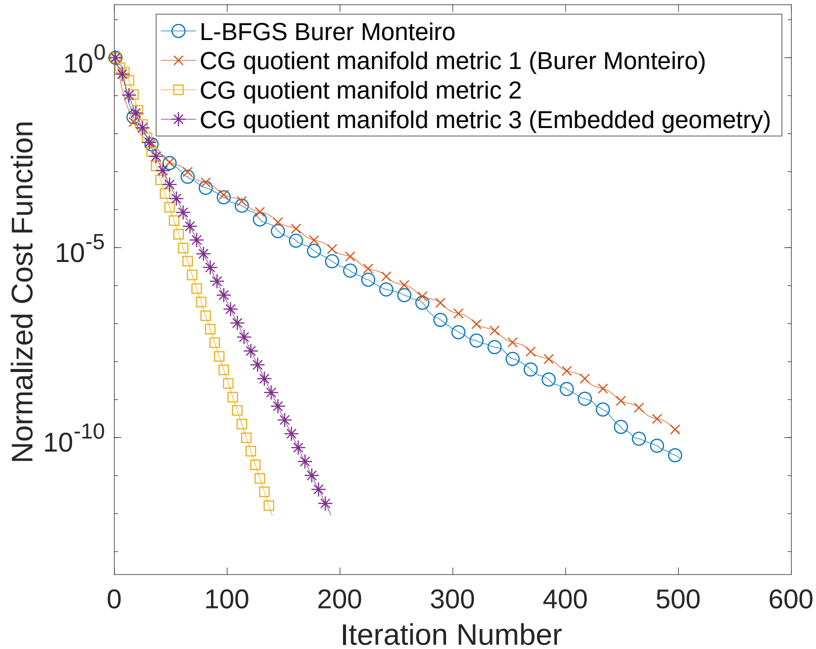


Figure 8: Interferometry recovery of a random 10 000-by-1000 F with 70% sampling. The algorithms are solved on the rank 3 manifold

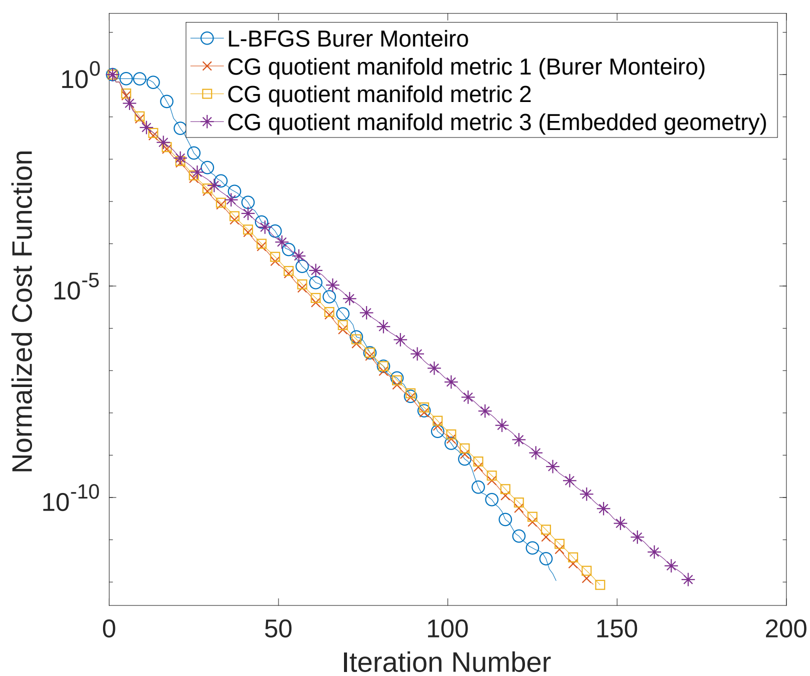


Figure 9: Same setting as in Figure 8 but on the rank 1 manifold.

9 Conclusion

In this paper, we have shown that the nonlinear conjugate gradient method on the unconstrained Burer–Monteiro formulation for Hermitian PSD fixed-rank constraints is equivalent to a Riemannian conjugate gradient method on a quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ with a specific metric g^1, retraction, and vector transport. We have also shown that the Riemannian conjugate gradient method on the embedded geometry of $\mathcal{H}_+^{n,p}$ is equivalent to a Riemannian conjugate gradient method on a quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ with a metric g^3, a special retraction, and a special vector transport. We have analyzed the condition numbers of the Riemannian Hessians on $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^i)$ for these metrics g^1, g^3 and another metric g^2 used in the literature. As notheworthy result, we have show that when the rank p of the optimization manifold is larger than the rank of the minimizer to the original PSD constrained minimization, the condition number of the Riemannian Hessian on $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^1)$ can be unbounded, which is consistent with the observation that the Burer–Monteiro approach often has a slower asymptotic convergence rate in numerical tests.

Appendix A Derivatives

See A.5 in [25] for more details in this section.

A.1 Fréchet derivatives

For any two finite-dimensional vector spaces $\mathcal{U}$ and $\mathcal{V}$ over $\mathbb{R}$, a mapping $F : \mathcal{U} \to \mathcal{V}$ is *Fréchet differentiable* at $x \in \mathcal{U}$ if there exists a linear operator

$$\begin{aligned} \mathrm{D}F(x): \quad \mathcal{U} &\to \mathcal{V} \\ h &\mapsto \mathrm{D}F(x)[h] \end{aligned}$$

such that

$$F(x+h) = F(x) + \mathrm{D}F(x)[h] + o(\|h\|).$$

The operator $\mathrm{D}F(x)$ is called the *Fréchet differential* and $\mathrm{D}F(x)[h]$ is called the *directional derivative* of F at x along h. The derivative satisfies the chain rule

$$\mathrm{D}(f \circ g)(x)[h] = \mathrm{D}f(g(x))[\mathrm{D}g(x)[h]].$$

For a smooth real-valued function $f : \mathcal{U} \to \mathbb{R}$, the *Fréchet gradient* of f at x, denoted by $\nabla f(x)$, is the unique element in $\mathcal{U}$ satisfying

$$\langle \nabla f(x), h \rangle_{\mathcal{U}} = \mathrm{D}f(x)[h], \quad \forall h \in \mathcal{U}, \tag{100}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{U}}$ is the inner product in $\mathcal{U}$.

In particular, consider $\mathcal{U} = \mathbb{C}^{n\times n}$ as a vector space over $\mathbb{R}$ with the standard inner product $\langle X, Y \rangle_{\mathbb{C}^{n\times n}} = \Re(\mathrm{tr}(X^*Y))$. Then the expression

$$\nabla f(X) = \frac{\partial f(X)}{\partial \Re_X} + \mathbb{i}\frac{\partial f(X)}{\partial \Im_X} \tag{101}$$

defines a Fréchet gradient. To see this, view X as $(\Re(X), \Im(X))$ and apply the multivariate Taylor theorem to f:

$$\begin{aligned} &|f(X+h) - f(X) - \langle \nabla f(X), h \rangle_{\mathbb{C}^{n\times n}}| = \\ &\left| f\left(\Re(X) + \Re(h), \Im(X) + \Im(h)\right) - f(\Re(X), \Im(X)) - \left(\left\langle \frac{\partial f}{\partial \Re(X)}, \Re(h) \right\rangle + \left\langle \frac{\partial f}{\partial \Im(X)}, \Im(h) \right\rangle \right) \right|, \end{aligned}$$

where X and h are n-by-n complex matrices. The last line is $o(\|h\|_{\mathbb{C}^{n\times n}})$ due to the multivariate Taylor theorem.

