

A BRIEF INTRODUCTION TO MANIFOLD OPTIMIZATION

JIANG HU*, XIN LIU[†], ZAIWEN WEN[‡], AND YAXIANG YUAN[§]

Abstract. Manifold optimization is ubiquitous in computational and applied mathematics, statistics, engineering, machine learning, physics, chemistry and etc. One of the main challenges usually is the non-convexity of the manifold constraints. By utilizing the geometry of manifold, a large class of constrained optimization problems can be viewed as unconstrained optimization problems on manifold. From this perspective, intrinsic structures, optimality conditions and numerical algorithms for manifold optimization are investigated. Some recent progress on the theoretical results of manifold optimization are also presented.

1. Introduction. Manifold optimization is concerned with the following optimization problem

$$(1.1) \quad \min_{x \in \mathcal{M}} f(x),$$

where \mathcal{M} is a Riemannian manifold and f is a real-valued function on \mathcal{M} , which can be non-smooth. If additional constraints other than the manifold constraint are involved, we can add in f an indicator function of the feasible set of these additional constraints. Hence, (1.1) covers a general formulation for manifold optimization. In fact, manifold optimization has been widely used in computational and applied mathematics, statistics, machine learning, data science, material science and so on. The existence of the manifold constraint is one of the main difficulties in algorithmic design and theoretical analysis.

Notations. Let \mathbb{R} and \mathbb{C} be the set of real and complex numbers. For a matrix $X \in \mathbb{C}^{n \times p}$, \bar{X} , X^* , $\Re X$ and $\Im X$ are its complex conjugate, complex conjugate transpose, real and imaginary parts, respectively. Let \mathbb{S}^n be the set of all n -by- n real symmetric matrices. For a matrix $M \in \mathbb{C}^{n \times n}$, $\text{diag}(M)$ is a vector in \mathbb{C}^n formulated by the diagonal elements of M . For a vector $c \in \mathbb{C}^n$, $\text{Diag}(c)$ is an n -by- n diagonal matrix with the elements of c on the diagonal. For a differentiable function f on \mathcal{M} , let $\text{grad} f(x)$ and $\text{Hess} f(x)$ be its Riemannian gradient and Hessian at x , respectively. If f can be extended to the ambient Euclidean space, we denote its Euclidean gradient and Hessian by $\nabla f(x)$ and $\nabla^2 f(x)$, respectively.

This paper is organized as follows. In section 2, various kinds of applications of manifold optimization are presented. We review geometry on manifolds, optimality

*Beijing International Center for Mathematical Research, Peking University, China (jianghu@pku.edu.cn)

[†]State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, and University of Chinese Academy of Sciences, China (email: liuxin@lsec.cc.ac.cn). Research supported in part by NSFC grants 11622112 and 11688101, the National Center for Mathematics and Interdisciplinary Sciences, CAS, and Key Research Program of Frontier Sciences QYZDJ-SSW-SYS010, CAS.

[‡]Beijing International Center for Mathematical Research, Peking University, China (wenzw@pku.edu.cn). Research supported in part by the NSFC grants 11421101 and 11831002, and by the National Basic Research Project under the grant 2015CB856002.

[§]State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China (yyx@lsec.cc.ac.cn). Research supported in part by NSFC grants 11331012 and 11461161005.

conditions as well as state-of-the-art algorithms for manifold optimization in [section 3](#). For some selected practical applications in [section 2](#), a few theoretical results based on manifold optimization are introduced in [section 4](#).

2. Applications of manifold optimization. In this section, we introduce applications of manifold optimization in p -harmonic flow, max-cut problems, phase retrieval, eigenvalue problem, electronic structure calculations, Bose-Einstein condensates, cryo-electron microscopy (Cryo-EM), combinatorial optimization, deep learning and etc.

2.1. P -harmonic flow. P -harmonic flow is used in the color image recovery and medical image analysis. For instance, in medical image analysis, the human brain is often mapped to a unit sphere via a conformal mapping, see [Figure 1](#). By

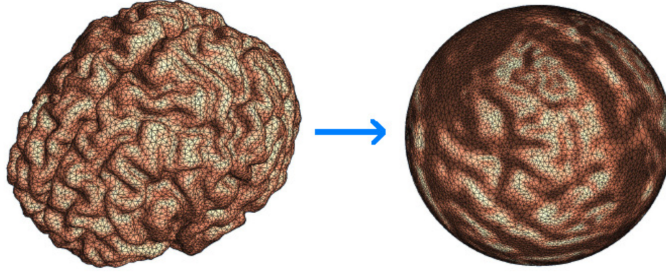


FIG. 1. conformal mapping between the human brain and the unit sphere [\[57\]](#).

establishing a conformal mapping between an irregular surface and the unit sphere, we can handle the complicated surface with the simple parameterizations of the unit sphere. Here, we focus on the conformal mapping between genus-0 surfaces. From [\[77\]](#), a diffeomorphic map between two genus-0 surfaces \mathcal{N}_1 and \mathcal{N}_2 is conformal if and only if it is a local minimizer of the corresponding harmonic energy. Hence, one effective way to compute the conformal mapping between two genus-0 surfaces is to minimize the harmonic energy of the map. Before introducing the harmonic energy minimization model and the diffeomorphic mapping, we review some related concepts on manifold. Let $\phi_{\mathcal{N}_1}(x^1, x^2) : \mathbb{R}^2 \rightarrow \mathcal{N}_1 \subset \mathbb{R}^3$, $\phi_{\mathcal{N}_2}(x^1, x^2) : \mathbb{R}^2 \rightarrow \mathcal{N}_2 \subset \mathbb{R}^3$ be the local coordinates on \mathcal{N}_1 and \mathcal{N}_2 , respectively. The first fundamental form on \mathcal{N}_1 is $g = \sum_{ij} g_{ij} dx^i dx^j$, where $g_{ij} = \frac{\partial \phi_{\mathcal{N}_1}}{\partial x^i} \cdot \frac{\partial \phi_{\mathcal{N}_1}}{\partial x^j}$, $i, j = 1, 2$. The first fundamental form on \mathcal{N}_2 is $h = \sum_{ij} h_{ij} dx^i dx^j$, where $h_{ij} = \frac{\partial \phi_{\mathcal{N}_2}}{\partial x^i} \cdot \frac{\partial \phi_{\mathcal{N}_2}}{\partial x^j}$, $i, j = 1, 2$. Given a smooth map $f : \mathcal{N}_1 \rightarrow \mathcal{N}_2$, whose local coordinate representation is $f(x^1, x^2) = (f_1(x^1, x^2), f_2(x^1, x^2))$, the density of the harmonic energy of f is

$$e(f) = \|df\|^2 = \sum_{i,j=1,2} g^{ij} \langle f_* \partial_{x^i}, f_* \partial_{x^j} \rangle_h,$$

where (g^{ij}) is the inverse of (g_{ij}) and the inner product between $f_* \partial_{x^i}$ and $f_* \partial_{x^j}$ is defined as:

$$\langle f_* \partial_{x^i}, f_* \partial_{x^j} \rangle_h = \left\langle \sum_{m=1}^2 \frac{\partial f_m}{\partial x^i} \partial_{y_m}, \sum_{n=1}^2 \frac{\partial f_n}{\partial x^j} \partial_{y_n} \right\rangle_h = \sum_{m,n=1}^2 h_{mn} \frac{\partial f_m}{\partial x^i} \frac{\partial f_n}{\partial x^j}.$$

61 This also defines a new Riemannian metric on \mathcal{N}_1 , $f^*(h)(\vec{v}_1, \vec{v}_2) := \langle f_*(\vec{v}_1), f_*(\vec{v}_2) \rangle_h$,
 62 which is called the pullback metric induced by f and h . Denote by $\mathbb{S}(\mathcal{N}_1, \mathcal{N}_2)$ the set
 63 of smooth maps between \mathcal{N}_1 and \mathcal{N}_2 . Then the harmonic flow minimization problem
 64 solves

$$65 \quad \min_{f \in \mathbb{S}(\mathcal{N}_1, \mathcal{N}_2)} \mathbf{E}(f) = \frac{1}{2} \int_{\mathcal{N}_1} e(f) d\mathcal{N}_1,$$

66 where $\mathbf{E}(f)$ is called the harmonic energy of f . Stationary points of \mathbf{E} are the harmonic
 67 maps from \mathcal{N}_1 to \mathcal{N}_2 . In particular, If $\mathcal{N}_2 = \mathbb{R}^2$, the conformal map $f = (f_1, f_2)$ is
 68 two harmonic functions defined on \mathcal{N}_1 . If we consider a p -harmonic map from n
 69 dimensional manifold \mathcal{M} to n dimensional sphere $S^n \subset \mathbb{R}^{n+1}$, the p -harmonic energy
 70 minimization problem can be written as

$$71 \quad \min_{\vec{F}(x)=(f_1(x), \dots, f_{n+1}(x))} \mathbf{E}_p(\vec{F}) = \frac{1}{p} \int_{\mathcal{M}} \left(\sum_{k=1}^{n+1} \|\nabla_{\mathcal{M}} f_k\|^2 \right)^{p/2} d\mathcal{M}$$

$$\text{s.t.} \quad \vec{F}(x) \in S^n, \quad \forall x \in \mathcal{M}.$$

72 **2.2. Max cut.** Given a graph $G = (V, E)$ with a set of n vertexes V ($|V| = n$)
 73 and a set of edges E . Denote by the weight matrix $W = (w_{ij})$. The max-cut problem
 74 is to split V into two nonempty sets $(S, V \setminus S)$ such that the total weights of edges in
 75 the cut is maximized. For each vertex $i = 1, \dots, n$, we define $x_i = 1$ if $i \in S$ and -1
 76 otherwise. The maxcut problem can be written as

$$77 \quad (2.1) \quad \max_{x \in \mathbb{R}^n} \frac{1}{2} \sum_{i < j} w_{ij} (1 - x_i x_j), \text{ s.t. } x_i^2 = 1, \quad i = 1, \dots, n.$$

78 It is NP-hard. By relaxing the rank-1 constraint xx^\top to a positive semidefinite ma-
 79 trix X and further neglecting the rank-1 constraint on X , we obtain the following
 80 semidefinite program (SDP)

$$81 \quad (2.2) \quad \max_{X \succeq 0} \text{tr}(CX), \text{ s.t. } X_{ii} = 1, \quad i = 1, \dots, n,$$

82 where C is the graph Laplacian matrix divided by 4, i.e., $C = -\frac{1}{4}(\text{diag}(We) - W)$. If
 83 we decompose $X = V^\top V$ with $V := [V_1, \dots, V_n] \in \mathbb{R}^{p \times n}$, a nonconvex relaxation of
 84 (2.1) is

$$85 \quad (2.3) \quad \max_{V=[V_1, \dots, V_n]} \text{tr}(CV^\top V), \text{ s.t. } \|V_i\|_2 = 1, \quad i = 1, \dots, n.$$

86 It is an optimization problem over multiple spheres.

87 **2.3. Low-rank nearest correlation estimation.** Given a symmetric matrix
 88 $C \in \mathbb{S}^n$ and a non-negative symmetric weight matrix $H \in \mathbb{S}^n$, this problem is to find
 89 a correlation matrix X of low rank such that the distance weighted by H between X
 90 and C is minimized:

$$91 \quad (2.4) \quad \min_{X \succeq 0} \frac{1}{2} \|H \odot (X - C)\|_F^2, \text{ s.t. } X_{ii} = 1, \quad i = 1, \dots, n, \quad \text{rank}(X) \leq p.$$

Algorithms for solving (2.4) can be found in [80, 34]. Similar to the maxcut problem, we decompose the low-rank matrix X with $X = V^\top V$, in which $V = [V_1, \dots, V_n] \in \mathbb{R}^{p \times n}$. Therefore, problem (2.4) is converted to a quartic polynomial optimization problem over multiple spheres:

$$\min_{V \in \mathbb{R}^{p \times n}} \frac{1}{2} \|H \odot (V^\top V - C)\|_F^2, \text{ s.t. } \|V_i\|_2 = 1, i = 1, \dots, n.$$

2.4. Phase retrieval. Given some modules of a complex signal $x \in \mathbb{C}^n$ under linear measurements, a classic model for phase retrieval is to solve

$$(2.5) \quad \begin{aligned} &\text{find } x \in \mathbb{C}^n \\ &\text{s.t. } |Ax| = b, \end{aligned}$$

where $A \in \mathbb{C}^{m \times n}$ and $b \in \mathbb{R}^m$. This problem plays an important role in X-ray, crystallography imaging, diffraction imaging and microscopy. Problem (2.5) is equivalent to the following problem, which minimizes the phase variable y and signal variable x simultaneously:

$$(2.6) \quad \begin{aligned} &\min_{x \in \mathbb{C}^n, y \in \mathbb{C}^m} \|Ax - y\|_2^2 \\ &\text{s.t. } |y| = b. \end{aligned}$$

In [89], the problem above is rewritten as

$$(2.7) \quad \begin{aligned} &\min_{x \in \mathbb{C}^n, u \in \mathbb{C}^m} \frac{1}{2} \|Ax - \text{diag}\{b\}u\|_2^2 \\ &\text{s.t. } |u_i| = 1, i = 1, \dots, m. \end{aligned}$$

For a fixed phase u , the signal x can be represented by $x = A^\dagger \text{diag}\{b\}u$. Hence, problem (2.6) is converted to

$$(2.7) \quad \begin{aligned} &\min_{u \in \mathbb{C}^m} u^* M u \\ &\text{s.t. } |u_i| = 1, i = 1, \dots, m, \end{aligned}$$

where $M = \text{diag}\{b\}(I - AA^\dagger)\text{diag}\{b\}$ is positive definite. It can be regarded as a generalization of the maxcut problem to complex spheres.

