A Riemannian inexact Newton-CG method for constructing a nonnegative matrix with prescribed realizable spectrum

Zhi Zhao, Zheng-Jian Bai & Xiao-Qing Jin

Numerische Mathematik

ISSN 0029-599X
Volume 140
Number 4

Numer. Math. (2018) 140:827-855
DOI 10.1007/s00211-018-0982-2


Numerische Mathematik

Founded in 1959 by A. S. Householder, R. Sauer, E. Stiefel, and A. Walther

Editors-in-Chief: F. Brezzi, T.F. Chan and M. Griebel

Volume 140 Number 4 December 2018

Comprehensively covered by
Zentralblatt MATH
Mathematical Reviews, and Current Contents

Springer

Numerische Mathematik (2018) 140:827–855
https://doi.org/10.1007/s00211-018-0982-2

Numerische Mathematik

A Riemannian inexact Newton-CG method for constructing a nonnegative matrix with prescribed realizable spectrum

Zhi Zhao[1] · Zheng-Jian Bai[2] · Xiao-Qing Jin[3]

Received: 28 May 2017 / Revised: 26 March 2018 / Published online: 28 June 2018

Abstract
This paper is concerned with the inverse eigenvalue problem of finding a nonnegative matrix such that it has the prescribed realizable spectrum. We reformulate the inverse eigenvalue problem as an under-determined constrained nonlinear matrix equation over several matrix manifolds. Then we propose a Riemannian inexact Newton-CG method for solving the nonlinear matrix equation. The global and quadratic convergence of the proposed method is established under some assumptions. We also extend the proposed method to the case of prescribed entries. Finally, numerical experiments are reported to illustrate the efficiency of the proposed method.

Mathematics Subject Classification 65F18 · 65F15 · 15A18 · 58C15

The research of Z. Zhao is supported by the National Natural Science Foundation of China (No. 11601112). The research of Z.-J. Bai is partially supported by the National Natural Science Foundation of China (No. 11671337), the Natural Science Foundation of Fujian Province of China (No. 2016J01035), and the Fundamental Research Funds for the Central Universities (No. 20720180008). The research of X.-Q. Jin is supported by the research Grant MYRG2016-00077-FST from University of Macau..

✉ Zheng-Jian Bai
zjbai@xmu.edu.cn

Zhi Zhao
zhaozhi231@163.com

Xiao-Qing Jin
xqjin@umac.mo

1 Department of Mathematics,School of Sciences, Hangzhou Dianzi University, Hangzhou 310018, People's Republic of China

2 School of Mathematical Sciences and Fujian Provincial Key Laboratory on Mathematical Modeling & High Performance Scientific Computing, Xiamen University, Xiamen 361005, People's Republic of China

3 Department of Mathematics, University of Macau, Macao, People's Republic of China

1 Introduction

An n-by-n nonnegative matrix C is a real matrix whose entries are all greater than or equal to zero, i.e., $(C)_{ij} \geq 0$ for all $i, j = 1, \ldots, n$, where $(C)_{ij}$ denotes the (i, j)-entry of C. Nonnegative matrices arise in various applications such as Markov chains, linear complementarity problems, probabilistic algorithms, discrete distributions, categorical data, group theory, matrix scaling, and economics. See for instance [3,5,33,39] and the references therein.

The *nonnegative inverse eigenvalue problem* (NIEP) is to determine whether a given list of n complex numbers (counting multiplicities) is the spectrum of an n-by-n nonnegative matrix. The NIEP is an *inverse eigenvalue problem* and it is a classical unsolved problem in linear algebra.

A list which occurs as the spectrum of some nonnegative matrix is called a *realizable spectrum* or *realizable*, and the matrix that realizes the given list is called a *realizing matrix* [25]. The early work on the NIEP is due to Suleĭmanova [44], Karpelevič [27], and Perfect [36,37]. There has been much literature on the study of the NIEP since then. For more solvability conditions (i.e., realizability criteria) of the NIEP, one may refer to for instance [4,8,16,19,22,23,28,30,32,38,40–42]. For more comprehensive discussions on the NIEP, one may refer to [13,14,25,33,47] and the references therein. However, the real NIEP is NP-hard [7]. Also, as noted in [10, p. 18], the constructions of realizing matrices are not readily available in the literature.

The NIEP sparks the interest in constructing algorithms for finding a *realizing matrix* from a prescribed realizable spectrum. In this paper, we consider the inverse eigenvalue problem of constructing a nonnegative matrix with a prescribed realizable spectrum:

Problem I. *Given a realizable list of n complex numbers $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, find an n-by-n nonnegative matrix C such that its eigenvalues are $\lambda_1, \lambda_2, \ldots, \lambda_n$.*

There are a few numerical methods for solving **Problem I** such as constructive methods [26,35,43], the alternating projection method [34], isospectral gradient flow methods [9,11,12,15], and recursive algorithms [21,29] for the case where the prescribed realizable spectrum is a list of real numbers and satisfies some additional conditions.

Recently, results have been derived for Riemannian optimization methods for both eigenvalue problems and inverse eigenvalue problems. See for instance [1,2,46,48–50]. In this paper, we propose a Riemannian inexact Newton-CG method for solving **Problem I**. This is motivated by the recent two papers due to Dedieu, Priouret, and Malajovich [17] and Simonis [45]. In [17], by using the exponential map, Dedieu et al. presented Newton's method for finding zeros of a mapping from a Riemannian manifold to a linear space of the same dimension and the quadratic convergence was also investigated. In [45], Simonis gave some inexact Newton methods for solving an under-determined system of nonlinear equations over vector spaces. By using the real Schur decomposition of a real square matrix, we rewrite **Problem I** as an equivalent under-determined constrained nonlinear matrix equation over several matrix manifolds. Then we present a Riemannian inexact Newton-CG method for solving the under-determined constrained nonlinear matrix equation. Under some assumptions, the global and quadratic convergence property of the proposed method is established.

We also extend the proposed method to the case of prescribed entries. Numerical experiments show that the proposed method is more efficient than the alternating projection method in [34] and the Riemannian nonlinear conjugate gradient methods in [48,50].

Throughout this paper, we use the following notation. The symbols A^T and A^H denote the transpose and complex conjugate transpose of a matrix A, respectively. I_n denotes the identity matrix of order n. Let $\mathbb{R}^{n\times n}$ and $\mathbb{SR}^{n\times n}$ be the set of all n-by-n real matrices and the set of all n-by-n real symmetric matrices, respectively. Let $\mathbb{R}_+^{n\times n}$ and $\mathbb{SR}_+^{n\times n}$ denote the nonnegative orthants of $\mathbb{R}^{n\times n}$ and $\mathbb{SR}^{n\times n}$, respectively. $\|\cdot\|_F$ stands for the matrix Frobenius norm. Denote by $A \odot B$ and $[A, B] := AB - BA$ the Hadamard product and Lie Bracket of two n-by-n matrices A and B, respectively. Given a vector $\mathbf{a} \in \mathbb{R}^n$, Diag($\mathbf{a}$) denotes a diagonal matrix with $\mathbf{a}$ on its diagonal. Let vec(A) be the vectorization of a matrix A, i.e., a column vector obtained by stacking the columns of A on top of one another. Denote by tr(A) the sum of the diagonal entries of a square matrix A. Define the index set $\mathcal{N} := \{(i, j) \mid i, j = 1, \ldots, n\}$. Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite-dimensional vector spaces equipped with a scalar inner product $\langle\cdot,\cdot\rangle$ and its induced norm $\|\cdot\|$. Let $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$ be a linear operator such that $\mathcal{A}[x] \in \mathcal{Y}$ for all $x \in \mathcal{X}$ and the adjoint of $\mathcal{A}$ is denoted by $\mathcal{A}^*$. Define the operator norm of $\mathcal{A}$ by $|||\mathcal{A}||| := \sup\{\|\mathcal{A}[x]\| \mid x \in \mathcal{X} \text{ with } \|x\| = 1\}$.

The rest of this paper is organized as follows. In Sect. 2 we propose a Riemannian inexact Newton-CG method for solving **Problem I**. In Sect. 3 the global and quadratic convergence of our method is established under some assumptions. In Sect. 4, we discuss possible extensions. Finally, some numerical tests are reported in Sect. 5 and some concluding remarks are given in Sect. 6.

2 Riemannian inexact Newton-CG method

In this section, we first reformulate **Problem I** as a nonlinear matrix equation defined on a Riemannian product manifold. Then we propose a Riemannian inexact Newton-CG method for solving the nonlinear matrix equation.

2.1 Reformulation

For the two matrix sets $\mathbb{R}_+^{n\times n}$ and $\mathbb{R}^{n\times n}$ one has

$$\mathbb{R}_+^{n\times n} = \left\{ S \odot S \mid S \in \mathbb{R}^{n\times n} \right\}.$$

Since the prescribed n-tuple $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ is realizable, the set of prescribed eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ is closed under complex conjugation. Assume without loss of generality that

$$\begin{aligned} &\lambda_{2i-1} = a_i + b_i\sqrt{-1}, \quad \lambda_{2i} = a_i - b_i\sqrt{-1}, \\ &i = 1, \ldots, s; \quad \lambda_i \in \mathbb{R}, \quad i = 2s+1, \ldots, n, \end{aligned}$$

where $a_i, b_i \in \mathbb{R}$ with $b_i \neq 0$ for $i = 1, \ldots, s$. Let Λ be a block diagonal matrix defined by

$$\Lambda := \text{blkdiag}\big(\lambda_1^{[2]}, \ldots, \lambda_s^{[2]}, \lambda_{2s+1}, \ldots, \lambda_n\big),$$

where

$$\lambda_i^{[2]} := \begin{bmatrix} a_i & b_i \\ -b_i & a_i \end{bmatrix}, \quad i = 1, \ldots, s.$$

By using the real Schur decomposition for a real square matrix [24], the set of all isospectral matrices is defined by

$$\mathcal{M}(\Lambda) := \big\{X \in \mathbb{R}^{n\times n} \mid X = Q(\Lambda + V)Q^T, \ Q \in \mathcal{O}(n), V \in \mathcal{V}\big\}.$$

Here, $\mathcal{O}(n)$ represents the set of all n-by-n orthogonal matrices, i.e.,

$$\mathcal{O}(n) := \big\{Q \in \mathbb{R}^{n\times n} \mid Q^T Q = I_n\big\}.$$

The set $\mathcal{V}$ is defined by

$$\mathcal{V} := \big\{V \in \mathbb{R}^{n\times n} \mid V_{ij} = 0, \ (i, j) \in \mathcal{I}\big\},$$

where

$$\mathcal{I} := \big\{(i, j) \mid i \geq j \text{ or } \Lambda_{ij} \neq 0, \ i, j = 1, \ldots, n\big\} \subset \mathcal{N}.$$

Thus the NIEP has a solution if and only if $\mathcal{M}(\Lambda) \cap \mathbb{R}_+^{n\times n} \neq \emptyset$.

Since the prescribed n-tuple $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ is realizable, there exits at least one nonnegative matrix with the prescribed spectrum $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. Thus **Problem I** is equivalent to solving the following constrained nonlinear matrix equation

$$G(S, Q, V) = \mathbf{0}_{n\times n} \tag{2.1}$$

for $(S, Q, V) \in \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$, where $\mathbf{0}_{n\times n}$ is the zero matrix of order n and the smooth mapping $G : \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V} \to \mathbb{R}^{n\times n}$ is defined by

$$G(S, Q, V) := S \odot S - Q(\Lambda + V)Q^T, \quad (S, Q, V) \in \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}.$$

We point out that G is a smooth mapping from the product manifold $\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ to the linear space $\mathbb{R}^{n\times n}$. If we find a solution $(\overline{S}, \overline{Q}, \overline{V}) \in \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ to (2.1), then $\overline{C} := \overline{S} \odot \overline{S}$ is a solution to **Problem I**.

2.2 Riemannian inexact Newton-CG method

In [45], Simonis presented some inexact Newton methods for the under-determined system of nonlinear equations $F(\mathbf{x}) = \mathbf{0}_n$, where $F : \mathbb{R}^m \to \mathbb{R}^n$ is continuously differentiable $(m > n)$ and $\mathbf{0}_n$ is an n-vector of all zeros. Sparked by this, in this section, we propose a Riemannian inexact Newton-CG method for solving the nonlinear Eq. (2.1).

We first note that $\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ is a product manifold and as shown in "Appendix", the nonlinear matrix Eq. (2.1) is under-determined for all $n \geq 2$. It is easy to see that $\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ is an embedded submanifold of $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$ and then every tangent space $T_{(S,Q,V)}(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V})$, which is characterized as in "Appendix", can be regarded as a subspace of $T_{(S,Q,V)}(\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}) \simeq \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$, where "'$\simeq$" means the identification of two sets. Hence, the Riemannian metric of $\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ inherited from the standard inner product on $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$ is given by

$$
\begin{aligned}
& g_{(S,Q,V)}\big((\xi_1, \zeta_1, \eta_1), (\xi_2, \zeta_2, \eta_2)\big) \\
& \quad := \langle(\xi_1, \zeta_1, \eta_1), (\xi_2, \zeta_2, \eta_2)\rangle = \operatorname{tr}(\xi_1^T \xi_2) + \operatorname{tr}(\zeta_1^T \zeta_2) + \operatorname{tr}(\eta_1^T \eta_2),
\end{aligned} \tag{2.2}
$$

for all $(S, Q, V) \in \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ and

$$
(\xi_1, \zeta_1, \eta_1), (\xi_2, \zeta_2, \eta_2) \in T_{(S,Q,V)}\big(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}\big).
$$

In what follows, we denote by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ the Riemannian metric and its induced norm on $\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$, respectively.

