RIEMANNIAN OPTIMIZATION USING THREE DIFFERENT METRICS FOR HERMITIAN PSD FIXED-RANK CONSTRAINTS*

SHIXIN ZHENG†, WEN HUANG‡, BART VANDEREYCKEN§, AND XIANGXIONG ZHANG¶

Abstract. For optimization under a Hermitian positive semidefinite fixed-rank constraint, we consider three approaches including the simple Burer–Monteiro method, Riemannian optimization over a quotient manifold, and the embedded manifold, all of which can be represented via quotient geometry with three Riemannian metrics $g^i(\cdot,\cdot)$ $(i = 1, 2, 3)$. By taking the nonlinear conjugate gradient method (CG) as an example, we show that CG in the factor-based Burer–Monteiro approach is equivalent to Riemannian CG on the quotient geometry with the Bures-Wasserstein metric g^1. Riemannian CG on the quotient geometry with the metric g^3 is equivalent to Riemannian CG on the embedded geometry. For comparing the three approaches, we analyze the condition number of the Riemannian Hessian near the minimizer. Under certain assumptions, the condition number from the Bures-Wasserstein metric g^1 is significantly different from the other two metrics. Numerical tests show that the Burer–Monteiro CG method has a slower asymptotic convergence rate if the minimizer is rank deficient, which is consistent with the condition number analysis.

Key words. Riemannian optimization, Hermitian PSD fixed-rank matrices, embedded manifold, quotient manifold, Burer–Monteiro, conjugate gradient, Riemannian Hessian, Bures-Wasserstein

MSC codes. 65K05, 49Q99, 53B20, 65F55, 90C30

1. Introduction.

1.1. The Hermitian PSD low-rank constraints.

We are interested in methods for minimization with a positive semidefinite (PSD) low-rank constraint

$$\min_X f(X), \quad X \in \mathcal{H}_+^{n,p}, \tag{1.1}$$

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 (1.1) is often used for approximating solutions to a minimization with a convex PSD constraint:

$$\min_X f(X), \quad X \in \mathbb{C}^{n\times n}, X \succcurlyeq 0. \tag{1.2}$$

PSD constraints arise in semidefinite programming. If the solution of (1.2) is low rank, it is preferable to consider a low-rank representation of PSD matrices, e.g., real symmetric PSD fixed-rank matrices were used in [4, 28]. Since $X \in \mathcal{H}_+^{n,p}$ has a low-rank structure, its low-rank compact form has the complexity $O(np^2)$, which is smaller than the $O(n^2)$ storage when using $X \in \mathbb{C}^{n\times n}$. For many problems such as the PhaseLift problem [9, 8] and the interferometry recovery problem [18, 10], solving (1.1) can lead to a good approximate solution to (1.2) with compact storage and cost.

---

*

Funding: S.Z. and X.Z. are supported by NSF DMS-2208518. W.H. is partially supported by National Natural Science Foundation of China (No. 12001455). B.V. is partially supported by the Swiss National Science Foundation (grant 178752).

†Department of Mathematics, Purdue University, West Lafayette, USA (zheng513@purdue.edu).

‡Corresponding author, School of Mathematical Sciences, Xiamen University, Xiamen, China (wen.huang@xmu.edu.cn).

§Section of Mathematics, University of Geneva, Switzerland (bart.vandereycken@unige.ch).

¶Corresponding author, Department of Mathematics, Purdue University, West Lafayette, USA (zhan1966@purdue.edu).

1.2. The real inner product and induced gradient. Since $f(X)$ is real-valued, $f(X)$ does not have a complex derivative. 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 $\langle\cdot,\cdot\rangle_{\mathcal{E}}$. The Hilbert–Schmidt inner product for $\mathbb{R}^{m\times n}$ is $\langle A,B\rangle_{\mathbb{R}^{m\times n}} = \mathrm{tr}(A^TB)$. Let $\Re(A)$ and $\Im(B)$ represent the real and imaginary parts of $A\in\mathbb{C}^{m\times n}$. The real inner product for the real vector space $\mathbb{C}^{m\times n}$ is

$$\langle A,B\rangle_{\mathbb{C}^{m\times n}} := \Re(\mathrm{tr}(A^*B)), \tag{1.3}$$

where * denotes the conjugate transpose. The gradient of $f(X)$ w.r.t (1.3) is

$$\nabla f(X) = \frac{\partial f(X)}{\partial \Re(X)} + \mathbb{i}\frac{\partial f(X)}{\partial \Im(X)} \in \mathbb{C}^{m\times n}. \tag{1.4}$$

See [29] for a derivation of (1.4). For $f(X) = \frac{1}{2}\|\mathcal{A}(X)-b\|_F^2$ with a linear operator $\mathcal{A}$, (1.4) becomes $\nabla f(X) = \mathcal{A}^*(\mathcal{A}(X)-b)$, where $\mathcal{A}^*$ is the adjoint operator of $\mathcal{A}$.

1.3. Three different methodologies. We consider three methods for (1.1). The first approach, often called the Burer–Monteiro method [7, 6], is to solve

$$\min_{Y\in\mathbb{C}^{n\times p}} F(Y) := f(YY^*). \tag{1.5}$$

The gradient descent (GD) method is $Y_{k+1} = Y_k - \tau\nabla F(Y_k) = Y_k - \tau 2\nabla f(Y_kY_k^*)Y_k$, which is one of the simplest low-rank algorithms. The nonlinear conjugate gradient (CG) and quasi-Newton type methods, like L-BFGS [10], can also be easily used for (1.5). It is not clear in what sense it converges since $F(Y) = F(YO)$ for any $O\in\mathcal{O}_p$, where $\mathcal{O}_p$ denotes the set of unitary matrices of size $p\times p$.

To remove the ambiguity from $\mathcal{O}_p$, it is natural to consider the quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$, see [5, 17, 21, 13, 16], where $\mathbb{C}_*^{n\times p} = \{X\in\mathbb{C}^{n\times p} : \mathrm{rank}(X) = p\}$ denotes the noncompact Stiefel manifold.

Another natural approach is to consider Riemannian optimization algorithms on $\mathcal{H}_+^{n,p}$ as an embedded manifold in the Euclidean space $\mathbb{C}^{n\times n}$ [26, 25, 19]. We shall regard $\mathcal{H}_+^{n,p}\subset\mathbb{C}^{n\times n}$ as a manifold over $\mathbb{R}$ since $f(X)$ is real-valued.

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? For comparison, we 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.

It is common to explore different metrics in Riemannian optimization [1, 27, 23]. For any $Y\in\mathbb{C}_*^{n\times p}$, $A,B\in\mathbb{C}^{n\times p}$, we consider metrics $g_Y^i(\cdot,\cdot)$ for the total space $\mathbb{C}_*^{n\times p}$:

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

where $Skew(X) = (X-X^*)/2$. We have three metrics g^i for the quotient manifold induced from the submersion $\mathbb{C}_*^{n\times p} \longrightarrow \mathbb{C}_*^{n\times p}/\mathcal{O}_p$. The first metric is the Bures-Wasserstein metric [22, 21], the second metric is used in [16], and the embedded manifold approach corresponds to the third metric.

We will prove that the GD and CG methods for solving (1.5) are exactly equivalent to the Riemannian GD and CG methods on $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^1)$ with a specific vector transport. We will also prove that GD and the CG methods using the embedded geometry of $\mathcal{H}_+^{n,p}$ are equivalent to GD and CG methods on $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^3)$.

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. We will analyze and compare the condition numbers of the Riemannian Hessian using these three different metrics by estimating their Rayleigh quotient.

1.5. Contributions and 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.

The first major contribution is the equivalence between the CG method for (1.5) and the CG method on $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^1)$ for solving (1.1). Thus the convergence of the simple Burer–Monteiro approach can be understood in the context of Riemannian optimization on the quotient manifold with the Bures-Wasserstein metric.

In Section 6, we analyze the condition number of the Riemannian Hessian on the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^i)$ near the minimizer, which is another contribution. Our analysis is also consistent with empirical observation of the performance of different methods in numerical tests in Section 7. Section 8 are concluding remarks.

2. Notation. 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 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 $\operatorname{diag}(M)$ be the n-by-1 vector that is the diagonal of the n-by-n matrix M. Given a vector v, $\operatorname{Diag}(v)$ is a square matrix with its ith diagonal entry equal to v_i. Given a matrix A, $\operatorname{tr}(A)$ 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 singular values. The compact singular value decomposition (SVD) of X is denoted by $X = U\Sigma U^*$ and $\Sigma = \operatorname{Diag}(\sigma)$ with singular values $\sigma_1 \geq \cdots \geq \sigma_p > 0$.

In this paper, all manifolds of complex matrices are viewed as manifolds over $\mathbb{R}$. Given a Euclidean space $\mathcal{E}$, the inner product on $\mathcal{E}$ is denoted by $\langle .,. \rangle_{\mathcal{E}}$. Specifically, $\langle A, B\rangle_{\mathbb{R}^{m\times n}} = \operatorname{tr}(A^T B)$ for $A, B \in \mathbb{R}^{m\times n}$ and $\langle A, B\rangle_{\mathbb{C}^{m\times n}} = \Re(\operatorname{tr}(A^* B))$ for $A, B \in \mathbb{C}^{m\times n}$ denote 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 $\mathcal{S}_+^{n,p} = \{X \in \mathbb{R}^{n\times n} : X \succcurlyeq 0, \operatorname{rank}(X) = p\}$ in [26]. Such an extension is not entirely obvious since $\mathcal{H}_+^{n,p}$ is treated as a real manifold and (1.3) is not the complex Hilbert–Schmidt inner product. Nonetheless, all proofs can be done following [26], thus we only state the results. Omitted proofs can be found in [29].

3.1. Tangent space. First we show that $\mathcal{H}_+^{n,p}$ is a smooth embedded submanifold of $\mathbb{C}^{n\times n}$ following the case of $\mathcal{S}_+^{n,p}$ in [26, Prop. 2.1], [12, Prop. 2.1] and [11, Chap. 5]. The tangent space of $\mathcal{H}^{n,p}$ follows the argument in [25, Proposition 2.1].

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 $2np - p^2$.*

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

$$T_X\mathcal{H}_+^{n,p} = \left\{ \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} H & K^* \\ K & 0 \end{bmatrix} \begin{bmatrix} U^* \\ U_\perp^* \end{bmatrix}, \quad H = H^* \in \mathbb{C}^{p\times p}, K \in \mathbb{C}^{(n-p)\times p} \right\}.$$

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 $\mathbb{C}^{n\times n}$,

$$g_X(\zeta_1, \zeta_2) = \langle \zeta_1, \zeta_2 \rangle_{\mathbb{C}^{n\times n}} = \Re(\mathrm{tr}(\zeta_1^*\zeta_2)), \quad \zeta_1, \zeta_2 \in T_X\mathcal{H}_+^{n,p}. \tag{3.1}$$

The *Riemannian gradient* of f at X is the projection of $\nabla f(X)$ onto $T_X\mathcal{H}_+^{n,p}$ [2]:

$$\mathrm{grad}\, f(X) = P_X^t(\nabla f(X)),$$

where P_X^t is the orthogonal projection onto $T_X\mathcal{H}_+^{n,p}$, given by the following theorem.

THEOREM 3.3. *Let $X = YY^* = U\Sigma U^*$ be the compact SVD for $X \in \mathcal{H}_+^{n,p}$ with $Y \in \mathbb{C}_*^{n\times p}$. For a complex matrix Z, the orthogonal projection onto $T_X\mathcal{H}_+^{n,p}$ is*

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

REMARK 3.4. *We can write $P_X^t = P_X^s + P_X^p$ by introducing the two operators*

$$P_X^s : Z \mapsto P_U \frac{Z+Z^*}{2} P_U, \quad P_X^p : Z \mapsto P_{U_\perp} \frac{Z+Z^*}{2} P_U + P_U \frac{Z+Z^*}{2} P_{U_\perp}, \tag{3.2}$$

where $P_U = UU^*$ and $P_{U_\perp} = U_\perp U_\perp^*$.

3.3. A retraction by projection to the embedded manifold. A retraction is essentially a first-order approximation to the exponential map; see [2, Def. 4.1.1]. By [3, Props. 3.2 and 3.3], the truncated SVD $R_X(Z) := P_{\mathcal{H}_+^{n,p}}(X+Z) = \sum_{i=1}^p \sigma_i(X+Z)v_i v_i^*$ is a retraction on $\mathcal{H}_+^{n,p}$, where v_i is the singular vector of $X+Z$ corresponding to the ith largest singular value $\sigma_i(X+Z)$. We remark that such a retraction can be compactly implemented, see Section 5 and [29] for implementation details.