Now let $g : \mathbb{C}^{n\times p} \to \mathbb{C}^{n\times n} : Y \mapsto YY^*$ and the the inner product on $\mathbb{C}^{n\times p}$ as the standard inner product $\langle A, B \rangle_{\mathbb{C}^{n\times p}} = \Re(\mathrm{tr}(A^*B))$. Then the Fréchet gradient of $q := f \circ g$ satisfies

$$q'(Y) = 2Herm(\nabla f(YY^*))Y. \tag{102}$$

Indeed, by the chain rule of Fréchet derivative we have

$$\mathrm{D}\, q(Y)[h] = \mathrm{D}\, f(g(Y))\left[\mathrm{D}\, g(Y)[h]\right], \quad \forall h \in \mathbb{C}^{n\times p}. \tag{103}$$

Hence

$$\langle q'(Y), h \rangle_{\mathbb{C}^{n\times p}} = \langle \nabla f(YY^*), \mathrm{D}\, g(Y)[h] \rangle_{\mathbb{C}^{n\times n}}. \tag{104}$$

One can check by definition that $\mathrm{D}\, g(Y)[h] = Yh^* + hY^*$. Hence

$$\langle q'(Y), h \rangle_{\mathbb{C}^{n\times p}} = \langle \nabla f(YY^*), Yh^* + hY^* \rangle_{\mathbb{C}^{n\times n}} = \langle 2Herm(\nabla f(YY^*))Y, h \rangle_{\mathbb{C}^{n\times p}}. \tag{105}$$

This proves (102).

Theorem A.1. *The Fréchet gradient of* $f(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_F^2$ *for a linear operator* $\mathcal{A}$ *is given by*

$$\nabla f(X) = \mathcal{A}^*(\mathcal{A}(X) - b). \tag{106}$$

Proof. We know by the definition of Fréchet gradient (see A.1) that

$$\nabla f(X) = \frac{\partial f}{\partial \Re_X} + \mathbb{i}\frac{\partial f}{\partial \Im_X}, \tag{107}$$

Now for $f(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|^2 = \frac{1}{2}\langle \mathcal{A}(X) - b, \mathcal{A}(X) - b \rangle$, by the linearity of $\mathcal{A}$, we have

$$\nabla f(X) \quad = \quad \frac{1}{2}\frac{\partial}{\partial X}\langle \mathcal{A}(X) - b, \mathcal{A}(\Delta) - b \rangle\Big|_{\Delta = X} + \frac{1}{2}\frac{\partial}{\partial \Delta}\langle \mathcal{A}(X) - b, \mathcal{A}(\Delta) - b \rangle\Big|_{\Delta = X}$$

$$
\begin{aligned}
&= \frac{1}{2}\frac{\partial}{\partial X}\langle \mathcal{A}(X)-b, \mathcal{A}(\Delta)-b\rangle\Big|_{\Delta=X} + \frac{1}{2}\frac{\partial}{\partial \Delta}\langle \mathcal{A}(\Delta)-b, \mathcal{A}(X)-b\rangle\Big|_{\Delta=X} \\
&= \frac{1}{2}\frac{\partial}{\partial X}\langle \mathcal{A}(X)-b, \mathcal{A}(\Delta)-b\rangle_{\mathbb{C}^{n\times n}}\Big|_{\Delta=X} + \frac{1}{2}\frac{\partial}{\partial \Delta}\langle \mathcal{A}(\Delta)-b, \mathcal{A}(X)-b\rangle_{\mathbb{C}^{n\times n}}\Big|_{\Delta=X} \\
&= \frac{1}{2}\frac{\partial}{\partial X}\langle X, \mathcal{A}^*(\mathcal{A}(\Delta)-b)\rangle_{\mathbb{C}^{n\times n}}\Big|_{\Delta=X} + \frac{1}{2}\frac{\partial}{\partial \Delta}\langle \Delta, \mathcal{A}^*(\mathcal{A}(X)-b)\rangle_{\mathbb{C}^{n\times n}}\Big|_{\Delta=X} \\
&= \frac{1}{2}\frac{\partial}{\partial X}\left(\langle \Re(X), \Re(\mathcal{A}^*(\mathcal{A}(\Delta)-b))\rangle + \langle \Im(X), \Im(\mathcal{A}^*(\mathcal{A}(\Delta)-b))\rangle\right)\Big|_{\Delta=X} \\
&\quad + \frac{1}{2}\frac{\partial}{\partial \Delta}\left(\langle \Re(\Delta), \Re(\mathcal{A}^*(\mathcal{A}(X)-b))\rangle + \langle \Im(\Delta), \Im(\mathcal{A}^*(\mathcal{A}(X)-b))\rangle\right)\Big|_{\Delta=X} \\
&= \frac{1}{2}\left(\frac{\partial}{\partial \Re(X)} + \mathbb{i}\frac{\partial}{\partial \Im(X)}\right)\left(\langle \Re(X), \Re(\mathcal{A}^*(\mathcal{A}(\Delta)-b))\rangle + \langle \Im(X), \Im(\mathcal{A}^*(\mathcal{A}(\Delta)-b))\rangle\right)\Big|_{\Delta=X} \\
&\quad + \frac{1}{2}\left(\frac{\partial}{\partial \Re(\Delta)} + \mathbb{i}\frac{\partial}{\partial \Im(\Delta)}\right)\left(\langle \Re(\Delta), \Re(\mathcal{A}^*(\mathcal{A}(X)-b))\rangle + \langle \Im(\Delta), \Im(\mathcal{A}^*(\mathcal{A}(X)-b))\rangle\right)\Big|_{\Delta=X} \\
&= \frac{1}{2}(\Re(\mathcal{A}^*(\mathcal{A}(\Delta)-b)) + \mathbb{i}\Im(\mathcal{A}^*(\mathcal{A}(\Delta)-b)))\Big|_{\Delta=X} \\
&\quad + \frac{1}{2}(\Re(\mathcal{A}^*(\mathcal{A}(X)-b)) + \mathbb{i}\Im(\mathcal{A}^*(\mathcal{A}(X)-b)))\Big|_{\Delta=X} \\
&= \mathcal{A}^*(\mathcal{A}(X)-b).
\end{aligned}
$$

□

A.2 Hessian

For a Euclidean space $\mathcal{E}$ and a twice-differentiable, real-valued function f on $\mathcal{E}$, the *Fréchet Hessian operator* of f at x is the unique symmetric operator $\nabla^2 f(x) : \mathcal{E} \to \mathcal{E}$ defined by

$$\nabla^2 f(x)[h] = \mathrm{D}\,(f')(x)[h] \tag{108}$$

for all $h \in \mathcal{E}$.

A.3 Taylor's formula

Let $\mathcal{E}$ be finite-dimensional Euclidean space. Let f be a twice-differentiable real-valued function on an open convex domain $\Omega \subset \mathcal{E}$. Then for all x and $x + h \in \Omega$,

$$f(x+h) = f(x) + \langle \nabla f(x), h\rangle_{\mathcal{E}} + \frac{1}{2}\left\langle \nabla^2 f(x)[h], h\right\rangle_{\mathcal{E}} + O\left(\|h\|_{\mathcal{E}}^3\right). \tag{109}$$

Appendix B Embedded manifold $\mathcal{H}_+^{n,p}$

The geometry of the real case, i.e., $\mathcal{S}_+^{n,p}$ has been explored in [12]. However, it is not straightforward to extend these results directly to the complex case. Although the methods of proofs of the complex case turn out to be similar to the real case, we still need to provide. In this paper, recall that a complex matrix manifold is viewed as a manifold over $\mathbb{R}$ instead of $\mathbb{C}$. One way is to identify a complex matrix with the pair of its real and imaginary part; another way is to identify the matrix with its *realification*.

Definition B.1 (Realification)**.** *The realification is an injective mapping $\mathcal{R} : \mathbb{C}^{n\times n} \to \mathbb{R}^{2n\times 2n}$ defined by replacing each entry A_{ij} of $A \in \mathbb{C}^{n\times n}$ by the 2×2 matrix $\begin{bmatrix}\Re(A_{ij}) & -\Im(A_{ij})\\ \Im(A_{ij}) & \Re(A_{ij})\end{bmatrix}$. It can be shown that $\mathcal{R}$ preserves the algebraic structure:*

- $\mathcal{R}(A+B) = \mathcal{R}(A) + \mathcal{R}(B)$
- $\mathcal{R}(AB) = \mathcal{R}(A)\mathcal{R}(B)$
- $\mathcal{R}(aA) = a\mathcal{R}(A) \quad \forall a \in \mathbb{R}$

- $\mathcal{R}(I) = I$
- $\mathcal{R}(A^*) = (\mathcal{R}(A))^T$

Hence $A \in \mathbb{C}^{n\times n}$ is invertible if and only if $\mathcal{R}(A)$ is invertible.