Next, we propose a Riemannian inexact Newton-CG method for solving the under-determined matrix Eq. (2.1). As in [17], one may propose the following geometric Newton method: Given the current iterate $X^k := (S^k, Q^k, V^k) \in \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$, solve the Newton equation

$$
\mathrm{D}G(X^k)[\varDelta X^k] = -G(X^k) \tag{2.3}
$$

for $\varDelta X^k := (\varDelta S^k, \varDelta Q^k, \varDelta V^k) \in T_{X^k}(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V})$ and set

$$
X^{k+1} := R_{X^k}(\varDelta X^k),
$$

where $\mathrm{D}G(X^k)$ is the differential of G at X^k and R is a retraction on $\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$. For the explicit expressions of $\mathrm{D}G(\cdot)$ and R, see "Appendix".

Based on the analysis in section 6.11 of [31], we conclude that the linear operator Eq. (2.3) has a solution if and only if

$$
\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^\dagger[-G(X^k)] = -G(X^k), \tag{2.4}
$$

where $(\mathrm{D}G(X^k))^\dagger$ means the pseudoinverse of the linear operator $\mathrm{D}G(X^k)$ [31, p. 163]. Note that the Newton Eq. (2.3) is an under-determined linear system. If it is solvable, i.e., condition (2.4) is satisfied, then it has infinitely many solutions. In this case, the minimum norm solution of (2.3) is given by [31, pp. 163–164]:

$$\Delta X^k = -(\mathrm{D}G(X^k))^\dagger G(X^k).$$

In particular, if $\mathrm{D}G(X^k)$ is surjective, then we have [31, p. 165]:

$$(\mathrm{D}G(X^k))^\dagger = (\mathrm{D}G(X^k))^* \circ \big(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^*\big)^{-1}, \tag{2.5}$$

where $(\mathrm{D}G(X^k))^*$ is the adjoint of $\mathrm{D}G(X^k)$ with respect to the Riemannian metric $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ [31, p. 152]. For the explicit expression of $(\mathrm{D}G(\cdot))^*$, see "Appendix". Thus, if $\mathrm{D}G(X^k)$ is surjective, then condition (2.4) holds and thus one may solve the following normal equation

$$\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^*[\Delta Z] = -G(X^k), \quad \text{s.t.} \quad \Delta Z^k \in T_{G(X^k)}\mathbb{R}^{n\times n} \tag{2.6}$$

for the minimum norm solution $\Delta X^k = (\mathrm{D}G(X^k))^*[\Delta Z^k] \in T_{X^k}(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V})$. Therefore, the conjugate gradient (CG) method [24] can be used to solve the self-adjoint and positive definite Eq. (2.6).

If $\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^*$ is singular, then the equality in (2.5) does not hold. For the pseudoinverse $(\mathrm{D}G(X^k))^\dagger$, we have [31, p. 167]

$$(\mathrm{D}G(X^k))^\dagger = \lim_{\sigma\to 0^+} (\mathrm{D}G(X^k))^* \circ \big(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^* + \sigma \mathrm{id}_k\big)^{-1}, \tag{2.7}$$

where id_k denotes the identity operator on $T_{G(X^k)}\mathbb{R}^{n\times n}$. Instead of (2.6), one may solve the following perturbed normal equation

$$\big(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^* + \overline{\sigma}\mathrm{id}_k\big)[\Delta Z^k] = -G(X^k)$$

for $\Delta Z^k \in T_{G(X^k)}\mathbb{R}^{n\times n} \simeq \mathbb{R}^{n\times n}$, where $\overline{\sigma} > 0$ is a prescribed constant.

Based on the above discussion, we propose the following Riemannian inexact Newton-CG algorithm for solving the nonlinear matrix Eq. (2.1).

Algorithm 1 (Riemannian inexact Newton-CG method)

Step 0. *Choose an initial point* $X^0 \in \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$, $\overline{\epsilon} > 0$, $\overline{\sigma}_{\max}, \overline{\eta}_{\max} \in [0, 1)$, $t \in (0, 1)$, $0 < \theta_{\min} < \theta_{\max} < 1$. *Let* $k := 0$.

Step 1. *If* $\|G(X^k)\|_F < \overline{\epsilon}$, *stop.*

Step 2. *Apply the CG method to find an approximate solution* $\Delta Z^k \in T_{G(X^k)}\mathbb{R}^{n\times n}$ *to*

$$\big(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^* + \overline{\sigma}_k\mathrm{id}_k\big)[\Delta Z^k] = -G(X^k), \tag{2.8}$$

such that

$$\|(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^* + \overline{\sigma}_k\mathrm{id}_k)[\Delta Z^k] + G(X^k)\big\|_F \le \overline{\eta}_k\|G(X^k)\|_F, \tag{2.9}$$

and

$$\|\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^*[\Delta Z^k] + G(X^k)\|_F < \|G(X^k)\|_F, \tag{2.10}$$

where $\overline{\sigma}_k := \min\{\overline{\sigma}_{\max}, \|G(X^k)\|_F\}$, $\overline{\eta}_k := \min\{\overline{\eta}_{\max}, \|G(X^k)\|_F\}$. *Let*

$$\widehat{\Delta X}^k := (\mathrm{D}G(X^k))^*[\Delta Z^k] \quad \text{and} \quad \widehat{\eta}_k := \frac{\|\mathrm{D}G(X^k)[\widehat{\Delta X}^k] + G(X^k)\|_F}{\|G(X^k)\|_F}. \tag{2.11}$$

Step 3. *Evaluate* $G\big(R_{X^k}(\widehat{\Delta X}^k)\big)$. *Set* $\eta_k = \widehat{\eta}_k$ *and* $\Delta X^k = \widehat{\Delta X}^k$.
Repeat *until* $\|G\big(R_{X^k}(\Delta X^k)\big)\|_F \leq (1 - t(1 - \eta_k))\|G(X^k)\|_F$.
Choose $\theta \in [\theta_{\min}, \theta_{\max}]$.
Replace ΔX^k *by* $\theta \Delta X^k$ *and* η_k *by* $1 - \theta(1 - \eta_k)$.
end (Repeat)
Set

$$X^{k+1} := R_{X^k}(\Delta X^k).$$

Step 4. *Replace* k *by* $k + 1$ *and go to Step 1.*

We now make several remarks on Algorithm 1. We first note that the new iterate for Newton's method in [17] is updated by using the exponential map while in Algorithm 1, the new iterate X^{k+1} is updated by using a retraction on $\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ defined as in "Appendix" instead of the exponential map on $\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$, which is in general computationally costly [1, p. 59 and p. 103]. In addition, in Step 3 of Algorithm 1, one needs to choose a scaling factor $\theta \in [\theta_{\min}, \theta_{\max}]$. As in [45], one may choose θ by employing the quadratic backtracking method (see also [18]). Let

$$T(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}) = \cup_{X \in \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}} T_X(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V})$$

be the tangent bundle of $\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ [1, p. 36]. The pullback $\widehat{G}$ of G is a smooth mapping from $T(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V})$ to $\mathbb{R}^{n\times n}$ defined by

$$\widehat{G}(\xi) := G(R(\xi)), \quad \forall \xi \in T(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}). \tag{2.12}$$

The restriction of $\widehat{G}$ on $T_X(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V})$ for $X \in \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ is defined by

$$\widehat{G}_X(\xi_X) = G(R_X(\xi_X)), \quad \forall \xi_X \in T_X(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}). \tag{2.13}$$

Then one has

$$\mathrm{D}G(X) = \mathrm{D}\widehat{G}_X(0_X), \quad \forall X \in \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}, \tag{2.14}$$

where 0_X is the origin of $T_X(\mathbb{R}^{n\times n}\times\mathcal{O}(n)\times\mathcal{V})$. Next, we find an approximate minimizer of the cost function

$$u(\theta) := \|G(R_{X^k}(\theta\Delta X^k))\|_F^2 = \|\widehat{G}_{X^k}(\theta\Delta X^k)\|_F^2.$$

Define a quadratic polynomial by

$$q(\theta) := (u(1)-u(0)-u'(0))\theta^2 + u'(0)\theta + u(0),$$

where

$$u(0) = \|\widehat{G}_{X^k}(0_{X^k})\|_F^2 = \|G(X^k)\|_F^2, \quad u(1) = \|\widehat{G}_{X^k}(\Delta X^k)\|_F^2 = \|G(R_{X^k}(\Delta X^k))\|_F^2,$$
$$u'(0) = 2\langle \mathrm{D}\widehat{G}_{X^k}(0_{X^k})[\Delta X^k], \widehat{G}_{X^k}(0_{X^k})\rangle = 2\langle \mathrm{D}G(X^k)[\Delta X^k], G(X^k)\rangle.$$

Obviously, the values of $u(0)$ and $u(1)$ have been evaluated in Algorithm 1 and it is not so complicated to compute $u'(0)$. It is easy to check that

$$q'(\theta) = 2(u(1)-u(0)-u'(0))\theta + u'(0) \quad \text{and} \quad q''(\theta) = 2(u(1)-u(0)-u'(0)).$$

If $q''(\theta)\le 0$, then the quadratic polynomial q is concave and we choose $\theta=\theta_{\max}$. If $q''(\theta)>0$, then the minimizer of q is reached at the point θ satisfying $q'(\theta)=0$, i.e.,

$$\theta = \frac{-u'(0)}{2(u(1)-u(0)-u'(0))}.$$

Since we require $\theta\in[\theta_{\min},\theta_{\max}]$, the approximate minimizer θ of u is given by

$$\theta = \min\left\{\max\left\{\theta_{\min}, \frac{-u'(0)}{2(u(1)-u(0)-u'(0))}\right\}, \theta_{\max}\right\}.$$

3 Convergence analysis

In this section, we establish the global and quadratic convergence of Algorithm 1. Notice that $\mathbb{R}^{n\times n}$ and $\mathcal{V}$ are two linear matrix manifolds and $\mathcal{O}(n)$ is a compact manifold. For the retraction R on $\mathbb{R}^{n\times n}\times\mathcal{O}(n)\times\mathcal{V}$ defined as in "Appendix", there exist two scalars $\nu>0$ and $\mu_\nu>0$ such that [1, p. 149]

$$\nu\|\Delta X\| \ge \mathrm{dist}\big(X, R_X(\Delta X)\big), \tag{3.1}$$

for all $X := (S, Q, V)\in\mathbb{R}^{n\times n}\times\mathcal{O}(n)\times\mathcal{V}$ and

$$\Delta X := (\Delta S, \Delta Q, \Delta V)\in T_X(\mathbb{R}^{n\times n}\times\mathcal{O}(n)\times\mathcal{V}) \quad \text{with} \quad \|\Delta X\|\le\mu_\nu, \tag{3.2}$$

where "dist" means the Riemannian distance on $\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$. As in [20], a point $X \in \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ is called a *stationary point* of $\|G\|_F$ if

$$\|G(X)\|_F \leq \|G(X) + \mathrm{D}G(X)[\Delta X]\|_F, \quad \forall \Delta X \in T_X(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}).$$

Thus X is a stationary point if and only if grad $\left(1/2\|G(X)\|_F^2\right) = 0_X$, i.e.,

$$\operatorname{grad}\left(\frac{1}{2}\|G(X)\|_F^2\right) = (\mathrm{D}G(X))^*[G(X)] = 0_X. \tag{3.3}$$

To establish the global and quadratic convergence of Algorithm 1, we need the following assumption.

Assumption 1 Suppose Algorithm 1 does not break down, $\sum_{k=0}^{\infty}(1-\eta_k)$ is divergent and $\mathrm{D}G(\overline{X}) : T_{\overline{X}}(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}) \to T_{G(\overline{X})}\mathbb{R}^{n\times n}$ is surjective, where $\overline{X} \in \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ is an accumulation point of the sequence $\{X^k\}$ generated by Algorithm 1.

The next theorems present the global and quadratic convergence result of Algorithm 1.

Theorem 2 *Suppose Assumption 1 is satisfied. Let $\overline{X}$ be an accumulation point of the sequence $\{X^k\}$ generated by Algorithm 1. Then the sequence $\{X^k\}$ converges to $\overline{X}$ and $G(\overline{X}) = \mathbf{0}_{n\times n}$.*

Theorem 3 *Suppose Assumption 1 is satisfied. Let $\overline{X}$ be an accumulation point of the sequence $\{X^k\}$ generated by Algorithm 1. Then the sequence $\{X^k\}$ converges to $\overline{X}$ quadratically.*

The following result gives the surjectivity conditions of $\mathrm{D}G(\cdot)$ at $\overline{X}$, where $\overline{X}$ is an accumulation point of the sequence $\{X^k\}$ generated by Algorithm 1.