3.4. Vector transport. A vector transport is a mapping that transports a tangent vector from one tangent space to another tangent space. See [2, Def. 8.1.1]. The vector transport of $\mathcal{H}_+^{n,p}$ that we use is derived from the vector transport by projection. Let $\xi_X, \eta_X \in T_X\mathcal{H}_+^{n,p}$ and let R be a retraction on $\mathcal{H}_+^{n,p}$. By [2, section 8.1.3], the projection of one tangent vector onto another tangent space is a vector transport:

$$\mathcal{T}_{\eta_X}\xi_X := P_{R_X(\eta_X)}^t \xi_X, \tag{3.3}$$

where P_Z^t is the projection operator onto $T_Z\mathcal{H}_+^{n,p}$ with $Z = R_X(\eta_X)$. Namely, we first apply the retraction R_X to η_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 (3.3) 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 $\nu_1 = \begin{bmatrix} U_1 & U_{1\perp} \end{bmatrix} \begin{bmatrix} H_1 & K_1^* \\ K_1 & 0 \end{bmatrix} \begin{bmatrix} U_1^* \\ U_{1\perp}^* \end{bmatrix} = U_1H_1U_1^* + U_{1\perp}K_1U_1^* + U_1K_1^*U_{1\perp}^* \in T_{X_1}\mathcal{H}_+^{n,p}$, we need to determine H_2 and K_2 of the transported tangent vector $\nu_2 =$

$\begin{bmatrix} U_2 & U_{2\perp} \end{bmatrix} \begin{bmatrix} H_2 & K_2^* \\ K_2 & 0 \end{bmatrix} \begin{bmatrix} U_2^* \\ U_{2\perp}^* \end{bmatrix} \in T_{X_2}\mathcal{H}_+^{n,p}$. By the projection formula (3.2), we have $\nu_2 = P_{X_2}^t(\nu_1) = \begin{bmatrix} U_2 & U_{2\perp} \end{bmatrix} \begin{bmatrix} U_2^*\nu_1 U_2 & U_2^*\nu_1 U_{2\perp} \\ U_{2\perp}^*\nu_1 U_2 & 0 \end{bmatrix} \begin{bmatrix} U_2^* \\ U_{2\perp}^* \end{bmatrix}$, where

$$H_2 = U_2^*\nu_1 U_2 = U_2^* U_1 H_1 U_1^* U_2 + U_2^* U_{1\perp} K_1 U_1^* U_2 + U_2^* U_1 K_1^* U_{1\perp}^* U_2, \text{ and}$$
$$K_2 = U_{2\perp}^*\nu_1 U_2 = U_{2\perp}^* U_1 H_1 U_1^* U_2 + U_{2\perp}^* U_{1\perp} K_1 U_1^* U_2 + U_{2\perp}^* U_1 K_1^* U_{1\perp}^* U_2.$$

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

$$H_2 = U_2^* U_1 H_1 U_1^* U_2, \quad K_2 = U_{2\perp}^* U_{1\perp} K_1 U_1^* U_2. \tag{3.4a}$$

One can verify that (3.4) is a vector transport by parallelization in [14]. In numerical tests, we have observed that the nonlinear CG 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 (3.3). 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 $\xi_{X_1} = \begin{bmatrix} U_1 & U_{1\perp} \end{bmatrix} \begin{bmatrix} H_1 & K_1^* \\ K_1 & 0 \end{bmatrix} \begin{bmatrix} U_1^* \\ U_{1\perp}^* \end{bmatrix} \in T_{X_1}\mathcal{H}_+^{n,p}$, then compute
$$\mathcal{T}_{\eta_{X_1}}\xi_{X_1} = \begin{bmatrix} U_2 & U_{2\perp} \end{bmatrix} \begin{bmatrix} H_2 & K_2^* \\ K_2 & 0 \end{bmatrix} \begin{bmatrix} U_2^* \\ U_{2\perp}^* \end{bmatrix} \in T_{X_2}\mathcal{H}_+^{n,p}. \tag{3.4b}$$

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 w.r.t (1.3), see [29]. The *Riemannian Hessian* (see [2, definition 5.5.1]) of f at X, is denoted by $\operatorname{Hess} f(X)$, where f is viewed as a function on the manifold $\mathcal{H}_+^{n,p}$ with metric (3.1).

The following proposition gives the Riemannian Hessian of f. The proof follows similar ideas as in [28, Prop. 5.10] and [24, Prop. 2.3]. We leave the outline of the proof in Appendix A.1.

PROPOSITION 3.5. *Let $f(X)$ be a real-valued function defined on $\mathcal{H}_+^{n,p}$ with metric* (3.1). *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*

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

where $\cdot^\dagger$ denotes the pseudo-inverse operator, $\xi_X^s = P_X^s(\xi_X)$, $\xi_X^p = P_X^p(\xi_X)$, and P_X^t and P_X^p are defined in (3.2).

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 $\mathcal{H}_+^{n\times p}$ is diffeomorphic to $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ as will be shown below. The smooth Lie group action of $\mathcal{O}_p$ on $\mathbb{C}_*^{n\times p}$ 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 $Y_1 = Y_2 O$. The set $\mathbb{C}_*^{n\times p}$ is called the *total space* of $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$.

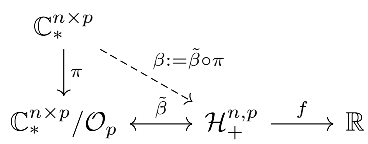
Denote the natural projection as

$$\pi : \mathbb{C}_*^{n\times p} \to \mathbb{C}_*^{n\times p}/\mathcal{O}_p.$$

The equivalence class of Y is denoted as $[Y] = \pi^{-1}(\pi(Y)) = \{YO | O \in \mathcal{O}_p\}$. Define $h(\pi(Y)) = f(YY^*)$, then (1.1) is equivalent to

$$\min_{\pi(Y)} h(\pi(Y)), \quad \pi(Y) \in \mathbb{C}_*^{n\times p}/\mathcal{O}_p. \tag{4.1}$$

Define a map $\beta : \mathbb{C}_*^{n\times p} \to \mathcal{H}_+^{n,p}$ with $\beta(Y) = YY^*$. 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 $\beta = \tilde{\beta} \circ \pi$ [2, Section 3.4.2]. One can easily check that $\tilde{\beta}$ is a bijection. For any f on $\mathcal{H}_+^{n,p}$, there is a function F defined on $\mathbb{C}_*^{n\times p}$ that induces f: for any $X = YY^* \in \mathcal{H}_+^{n,p}$, $F(Y) := f \circ \beta(Y) = f(YY^*)$, which is summarized in the diagram:



The next theorems follow from [20, Cor. 21.6; Thm. 21.10], and [21, Prop. A.7].

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

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

4.1. Vertical space, three Riemannian metrics, and horizontal spaces. The equivalence class $[Y]$ is an embedded submanifold of $\mathbb{C}_*^{n\times p}$ [2, Prop. 3.4.4]. Therefore, the tangent space of $[Y]$ at Y is 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 $Y \in [Y] = \{YO | O \in \mathcal{O}_p\}$, defined as the tangent space of $[Y]$ at Y, is $\mathcal{V}_Y = \{Y\Omega | \Omega^* = -\Omega, \Omega \in \mathbb{C}^{p\times p}\}$.*

With a *Riemannian metric* g of the total space $\mathbb{C}_*^{n\times p}$, we can define the orthogonal complement in $T_Y\mathbb{C}_*^{n\times p}$ of $\mathcal{V}_Y$. In other words, we choose the *horizontal distribution* as orthogonal complement w.r.t. Riemannian metric g, see [2, Section 3.5.8]. This orthogonal complement to $\mathcal{V}_Y$ is called *horizontal space* at Y and is denoted by $\mathcal{H}_Y$:

$$T_Y\mathbb{C}_*^{n\times p} = \mathcal{H}_Y \oplus \mathcal{V}_Y. \tag{4.2}$$

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.

There exist more than one choice of Riemannian metric on $\mathbb{C}_*^{n\times p}$. Metrics do not affect the vertical space but generally result in different horizontal spaces.

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

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

PROPOSITION 4.4. *Under metric g^1, the horizontal space at Y satisfies*

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

g^1 is also called the Bures-Wasserstein metric [22] for the quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$. One can show that g^1 is also consistent with the Bures-Wasserstein metric defined for Hermitian positive-definite matrices, see [29] for details.

4.1.2. The second quotient metric. A metric used in [16, 13] is defined by

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

PROPOSITION 4.5. *Under metric g^2, the horizontal space at Y satisfies*

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

4.1.3. The third quotient metric. The third metric for is induced by the diffeomorphism between $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ and the embedded geometry of $\mathcal{H}_+^{n,p}$. We first use the 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 $A = A^{\mathcal{V}} + A^{\mathcal{H}^2}$, $A^{\mathcal{V}} \in \mathcal{V}_Y$, $A^{\mathcal{H}^2} \in \mathcal{H}_Y^2$. Now define g^3 as

$$\begin{aligned} g_Y^3(A,B) &:= \left\langle \mathrm{D}\,\beta(Y)[A^{\mathcal{H}^2}], \mathrm{D}\,\beta(Y)[B^{\mathcal{H}^2}]\right\rangle_{\mathbb{C}^{n\times n}} + g_Y^2\left(A^{\mathcal{V}}, B^{\mathcal{V}}\right) \\ &= \langle YA^* + AY^*, YB^* + BY^*\rangle_{\mathbb{C}^{n\times n}} + \left\langle Y Skew\left((Y^*Y)^{-1}Y^*A\right)Y^*, Y Skew\left((Y^*Y)^{-1}Y^*B\right)Y^*\right\rangle_{\mathbb{C}^{n\times n}}. \end{aligned}$$

It is straightforward to verify that g^3 defined above is a Riemannian metric. With the definition (1.3), we have

$$\text{(4.3)} \quad \forall A,B \in A^{\mathcal{H}^2}, \quad g_Y^3(A,B) = \langle YA^* + AY^*, YB^* + BY^*\rangle_{\mathbb{C}^{n\times n}} = 2\langle AY^*Y + YA^*Y, B\rangle_{\mathbb{C}^{n\times p}}.$$

PROPOSITION 4.6. *Under metric g^3, the horizontal space at Y is the same as $\mathcal{H}_Y^2$:*

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

4.2. Projections onto vertical space and horizontal space. Due to the direct sum property (4.2), for $\mathcal{H}_Y^i$, there exist projection operators for any $A \in T_Y\mathbb{C}_*^{n\times p}$ to $\mathcal{H}_Y^i$ as $A = P_Y^{\mathcal{V}}(A) + P_Y^{\mathcal{H}^i}(A)$. We note that the operator $P_Y^{\mathcal{V}}$ depends on g^i but $\mathcal{V}$ is independent of g^i. It is straightforward to verify the following formulae.

PROPOSITION 4.7. *For g^1, $P_Y^{\mathcal{V}}(A) = Y\Omega$, $P_Y^{\mathcal{H}^1}(A) = A - Y\Omega$, where Ω is the skew-Hermitian matrix that solves the Lyapunov equation $\Omega Y^*Y + Y^*Y\Omega = Y^*A - A^*Y$. For g^2, we have $P_Y^{\mathcal{V}}(A) = Y Skew\left((Y^*Y)^{-1}Y^*A\right)$, and*

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

For g^3, we have $P_Y^{\mathcal{V}}(A) = Y Skew((Y^*Y)^{-1}Y^*A)$, and

$$P_Y^{\mathcal{H}^3}(A) = A - P_Y^{\mathcal{V}}(A) = Y Herm\left((Y^*Y)^{-1}Y^*A\right) + Y_\perp Y_\perp^* A.$$

4.3. $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ as a Riemannian quotient manifold. First, we show in the following lemma the relationship between the horizontal lifts of the quotient tangent vector $\xi_{\pi(Y)}$ lifted at different representatives in $[Y]$. A proof based on metric g^1 for $\mathcal{S}_+^{n,p}$ is given in [21, Prop. A.8], and [16, Lemma 5.1] proves the result for metric g^2. The proof for g^3 can be found in [29].

LEMMA 4.8. *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*

$$\bar{\eta}_{YO} = \bar{\eta}_Y O, \quad \forall O \in \mathcal{O}_p.$$

Recall from [2, Section 3.6.2] that 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

$$g_{\pi(Y)}(\xi_{\pi(Y)}, \zeta_{\pi(Y)}) := g_Y(\bar{\xi}_Y, \bar{\zeta}_Y) \tag{4.4}$$

defines a Riemannian metric on the quotient manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$. By Lemma 4.8, it is straightforward to verify that each Riemannian metric g^i on $\mathbb{C}_*^{n\times p}$ 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 (4.4) 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 $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$.

4.4. Riemannian gradient. Given a smooth real-valued function f on $\mathcal{H}_+^{n,p}$, recall that a corresponding cost function h is defined on $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ satisfying (4.1). The next theorem shows that the horizontal lift of $\operatorname{grad} h(\pi(Y))$ can be obtained from the Riemannian gradient of F. Its proof can be found in [2, Section 3.6.2].