$\mathbb{C}^{n\times n}$ is a vector space of dimension $2n^2$ over $\mathbb{R}$, and thus a smooth manifold of dimension $2n^2$. We will show in Theorem 3.1 that $\mathcal{H}_+^{n,p}$ is a smooth embedded submanifold of $\mathbb{C}^{n\times n}$ of dimension $2np - p^2$.

Lemma B.1. *Let $\mathrm{GL}(n, \mathbb{C})$ be the general linear group viewed as a real Lie group. Then it is a semialgebraic set.*

Proof. Recall that a subset of $\mathbb{R}^m$ is a *semialgebraic* set if it can be obtained by finitely many intersections, union and set differences starting from sets of the from $\{x \in \mathbb{R}^m : P(x) > 0\}$ with P a polynomial on $\mathbb{R}^m$ [23, Appendix B]. Since $\mathrm{GL}(n, \mathbb{C})$ is viewed as a real Lie group, $\mathrm{GL}(n, \mathbb{C})$ is understood as a subset of $\mathrm{GL}(2n, \mathbb{R})$ through realification. It can be shown that

$$\mathrm{GL}(n, \mathbb{C}) = \{X \in \mathrm{GL}(2n, \mathbb{R}) : XJ = JX\}, \quad \text{with } J = \mathcal{R}(\mathrm{i}I). \tag{110}$$

We know that $\mathrm{GL}(2n, \mathbb{R})$ is a semialgebraic set since it is the non-vanishing points of determinant; and $\{X \in \mathbb{R}^{2n\times 2n} : XJ = JX\}$ is also a semialgebraic set by definition. Hence $\mathrm{GL}(n, \mathbb{C})$ is a semialgebraic set. □

B.1 Riemannian Hessian operator

Let f be a smooth real-valued function on $\mathcal{H}_+^{n,p}$. In this section we derive the Riemannian Hessian operator of f.

By [26, section 4] we know that R defined in (28) is a second-order retraction. One can also see this from the following remark.

Remark B.2. *Since $\mathcal{H}_+^{n,p}$ is a Riemannian submanifold of the Euclidean space $\mathbb{C}^{n\times n}$, the Riemannian connection on $\mathcal{H}_+^{n,p}$ satisfies*

$$\nabla_{\eta_X}\xi = P_X^t(D\,\xi(X)[\eta_X]), \tag{111}$$

On other words, it is a classical directional derivative followed by an orthogonal projection to tangent space.(See [25, Proposition 5.3.2])

The definition of a second-order retraction in [26, Equation 2.3] is equivalent with the definition in [25, Proposition 5.5.5] as one can simply check the following. For all $\xi_X \in T_X\mathcal{H}_+^{n,p}$, we have

$$\frac{D^2}{dt^2}R(t\xi_X)\bigg|_{t=0} = \frac{D}{dt}\left[\frac{d}{dt}R(t\xi_X)\right]\bigg|_{t=0} = \nabla_{\xi_X}\left(\frac{d}{dt}R(t\xi_X)\right) \tag{112}$$

$$= P_X^t\left(\frac{d^2}{dt^2}\bigg|_{t=0}R(t\xi_X)\right) = 0. \tag{113}$$

Proposition 5.5.5 in [25] states that if R is a second-order retraction, then the Riemannian Hessian of f can be computed in the following nice way:

$$\mathrm{Hess}\, f(X) = \mathrm{Hess}\,(f \circ R_X)(0_X). \tag{114}$$

Notice that now $f \circ R_X$ is a smooth function defined on a vector space. Hence, we obtain

$$g_X\left(\mathrm{Hess}\, f(X)[\xi_X], \xi_X\right) = \frac{d^2}{dt^2}f(R_X(t\xi_X))\bigg|_{t=0}. \tag{115}$$

However, it is difficult to obtain a second-order derivative of $f \circ R_X$ using the retraction R_X defined in (28). The references [4] and [10] proposed a method to compute $\mathrm{Hess}\, f(X)$ by constructing a second-order retraction $R^{(2)}$ that has a second-order series expansion which makes it simple to derive a series expansion of $f \circ R_X^{(2)}$ up to second order and thus obtain the Hessian of f. We will summarize the derivation below.

Lemma B.3. *For any $X \in \mathcal{H}_+^{n,p}$ with $X^\dagger$ the pseudoinverse, the mapping $R_X^{(2)} : T_X\mathcal{H}_+^{n,p} \to \mathcal{H}_+^{n,p}$ given by*

$$\xi_X \mapsto wX^\dagger w^*, \text{ with } w = X + \frac{1}{2}\xi_X^s + \xi_X^p - \frac{1}{8}\xi_X^s X^\dagger \xi_X^s - \frac{1}{2}\xi_X^p X^\dagger \xi_X^s, \tag{116}$$

where $\xi_X^s = P_X^s(\xi_X)$ and $\xi_X^p = P_X^p(\xi_X)$ (see 26) is a second-order retraction on $\mathcal{H}_+^{n,p}$. Moreover we have

$$R_X^{(2)}(\xi_X) = X + \xi_X + \xi_X^p X^\dagger \xi_X^p + O(\|\xi_X\|^3). \tag{117}$$

Proof. It follows the same proof of [4, Proposition 5.10] . ☐

From this the Riemannian Hessian operator of f can be computed in essentially the same way as in [28, Section A.2] but applied to the general cost function $f(X)$. Consider the Taylor expansion of $\hat{f}_X^{(2)} := f \circ R_X^{(2)}$, which is a real-valued function on a vector space. We get

$$\begin{aligned}
\hat{f}_X^{(2)}(\xi_X) &= f(R_X^{(2)}(\xi_X)) \\
&= f\left(X + \xi_X + \xi_X^p X^\dagger \xi_X^p + O(\|\xi_X\|^3)\right) \\
&= f(X) + \left\langle \nabla f(X), \xi_X + \xi_X^p X^\dagger \xi_X^p \right\rangle_{\mathbb{C}^{n\times n}} + \frac{1}{2}\left\langle \nabla^2 f(X)[\xi_X + \xi_X^p X^\dagger \xi_X^p], \xi_X + \xi_X^p X^\dagger \xi_X^p \right\rangle_{\mathbb{C}^{n\times n}} + O(\|\xi_X\|^3) \\
&= f(X) + \langle \nabla f(X), \xi_X \rangle_{\mathbb{C}^{n\times n}} + \left\langle \nabla f(X), \xi_X^p X^\dagger \xi_X^p \right\rangle_{\mathbb{C}^{n\times n}} + \frac{1}{2}\left\langle \nabla^2 f(X)[\xi_X], \xi_X \right\rangle_{\mathbb{C}^{n\times n}} + O(\|\xi_X\|^3).
\end{aligned}$$

We can immediately recognize the first order term and the second order term that contribute to the Riemannian gradient and Hessian, respectively. That is,

$$g_X(\operatorname{grad} f(X), \xi_X) = \langle \nabla f(X), \xi_X \rangle_{\mathbb{C}^{n\times n}}, \tag{118}$$

$$g_X(\operatorname{Hess} f(X)[\xi_X], \xi_X) = \underbrace{2\left\langle \nabla f(X), \xi_X^p X^\dagger \xi_X^p \right\rangle_{\mathbb{C}^{n\times n}}}_{f_1 := \langle \mathcal{H}_1(\xi_X), \xi_X \rangle_{\mathbb{C}^{n\times n}}} + \underbrace{\left\langle \nabla^2 f(X)[\xi_X], \xi_X \right\rangle_{\mathbb{C}^{n\times n}}}_{f_2 := \langle \mathcal{H}_2(\xi_X), \xi_X \rangle_{\mathbb{C}^{n\times n}}}. \tag{119}$$