Theorem 4 *Suppose Algorithm 1 generates an infinite sequence $\{X^k := (S^k, Q^k, V^k)\}$. Let $\overline{X} := (\overline{S}, \overline{Q}, \overline{V}) \in \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ be an accumulation point of $\{X^k\}$. Then the linear operator $\mathrm{D}G(\overline{X})$ is surjective if and only if*

$$\operatorname{null}\left(\begin{bmatrix} \operatorname{Diag}\big(\operatorname{vec}(\overline{S})\big) \\ (I_{n^2} - \widehat{P})\big((\overline{S}\odot\overline{S})\otimes I_n - I_n \otimes (\overline{S}\odot\overline{S})^T\big) \\ \operatorname{Diag}\big(\operatorname{vec}(W)\big)(\overline{Q}\otimes\overline{Q})^T \end{bmatrix}\right) = \{\mathbf{0}_{n^2}\}, \tag{3.4}$$

where $W \in \mathbb{R}^{n\times n}$ is defined by (A.1) in "Appendix" and $\widehat{P} \in \mathbb{R}^{n^2\times n^2}$ is the vectorized transpose matrix such that

$$\operatorname{vec}(A^T) = \widehat{P}\operatorname{vec}(A), \quad \forall A \in \mathbb{R}^{n\times n}.$$

In the rest of this section, we show the global and quadratic convergence of Algorithm 1 and the surjectivity of $\mathrm{D}G(\cdot)$, respectively.

3.1 Global convergence

To prove the global convergence of Algorithm 1, we need some preliminary lemmas.

Lemma 1 *Let X^k be the current iterate generated by Algorithm 1. If X^k is not a stationary point of $\|G\|_F$, then conditions (2.9) and (2.10) are attainable by solving (2.8) accurately enough.*

Proof Since the operator $\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^* + \overline{\sigma}_k \mathrm{id}_k$ is positive definite, the CG method is convergent for the linear Eq. (2.8). Thus condition (2.9) can be satisfied if the number of CG iterations is large enough. Let $\overline{\Delta Z}^k$ be the exact solution of (2.8), i.e.,

$$\overline{\Delta Z}^k := -\big(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^* + \overline{\sigma}_k \mathrm{id}_k\big)^{-1}[G(X^k)].$$

Define $\overline{\Delta X}^k := (\mathrm{D}G(X^k))^*[\overline{\Delta Z}^k]$. Thus,

$$\begin{aligned}
&\mathrm{D}G(X^k)[\overline{\Delta X}^k] + G(X^k)\\
&\quad = \mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^*[\overline{\Delta Z}_k] + G(X^k)\\
&\quad = -\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^* \circ \big(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^* + \overline{\sigma}_k \mathrm{id}_k\big)^{-1}[G(X^k)] + G(X^k)\\
&\quad = \Big(\mathrm{id}_k - \mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^* \circ \big(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^* + \overline{\sigma}_k \mathrm{id}_k\big)^{-1}\Big)[G(X^k)].
\end{aligned} \tag{3.5}$$

Since X^k is not a stationary point of $\|G\|_F$, it follows from (3.3) that

$$G(X^k) \notin \mathrm{null}\big((\mathrm{D}G(X))^*\big). \tag{3.6}$$

Now using (3.5) and (3.6) we find

$$\|\mathrm{D}G(X^k)[\overline{\Delta X}^k] + G(X^k)\|_F < \|G(X^k)\|_F.$$

This shows that condition (2.10) is satisfied if the number of CG iterations is large enough. □

For the iterate $\widehat{\Delta X}^k$ generated by Algorithm 1, we have the following estimate.

Lemma 2 *Let X^k be the current iterate generated by Algorithm 1. If X^k is not a stationary point of $\|G\|_F$, then*

$$\|\widehat{\Delta X}^k\| \le (1 + \overline{\eta}_k)\|\|(\mathrm{D}G(X^k))^\dagger\|\| \cdot \|G(X^k)\|_F.$$

Proof Since X^k is not a stationary point of $\|G\|_F$, it follows from Lemma 1 that conditions (2.9) and (2.10) are satisfied if one finds an approximate solution to (2.8) with sufficient accuracy. Let

$$J(X^k) := \mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^* + \overline{\sigma}_k \mathrm{id}_k \quad \text{and} \quad V(X^k) := G(X^k) + J(X^k)[\Delta Z^k].$$

From (2.9) we have

$$\|V(X^k)\|_F \le \overline{\eta}_k \|G(X^k)\|_F. \tag{3.7}$$

By using (2.11), $\widehat{\Delta X}^k = (\mathrm{D}G(X^k))^*[\Delta Z^k]$ and thus $\widehat{\Delta X}^k \perp \mathrm{null}(\mathrm{D}G(X^k))$. Using Lemma 4 in page 17 of [45], we have

$$\|\widehat{\Delta X}^k\| = \|(\mathrm{D}G(X^k))^\dagger \circ \mathrm{D}G(X^k)[\widehat{\Delta X}^k]\|. \tag{3.8}$$

It follows from (2.11), (3.7), and (3.8) that

$$\begin{aligned}
\|\widehat{\Delta X}^k\| &= \|(\mathrm{D}G(X^k))^\dagger \circ \mathrm{D}G(X^k)[\widehat{\Delta X}^k]\| \\
&= \|(\mathrm{D}G(X^k))^\dagger \circ \mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^*[\Delta Z^k]\| \\
&\le |\!|\!|(\mathrm{D}G(X^k))^\dagger|\!|\!| \cdot \|\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^*[\Delta Z^k]\|_F \\
&\le |\!|\!|(\mathrm{D}G(X^k))^\dagger|\!|\!| \cdot \|\big(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^* + \overline{\sigma}_k \mathrm{id}_k\big)[\Delta Z^k]\|_F \\
&= |\!|\!|(\mathrm{D}G(X^k))^\dagger|\!|\!| \cdot \|J(X^k)[\Delta Z^k]\|_F \\
&= |\!|\!|(\mathrm{D}G(X^k))^\dagger|\!|\!| \cdot \|V(X^k) - G(X^k)\|_F \\
&\le |\!|\!|(\mathrm{D}G(X^k))^\dagger|\!|\!| \cdot \big(\|V(X^k)\|_F + \|G(X^k)\|_F\big) \\
&= (1 + \overline{\eta}_k)|\!|\!|(\mathrm{D}G(X^k))^\dagger|\!|\!| \cdot \|G(X^k)\|_F.
\end{aligned}$$

□

For the upper bound of the iterate $\widehat{\eta}_k$ generated by Algorithm 1, we have the following result.

Lemma 3 *Let X^k be the current iterate generated by Algorithm 1. If X^k is not a stationary point of $\|G\|_F$, then*

$$\widehat{\eta}_k \le \frac{\overline{\sigma}_k}{\lambda_{\min}\big(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^*\big) + \overline{\sigma}_k} + \overline{\eta}_k \quad \text{and} \quad \widehat{\eta}_k < 1, \tag{3.9}$$

where $\lambda_{\min}(\cdot)$ means the smallest eigenvalue of a positive semidefinite linear operator.

Proof Note that X^k is not a stationary point of $\|G\|_F$. Lemma 1 tells us that conditions (2.9) and (2.10) are satisfied if one finds an approximate solution to (2.8) with sufficient accuracy. Let $J(X^k)$ and $V(X^k)$ be defined as in Lemma 2. Using (2.11) and (3.7) we see that

$$\begin{aligned}
&\|G(X^k) + \mathrm{D}G(X^k)[\widehat{\Delta X}^k]\|_F \\
&\quad = \|G(X^k) + \mathrm{D}G(X^k)\big[(\mathrm{D}G(X^k))^*[\Delta Z_k]\big]\|_F \\
&\quad = \|G(X^k) + \big(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^*\big) \circ (J(X^k))^{-1}[V(X^k) - G(X^k)]\|_F \\
&\quad \le |\!|\!|\mathrm{id} - \big(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^*\big) \circ (J(X^k))^{-1}|\!|\!| \cdot \|G(X^k)\|_F \\
&\qquad + |\!|\!|\big(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^*\big) \circ (J(X^k))^{-1}|\!|\!| \cdot \|V(X^k)\|_F \\
&\quad \le \left(\frac{\overline{\sigma}_k}{\lambda_{\min}\big(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^*\big) + \overline{\sigma}_k} + \overline{\eta}_k\right)\|G(X^k)\|_F.
\end{aligned}$$

This, together with (2.10), yields (3.9). □

For the repeat-loop in Algorithm 1, we have the following lemma.

Lemma 4 *Let X^k be the current iterate generated by Algorithm 1. If X^k is not a stationary point of $\|G\|_F$, then the repeat-loop terminates in finite steps with ΔX^k and η_k satisfying*

$$\begin{cases} \|G(X^k) + \mathrm{D}G(X^k)[\Delta X^k]\|_F \leq \eta_k \|G(X^k)\|_F, \\ \|G(X^{k+1})\|_F \leq \big(1 - t(1-\eta_k)\big)\|G(X^k)\|_F. \end{cases} \tag{3.10}$$

Proof Note that X^k is not a stationary point of $\|G\|_F$. By Lemma 1, conditions (2.9) and (2.10) are satisfied if one finds an approximate solution to (2.8) with sufficient accuracy. By using Lemma 3, $\widehat{\eta}_k < 1$. In the repeat-loop, the search direction ΔX^k is scaled by some $\theta_j \in [\theta_{\min}, \theta_{\max}]$ at the j-th step. Hence, at the m-th step of the repeat-loop,

$$\Delta X^k = \prod_{j=1}^{m} \theta_j \widehat{\Delta X}^k \quad \text{and} \quad \eta_k = 1 - \prod_{j=1}^{m} \theta_j (1 - \widehat{\eta}_k).$$

It is clear that

$$\Theta_m := \prod_{j=1}^{m} \theta_j \leq \prod_{j=1}^{m} \theta_{\max} = \theta_{\max}^m, \tag{3.11}$$

where $\Theta_m := 1$ for $m = 0$.

We note that G is continuously differentiable and $0 < \theta_{\max} < 1$. Using (2.13) and (2.14) we obtain for all m sufficiently large,

$$\|G\big(R_{X^k}(\Theta_m \widehat{\Delta X}^k)\big) - G(X^k) - \mathrm{D}G(X^k)[\Theta_m \widehat{\Delta X}^k]\|_F \leq \epsilon_k \|\Theta_m \widehat{\Delta X}^k\|,$$

and thus

$$\|\widehat{G}_{X^k}(\Theta_m \widehat{\Delta X}^k) - \widehat{G}_{X^k}(0_{X^k}) - \mathrm{D}\widehat{G}_{X^k}(0_{X^k})[\Theta_m \widehat{\Delta X}^k]\|_F \leq \epsilon_k \|\Theta_m \widehat{\Delta X}^k\|, \tag{3.12}$$

where $\epsilon_k := ((1-t)(1-\widehat{\eta}_k))/((1+\overline{\eta}_{\max})\|\|(\mathrm{D}G(X^k))^\dagger\|\|)$.

We now show that the repeat-loop terminates in finite steps. Let $\widehat{m}$ be the smallest integer such that (3.12) holds. Let $\Delta X^k := \Theta_{\widehat{m}} \widehat{\Delta X}^k$. From (2.11), (2.13), and (2.14), it follows that

$$\begin{aligned} &\|G(X^k) + \mathrm{D}G(X^k)[\Delta X^k]\|_F \\ &\quad = \|\widehat{G}_{X^k}(0_{X^k}) + \mathrm{D}\widehat{G}_{X^k}(0_{X^k})[\Delta X^k]\|_F \\ &\quad = \|(1-\Theta_{\widehat{m}})\widehat{G}_{X^k}(0_{X^k}) + \Theta_{\widehat{m}}\widehat{G}_{X^k}(0_{X^k}) + \Theta_{\widehat{m}}\mathrm{D}\widehat{G}_{X^k}(0_{X^k})[\widehat{\Delta X}^k]\|_F \\ &\quad \leq (1-\Theta_{\widehat{m}})\|\widehat{G}_{X^k}(0_{X^k})\|_F + \Theta_{\widehat{m}}\|\widehat{G}_{X^k}(0_{X^k}) + \mathrm{D}\widehat{G}_{X^k}(0_{X^k})[\widehat{\Delta X}^k]\|_F \end{aligned}$$

$$
\begin{aligned}
&= (1-\Theta_{\widehat{m}})\|\widehat{G}_{X^k}(0_{X^k})\|_F + \Theta_{\widehat{m}}\widehat{\eta}_k\|\widehat{G}_{X^k}(0_{X^k})\|_F \\
&= \big(1-\Theta_{\widehat{m}}+\Theta_{\widehat{m}}\widehat{\eta}_k\big)\|\widehat{G}_{X^k}(0_{X^k})\|_F \\
&= \big(1-\Theta_{\widehat{m}}(1-\widehat{\eta}_k)\big)\|G(X^k)\|_F \\
&= \eta_k\|G(X^k)\|_F. \qquad (3.13)
\end{aligned}
$$