THEOREM 4.9. *The horizontal lift of the gradient of h at $\pi(Y)$ is the Riemannian gradient of F at Y. That is,*

$$\overline{\operatorname{grad} h(\pi(Y))}_Y = \operatorname{grad} F(Y).$$

Therefore, $\operatorname{grad} F(Y)$ is always in $\mathcal{H}_Y$.

The next proposition summarizes the expression of $\operatorname{grad} F(Y)$ under different metrics. The proof is by simple calculation and definition of each metric, which can be found in [29].

PROPOSITION 4.10. *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*

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

where ∇f denotes the gradient (1.4) *and $P_Y = Y(Y^*Y)^{-1}Y^*$.*

4.5. Retraction. The retraction on $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ 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 $\tau > 0$,

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

is a retraction on $\mathbb{C}_*^{n\times p}$ if $Y + \tau A$ remains full rank, which is ensured for small enough τ. Lemma 4.8 indicates that $\overline{R}$ satisfies the conditions of [2, Prop. 4.1.3], implying

$$R_{\pi(Y)}(\tau\eta_{\pi(Y)}) := \pi(\overline{R}_Y(\tau\overline{\eta}_Y)) = \pi(Y + \tau\overline{\eta}_Y) \tag{4.5}$$

defines a retraction on the manifold $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ for a small step size $\tau > 0$.

4.6. Vector transport. A vector transport on $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ is projection to horizontal space (see [2, Section 8.1.2]):

$$\overline{\left(\mathcal{T}_{\eta_{\pi(Y)}}\xi_{\pi(Y)}\right)}_{Y+\overline{\eta}_Y} := P^{\mathcal{H}}_{Y+\overline{\eta}_Y}(\overline{\xi}_Y). \tag{4.6}$$

It can be shown that this vector transport is actually the differential of the retraction R defined in (4.5). Denote $Y_2 = Y_1 + \overline{\eta}_{Y_1}$. Base on the projection formula in Section 4.2, the explicit formula of (4.6) using different Riemannian metrics is then

$$\overline{\left(\mathcal{T}_{\eta_{\pi(Y_1)}}\xi_{\pi(Y_1)}\right)}_{Y_1+\overline{\eta}_{Y_1}} = \begin{cases} \overline{\xi}_{Y_1} - Y_2\Omega, & \text{for } g^1, \\ Y_2 Herm((Y_2^*Y_2)^{-1}Y_2^*\overline{\xi}_{Y_1}) + Y_{2\perp}Y_{2\perp}^*\overline{\xi}_{Y_1}, & \text{for } g^2 \text{ or } g^3. \end{cases}$$

4.7. Riemannian Hessian operator. Recall that the function h on $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ is defined in (4.1). The Riemannian Hessian of h under the three different metrics g^i can be given as follows. The proofs are given in Appendix B.1.

PROPOSITION 4.11. *Using g^1, the Riemannian Hession of h is given by*

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

PROPOSITION 4.12. *Using g^2, the Riemannian Hession of h is given by*

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

PROPOSITION 4.13. *Using g^3, the Riemannian Hession of h is given by*

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

5. The Riemannian conjugate gradient method. We only consider the Riemannian CG (RCG) described as Algorithm 1 in [25] with the geometric variant of Polak–Ribiére (PR+). Note that it is possible to explore other methods such as LRBFGS in [15]. We choose RCG since RCG is easier to implement and performs well on a wide variety of problems.

We focus on establishing two equivalences in algorithms. First, we show that the Burer–Monteiro CG method, i.e. CG solving (1.5), is equivalent to RCG on $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^1)$ with the retraction (4.5) and vector transport (4.6). Second, we show that RCG on the embedded manifold $\mathcal{H}_+^{n,p}$ is equivalent to RCG $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^3)$ with a specific retraction (5.3) and vector transport (5.4) given later.

Let $\mathcal{T}_{X_{k-1}\to X_k}$ denote a vector transport that maps from $T_{X_{k-1}}\mathcal{H}_+^{n,p}$ to $T_{X_k}\mathcal{H}_+^{n,p}$:

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

where R_X^{-1} exists locally for every $X \in \mathcal{H}_+^{n,p}$. Hence $\mathcal{T}_{X_{k-1}\to X_k}$ should be understood locally in the sense that X_{k-1} is sufficiently close to X_k (see [24, Section 2.4]). Similarly, $\mathcal{T}_{Y_{k-1}\to Y_k}$ denotes a vector transport that maps from $\mathcal{H}_{Y_{k-1}}$ to $\mathcal{H}_{Y_k}$:

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

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 summarize two RCG algorithms in Algorithm 5.1 and Algorithm 5.2 below. Algorithm 5.1 is the RCG on the embedded manifold for solving (1.1) and Algorithm 5.2 is the RCG on the quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^i)$ for solving (4.1). The explicit constants 0.0001 and 0.5 in the Armijo backtracking are chosen for convenience.

Algorithm 5.1 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 5.3
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 5.4

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

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

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

7: Obtain the new iterate by retraction
   $X_{k+1} = R_{X_k}(0.5^m t_k \eta_k)$ ▷ See Algorithm 5.5
8: **end for**

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

THEOREM 5.1. *Using retraction (4.5), vector transport (4.6) and metric g^1, Algorithm 5.2 is equivalent to the conjugate gradient method solving* (1.5) *in the sense that they produce exactly the same iterates if started from the same initial point.*

Proof. First of all, for g^1, the Riemannian gradient of F at Y is $\operatorname{grad} F(Y) = 2\nabla f(YY^*)Y$, which is equal to the gradient of $F(Y) = f(YY^*)$ at Y. Since vector transport is the orthogonal projection to the horizontal space, the β_k of PR_+ used in Riemannian CG becomes

$$\beta_k = \frac{g^1_{Y_k}\left(\operatorname{grad} F(Y_k), \operatorname{grad} F(Y_k) - P^{\mathcal{H}^1}_{Y_k}(\operatorname{grad} F(Y_{k-1}))\right)}{g^1_{Y_{k-1}}\left(\operatorname{grad} F(Y_{k-1}), \operatorname{grad} F(Y_{k-1})\right)}. \tag{5.1}$$

Now observe that

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

and g^1 is equivalent to the classical inner product for $\mathbb{C}^{n\times p}$. Hence β_k computed by (5.1) is equal to β_k of PR_+ in conjugate gradient for (1.5).

Algorithm 5.2 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 $\eta_0 = 0$

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

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

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

4: Compute a conjugate direction by PR_+ and vector transport
$\eta_k = -\xi_k + \beta_k \mathcal{T}_{Y_{k-1}\to Y_k}(\eta_{k-1})$

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

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

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

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

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

8: **end for**

Since $\eta_1 = -\operatorname{grad} F(Y_1) = -\nabla F(Y_1)$, Burer–Monteiro CG coincides with RCG 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

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

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

$$(Y_k^* Y_k)\Omega + \Omega(Y_k^* Y_k) = Y_k^* \eta_{k-1} - \eta_{k-1}^* Y_k \tag{5.2}$$

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 shows that the RHS of (5.2) satisfies

$$\begin{aligned} Y_2^* \eta_1 - \eta_1^* Y_2 &= -Y_2^* \xi_1 + \xi_1^* Y_2 = -(Y_1 - c\xi_1)^* \xi_1 + \xi_1^*(Y_1 - c\xi_1) = \xi_1^* Y_1 - Y_1^* \xi_1 \\ &= Y_1^*(2\nabla f(Y_1 Y_1^*))Y_1 - Y_1^*(2\nabla f(Y_1 Y_1^*))Y_1 = 0. \end{aligned}$$

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 (5.2) 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

$$Y_{k+1}^* \eta_k - \eta_k^* Y_{k+1} = (Y_k + c\eta_k)^* \eta_k - \eta_k^*(Y_k + c\eta_k) = Y_k^* \eta_k - \eta_k^* Y_k$$

$$= Y_k^* \left(-\xi_k + \beta_k P_{Y_k}^{\mathcal{H}^1}(\eta_{k-1})\right) - \left(-\xi_k + \beta_k P_{Y_k}^{\mathcal{H}^1}(\eta_{k-1})\right)^* Y_k$$
$$= Y_k^*(-\xi_k + \beta_k \eta_{k-1}) - (-\xi_k + \beta_k \eta_{k-1})^* Y_k$$
$$= -Y_k^* \xi_k + \xi_k^* Y_k = -Y_k^*(2\nabla f(Y_k Y_k^*))Y_k + Y_k^*(2\nabla f(Y_k Y_k^*))Y_k = 0.$$

So $P_{Y_{k+1}}^{\mathcal{H}^1}(\eta_k) = \eta_k$ also holds, thus RCG is equivalent to Burer–Monteiro CG. ∎

Since $\beta_k \equiv 0$ gives the gradient descent, the same proof above gives Theorem 5.2.

THEOREM 5.2. *Using retraction (4.5) and metric g^1, the Riemannian gradient descent is equivalent to the Burer–Monteiro gradient descent method with suitable step size* (1.3) *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 5.1 is equivalent to Algorithm 5.2 with Riemannian metric g^3, a specific retraction (5.3) and a specific vector transport (5.4). 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 $\mathcal{H}_+^{n,p}$.

Since $\tilde{\beta}$ is a diffeomorphism between $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ and $\mathcal{H}_+^{n,p}$, $D\tilde{\beta}(\pi(Y))[\cdot]$ defines an isomorphism between the 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 $L_{\pi(Y)}$. The following lemma can be verified by straightforward computation, see [29].

LEMMA 5.3. *For $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^3)$, the Riemannian gradient of f and h is related by $(D\tilde{\beta})(\pi(Y))[\operatorname{grad} h(\pi(Y))] = \operatorname{grad} f(YY^*)$ and*

$$L_{\pi(Y)}(\operatorname{grad} h(\pi(Y))) = \operatorname{grad} f(\tilde{\beta}(\pi(Y))).$$

In Algorithm 5.1, 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 is a retraction R^Q and a vector transport $\mathcal{T}^Q$, (with the superscript Q denoting *Quotient*), on the Riemannian quotient manifold $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^3)$, such that Algorithm 5.2 is equivalent to Algorithm 5.1. The idea is again to use the diffeomorphism $\tilde{\beta}$ and the isomorphism $L_{\pi(Y)}$. We give the desired expression of R^Q and $\mathcal{T}^Q$ as follows.

$$R_{\pi(Y)}^Q(\xi_{\pi(Y)}) := \tilde{\beta}^{-1}\left(R_{\tilde{\beta}(\pi(Y))}^E\left(L(\xi_{\pi(Y)})\right)\right), \tag{5.3}$$

$$\mathcal{T}_{\eta_{\pi(Y)}}^Q(\xi_{\pi(Y)}) := L_{\pi(Y_2)}^{-1}\left(\mathcal{T}_{L(\eta_{\pi(Y)})}^E\left(L(\xi_{\pi(Y)})\right)\right), \tag{5.4}$$

where $\pi(Y_2)$ is in $\mathbb{C}_*^{n\times p}/\mathcal{O}_p$ such that $\tilde{\beta}(\pi(Y_2))$ denotes the foot of the tangent vector $\mathcal{T}_{L(\eta_{\pi(Y)})}^E\left(L(\xi_{\pi(Y)})\right)$.

Now it remains to show that R^Q defined in (5.3) is indeed a retraction and $\mathcal{T}^Q$ defined in (5.4) is indeed a vector transport.

LEMMA 5.4. *R^Q defined in (5.3) is a retraction.*

Proof. First it is easy to see that $R_{\pi(Y)}^Q(0_{\pi(Y)}) = \pi(Y)$. Then we also have for all $v_{\pi(Y)} \in T_{\pi(Y)}\mathbb{C}_*^{n\times p}/\mathcal{O}_p$, $\mathrm{D}\, R_{\pi(Y)}^Q(0_{\pi(Y)})[\cdot]$ is an identity map because

$$\mathrm{D}\, R_{\pi(Y)}^Q(0_{\pi(Y)})[v_{\pi(Y)}] = (\mathrm{D}\,\tilde{\beta}^{-1})(\tilde{\beta}(\pi(Y))\left[\mathrm{D}\, R_{\tilde{\beta}(\pi(Y))}^E(0)\left[\mathrm{D}\, L(0)\left[v_{\pi(Y)}\right]\right]\right]$$

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

$$= (\mathrm{D}\,\tilde{\beta}^{-1})(\tilde{\beta}(\pi(Y))\left[L(v_{\pi(Y)})\right] = \left(\mathrm{D}\,\tilde{\beta}(\pi(Y))\right)^{-1}\left[L(v_{\pi(Y)})\right] = L^{-1}(L(v_{\pi(Y)})) = v_{\pi(Y)}$$

LEMMA 5.5. *$\mathcal{T}^E$ defined in (5.4) 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 $\pi(Y_2)$. ∎

We also need the initial step size to match the one in step 5 of Algorithm 5.2. We simply replace the original initial step size t_k by $t_k = \arg\min_t f(Y_kY_k^* + t(Y_k\eta_k^* + \eta_kY_k^*))$.