If we denote $X = uu^*$, (2.7) can also be modelled as the following SDP problem [21]

$$\min \text{tr}(MX) \quad \text{s.t. } X \succeq 0, \text{ rank}(X) = 1,$$

which can be further relaxed as

$$\min \text{tr}(MX) \quad \text{s.t. } \text{rank}(X) = 1,$$

whose constraint is a manifold.

2.5. Bose-Einstein condensates. In Bose-Einstein condensates (BEC), the total energy functional is defined as

$$E(\psi) = \int_{\mathbb{R}^d} \left[\frac{1}{2} |\nabla \psi(w)|^2 + V(w) |\psi(w)|^2 + \frac{\beta}{2} |\psi(w)|^4 - \Omega \bar{\psi}(w) L_z(w) \right] dw,$$

where $w \in \mathbb{R}^d$ is the spatial coordinate vector, $\bar{\psi}$ is the complex conjugate of ψ , $L_z = -i(x\partial - y\partial x)$, $V(w)$ is an external trapping potential, and β, Ω are given constants. The ground state of BEC is defined as the minimizer of the following optimization problem

$$\min_{\phi \in S} E(\phi),$$

where the spherical constraint S is

$$S = \left\{ \phi : E(\phi) \leq \infty, \int_{\mathbb{R}^d} |\phi(w)|^2 dw = 1 \right\}.$$

The Euler-Lagrange equation of this problem is to find $(\mu \in \mathbb{R}, \phi(w))$ such that

$$\mu\phi(w) = -\frac{1}{2}\nabla^2\phi(w) + V(w)\phi(w) + \beta|\phi(w)|^2\phi(w) - \Omega L_z\phi(w), \quad \xi \in \mathbb{R}^d,$$

and

$$\int_{\mathbb{R}^d} |\phi(w)|^2 dw = 1.$$

Utilizing some proper discretization, such as finite difference, sine pseudospectral and Fourier pseudospectral methods, we obtain a discretized BEC problem

$$\min_{x \in \mathbb{C}^M} f(x) := \frac{1}{2}x^*Ax + \frac{\beta}{2} \sum_{j=1}^M |x_j|^4, \quad \text{s.t.} \quad \|x\|_2 = 1,$$

where $M \in \mathbb{N}$, β are given constants and $A \in \mathbb{C}^{M \times M}$ is Hermitian. Consider the case that x and A are real. Since $x^\top x = 1$, multiplying the quadratic term of the objective function by $x^\top x$, we obtain the following equivalent problem

$$\begin{cases} \min_{x \in \mathbb{R}^M} f(x) = \frac{1}{2}x^{*\top}Axx^\top x + \frac{\beta}{2} \sum_{i=1}^M |x_i|^4 \\ \text{s.t.} \quad \|x\|_2 = 1. \end{cases}$$

The problem above can be also regarded as the best rank-1 tensor approximation of a fourth-order tensor \mathcal{F} [39], with

$$\mathcal{F}_{\pi(i,j,k,l)} = \begin{cases} a_{kl}/4, & i = j = k \neq l, \\ a_{kl}/12, & i = j, i \neq k, i \neq l, k \neq l, \\ (a_{ii} + a_{kk})/12, & i = j \neq k = l, \\ a_{ii}/2 + \beta/4, & i = j = k = l, \\ 0, & \text{otherwise.} \end{cases}$$

For the complex case, we can obtain a best rank-1 complex tensor approximation problem by a similar fashion. Therefore, BEC is an polynomial optimization problem over single sphere.

2.6. Cryo-EM. The Cryo-EM problem is to reconstruct a three-dimensional object from a series of two-dimensional projected images $\{P_i\}$ of the object. A classic model formulates it into an optimization problem over multiple orthogonality constraints [81] to compute the N corresponding directions $\{\tilde{R}_i\}$ of $\{P_i\}$, see Figure 2.

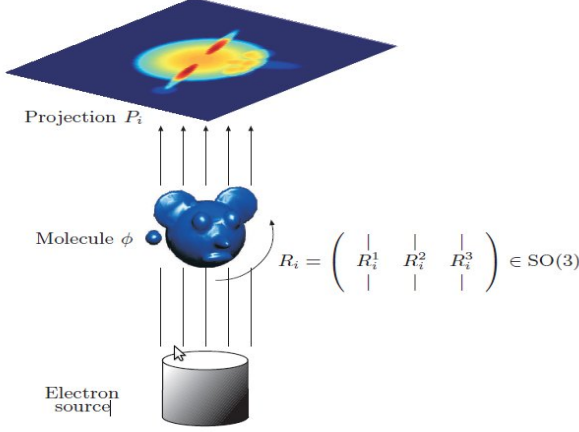


FIG. 2. Recover the 3-D structure from 2-D projections [81].

Each $\tilde{R}_i \in \mathbb{R}^{3 \times 3}$ is a three-dimensional rotation, i.e., $\tilde{R}_i^\top \tilde{R}_i = I_3$ and $\det(\tilde{R}_i) = 1$. Let $\tilde{c}_{ij} = (x_{ij}, y_{ij}, 0)$ be the common line of P_i and P_j (viewed in P_i). If the data are exact, it follows from the Fourier projection-slice theorem [81], the common lines coincide, i.e.,

$$\tilde{R}_i \tilde{c}_{ij} = \tilde{R}_j \tilde{c}_{ji}.$$

Since the third column of \tilde{R}_i^3 can be represented by the first two columns \tilde{R}_i^1 and \tilde{R}_i^2 as $\tilde{R}_i^3 = \pm \tilde{R}_i^1 \times \tilde{R}_i^2$, the rotations $\{\tilde{R}_i\}$ can be compressed as a 3-by-2 matrix. Therefore, the corresponding optimization problem is

$$(2.8) \quad \min_{R_i} \sum_{i=1}^N \rho(R_i c_{ij}, R_j c_{ji}), \quad \text{s.t.} \quad R_i^\top R_i = I_2, R_i \in \mathbb{R}^{3 \times 2},$$

where ρ is a function to measure the distance between two vectors, R_i are the first two columns of \tilde{R}_i and c_{ij} are the first two entries of \tilde{c}_{ij} . In [81], the distance function is set as $\rho(u, v) = \|u - v\|_2^2$. An eigenvector relaxation and SDP relaxation are also presented in [81].

2.7. Linear eigenvalue problem. Linear eigenvalue decomposition and singular value decomposition are the special cases of optimization with orthogonality constraints. Linear eigenvalue problem can be written as

$$(2.9) \quad \min_{X \in \mathbb{R}^{n \times p}} \text{tr}(X^\top A X), \quad \text{s.t.} \quad X^\top X = I,$$

where $A \in \mathbb{S}^n$ is given. Applications from low rank matrix optimization, data mining, principal component analysis and high dimensionality reduction techniques often need to deal with large-scale dense matrices or matrices with some special structures. Although modern computers are developing rapidly, most of the current eigenvalue and singular value decomposition softwares are limited by the traditional design and implementation. In particular, the efficiency may not be significantly improved when

working with thousands of CPU cores. From the perspective of optimization, a series of fast algorithms for solving (2.9) was proposed in [65, 64, 93, 96], whose essential parts can be divided into two steps, updating a subspace to approximate the eigenvector space better and extracting eigenvectors by the Rayleigh-Ritz (RR) process. The main numerical algebraic technique for updating subspaces is usually based on the Krylov subspace, which constructs a series of orthogonal bases sequentially. In [93], the authors propose an equivalent unconstrained penalty function model

$$\min_{X \in \mathbb{R}^{n \times p}} f_\mu(X) := \frac{1}{2} \text{tr}(X^\top A X) + \frac{\mu}{4} \|X^\top X - I\|_F^2,$$

where μ is a parameter. By choosing an appropriate finite large μ , the authors established its equivalence with (2.9). When μ is chosen properly, the number of saddle points of this model is less than that of (2.9). More importantly, the model allows one to design an algorithm that uses only matrix-matrix multiplication. A Gauss-Newton algorithm for calculating low rank decomposition is developed in [65]. When the matrix to be decomposed is of low rank, this algorithm can be more effective while its complexity is similar to the gradient method but with Q linear convergence. Because the bottleneck of many current iterative algorithms is the RR procedure of the eigenvalue decomposition of smaller dense matrices, the authors of [96] proposed a unified augmented subspace algorithmic framework. Each step iteratively solves a linear eigenvalue problem:

$$Y = \arg \min_{X \in \mathbb{R}^{n \times p}} \{\text{tr}(X^\top A X) : X^\top X = I, X \in \mathcal{S}\},$$

where $\mathcal{S} := \text{span}\{X, AX, A^2X, \dots, A^kX\}$ with a small k (which can be far less than p). By combining with the polynomial acceleration technique and deflation in classical eigenvalue calculations, it needs only one RR procedure theoretically to reach a high accuracy.

When the problem dimension reaches the magnitude of $O(10^{42})$, the scale of data storage far exceeds the extent that traditional algorithms can handle. In [108], the authors consider to use a low-rank tensor format to express data matrices and eigenvectors. Let $N = n_1 n_2 \dots n_d$ with positive integer n_1, \dots, n_d . A vector $u \in \mathbb{R}^N$ can be reshaped as a tensor $\mathbf{u} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, whose entries $u_{i_1 i_2 \dots i_d}$ are aligned in reverse lexicographical order, $1 \leq i_\mu \leq n_\mu, \mu = 1, 2, \dots, d$. A tensor \mathbf{u} can be written as the TT format if its entries can be represented by

$$u_{i_1 i_2 \dots i_d} = U_1(i_1) U_2(i_2) \dots U_d(i_d),$$

where $U_\mu(i_\mu) \in \mathbb{R}^{r_{\mu-1} \times r_\mu}, i_\mu = 1, 2, \dots, n_\mu$ and fixed dimensions $r_\mu, \mu = 0, 1, \dots, d$ with $r_0 = r_d = 1$. In fact, the components $r_\mu, \mu = 1, \dots, d-1$ are often equal to a value r (r is then called the TT-rank). Hence, a vector u of dimension $\mathcal{O}(n^d)$ can be stored with $\mathcal{O}(d n r^2)$ entries if the corresponding tensor \mathbf{u} has a TT format. A graphical representation of \mathbf{u} can be seen in Figure 3. The eigenvalue problem can be solved based on the subspace algorithm. By utilizing the alternating direction method with suitable truncations, the performance of the algorithm can be further improved.

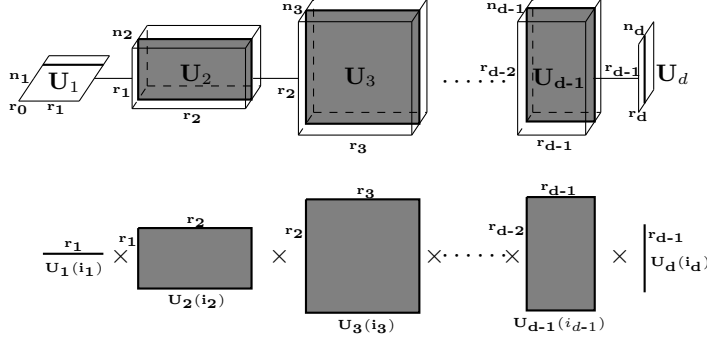


FIG. 3. Graphical representation of a TT tensor of order d with cores \mathbf{U}_μ , $\mu = 1, 2, \dots, d$. The first row is \mathbf{u} , the second row are its entries $u_{i_1 i_2 \dots i_d}$.

The online singular value/eigenvalue decomposition appears in principal component analysis (PCA). The traditional PCA first reads the data and then performs eigenvalue decompositions on the sample covariance matrices. If the data is updated, the principal component vectors need be investigated again based on the new data. Unlike traditional PCA, the online PCA reads the samples one by one and updates the principal component vector in an iterative way, which is essentially a random iterative algorithm of the maximal trace optimization problem. As the sample grows, the online PCA algorithm leads to more accurate main components. An online PCA is proposed and analyzed in [70]. It is proved that the convergence rate is $O(1/n)$ with high probability. A linear convergent VR-PCA algorithm is investigated in [79]. In [59], the scheme in [70] is further proved that under the assumption of Subgaussian's stochastic model, the convergence speed of the algorithm can reach the minimal bound of the information, and the convergence speed is near-global.