This, together with Lemmas 2 and 3, (2.11), and (3.12), yields

$$
\begin{aligned}
\|G(X^{k+1})\|_F &= \|\widehat{G}_{X^k}(\Delta X^k)\|_F \\
&\le \|\widehat{G}_{X^k}(0_{X^k}) + \mathrm{D}\widehat{G}_{X^k}(0_{X^k})[\Delta X^k]\|_F \\
&\quad + \|\widehat{G}_{X^k}(\Delta X^k) - \widehat{G}_{X^k}(0_{X^k}) - \mathrm{D}\widehat{G}_{X^k}(0_{X^k})[\Delta X^k]\|_F \\
&\le \eta_k\|\widehat{G}_{X^k}(0_{X^k})\|_F + \epsilon_k\Theta_{\widehat{m}}\|\widehat{\Delta X}^k\|_F \\
&\le \eta_k\|\widehat{G}_{X^k}(0_{X^k})\|_F + \epsilon_k\Theta_{\widehat{m}}(1+\overline{\eta}_k)|||(\mathrm{D}G(X^k))^\dagger|||\cdot\|G(X^k)\|_F \\
&\le \eta_k\|\widehat{G}_{X^k}(0_{X^k})\|_F + \epsilon_k\Theta_{\widehat{m}}(1+\overline{\eta}_{\max})|||(\mathrm{D}G(X^k))^\dagger|||\cdot\|G(X^k)\|_F \\
&= \big(\eta_k + \epsilon_k\Theta_{\widehat{m}}(1+\overline{\eta}_{\max})|||(\mathrm{D}G(X^k))^\dagger|||\big)\|G(X^k)\|_F \\
&= \left(\eta_k + \Theta_{\widehat{m}}\frac{(1-t)(1-\widehat{\eta}_k)}{(1+\overline{\eta}_{\max})|||(\mathrm{D}G(X^k))^\dagger|||}(1+\overline{\eta}_{\max})|||(\mathrm{D}G(X^k))^\dagger|||\right)\|G(X^k)\|_F \\
&= \big(\eta_k + \Theta_{\widehat{m}}(1-t)(1-\widehat{\eta}_k)\big)\|G(X^k)\|_F \\
&= \Big(\eta_k + 1 - \big(1-\Theta_{\widehat{m}}(1-\widehat{\eta}_k)\big) - t + t\big(1-\Theta_{\widehat{m}}(1-\widehat{\eta}_k)\big)\Big)\|G(X^k)\|_F \\
&= \big(\eta_k + 1 - \eta_k - t + t\eta_k\big)\|G(X^k)\|_F \\
&= \big(1 - t(1-\eta_k)\big)\|G(X^k)\|_F. \qquad (3.14)
\end{aligned}
$$

□

From Lemmas 1 and 4, Algorithm 1 never breaks down if X^k is not a stationary point of $\|G\|_F$.

We now give the proof of Theorem 2.

Proof of Theorem 2 By hypothesis, $\mathrm{D}G(\overline{X})$ is surjective. In addition, G is continuously differentiable. Define $\overline{\lambda}_{\min} := \lambda_{\min}\big(\mathrm{D}G(X^k)\circ(\mathrm{D}G(X^k))^*\big)$. Thus, there exists a sufficiently small constant $\overline{\delta} > 0$ such that for any X in a ball $B_{\overline{\delta}}(\overline{X})$ of $\overline{X}$, the linear operator $\mathrm{D}G(X)$ is surjective and

$$
|||(\mathrm{D}G(X))^\dagger||| \le 2|||(\mathrm{D}G(\overline{X}))^\dagger|||, \quad \lambda_{\min}(\mathrm{D}G(X)\circ(\mathrm{D}G(X))^*) \ge \frac{1}{2}\overline{\lambda}_{\min} > 0. \qquad (3.15)
$$

By assumption, $\sum_{k=0}^{\infty}(1-\eta_k)$ is divergent. This, together with (3.10), implies that

$$
\begin{aligned}
\|G(X^{k+1})\|_F &\le \big(1-t(1-\eta_k)\big)\|G(X^k)\|_F \le \|G(X^0)\|_F\prod_{0\le l\le k}\big(1-t(1-\eta_l)\big) \\
&\le \|G(X^0)\|_F\exp\Big(-t\sum_{0\le l\le k}(1-\eta_l)\Big) \to 0, \quad k\to\infty.
\end{aligned}
$$

□

Thus,

$$\lim_{k\to\infty} \|G(X^{k+1})\|_F = 0. \tag{3.16}$$

Since $\overline{X}$ is an accumulation point of $\{X^k\}$ and G is continuously differentiable, it follows from (3.16) that $G(\overline{X}) = \mathbf{0}_{n\times n}$. By using the definitions of $\overline{\sigma}_k$ and $\overline{\eta}_k$ in Algorithm 1 and (3.9), we get that for any $X^k \in B_{\overline{\delta}}(\overline{X})$,

$$\begin{aligned}\widehat{\eta}_k &\le \frac{\overline{\sigma}_k}{\lambda_{\min}\big(DG(X^k)\circ(DG(X^k))^*\big)+\overline{\sigma}_k}+\overline{\eta}_k\\ &\le \frac{1}{\frac{1}{2}\overline{\lambda}_{\min}+\overline{\sigma}_k}\overline{\sigma}_k+\overline{\eta}_k \le \Big(\frac{2}{\overline{\lambda}_{\min}}+1\Big)\|G(X^k)\|_F.\end{aligned} \tag{3.17}$$

According to (3.16) and (3.17), for a given constant $0 < \widehat{\eta}_{\max} < 1$, there exists a constant $0 < \widehat{\delta} < \overline{\delta}$ such that

$$\widehat{\eta}_k \le \widehat{\eta}_{\max} < 1, \quad \forall X^k \in B_{\widehat{\delta}}(\overline{X}). \tag{3.18}$$

Let $\epsilon := ((1-t)(1-\widehat{\eta}_{\max}))/(2((1+\overline{\eta}_{\max})|||\mathrm{D}G(\overline{X}))^\dagger|||)$. Then there exist two constants $\delta_1 > 0$ and $\mu_1 > 0$ such that

$$\|\widehat{G}_X(\Delta X) - \widehat{G}_X(0_X) - \mathrm{D}\widehat{G}_X(0_X)[\Delta X]\|_F \le \epsilon\|\Delta X\| \tag{3.19}$$

for all $X \in B_{\delta_1}(\overline{X})$ and $\|\Delta X\| \le \mu_1$. Let $\delta = \min\{\widehat{\delta}, \delta_1\}$. Since $\overline{X}$ is an accumulation point of the sequence $\{X^k\}$, there exist infinitely many k such that $X^k \in B_\delta(\overline{X})$. Let $\widehat{m}$ be the smallest integer such that

$$2\theta_{\max}^{\widehat{m}}(1+\overline{\eta}_{\max})|||(\mathrm{D}G(\overline{X}))^\dagger|||\cdot\|G(X^0)\|_F < \mu_1.$$

Let $\Theta_{\widehat{m}} := \prod_{i=1}^{\widehat{m}}\theta_i$. By using Lemma 2, (3.15), and the above inequality we obtain

$$\begin{aligned}\|\Theta_{\widehat{m}}\widehat{\Delta X}^k\| &\le \theta_{\max}^{\widehat{m}}\|\widehat{\Delta X}^k\|\\ &\le \theta_{\max}^{\widehat{m}}(1+\overline{\eta}_k)|||(\mathrm{D}G(X^k))^\dagger|||\cdot\|G(X^k)\|_F\\ &\le 2\theta_{\max}^{\widehat{m}}(1+\overline{\eta}_{\max})|||(\mathrm{D}G(\overline{X}))^\dagger|||\cdot\|G(X^0)\|_F\\ &< \mu_1\end{aligned} \tag{3.20}$$

for $X^k \in B_\delta(\overline{X})$. This, together with (3.19), yields

$$\|\widehat{G}_{X^k}(\Theta_{\widehat{m}}\widehat{\Delta X}^k) - \widehat{G}_{X^k}(0_{X^k}) - \mathrm{D}\widehat{G}_{X^k}(0_{X^k})[\Theta_{\widehat{m}}\widehat{\Delta X}^k]\|_F \le \epsilon\|\Theta_{\widehat{m}}\widehat{\Delta X}^k\|$$

for $X^k \in B_\delta(\overline{X})$. By using Lemma 4, the repeat-loop terminates in at most $\widehat{m}$ steps. Then for any $X^k \in B_\delta(\overline{X})$,

$$1-\eta_k = \Theta_{\widehat{m}}(1-\widehat{\eta}_k) \ge \theta_{\min}^{\widehat{m}}(1-\widehat{\eta}_{\max}) > 0. \tag{3.21}$$

Next, we show that $\{X^k\}$ converges to $\overline{X}$. We know from (3.18), (3.20), and (3.21) that for $X^k \in B_\delta(\overline{X})$,

$$\begin{aligned}\|\Delta X^k\| &\leq \theta_{\max}^{\widehat{m}} \|\widehat{\Delta X}^k\| \\ &\leq 2\theta_{\max}^{\widehat{m}}(1+\overline{\eta}_{\max})\|\|(\mathrm{D}G(\overline{X}))^\dagger\|\| \cdot \|G(X^k)\|_F \\ &\leq \frac{2\theta_{\max}^{\widehat{m}}(1+\overline{\eta}_{\max})\|\|(\mathrm{D}G(\overline{X}))^\dagger\|\|}{\theta_{\min}^{\widehat{m}}(1-\widehat{\eta}_{\max})}(1-\eta_k)\|G(X^k)\|_F. \end{aligned} \tag{3.22}$$

By using (3.16) and (3.22) we find

$$\lim_{k\to\infty} \|\Delta X^k\| = 0. \tag{3.23}$$

Thus for all k sufficiently large with $X^k \in B_\delta(\overline{X})$,

$$\|\Delta X^k\| \leq \mu_\nu, \tag{3.24}$$

where μ_ν is the constant given in (3.2).

We now show that the sequence $\{X^k\}$ converges to $\overline{X}$. By contradiction, suppose the sequence $\{X^k\}$ does not converge to $\overline{X}$. Then there exist infinitely many k such that $X^k \notin B_\delta(\overline{X})$. Since $\overline{X}$ is an accumulation point of $\{X^k\}$, there exist two index sets $\{m_j\}$ and $\{n_j\}$ such that $\lim_{j\to\infty} X^{m_j} = \overline{X}$, and for each j,

$$\begin{cases} X^{m_j} \in B_\delta(\overline{X}), & X^{m_j+i} \in B_\delta(\overline{X}), \quad i = 0, \ldots, n_j - 1, \\ X^{m_j+n_j} \notin B_\delta(\overline{X}), & m_j + n_j < m_{j+1}. \end{cases}$$

Thus, it follows from (3.1), (3.10), (3.22), and (3.24) that

$$\begin{aligned} \frac{\delta}{2} &\leq \mathrm{dist}(X^{m_j+n_j}, X^{m_j}) \leq \sum_{k=m_j}^{m_j+n_j-1} \mathrm{dist}(X^{k+1}, X^k) \\ &= \sum_{k=m_j}^{m_j+n_j-1} \mathrm{dist}\big(R_{X^k}(\Delta X^k), X^k\big) \leq \sum_{k=m_j}^{m_j+n_j-1} \nu\|\Delta X^k\| \\ &\leq \sum_{k=m_j}^{m_j+n_j-1} \nu \frac{2\theta_{\max}^{\widehat{m}}(1+\overline{\eta}_{\max})\|\|(\mathrm{D}G(\overline{X}))^\dagger\|\|}{\theta_{\min}^{\widehat{m}}(1-\widehat{\eta}_{\max})}(1-\eta_k)\|G(X^k)\|_F \\ &\leq \sum_{k=m_j}^{m_j+n_j-1} \frac{2\nu\theta_{\max}^{\widehat{m}}(1+\overline{\eta}_{\max})\|\|(\mathrm{D}G(\overline{X}))^\dagger\|\|}{\theta_{\min}^{\widehat{m}}(1-\widehat{\eta}_{\max})} \times \frac{\|G(X^k)\|_F - \|G(X^{k+1})\|_F}{t} \\ &= \frac{2\nu\theta_{\max}^{\widehat{m}}(1+\overline{\eta}_{\max})\|\|(\mathrm{D}G(\overline{X}))^\dagger\|\|}{t\theta_{\min}^{\widehat{m}}(1-\widehat{\eta}_{\max})}\big(\|G(X^{m_j})\|_F - \|G(X^{m_j+n_j})\|_F\big) \end{aligned}$$

$$\le \frac{2\nu\theta_{\max}^{\widehat{m}}(1+\overline{\eta}_{\max})\|\|(\mathrm{D}G(\overline{X}))^{\dagger}\|\|}{t\theta_{\min}^{\widehat{m}}(1-\widehat{\eta}_{\max})}\left(\|G(X^{m_j})\|_F - \|G(X^{m_{j+1}})\|_F\right)$$
$$\to 0, \quad \text{as} \quad j \to \infty,$$

since $X^{m_j} \to \overline{X}$ as $j \to \infty$. This is a contradiction. Therefore, the sequence $\{X^k\}$ converges to $\overline{X}$. □

3.2 Quadratic convergence

We show the quadratic convergence of Algorithm 1. First, we have the following result on the backtracking line search procedure.