This value of t_k now is equivalent to the initial step size in Step 5 of Algorithm 5.1. This gives us the following result:

THEOREM 5.6. *With the newly constructed initial step size, retraction, and vector transport in this subsection, Algorithm 5.2 for solving* (4.1) *is equivalent to Algorithm 5.1 solving* (1.1) *in the sense that they produce exactly the same iterates.*

5.3. Implementation details.

The algorithms in this paper can be used for any smooth $f(X)$ in (1.1). For large n, however, it is advisable to avoid using $\nabla f(X) \in \mathbb{C}^{n\times n}$ explicitly. Instead, we compute the matrix-vector multiplications $\nabla f(X)U$. For example, in the PhaseLift problem [9], these matrix-vector multiplications can be implemented via the FFT at a cost of $O(pn\log n)$ when $U \in \mathbb{C}^{n\times p}$, see [16]. We give some detailed implementation in Algorithms 5.1 and 5.2. When counting flops, we assume that $\nabla f(X)U \in \mathbb{C}^{n\times p}$ can be computed in $spn\log n$ flops with s small.

Algorithm 5.3 Calculate the Riemannian gradient grad $f(X)$

Require: $X = U\Sigma U^* \in \mathcal{H}^{n,p}_+$
Ensure: grad $f(X) = UHU^* + U_pU^* + UU_p^* \in T_X\mathcal{H}^{n,p}_+$
- $T \leftarrow \nabla f(X)U$ ▷ # $spn\log n$ flops
- $H \leftarrow U^*T$ ▷ # $np(2p-1)$ flops
- $U_p \leftarrow T - UH$ ▷ # $np(2p-1) + np$ flops

6. Estimates of Rayleigh quotient for Riemannian Hessians.

In many applications, (1.1) or (4.1) is often used for solving (1.2). Even if the global minimizer of (1.2) has a known rank r, one might consider solving (1.1) or (4.1) for Hermitian PSD matrices with fixed rank $p > r$. For instance, in PhaseLift [9] and interferometry recovery [10], the minimizer to (1.2) is 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 [10, 16]. If $p > r$, then an algorithm that solves (1.1) or (4.1) 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 $k \to \infty$.

Algorithm 5.4 Calculate the vector transport $P^t_{X_2}(\nu)$

Require: $X_1 = U_1\Sigma_1 U_1^*$, $X_2 = U_2\Sigma_2 U_2^*$ and tangent vector $\nu = U_1 H_1 U_1^* + U_{p1}U_1^* + U_1 U_{p1}^* \in T_{X_1}\mathcal{H}_+^{n,p}$.

Ensure: $P^t_{X_2}(\nu) = U_2 H_2 U_2^* + U_{p2}U_2^* + U_2 U_{p2}^*$

$A \leftarrow U_1^* U_2$ ▷ # $np(2p-1)$ flops

$H_2^{(1)} \leftarrow A^* H_1 A, \quad U_p^{(1)} \leftarrow U_1(H_1 A)$ ▷ # $3p^2(2p-1) + np(2p-1)$ flops

$H_2^{(2)} \leftarrow U_2^* U_{p1} A, \quad U_p^{(2)} \leftarrow U_{p1} A$ ▷ # $p^2(2n-1) + 2np(2p-1)$ flops

$H_2^{(3)} \leftarrow H_2^{(2)*}, \quad U_p^{(3)} \leftarrow U_1({U_1}_p^* U_2)$ ▷ # $np(2p-1) + p^2(2n-1)$ flops

$H_2 \leftarrow H_2^{(1)} + H_2^{(2)} + H_2^{(3)}$ ▷ # $2p^2$ flops

$U_{p2} \leftarrow U_p^{(1)} + U_p^{(2)} + U_p^{(3)}, \quad U_{p2} \leftarrow U_{p2} - U_2(U_2^* U_{p2})$ ▷ # $3np + np(2p-1) + p^2(2n-1)$ flops

Algorithm 5.5 Calculate the retraction $R_X(Z) = P_{\mathcal{H}_+^{n,p}}(X+Z)$

Require: $X = U\Sigma U^* \in \mathcal{H}_+^{n,p}$, tangent vector $Z = UHU^* + U_p U^* + UU_p^*$.

Ensure: $R_X(Z) = U_+\Sigma_+ U_+^*$.

$(Q,R) \leftarrow \mathrm{qr}(U_p, 0) \quad M \leftarrow \begin{bmatrix} \Sigma + H & R^* \\ R & 0 \end{bmatrix}$ ▷ # $20np^2$ flops

$[V,S] \leftarrow \mathrm{eig}(M)$ ▷ $O(p^3)$ flops

$\Sigma+ \leftarrow S(1:p, 1:p), \quad U_+ \leftarrow \begin{bmatrix} U & Q \end{bmatrix} V(:, 1:p)$ ▷ # $np(4p-1)$ flops

In this section, we analyze the eigenvalues of the Riemannian Hessian near the global minimizer. We will obtain upper and lower bounds of the Rayleigh quotient at $X = YY^*$ (or $\pi(Y)$) that is close to the global minimizer $\hat{X} = \hat{Y}\hat{Y}^*$ (or $\pi(\hat{Y})$).

6.1. The Rayleigh quotient estimates.

DEFINITION 6.1. *The Rayleigh quotient of the Riemannian Hessian of f on $(\mathcal{H}_+^{n,p}, g)$ is defined by $\rho^E(X,\zeta_X) = \frac{g_X(\mathrm{Hess}\, f(X)[\zeta_X],\zeta_X)}{g_X(\zeta_X,\zeta_X)}, \forall \zeta_X \in T_X\mathcal{H}_+^{n,p}$. The Rayleigh quotient of the Riemannian Hessian of h on $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^i)$ is defined by $\rho^i(\pi(Y),\xi_{\pi(Y)}) = \frac{g^i_{\pi(Y)}(\mathrm{Hess}\, h(\pi(Y))[\xi_{\pi(Y)}],\xi_{\pi(Y)})}{g^i_{\pi(Y)}(\xi_{\pi(Y)},\xi_{\pi(Y)})}, \quad \forall \xi_{\pi(Y)} \in T_{\pi(Y)}\mathbb{C}_*^{n\times p}/\mathcal{O}_p$. If the Rayleigh quotient has a lower bound a and an upper bound b, then we define $\frac{b}{a}$ as an upper bound on the condition number of the Riemannian Hessian.*

By the expressions of Riemannian Hessian, we have

$$\rho^E(X,\zeta_X) = \frac{\left\langle \nabla^2 f(X)[\zeta_X],\zeta_X \right\rangle_{\mathbb{C}^{n\times n}}}{g_X(\zeta_X,\zeta_X)} + \frac{g_X\left(P_X^p\left(\nabla f(X)(X^\dagger \zeta_X^p)^* + (\zeta_X^p X^\dagger)^*\nabla f(X)\right),\zeta_X\right)}{g_X(\zeta_X,\zeta_X)}.$$

$$\rho^1(\pi(Y),\xi_{\pi(Y)}) = \frac{\left\langle \nabla^2 f(YY^*)[Y\bar{\xi}_Y^* + \bar{\xi}_Y Y^*], Y\bar{\xi}_Y^* + \bar{\xi}_Y Y^* \right\rangle_{\mathbb{C}^{n\times n}}}{g_Y^1(\bar{\xi}_Y,\bar{\xi}_Y)} + \frac{g_Y^1(2\nabla f(YY^*)\bar{\xi}_Y,\bar{\xi}_Y)}{g_Y^1(\bar{\xi}_Y,\bar{\xi}_Y)}.$$

$$\rho^2(\pi(Y),\xi_{\pi(Y)}) = \frac{\left\langle \nabla^2 f(YY^*)[Y\bar{\xi}_Y^* + \bar{\xi}_Y Y^*], Y\bar{\xi}_Y^* + \bar{\xi}_Y Y^* \right\rangle_{\mathbb{C}^{n\times n}}}{g_Y^2(\bar{\xi}_Y,\bar{\xi}_Y)} + \frac{\left\langle \nabla f(YY^*)P_Y^\perp\bar{\xi}_Y,\bar{\xi}_Y \right\rangle_{\mathbb{C}^{n\times p}}}{g_Y^2(\bar{\xi}_Y,\bar{\xi}_Y)}$$
$$+ \frac{\left\langle P_Y^\perp\nabla f(YY^*)\bar{\xi}_Y,\bar{\xi}_Y \right\rangle_{\mathbb{C}^{n\times p}}}{g_Y^2(\bar{\xi}_Y,\bar{\xi}_Y)} + \frac{\left\langle Y\bar{\xi}_Y^*\bar{\xi}_Y, 2\nabla f(YY^*)Y(Y^*Y)^{-1} \right\rangle_{\mathbb{C}^{n\times p}}}{g_Y^2(\bar{\xi}_Y,\bar{\xi}_Y)} - \frac{\left\langle \bar{\xi}_Y Y^*\bar{\xi}_Y, 2\nabla f(YY^*)Y(Y^*Y)^{-1} \right\rangle_{\mathbb{C}^{n\times p}}}{g_Y^2(\bar{\xi}_Y,\bar{\xi}_Y)}.$$

$$\rho^3(\pi(Y),\xi_{\pi(Y)}) = \frac{\left\langle \nabla^2 f(YY^*)[Y\bar{\xi}_Y^* + \bar{\xi}_Y Y^*], Y\bar{\xi}_Y^* + \bar{\xi}_Y Y^* \right\rangle_{\mathbb{C}^{n\times n}}}{g_Y^3(\bar{\xi}_Y,\bar{\xi}_Y)} + \frac{g_Y^3((I-P_Y)\nabla f(YY^*)(I-P_Y)\bar{\xi}_Y(Y^*Y)^{-1},\bar{\xi}_Y)}{g_Y^3(\bar{\xi}_Y,\bar{\xi}_Y)}.$$

Observe that the leading terms in the above Rayleigh quotients take similar forms: the numerator involves the Hessian $\nabla^2 f$, and the denominator is the induced norm of tangent vector from the respective Riemannian metric. We call the leading term *second order term* (SOT) as it involves Hessian of f as the second order information of f and we call the other terms that follow the leading term *first order terms* (FOTs) as they only contain the first order gradient.

We assume that the Hessian $\nabla^2 f$ is well conditioned on the tangent space:

ASSUMPTION 6.1. *For a fixed $\epsilon > 0$, there exists constants $A > 0$ and $B > 0$ such that for all X with $\left\| X - \hat{X} \right\|_F < \epsilon$, the following inequality holds for all $\zeta_X \in T_X \mathcal{H}_+^{n,p}$.*

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

Observe that this assumption is always satisfied for sufficiently small ϵ when f is smooth and $\hat{X}$ is a nondegenerate minimizer of f. However, the condition number B/A might be large in general. An important case for which this assumption holds is $f(X) = \frac{1}{2} \left\| X - H \right\|_F^2$ with H being a given Hermitian PSD matrix. In this case, $\nabla^2 f(X)$ is the identity operator thus $A = B = 1$.

Under Assumption 6.1, we get bounds of the SOT in $\rho^E(X, \zeta_X)$ as:

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

For quotient manifold, since $Y \bar{\xi}_Y^* + \bar{\xi}_Y Y^* \in T_{YY^*} \mathcal{H}_+^{n,p}$, under Assumption 6.1, we get

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

So the estimates of SOT for quotient manifold reduces to analyzing $\frac{\left\| Y\bar{\xi}_Y^* + \bar{\xi}_Y Y^* \right\|_F^2}{g_Y^i\left(\bar{\xi}_Y, \bar{\xi}_Y\right)}$. We denote its infimum and supremum by

$$C_{\pi(Y)}^i := \inf_{\xi_{\pi(Y)} \in T_{\pi(Y)} \mathbb{C}_*^{n\times p}/\mathcal{O}_p} \frac{\left\| Y\bar{\xi}_Y^* + \bar{\xi}_Y Y^* \right\|_F^2}{g_Y^i(\bar{\xi}_Y, \bar{\xi}_Y)}, D_{\pi(Y)}^i := \sup_{\xi_{\pi(Y)} \in T_{\pi(Y)} \mathbb{C}_*^{n\times p}/\mathcal{O}_p} \frac{\left\| Y\bar{\xi}_Y^* + \bar{\xi}_Y Y^* \right\|_F^2}{g_Y^i(\bar{\xi}_Y, \bar{\xi}_Y)}.$$

The subscript is used to emphasize that the infimum and supremum are dependent on $\pi(Y)$. The next lemma characterizes these infimum and supremum.