2.8. Nonlinear eigenvalue problem. The nonlinear eigenvalue problems from electronic structure calculations are another important source of problems with orthogonality constraints, such as Kohn-Sham (KS) and Hartree-Fock (HF) energy minimization problems. By properly discretizing, KS energy functional can be expressed as

$$E_{\text{ks}}(X) := \frac{1}{4} \text{tr}(X^* L X) + \frac{1}{2} \text{tr}(X^* V_{\text{ion}} X) + \frac{1}{2} \sum_l \sum_i \zeta_l |x_i^* w_l|^2 + \frac{1}{4} \rho^\top L^\dagger \rho + \frac{1}{2} e^\top \epsilon_{\text{xc}}(\rho),$$

where $X \in \mathbb{C}^{n \times p}$ satisfies $X^* X = I_p$, $\rho = \text{diag}(X X^*)$ is the charge density and $\mu_{\text{xc}}(\rho) := \frac{\partial \epsilon_{\text{xc}}(\rho)}{\partial \rho}$ and e are vectors in \mathbb{R}^n with elements all of ones. More specifically, L is a finite dimensional representation of the Laplacian operator, V_{ion} is a constant example, w_l represents a discrete reference projection function, ζ_l is a constant of ± 1 , ϵ_{xc} is used to characterize exchange-correlation energy. With the KS energy functional, the KS energy minimization problem is defined as

$$\min_{X \in \mathbb{C}^{n \times p}} E_{\text{ks}}(X) \quad \text{s.t.} \quad X^* X = I_p.$$

Compared to the KS density functional theory, the HF theory can provide a more accurate model. Specifically, it introduces a Fock exchange operator, which is a fourth-

order tensor by some discretization, $\mathcal{V}(\cdot) : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{n \times n}$. The corresponding Fock energy can be expressed as

$$E_f := \frac{1}{4} \langle \mathcal{V}(XX^*)X, X \rangle = \frac{1}{4} \langle \mathcal{V}(XX^*), XX^* \rangle.$$

The HF energy minimization problem is then

$$(2.10) \quad \min_{X \in \mathbb{C}^{n \times p}} E_{\text{hf}}(X) := E_{\text{ks}}(X) + E_f(X) \quad \text{s.t.} \quad X^*X = I_p.$$

The first-order optimality conditions of KS and HF energy minimization problems correspond to two different nonlinear eigenvalue problems. Taking KS energy minimization as an example, the first-order optimality condition is

$$(2.11) \quad H_{\text{ks}}(\rho)X = X\Lambda, \quad X^*X = I_p,$$

where $H_{\text{ks}}(X) := \frac{1}{2}L + V_{\text{ion}} + \sum_l \zeta_l w_l w_l^* + \text{diag}((\Re L^\dagger)\rho) + \text{diag}(\mu_{\text{xc}}(\rho)^*e)$ and Λ are diagonal matrices. The equation (2.11) is also called the KS equation. The nonlinear eigenvalue problem aims to find some orthogonal eigenvectors satisfying (2.11), while the optimization problem with orthogonality constraints minimizes the objective function under the same constraints. These two problems are connected by the optimality condition and both describe the steady state of the physical system.

The most widely used algorithm for solving the KS equation is the so-called self-consistent field iteration (SCF), which is to solve the following linear eigenvalue problems repeatedly

$$(2.12) \quad H_{\text{ks}}(\rho_k)X_{k+1} = X_{k+1}\Lambda_{k+1}, \quad X_{k+1}^*X_{k+1} = I_p,$$

where $\rho_k = \text{diag}(X_k X_k^*)$. In practice, to accelerate the convergence, we often replace the charge density ρ_k by a linear combination of the previously existing m charge densities

$$\rho_{\text{mix}} = \sum_{j=0}^{m-1} \alpha_j \rho_{k-j}.$$

In the above expression, $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_{m-1})$ is the solution to the following minimization problem:

$$\min_{\alpha^\top e = 1} \|R\alpha\|^2,$$

where $R = (\Delta\rho_k, \Delta\rho_{k-1}, \dots, \Delta\rho_{k-m+1})$, $\delta_j = \rho_j - \rho_{j-1}$ and e is a n -dimensional vector of all entries ones. After obtaining ρ_{mix} , we replace $H_{\text{ks}}(\rho_k)$ in (2.12) with $H_{\text{ks}}(\rho_{\text{mix}})$ and execute the iteration (2.12). This technique is called charge mixing. For more details, one can refer to [72, 73, 84].

Since SCF may not converge, many researchers have recently developed optimization algorithms for the electronic structure calculation that can guarantee convergence. In [111], the manifold gradient method is directly extended to solve the KS minimum problem. The algorithm complexity is mainly from the calculation of the total energy and its gradient calculation, and the projection on the Stiefel manifold. Its complexity at each step is much lower than the linear eigenvalue problem and it is easy to be parallelized. Extensive numerical experiments based on the software

277 packages Octopus and RealSPACES show that the algorithm is often more efficient
 278 than SCF. In fact, the iteration (2.12) of SCF can be understood as an approximate
 279 Newton algorithm in the sense that the complicated part of the Hessian of the total
 280 energy is not considered:

$$281 \quad \min_{X \in \mathbb{C}^{n \times p}} q(X) := \frac{1}{2} \text{tr}(X^* H_{\text{ks}}(\rho_k) X) \quad \text{s.t.} \quad X^* X = I_p.$$

282 Since $q(X)$ is only a local approximation model of $E_{\text{ks}}(X)$, there is no guarantee that
 283 the above model ensures a sufficient decrease of $E_{\text{ks}}(X)$.

284 An explicit expression of the complicated part of the Hessian matrix is derived in
 285 [92]. Although this part is not suitable for an explicit storage, its operation with a
 286 vector is simple and feasible. Hence, the full Hessian matrix can be used to improve
 287 the reliability of Newton's method. By adding regularization terms, the global con-
 288 vergence is also guaranteed. A few other related works include [27, 113, 110, 33, 55].

289 The ensemble-based density functional theory is especially important when the
 290 spectrum of the Hamiltonian matrix has no significant gaps. The KS energy min-
 291 imization model is modified by allowing the charge density to contain more wave
 292 functions. Specifically, $\rho(r) = \sum_{i=1}^p f_i |\psi_i(r)|^2$ where $p \geq p_e$ and the fraction occu-
 293 pation $0 \leq f_i \leq 1$ is to ensure that the total charge density of the total orbit is p ,
 294 i.e., $\sum_{i=1}^p f_i = p_e$. To calculate the fractional occupancy, the energy functional in the
 295 ensemble model introduces a temperature T associated with an entropy $\alpha R(f)$, where

$$296 \quad \alpha := \kappa_B T, \quad \kappa_B \text{ is the Boltzmann constant, } R(f) = \sum_{i=1}^p s(f_i),$$

$$297 \quad s(t) = \begin{cases} t \ln t + (1-t) \ln(1-t), & 0 < t < 1, \\ 0, & \text{otherwise.} \end{cases}$$

298 This method is often referred as the KS energy minimization model with temperature
 299 or the ensemble KS energy minimization model (EDFT). Similar to the KS energy
 300 minimization model, by using the appropriate discretization, the wavefunction can be
 301 represented with $X = [x_1, \dots, x_p] \in \mathbb{C}^{n \times p}$. The discretized charge density in EDFT
 302 can be written as

$$303 \quad \rho(X, f) := \text{diag}(X \text{diag}(f) X^*).$$

304 Obviously, $\rho(X, f)$ is real. The corresponding discretized energy functional is

$$305 \quad M(X, f) = \text{tr}(\text{diag}(f) X^* A X) + \frac{1}{2} \rho^\top L^\dagger \rho + e^\top \epsilon_{\text{xc}}(\rho) + \alpha R(f).$$

306 The discretized EDFT model is

$$307 \quad (2.13) \quad \begin{aligned} & \min_{X \in \mathbb{C}^{n \times p}, f \in \mathbb{R}^p} M(X, f) \\ & \text{s.t.} \quad X^* X = I_p, \\ & \quad e^\top f = p_e \quad 0 \leq f \leq 1. \end{aligned}$$

308 Although SCF can be generalized to this model, its convergence is still not guaranteed.
 309 An equivalent simple model with only one-ball constraint is proposed in [86]. It is

solved by a proximal gradient method such that the terms other than the entropy function term are linearized. An explicit solution of the subproblem is then derived and the convergence of the algorithm is established.

2.9. Approximation models for integer programming. Many optimization problems arising from data analysis are NP-hard integer programmings. Spherical constraints and orthogonal constraints are often used to obtain approximate solutions with high quality. Consider optimization problem over the permutation matrices:

$$\min_{X \in \Pi_n} f(X),$$

where $f(X) : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ is differentiable, and Π_n is a collection of n -order permutation matrices

$$\Pi_n := \{X \in \mathbb{R}^{n \times n} : Xe = X^\top e = e, X_{ij} \in \{0, 1\}\}.$$

This constraint is equivalent to

$$\Pi_n := \{X \in \mathbb{R}^{n \times n} : X^\top X = I_n, X \geq 0\}.$$

It is proved in [49] that it is equivalent to an L_p -regularized optimization problem over the doubly stochastic matrices, which is much simpler than the original problem. An estimation of the lower bound of the non-zero elements at the stationary points are presented. Combining with the cut plane method, a novel gradient-type algorithm with negative proximal terms is also proposed.

Given k communities S_1, S_2, \dots, S_k and the set of partition matrix P_n^k , where the partition matrix $X \in P_n^k$ means $X_{ij} = 1, i, j \in S_t, t \in \{1, \dots, k\}$, otherwise $X_{ij} = 0$. Let A be the adjacency matrix of the network, $d_i = \sum_j A_{ij}, i \in \{1, \dots, n\}$ and $\lambda = 1/\|d\|_2$. Define the matrix $C := -(A - \lambda dd^\top)$. The community detection problem in social networks is to find a partition matrix to maximize the modularity function under the stochastic block model:

$$(2.14) \quad \min_X \langle C, X \rangle \quad \text{s.t. } X \in P_n^k.$$

A SDP relaxation of (2.14) is

$$\begin{aligned} \min_X \quad & \langle C, X \rangle \\ \text{s.t.} \quad & X_{ii} = 1, i = 1, \dots, n, \\ & 0 \leq X_{ij} \leq 1, \forall i, j, \\ & X \succeq 0. \end{aligned}$$

A sparse and low-rank completely positive relaxation technique is further investigated in [106] to transform the model into an optimization problem over multiple non-negative spheres:

$$\begin{aligned} \min_{U \in \mathbb{R}^{n \times k}} \quad & \langle C, UU^\top \rangle \\ \text{s.t.} \quad & \|u_i\|_2 = 1, i = 1, \dots, n, \\ & \|u_i\|_0 \leq p = 1, i = 1, \dots, n, \\ & U \geq 0, \end{aligned}$$

where $1 \leq p \leq r$ is usually taken as a small number so that U can be stored for large-scale data sets. The equivalence to the original problem is proved theoretically and an efficient row-by-row type block coordinate descent method is proposed. In order to quickly solve network problems whose dimension is more than 10 million, an asynchronous parallel algorithm is further developed.

2.10. Deep learning. Batch normalization is a very popular technique in deep neural networks. It avoids internal covariance translation by normalizing the input of each neuron. The space formed by its corresponding coefficient matrix can be regarded as a Riemannian manifold. For a deep neural network, batch normalization usually involves input processing before the nonlinear activation function. Define x and w as the outputs of the previous layer and the parameter vector for the current neuron, the batch normalization of $z := w^\top x$ can be written as

$$\text{BN}(z) = \frac{z - \mathbf{E}(z)}{\text{Var}(z)} = \frac{w^\top (x - \mathbf{E}(x))}{\sqrt{w^\top R_{xx} w}} = \frac{u^\top (x - \mathbf{E}(x))}{\sqrt{u^\top R_{xx} u}},$$

where $u := w/|w|$, $\mathbf{E}(z)$ is the expectation of random variable z and R_{xx} are the covariance matrices of x . From the definition, we have $\text{BN}(w^\top x) = \text{BN}(u^\top x)$ and

$$\frac{\partial \text{BN}(w^\top x)}{\partial x} = \frac{\partial \text{BN}(u^\top x)}{\partial x}, \quad \frac{\partial \text{BN}(z)}{\partial w} = \frac{1}{w} \frac{\partial \text{BN}(z)}{\partial u}.$$

Therefore, the use of the batch standardization ensures that the model does not explode with large learning rates and that the gradient is invariant to linear scaling during propagation.

Since $\text{BN}(cw^\top x) = \text{BN}(w^\top x)$ holds for any constant c , the optimization problem for deep neural networks using batch normalization can be written as

$$\min_{X \in \mathcal{M}} L(X), \quad \mathcal{M} = S^{n_1-1} \times \dots \times S^{n_m-1} \times R^l,$$

where $L(X)$ is the loss function, S^{n-1} is a sphere in \mathbb{R}^n (can also be viewed as a Grassmann manifold), n_1, \dots, n_m are the dimensions of the weight vectors, m is the number of weight vectors, and l is the number of remaining parameters to be decided, including deviations and other weight parameters. For more information, we refer to [26].

2.11. Sparse PCA. In the traditional PCA, the obtained principle eigenvectors are usually not sparse, which leads to high computational cost for computing the principle components. Sparse PCA [51] wants to find principle eigenvectors with few non-zero elements. The mathematical formulation is

$$(2.16) \quad \begin{aligned} \min_{X \in \mathbb{R}^{n \times p}} \quad & -\text{tr}(X^\top A^\top A X) + \rho \|X\|_1 \\ \text{s.t.} \quad & X^\top X = I_p, \end{aligned}$$

where $\|X\|_1 = \sum_{ij} |X_{ij}|$ and $\rho > 0$ is a trade-off parameter. When $\rho = 0$, this reduces to the traditional PCA problem. For $\rho > 0$, the term $\|X\|_1$ plays a role to promote sparsity. Problem (2.16) is a non-smooth optimization problem on the Stiefel manifold.