Lemma 5 *Suppose Assumption 1 is satisfied. Let $\overline{X}$ be an accumulation point of the sequence $\{X^k\}$ generated by Algorithm 1. Then $\eta_k = \widehat{\eta}_k$ and $\Delta X^k = \widehat{\Delta X}^k$ for all k sufficiently large.*

Proof We note that G is continuously differentiable. By hypothesis, $\mathrm{D}G(\overline{X})$ is surjective. Theorem 2 tells us that the sequence $\{X^k\}$ converges to $\overline{X}$ with $G(\overline{X}) = \mathbf{0}_{n\times n}$. From (3.15), it follows that for all k sufficiently large, $\mathrm{D}G(X^k)$ is surjective and satisfies

$$\|\|(\mathrm{D}G(X^k))^{\dagger}\|\| \le 2\|\|(\mathrm{D}G(\overline{X}))^{\dagger}\|\|.$$

From Lemma 2 and the definition of $\overline{\eta}_k$ in Algorithm 1, we have that for all k sufficiently large,

$$\begin{aligned}
\|\widehat{\Delta X}^k\| &\le (1+\overline{\eta}_k)\|\|(\mathrm{D}G(X^k))^{\dagger}\|\| \cdot \|G(X^k)\|_F \\
&\le (1+\overline{\eta}_{\max})\|\|(\mathrm{D}G(X^k))^{\dagger}\|\| \cdot \|G(X^k)\|_F \\
&\le 2(1+\overline{\eta}_{\max})\|\|(\mathrm{D}G(\overline{X}))^{\dagger}\|\| \cdot \|G(X^k)\|_F .
\end{aligned}$$

By using (3.16) and the above inequality, we obtain $\lim_{k\to\infty}\|\widehat{\Delta X}^k\| = 0$. Hence, for all k sufficiently large, it holds that

$$\|\widehat{G}_{X^k}(\widehat{\Delta X}^k) - \widehat{G}_{X^k}(0_{X^k}) - \mathrm{D}\widehat{G}_{X^k}(0_{X^k})[\widehat{\Delta X}^k]\|_F \le \epsilon\|\widehat{\Delta X}^k\| \le \epsilon_k\|\widehat{\Delta X}^k\|, \tag{3.25}$$

where the condition $\epsilon \le \epsilon_k$ is used with ϵ_k and ϵ being defined in (3.12) and (3.19). Using (3.25), condition (3.12) is satisfied for $m = 0$ and $\Theta_m = 1$, where Θ_m is defined in (3.11). By using Lemma 4 the inequalities in (3.13) and (3.14) are satisfied for $\eta_k = \widehat{\eta}_k$ and $\Delta X^k = \widehat{\Delta X}^k$. This shows that $\eta_k = \widehat{\eta}_k$ and $\Delta X^k = \widehat{\Delta X}^k$ for all k sufficiently large. □

We now give the proof of Theorem 3.

Proof of Theorem 3 By using Theorem 2 and Lemma 5 we know that the sequence $\{X^k\}$ converges to $\overline{X}$ with $G(\overline{X}) = \mathbf{0}_{n\times n}$ and $\eta_k = \widehat{\eta}_k$ and $\Delta X^k = \widehat{\Delta X}^k$ for all k sufficiently large with $\|\Delta X^k\| = \|\widehat{\Delta X}^k\| \to 0$ as $k \to \infty$. We note that G is continuously differentiable and $\mathrm{D}G(\overline{X})$ is surjective by hypothesis. Let $\overline{\lambda}_{\min} = \lambda_{\min}(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^*)$. From (3.15), we know that for all k sufficiently large, $\mathrm{D}G(X^k)$ is surjective and

$$|||(\mathrm{D}G(X^k))^\dagger||| \le 2|||(\mathrm{D}G(\overline{X}))^\dagger|||,\ \lambda_{\min}(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^*) \ge \frac{1}{2}\overline{\lambda}_{\min} > 0. \tag{3.26}$$

□

Moreover, there exist two constants $L_1, L_2 > 0$ such that for all k sufficiently large,

$$\begin{cases} \|G(X^k) - G(\overline{X})\|_F \le L_1 \mathrm{dist}(X^k, \overline{X}), \\ \|\widehat{G}_{X^k}(\Delta X^k) - \widehat{G}_{X^k}(0_{X^k}) - \mathrm{D}\widehat{G}_{X^k}(0_{X^k})[\Delta X^k]\|_F \le L_2\|\Delta X^k\|^2, \\ \mathrm{dist}\big(X^k, R_{X^k}(\Delta X^k)\big) \le \nu\|\Delta X^k\|, \end{cases} \tag{3.27}$$

where ν is the constant given in (3.1).

It follows from (3.9), (3.26), (3.27), and the definition of $\overline{\sigma}_k$ and $\overline{\eta}_k$ in Algorithm 1 that for all k sufficiently large,

$$\begin{aligned} \widehat{\eta}_k &\le \frac{\overline{\sigma}_k}{\lambda_{\min}\big(\mathrm{D}G(X^k) \circ (\mathrm{D}G(X^k))^*\big) + \overline{\sigma}_k} + \overline{\eta}_k \\ &\le \frac{1}{\frac{1}{2}\overline{\lambda}_{\min} + \overline{\sigma}_k}\overline{\sigma}_k + \overline{\eta}_k \le \frac{2}{\overline{\lambda}_{\min}}\|G(X^k)\|_F + \|G(X^k)\|_F \\ &\le \frac{2+\overline{\lambda}_{\min}}{\overline{\lambda}_{\min}} L_1 \mathrm{dist}(X^k, \overline{X}) \equiv c_1 \mathrm{dist}(X^k, \overline{X}), \end{aligned} \tag{3.28}$$

where $c_1 := (L_1(2+\overline{\lambda}_{\min}))/\overline{\lambda}_{\min}$. From Lemmas 2 and 3, (2.11), (3.26), (3.27), and (3.28), we have for all k sufficiently large that

$$\begin{aligned} &\|G(X^{k+1})\|_F \\ &= \|G(X^{k+1}) - G(X^k) - \mathrm{D}G(X^k)[\Delta X^k] + G(X^k) + \mathrm{D}G(X^k)[\Delta X^k]\|_F \\ &\le \|\widehat{G}_{X^k}(\Delta X^k) - \widehat{G}_{X^k}(0_{X^k}) - \mathrm{D}\widehat{G}_{X^k}(0_{X^k})[\Delta X^k]\|_F \\ &\quad + \|\widehat{G}_{X^k}(0_{X^k}) + \mathrm{D}\widehat{G}_{X^k}(0_{X^k})[\Delta X^k]\|_F \\ &\le L_2\|\Delta X^k\|^2 + \widehat{\eta}_k\|G(X^k)\|_F \\ &\le L_2\big((1+\overline{\eta}_k)|||(\mathrm{D}G(X^k))^\dagger||| \cdot \|G(X^k)\|\big)^2 + \widehat{\eta}_k\|G(X^k)\|_F \\ &\le L_2\big((1+\overline{\eta}_k)|||(\mathrm{D}G(X^k))^\dagger|||\big)^2\big(L_1\mathrm{dist}(X^k, \overline{X})\big)^2 + \widehat{\eta}_k L_1 \mathrm{dist}(X^k, \overline{X}) \\ &\le L_2\big(2(1+\overline{\eta}_{\max})L_1|||(\mathrm{D}G(\overline{X}))^\dagger|||\big)^2\big(\mathrm{dist}(X^k, \overline{X})\big)^2 + c_1 L_1\big(\mathrm{dist}(X^k, \overline{X})\big)^2 \\ &\equiv c_2\big(\mathrm{dist}(X^k, \overline{X})\big)^2, \end{aligned} \tag{3.29}$$

where $c_2 := L_2\big(2(1+\overline{\eta}_{\max})L_1|||(\mathrm{D}G(\overline{X}))^\dagger|||\big)^2 + c_1 L_1$. By using (3.28), for a given constant $0 < \widetilde{\eta}_{\max} < 1$, it holds that

$$\widehat{\eta}_k \le \widetilde{\eta}_{\max} < 1 \tag{3.30}$$

for all k sufficiently large. From Lemma 2, (3.26), (3.27), (3.29), and (3.30), it follows that for all k sufficiently large,

$$\begin{aligned}
\mathrm{dist}(X^{k+1}, \overline{X}) &\le \sum_{j=k+1}^{\infty} \mathrm{dist}(X^j, X^{j+1}) = \sum_{j=k+1}^{\infty} \mathrm{dist}\big(X^j, R_{X^j}(\Delta X^j)\big) \\
&\le \sum_{j=k+1}^{\infty} \nu\|\Delta X^j\| \le \sum_{j=k+1}^{\infty} 2\nu(1+\overline{\eta}_{\max})|||(\mathrm{D}G(\overline{X}))^\dagger||| \cdot \|G(X^j)\|_F \\
&= 2\nu(1+\overline{\eta}_{\max})|||(\mathrm{D}G(\overline{X}))^\dagger||| \sum_{j=0}^{\infty} \big(1-t(1-\widehat{\eta}_k)\big)^j \|G(X^{k+1})\|_F \\
&\le 2\nu(1+\overline{\eta}_{\max})|||(\mathrm{D}G(\overline{X}))^\dagger||| \sum_{j=0}^{\infty} \big(1-t(1-\widetilde{\eta}_{\max})\big)^j \|G(X^{k+1})\|_F \\
&= \frac{2\nu(1+\overline{\eta}_{\max})|||(\mathrm{D}G(\overline{X}))^\dagger|||}{t(1-\widetilde{\eta}_{\max})} \|G(X^{k+1})\|_F \\
&\equiv c_3\big(\mathrm{dist}(X^k, \overline{X})\big)^2,
\end{aligned}$$

where $c_3 := (2c_2\nu(1+\overline{\eta}_{\max})|||(\mathrm{D}G(\overline{X}))^\dagger|||)/(t(1-\widetilde{\eta}_{\max}))$. This completes the proof. □

3.3 Surjectivity conditions of $\mathrm{D}G(\cdot)$

We give the proof of Theorem 4.

Proof of Theorem 4 We note that $T_{G(\overline{X})}\mathbb{R}^{n\times n} = \mathrm{im}(\mathrm{D}G(\overline{X})) \oplus \mathrm{im}(\mathrm{D}G(\overline{X}))^\perp$ and $\mathrm{im}(\mathrm{D}G(\overline{X}))^\perp = \ker\big((\mathrm{D}G(\overline{X}))^*\big)$, where $\mathrm{im}(\mathrm{D}G(\overline{X}))$ and $\ker\big((\mathrm{D}G(\overline{X}))^*\big)$ denote the image of $\mathrm{D}G(\overline{X})$ and the kernel of $(\mathrm{D}G(\overline{X}))^*$, respectively. Thus, the linear operator $\mathrm{D}G(\overline{X})$ is surjective if and only if $\ker((\mathrm{D}G(\overline{X}))^*) = \{\mathbf{0}_{n\times n}\}$. □

We now derive a sufficient and necessary condition for $\ker\big((\mathrm{D}G(\overline{X}))^*\big) = \{\mathbf{0}_{n\times n}\}$. Let $\Delta Z \in T_{G(\overline{X})}\mathbb{R}^{n\times n}$ be such that $(\mathrm{D}G(\overline{X}))^*[\Delta Z] = 0_{\overline{X}}$. Using the expression of $\mathrm{D}G(\cdot)^*$ given in "Appendix", it follows that $\ker\big((\mathrm{D}G(\overline{X}))^*\big) = \{\mathbf{0}_{n\times n}\}$ if and only if the following equation

$$\begin{cases}
\overline{S} \odot \Delta Z = \mathbf{0}_{n\times n}, \\
[\overline{Q}(\Lambda+\overline{V})(\overline{Q})^T, (\Delta Z)^T] + [\overline{Q}(\Lambda+\overline{V})^T(\overline{Q})^T, \Delta Z] = \mathbf{0}_{n\times n}, \\
W \odot ((\overline{Q})^T \Delta Z \overline{Q}) = \mathbf{0}_{n\times n}
\end{cases}$$

has a unique solution equal to zero, i.e., $\Delta Z = \mathbf{0}_{n\times n}$ or

$$\begin{cases} \mathrm{Diag}\big(\mathrm{vec}(\overline{S})\big)\mathrm{vec}(\Delta Z) = \mathbf{0}_{n^2}, \\ (I_{n^2} - \widehat{P})(\overline{Q} \otimes \overline{Q})\big((\Lambda + \overline{V}) \otimes I_n - I_n \otimes (\Lambda + \overline{V})^T\big)(\overline{Q} \otimes \overline{Q})^T \mathrm{vec}(\Delta Z) = \mathbf{0}_{n^2}, \\ \mathrm{Diag}\big(\mathrm{vec}(W)\big)(\overline{Q} \otimes \overline{Q})^T \mathrm{vec}(\Delta Z) = \mathbf{0}_{n^2} \end{cases}$$

has a unique solution equal to zero, i.e., $\mathrm{vec}(\Delta Z) = \mathbf{0}_{n^2}$, where the relation

$$\widehat{P}(A \otimes B) = (B \otimes A)\widehat{P}, \quad \forall A, B \in \mathbb{R}^{n\times n}$$

is used [6, p. 448]. This establishes the theorem. □

We have some remarks about Theorem 4.