LEMMA 6.1. *Let $YY^* = U\Sigma U^*$ denote the compact SVD of YY^* and denote the i-th diagonal entry of Σ by σ_i with $\sigma_1 \geq \cdots \geq \sigma_p > 0$. Then the following estimates for the infimum $C_{\pi(Y)}^i$ and the supremum $D_{\pi(Y)}^i$ of $\frac{\left\| Y\bar{\xi}_Y^* + \bar{\xi}_Y Y^* \right\|_F^2}{g_Y^i\left(\bar{\xi}_Y, \bar{\xi}_Y\right)}$ hold: $C_{\pi(Y)}^1 = 2\sigma_p, 2\sigma_1 \leq D_{\pi(Y)}^1 \leq 2\left(\frac{\sigma_1^2}{\sigma_p} + \sigma_1\right)$; $C_{\pi(Y)}^2 = 2, D_{\pi(Y)}^2 = 4$; and $C_{\pi(Y)}^3 = D_{\pi(Y)}^3 = 1$.*

Proof. It is straightforward to see $C_{\pi(Y)}^3 = D_{\pi(Y)}^3 = 1$ by the definition of g^3. For metric 2, write $\bar{\xi}_Y = YS + Y_\perp K$ for some $S = S^* \in \mathbb{C}^{p\times p}$ and $K \in \mathbb{C}^{n\times p}$. We have

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

Hence it is easy to see $C_{\pi(Y)}^2 = 2$ when S is zero matrix and $D_{\pi(Y)}^2 = 4$ when YSY^* is nonzero and K is zero matrix. For metric 1, by its horizontal space, we can write

$\overline{\xi}_Y = Y(Y^*Y)^{-1}S + Y_\perp K$ for some $S = S^* \in \mathbb{C}^{p\times p}$ and $K \in \mathbb{C}^{n\times p}$. Notice that the SVD of Y can be given as $Y = U\Sigma^{\frac{1}{2}}V^*$. Let $\bar{S} = V^*SV$ and $\bar{K} = KV$, and $\bar{K}_i$ be the i-th column of $\bar{K}$, then

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

$$= \frac{\left\|\Sigma^{-\frac{1}{2}}\bar{S}\Sigma^{\frac{1}{2}} + \Sigma^{\frac{1}{2}}\bar{S}\Sigma^{-\frac{1}{2}}\right\|_F^2 + 2\left\|\bar{K}\Sigma^{\frac{1}{2}}\right\|_F^2}{\left\|\Sigma^{-\frac{1}{2}}\bar{S}\right\|_F^2 + \left\|\bar{K}\right\|_F^2} = \frac{2\sum\limits_{i,j=1}^{p}\frac{\sigma_j}{\sigma_i}\left|\bar{S}_{ij}\right|^2 + 2\sum\limits_{i,j=1}^{p}\left|\bar{S}_{ij}\right|^2 + 2\sum\limits_{i=1}^{p}\sigma_i\|K_i\|_F^2}{\sum\limits_{i,j=1}^{p}\frac{|\bar{S}_{ij}|^2}{\sigma_i} + \sum\limits_{i=1}^{p}\left\|\bar{K}_i\right\|_F^2},$$

where symmetry $\bar{S}^* = \bar{S}$ is used in the last step. The lower bound is given by

$$\frac{2\sum\limits_{i,j=1}^{p}\frac{\sigma_j}{\sigma_i}\left|\bar{S}_{ij}\right|^2 + 2\sum\limits_{i,j=1}^{p}\left|\bar{S}_{ij}\right|^2 + 2\sum\limits_{i=1}^{p}\sigma_i\left\|\bar{K}_i\right\|_F^2}{\sum\limits_{i,j=1}^{p}\frac{|\bar{S}_{ij}|^2}{\sigma_i} + \sum\limits_{i=1}^{p}\left\|\bar{K}_i\right\|_F^2} \geq \frac{2\left(\frac{\sigma_p}{\sigma_1}+1\right)\sum\limits_{i,j=1}^{p}\left|\bar{S}_{ij}\right|^2 + 2\sigma_p\sum\limits_{i=1}^{p}\left\|\bar{K}_i\right\|_F^2}{\frac{1}{\sigma_p}\sum\limits_{i,j=1}^{p}\left|\bar{S}_{ij}\right|^2 + \sum\limits_{i=1}^{p}\left\|\bar{K}_i\right\|_F^2}$$

$$= \frac{2\left(\frac{\sigma_p^2}{\sigma_1} + \sigma_p\right)\sum\limits_{i,j=1}^{p}\left|\bar{S}_{ij}\right|^2 + 2\sigma_p^2\sum\limits_{i=1}^{p}\left\|\bar{K}_i\right\|_F^2}{\sum\limits_{i,j=1}^{p}\left|\bar{S}_{ij}\right|^2 + \sigma_p\sum\limits_{i=1}^{p}\left\|\bar{K}_i\right\|_F^2} \geq 2\sigma_p.$$

This lower bound is sharp as one can choose $S = 0$ and K with $\left\|\bar{K}_p\right\|_F = 1$ and $\left\|\bar{K}_i\right\|_F = 0$ for $i < p$. We have the upper bound as follows.

$$\frac{2\sum\limits_{i,j=1}^{p}\frac{\sigma_j}{\sigma_i}\left|\bar{S}_{ij}\right|^2 + 2\sum\limits_{i,j=1}^{p}\left|\bar{S}_{ij}\right|^2 + 2\sum\limits_{i=1}^{p}\sigma_i\left\|\bar{K}_i\right\|_F^2}{\sum\limits_{i,j=1}^{p}\frac{|\bar{S}_{ij}|^2}{\sigma_i} + \sum\limits_{i=1}^{p}\left\|\bar{K}_i\right\|_F^2} \leq \frac{2\left(\frac{\sigma_1}{\sigma_p}+1\right)\sum\limits_{i,j=1}^{p}\left|\bar{S}_{ij}\right|^2 + 2\sigma_1\sum\limits_{i=1}^{p}\left\|\bar{K}_i\right\|_F^2}{\frac{1}{\sigma_1}\sum\limits_{i,j=1}^{p}\left|\bar{S}_{ij}\right|^2 + \sum\limits_{i=1}^{p}\left\|\bar{K}_i\right\|_F^2}$$

$$= \frac{2\left(\frac{\sigma_1^2}{\sigma_p} + \sigma_1\right)\sum\limits_{i,j=1}^{p}\left|\bar{S}_{ij}\right|^2 + 2\sigma_1^2\sum\limits_{i=1}^{p}\left\|\bar{K}_i\right\|_F^2}{\sum\limits_{i,j=1}^{p}\left|\bar{S}_{ij}\right|^2 + \sigma_1\sum\limits_{i=1}^{p}\left\|\bar{K}_i\right\|_F^2} < 2\left(\frac{\sigma_1^2}{\sigma_p} + \sigma_1\right),$$

where the last inequality is obtained by the range of the rational function $f(x, y) = \frac{ax+by}{x+dy}$ with $a = 2\left(\frac{\sigma_1^2}{\sigma_p} + \sigma_1\right), b = 2\sigma_1^2$ and $d = \sigma_1$ on $\{(x, y)|x \geq 0, y \geq 0, xy \neq 0\}$.

This upper bound $2\left(\frac{\sigma_1^2}{\sigma_p} + \sigma_1\right)$ may not be the supremum as the inequalities are not sharp. However, we can show that $D^1_{\pi(Y)} \geq 2\sigma_1$. To see this, choose $\bar{S} = 0$ and K with $\left\|\bar{K}_1\right\|_F = 1$ and $\left\|\bar{K}_i\right\|_F = 0$ for $i > 1$. Then (6.1) reaches the value $2\sigma_1$. Hence the supremum must be at least $2\sigma_1$. So we have

$$2\sigma_1 \leq D^1_{\pi(Y)} \leq 2\left(\frac{\sigma_1^2}{\sigma_p} + \sigma_1\right). \tag{6.1}$$

□

Next we estimate the FOTs in Rayleigh quotient.

Lemma 6.2. *Let* $X = YY^*$ *for any* $Y \in \pi^{-1}(\pi(Y))$ *with* $X \in \mathcal{H}_+^{n,p}$ *and* $\pi(Y) \in \mathbb{C}_*^{n\times p}/\mathcal{O}_p$. *Let* $U\Sigma U^*$ *be the compact SVD of* X *and denote the ith diagonal entry of* Σ *with* $\sigma_1 \geq \cdots \geq \sigma_p > 0$.

1. *For the embedded manifold we have* $|FOT| \leq \frac{2}{\sigma_p} \|\nabla f(X)\|$.
2. *For the quotient manifold with metric* g^1 *we have* $|FOT| \leq 2\|\nabla f(YY^*)\|$.
3. *For the quotient manifold with* g^2 *we have* $|FOTs| \leq \frac{4(\sqrt{p}+1)}{\sigma_p} \|\nabla f(YY^*)\|$.
4. *For the quotient manifold with* g^3 *we have* $|FOTs| \leq \frac{1}{\sigma_p} \|\nabla f(YY^*)\|$.

Proof. We will use $\|B^*A^*\|_F = \|AB\|_F \leq \|A\|\|B\|_F \leq \|A\|_F\|B\|_F$ where $\|A\|$ is the spectral norm. If X is Hermitian, $\|AX\|_F = \|XA^*\|_F \leq \|X\|\|A^*\|_F = \|X\|\|A\|_F$.

For the embedded manifold, recall that $\xi_X^s = P_X^s(\xi_X)$ and $\xi_X^p = P_X^p(\xi_X)$ and P_X^t and P_X^p are defined in (3.2), and the bound for the FOT is given by

$$\frac{\left|g_X\left(P_X^p\left(\nabla f(X)(X^\dagger \zeta_X^p)^* + (\zeta_X^p X^\dagger)^* \nabla f(X)\right), \zeta_X\right)\right|}{g_X(\zeta_X, \zeta_X)} = \frac{\left|\left\langle P_X^p\left(\nabla f(X)\zeta_X^p X^\dagger + X^\dagger \zeta_X^p \nabla f(X)\right), \zeta_X\right\rangle_{\mathbb{C}^{n\times n}}\right|}{\langle \zeta_X, \zeta_X\rangle_{\mathbb{C}^{n\times n}}}$$

$$\leq \frac{\left|\left\langle P_X^p\left(\nabla f(X)\zeta_X^p X^\dagger\right), \zeta_X\right\rangle_{\mathbb{C}^{n\times n}}\right|}{\langle \zeta_X, \zeta_X\rangle_{\mathbb{C}^{n\times n}}} + \frac{\left|\left\langle P_X^p\left(X^\dagger \zeta_X^p \nabla f(X)\right), \zeta_X\right\rangle_{\mathbb{C}^{n\times n}}\right|}{\langle \zeta_X, \zeta_X\rangle_{\mathbb{C}^{n\times n}}}$$

$$\leq 2\frac{\|\nabla f(X)\zeta_X^p X^\dagger\|_F \|\zeta_X\|_F}{\langle \zeta_X, \zeta_X\rangle_{\mathbb{C}^{n\times n}}} \leq 2\frac{\|\nabla f(X)\|\|\zeta_X^p X^\dagger\|_F \|\zeta_X\|_F}{\langle \zeta_X, \zeta_X\rangle_{\mathbb{C}^{n\times n}}} \leq 2\frac{\|\nabla f(X)\|\|X^\dagger\|\|\zeta_X^p\|_F \|\zeta_X\|_F}{\langle \zeta_X, \zeta_X\rangle_{\mathbb{C}^{n\times n}}}$$

$$\leq \frac{2\|\nabla f(X)\|\left\|X^\dagger\right\|\|\zeta_X\|_F^2}{\langle \zeta_X, \zeta_X\rangle_{\mathbb{C}^{n\times n}}} = 2\|\nabla f(X)\|\left\|X^\dagger\right\| = \frac{2}{\sigma_p}\|\nabla f(X)\|.$$

For quotient manifold with g^1, the FOT is bounded by

$$\frac{\left|g_Y^1(2\nabla f(YY^*)\bar{\xi}_Y, \bar{\xi}_Y)\right|}{g_Y^1(\bar{\xi}_Y, \bar{\xi}_Y)} = \frac{\left|\left\langle 2\nabla f(YY^*)\bar{\xi}_Y, \bar{\xi}_Y\right\rangle_{\mathbb{C}^{n\times p}}\right|}{\left\langle \bar{\xi}_Y, \bar{\xi}_Y\right\rangle_{\mathbb{C}^{n\times p}}} \leq \frac{2\|\nabla f(YY^*)\|\left\|\bar{\xi}_Y\right\|_F^2}{\left\langle \bar{\xi}_Y, \bar{\xi}_Y\right\rangle_{\mathbb{C}^{n\times p}}} = 2\|\nabla f(YY^*)\|.$$