2.12. Low-rank matrix completion.

The low-rank matrix completion problem has important applications in computer vision, pattern recognitions, statistics, etc. It can be formulated as

$$(2.17) \quad \begin{aligned} \min_X \quad & \text{rank}(X) \\ \text{s.t.} \quad & X_{ij} = A_{ij}, (i, j) \in \Omega, \end{aligned}$$

where X is the matrix that we want to recover (some of its entries are known) and Ω is the index set of observed entries. Due to the difficulty of the rank, a popular approach is to relax it into a convex model using the nuclear norm. The equivalence between this convex problem and the nonconvex problem (2.17) is ensured under certain conditions. Another way is to use a low rank decomposition on X and then solve the corresponding unconstrained optimization problem [95]. If the rank of the ground-truth matrix A is known, an alternative model for a fixed-rank matrix completion is

$$(2.18) \quad \min_{X \in \mathbb{R}^{n \times p}} \|\mathbf{P}_\Omega(X - A)\|_F^2, \text{ s.t. } \text{rank}(X) = r,$$

where \mathbf{P}_Ω is a projection with $\mathbf{P}_\Omega(X)_{ij} = X_{ij}$, $(i, j) \in \Omega$ and 0 otherwise, and $r = \text{rank}(A)$. The set $\text{Fr}(m, n, r) := \{X \in \mathbb{R}^{m \times n} : \text{rank}(X) = r\}$ is a matrix manifold, called fixed-rank manifold. The related geometry is analyzed in [87]. Consequently, problem (2.18) can be solved by optimization algorithms on manifold. Problem (2.18) can deal with Gaussian noise properly. For data sets with a few outliers, the robust low-rank matrix completion problem (with the prior knowledge r) considers:

$$(2.19) \quad \min_{X \in \mathbb{R}^{n \times p}} \|\mathbf{P}_\Omega(X - A)\|_1, \text{ s.t. } \text{rank}(X) = r,$$

where $\|X\|_1 = \sum_{i,j} |X_{ij}|$. Problem (2.19) is a non-smooth optimization problem on the fixed-rank matrix manifold. For some related algorithms for (2.18) and (2.19), the readers can refer to [91, 23].

2.13. Sparse blind deconvolution.

Blind deconvolution is to recover a convolution kernel $a_0 \in \mathbb{R}^k$ and signal $x_0 \in \mathbb{R}^m$ from their convolution

$$y = a_0 \otimes x_0,$$

where $y \in \mathbb{R}^m$. Since there are infinitely many pairs (a_0, x_0) satisfying this condition, this problem is often ill-conditioned. To overcome this issue, some regularization terms and extra constraints are necessary. The sphere-constrained sparse blind deconvolution reformulate the problem as

$$(2.20) \quad \min_{a, x} \|y - a \otimes x\|_2^2 + \mu \|x\|_1, \text{ s.t. } \|a\|_2 = 1,$$

where μ is a parameter to control the sparsity of the signal x . This is a non-smooth optimization problem on the product manifold of a sphere and \mathbb{R}^m . Some related background and the corresponding algorithms can be found in [112].

2.14. Non-negative PCA. Since the principle eigenvectors obtained by the traditional PCA may not be sparse, one can enforce the sparsity by adding non-negativity constraints. The problem is formulated as

$$(2.20) \quad \min_{X \in \mathbb{R}^{n \times p}} \text{tr}(X^\top A A^\top X) \text{ s.t. } X^\top X = I_p, X \geq 0,$$

where $A = [a_1, \dots, a_k] \in \mathbb{R}^{n \times k}$ are given data points. Under the constraints, the variable X has at most one non-zero element in each row. This actually helps to guarantee the sparsity of the principle eigenvectors. Problem (2.20) is an optimization problem with manifold and non-negative constraints. Some related information can be found in [102, 68].

2.15. K -means clustering. K -means clustering is a fundamental problem in data mining. Given n data points (x_1, x_2, \dots, x_n) where each data point is a d -dimensional vector, k -means is to partition them into k clusters $S := \{S_1, S_2, \dots, S_k\}$ such that the within-cluster sum of squares is minimized. Each data point belongs to the cluster with the nearest mean. The mathematical form is

$$(2.21) \quad \min_S \sum_{i=1}^k \sum_{x \in S_i} \|x - c_i\|^2,$$

where $c_i = \frac{1}{\text{card}(S_i)} \sum_{x \in S_i} x$ is the center of i -th cluster and $\text{card}(S_i)$ is the cardinality of S_i . Equivalently, problem (2.21) can be written as [24, 61, 99]:

$$(2.22) \quad \begin{aligned} \min_{Y \in \mathbb{R}^{n \times k}} \quad & \text{tr}(Y^\top D Y) \\ \text{s.t.} \quad & Y Y^\top \mathbf{1} = \mathbf{1}, \\ & Y^\top Y = I_k, Y \geq 0, \end{aligned}$$

where $D_{ij} := \|x_i - x_j\|^2$ is the squared Euclidean distance matrix. Problem (2.22) is a minimization over the Stiefel manifold with linear constraints and non-negative constraints.

3. Algorithms for manifold optimization. In this section, we introduce a few state-of-the-art algorithms for optimization problems on Riemannian manifold. Let us start from the concepts of manifold optimization.

3.1. Preliminaries on Riemannian manifold. A d -dimensional manifold \mathcal{M} is a Hausdorff and second-countable topological space, which is homeomorphic to the d -dimensional Euclidean space locally via a family of charts. When the transition maps of intersecting charts are smooth, the manifold \mathcal{M} is called a smooth manifold. Intuitively, the tangent space $T_x \mathcal{M}$ at a point x of a manifold \mathcal{M} is the set of the tangent vectors of all the curves at x . Mathematically, a tangent vector ξ_x to \mathcal{M} at x is a mapping such that there exists a curve γ on \mathcal{M} with $\gamma(0) = x$, satisfying

$$\xi_x u := \dot{\gamma}(0)u \triangleq \left. \frac{d(u(\gamma(t)))}{dt} \right|_{t=0}, \quad \forall u \in \mathfrak{S}_x(\mathcal{M}),$$

where $\mathfrak{S}_x(\mathcal{M})$ is the set of all real-valued functions f defined in a neighborhood of x in \mathcal{M} . Then, the tangent space $T_x \mathcal{M}$ to \mathcal{M} is defined as the set of all tangent

vectors to \mathcal{M} at x . If \mathcal{M} is equipped with a smoothly varied inner product $g(\cdot, \cdot) := \langle \cdot, \cdot \rangle_x$ on the tangent space, then (\mathcal{M}, g) is a Riemannian manifold. In practice, different Riemannian metrics may be investigated to design efficient algorithms. The Riemannian gradient $\text{grad } f(x)$ of a function f at x is a unique vector in $T_x \mathcal{M}$ satisfying

$$\langle \text{grad } f(x), \xi \rangle_x = Df(x)[\xi], \quad \forall \xi \in T_x \mathcal{M},$$

where $Df(x)[\xi]$ is the derivative of $f(\gamma(t))$ at $t = 0$, $\gamma(t)$ is any curve on the manifold that satisfies $\gamma(0) = x$ and $\dot{\gamma}(0) = \xi$. The Riemannian Hessian $\text{Hess } f(x)$ is a mapping from the tangent space $T_x \mathcal{M}$ to the tangent space $T_x \mathcal{M}$:

$$\text{Hess } f(x)[\xi] := \tilde{\nabla}_\xi \text{grad } f(x)$$

where $\tilde{\nabla}$ is the Riemannian connection [2]. For a function f defined on the manifold, if it can be extended to the ambient Euclidean space $\mathbb{R}^{n \times p}$, we have its Riemannian gradient $\text{grad } f$ and Riemannian Hessian $\text{Hess } f$:

$$(3.1) \quad \begin{aligned} \text{grad } f(X) &= \mathbf{P}_{T_x \mathcal{M}}(\nabla f(X)), \\ \text{Hess } f(X)[U] &= \mathbf{P}_{T_x \mathcal{M}}(D \text{grad } f(X)[U]) \end{aligned}$$

where D is the Euclidean derivative. More detailed information on the related backgrounds can be found in [2].

We next briefly introduce some typical manifolds.

- Sphere $\text{Sp}(n-1) := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$. Let $x(t)$ with $x(0) = x$ be a curve on sphere, i.e., $x(t)^\top x(t) = 1$ for all t . Taking the derivatives with respect to t , we have

$$\dot{x}(t)^\top x(t) + x(t)^\top \dot{x}(t) = 0.$$

At $t = 0$, we have $\dot{x}(0)x + x^\top \dot{x}(0) = 0$. Hence, the tangent space is

$$T_x \text{Sp}(n-1) = \{z : z^\top x = 0\}.$$

The projection operator is defined as

$$\mathbf{P}_{T_x \text{Sp}(n-1)}(z) = (I - xx^\top)z.$$

For a function defined on $\text{Sp}(n-1)$ with respect to the Euclidean metric $g_x(u, v) = u^\top v$, $u, v \in T_x \text{Sp}(n-1)$, its Riemannian gradient and Hessian at x can be represented by

$$\begin{aligned} \text{grad } f(x) &= \mathbf{P}_{T_x \text{Sp}(n-1)}(\nabla f(x)), \\ \text{Hess } f(x)[u] &= \mathbf{P}_{T_x \text{Sp}(n-1)}(\nabla^2 f(x)[u] - ux^\top \nabla f(x)), \quad u \in T_x \text{Sp}(n-1). \end{aligned}$$

- Stiefel manifold $\text{St}(n, p) := \{X \in \mathbb{R}^{n \times p} : X^\top X = I_p\}$. By a similar calculation as the spherical case, we have its tangent space:

$$T_X \text{St}(n, p) = \{Z : Z^\top X + X^\top Z = 0\}.$$

The projection operator onto $T_X \text{St}(n, p)$ is

$$\mathbf{P}_{T_X \text{St}(n, p)}(Z) = Z - X \text{sym}(X^\top Z),$$

where $\text{sym}(Z) := (Z + Z^\top)/2$. Given a function defined on $\text{St}(n, p)$ with respect to the Euclidean metric $g_X(U, V) = \text{tr}(U^\top V)$, $U, V \in T_X \text{St}(n, p)$, its Riemannian gradient and Hessian at X can be represented by

$$\begin{aligned} \text{grad } f(X) &= \mathbf{P}_{T_X \text{St}(n, p)}(\nabla f(X)), \\ \text{Hess } f(X)[U] &= \mathbf{P}_{T_X \text{St}(n, p)}(\nabla^2 f(X)[U] - U \text{sym}(X^\top \nabla f(X))), \quad U \in T_X \text{St}(n, p). \end{aligned}$$

- Oblique manifold $\text{Ob}(n, p) := \{X \in \mathbb{R}^{n \times p} \mid \text{diag}(X^\top X) = e\}$. Its tangent space is

$$T_X \text{Ob}(n, p) = \{Z : \text{diag}(X^\top Z) = 0\}.$$

The projection operator onto $T_X \text{Ob}(n, p)$ is

$$\mathbf{P}_{T_X \text{Ob}(n, p)} = Z - X \text{Diag}(\text{diag}(X^\top Z)).$$

Given a function defined on $\text{Ob}(n, p)$ with respect to the Euclidean metric, its Riemannian gradient and Hessian at X can be represented by

$$\begin{aligned} \text{grad } f(X) &= \mathbf{P}_{T_X \text{Ob}(n, p)}(\nabla f(X)), \\ \text{Hess } f(X)[U] &= \mathbf{P}_{T_X \text{Ob}(n, p)}(\nabla^2 f(X)[U] - U \text{Diag}(\text{diag}(X^\top \nabla f(X)))), \end{aligned}$$

with $U \in T_X \text{Ob}(n, p)$.

- Grassmann manifold $\text{Grass}(n, p) := \{\text{span}(X) : X \in \mathbb{R}^{n \times p}, X^\top X = I_p\}$. It denotes the set of all p -dimensional subspaces of \mathbb{R}^n . This manifold is different from other manifolds mentioned above. It is a quotient manifold since each element is an equivalent class of $n \times p$ matrices. From the definition of $\text{Grass}(p, n)$, the equivalence relation \sim is defined as

$$X \sim Y \Leftrightarrow \exists Q \in \mathbb{R}^{p \times p} \text{ with } Q^\top Q = Q Q^\top = I, \text{ s.t. } Y = XQ.$$

Its element is of the form

$$[X] := \{Y \in \mathbb{R}^{n \times p} : Y^\top Y = I, Y \sim X\}.$$

Then $\text{Grass}(n, p)$ is a quotient manifold of $\text{St}(n, p)$, i.e., $\text{St}(n, p)/\sim$. Due to this equivalence, a tangent vector ξ of $T_X \text{Grass}(n, p)$ may have many different representations in its equivalence class. To find the unique representation, a horizontal space [2, Section 3.5.8] is introduced. For a given $X \in \mathbb{R}^{n \times p}$ with $X^\top X = I_p$, the horizontal space is

$$\mathcal{H}_X \text{Grass}(n, p) = \{Z : Z^\top X = 0\}.$$

Here, a function of the horizontal space is similar to the tangent space when computing the Riemannian gradient and Hessian. We have the projection onto the horizontal space

$$\mathbf{P}_{\mathcal{H}_X \text{Grass}(n, p)}(Z) = Z - X X^\top Z.$$

Given a function defined on $\text{Grass}(n, p)$ with respect to the Euclidean metric $g_X = \text{tr}(U^\top V)$, $U, V \in \mathcal{H}_X \text{Grass}(n, p)$, its Riemannian gradient and Hessian at X can be represented by

$$\begin{aligned} \text{grad } f(X) &= \mathbf{P}_{\mathcal{H}_X \text{Grass}(n, p)}(\nabla f(X)), \\ \text{Hess } f(X)[U] &= \mathbf{P}_{\mathcal{H}_X \text{Grass}(n, p)}(\nabla^2 f(X)[U] - U X^\top \nabla f(X)), \quad U \in T_X \text{Grass}(n, p). \end{aligned}$$