The first equation immediately gives us

$$\operatorname{grad} f(X) = P_X^t(\nabla f(X)). \tag{120}$$

For the second equation, the inner product of the Riemannian Hessian consists of the sum of f_1 and f_2; and the Riemannian Hessian operator is the sum of two operators $\mathcal{H}_1$ and $\mathcal{H}_2$. Since ξ_X is already separated in f_2, the contribution to Riemannian Hessian from $\mathcal{H}_2$ is readily given by

$$\mathcal{H}_2(\xi_X) = P_X^t(\nabla^2 f(X)[\xi_X]). \tag{121}$$

Now, we still need to separate ξ_X in f_1 to see the contribution to Riemannian Hessian from $\mathcal{H}_1$. Since we can choose to bring over $\xi_X^p X^\dagger$ or $X^\dagger \xi_X^p$, we write $\mathcal{H}_1(\xi_X)$ as the linear combination of both:

$$f_1 = 2c\left\langle \nabla f(X)(X^\dagger \xi_X^p)^*, \xi_X^p \right\rangle_{\mathbb{C}^{n\times n}} + 2(1-c)\left\langle (\xi_X^p X^\dagger)^* \nabla f(X), \xi_X^p \right\rangle_{\mathbb{C}^{n\times n}}. \tag{122}$$

Operator $\mathcal{H}_1$ is clearly linear. Since $\mathcal{H}_1$ is symmetric, we must have $\langle \mathcal{H}_1(\xi_X), \nu_X \rangle_{\mathbb{C}^{n\times n}} = \langle \nu_X, \mathcal{H}_1(\xi_X) \rangle_{\mathbb{C}^{n\times n}}$ for all ν_X. Hence we must have $c = \frac{1}{2}$ and we obtain

$$\mathcal{H}_1(\xi_X) = P_X^p\left(\nabla f(X)(X^\dagger \xi_X^p)^* + (\xi_X^p X^\dagger)^* \nabla f(X)\right). \tag{123}$$

Putting $\mathcal{H}_1$ and $\mathcal{H}_2$ together, we obtain

$$\operatorname{Hess} f(X)[\xi_X] = P_X^t(\nabla^2 f(X)[\xi_X]) + P_X^p\left(\nabla f(X)(X^\dagger \xi_X^p)^* + (\xi_X^p X^\dagger)^* \nabla f(X)\right). \tag{124}$$

Appendix C Quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$

C.1 Calculations for the Riemannian Hessian

In this section, we outline the computations of the Riemannian Hessian operators of the cost function h defined on $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ under the three different metrics g^i.

Definition C.1. *[25, Definition 5.5.1] Given a real-valued function f on a Riemannian manifold $\mathcal{M}$, the Riemannian Hessian of f at a point x in $\mathcal{M}$ is the linear mapping Hess $f(x)$ of $T_x\mathcal{M}$ into itself defined by*

$$\textit{Hess}\, f(x)[\xi_x] = \nabla_{\xi_x} \textit{grad}\, f(x) \tag{125}$$

for all ξ_x in $T_x\mathcal{M}$, where ∇ is the Riemannian connection on $\mathcal{M}$.

Lemma C.1. *The Riemannian Hessian of $h : \mathbb{C}_*^{n\times p}/\mathcal{O}_p \mapsto \mathbb{R}$ is related to the Riemannian Hessian of $F : \mathbb{C}_*^{n\times p} \mapsto \mathbb{R}$ in the following way:*

$$\overline{\left(\textit{Hess}\, h(\pi(Y))[\xi_{\pi(Y)}]\right)}_Y = P_Y^{\mathcal{H}}\left(\textit{Hess}\, F(Y)[\overline{\xi}_Y]\right), \tag{126}$$

where $\overline{\xi}_Y$ is the horizontal lift of $\xi_{\pi(Y)}$ at Y.

Proof. The result follows from [25, Proposition 5.3.3] and the definition of the Riemannian Hessian. ☐

C.1.1 Riemannian Hessian for the metric g^1

Using the Riemannian metric g^1, $\mathbb{C}_*^{n\times p}$ is a Riemannian submanifold of a Euclidean space. By [25, Proposition 5.3.2], the Riemannian connection on $\mathbb{C}_*^{n\times p}$ is classical the directional derivative

$$\nabla_{\eta_Y}\xi = \mathrm{D}\,\xi(Y)[\eta_Y]. \tag{127}$$

Recall that for g^1, $\operatorname{grad} F(Y) = (\nabla f(YY^*) + \nabla f(YY^*)^*)Y$. Hence, the Riemannian Hessian of F at Y is given by

$$\begin{aligned}
\operatorname{Hess} F(Y)[\xi_Y] &= \nabla_{\xi_Y} \operatorname{grad} F & (128)\\
&= \mathrm{D}\operatorname{grad} F(Y)[\xi_Y] & (129)\\
&= 2Herm\{\nabla^2 f(YY^*)[Y\xi_Y^* + \xi_Y Y^*]\}Y + 2Herm(\nabla f(YY^*))\xi_Y. & (130)
\end{aligned}$$

The last line is by product rule and chain rule of differential. Therefore we obtain

$$\overline{\left(\operatorname{Hess} h(\pi(Y))[\xi_{\pi(Y)}]\right)}_Y = P_Y^{\mathcal{H}^1}\left(2Herm\{\nabla^2 f(YY^*)[Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*]\}Y + 2Herm(\nabla f(YY^*))\overline{\xi}_Y\right). \tag{131}$$

C.1.2 Riemannian Hessian under metric g^2

First, for any Riemannian metric g, g satisfies the Koszul formula

$$\begin{aligned}
2g_x(\nabla_{\xi_x}\lambda, \eta_x) &= \xi_x g(\lambda,\eta) + \lambda_x g(\eta,\xi) - \eta_x g(\xi,\lambda)\\
&\quad - g_x(\xi_x, [\lambda,\eta]_x) + g_x(\lambda_x, [\eta,\xi]_x) + g_x(\eta, [\xi,\lambda]_x)\\
&= \mathrm{D}\,g(\lambda,\eta)(x)[\xi_x] + \mathrm{D}\,g(\eta,\xi)(x)[\lambda_x] - \mathrm{D}\,g(\xi,\lambda)(x)[\eta_x]\\
&\quad - g_x(\xi_x, [\lambda,\eta]_x) + g_x(\lambda_x, [\eta,\xi]_x) + g_x(\eta, [\xi,\lambda]_x),
\end{aligned}$$

where the *Lie bracket* $[\cdot,\cdot]$ is defined in [25].

In particular, for g^2 the above Koszul formula turns into

$$\begin{aligned}
2g_Y^2(\nabla_{\xi_Y}\lambda, \eta_Y) &= \mathrm{D}\,g^2(\lambda,\eta)(Y)[\xi_Y] + \mathrm{D}\,g^2(\eta,\xi)(Y)[\lambda_Y] - \mathrm{D}\,g^2(\xi,\lambda)(Y)[\eta_Y] & (132)\\
&\quad - g_Y^2(\xi_Y, [\lambda,\eta]_Y) + g_Y^2(\lambda_Y, [\eta,\xi]_Y) + g_Y^2(\eta, [\xi,\lambda]_Y). & (133)
\end{aligned}$$

Recall that $g^2(\lambda,\eta)(Y) = \Re(tr(Y^*Y\lambda_Y^*\eta_Y))$. Hence, the first term in the above sum equals

$$\mathrm{D}\,g^2(\lambda,\eta)(Y)[\xi_Y] = g_Y^2(\mathrm{D}\,\lambda(Y)[\xi_Y], \eta_Y) + g_Y^2(\lambda_Y, \mathrm{D}\,\eta(Y)[\xi_Y]) + \Re(tr(\xi_Y^* Y\lambda_Y^*\eta_Y)) + \Re(tr(Y^*\xi_Y\lambda_Y^*\eta_Y)). \tag{134}$$

Following [25, Section 5.3.4], since $\mathbb{C}_*^{n\times p}$ is an open subset of $\mathbb{C}^{n\times p}$, we also have

$$[\lambda,\eta]_Y = \mathrm{D}\,\eta(Y)[\lambda_Y] - \mathrm{D}\,\lambda(Y)[\eta_Y]. \tag{135}$$