For quotient manifold with g^2, the FOTs contains four terms and we estimate each term separately. Notice that the SVD of Y can be given as $Y = U\Sigma^{\frac{1}{2}}V^*$. Let $\bar{S} = V^*SV$ and $\bar{K} = KV$, and $\bar{K}_i$ be the i-th column of $\bar{K}$. For the first summand we have

$$\frac{\left|\left\langle \nabla f(YY^*)P_Y^\perp \bar{\xi}_Y, \bar{\xi}_Y\right\rangle_{\mathbb{C}^{n\times p}}\right|}{g_Y^2(\bar{\xi}_Y, \bar{\xi}_Y)} = \frac{\left|\left\langle \nabla f(YY^*)P_Y^\perp \bar{\xi}_Y, \bar{\xi}_Y\right\rangle_{\mathbb{C}^{n\times p}}\right|}{\left\langle \bar{\xi}_Y Y^*, \bar{\xi}_Y Y^*\right\rangle_{\mathbb{C}^{n\times n}}} \leq \frac{\|\nabla f(YY^*)\|\left\|\bar{\xi}_Y\right\|_F^2}{\left\langle \bar{\xi}_Y Y^*, \bar{\xi}_Y Y^*\right\rangle_{\mathbb{C}^{n\times n}}}.$$

$$= \frac{\|YS\|_F^2 + \|K\|_F^2}{\|YSY^*\|_F^2 + \|KY^*\|_F^2}\|\nabla f(YY^*)\| \leq \left(\frac{\|YS\|_F^2}{\|YSY^*\|_F^2} + \frac{\|K\|_F^2}{\|KY^*\|_F^2}\right)\|\nabla f(YY^*)\|$$

$$= \left(\frac{\left\|\sqrt{\Sigma}\bar{S}\right\|_F^2}{\left\|\sqrt{\Sigma}\bar{S}\sqrt{\Sigma}\right\|_F^2} + \frac{\|\bar{K}\|_F^2}{\left\|\bar{K}\sqrt{\Sigma}\right\|_F^2}\right)\|\nabla f(YY^*)\| \leq \frac{2}{\sigma_p}\|\nabla f(YY^*)\|.$$

Similarly, we have the second term: $\frac{\left|\left\langle P_Y^\perp \nabla f(YY^*)\bar{\xi}_Y, \bar{\xi}_Y\right\rangle_{\mathbb{C}^{n\times p}}\right|}{g_Y^2(\bar{\xi}_Y, \bar{\xi}_Y)} \leq \frac{2}{\sigma_p}\|\nabla f(YY^*)\|$.

For the third term, with the fact $\|A^*A\|_F = \|A\|_F^2$, we have

$$\frac{\left|\left\langle Y\bar{\xi}_Y^*\bar{\xi}_Y, 2\nabla f(YY^*)Y(Y^*Y)^{-1}\right\rangle_{\mathbb{C}^{n\times p}}\right|}{g_Y^2(\bar{\xi}_Y, \bar{\xi}_Y)} = \frac{\left|\left\langle Y\bar{\xi}_Y^*\bar{\xi}_Y Y^*, 2\nabla f(YY^*)Y(Y^*Y)^{-2}Y^*\right\rangle_{\mathbb{C}^{n\times n}}\right|}{g_Y^2(\bar{\xi}_Y, \bar{\xi}_Y)}$$

$$\leq \frac{\left\|Y\bar{\xi}_Y^*\bar{\xi}_Y Y^*\right\|_F \left\|2\nabla f(YY^*)Y(Y^*Y)^{-2}Y^*\right\|_F}{g_Y^2(\bar{\xi}_Y, \bar{\xi}_Y)} \leq \frac{\left\|\bar{\xi}_Y Y^*\right\|_F^2 \|2\nabla f(YY^*)\|\left\|Y(Y^*Y)^{-2}Y^*\right\|_F}{g_Y^2(\bar{\xi}_Y, \bar{\xi}_Y)}$$

$$= 2\left\|Y(Y^*Y)^{-2}Y^*\right\|_F \|\nabla f(YY^*)\| \leq \frac{2\sqrt{p}}{\sigma_p}\|\nabla f(YY^*)\|.$$

Similarly, we can bound the fourth term: $\frac{\left|\left\langle \bar{\xi}_Y Y^* \bar{\xi}_Y, 2\nabla f(YY^*)Y(Y^*Y)^{-1}\right\rangle\right|_{\mathbb{C}^{n\times p}}}{g_Y^2(\bar{\xi}_Y,\bar{\xi}_Y)} \leq \frac{2\sqrt{p}}{\sigma_p}\left\|\nabla f(YY^*)\right\|$.

Thus, for the quotient manifold with g^2 we have $|\text{FOTs}| \leq \frac{4(\sqrt{p}+1)}{\sigma_p}\left\|\nabla f(YY^*)\right\|$.

For g^3, recall that $P_Y^\perp = I - P_Y = I - Y(Y^*Y)^{-1}Y^*$, with the property (4.3) and the fact $(I-P_Y)^*Y = 0$, the FOT can be bounded as follows:

$$|\text{FOT}| = \frac{\left|g_Y^3((I-P_Y)\nabla f(YY^*)(I-P_Y)\bar{\xi}_Y(Y^*Y)^{-1},\bar{\xi}_Y)\right|}{g_Y^3(\bar{\xi}_Y,\bar{\xi}_Y)} = \frac{2\left|\left\langle P_Y^\perp \nabla f(YY^*)P_Y^\perp \bar{\xi}_Y,\bar{\xi}_Y\right\rangle_{\mathbb{C}^{n\times p}}\right|}{g_Y^3(\bar{\xi}_Y,\bar{\xi}_Y)}$$

$$= \frac{2\left|\langle \nabla f(YY^*)Y_\perp K, Y_\perp K\rangle_{\mathbb{C}^{n\times p}}\right|}{\left\|Y\bar{\xi}_Y^* + \bar{\xi}_Y Y^*\right\|_F^2} = \frac{2\left|\langle \nabla f(YY^*)Y_\perp K, Y_\perp K\rangle_{\mathbb{C}^{n\times p}}\right|}{\left\|2YSY^* + Y_\perp KY^* + YK^*Y_\perp^*\right\|_F^2} = \frac{2\left|\langle \nabla f(YY^*)Y_\perp K, Y_\perp K\rangle_{\mathbb{C}^{n\times p}}\right|}{\|2YSY^*\|_F^2 + \|Y_\perp KY^*\|_F^2 + \left\|YK^*Y_\perp^*\right\|_F^2}$$

$$= \frac{\left|\langle \nabla f(YY^*)Y_\perp K, Y_\perp K\rangle_{\mathbb{C}^{n\times p}}\right|}{2\|YSY^*\|_F^2 + \|Y_\perp KY^*\|_F^2} \leq \frac{\left|\langle \nabla f(YY^*)Y_\perp K, Y_\perp K\rangle_{\mathbb{C}^{n\times p}}\right|}{\|Y_\perp KY^*\|_F^2} \leq \frac{\|Y_\perp K\|_F^2}{\|Y_\perp KY^*\|_F^2}\left\|\nabla f(YY^*)\right\| \leq \frac{1}{\sigma_p}\left\|\nabla f(YY^*)\right\|.$$

With Lemma 6.2 and Lemma 6.1, we summarize the main result as follows.

THEOREM 6.3. *Let $\hat{X} = \hat{Y}\hat{Y}^*$ be the global minimizer of (1.2) with rank $r \leq p$. For $X = YY^* = U\Sigma U^*$ with singular values σ_i near $\hat{X}$ where $Y \in \mathbb{C}_*^{n\times p}$, under the Assumption 6.1, for any arbitrary tangent vectors ζ_X and $\xi_{\pi(Y)}$, the following hold:*

1. $A - \frac{2}{\sigma_p}\left\|\nabla f(X)\right\| \leq \rho^E(X,\zeta_X) \leq B + \frac{2}{\sigma_p}\left\|\nabla f(X)\right\|$,
2. $2A\sigma_p - 2\left\|\nabla f(YY^*)\right\| \leq \rho^1(\pi(Y),\xi_{\pi(Y)}) \leq B\cdot D^1_{\pi(Y)} + 2\left\|\nabla f(YY^*)\right\|$,
3. $2A - \frac{4(\sqrt{p}+1)}{\sigma_p}\left\|\nabla f(YY^*)\right\| \leq \rho^2(\pi(Y),\xi_{\pi(Y)}) \leq 4B + \frac{4(\sqrt{p}+1)}{\sigma_p}\left\|\nabla f(YY^*)\right\|$,
4. $A - \frac{1}{\sigma_p}\left\|\nabla f(YY^*)\right\| \leq \rho^3(\pi(Y),\xi_{\pi(Y)}) \leq B + \frac{1}{\sigma_p}\left\|\nabla f(YY^*)\right\|$,

where $D^1_{\pi(Y)}$ satisfies $2\sigma_1 \leq D^1_{\pi(Y)} \leq 2\left(\frac{\sigma_1^2}{\sigma_p} + \sigma_1\right)$. In particular, if $\hat{X} = \hat{Y}\hat{Y}^*$ has rank p, we have the following limits, where $X \to \hat{X}$ and $\pi(Y) \to \pi(\hat{Y})$ are taken in the sense of $\left\|X - \hat{X}\right\|_F \to 0$ and $\left\|YY^* - \hat{Y}\hat{Y}^*\right\|_F \to 0$:

1. $A - \frac{2}{\hat{\sigma}_p}\left\|\nabla f(\hat{X})\right\| \leq \lim_{X\to\hat{X}} \rho^E(X,\xi_X) \leq B + \frac{2}{\hat{\sigma}_p}\left\|\nabla f(\hat{X})\right\|$,
2. $2A\hat{\sigma}_p - 2\left\|\nabla f(\hat{X})\right\| \leq \lim_{\pi(Y)\to\pi(\hat{Y})} \rho^1(\pi(Y),\xi_{\pi(Y)}) \leq B\cdot D^1_{\pi(\hat{Y})} + 2\left\|\nabla f(\hat{X})\right\|$,
3. $2A - \frac{4(\sqrt{p}+1)}{\hat{\sigma}_p}\left\|\nabla f(\hat{X})\right\| \leq \lim_{\pi(Y)\to\pi(\hat{Y})} \rho^2(\pi(Y),\xi_{\pi(Y)}) \leq 4B + \frac{4(\sqrt{p}+1)}{\hat{\sigma}_p}\left\|\nabla f(\hat{X})\right\|$,
4. $A - \frac{1}{\hat{\sigma}_p}\left\|\nabla f(\hat{X})\right\| \leq \lim_{\pi(Y)\to\pi(\hat{Y})} \rho^3(\pi(Y),\xi_{\pi(Y)}) \leq B + \frac{1}{\hat{\sigma}_p}\left\|\nabla f(\hat{X})\right\|$,

where $D^1_{\pi(\hat{Y})}$ satisfies $2\hat{\sigma}_1 \leq D^1_{\pi(\hat{Y})} \leq 2\left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_p} + \hat{\sigma}_1\right)$.

REMARK 6.4. *If we also assume $\nabla f(\hat{X}) = 0$, then the limits above can be further simplified. Though $\nabla f(\hat{X}) = 0$ may not be true in general, it holds for all numerical examples considered in this paper, where the cost function takes the form $f(X) = \frac{1}{2}\left\|\mathcal{A}(X) - b\right\|_F^2$, and the minimizer $\hat{X}$ for (1.1) or (1.2) satisfies $f(\hat{X}) = 0$. Thus $\hat{X}$ is also the minimizer for minimizing $f(X)$ over all $X \in \mathbb{C}$, which implies $\nabla f(\hat{X}) = 0$.*

REMARK 6.5. *Under the assumption $\nabla f(\hat{X}) = 0$, the limit of the condition number for the Bures-Wasserstein metric g^1 depends on the condition number of the minimizer $\hat{X}$. This reflects a significant difference between g^1 and the other two metrics.*

6.2. The Rayleigh quotient for a rank-deficient minimizer. Next, we consider the rank deficient case $p > r$ where r is the rank of the minimizer $\hat{X}$, i.e., the minimizer $\hat{X}$ lies on the boundary of the constraint manifold. Under the Assumption $\nabla f(\hat{X}) = 0$, any convergent algorithm that solves (1.1) or (4.1) will generate a sequence such that both $\sigma_{r+1}, \cdots, \sigma_p$ and $\nabla f(X)$ will vanish as $X \to \hat{X}$. We make one more assumption for a simpler quantification of the lower and upper bounds of Rayleigh quotient near the minimizer.

Assumption 6.2. *For a sequence* $\{X_k\}$ *with* $X_k \in \mathcal{H}_+^{n,p}$ *(or* $\pi(Y_k) \in \mathbb{C}_*^{n\times p}/\mathcal{O}_p$ *) that converges to the minimizer* $\hat{X}$ *(or* $\pi(\hat{Y})$*), let* $(\sigma_p)_k$ *be the smallest nonzero singular value of* $X_k = Y_k Y_k^*$*, assume the following limits hold.*

1. *For the embedded manifold,* $\lim_{k\to\infty} \frac{2}{(\sigma_p)_k} \|\nabla f(X_k)\| \leq \frac{A}{2}$.
2. *For the quotient manifold with metric* g^1, $\lim_{k\to\infty} \frac{1}{(\sigma_p)_k} \|\nabla f(Y_k Y_k^*)\| \leq \frac{A}{2}$.
3. *For the quotient manifold with metric* g^2, $\lim_{k\to\infty} \frac{4(\sqrt{p}+1)}{(\sigma_p)_k} \|\nabla f(Y_k Y_k^*)\| \leq A$.
4. *For the quotient manifold with metric* g^3, $\lim_{k\to\infty} \frac{1}{(\sigma_p)_k} \|\nabla f(Y_k Y_k^*)\| \leq \frac{A}{2}$.