- Fixed-rank manifold $\text{Fr}(n, p, r) := \{X \in \mathbb{R}^{n \times p} : \text{rank}(X) = r\}$ is a set of all $n \times p$ matrices of rank r . Using the singular value decomposition (SVD), this manifold can be represented equivalently by

$$\text{Fr}(n, p, r) = \{U\Sigma V^\top : U \in \text{St}(n, r), V \in \text{St}(p, r), \Sigma = \text{diag}(\sigma_i)\},$$

where $\sigma_1 \geq \dots \geq \sigma_k > 0$. Its tangent space at $X = U\Sigma V^\top$ is

$$\begin{aligned} T_X \text{Fr}(n, p, r) &= \left\{ [U, U_\perp] \begin{pmatrix} \mathbb{R}^{r \times r} & \mathbb{R}^{r \times (p-r)} \\ \mathbb{R}^{(n-r) \times r} & 0_{(n-r) \times (p-r)} \end{pmatrix} [V, V_\perp]^\top \right\} \\ &= \{UMV^\top + U_p V^\top + UV_p^\top : M \in \mathbb{R}^{r \times r}, \\ &\quad U_p \in \mathbb{R}^{n \times r}, U_p U = 0, V_p \in \mathbb{R}^{p \times r}, V_p^\top V = 0\}, \end{aligned}$$

where U_\perp and V_\perp are the orthogonal complements of U and V , respectively. The projection operator onto the tangent space is

$$\mathbf{P}_{T_X \text{Fr}(n, p, r)}(Z) = P_U Z P_V + P_U^\perp Z P_V + P_U Z P_V^\perp,$$

where $P_U = UU^\top$ and $P_U^\perp = I - P_U$. Comparing the representation with (3.2), we have

$$M(Z; X) := U^\top Z X, U_p(Z; X) = P_U^\perp Z V, V_p(Z; X) = P_V^\perp Z^\top U.$$

Given a function defined on $\text{Fr}(n, p, r)$ with respect to the Euclidean metric $g_X(U, V) = \text{tr}(U^\top V)$, its Riemannian gradient and Hessian at $X = U\Sigma V$ can be represented by

$$\begin{aligned} \text{grad} f(X) &= \mathbf{P}_{T_X \text{Fr}(n, p, r)}(\nabla f(X)), \\ \text{Hess} f(X)[H] &= U \hat{M} V^\top + \hat{U}_p V^\top + U \hat{V}_p^\top, H \in T_X \text{Fr}(n, p, r), \end{aligned}$$

where $\hat{M} = M(\nabla^2 f(X)[H], X)$, $\hat{U}_p = U_p(\nabla^2 f(X)[H]; X) + P_U^\perp \nabla f(X) V_p(H; X) / \Sigma$, $\hat{V}_p = V_p(\nabla^2 f(X)[H]; X) + P_V^\perp \nabla f(X) U_p(H; X) / \Sigma$.

- The set of symmetric positive definite matrices, i.e., $\text{SPD}(n) = \{X \in \mathbb{R}^{n \times n} : X^\top = X, X \succ 0\}$ is a manifold. Its tangent space at X is

$$T_X \text{SPD}(n) = \{Z : Z^\top = Z\}.$$

We have the projection onto $T_X \text{SPD}(n)$:

$$\mathbf{P}_{T_X \text{SPD}(n)}(Z) = (Z^\top + Z)/2.$$

Given a function defined on $\text{SPD}(n, p)$ with respect to the Euclidean metric $g_X(U, V) = \text{tr}(U^\top V)$, $U, V \in T_X \text{SPD}(n)$, its Riemannian gradient and Hessian at X can be represented by

$$\begin{aligned} \text{grad} f(X) &= \mathbf{P}_{T_X \text{SPD}(n)}(\nabla f(X)), \\ \text{Hess} f(X)[U] &= \mathbf{P}_{T_X \text{SPD}(n)}(\nabla^2 f(X)[U]), U \in T_X \text{SPD}(n). \end{aligned}$$

- The set of rank- r symmetric positive semidefinite matrices, i.e., $\text{FrPSD}(n, r) = \{X \in \mathbb{R}^{n \times n} : X = X^\top, X \succeq 0, \text{rank}(X) = r\}$. This manifold can be reformulated as

$$\text{FrPSD}(n, r) = \{YY^\top : Y \in \mathbb{R}^{n \times r}, \text{rank}(Y) = k\},$$

which is a quotient manifold. The horizontal space at Y is

$$T_Y \mathcal{H}_{\text{FrPSD}(n, r)} = \{Z \in \mathbb{R}^{n \times r} : Z^\top Y = Y^\top Z\}.$$

We have the projection operator onto $T_Y \mathcal{H}_{\text{FrPSD}(n, r)}$

$$\mathbf{P}_{T_Y \mathcal{H}_{\text{FrPSD}(n, r)}}(Z) = Z - Y\Omega,$$

where the skew-symmetric matrix Ω is the unique solution of the Sylvester equation $\Omega(Y^\top Y) + (Y^\top Y)\Omega = Y^\top Z - Z^\top Y$. Given a function f with respect to the Euclidean metric $g_Y(U, V) = \text{tr}(U^\top V)$, $U, V \in T_Y \mathcal{H}_{\text{FrPSD}(n, r)}$, its Riemannian gradient and Hessian can be represented by

$$\text{grad} f(Y) = \nabla f(Y),$$

$$\text{Hess} f(X)[U] = \mathbf{P}_{T_Y \mathcal{H}_{\text{FrPSD}(n, r)}}(\nabla^2 f(Y)[U]), \quad U \in T_Y \mathcal{H}_{\text{FrPSD}(n, r)}.$$

3.2. Optimality conditions. We next present the optimality conditions for manifold optimization problem in the following form

$$\begin{aligned} (3.3) \quad & \min_{x \in \mathcal{M}} f(x), \\ & \text{s.t.} \quad c_i(x) = 0, \quad i \in \mathcal{E} := \{1, \dots, \ell\} \\ & \quad c_i(x) \geq 0, \quad i \in \mathcal{I} := \{\ell + 1, \dots, m\}, \end{aligned}$$

where \mathcal{E} and \mathcal{I} denote the index sets of equality constraints and inequality constraints, respectively, and $c_i : \mathcal{M} \rightarrow \mathbb{R}$, $i \in \mathcal{E} \cup \mathcal{I}$ are smooth functions on \mathcal{M} . We mainly adopt the notions in [100]. By keeping the manifold constraint, the Lagrangian function of (3.3) is

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(x), \quad x \in \mathcal{M},$$

where λ_i , $i \in \mathcal{E} \cup \mathcal{I}$ are the Lagrangian multipliers. Here, we notice that the domain of \mathcal{L} is on the manifold \mathcal{M} . Let $\mathcal{A}(x) := \mathcal{E} \cup \{i \in \mathcal{I} : c_i(x) = 0\}$. Then the linear independence constraint qualifications (LICQ) for problem (3.3) holds at x if and only if

$$\text{grad} c_i(x), \quad i \in \mathcal{A}(x) \text{ is linear independent on } T_x \mathcal{M}.$$

Then the first-order necessary conditions can be described as follows.

THEOREM 1 (First-order necessary optimality conditions (KKT conditions)).

Suppose that x^ is a local minima of (3.3) and that the LICQ holds at x^* , then there exist Lagrangian multipliers $\lambda_i^*, i \in \mathcal{E} \cup \mathcal{I}$ such that the following KKT conditions hold:*

$$\begin{aligned} (3.4) \quad & \text{grad} f(x^*) + \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i^* \text{grad} c_i(x^*) = 0, \\ & c_i(x^*) = 0, \quad \forall i \in \mathcal{E}, \\ & c_i(x^*) \geq 0, \quad \lambda_i^* \geq 0, \quad \lambda_i^* c_i(x^*) = 0, \quad \forall i \in \mathcal{I}. \end{aligned}$$

Let x^* and λ_i^* , $i \in \mathcal{E} \cup \mathcal{I}$ be one of the solution of the KKT conditions (3.4). Similar to the case without the manifold constraint, we define a critical cone $\mathcal{C}(x^*, \lambda^*)$ as

$$w \in \mathcal{C}(x^*, \lambda^*) \Leftrightarrow \begin{cases} w \in T_{x^*} \mathcal{M}, \\ \langle \text{grad } c_i(x^*), w \rangle = 0, \forall i \in \mathcal{E}, \\ \langle \text{grad } c_i(x^*), w \rangle = 0, \forall i \in \mathcal{A}(x^*) \cap \mathcal{I} \text{ with } \lambda_i^* > 0, \\ \langle \text{grad } c_i(x^*), w \rangle \geq 0, \forall i \in \mathcal{A}(x^*) \cap \mathcal{I} \text{ with } \lambda_i^* = 0. \end{cases}$$

Then we have the following second-order necessary and sufficient conditions.

THEOREM 2 (Second-order optimality conditions).

- *Second-order necessary conditions:*
Suppose that x^* is a local minima of (3.3) and the LICQ holds at x^* . Let λ^* be the multipliers such that the KKT conditions (3.4) hold. Then we have

$$\langle \text{Hess}_x \mathcal{L}(x^*, \lambda^*)[w], w \rangle \geq 0, \forall w \in \mathcal{C}(x^*, \lambda^*),$$

where $\text{Hess}_x \mathcal{L}(x^*, \lambda^*)$ is the Riemannian Hessian of \mathcal{L} with respect to x at (x^*, λ^*) .

- *Second-order sufficient conditions:*
Suppose that x^* and λ^* satisfy the KKT conditions (3.4). If we further have

$$\langle \text{Hess}_x \mathcal{L}(x^*, \lambda^*)[w], w \rangle > 0, \forall w \in \mathcal{C}(x^*, \lambda^*), w \neq 0,$$

then x^* is a strict local minima of (3.4).

Suppose that we have only the manifold constraint, i.e., $\mathcal{E} \cup \mathcal{I}$ is empty. For a smooth function f on the manifold \mathcal{M} , the optimality conditions take a similar form to the Euclidean unconstrained case. Specifically, if x^* is a first-order stationary point, then it holds that

$$\text{grad } f(x^*) = 0.$$

If x^* is a second-order stationary point, then

$$\text{grad } f(x^*) = 0, \quad \text{Hess } f(x^*) \succeq 0.$$

If x^* satisfies

$$\text{grad } f(x^*) = 0, \quad \text{Hess } f(x^*) \succ 0,$$

then x^* is a strict local minimum. For more details, we refer the reader to [100].

3.3. First-order type algorithms. From the perspective of Euclidean constrained optimization problems, there are many standard algorithms which can solve this optimization problem on manifold. However, since the intrinsic structure of manifolds is not considered, these algorithms may not be effective in practice. By doing curvilinear search along the geodesic, a globally convergent gradient descent method is proposed in [32]. For Riemannian conjugate gradient (CG) methods [82], the parallel translation is used to construct the conjugate directions. Due to the difficulty of calculating geodesics (exponential maps) and parallel translations, computable retraction and vector transport operators are proposed to approximate the exponential

map and the parallel translation [2]. Therefore, more general Riemannian gradient descent methods and CG methods together with convergence analysis are obtained in [2]. These algorithms have been successfully applied to various applications [87, 54]. Numerical experiments exhibit the advantage of using geometry of the manifold. A proximal Riemannian gradient method is proposed in [41]. Specifically, the objective function is linearized using the first-order Taylor expansion on manifold and a proximal term is added. The original problem is then transformed into a series of projection problems on the manifold. For general manifolds, the existence and uniqueness of the projection operator can not be guaranteed. But when the given manifold satisfies certain differentiable properties, the projection operator is always locally well-defined and is also a specific retraction operator [3]. Therefore, in this case, the proximal Riemannian gradient method coincides with the Riemannian gradient method. By generalizing the adaptive gradient method in [30], an adaptive gradient method on manifold is also presented in [41]. In particular, optimization over Stiefel manifold is an important special case of Riemannian optimization. Various efficient retraction operators, vector transport operators and Riemannian metric have been investigated to construct more practical gradient descent and CG methods [94, 48, 114]. Non-retraction based first-order methods are also developed in [33].

We next present a brief introduction of first-order algorithms for manifold optimization. Let us start with the retraction operator R . It is a smooth mapping from the tangent bundle $T\mathcal{M} := \cup_{x \in \mathcal{M}} T_x \mathcal{M}$ to \mathcal{M} , and satisfies

- $R_x(0_x) = x$, 0_x is the zero element in the cut space $T_x \mathcal{M}$,
- $DR_x(0_x)[\xi] = \xi$, $\forall \xi \in T_x \mathcal{M}$,

where R_x is the retraction operator R at x . The well-posedness of the retraction operator is shown in Section 4.1.3 of [2]. The retraction operator provides an efficient way to pull the points from the tangent space back onto the manifold. Let $\xi_k \in T_x \mathcal{M}$ be a descent direction, i.e., $\langle \text{grad} f(x_k), \xi_k \rangle < 0$. Another important concept on manifold is the vector transport operator \mathcal{T} . It is a smooth mapping from the product of tangent bundles $T\mathcal{M} \oplus T\mathcal{M}$ to the tangent bundle $T\mathcal{M}$, and satisfies the following properties.