Summarizing, we get

$$\begin{aligned}
2g_Y^2(\nabla_{\xi_Y}\lambda, \eta_Y) &= \mathrm{D}\,g^2(\lambda,\eta)(Y)[\xi_Y] + \mathrm{D}\,g^2(\eta,\xi)(Y)[\lambda_Y] - \mathrm{D}\,g^2(\xi,\lambda)(Y)[\eta_Y]\\
&\quad - g^2(\xi_Y, \mathrm{D}\,\eta(Y)[\lambda_Y] - \mathrm{D}\,\lambda(Y)[\eta_Y])\\
&\quad + g^2(\lambda_Y, \mathrm{D}\,\xi(Y)[\eta_Y] - \mathrm{D}\,\eta(Y)[\xi_Y])\\
&\quad + g^2(\eta_Y, \mathrm{D}\,\lambda(Y)[\xi_Y] - \mathrm{D}\,\xi(Y)[\lambda_Y])\\
&= 2g_Y^2(\eta_Y, \mathrm{D}\,\lambda(Y)[\xi_Y])\\
&\quad + \Re(tr(\eta_Y^*(\lambda_Y(\xi_Y^*Y + Y^*\xi_Y) + \xi_Y(Y^*\lambda_Y + \lambda_Y^*Y) - Y\lambda_Y^*\xi_Y - Y\xi_Y^*\lambda_Y)))\\
&= 2g_Y^2(\eta_Y, \mathrm{D}\,\lambda(Y)[\xi_Y])\\
&\quad + g_Y^2(\eta_Y, (\lambda_Y(\xi_Y^*Y + Y^*\xi_Y) + \xi_Y(Y^*\lambda_Y + \lambda_Y^*Y) - Y\lambda_Y^*\xi_Y - Y\xi_Y^*\lambda_Y)(Y^*Y)^{-1}).
\end{aligned}$$

We therefore obtain a closed-form expression for Riemannian connection on $\mathbb{C}_*^{n\times p}$ for g^2:

$$\nabla_{\xi_Y}\lambda = \mathrm{D}\,\lambda(Y)[\xi_Y] + \frac{1}{2}\left(\lambda_Y(\xi_Y^*Y + Y^*\xi_Y) + \xi_Y(Y^*\lambda_Y + \lambda_Y^*Y) - Y\lambda_Y^*\xi_Y - Y\xi_Y^*\lambda_Y\right)(Y^*Y)^{-1}. \tag{136}$$

Recall that for the Riemannian metric g^2, we have $\operatorname{grad} F(Y) = (\nabla f(YY^*) + \nabla f(YY^*)^*)Y(Y^*Y)^{-1}$. Hence we have

$$\operatorname{Hess} F(Y)[\xi_Y] = \nabla_{\xi_Y} \operatorname{grad} F(Y)$$

$$
\begin{aligned}
&= \mathrm{D}_Y \mathrm{grad}\, F(Y)[\xi_Y] \\
&\quad + \frac{1}{2}\{\mathrm{grad}\, F(Y)(\xi_Y^* Y + Y^*\xi_Y) + \xi_Y(Y^*\mathrm{grad}\, F(Y) + \mathrm{grad}\, F(Y)^* Y) - \\
&\quad Y \mathrm{grad}\, F(Y)^*\xi_Y - Y\xi_Y^* \mathrm{grad}\, F(Y)\}(Y^*Y)^{-1} \\
&= 2Herm\{\nabla^2 f(YY^*)[Y\xi_Y^* + \xi_Y Y^*]\}Y(Y^*Y)^{-1} + 2Herm(\nabla f(YY^*))\xi_Y(Y^*Y)^{-1} \\
&\quad -2Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}(Y^*\xi_Y + \xi_Y^* Y)(Y^*Y)^{-1} \\
&\quad +Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}(Y^*\xi_Y + \xi_Y^* Y)(Y^*Y)^{-1} \\
&\quad +\xi_Y\{Y^*Herm(\nabla f(YY^*))Y(Y^*Y)^{-1} + (Y^*Y)^{-1}Y^*Herm(\nabla f(YY^*))Y\}(Y^*Y)^{-1} \\
&\quad -\{Y(Y^*Y)^{-1}Y^*Herm(\nabla f(YY^*))\xi_Y + Y\xi_Y^* Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\}(Y^*Y)^{-1} \\
&= 2Herm\{\nabla^2 f(YY^*)[Y\xi_Y^* + \xi_Y Y^*]\}Y(Y^*Y)^{-1} + 2Herm(\nabla f(YY^*))\xi_Y(Y^*Y)^{-1} \\
&\quad -Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}(Y^*\xi_Y + \xi_Y^* Y)(Y^*Y)^{-1} \\
&\quad +\xi_Y\{Y^*Herm(\nabla f(YY^*))Y(Y^*Y)^{-1} + (Y^*Y)^{-1}Y^*Herm(\nabla f(Y^*Y))Y\}(Y^*Y)^{-1} \\
&\quad -\{Y(Y^*Y)^{-1}Y^*Herm(\nabla f(YY^*))\xi_Y + Y\xi_Y^* Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\}(Y^*Y)^{-1} \\
&= 2Herm\{\nabla^2 f(YY^*)[Y\xi_Y^* + \xi_Y Y^*]\}Y(Y^*Y)^{-1} + 2Herm(\nabla f(YY^*))\xi_Y(Y^*Y)^{-1} \\
&\quad -Herm(\nabla f(YY^*))P_Y\xi_Y(Y^*Y)^{-1} - Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\xi_Y^* Y(Y^*Y)^{-1} \\
&\quad +\xi_Y Y^*Herm(\nabla f(YY^*))Y(Y^*Y)^{-2} + \xi_Y(Y^*Y)^{-1}Y^*Herm(\nabla f(YY^*))Y(Y^*Y)^{-1} \\
&\quad -P_Y Herm(\nabla f(YY^*))\xi_Y(Y^*Y)^{-1} - Y\xi_Y^* Herm(\nabla f(YY^*))Y(Y^*Y)^{-2} \\
&= 2Herm\{\nabla^2 f(YY^*)[Y\xi_Y^* + \xi_Y Y^*]\}Y(Y^*Y)^{-1} \\
&\quad +Herm(\nabla f(YY^*))\xi_Y(Y^*Y)^{-1} - Herm(\nabla f(YY^*))P_Y\xi_Y(Y^*Y)^{-1} \\
&\quad +Herm(\nabla f(YY^*))\xi_Y(Y^*Y)^{-1} - P_Y Herm(\nabla f(Y^*Y))\xi_Y(Y^*Y)^{-1} \\
&\quad +2skew(\xi_Y Y^*)Herm(\nabla f(YY^*))Y(Y^*Y)^{-2} \\
&\quad +2skew\{\xi_Y(Y^*Y)^{-1}Y^*Herm(\nabla f(YY^*))\}Y(Y^*Y)^{-1} \\
&= 2Herm\{\nabla^2 f(YY^*)[Y\xi_Y^* + \xi_Y Y^*]\}Y(Y^*Y)^{-1} \\
&\quad +Herm(\nabla f(YY^*))P_Y^\perp\xi_Y(Y^*Y)^{-1} + P_Y^\perp Herm(\nabla f(YY^*))\xi_Y(Y^*Y)^{-1} \\
&\quad +2skew(\xi_Y Y^*)Herm(\nabla f(YY^*))Y(Y^*Y)^{-2} \\
&\quad +2skew\{\xi_Y(Y^*Y)^{-1}Y^*Herm(\nabla f(YY^*))\}Y(Y^*Y)^{-1}.
\end{aligned}
$$

To conclude, we obtain

$$
\begin{aligned}
\overline{\left(\mathrm{Hess}\, h(\pi(Y))[\eta_{\pi(Y)}]\right)}_Y = & \; P_Y^{\mathcal{H}^2}\Big\{2Herm\{\nabla^2 f(YY^*)[Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*]\}Y(Y^*Y)^{-1} \\
& +Herm(\nabla f(YY^*))P_Y^\perp\overline{\xi}_Y(Y^*Y)^{-1} + P_Y^\perp Herm(\nabla f(YY^*))\overline{\xi}_Y(Y^*Y)^{-1} \\
& +2skew(\overline{\xi}_Y Y^*)Herm(\nabla f(YY^*))Y(Y^*Y)^{-2} \\
& + 2skew\{\overline{\xi}_Y(Y^*Y)^{-1}Y^*Herm(\nabla f(YY^*))\}Y(Y^*Y)^{-1}\Big\}.
\end{aligned}
$$