If $\hat{X}$ has rank $r < p$ and $\{X_k\}$ is a sequence that satisfies Assumption 6.2, then Theorem 6.3 implies

1. For the embedded manifold we have $\frac{A}{2} \leq \lim_{k\to\infty} \rho^E(X_k, \xi_{X_k}) \leq B + \frac{A}{2}$.
2. $A \leq \lim_{k\to\infty} \frac{\rho^1(\pi(Y_k),\xi_{\pi(Y_k)})}{(\sigma_p)_k} \leq B \lim_{k\to\infty} \frac{D^1_{\pi(Y_k)}}{(\sigma_p)_k} + 2A$,
3. $A \leq \lim_{k\to\infty} \rho^2(\pi(Y_k), \xi_{\pi(Y_k)}) \leq 4B + A$,
4. $\frac{A}{2} \leq \lim_{k\to\infty} \rho^3(\pi(Y_k), \xi_{\pi(Y_k)}) \leq B + \frac{A}{2}$,

where $\lim\limits_{k\to\infty} \frac{D^1_{\pi(Y_k)}}{(\sigma_p)_k} \geq \lim\limits_{k\to\infty} \frac{2(\sigma_1)_k}{(\sigma_p)_k} = +\infty$ since $\sigma_p \to \hat{\sigma}_p = 0$.

Notice that the condition number in Bures-Wassertein metric g^1 is fundamentally different from the other ones since it is the only metric that blows up.

7. Numerical experiments.

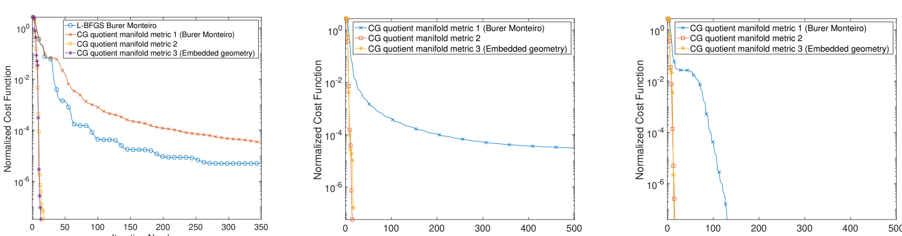
We compare the following four algorithms:

1. RCG on $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^1)$, i.e., Algorithm 5.2 with metric g^1. This algorithm is equivalent to Burer–Monteiro CG, that is, CG applied directly to (1.5).
2. RCG on $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^2)$, i.e., Algorithm 5.2 with metric g^2 in [16].
3. RCG on $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^3)$, i.e., Algorithm 5.2 with metric g^3.
4. Burer–Monteiro L-BFGS method, i.e., L-BFGS directly applied to (1.5).

7.1. Eigenvalue problem.

For a Hermitian PSD matrix H, its top p eigenvalues and associated eigenvectors can be found by solving $\min \frac{1}{2}\|X - H\|_F^2$ with $X \in \mathcal{H}_+^{n,p}$. It is easy to verify that $\nabla f(X) = X - H$ and $\nabla^2 f(X)$ is the identity map.

We consider random Hermitian PSD matrices H of size 50 000-by-50 000 with different ranks $r = 10$ or $r = 15$. See the performance of the algorithms on the manifold with rank $p = 15$ in Figure 1, in which we can see the slowness of Burer-Monteiro methods corresponding to Bures-Wasserstein metric g^1 is consistent with condition number analysis in the previous section.

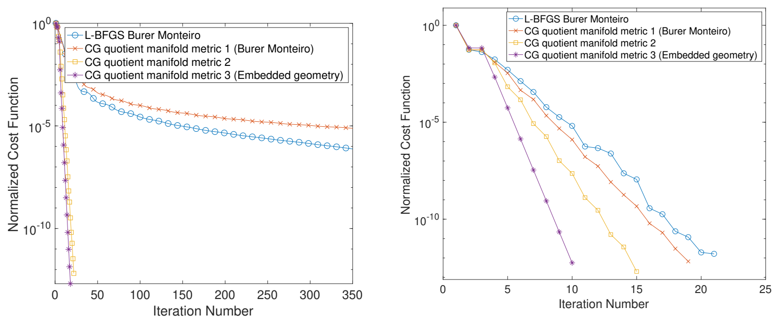


(a) $r = 10$ and $p = 15$ (b) $r = p = 15$ and $\frac{\hat{\sigma}_1}{\hat{\sigma}_p} = 10^6$ (c) $r = p = 15$ and $\frac{\hat{\sigma}_1}{\hat{\sigma}_p} = 10^3$

Fig. 1. *Eigenvalue problem: minimizer has rank* r*, solved on the rank* p *manifold. Burer-Monteiro methods (Bures-Wasserstein metric* g^1*) become slower either when the minimizer has a rank* $r < p$ *or when minimizer* $\hat{X}$ *has a larger condition number* $\frac{\hat{\sigma}_1}{\hat{\sigma}_p}$.

7.2. Matrix completion. We consider a Hermitian matrix completion problem for a given $H \in \mathcal{H}_+^{n,p}$: $\min \frac{1}{2} \|P_\Omega(X-A)\|_F^2, X \in \mathcal{H}_+^{n,p}$, where P_Ω is a sampling operator. We have $\nabla f(X) = P_\Omega(X-A), \quad \nabla^2 f(X)[\zeta_X] = P_\Omega(\zeta_X), \quad \zeta_X \in \mathbb{C}^{n\times n}$.

We consider a Hermitian PSD matrix $H \in \mathbb{C}^{n\times n}$ with $n = 10\,000$ with rank $r = 25$ and P_Ω a random 90% sampling operator. The initial guess is the same random matrix for all four algorithms. In Figure 2, we see that the simpler Burer–Monteiro approach, including the L-BFGS method and the CG method with Bures-Wasserstein metric g^1, is significantly slower for the rank deficient case $r < p$, which is consistent with the Hessian analysis in the previous section.



(a) $r = 25$ and $p = 30$. (b) $r = p = 25$.

FIG. 2. *Matrix completion: minimizer has rank r, solved on the rank p manifold. When $r < p$, Burer-Monteiro methods (Bures-Wasserstein metric g^1) are significantly slower.*

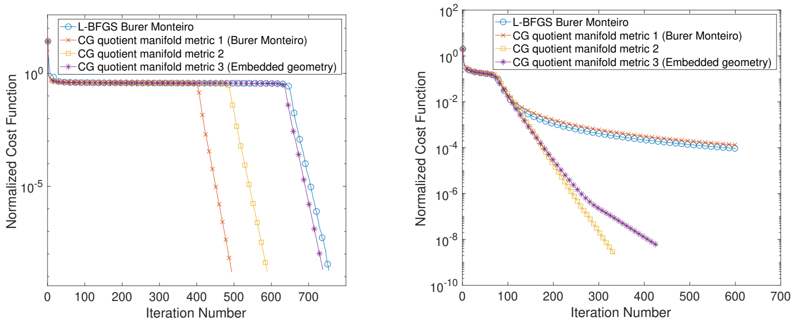
7.3. The PhaseLift problem. We consider the phase retrieval problem as described in [9]. The setup is the same as described in [16]. The cost function can be written as $f(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_F^2$. Straightforward calculation shows

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

For the numerical experiments, we take the phase retrieval problem for a complex gold ball image of size 256×256 as in [16]. Thus $n = 256^2 = 65,536$ in (1.2) or (1.1). We consider the operator $\mathcal{A}$ that corresponds to 6 Gaussian random masks. Hence, the size of b is $6n = 393,216$. Remark that the problem is easier to solve with more masks.

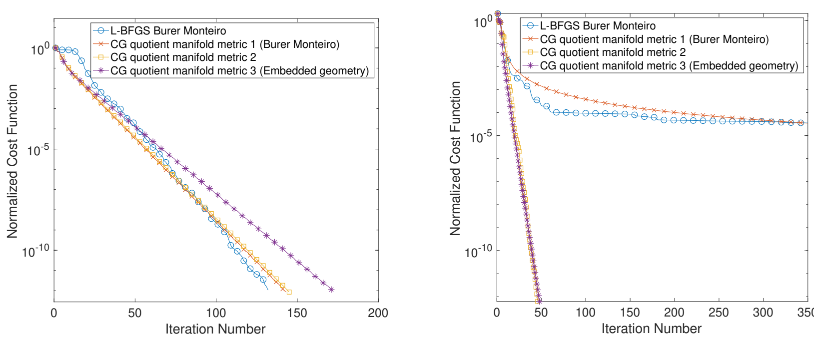
We first test the algorithms with the same random initial guess on the rank-1 and rank-3 manifolds. The results are shown in Figure 3. The initial guess is randomly generated. First, we observe that the nonconvex lifting solving it on rank-p manifold with $p > 1$ can accelerate the convergence, even though the minimizer is always rank-1. 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 $p > r$, we can see that the Burer–Monteiro approach has slower asymptotic convergence rates.

7.4. Interferometry recovery problem. We consider solving the interferometry recovery problem described in [10], given by $\min f(X) = \frac{1}{2} \|P_\Omega(FXF^* - dd^*)\|_F^2$, $X \in \mathcal{H}_+^{n,p}$, where P_Ω is a sparse and symmetric sampling operator, and $F \in \mathbb{C}^{m\times n}$. We solve an interferometry problem with a randomly generated $F \in \mathbb{C}^{10\,000\times 1000}$.



(a) Rank $p = 1$ manifold (b) Rank $p = 3$ manifold

Fig. 3. *Phase retrieval of a complex image: minimizer has rank $r = 1$. Nonconvex lifting on manifolds of rank-p with $p > r$ can accelerate convergence, but Burer-Monteiro methods (Bures-Wasserstein metric g^1) has an obvious slower asymptotic convergence rate when $p > r$.*



(a) Rank $p = 1$ manifold. (b) Rank $p = 3$ manifold.

Fig. 4. *Interferometry recovery: minimizer has rank $r = 1$. When the minimizer is rank deficient $r < p$, Burer-Monteiro methods (Bures-Wasserstein metric g^1) are significantly slower.*

Hence $n = 1000$ in (1.2) or (1.1). The sampling operator Ω is also randomly generated, with 1% density. In Figure 4, when $p = 3$ and $r = 1$, we can see that the Burer–Monteiro approach has slower asymptotic convergence rates.

8. Conclusion.

We have shown that the CG method on the Burer–Monteiro formulation for Hermitian PSD fixed-rank constraints is equivalent to a Riemannian CG method on a quotient manifold with the Bures-Wasserstein metric g^1. We have analyzed the condition numbers of the Riemannian Hessians on $(\mathbb{C}_*^{n\times p}/\mathcal{O}_p, g^i)$ for three metrics. We have shown 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 or Bures-Wasserstein metric often has a slower asymptotic convergence rate in numerical tests.

A. Embedded manifold $\mathcal{H}_+^{n,p}$.

A.1. Riemannian Hessian operator. By [3, section 4], the retraction R defined by projection is a second-order retraction. Proposition 5.5.5 in [2] states that if

R is a second-order retraction, then the Riemannian Hessian of f can be computed by $\operatorname{Hess} f(X) = \operatorname{Hess}(f \circ R_X)(0_X)$. Thus $g_X(\operatorname{Hess} f(X)[\xi_X], \xi_X) = \frac{d^2}{dt^2} f(R_X(t\xi_X))\Big|_{t=0}$. In [28] and [25], a method was proposed to compute $\operatorname{Hess} f(X)$ 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. Following [28, Proposition 5.10], we have

LEMMA A.1. *$\forall X \in \mathcal{H}_+^{n,p}$, the mapping $R_X^{(2)} : T_X\mathcal{H}_+^{n,p} \to \mathcal{H}_+^{n,p}$*

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

is a second-order retraction on $\mathcal{H}_+^{n,p}$, where $X^\dagger$ is the pseudoinverse, $\xi_X^s = P_X^s(\xi_X)$ and $\xi_X^p = P_X^p(\xi_X)$ as defined in (3.2). Moreover, we have

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

From this the Riemannian Hessian operator of f can be computed in essentially the same way as in [24, Section A.2] but applied to the general cost function $f(X)$ instead of a least square cost function. 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

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

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

$$\begin{aligned}
&g_X(\operatorname{grad} f(X), \xi_X) = \langle \nabla f(X), \xi_X \rangle_{\mathbb{C}^{n\times n}} \Rightarrow \operatorname{grad} f(X) = P_X^t(\nabla f(X)), \\
&g_X(\operatorname{Hess} f(X)[\xi_X], \xi_X) = \underbrace{2\langle \nabla f(X), \xi_X^p X^\dagger \xi_X^p \rangle_{\mathbb{C}^{n\times n}}}_{f_1 := \langle \mathcal{H}_1(\xi_X), \xi_X \rangle_{\mathbb{C}^{n\times n}}} + \underbrace{\langle \nabla^2 f(X)[\xi_X], \xi_X \rangle_{\mathbb{C}^{n\times n}}}_{f_2 := \langle \mathcal{H}_2(\xi_X), \xi_X \rangle_{\mathbb{C}^{n\times n}}}.
\end{aligned}$$

Since ξ_X is already separated in f_2, the contribution to Riemannian Hessian from $\mathcal{H}_2$ is readily given by $\mathcal{H}_2(\xi_X) = P_X^t(\nabla^2 f(X)[\xi_X])$.