Remark 1 Let $\overline{X} := (\overline{S}, \overline{Q}, \overline{V}) \in \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ be an accumulation point of the sequence $\{X^k := (S^k, Q^k, V^k)\}$ generated by Algorithm 1. Define

$$\begin{cases} J_{\overline{S}} := \mathrm{Diag}\big(\mathrm{vec}(\overline{S})\big) \in \mathbb{R}^{n^2\times n^2}, \\ J_{\overline{Q}} := (I_{n^2} - \widehat{P})\big((\overline{S} \odot \overline{S}) \otimes I_n - I_n \otimes (\overline{S} \odot \overline{S})^T\big) \in \mathbb{R}^{n^2\times n^2}, \\ J_{\overline{V}} := \mathrm{Diag}\big(\mathrm{vec}(W)\big)(\overline{Q} \otimes \overline{Q})^T \in \mathbb{R}^{n^2\times n^2} \end{cases} \tag{3.31}$$

and

$$J_{\overline{X}} := \Big[\, J_{\overline{S}}^T \;\; J_{\overline{Q}}^T \;\; J_{\overline{V}}^T \,\Big]^T \in \mathbb{R}^{3n^2\times n^2}. \tag{3.32}$$

By Theorem 4, the linear operator $\mathrm{D}G(\overline{X})$ is surjective if and only if the matrix $J_{\overline{X}}$ is of full column rank, i.e., $\mathrm{rank}(J_{\overline{X}}) = n^2$. Based on Fact 7.5.2 in [6], we can obtain

$$\mathrm{rank}\big((\overline{S} \odot \overline{S}) \otimes I_n - I_n \otimes (\overline{S} \odot \overline{S})^T\big) = \dim\big(\{[(\overline{S} \odot \overline{S})^T, A] \mid A \in \mathbb{R}^{n\times n}\}\big). \tag{3.33}$$

Based on (3.33), the rank of the matrix $J_{\overline{Q}}$ depends on the property of the matrix $(\overline{S} \odot \overline{S})^T$. Since $\mathrm{Diag}\big(\mathrm{vec}(\overline{S})\big)$ and $\mathrm{Diag}\big(\mathrm{vec}(W)\big)$ are two diagonal matrices and $\overline{Q} \otimes \overline{Q}$ is an orthogonal matrix, it follows from (3.31) and the definition of W that

$$\begin{cases} \mathrm{rank}(J_{\overline{S}}) = \mathrm{rank}\big(\mathrm{vec}(\overline{S})\big) = \text{number of nonzero elements of } S, \\ \mathrm{rank}(J_{\overline{V}}) = \mathrm{rank}\big(\mathrm{vec}(W)\big) = \text{cardinality of the index subset } \mathcal{J} \geq \frac{(n-1)(n-2)}{2}. \end{cases} \tag{3.34}$$

In addition, the number of nonzero row vectors of the matrix $J_{\overline{V}}$ is also equal to the cardinality of the index subset $\mathcal{J}$. In general, the matrix $J_{\overline{Q}}$ also has many nonzero row vectors. From (3.32) we see that if the matrix $\overline{S}$ is not very sparse, then it is likely that $J_{\overline{X}}$ has a nonzero n^2-by-n^2 principal minor and thus $J_{\overline{X}}$ is of full column rank.

Specially, if the matrix $\overline{S}$ only has nonzero elements, it follows from (3.34) that the matrix $J_{\overline{S}}$ is of full rank and thus $J_{\overline{X}}$ is of full column rank.

Remark 2 Instead of the solution of (2.1), one may solve **Problem I** by finding a global solution to the following nonlinear minimization problem:

$$\begin{array}{ll} \min & \phi(S, Q, V) := \frac{1}{2}\|G(S, Q, V)\|_F^2 \\ \text{subject to (s.t.)} & (S, Q, V) \in \mathbb{R}^{n \times n} \times \mathcal{O}(n) \times \mathcal{V}. \end{array} \tag{3.35}$$

Using Lemma 7.1 in [20], (2.10), and Lemmas 1 and 4, it follows that the search direction $\widehat{\Delta X}^k$ is a descent direction of ϕ if X^k is not a stationary point of $\|G\|_F$. Thus one may see Algorithm 1 as a Riemannian optimization method for solving the optimization problem (3.35) in the sense that it provides a search direction ΔX^k with a line search as in Step 3. Since the objective function ϕ defined in (3.35) is not a Riemannian convex function, ϕ may have different stationary points, which are local minimizers, local maximizers, or saddle points. Therefore, the convergence of Algorithm 1 depends on the choice of the starting point. To guarantee that Algorithm 1 converges to a global minimizer of ϕ, the starting point needs to be contained in an appropriate neighborhood of a global minimizer.

We see that the global convergence of Algorithm 1 is guaranteed by Assumption 1. However, Algorithm 1 may fail to converge if Assumption 1 is not satisfied. To increase the possibility of the global convergence of Algorithm 1, one may first generate a starting point by using the alternating projection method proposed by Orsi [34] such that the objective function ϕ in (3.35) is small enough, then apply Algorithm 1 to find a solution to (2.1). This hybrid strategy may increase the efficiency of Algorithm 1.

4 Extensions

We extend the proposed Riemannian inexact Newton-CG method to the inverse eigenvalue problem of finding a nonnegative matrix with prescribed realizable spectrum and prescribed entries. This inverse problem is stated as follows:

Problem II. *Given a realizable list of n complex numbers* $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, *find an* n-by-n *real nonnegative matrix* C *such that its eigenvalues are* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *and*

$$(C)_{ij} = (C_a)_{ij}, \quad \forall (i, j) \in \mathcal{L}, \tag{4.1}$$

where $\mathcal{L} \subset \mathcal{N}$ *is a given index subset and* C_a *is any given n-by-n nonnegative matrix such that* $\{(C_a)_{ij} \mid (i, j) \in \mathcal{L}\}$ *are prescribed entries.*

In the following discussion, we assume that the prescribed entries characterized by (4.1) are realizable for the realizable spectrum $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, i.e., there exists an n-by-n nonnegative matrix which has both the prescribed spectrum and the prescribed entries. Define the matrix $\widehat{U} \in \mathbb{R}^{n \times n}$ by $(\widehat{U})_{ij} = 1$, if $(i, j) \in \mathcal{L}$; 0, otherwise. Let the matrix $\widehat{C}_a \in \mathbb{R}^{n \times n}$ be defined by $\widehat{C}_a := \widehat{U} \odot C_a$. Also, define a set $\mathcal{Z}$ by

$$\mathcal{Z} := \{S \in \mathbb{R}^{n \times n} \mid \widehat{U} \odot S = \mathbf{0}_{n \times n}\}.$$

Then **Problem II** is to solve the following nonlinear equation

$$H(S, Q, V) = \mathbf{0}_{n\times n} \tag{4.2}$$

for $(S, Q, V) \in \mathcal{Z} \times \mathcal{O}(n) \times \mathcal{V}$, where $H : \mathcal{Z} \times \mathcal{O}(n) \times \mathcal{V} \to \mathbb{R}^{n\times n}$ is defined by

$$H(S, Q, V) = \widehat{C}_a + S \odot S - Q(\Lambda + V)Q^T, \quad (S, Q, V) \in \mathcal{Z} \times \mathcal{O}(n) \times \mathcal{V}.$$

Obviously, H is smooth mapping from the product manifold $\mathcal{Z} \times \mathcal{O}(n) \times \mathcal{V}$ to the linear space $\mathbb{R}^{n\times n}$.

We note that the dimension of $\mathcal{Z} \times \mathcal{O}(n) \times \mathcal{V}$ is given by

$$\dim(\mathcal{Z} \times \mathcal{O}(n) \times \mathcal{V}) = n^2 - |\mathcal{L}| + \frac{n(n-1)}{2} + |\mathcal{J}|.$$

We point out that the nonlinear equation $H(S, Q, V) = \mathbf{0}_{n\times n}$ is under-determined over $\mathcal{Z} \times \mathcal{O}(n) \times \mathcal{V}$ if the problem size n is large and the number $|\mathcal{L}|$ of prescribed entries is small. We also remark that, if $(\overline{S}, \overline{Q}, \overline{V}) \in \mathcal{Z} \times \mathcal{O}(n) \times \mathcal{V}$ is a solution to $H(S, Q, V) = \mathbf{0}_{n\times n}$, then $\overline{C} := \widehat{C}_a + \overline{S} \odot \overline{S}$ is a solution to **Problem II**. Alternatively, one may solve **Problem II** by finding a global solution to the following nonlinear minimization problem:

$$\begin{aligned} &\min\ \psi(S, Q, V) := \frac{1}{2}\|H(S, Q, V)\|_F^2 \\ &\text{s.t.}\quad (S, Q, V) \in \mathcal{Z} \times \mathcal{O}(n) \times \mathcal{V}. \end{aligned} \tag{4.3}$$

One may apply Algorithm 1 to solving the nonlinear Eq. (4.2). Under some assumptions as in Assumption 1, the global and quadratic convergence can be established in a similar way as in Sect. 3.

5 Numerical tests

In this section, we report the numerical performance of Algorithm 1 for solving **Problem I** and **Problem II** via solving the nonlinear Eqs. (2.1) and (4.2). All the numerical tests are carried out by using MATLAB 7.1 running on a workstation with a Intel Xeon CPU E5-2687W at 3.10 GHz and 32 GB of RAM. To illustrate the efficiency of our algorithm, we compare Algorithm 1 with the alternating projection method [34], the Riemannian Fletcher-Reeves conjugate gradient method (RFR) [48] and the geometric Polak-Ribière-Polyak-based nonlinear conjugate gradient method (GPRP) [50].

The two Riemannian conjugate gradient methods RFR and GPRP in [48,50] are used to solve the Riemannian optimization problems (3.35) and (4.3). The alternating projection method in [34] is employed to solve **Problem I**:

$$\text{Find } C \in \mathcal{P} \cap \mathbb{R}_+^{n\times n} \tag{5.1}$$

and **Problem II**:

$$\text{Find } C \in \mathcal{P} \cap \mathcal{Q}, \tag{5.2}$$

where

$$\mathcal{P} = \{A \in \mathbb{C}^{n\times n} \mid A = UTU^H \text{ for some unitary matrix } U \text{ and some } T \in \mathcal{T}\}$$

and

$$\mathcal{Q} = \{C \in \mathbb{R}_+^{n\times n} \mid (C)_{ij} = (C_a)_{ij} \text{ for all } (i, j) \in \mathcal{L}\}.$$

Here, $\mathcal{T} = \{T \in \mathbb{C}^{n\times n} \mid T \text{ is upper triangular with spectrum } \{\lambda_1, \lambda_2, \ldots, \lambda_n\}\}$. The associated alternating projection algorithm for solving problem (5.1) (problem (5.2), respectively) is stated as follows.

Algorithm 2 (Alternating projection algorithm)

Step 0. *Choose an initial point* $C^0 \in \mathbb{R}_+^{n\times n}$ ($C^0 \in \mathcal{Q}$, *respectively*). *Let* $k := 0$.
Step 1. *Calculate a Schur decomposition of* $C^k = U^k T^k (U^k)^H$.
Step 2. *Set* $Y^{k+1} = P_{\mathcal{P}}(U^k, T^k)$, *where* $P_{\mathcal{P}}(U^k, T^k)$ *is defined as in* [34, Definition 4.2].
Step 3. *Set* $C^{k+1} = P_{\mathbb{R}_+^{n\times n}}(Y^{k+1})$ ($C^{k+1} = P_{\mathcal{Q}}(Y^{k+1})$, *respectively*), *where* $P_{\mathbb{R}_+^{n\times n}}(Y^{k+1})$ *is the projection of* Y^{k+1} *onto* $\mathbb{R}_+^{n\times n}$.
Step 4. *Replace* k *by* $k+1$ *and go to Step 1.*

For Algorithm 1 for solving (2.1), Algorithm 2 for problem (5.1), and RFR and GPRP for problem (3.35), we randomly generate the starting points by the built-in functions `rand`, `schur`, and `svd`:

$$\begin{array}{c} S \odot S = \texttt{rand}\,(n, n), \quad S^0 = S \in \mathbb{R}^{n\times n}, \quad C^0 = S^0 \odot S^0, \\ \left[Q^0, V\right] = \texttt{schur}\,(S^0 \odot S^0, '\mathrm{real}'), \quad V^0 = W \odot V. \end{array} \tag{5.3}$$

For Algorithm 1 for solving (4.2), Algorithm 2 for problem (5.2), and RFR and GPRP for problem (4.3), the starting points are generated randomly as follows:

$$\begin{array}{c} S \odot S = \texttt{rand}\,(n, n), \quad S^0 = \widehat{U} \odot S \in \mathcal{Z}, \quad C^0 = \widehat{C}_a + S^0 \odot S^0, \\ \left[Q^0, V\right] = \texttt{schur}\,(\widehat{C}_a + S^0 \odot S^0, '\mathrm{real}'), \quad V^0 = W \odot V. \end{array} \tag{5.4}$$

For comparison purposes, the stopping criteria for Algorithm 1, Algorithm 2 for problems (5.1) and (5.2), and the two Riemannian conjugate gradient methods in [48,50] for problems (3.35) and (4.3) are set to be

$$\|G(X^k)\|_F < 10^{-8}, \quad \|H(X^k)\|_F < 10^{-8}, \quad \text{and} \quad \|C^k - Y^k\|_F < 10^{-8}.$$

In our numerical tests, we set $\overline{\sigma}_{\max} = 0.01$, $\overline{\eta}_{\max} = 0.1$, $\theta_{\min} = 0.1$, $\theta_{\max} = 0.9$, and $t = 10^{-4}$. The largest number of iterations in Algorithm 2 is set to be 100000.