C.1.3 Riemannian Hessian under metric g^3

Recall that the Riemannian metric g^3 on $\mathbb{C}_*^{n\times p}$ satisfies

$$
\begin{aligned}
g_Y^3(\xi_Y,\eta_Y) &= \tilde{g}_Y(\xi_Y,\eta_Y) + g_Y^2(P_Y^{\mathcal{V}}(\xi_Y), P_Y^{\mathcal{V}}(\eta_Y)) && (137)\\
&= 2\Re(tr(Y^*\xi_Y Y^*\eta_Y + Y^*Y\xi_Y^*\eta_Y)) + \Re(tr(YP_Y^{\mathcal{V}}(\xi_Y)^* P_Y^{\mathcal{V}}(\eta_Y)Y^*)) && (138)
\end{aligned}
$$

where

$$\tilde{g}_Y(\xi_Y,\eta_Y) = \langle Y\xi_Y^* + \xi_Y Y^*, Y\eta_Y^* + \eta_Y Y^*\rangle_{\mathbb{C}^{n\times n}}. \quad (139)$$

$$P_Y^{\mathcal{V}}(\lambda_Y) = Y skew((Y^*Y)^{-1}Y^*\lambda_Y). \quad (140)$$

Hence

$$
\begin{aligned}
& \mathrm{D}\, g^3(\lambda,\eta)(Y)[\xi_Y] \\
= \; & \tilde{g}_Y(\mathrm{D}\,\lambda(Y)[\xi_Y],\eta_Y) + \tilde{g}(\lambda_Y, D\eta(Y)[\xi_Y]) + 2\Re(tr(\xi_Y^*\lambda_Y Y^*\eta_Y + Y^*\lambda_Y\xi_Y^*\eta_Y + \xi_Y^* Y\lambda_Y^*\eta_Y + Y^*\xi_Y\lambda_Y^*\eta_Y))
\end{aligned}
$$

$$+g_Y^2(P_Y^{\mathcal{V}}(\lambda_Y), DP_Y^{\mathcal{V}}(\eta_Y)[\xi_Y]) + g^2(\mathrm{D}\, P_Y^{\mathcal{V}}(\lambda_Y)[\xi_Y], P_Y^{\mathcal{V}}(\eta_Y))$$
$$+\Re(tr(\xi_Y P_Y^{\mathcal{V}}(\lambda_Y)^* P_Y^{\mathcal{V}}(\eta_Y)Y^* + Y P_Y^{\mathcal{V}}(\lambda_Y)^* P_Y^{\mathcal{V}}(\eta_Y)\xi_Y^*)).$$

Suppose λ, η and ξ are horizontal vector fields, then many terms in the above equation vanish:

$$\begin{aligned}\mathrm{D}\, g^3(\lambda,\eta)(Y)[\xi_Y] &= \tilde{g}_Y(\mathrm{D}\,\lambda(Y)[\xi_Y],\eta_Y) + \tilde{g}_Y(\lambda_Y, \mathrm{D}\,\eta_Y[\xi_Y]) \\ &\quad +2\Re(tr(\xi_Y^*\lambda_Y Y^*\eta_Y + Y^*\lambda_Y\xi_Y^*\eta_Y + \xi_Y^* Y\lambda_Y^*\eta_Y + Y^*\xi_Y\lambda_Y^*\eta_Y)).\end{aligned}$$

Combining the above equation and the Koszul formul with ξ, η, λ horizontal vector fields, we obtain

$$\begin{aligned}
& 2g_Y^3(\nabla_{\xi_Y}\lambda, \eta_Y) \\
= \; & \mathrm{D}\, g^3(\lambda,\eta)(Y)[\xi_Y] + \mathrm{D}\, g^3(\eta,\xi)(Y)[\lambda_Y] - \mathrm{D}\, g^3(\xi,\lambda)(Y)[\eta_Y] \\
& -g_Y^3(\xi_Y, \mathrm{D}\,\eta(Y)[\lambda_Y] - \mathrm{D}\,\lambda(Y)[\eta_Y]) \\
& +g_Y^3(\lambda_Y, \mathrm{D}\,\xi(Y)[\eta_Y] - \mathrm{D}\,\eta(Y)[\xi_Y]) \\
& +g_Y^3(\eta_Y, \mathrm{D}\,\lambda(Y)[\xi_Y] - \mathrm{D}\,\xi(Y)[\lambda_Y]) \\
= \; & \tilde{g}_Y(\mathrm{D}\,\lambda(Y)[\xi_Y],\eta_Y) + \tilde{g}_Y(\lambda_Y, \mathrm{D}\,\eta(Y)[\xi_Y]) + 2\Re(tr(\xi_Y^*\lambda_Y Y^*\eta_Y + Y^*\lambda_Y\xi_Y^*\eta_Y + \xi_Y^* Y\lambda_Y^*\eta_Y + Y^*\xi_Y\lambda_Y^*\eta_Y)) \\
& +\tilde{g}_Y(\mathrm{D}\,\eta(Y)[\lambda_Y],\xi_Y) + \tilde{g}_Y(\eta_Y, \mathrm{D}\,\xi(Y)[\lambda_Y]) + 2\Re(tr(\lambda_Y^*\eta_Y Y^*\xi_Y + Y^*\eta_Y\lambda_Y^*\xi_Y + \lambda_Y^* Y\eta_Y^*\xi_Y + Y^*\lambda_Y\eta_Y^*\xi_Y)) \\
& -\tilde{g}_Y(\mathrm{D}\,\xi(Y)[\eta_Y],\lambda_Y) - \tilde{g}_Y(\xi_Y, \mathrm{D}\,\lambda(Y)[\eta_Y]) - 2\Re(tr(\eta_Y^*\xi_Y Y^*\lambda_Y + Y^*\xi_Y\eta_Y^*\lambda_Y + \eta_Y^* Y\xi_Y^*\lambda_Y + Y^*\eta_Y\xi_Y^*\lambda_Y)) \\
& -\tilde{g}_Y(\xi_Y, \mathrm{D}\,\eta(Y)[\lambda_Y]) + \tilde{g}_Y(\xi_Y, \mathrm{D}\,\lambda(Y)[\eta_Y]) \\
& +\tilde{g}_Y(\lambda_Y, \mathrm{D}\,\xi(Y)[\eta_Y]) - \tilde{g}_Y(\lambda_Y, \mathrm{D}\,\eta(Y)[\xi_Y]) \\
& +\tilde{g}_Y(\eta_Y, \mathrm{D}\,\lambda(Y)[\xi_Y]) - \tilde{g}_Y(\eta_Y, \mathrm{D}\,\xi(Y)[\lambda_Y]) \\
= \; & 2\tilde{g}_Y(\mathrm{D}\,\lambda(Y)[\xi_Y],\eta_Y) + 4\Re(tr(Y^*\xi_Y\lambda_Y^*\eta_Y + Y^*\lambda_Y\xi_Y^*\eta_Y)).
\end{aligned}$$

It follows that

$$g_Y^3(\nabla_{\xi_Y}\lambda, \eta_Y) = \tilde{g}_Y(\mathrm{D}\,\lambda(Y)[\xi_Y],\eta_Y) + 2\Re(tr(Y^*\xi_Y\lambda_Y^*\eta_Y + Y^*\lambda_Y\xi^*\eta_Y)). \tag{141}$$