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 $X^\dagger \xi_X^p$ to the first position of $\langle .,. \rangle_{\mathbb{C}^{n\times n}}$, we write $\mathcal{H}_1(\xi_X)$ as the linear combination of both:

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

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 tangent vector ν_X. Hence we must have $c = \frac{1}{2}$ and we obtain

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

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

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

B.1. Calculations for the Riemannian Hessian. By [2, Definition 5.5.1], the Riemannian Hessian of f at a point x in $\mathcal{M}$ is given by

$$\operatorname{Hess} f(x)[\xi_x] = \nabla_{\xi_x} \operatorname{grad} f(x), \quad \xi_x \in T_x\mathcal{M},$$

where ∇ is the Riemannian connection on $\mathcal{M}$. By [2, Proposition 5.3.3] and the definition of the Riemannian Hessian, we have

LEMMA B.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:*

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

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

B.1.1. Riemannian Hessian for the metric g^1. By [2, Proposition 5.3.2], the Riemannian connection on $\mathbb{C}_*^{n\times p}$ is the classical directional derivative $\nabla_{\eta_Y}\xi = \mathrm{D}\,\xi(Y)[\eta_Y]$. Recall that for g^1, $\operatorname{grad} F(Y) = 2\nabla f(YY^*)Y$. Thus

$$\operatorname{Hess} F(Y)[\xi_Y] = \nabla_{\xi_Y} \operatorname{grad} F = \mathrm{D} \operatorname{grad} F(Y)[\xi_Y] = 2\nabla^2 f(YY^*)[Y\xi_Y^* + \xi_Y Y^*]Y + 2\nabla f(YY^*)\xi_Y.$$

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

B.1.2. Riemannian Hessian under metric g^2. Any Riemannian metric g satisfies the Koszul formula

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

where $[\cdot,\cdot]$ is the *Lie bracket*. In particular, for g^2 the Koszul formula turns into

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

Recall that $g^2(\lambda,\eta)(Y) = \Re(\operatorname{tr}(Y^*Y\lambda_Y^*\eta_Y))$. The first term equals

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

Following [2, Section 5.3.4], since $\mathbb{C}_*^{n\times p}$ is an open subset of $\mathbb{C}^{n\times p}$, we also have $[\lambda,\eta]_Y = \mathrm{D}\,\eta(Y)[\lambda_Y] - \mathrm{D}\,\lambda(Y)[\eta_Y]$. Thus we get

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

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

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

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

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

B.1.3. Riemannian Hessian under metric g^3. Denote

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

Recall that the Riemannian metric g^3 on $\mathbb{C}_*^{n\times p}$ satisfies $g_Y^3(\xi_Y, \eta_Y) = \tilde{g}_Y(\xi_Y, \eta_Y) + g_Y^2(P_Y^{\mathcal{V}}(\xi_Y), P_Y^{\mathcal{V}}(\eta_Y))$. Hence $\mathrm{D}\, g^3(\lambda, \eta)(Y)[\xi_Y] =$

$$
\begin{aligned}
&\tilde{g}_Y(\mathrm{D}\,\lambda(Y)[\xi_Y], \eta_Y) + \tilde{g}(\lambda_Y, D\eta(Y)[\xi_Y]) + 2\Re(\mathrm{tr}(\xi_Y^*\lambda_Y Y^*\eta_Y + Y^*\lambda_Y\xi_Y^*\eta_Y + \xi_Y^* Y\lambda_Y^*\eta_Y + Y^*\xi_Y\lambda_Y^*\eta_Y))\\
&+ g_Y^2(P_Y^{\mathcal{V}}(\lambda_Y), DP_Y^{\mathcal{V}}(\eta_Y)[\xi_Y]) + g^2(\mathrm{D}\, P_Y^{\mathcal{V}}(\lambda_Y)[\xi_Y], P_Y^{\mathcal{V}}(\eta_Y)) + \Re(\mathrm{tr}(\xi_Y P_Y^{\mathcal{V}}(\lambda_Y)^* P_Y^{\mathcal{V}}(\eta_Y)Y^* + YP_Y^{\mathcal{V}}(\lambda_Y)^* P_Y^{\mathcal{V}}(\eta_Y)\xi_Y^*)).
\end{aligned}
$$

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

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

Combining it with the Koszul formula with ξ, η, λ horizontal vector fields, we obtain

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

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

Recall $\mathrm{Hess}\, F(Y)[\xi_Y] = \nabla_{\xi_Y}\mathrm{grad}\, F$. For ξ_Y being a horizontal vector we have

$$
\begin{aligned}
&g_Y^3(\mathrm{Hess}\, F(Y)[\xi_Y], \eta_Y) = g_Y^3(\nabla_{\xi_Y}\mathrm{grad}\, F, \eta_Y)\\
&= \tilde{g}(\eta_Y, \mathrm{D}\,\mathrm{grad}\, F(Y)[\xi_Y]) + 2\Re(\mathrm{tr}(Y^*\xi_Y\mathrm{grad}\, F(Y)^*\eta_Y + Y^*\mathrm{grad}\, F(Y)\xi_Y^*\eta_Y))\\
&= \tilde{g}(\eta_Y, \mathrm{D}\,\mathrm{grad}\, F(Y)[\xi_Y]) + \Re(\mathrm{tr}((Y\eta_Y^* + \eta_Y Y^*)(\mathrm{grad}\, F(Y)\xi_Y^* + \xi_Y\mathrm{grad}\, F(Y)^*)))\\
&= \tilde{g}(\eta_Y, \mathrm{D}\,\mathrm{grad}\, F(Y)[\xi_Y]) + \tilde{g}\left(\eta_Y, \left(I - \tfrac{1}{2}P_Y\right)(\mathrm{grad}\, F(Y)\xi_Y^* + \xi_Y\mathrm{grad}\, F(Y)^*)Y(Y^*Y)^{-1}\right).
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{D}\,\mathrm{grad}\, F(Y)[\xi_Y] = \left(I - \frac{1}{2}P_Y\right)\nabla^2 f(YY^*)[Y\xi_Y^* + \xi_Y Y^*]Y(Y^*Y)^{-1}\\
&-\tfrac{1}{2}(\mathrm{D}\,(P_Y)[\xi_Y])\nabla f(YY^*)Y(Y^*Y)^{-1} + \left(I - \tfrac{1}{2}P_Y\right)\nabla f(YY^*)\mathrm{D}\,(Y(Y^*Y)^{-1})[\xi_Y],
\end{aligned}
$$

where we have

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

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

Combining these equations we have

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

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

$$\operatorname{Hess} F(Y)[\xi_Y] = \left(I - \tfrac{1}{2}P_Y\right)\nabla^2 f(YY^*)[Y\xi_Y^* + \xi_Y Y^*]Y(Y^*Y)^{-1} + (I - P_Y)\nabla f(YY^*)(I - P_Y)\xi_Y(Y^*Y)^{-1}.$$

References.

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

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

[3] P.-A. Absil and J. Malick, *Projection-like Retractions on Matrix Manifolds*, SIAM Journal on Optimization, 22 (2012), pp. 135–158.

[4] S. Bonnabel, G. Meyer, and R. Sepulchre, *Adaptive filtering for estimation of a low-rank positive semidefinite matrix*, in Proc. of the 19th International Symposium on Mathematical Theory of Networks and Systems, 2010.

[5] S. Bonnabel and R. Sepulchre, *Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank*, SIAM Journal on Matrix Analysis and Applications, 31 (2010), pp. 1055–1070.

[6] N. Boumal, V. Voroninski, and A. S. Bandeira, *Deterministic Guarantees for Burer-Monteiro Factorizations of Smooth Semidefinite Programs*, Communications on Pure and Applied Mathematics, 73 (2020), pp. 581–608.

[7] S. Burer and R. D. Monteiro, *Local minima and convergence in low-rank semidefinite programming*, Mathematical Programming, 103 (2005), pp. 427–444.

[8] E. J. Candes, X. Li, and M. Soltanolkotabi, *Phase retrieval via wirtinger flow: Theory and algorithms*, IEEE Transactions on Information Theory, 61 (2015), pp. 1985–2007.

[9] E. J. Candes, T. Strohmer, and V. Voroninski, *Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming*, Communications on Pure and Applied Mathematics, 66 (2013), pp. 1241–1274.

[10] L. Demanet and V. Jugnon, *Convex recovery from interferometric measurements*, IEEE Transactions on Computational Imaging, 3 (2017), pp. 282–295.

[11] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*, Springer Science & Business Media, Dec. 2012.

[12] U. Helmke and M. A. Shayman, *Critical points of matrix least squares distance functions*, Linear Algebra and its Applications, 215 (1995), pp. 1–19.
[13] W. Huang, *Optimization algorithms on Riemannian manifolds with applications*, PhD thesis, The Florida State University, 2013.
[14] W. Huang, P.-A. Absil, and K. A. Gallivan, *Intrinsic representation of tangent vectors and vector transport on matrix manifolds*, Numerische Mathematik, 136 (2017), pp. 523–543.
[15] W. Huang, K. A. Gallivan, and P.-A. Absil, *A broyden class of quasi-newton methods for riemannian optimization*, SIAM Journal on Optimization, 25 (2015), pp. 1660–1685.
[16] W. Huang, K. A. Gallivan, and X. Zhang, *Solving phaselift by low-rank riemannian optimization methods for complex semidefinite constraints*, SIAM Journal on Scientific Computing, 39 (2017), pp. B840–B859.
[17] M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre, *Low-rank optimization on the cone of positive semidefinite matrices*, SIAM Journal on Optimization, 20 (2010), pp. 2327–2351.
[18] V. Jugnon and L. Demanet, *Interferometric inversion: a robust approach to linear inverse problems*, in 2013 SEG Annual Meeting, OnePetro, 2013.
[19] D. Kressner, M. Steinlechner, and B. Vandereycken, *Low-rank tensor completion by Riemannian optimization*, BIT Numerical Mathematics, 54 (2014), pp. 447–468.
[20] J. M. Lee, *Introduction to Smooth Manifolds*, vol. 218 of Graduate Texts in Mathematics, Springer New York, New York, NY, 2012, https://doi.org/10.1007/978-1-4419-9982-5.
[21] E. 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 (2020), pp. 171–198.
[22] E. Massart, J. M. Hendrickx, and P.-A. Absil, *Curvature of the manifold of fixed-rank positive-semidefinite matrices endowed with the Bures–Wasserstein metric*, in Geometric Science of Information: 4th International Conference, GSI 2019, Toulouse, France, August 27–29, 2019, Proc., Springer, 2019, pp. 739–748.
[23] B. Mishra, *A Riemannian approach to large-scale constrained least-squares with symmetries*, PhD thesis, Universite de Liege, Liege, Belgique, 2014.
[24] B. Vandereycken, *Low-rank matrix completion by Riemannian optimization—extended version*, arXiv:1209.3834 [math], (2012).
[25] B. Vandereycken, *Low-rank matrix completion by riemannian optimization*, SIAM Journal on Optimization, 23 (2013), pp. 1214–1236.
[26] B. Vandereycken, P.-A. Absil, and S. Vandewalle, *Embedded geometry of the set of symmetric positive semidefinite matrices of fixed rank*, in 2009 IEEE/SP 15th Workshop on Statistical Signal Processing, Cardiff, United Kingdom, Aug. 2009, IEEE, pp. 389–392.
[27] 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 (2013), pp. 481–514.
[28] B. Vandereycken and S. Vandewalle, *A riemannian optimization approach for computing low-rank solutions of lyapunov equations*, SIAM Journal on Matrix Analysis and Applications, 31 (2010), pp. 2553–2579.
[29] S. Zheng, W. Huang, B. Vandereycken, and X. Zhang, *Riemannian optimization using three different metrics for hermitian psd fixed-rank constraints: an extended version*, 2022, https://arxiv.org/abs/2204.07830.