- There exists a retraction R associated with \mathcal{T} , i.e.,

$$\mathcal{T}_{\eta_x} \xi_x = \left. \frac{d}{dt} R_x(\eta_x + t\xi_x) \right|_{t=0}.$$

- $\mathcal{T}_{0_x} \xi_x = \xi_x$ for all $x \in \mathcal{M}$ and $\xi_x \in T_x \mathcal{M}$.
- $\mathcal{T}_{\eta_x}(a\xi_x + b\zeta_x) = a\mathcal{T}_{\eta_x} \xi_x + b\mathcal{T}_{\eta_x} \zeta_x$.

The vector transport is a generalization of the parallel translation [2, Section 5.4].

The general feasible algorithm framework on the manifold can be expressed as

$$(3.5) \quad x_{k+1} = R_{x_k}(t_k \xi_k),$$

where t_k is a well-chosen step size. Similar to the line search method in Euclidean space, the step size t_k can be obtained by the curvilinear search on the manifold. Here, we take the Armijo search as an example. Given $\rho, \delta \in (0, 1)$, the monotone

and nonmonotone search try to find the smallest integer h to such that

$$(3.6) \quad f(R_{x_k}(t_k \xi_k)) \leq f(x_k) + \rho t_k \langle \text{grad } f(x_k), \xi_k \rangle_{x_k},$$

$$(3.7) \quad f(R_{x_k}(t_k \xi_k)) \leq C_k + \rho t_k \langle \text{grad } f(x_k), \xi_k \rangle_{x_k},$$

respectively, where $\langle \text{grad } f(x_k), \xi_k \rangle_{x_k} := g_{x_k}(\text{grad } f(x_k), \xi_k)$, $t_k = \gamma_k \delta^h$ and γ_k is an initial step size. The reference value C_{k+1} is a convex combination of C_k and $f(x_{k+1})$ and is calculated via $C_{k+1} = (\varrho Q_k C_k + f(x_{k+1}))/Q_{k+1}$, where $C_0 = f(x_0)$, $Q_{k+1} = \varrho Q_k + 1$ and $Q_0 = 1$. From the Euclidean optimization, we know that the Barzilai-Borwein (BB) step size often accelerates the convergence. The BB step size can be generalized to Riemannian manifold [41] as

$$(3.8) \quad \gamma_k^{(1)} = \frac{\langle s_{k-1}, s_{k-1} \rangle_{x_k}}{|\langle s_{k-1}, v_{k-1} \rangle_{x_k}|} \quad \text{or} \quad \gamma_k^{(2)} = \frac{|\langle s_{k-1}, v_{k-1} \rangle_{x_k}|}{\langle v_{k-1}, v_{k-1} \rangle_{x_k}},$$

where

$$s_{k-1} = -t_{k-1} \cdot \mathcal{T}_{x_{k-1} \rightarrow x_k}(\text{grad } f(x_{k-1})), \quad v_{k-1} = \text{grad } f(x_k) + t_{k-1}^{-1} \cdot s_{k-1},$$

and $\mathcal{T}_{x_{k-1} \rightarrow x_k} : T_{x_{k-1}} \mathcal{M} \mapsto T_{x_k} \mathcal{M}$ denotes an appropriate vector transport mapping connecting x_{k-1} and x_k ; see [2, 47]. When \mathcal{M} is a submanifold of an Euclidean space, the Euclidean differences $s_{k-1} = x_k - x_{k-1}$ and $v_{k-1} = \text{grad } f(x_k) - \text{grad } f(x_{k-1})$ are an alternative choice if the Euclidean inner product is used in (3.8). This choice is often attractive since the vector transport is not needed [94, 41]. We note that the differences between first- and second-order algorithms are mainly due to their specific ways of acquiring ξ_k .

In practice, the computational cost and convergence behavior of different retraction operators differ a lot. Similarly, the vector transport plays an important role in CG methods and quasi-Newton methods (we will introduce them later). There are many studies on the retraction operators and vector transports. Here, we take the Stiefel manifold $\text{St}(n, p)$ as an example to introduce several different retraction operators at the current point X for a given step size τ and descent direction $-D$.

- Exponential map [31]

$$R_X^{\text{geo}}(-\tau D) = [X, Q] \exp \left(\tau \begin{bmatrix} -X^\top D & -R^\top \\ R & 0 \end{bmatrix} \right) \begin{bmatrix} I_p \\ 0 \end{bmatrix},$$

where $QR = -(I_n - XX^\top)D$ is the QR decomposition of $-(I_n - XX^\top)D$. This scheme needs to calculate an exponent of a $2p$ -by- $2p$ matrix and an QR decomposition of an n -by- p matrix. From [31], an explicit form of parallel translation is unknown.

- Cayley transform [92]

$$(3.9) \quad R_X^{\text{wy}}(-\tau D) = X - \tau U \left(I_{2p} + \frac{\tau}{2} V^\top U \right)^{-1} V^\top X,$$

where $U = [P_X D, X]$, $V = [X, -P_X D] \in \mathbb{R}^{n \times (2p)}$ with $P_X := (I - \frac{1}{2} X X^\top)$. When $p < n/2$, this scheme is much cheaper than the exponential map. The

associated vector transport is [114]

$$\mathcal{T}_{\eta_X}^{\text{wy}}(\xi_X) = \left(I - \frac{1}{2}W_{\eta_X}\right)^{-1} \left(I + \frac{1}{2}W_{\eta_X}\right) \xi_X, \quad W_Z = P_X \eta_X X - X \eta_X P_X,$$

- Polar decomposition [2]

$$R_X^{\text{pd}}(-\tau D) = (X - \tau D)(I_p + \tau^2 D^\top D)^{-1/2}.$$

The computational cost is lower than the Cayley transform, but the Cayley transform gives a better approximation to the exponential map. The associated vector transport is then defined as [42]

$$\mathcal{T}_{\eta_X}^{\text{pd}} \xi_X = Y \Omega + (I - Y Y^\top) \xi_X (Y^\top (X + \eta_X))^{-1},$$

where $Y = R_X \eta_X$ and $\text{vec}(\Omega) = (Y^\top (X + \eta_X)) \oplus (Y^\top (X + \eta_X))^{-1} \text{vec}(Y^\top \xi_X - \xi_X^\top Y)$ and \oplus is the Kronecker sum, i.e., $A \oplus B = A \otimes I + I \otimes B$ with Kronecker product \otimes . Numerical experiments show that more iterations may be required compared to the Cayley transform.

- QR decomposition

$$R_X^{\text{qr}}(-\tau D) = \text{qr}(X - \tau D).$$

It can be seen as an approximation of the polar decomposition. The main cost is the QR decomposition of a n -by- p matrix. The associated vector transport is defined as [2, Example 8.1.5]

$$\mathcal{T}_{\eta_X}^{\text{qr}} \xi_X = Y \rho_{\text{skew}}(Y^\top \xi_X (Y^\top (X + \eta_X))^{-1}) + (I - Y Y^\top) \xi_X (Y^\top (X + \eta_X))^{-1},$$

where $Y = R_X(\eta_X)$.

The vector transport above requires an associated retraction. Removing the dependence of the retraction, a new class of vector transports is introduced in [46]. Specifically, a jointly smooth operator $\mathcal{L}(x, y) : T_x \mathcal{M} \rightarrow T_y \mathcal{M}$ is defined. In addition, $\mathcal{L}(x, x)$ is required to be an identity for all x . For a d -dimensional submanifold \mathcal{M} of n -dimensional Euclidean space, two popular vector transports are defined by the projection [2, Section 8.1.3]

$$\mathcal{L}^{\text{pj}}(x, y) \xi_x = \mathbf{P}_{T_y \mathcal{M}}(\xi_x),$$

and by parallelization [46]

$$\mathcal{L}^{\text{pl}}(x, y) \xi_x = B_y B_x^\dagger \xi_x,$$

where $B : \mathcal{V} \rightarrow \mathbb{R}^{n \times d} : z \rightarrow B_z$ is a smooth tangent basis field defined on an open neighborhood \mathcal{V} of \mathcal{M} and B_z^\dagger is the pseudo-inverse of B_z . With the tangent basis B_z , we can also represent the vector transport mentioned above intrinsically, which sometimes reduces computational cost significantly [45].

To better understand Riemannian first-order algorithms, we present a Riemannian gradient method [41] in Algorithm 1. One can easily see that the difference to the Euclidean case is an extra retraction step.

The convergence of Algorithm 1 [40, Theorem 1] is given as follows.

Algorithm 1: Riemannian gradient method

```

1 Input  $x_0 \in \mathcal{M}$ . Set  $k = 0$ ,  $\gamma_{\min} \in [0, 1]$ ,  $\gamma_{\max} \geq 1$ ,  $C_0 = f(x_0)$ ,  $Q_0 = 1$ .
2 while  $\|\text{grad } f(x_k)\| \neq 0$  do
3   Compute  $\eta_k = -\text{grad } f(x_k)$ .
4   Calculate  $\gamma_k$  according to (3.8) and set  $\gamma_k = \max(\gamma_{\min}, \min(\gamma_k, \gamma_{\max}))$ .
   Then, compute  $C_k$ ,  $Q_k$  and find a step size  $t_k$  satisfying (3.7).
5   Set  $x_{k+1} \leftarrow R_{x_k}(t_k \eta_k)$ .
6   Set  $k \leftarrow k + 1$ .

```

THEOREM 3. Let $\{x_k\}$ be a sequence generated by Algorithm 1 using the non-monotone line search (3.7). Suppose that f is continuously differentiable on the manifold \mathcal{M} . Then, every accumulation point x_* of the sequence $\{x_k\}$ is a stationary point of problem (1.1), i.e., it holds $\text{grad } f(x_*) = 0$.

Proof. At first, by using $\langle \text{grad } f(x_k), \eta_k \rangle_{x_k} = -\|\text{grad } f(x_k)\|_{x_k}^2 < 0$ and applying [103, Lemma 1.1], we have $f(x_k) \leq C_k$ and $x_k \in \mathcal{L}$ for all $k \in \mathbb{N}$. Next, due to

$$\lim_{t \downarrow 0} \frac{(f \circ R_{x_k})(t\eta_k) - f(x_k)}{t} - \rho \langle \text{grad } f(x_k), \eta_k \rangle_{x_k} = \nabla f(R_{x_k}(0))^\top D R_{x_k}(0) \eta_k + \rho \|\text{grad } f(x_k)\|_{x_k}^2 = -(1 - \rho) \|\text{grad } f(x_k)\|_{x_k}^2 < 0,$$

there always exists a positive step size $t_k \in (0, \gamma_k]$ satisfying the monotone and non-monotone Armijo conditions (3.6) and (3.7), respectively. Now, let $x_* \in \mathcal{M}$ be an arbitrary accumulation point of $\{x_k\}$ and let $\{x_k\}_K$ be a corresponding subsequence that converges to x_* . By the definition of C_{k+1} and (3.6), we have

$$C_{k+1} = \frac{\varrho Q_k C_k + f(x_{k+1})}{Q_{k+1}} < \frac{(\varrho Q_k + 1)C_k}{Q_{k+1}} = C_k.$$

Hence, $\{C_k\}$ is monotonically decreasing and converges to some limit $\bar{C} \in \mathbb{R} \cup \{-\infty\}$. Using $f(x_k) \rightarrow f(x_*)$ for $K \ni k \rightarrow \infty$, we can infer $\bar{C} \in \mathbb{R}$ and thus, we obtain

$$\infty > C_0 - \bar{C} = \sum_{k=0}^{\infty} C_k - C_{k+1} \geq \sum_{k=0}^{\infty} \frac{\rho t_k \|\text{grad } f(x_k)\|_{x_k}^2}{Q_{k+1}}.$$

Due to $Q_{k+1} = 1 + \varrho Q_k = 1 + \varrho + \varrho^2 Q_{k-1} = \dots = \sum_{i=0}^k \varrho^i < (1 - \varrho)^{-1}$, this implies $\{t_k \|\text{grad } f(x_k)\|_{x_k}^2\} \rightarrow 0$. Let us now assume $\|\text{grad } f(x_*)\| \neq 0$. In this case, we have $\{t_k\}_K \rightarrow 0$ and consequently, by the construction of Algorithm 1, the step size $\delta^{-1} t_k$ does not satisfy (3.7), i.e., it holds

$$-\rho(\delta^{-1} t_k) \|\text{grad } f(x_k)\|_{x_k}^2 < f(R_{x_k}(\delta^{-1} t_k \eta_k)) - C_k \leq f(R_{x_k}(\delta^{-1} t_k \eta_k)) - f(x_k)$$

for all $k \in K$ sufficiently large. Since the sequence $\{\eta_k\}_K$ is bounded, the rest of the proof is now identical to the proof of [2, Theorem 4.3.1]. In particular, applying the mean value theorem in (3.10) and using the continuity of the Riemannian metric, we can easily derive a contradiction. We refer to [2] for more details. \square

3.4. Second-order type algorithms. A gradient-type algorithm usually is fast in the early iterations, but it often slows down or even stagnates when the generated iterations are close to an optimal solution. When a high accuracy is required, second-order type algorithms may have its advantage.

By utilizing the exact Riemannian Hessian and different retraction operators, Riemannian Newton methods, trust-region methods, adaptive regularized Newton method have been proposed in [85, 1, 2, 41]. When the second-order information is not available, the quasi-Newton type method becomes necessary. As in the Riemannian CG method, we need the vector transport operator to compare different tangent vectors from different tangent spaces. In addition, extra restrictions on the vector transport and the retraction are required for better convergence property or even convergence [74, 75, 78, 42, 44, 46, 43]. Non-vector-transport based quasi-Newton method is also explored in [38].