By definition, we have Hess $F(Y)[\xi_Y] = \nabla_{\xi_Y}\operatorname{grad} F$. By Lemma (C.1), it suffices to assume that ξ_Y is a horizontal vector in order to obtain the Hessian operator of h. Therefore,

$$\begin{aligned}
& g_Y^3(\text{Hess}\, F(Y)[\xi_Y], \eta_Y) && (142)\\
= \; & g_Y^3(\nabla_{\xi_Y}\operatorname{grad} F, \eta_Y) && (143)\\
= \; & \tilde{g}(\eta_Y, \mathrm{D}\operatorname{grad} F(Y)[\xi_Y]) + 2\Re(tr(Y^*\xi_Y \operatorname{grad} F(Y)^*\eta_Y + Y^*\operatorname{grad} F(Y)\xi_Y^*\eta_Y)) && (144)\\
= \; & \tilde{g}(\eta_Y, \mathrm{D}\operatorname{grad} F(Y)[\xi_Y]) + \Re(tr((Y\eta_Y^* + \eta_Y Y^*)(\operatorname{grad} F(Y)\xi_Y^* + \xi_Y \operatorname{grad} F(Y)^*))) && (145)\\
= \; & \tilde{g}(\eta_Y, \mathrm{D}\operatorname{grad} F(Y)[\xi_Y]) && (146)\\
& +\tilde{g}\left(\eta_Y, \left(I - \frac{1}{2}P_Y\right)(\operatorname{grad} F(Y)\xi_Y^* + \xi_Y \operatorname{grad} F(Y)^*)Y(Y^*Y)^{-1}\right). && (147)
\end{aligned}$$

Recall that for Riemannian metric g^3, we have $\operatorname{grad} F(Y) = \left(I - \frac{1}{2}P_Y\right) Herm\nabla f(Y^*Y)Y(Y^*Y)^{-1}$. Hence

$$\begin{aligned}
& \mathrm{D}\operatorname{grad} F(Y)[\xi_Y] && (148)\\
= \; & \left(I - \frac{1}{2}P_Y\right) Herm\{\nabla^2 f(Y^*Y)[Y\xi_Y^* + \xi_Y Y^*]\}Y(Y^*Y)^{-1} && (149)\\
& -\frac{1}{2}(\mathrm{D}\,(P_Y)[\xi_Y])Herm(\nabla f(Y^*Y))Y(Y^*Y)^{-1} && (150)\\
& +\left(I - \frac{1}{2}P_Y\right) Herm(\nabla f(Y^*Y))\mathrm{D}\,(Y(Y^*Y)^{-1})[\xi_Y], && (151)
\end{aligned}$$

where we have

$$\begin{aligned}
& \mathrm{D}\,(P_Y)[\xi_Y] = \mathrm{D}\,(Y(Y^*Y)^{-1}Y^*)[\xi_Y] && (152)\\
= \; & \xi_Y(Y^*Y)^{-1}Y^* - Y(Y^*Y)^{-1}(\xi_Y^*Y + Y^*\xi_Y)(Y^*Y)^{-1}Y^* + Y(Y^*Y)^{-1}\xi_Y^* && (153)
\end{aligned}$$

and

$$\mathrm{D}\,(Y(Y^*Y)^{-1})[\xi_Y] = \xi_Y(Y^*Y)^{-1} - Y(Y^*Y)^{-1}(\xi_Y^*Y + Y^*\xi_Y)(Y^*Y)^{-1}. \tag{154}$$

Combining these equations we have

$$
\begin{aligned}
& g^3_Y(\text{Hess}\, F(Y)[\xi_Y], \eta_Y) \\
= \;& \tilde{g}\left(\eta_Y, \left(I - \frac{1}{2}P_Y\right) Herm\{\nabla^2 f(Y^*Y)[Y\xi_Y^* + \xi_Y Y^*]\} Y(Y^*Y)^{-1}\right) \\
& -\tilde{g}\left(\eta_Y, \frac{1}{2}(\xi_Y(Y^*Y)^{-1}Y^* - Y(Y^*Y)^{-1}(\xi_Y^*Y + Y^*\xi_Y)(Y^*Y)^{-1}Y^* + Y(Y^*Y)^{-1}\xi_Y^*) Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\right) \\
& +\tilde{g}\left(\eta_Y, \left(I - \frac{1}{2}P_Y\right) Herm(\nabla f(YY^*)) \left(\xi_Y(Y^*Y)^{-1} - Y(Y^*Y)^{-1}(\xi_Y^*Y + Y^*\xi_Y)(Y^*Y)^{-1}\right)\right) \\
& +\tilde{g}\left(\eta_Y, \left(I - \frac{1}{2}P_Y\right)\left(\left(I - \frac{1}{2}P_Y\right) Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\xi_Y^*\right.\right. \\
& \left.\left. +\; \xi_Y(Y^*Y)^{-1}Y^* Herm(\nabla f(YY^*))\left(I - \frac{1}{2}P_Y\right)\right) Y(Y^*Y)^{-1}\right) \\
= \;& \tilde{g}\left(\eta_Y, \left(I - \frac{1}{2}P_Y\right) Herm\{\nabla^2 f(Y^*Y)[Y\xi_Y^* + \xi_Y Y^*]\} Y(Y^*Y)^{-1}\right) \\
& -\tilde{g}\left(\eta_Y, \frac{1}{2}(\xi_Y(Y^*Y)^{-1}Y^* - Y(Y^*Y)^{-1}(\xi_Y^*Y + Y^*\xi_Y)(Y^*Y)^{-1}Y^* + Y(Y^*Y)^{-1}\xi_Y^*) Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\right) \\
& +\tilde{g}\left(\eta_Y, \left(I - \frac{1}{2}P_Y\right) Herm(\nabla f(YY^*)) \left(\xi_Y(Y^*Y)^{-1} - Y(Y^*Y)^{-1}(\xi_Y^*Y + Y^*\xi_Y)(Y^*Y)^{-1}\right)\right) \\
& +\tilde{g}\left(\eta_Y, \left(I - \frac{3}{4}P_Y\right) Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\xi_Y^* Y(Y^*Y)^{-1}\right) \\
& +\tilde{g}\left(\eta_Y, \frac{1}{2}\left(I - \frac{1}{2}P_Y\right)\xi_Y(Y^*Y)^{-1}Y^* Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\right) \\
= \;& \tilde{g}\left(\eta_Y, \left(I - \frac{1}{2}P_Y\right) Herm\{\nabla^2 f(Y^*Y)[Y\xi_Y^* + \xi_Y Y^*]\} Y(Y^*Y)^{-1}\right) \\
& -\tilde{g}\left(\eta_Y, \frac{1}{2}\xi_Y(Y^*Y)^{-1}Y^* Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\right) \\
& -\tilde{g}\left(\eta_Y, \frac{1}{2}Y(Y^*Y)^{-1}\xi_Y^* Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\right) \\
& +\tilde{g}\left(\eta_Y, \frac{1}{2}Y(Y^*Y)^{-1}\xi_Y^* P_Y Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\right) \\
& +\tilde{g}\left(\eta_Y, \frac{1}{2}P_Y\xi_Y(Y^*Y)^{-1}Y^* Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\right) \\
& +\tilde{g}\left(\eta_Y, \left(I - \frac{1}{2}P_Y\right) Herm(\nabla f(YY^*)) \left((I - P_Y)\xi_Y(Y^*Y)^{-1} - Y(Y^*Y)^{-1}\xi_Y^* Y(Y^*Y)^{-1}\right)\right) \\
& +\tilde{g}\left(\eta_Y, \left(I - \frac{1}{2}P_Y\right) Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\xi_Y^* Y(Y^*Y)^{-1} - \frac{1}{4}P_Y Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\xi_Y^* Y(Y^*Y)^{-1}\right) \\
& +\tilde{g}\left(\eta_Y, \frac{1}{2}(I - P_Y)\xi_Y Y(Y^*Y)^{-1}Y^* Herm(\nabla f(YY^*))Y(Y^*Y)^{-1} + \frac{1}{4}P_Y\xi_Y(Y^*Y)^{-1}Y^* Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\right) \\
= \;& \tilde{g}\left(\eta_Y, \left(I - \frac{1}{2}P_Y\right) Herm\{\nabla^2 f(Y^*Y)[Y\xi_Y^* + \xi_Y Y^*]\} Y(Y^*Y)^{-1}\right) \\
& +\tilde{g}\left(\eta_Y, (I - P_Y) Herm(\nabla f(YY^*))(I - P_Y)\xi_Y(Y^*Y)^{-1}\right) \\
& +\tilde{g}\left(\eta_Y, \frac{1}{2}Y skew\left((Y^*Y)^{-1}Y\xi_Y(Y^*Y)^{-1}Y^* Herm(\nabla f(YY^*))Y(Y^*Y)^{-1}\right)\right) \\
& +\tilde{g}\left(\eta_Y, Y skew\left((Y^*Y)^{-1}Y^* Herm(\nabla f(YY^*))(I - P_Y)\xi_Y(Y^*Y)^{-1}\right)\right) \\
= \;& \tilde{g}\left(\eta_Y, \left(I - \frac{1}{2}P_Y\right) Herm\{\nabla^2 f(Y^*Y)[Y\xi_Y^* + \xi_Y Y^*]\} Y(Y^*Y)^{-1}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\tilde{g}\left(\eta_Y, (I-P_Y)Herm(\nabla f(YY^*))(I-P_Y)\xi_Y(Y^*Y)^{-1}\right)\\
= \;& g_Y^3\left(\eta_Y, \left(I-\frac{1}{2}P_Y\right)Herm\{\nabla^2 f(Y^*Y)[Y\xi_Y^* + \xi_Y Y^*]\}Y(Y^*Y)^{-1}\right.\\
&\left.+(I-P_Y)Herm(\nabla f(YY^*))(I-P_Y)\xi_Y(Y^*Y)^{-1}\right)
\end{aligned}
$$