Table 1 Numerical results of Example 1

Alg.	n	CT. (s)	IT.	NF.	NCG.	Res.	grad.
GPRP	10	0.0348	76.0	79.1		8.7×10^{-9}	1.9×10^{-8}
	20	0.0592	136.4	140.3		9.3×10^{-9}	3.1×10^{-8}
	50	0.3180	356.2	361.3		9.7×10^{-9}	6.8×10^{-8}
	80	1.1895	621.4	627.4		9.8×10^{-9}	7.6×10^{-8}
	100	2.1592	740.4	746.8		9.8×10^{-9}	8.5×10^{-8}
	150	6.7587	1152.1	1159.2		9.9×10^{-9}	1.1×10^{-7}
	200	14.284	1496.9	1504.8		9.9×10^{-9}	1.2×10^{-7}
RFR	10	0.1465	120.9	123.5		8.8×10^{-9}	2.6×10^{-8}
	20	0.2479	193.3	196.8		9.4×10^{-9}	4.1×10^{-8}
	50	0.4624	301.0	306.0		9.3×10^{-9}	7.2×10^{-8}
	80	1.0742	458.3	464.1		9.6×10^{-9}	1.1×10^{-7}
	100	1.7413	526.2	532.2		9.6×10^{-9}	1.1×10^{-7}
	150	6.1772	984.9	991.9		9.9×10^{-9}	9.9×10^{-8}
	200	11.223	1172.9	1180.0		9.8×10^{-9}	1.4×10^{-7}
Alg. 2	10	0.0480	43.2			4.2×10^{-9}	
	20	0.0595	44.3			3.8×10^{-9}	
	50	3.6164	746.4			5.1×10^{-9}	
	80	16.877	1447.6			2.7×10^{-9}	
	100	13.695	816.0			6.9×10^{-9}	
	150	05 m 25	8340.8			5.7×10^{-9}	
	200	24 m 09	18,961			0.1816*	
Alg. 1	10	0.0081	5.2	6.2	17.2	1.8×10^{-9}	2.5×10^{-9}
	20	0.0118	5.9	6.9	33.6	9.7×10^{-10}	4.4×10^{-9}
	50	0.0418	6.0	7.0	49.3	3.8×10^{-10}	3.2×10^{-9}
	80	0.1806	6.6	7.6	71.1	2.4×10^{-9}	3.6×10^{-8}
	100	0.3595	7.0	8.0	86.9	5.6×10^{-14}	5.6×10^{-13}
	150	0.8189	7.0	8.0	96.2	3.4×10^{-12}	9.0×10^{-11}
	200	1.6995	7.0	8.0	108.6	2.3×10^{-11}	8.8×10^{-10}

The largest number of outer iterations in Algorithm 1 is set to be 100 and the largest number of iterations in the CG method is set to be n^2.

For comparison purposes, we repeat our experiments over 10 different problems. In our numerical tests, ‘CT.’, IT.’, ‘NF.’, ‘NCG.’, ‘Res.’, and ‘grad.’ mean the averaged total computing time in seconds, the averaged number of iterations, the averaged number of function evaluations, the averaged number of inner CG iterations, the averaged residual $\|G(X^k)\|_F$, $\|H(X^k)\|_F$, or $\|C^k - Y^k\|_F$, and the averaged residual $\|\mathrm{grad}\,\phi(X^k)\|$ or $\|\mathrm{grad}\,\psi(X^k)\|$ at the final iterates of the corresponding algorithms accordingly.

Table 2 Numerical results of Example 2

Alg.	n	CT. (s)	IT.	NF.	NCG.	Res.	grad.
GPRP	10	0.0418	108.5	111.6		9.1×10^{-9}	1.7×10^{-8}
	20	0.0854	195.0	198.8		9.2×10^{-9}	2.7×10^{-8}
	50	0.4630	526.6	532.3		9.8×10^{-9}	4.7×10^{-8}
	80	1.6851	878.8	885.3		9.9×10^{-9}	6.2×10^{-8}
	100	3.2997	1129.9	1137.0		9.9×10^{-9}	7.1×10^{-8}
	150	10.016	1701.8	1709.8		9.9×10^{-9}	8.7×10^{-8}
	200	21.287	2200.0	2208.8		9.9×10^{-9}	1.0×10^{-7}
RFR	10	0.2405	205.6	207.8		9.6×10^{-9}	2.8×10^{-8}
	20	0.3589	285.6	288.8		9.6×10^{-9}	4.1×10^{-8}
	50	0.7371	475.4	480.4		9.6×10^{-9}	7.8×10^{-8}
	80	1.3239	560.9	566.9		9.7×10^{-9}	9.3×10^{-8}
	100	2.4316	733.0	739.3		9.7×10^{-9}	1.3×10^{-7}
	150	5.9154	946.8	954.4		9.8×10^{-9}	1.3×10^{-7}
	200	12.771	1335.2	1343.2		9.8×10^{-9}	1.5×10^{-7}
Alg. 2	10	0.0284	34.1			7.4×10^{-9}	
	20	0.0688	51.9			7.3×10^{-9}	
	50	6.6000	122.4			6.1×10^{-9}	
	80	2.2811	229.6			5.9×10^{-9}	
	100	4.5759	297.0			8.0×10^{-9}	
	150	23.701	643.2			8.2×10^{-9}	
	200	29 m 47	22723			0.3227*	
Alg. 1	10	0.0093	5.8	6.8	24.1	2.1×10^{-10}	1.8×10^{-10}
	20	0.0131	6.0	7.0	37.3	6.8×10^{-10}	1.1×10^{-9}
	50	0.0632	6.6	7.6	65.8	1.8×10^{-9}	1.3×10^{-8}
	80	0.2339	7.0	8.0	88.8	1.2×10^{-13}	1.7×10^{-12}
	100	0.3790	7.0	8.0	92.2	1.3×10^{-11}	2.2×10^{-10}
	150	0.8715	7.0	8.0	102.4	4.7×10^{-10}	1.3×10^{-8}
	200	2.1890	7.4	8.4	130.2	1.3×10^{-9}	3.5×10^{-8}

Example 1 We consider **Problem I** with varying n. Let $\widehat{C}$ be a random $n \times n$ nonnegative matrix with each entry generated from the uniform distribution on the interval $[0, 1]$. We choose the eigenvalues of $\widehat{C}$ as prescribed spectrum.

Example 2 We consider **Problem II** with varying n. Let $\widehat{C}$ be a random $n \times n$ nonnegative matrix with each entry generated from the uniform distribution on the interval $[0, 1]$. We choose the eigenvalues of $\widehat{C}$ as prescribed spectrum. Also, we choose the index subset $\mathcal{L} := \{(i, j) \mid 0.2 \leq (\widehat{C})_{ij} \leq 0.3,\ i, j = 1, \ldots, n\}$. The nonnegative matrix $C_a \in \mathbb{R}^{n \times n}$ with prescribed entries is defined by $(C_a)_{ij} := (\widehat{C})_{ij}$, if $(i, j) \in \mathcal{L}$; 0, otherwise.

Table 3 Numerical results of Example1

Alg.	n	CT.	IT.	NF.	NCG.	Res.	grad.
GPRP	400	04 m 02 s	2957	2966		9.9×10^{-9}	2.2×10^{-7}
	600	17 m 02 s	4103	4112		9.9×10^{-9}	1.0×10^{-7}
	800	38 m 56 s	4931	4941		9.9×10^{-9}	2.7×10^{-7}
	1000	01 h 19 m 26 s	5605	5615		9.9×10^{-9}	4.9×10^{-7}
RFR	400	04 m 21 s	3547	3555		9.9×10^{-9}	1.8×10^{-7}
	600	22 m 14 s	5680	5689		9.9×10^{-9}	1.7×10^{-7}
	800	01 h 00 m 28 s	8194	8203		9.9×10^{-9}	2.6×10^{-7}
	1000	02 h 17 m 06 s	10,288	10,298		9.9×10^{-9}	3.0×10^{-7}
Alg. 1	400	27.6 s	8.0	9.0	186.1	3.0×10^{-13}	7.0×10^{-12}
	600	01 m 16 s	8.0	9.0	195.0	5.4×10^{-13}	3.5×10^{-11}
	800	02 m 55 s	8.0	9.0	206.4	1.2×10^{-11}	2.9×10^{-9}
	1000	05 m 49 s	8.0	9.0	225.9	3.3×10^{-11}	3.9×10^{-9}

Table 4 Numerical resultsofExample2

Alg.	n	CT.	IT.	NF.	NCG.	Res.	grad.
GPRP	400	06 m 40 s	4594	4604		9.9×10^{-9}	1.6×10^{-7}
	600	22 m 06 s	5529	5541		9.9×10^{-9}	2.7×10^{-7}
	800	44 m 24 s	5260	5272		9.9×10^{-9}	4.4×10^{-7}
	1000	01 h 21 m 07 s	5399	5412		9.8×10^{-9}	6.7×10^{-7}
RFR	400	05 m 25 s	4036	4046		9.9×10^{-9}	1.4×10^{-7}
	600	21 m 42 s	5855	5866		9.9×10^{-9}	1.8×10^{-7}
	800	01 h 10 m 44 s	8979	8991		9.9×10^{-9}	2.0×10^{-7}
	1000	02 h 44 m 41 s	11,764	11,776		9.9×10^{-9}	2.3×10^{-7}
Alg. 1	400	26.6 s	8.0	9.0	183.2	1.8×10^{-12}	1.8×10^{-10}
	600	01 m 41 s	8.0	9.0	231.6	8.2×10^{-13}	5.3×10^{-11}
	800	03 m 38 s	8.0	9.0	233.4	4.5×10^{-10}	7.6×10^{-8}
	1000	06 m 34 s	8.0	9.0	236.3	6.7×10^{-9}	1.4×10^{-6}

Tables 1, 2 list numerical results for Examples 1, 2, where "*" means that the largest number of iterations is reached for some stating points.

We observe from Tables 1, 2 that Algorithm 2 behaviors better than GPRP and/or RFR in terms of computing time for small n (e.g., $n = 10, 20$) while GPRP and RFR work much better than Algorithm 2 in terms of computing time for $n \geq 50$. However, Algorithm 1 is the most effective in terms of computing time.

To further illustrate the efficiency of our algorithm, we report numerical results for Examples 1, 2 with various problem sizes. Tables 3, 4 display numerical results for Examples 1, 2.

We see from Tables 3, 4 that Algorithm 1, GPRP and RFR work for large problems while Algorithm 1 is more efficient than GPRP and RFR for large problems.

Finally, we point out that all algorithms may converge to different solutions for different starting points.

6 Conclusions

This paper is concerned with the inverse eigenvalue problem of finding a nonnegative matrix from the prescribed realizable spectrum. The inverse problem is rewritten as an under-determined constrained nonlinear matrix equation over several matrix manifolds. Then a Riemannian inexact Newton-CG method is proposed for solving the constrained nonlinear matrix equation. The global and quadratic convergence of the proposed geometric method is established under some assumptions. Our method is also extended to the case of prescribed entries. Numerical tests illustrate the efficiency of the proposed geometric algorithm. From our numerical tests, we observe that, for large problems, most of our computing time is spent on the CG method for solving (2.8). It would improve the efficiency if one can find a good preconditioner for (2.8), which needs further study.

Acknowledgements We are very grateful to the editor and the referees for their valuable comments and suggestions, which have considerably improved this paper.

Appendix

In this appendix, we establish some basic properties of the product manifold $\mathbb{R}^{n\times n}\times\mathcal{O}(n)\times\mathcal{V}$ and the differential of G defined in (2.1). We first show that the nonlinear matrix Eq. (2.1) is under-determined for all $n\geq 2$. The dimension of $\mathbb{R}^{n\times n}\times\mathcal{O}(n)\times\mathcal{V}$ is given by

$$\dim(\mathbb{R}^{n\times n}\times\mathcal{O}(n)\times\mathcal{V})=n^2+\frac{n(n-1)}{2}+|\mathcal{J}|,$$

where $\mathcal{J}$ is the complementary index set of $\mathcal{I}$ with respect to the index set $\mathcal{N}$, and $|\mathcal{J}|$ is the cardinality of $\mathcal{J}$. Thus

$$\dim(\mathbb{R}^{n\times n}\times\mathcal{O}(n)\times\mathcal{V})>\dim\mathbb{R}^{n\times n}\quad\text{for } n\geq 2.$$

Hence, (2.1) is under-determined for all $n\geq 2$.