3.4.1. Riemannian trust-region method. One of the popular second-order algorithms is a Riemannian trust-region (RTR) algorithm [1, 2]. At the k -th iteration x_k , by utilizing the Taylor expansion on manifold, RTR constructs the following subproblem on the Tangent space:

$$(3.11) \quad \min_{\xi \in T_{x_k} \mathcal{M}} m_k(\xi) := \langle \text{grad } f(x_k), \xi \rangle_{x_k} + \frac{1}{2} \langle \text{Hess } f(x_k)[\xi], \xi \rangle_{x_k}, \quad \text{s.t. } \|\xi\|_{x_k} \leq \Delta_k,$$

where Δ_k is the trust-region radius. In [69], extensive methods for solving (3.11) are summarized. Among them, the Steihaug CG method, also named as truncated CG method, is most popular due to its good properties and relatively cheap computational cost. By solving this trust-region subproblem, we obtain a direction $\xi_k \in T_{x_k} \mathcal{M}$ satisfying the so-called Cauchy decrease. Then a trial point is computed as $z_k = R_{x_k}(\xi_k)$, where the step size is chosen as 1. To determine the acceptance of z_k , we compute the ratio between the actual reduction and the predicted reduction

$$(3.12) \quad \rho_k := \frac{f(x_k) - f(R_{x_k}(\xi_k))}{m_k(0) - m_k(\xi_k)}.$$

When ρ_k is greater than some given parameter $0 < \eta_1 < 1$, z_k is accepted. Otherwise, z_k is rejected. To avoid the algorithm stagnating at some feasible point and promote the efficiency as well, the trust-region radius is also updated based on ρ_k . The full algorithm is presented in Algorithm 2.

For the global convergence, the following assumptions are necessary for second-order type algorithms on manifold.

ASSUMPTION 4. (a). The function f is continuous differentiable and bounded from below on the level set $\{x \in \mathcal{M} : f(x) \leq f(x_0)\}$.

(b). There exists a constant $\beta_{\text{Hess}} > 0$ such that

$$\|\text{Hess } f(x_k)\| \leq \beta_{\text{Hess}}, \quad \forall k = 0, 1, 2, \dots$$

Algorithm 2 also requires a Lipschitz type continuous property on the objective function f [2, Definition 7.4.1].

Algorithm 2: Riemannian trust-region method

1 Input: Initial guess $x_0 \in \mathcal{M}$ and parameters $\bar{\Delta} > 0$, $\Delta_0 \in (0, \bar{\Delta})$, $\rho' \in [0, \frac{1}{4})$.
2 Output: Sequences of iterates $\{x_k\}$ and related information.
3 for $k = 0, 1, 2, \dots$ **do**
4 Use the truncated CG method to obtain ξ_k by solving (3.11).
5 Compute the ratio ρ_k in (3.12).
6 **if** $\rho_k < \frac{1}{4}$ **then** $\Delta_{k+1} = \frac{1}{4}\Delta_k$;
7 **else if** $\rho_k > \frac{3}{4}$ **and** $\|\xi_k\| = \Delta_k$ **then** $\Delta_{k+1} = \min(2\Delta_k, \bar{\Delta})$;
8 **else** $\Delta_{k+1} = \Delta_k$;
9 **if** $\rho_k > \rho'$ **then** $x_{k+1} = R_{x_k}(\xi_k)$;
10 **else** $x_{k+1} = x_k$;

785 ASSUMPTION 5. *There exists two constants $\beta_{RL} > 0$ and $\delta_{RL} > 0$ such that for*
 786 *all $x \in \mathcal{M}$ and $\xi \in T_x \mathcal{M}$ with $\|\xi\| = 1$,*

$$787 \quad \left| \frac{d}{dt} f \circ R_x(t\xi) \Big|_{t=\tau} - \frac{d}{dt} f \circ R_x(t\xi) \Big|_{t=0} \right| \leq \tau \beta_{RL}, \forall \tau \leq \delta_{RL}.$$

788 Then the global convergence to a stationary point [2, Theorem 7.4.2] is presented as
 789 follows.

790 THEOREM 6. *Let $\{x_k\}$ be a sequence generated by Algorithm 2. Suppose that*
 791 *Assumptions 4 and 5 holds, then*

$$792 \quad \liminf_{k \rightarrow \infty} \|\text{grad } f(x_k)\| = 0.$$

793 By further assuming the Lipschitz continuous property of the Riemannian gradient
 794 [2, Definition 7.4.3] and some isometric property of the Retraction operator R [2,
 795 Equation (7.25)], the convergence of the whole sequence is proved [2, Theorem 7.4.4].
 796 The locally superlinear convergence rate of RTR and its related assumptions can be
 797 found in [2, Section 7.4.2].

798 **3.4.2. Adaptive regularized Newton method.** From the perspective of Eu-
 799 clidean approximation, an adaptive regularized Newton algorithm (ARNT) is pro-
 800 posed for specific and general manifold optimization problems [92, 97, 41]. In the
 801 subproblem, the objective function is constructed by the second-order Taylor expan-
 802 sion in the Euclidean space and an extra regularization term, while the manifold
 803 constraint is kept. Specifically, the mathematical formulation is

$$804 \quad (3.13) \quad \min_{x \in \mathcal{M}} \quad \hat{m}_k(x) := \langle \nabla f(x), x - x_k \rangle + \frac{1}{2} \langle H_k[x - x_k], x - x_k \rangle + \frac{\sigma_k}{2} \|x - x_k\|^2,$$

805 where H_k is the Euclidean Hessian or its approximation. From the definition of
 806 Riemannian gradient and Hessian, we have

$$807 \quad (3.14) \quad \begin{aligned} \text{grad } \hat{m}_k(x_k) &= \text{grad } f(x_k) \\ \text{Hess } \hat{m}_k(x_k)[U] &= \mathbf{P}_{T_{x_k} \mathcal{M}}(H_k[U]) + \mathfrak{W}_{x_k}(U, \mathbf{P}_{T_{x_k} \mathcal{M}}^\perp(\nabla f(x_k))) + \sigma_k U, \end{aligned}$$

Algorithm 3: A modified CG method for solving subproblem (3.13)

```

1 Set  $T > 0$ ,  $\theta > 1$ ,  $\epsilon \geq 0$ ,  $\eta_0 = 0$ ,  $r_0 = \text{grad } m_k(x_k)$ ,  $p_0 = -r_0$ , and  $i = 0$ .
2 while  $i \leq n - 1$  do
3   Compute  $\pi_i = \langle p_i, \text{Hess } \hat{m}_k(x_k)[p_i] \rangle_{x_k}$ .
4   if  $\pi_i / \langle p_i, p_i \rangle_{x_k} \leq \epsilon$  then
5     if  $i = 0$  then set  $s_k = -p_0$ ,  $d_k = 0$ ;
6     else set  $s_k = \eta_i$ ,
7     if  $\pi_i / \langle p_i, p_i \rangle_{x_k} \leq -\epsilon$  then  $d_k = p_i$ , set  $\sigma_{est} = |\pi_i| / \langle p_i, p_i \rangle_{x_k}$ ;
8     else  $d_k = 0$ ;
9     break;
10  Set  $\alpha_i = \langle r_i, r_i \rangle_{x_k} / \pi_i$ ,  $\eta_{i+1} = \eta_i + \alpha_i p_i$ , and
     $r_{i+1} = r_i + \alpha_i \text{Hess } \hat{m}_k(x_k)[p_i]$ .
11  if  $\|r_{i+1}\|_{x_k} \leq \min\{\|r_0\|_{x_k}^\theta, T\}$  then
12    choose  $s_k = \eta_{i+1}$ ,  $d_k = 0$ ; break;
13  Set  $\beta_{i+1} = \langle r_{i+1}, r_{i+1} \rangle_{x_k} / \langle r_i, r_i \rangle_{x_k}$  and  $p_{i+1} = -r_{i+1} + \beta_{i+1} p_i$ .
14   $i \leftarrow i + 1$ .
15 Update  $\xi_k$  according to (3.15).

```

where $U \in T_{x_k} \mathcal{M}$, $\mathbf{P}_{T_{x_k} \mathcal{M}}^\perp := I - \mathbf{P}_{T_{x_k} \mathcal{M}}$ is the projection onto the normal space and the Weingarten map $\mathfrak{W}_x(\cdot, v)$ with $v \in T_{x_k} \mathcal{M}$ is a symmetric linear operator which is related to the second fundamental form of \mathcal{M} . To solve (3.13), a modified CG method is proposed in [41] to solve the Riemannian Newton equation at x_k ,

$$\text{grad } \hat{m}_k(x_k) + \text{Hess } \hat{m}_k(x_k)[\xi_k] = 0.$$

Since $\text{Hess } \hat{m}_k(x_k)$ may not be positive definite, CG may be terminated if a direction with negative curvature, says d_k , is encountered. Different from the truncated CG method used in RTR, a linear combination of s_k (the output of the truncated CG method) and the negative curvature direction d_k is used to construct a descent direction

$$(3.15) \quad \xi_k = \begin{cases} s_k + \tau_k d_k & \text{if } d_k \neq 0, \\ s_k & \text{if } d_k = 0, \end{cases} \quad \text{with} \quad \tau_k := \frac{\langle d_k, \text{grad } m_k(x_k) \rangle_{x_k}}{\langle d_k, \text{Hess } m_k(x_k)[d_k] \rangle_{x_k}}.$$

A detailed description on the modified CG method is presented in Algorithm 3. Then, Armijo search along ξ_k is adopted to obtain a trial point z_k . After obtaining z_k , we compute the following ratio between the actual reduction and the predicted reduction,

$$(3.16) \quad \rho_k = \frac{f(z_k) - f(x_k)}{m_k(z_k)}.$$

If $\rho_k \geq \eta_1 > 0$, then the iteration is successful and we set $x_{k+1} = z_k$; otherwise, the iteration is not successful and we set $x_{k+1} = x_k$, i.e.,

$$(3.17) \quad x_{k+1} = \begin{cases} z_k, & \text{if } \rho_k \geq \eta_1, \\ x_k, & \text{otherwise.} \end{cases}$$

Algorithm 4: An Adaptive Regularized Newton Method

- 1 Choose a feasible initial point $x_0 \in \mathcal{M}$ and an initial regularization parameter $\sigma_0 > 0$. Choose $0 < \eta_1 \leq \eta_2 < 1$, $0 < \gamma_0 < 1 < \gamma_1 \leq \gamma_2$. Set $k := 0$.
 - 2 **while** *stopping conditions not met* **do**
 - 3 Compute a new trial point z_k by doing Armijo search along ξ_k obtained by Algorithm 3.
 - 4 Compute the ratio ρ_k via (3.16).
 - 5 Update x_{k+1} from the trial point z_k based on (3.17).
 - 6 Update σ_k according to (3.18).
 - 7 $k \leftarrow k + 1$.
-

826 The regularization parameter σ_{k+1} is updated as follows

$$827 \quad (3.18) \quad \sigma_{k+1} \in \begin{cases} (0, \gamma_0 \sigma_k] & \text{if } \rho_k \geq \eta_2, \\ [\gamma_0 \sigma_k, \gamma_1 \sigma_k] & \text{if } \eta_1 \leq \rho_k < \eta_2, \\ [\gamma_1 \sigma_k, \gamma_2 \sigma_k] & \text{otherwise,} \end{cases}$$

828 where $0 < \eta_1 \leq \eta_2 < 1$ and $0 < \gamma_0 < 1 < \gamma_1 \leq \gamma_2$. These parameters determine how
829 aggressively the regularization parameter is adjusted when an iteration is successful
830 or unsuccessful. Putting these features together, we obtain Algorithm 4, which is
831 dubbed as ARNT.

832 We next present the convergence property of Algorithm 4 with the exact Euclidean
833 Hessian (i.e., $H_k = \nabla^2 f(x_k)$) starting from a few assumptions.

834 ASSUMPTION 7. Let $\{x_k\}$ be generated by Algorithm 4 the exact Euclidean Hes-
835 sian.

836 (A.1) The gradient ∇f is Lipschitz continuous on the convex hull of the manifold
837 \mathcal{M} – denoted by $\text{conv}(\mathcal{M})$, i.e., there exists $L_f > 0$ such that

$$838 \quad \|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\|, \quad \forall x, y \in \text{conv}(\mathcal{M}).$$

839 (A.2) There exists $\kappa_g > 0$ such that $\|\nabla f(x_k)\| \leq \kappa_g$ for all $k \in \mathbb{N}$.

840 (A.3) There exists $\kappa_H > 0$ such that $\|\nabla^2 f(x_k)\| \leq \kappa_H$ for all $k \in \mathbb{N}$.

841 (A.4) Suppose there exists $\underline{\omega} > 0$, $\overline{\omega} \geq 1$ such that $\underline{\omega}$ and $\overline{\omega}$

$$842 \quad \underline{\omega} \|\xi\|_2 \leq g_{x_k}(\xi, \xi) \leq \overline{\omega} \|\xi\|^2, \quad \xi \in T_{x_k} \mathcal{M},$$

843 for all $k \in \mathbb{N}$.

844 We note that the assumptions (A.2) and (A.4) hold if f is continuous differentiable
845 and the level set $\{x \in \mathcal{M} : f(x) \leq f(x_0)\}$ is compact.

846 The global convergence to an stationary point can be obtained.

847 THEOREM 8. Suppose that Assumptions 4 and 7 hold. Then, either

$$848 \quad \text{grad } f(x_\ell) = 0 \text{ for some } \ell \geq 0 \quad \text{or} \quad \liminf_{k \rightarrow \infty} \|\text{grad } f(x_k)\|_{x_k} = 0.$$

For the local convergence rate, we make the following assumptions.