Hence for $\xi_Y \in \mathcal{H}_Y$, we have

$$
\begin{aligned}
&\text{Hess}\, F(Y)[\xi_Y] && (155)\\
= \;& \left(I-\frac{1}{2}P_Y\right)Herm\{\nabla^2 f(Y^*Y)[Y\xi_Y^* + \xi_Y Y^*]\}Y(Y^*Y)^{-1} && (156)\\
&+(I-P_Y)Herm(\nabla f(YY^*))(I-P_Y)\xi_Y(Y^*Y)^{-1} && (157)
\end{aligned}
$$

To summarize, we obtain

$$
\begin{aligned}
\overline{\left(\text{Hess}\, h(\pi(Y))[\eta_{\pi(Y)}]\right)}_Y &= P_Y^{\mathcal{H}^3}(\text{Hess}\, F(Y)[\overline{\xi}_Y])\\
&= \left(I-\frac{1}{2}P_Y\right)Herm\{\nabla^2 f(Y^*Y)[Y\overline{\xi}_Y^* + \overline{\xi}_Y Y^*]\}Y(Y^*Y)^{-1}\\
&\quad +(I-P_Y)Herm(\nabla f(YY^*))(I-P_Y)\overline{\xi}_Y(Y^*Y)^{-1}.
\end{aligned}
$$

References

[1] Estelle Massart and P.-A. Absil. Quotient Geometry with Simple Geodesics for the Manifold of Fixed-Rank Positive-Semidefinite Matrices. *SIAM Journal on Matrix Analysis and Applications*, 41(1):171–198, January 2020.

[2] B. Vandereycken, P.-A. Absil, and S. Vandewalle. A Riemannian geometry with complete geodesics for the set of positive semidefinite matrices of fixed rank. *IMA Journal of Numerical Analysis*, 33(2):481–514, April 2013.

[3] Silvère Bonnabel, Gilles Meyer, and Rodolphe Sepulchre. Adaptive filtering for estimation of a low-rank positive semidefinite matrix. In *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems*, 2010.

[4] Bart Vandereycken and Stefan Vandewalle. A Riemannian Optimization Approach for Computing Low-Rank Solutions of Lyapunov Equations. *SIAM Journal on Matrix Analysis and Applications*, 31(5):2553–2579, January 2010.

[5] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[6] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[7] Vincent Jugnon and Laurent Demanet. Interferometric inversion: a robust approach to linear inverse problems. In *2013 SEG Annual Meeting*. OnePetro, 2013.

[8] Laurent Demanet and Vincent Jugnon. Convex recovery from interferometric measurements. *IEEE Transactions on Computational Imaging*, 3(2):282–295, 2017.

[9] Samuel Burer and Renato D.C. Monteiro. Local Minima and Convergence in Low-Rank Semidefinite Programming. *Mathematical Programming*, 103(3):427–444, July 2005.

[10] Bart Vandereycken. Low-Rank Matrix Completion by Riemannian Optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, January 2013.

[11] Daniel Kressner, Michael Steinlechner, and Bart Vandereycken. Low-rank tensor completion by Riemannian optimization. *BIT Numerical Mathematics*, 54(2):447–468, June 2014.

[12] Bart Vandereycken, P.-A. Absil, and Stefan Vandewalle. Embedded geometry of the set of symmetric positive semidefinite matrices of fixed rank. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 389–392, Cardiff, United Kingdom, August 2009. IEEE.

[13] Wen Huang, K. A. Gallivan, and Xiangxiong Zhang. Solving PhaseLift by Low-Rank Riemannian Optimization Methods for Complex Semidefinite Constraints. *SIAM Journal on Scientific Computing*, 39(5):B840–B859, January 2017.

[14] Nicolas Boumal, Vladislav Voroninski, and Afonso S Bandeira. Deterministic Guarantees for Burer-Monteiro Factorizations of Smooth Semidefinite Programs. *Communications on Pure and Applied Mathematics*, 73(3):581–608, 2020.

[15] Silvère Bonnabel and Rodolphe Sepulchre. Riemannian Metric and Geometric Mean for Positive Semidefinite Matrices of Fixed Rank. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1055–1070, January 2010. Publisher: Society for Industrial and Applied Mathematics.

[16] M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre. Low-Rank Optimization on the Cone of Positive Semidefinite Matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, January 2010.

[17] Wen Huang. *Optimization algorithms on Riemannian manifolds with applications*. PhD thesis, The Florida State University, 2013.

[18] P.-A. Absil, M. Ishteva, L. De Lathauwer, and S. Van Huffel. A geometric Newton method for Oja's vector field. *Neural Computation*, 21(5):1415–1433, May 2009. arXiv: 0804.0989.

[19] Bamdev Mishra. *A Riemannian approach to large-scale constrained least-squares with symmetries*. PhD thesis, Universite de Liege, Liege, Belgique, 2014.

[20] Uwe Helmke and Mark A. Shayman. Critical points of matrix least squares distance functions. *Linear Algebra and its Applications*, 215:1–19, January 1995.

[21] Uwe Helmke and John B. Moore. *Optimization and Dynamical Systems*. Springer Science & Business Media, December 2012. Google-Books-ID: yR7vBwAAQBAJ.

[22] Michel Coste. AN INTRODUCTION TO SEMIALGEBRAIC GEOMETRY. *Citeseer*, page 26, 2000.

[23] Christopher G. Gibson. *Singular points of smooth mappings*. Number 25 in Research notes in mathematics. Pitman, London ; San Francisco, 1979.

[24] John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer New York, New York, NY, 2012.

[25] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, N.J. ; Woodstock, 2008. OCLC: ocn174129993.

[26] P.-A. Absil and Jérôme Malick. Projection-like Retractions on Matrix Manifolds. *SIAM Journal on Optimization*, 22(1):135–158, January 2012.

[27] W. Huang, P.-A. Absil, and K. A. Gallivan. Intrinsic representation of tangent vectors and vector transport on matrix manifolds. *Numerische Mathematik*, 136(2):523–543, 2017.

[28] Bart Vandereycken. Low-rank matrix completion by Riemannian optimization—extended version. *arXiv:1209.3834 [math]*, September 2012. arXiv: 1209.3834.

[29] F. Brickell and R. S. Clark. *Differentiable Manifolds: an introduction*. Van Nostrand Reinhold, 1970. Google-Books-ID: 25EJNQEACAAJ.

[30] Wen Huang, K. A. Gallivan, and P.-A. Absil. A Broyden Class of Quasi-Newton Methods for Riemannian Optimization. *SIAM Journal on Optimization*, 25(3):1660–1685, January 2015.

[31] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.