The tangent space of $\mathbb{R}^{n\times n}\times\mathcal{O}(n)\times\mathcal{V}$ at a point $(S,Q,V)\in\mathbb{R}^{n\times n}\times\mathcal{O}(n)\times\mathcal{V}$ is given by

$$T_{(S,Q,V)}\big(\mathbb{R}^{n\times n}\times\mathcal{O}(n)\times\mathcal{V}\big)=T_S\mathbb{R}^{n\times n}\times T_Q\mathcal{O}(n)\times T_V\mathcal{V}.$$

Here, $T_S\mathbb{R}^{n\times n}$, $T_Q\mathcal{O}(n)$, and $T_V\mathcal{V}$ are the tangent spaces of $\mathbb{R}^{n\times n}$, $\mathcal{O}(n)$, and $\mathcal{V}$ at $S \in \mathbb{R}^{n\times n}$, $Q \in \mathcal{O}(n)$, and $V \in \mathcal{V}$ accordingly, which are given by [1, p. 42]:

$$T_S\mathbb{R}^{n\times n} = \mathbb{R}^{n\times n}, \quad T_Q\mathcal{O}(n) = \left\{Q\Omega \mid \Omega^T = -\Omega,\ \Omega \in \mathbb{R}^{n\times n}\right\}, \quad T_V\mathcal{V} = \mathcal{V}.$$

A retraction R on $\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ is given by

$$R_{(S,Q,V)}(\xi_S, \zeta_Q, \eta_V) = \big(R_S(\xi_S), R_Q(\zeta_Q), R_V(\eta_V)\big)$$

for all $(S, Q, V) \in \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ and $(\xi_S, \eta_Q, \gamma_V) \in T_{(S,Q,V)}\big(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}\big)$, where R_S, R_Q, and R_V are the retractions on $\mathbb{R}^{n\times n}$, $\mathcal{O}(n)$, and $\mathcal{V}$ accordingly, which may take the following form:

$$\begin{cases} R_S(\xi_S) = S + \xi_S, & \text{for } \xi_S \in T_S\mathbb{R}^{n\times n}, \\ R_Q(\zeta_Q) = \mathrm{qf}(Q + \zeta_Q), & \text{for } \zeta_Q \in T_Q\mathcal{O}(n), \\ R_V(\eta_V) = V + \eta_V, & \text{for } \eta_V \in T_V\mathcal{V}. \end{cases}$$

Here, $\mathrm{qf}(A)$ means the Q factor of the QR decomposition of a nonsingular matrix $A \in \mathbb{R}^{n\times n}$ in the form of $A = Q\widetilde{R}$ with $Q \in \mathcal{O}(n)$ and $\widetilde{R}$ being an upper triangular matrix with strictly positive diagonal entries. For other choices of retractions on $\mathcal{O}(n)$, one may refer to [1, pp. 58–59].

We now establish the differential of G. By simple calculation, the differential $\mathrm{D}G(S, Q, V) : T_{(S,Q,V)}\big(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}\big) \to T_{G(S,Q,V)}\mathbb{R}^{n\times n} \simeq \mathbb{R}^{n\times n}$ of G at $(S, Q, V) \in \mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V}$ is determined by

$$\mathrm{D}G(S, Q, V)[(\Delta S, \Delta Q, \Delta V)] = 2S \odot \Delta S + [Q(\Lambda + V)Q^T, \Delta Q Q^T] - Q\Delta V Q^T$$

for all $(\Delta S, \Delta Q, \Delta V) \in T_{(S,Q,V)}(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V})$. On the other hand, with respect to the Riemannian metric defined in (2.2), the adjoint $(\mathrm{D}G(S, Q, V))^* : T_{G(S,Q,V)}\mathbb{R}^{n\times n} \to T_{(S,Q,V)}(\mathbb{R}^{n\times n} \times \mathcal{O}(n) \times \mathcal{V})$ of $\mathrm{D}G(S, Q, V)$ is determined by

$$\begin{aligned} &(\mathrm{D}G(S, Q, V))^*[\Delta Z] \\ &\quad = ((\mathrm{D}G(S, Q, V))^*_1[\Delta Z], (\mathrm{D}G(S, Q, V))^*_2[\Delta Z], (\mathrm{D}G(S, Q, V))^*_3[\Delta Z]) \end{aligned}$$

for all $\Delta Z \in T_{G(S,Q,V)}\mathbb{R}^{n\times n}$, where for each $\Delta Z \in T_{G(S,Q,V)}\mathbb{R}^{n\times n}$,

$$\begin{cases} (\mathrm{D}G(S, Q, V))^*_1[\Delta Z] = 2S \odot \Delta Z, \\ (\mathrm{D}G(S, Q, V))^*_2[\Delta Z] = \dfrac{1}{2}\big([Q(\Lambda + V)Q^T, (\Delta Z)^T] + [Q(\Lambda + V)^T Q^T, \Delta Z]\big)Q, \\ (\mathrm{D}G(S, Q, V))^*_3[\Delta Z] = -W \odot \big(Q^T \Delta Z Q\big). \end{cases}$$

Here, $W \in \mathbb{R}^{n\times n}$ is defined by

$$W_{ij} = \begin{cases} 0, & \text{if } (i, j) \in \mathcal{I}, \\ 1, & \text{otherwise}. \end{cases} \tag{A.1}$$

Analogously, we can establish the tangent space of the product manifold $\mathcal{Z} \times \mathcal{O}(n) \times \mathcal{V}$, a retraction on $\mathcal{Z} \times \mathcal{O}(n) \times \mathcal{V}$, and the differential of H defined in (4.2) and its adjoint. Here, $\mathcal{Z} \times \mathcal{O}(n) \times \mathcal{V}$ is equipped with the Riemannian metric defined as in (2.2).

References

1. Absil, P.-A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2008)
2. Baker, C.G.: Riemannian Manifold Trust-Region Methods with Applications to Eigenproblems. Ph.D. thesis. School of Computational Science, Florida State University, Tallahassee (2008)
3. Bapat, R.B., Raghavan, T.E.S.: Nonnegative Matrices and Applications. Cambridge University Press, Cambridge (1997)
4. Barrett, W.W., Johnson, C.R.: Possible spectra of totally positive matrices. Linear Algebra Appl. **62**, 231–233 (1984)
5. Berman, A., Plemmons, R.J.: Nonnegative Matrices in the Mathematical Sciences. Academic Press, New York (1979)
6. Bernstein, D.: Matrix Mathematics—Theory, Facts, and Formulas, 2nd edn. Princeton University Press, Princeton (2009)
7. Borobia, A., Canogar, R.: The real nonnegative inverse eigenvalue problem is NP-hard. Linear Algebra Appl. **522**, 127–139 (2017)
8. Boyle, M., Handelman, D.: The spectra of nonnegative matrices via symbolic dynamics. Ann. Math. **133**, 249–316 (1991)
9. Chen, X., Liu, D.L.: Isospectral flow method for nonnegative inverse eigenvalue problem with prescribed structure. J. Comput. Appl. Math. **235**, 3990–4002 (2011)
10. Chu, M.T.: Inverse eigenvalue problems. SIAM Rev. **40**, 1–39 (1998)
11. Chu, M.T., Diele, F., Sgura, I.: Gradient flow method for matrix completion with prescribed eigenvalues. Linear Algebra Appl. **379**, 85–112 (2004)
12. Chu, M.T., Driessel, K.R.: Constructing symmetric nonnegative matrices with prescribed eigenvalues by differential equations. SIAM J. Math. Anal. **22**, 1372–1387 (1991)
13. Chu, M.T., Golub, G.H.: Structured inverse eigenvalue problems. Acta Numer. **11**, 1–71 (2002)
14. Chu, M.T., Golub, G.H.: Inverse Eigenvalue Problems: Theory, Algorithms, and Applications. Oxford University Press, Oxford (2005)
15. Chu, M.T., Guo, Q.: A numerical method for the inverse stochastic spectrum problem. SIAM J. Matrix Anal. Appl. **19**, 1027–1039 (1998)
16. de Oliveira, G.N.: Nonnegative matrices with prescribed spectrum. Linear Algebra Appl. **54**, 117–121 (1983)
17. Dedieu, J.P., Priouret, P., Malajovich, G.: Newton's method on Riemannian manifolds: covariant alpha theory. IMA J. Numer. Anal. **23**, 395–419 (2003)
18. Dennis, J.E., Schnabel, R.B.: Numerical Methods for Unconstrained Optimization and Nonlinear Equations. SIAM, Philadelphia (1996)
19. Egleston, P.D., Lenker, T.D., Narayan, S.K.: The nonnegative inverse eigenvalue problem. Linear Algebra Appl. **379**, 475–490 (2004)
20. Eisenstat, S.C., Walker, H.F.: Globally convergent inexact Newton methods. SIAM J. Optim. **4**, 392–422 (1994)
21. Ellard, R., Migoc, H.: Connecting sufficient conditions for the symmetric nonnegative inverse eigenvalues problem. Linear Algebra Appl. **498**, 521–552 (2016)
22. Fiedler, M.: Eigenvalues of nonnegative symmetric matrices. Linear Algebra Appl. **9**, 119–142 (1974)
23. Friedland, S., Melkman, A.A.: On the eigenvalues of nonnegative Jacobi matrices. Linear Algebra Appl. **25**, 239–254 (1979)
24. Golub, G.H., Van Loan, C.F.: Matrix Computations, 4th edn. Johns Hopkins University Press, Baltimore (2013)
25. Johnson, C.R., Marijuán, C., Paparella, P., Pisonero, M.: The NIEP. arXiv:1703.10992 (2017)
26. Johnson, C.R., Paparella, P.: Perron spectratopes and the real nonnegative inverse eigenvalue problem. Linear Algebra Appl. **493**, 281–300 (2016)

27. Karpelevič, F.I.: On the characteristic roots of matrices with nonnegative elements. Izv. Akad. Nauk SSSR Ser. Mat. **15**, 361–383 (1951). **(in Russian)**
28. Laffey, T.J., Šmigoc, H.: Nonnegative realization of spectra having negative real parts. Linear Algebra Appl. **416**, 148–159 (2006)
29. Lin, M.M.: Fast recursive algorithm for constructing nonnegative matrices with prescribed real eigenvalues. Appl. Math. Comput. **256**, 582–590 (2015)
30. Loewy, R., London, D.: A note on an inverse problems for nonnegative matrices. Linear Multilinear Algebra **6**, 83–90 (1978)
31. Luenberger, D.G.: Optimization by Vector Space Methods. Wiley, New York (1969)
32. Marijuán, C., Pisonero, M., Soto, R.L.: A map of sufficient conditions for the symmetric nonnegative inverse eigenvalue problem. Linear Algebra Appl. **530**, 344–365 (2017)
33. Minc, H.: Nonnegative Matrices. Wiley, New York (1988)
34. Orsi, R.: Numerical methods for solving inverse eigenvalue problems for nonnegative matrices. SIAM J. Matrix Anal. Appl. **28**, 190–212 (2006)
35. Paparella, P.: Realizing Suleimanova-type spectra via permutative matrices. Electron. J. Linear Algebra. **31**, 306–312 (2016)
36. Perfect, H.: Methods of constructing certain stochastic matrices. Duke Math. J. **20**, 395–404 (1953)
37. Perfect, H.: Methods of constructing certain stochastic matrices. II. Duke Math. J. **22**, 305–311 (1955)
38. Reams, R.: An inequality for nonnegative matrices and the inverse eigenvalue problem. Linear Multilinear Algebra **41**, 367–375 (1996)
39. Senata, E.: Non-negative Matrices and Markov Chains, 2nd edn. Springer, New York (2006)
40. Soto, R.L.: Existence and construction of nonnegative matrices with prescribed spectrum. Linear Algebra Appl. **369**, 169–184 (2003)
41. Soto, R.L.: Realizability criterion for the symmetric nonnegative inverse eigenvalue problem. Linear Algebra Appl. **416**, 783–794 (2006)
42. Soto, R.L.: A family of realizability criteria for the real and symmetric nonnegative inverse eigenvalue problem. Numer. Linear Algebra Appl. **20**, 336–348 (2013)
43. Soules, G.W.: Constructing symmetric nonnegative matrices. Linear Multilinear Algebra **13**, 241–251 (1983)
44. Suleĭmanova, H.R.: Stochastic matrices with real characteristic numbers. Doklady Akad. Nauk SSSR (NS) **66**, 343–345 (1949)
45. Simonis, J.P.: Inexact Newton Methods Applied to Under-Determined Systems. Ph.D. thesis. Department of Mathematical Science, Worcester Polytechnic Institute (2006)
46. Smith, S.T.: Optimization techniques on Riemannian manifolds. Fields Inst. Commun. **3**, 113–136 (1994)
47. Xu, S.F.: An Introduction to Inverse Algebraic Eigenvalue Problems. Beijing: Friedr. Vieweg & Sohn, Braunschweig (1998)
48. Yao, T.T., Bai, Z.J., Zhao, Z., Ching, W.K.: A Riemannian Fletcher-Reeves conjugate gradient method for doubly stochastic inverse eigenvalue problems. SIAM J. Matrix Anal. Appl. **37**, 215–234 (2016)
49. Zhao, Z., Bai, Z.J., Jin, X.Q.: A Riemannian Newton algorithm for nonlinear eigenvalue problems. SIAM J. Matrix Anal. Appl. **36**, 752–774 (2015)
50. Zhao, Z., Jin, X.Q., Bai, Z.J.: A geometric nonlinear conjugate gradient method for stochastic inverse eigenvalue problems. SIAM J. Numer. Anal. **54**, 2015–2035 (2016)