ASSUMPTION 9. Let $\{x_k\}$ be generated by Algorithm 4.

(B.1) There exists $\beta_R, \delta_R > 0$ such that

$$\left\| \frac{D}{dt} \frac{d}{dt} R_x(t\xi) \right\|_x \leq \beta_R$$

for all $x \in \mathcal{M}$, all $\xi \in T_x \mathcal{M}$ with $\|\xi\|_x = 1$ and all $t < \delta_R$.

(B.2) The sequence $\{x_k\}$ converges to x_* .

(B.3) The Euclidean Hessian $\nabla^2 f$ is continuous on $\text{conv}(\mathcal{M})$.

(B.4) The Riemannian Hessian $\text{Hess } f$ is positive definite at x_* and the constant ϵ in Algorithm 3 is set to zero.

(B.5) H_k is a good approximation of the Euclidean Hessian $\nabla^2 f$, i.e., it holds

$$\|H_k - \nabla^2 f(x_k)\| \rightarrow 0, \quad \text{whenever} \quad \|\text{grad } f(x_k)\|_{x_k} \rightarrow 0.$$

Then we have the following results on the local convergence rate.

THEOREM 10. Suppose that the conditions (B.1)–(B.5) in Assumption 9 are satisfied. Then, the sequence $\{x_k\}$ converges q -superlinearly to x_* .

The detailed convergence analysis can be found in [41].

3.4.3. Quasi-Newton type methods.

When the Riemannian Hessian $\text{Hess } f(x)$ is computationally expensive or even not available, quasi-Newton-type methods turn out to be an attractive approach. In literatures [74, 75, 78, 42, 44, 46, 43], extensive variants of quasi-Newton methods are proposed. Here, we take the Riemannian Broyden-Fletcher-Goldfarb-Shanno (BFGS) as an example to show the general idea of quasi-Newton methods on Riemannian manifold. Similar to the quasi-Newton method in the Euclidean space, an approximation \mathcal{B}_{k+1} should satisfy the following secant equation

$$\mathcal{B}_{k+1} s_k = y_k,$$

where $s_k = \mathcal{T}_{S_{\alpha_k \xi_k}} \alpha_k \xi_k$ and $y_k = \beta_k^{-1} \text{grad } f(x_{k+1}) - \mathcal{T}_{S_{\alpha_k \xi_k}} \text{grad } f(x_k)$ with parameter β_k . Here, α_k and ξ_k is the step size and the direction used in the k -th iteration. $\mathcal{T}_{S_{\alpha_k \xi_k}}$ is an isometric vector transport operator associated with the retraction R , i.e.,

$$\langle \mathcal{T}_{S_{\xi_x}} u_x, \mathcal{T}_{S_{\xi_x}} v_x \rangle_{R_x(\xi_x)} = \langle u_x, v_x \rangle_x.$$

Additionally, \mathcal{T}_S should satisfy the following locking condition,

$$\mathcal{T}_{S_{\xi_k}} \xi_k = \beta_k \mathcal{T}_{R_{\xi_k}} \xi_k, \quad \beta_k = \frac{\|\xi_k\|_{x_k}}{\|\mathcal{T}_{R_{\xi_k}} \xi_k\|_{R_{x_k}(\xi_k)}},$$

where $\mathcal{T}_{R_{\xi_k}} \xi_k = \frac{d}{dt} R_{x_k}(t\xi_k) |_{t=1}$. Then, the scheme of the Riemannian BFGS is

$$(3.19) \quad \mathcal{B}_{k+1} = \hat{\mathcal{B}}_k - \frac{\hat{\mathcal{B}}_k s_k (\hat{\mathcal{B}}_k s_k)^{\flat}}{(\hat{\mathcal{B}}_k s_k)^{\flat} s_k} + \frac{y_k y_k^{\flat}}{y_k^{\flat} s_k},$$

where $\hat{\mathcal{B}}_k = \mathcal{T}_{S_{\alpha_k \xi_k}} \alpha_k \xi_k \circ \mathcal{B}_k \circ (\mathcal{T}_{S_{\alpha_k \xi_k}} \alpha_k \xi_k)^{-1}$ is from $T_{x_{k+1}} \mathcal{M}$ to $T_{x_{k+1}} \mathcal{M}$. With this choice of β_k and the isometric property of \mathcal{T}_S , we can guarantee the positive

Algorithm 5: Riemannian BFGS method

1 Input: Initial guess $x_0 \in \mathcal{M}$, isometric vector transport \mathcal{T}_S associated with the retraction R , initial Riemannian Hessian approximation $\mathcal{B}_0 : T_{x_0}\mathcal{M} \rightarrow T_{x_0}\mathcal{M}$, which is symmetric positive definite, Wolfe condition parameters $0 < c_1 < \frac{1}{2} < c_2 < 1$.
2 for $k = 0, 1, 2, \dots$ **do**
3 Solve $\mathcal{B}_k \xi_k = -\text{grad } f(x_k)$ to get ξ_k .
4 Obtain x_{k+1} by doing a Wolfe search along ξ_k , i.e., finding $\alpha_k > 0$ such that the following two conditions are satisfied

$$f(R_{x_k}(\alpha_k \xi_k)) \leq f(x_k) + c_1 \alpha_k \langle \text{grad } f(x_k), \xi_k \rangle_{x_k},$$

$$\frac{d}{dt} f(R_{x_k}(t \xi_k)) \big|_{t=\alpha_k} \geq c_2 \frac{d}{dt} f(R_{x_k}(t \xi_k)) \big|_{t=0}.$$

5 Set $x_k = R_{x_k}(\alpha_k \xi_k)$.
6 Update \mathcal{B}_{k+1} by (3.19).

definiteness of \mathcal{B}_{k+1} . After obtaining the new approximation \mathcal{B}_{k+1} , the Riemannian BFGS method solves the following linear system

$$\mathcal{B}_{k+1} \xi_{k+1} = -\text{grad } f(x_{k+1})$$

to get ξ_k . The detailed algorithm is presented in Algorithm 5. The choice of $\beta_k = 1$ can also guarantee the convergence but with more strict assumptions. One can refer to [46] for the convergence analysis. Since the computation of differentiated retraction may be costly, authors in [43] investigate another way to preserve the positive definiteness of the BFGS scheme. Meanwhile, the Wolfe search is replaced by the Armijo search. As a result, the differentiated retraction can be avoided and the convergence analysis is presented as well.

The aforementioned quasi-Newton methods rely on the vector transport operator. When the vector transport operation is computationally costly, these methods may be less competitive. Noticing the structure of the Riemannian Hessian $\text{Hess } f(x_k)$, i.e.,

$$\text{Hess } f(x_k)[U] = \mathbf{P}_{T_{x_k}\mathcal{M}}(\nabla^2 f(x_k)[U]) + \mathfrak{W}_{x_k}(U, \mathbf{P}_{T_{x_k}\mathcal{M}}^\perp(\nabla f(x_k))), \quad U \in T_{x_k}\mathcal{M},$$

where the second term $\mathfrak{W}_{x_k}(U, \mathbf{P}_{T_{x_k}\mathcal{M}}^\perp(\nabla f(x_k)))$ is often much cheaper than the first term $\mathbf{P}_{T_{x_k}\mathcal{M}}(\nabla^2 f(x_k)[U])$. Similar to the quasi-Newton methods in unconstrained nonlinear least square problems [52] [83, Chapter 7], we can focus on the construction of an approximation of the Euclidean Hessian $\nabla^2 f(x_k)$ and use exact formulations of remaining parts. Furthermore, if the Euclidean Hessian itself consists of cheap and expensive parts, i.e.,

$$(3.20) \quad \nabla^2 f(x_k) = \mathcal{H}^c(x_k) + \mathcal{H}^e(x_k),$$

where the computational cost of $\mathcal{H}^e(x_k)$ is much more expensive than $\mathcal{H}^c(x_k)$, an

approximation of $\nabla^2 f(x_k)$ is constructed as

$$(3.21) \quad H_k = \mathcal{H}^c(x_k) + C_k,$$

where C_k is an approximation of $\mathcal{H}^e(x_k)$ obtained by a quasi-Newton method in the ambient Euclidean space. If an objective function f is not equipped with the structure (3.20), H_k is a quasi-Newton approximation of $\nabla^2 f(x_k)$. In the construction of the quasi-Newton approximation, a Nyström approximation technique [38, Section 2.3] is explored, which turns to be a better choice than the BB type initialization [69, Chapter 6]. Since the quasi-Newton approximation is constructed in the ambient Euclidean space, the vector transport is not necessary. Then, subproblem (3.13) is constructed with H_k . From the expression of the Riemannian Hessian $\text{Hess } \hat{m}_k$ in (3.14), we see that subproblem (3.13) gives us a way to approximate the Riemannian Hessian when an approximation H_k to the Euclidean Hessian is available. The same procedures of ARNT can be utilized for (3.13) with the approximate Euclidean Hessian H_k . An adaptive structured quasi-Newton method given in [38] is presented in Algorithm 6.

Algorithm 6: A structured quasi-Newton method

- 1 Input an initial guess $X^0 \in \mathcal{M}$. Choose $\tau_0 > 0$, $0 < \eta_1 \leq \eta_2 < 1$, $1 < \gamma_1 \leq \gamma_2$. Set $k = 0$.
 - 2 **while** *stopping conditions not met* **do**
 - 3 Check the structure of $\nabla^2 f(x_k)$ to see if it can be written in a form as (3.20).
 - 4 Construct an approximation H_k by utilizing a quasi-Newton method.
 - 5 Construct and solve the subproblem (3.13) (by using the modified CG method or the Riemannian gradient type method) to obtain a new trial point z_k .
 - 6 Compute the ratio ρ_k via (3.12).
 - 7 Update x_{k+1} from the trial point z_k based on (3.17).
 - 8 Update τ_k according to (3.18).
 - 9 $k \leftarrow k + 1$.
-

To explain the differences between the two quasi-Newton algorithms more straightforwardly, we take the HF total energy minimization problem (2.10) as an example. From the calculation in [38], we have the Euclidean gradients

$$\nabla E_{\text{ks}}(X) = H_{\text{ks}}(X)X, \quad \nabla E_{\text{hf}}(X) = H_{\text{hf}}(X)X,$$

where $H_{\text{ks}}(X) := \frac{1}{2}L + V_{\text{ion}} + \sum_l \zeta_l w_l w_l^* + \text{Diag}((\Re L^\dagger)\rho) + \text{Diag}(\mu_{\text{xc}}(\rho)^* e)$ and $H_{\text{hf}}(X) = H_{\text{ks}}(X) + \mathcal{V}(XX^*)$. The Euclidean Hessian of E_{ks} and E_{f} along a matrix $U \in \mathbb{C}^{n \times p}$ are

$$\begin{aligned} \nabla^2 E_{\text{ks}}(X)[U] &= H_{\text{ks}}(X)U + \text{Diag} \left((\Re L^\dagger + \frac{\partial^2 \epsilon_{\text{xc}}}{\partial \rho^2} e)(\bar{X} \odot U + X \odot \bar{U})e \right) X, \\ \nabla^2 E_{\text{f}}(X)[U] &= \mathcal{V}(XX^*)U + \mathcal{V}(XU^* + UX^*)X. \end{aligned}$$

Since $\nabla^2 E_f(X)$ is significantly more expensive than $\nabla^2 E_{\text{ks}}(X)$, we only need to approximate $\nabla^2 E_f(X)$. The differences $X_k - X_{k-1}$, $\nabla E_f(X_k) - \nabla E_f(X_{k-1})$ are computed. Then, a quasi-Newton approximation C_k of $\nabla^2 E_f$ is obtained without requiring vector transport. By adding the exact formulation of $\nabla^2 E_{\text{ks}}(X_k)$, we have an approximation H_k , i.e.,

$$H_k = \nabla^2 E_{\text{ks}} + C_k.$$

A Nyström approximation for C_k is also investigated. Note that the spectrum of $\nabla^2 E_{\text{ks}}(X)$ dominates the spectrum of $\nabla^2 E_f(X)$. The structured approximation H_k is more reliable than a direct quasi-Newton approximate to $\nabla^2 E_{\text{hf}}(X)$ because the spectrum of $\nabla^2 E_{\text{ks}}$ is inherited from the exact form. The Remaining procedure is to solve subproblem (3.13) to update X_k .

3.5. Stochastic algorithms. For problems arising from machine learning, the objective function f is often a summation of a finite number of functions $f_i, i = 1, \dots, m$

$$f(x) = \sum_{i=1}^m f_i(x).$$

For unconstrained situations, there are many efficient algorithms, such as Adam, Adagrad, RMSProp, Adelta, SVRG, etc. One can refer to [58]. For the case with manifold constraints, combining with retraction operators and vector transport operator, these algorithms can be well generalized. However, in the implementation, due to the considerations of the computational costs of different parts, they may have different versions. Riemannian stochastic gradient method is first developed in [14]. Later, a class of first-order methods and their accelerations are investigated for geodesically convex optimization in [105, 66]. With the help of parallel translation or vector transport, Riemannian SVRG methods are generalized in [104, 76]. In consideration of the computational cost of the vector transport, non-vector transport based Riemannian SVRG is proposed in [50]. Since an intrinsic coordinate system is absent, the coordinate-wise update on manifold should be further investigated. A compromised approach for Riemannian adaptive optimization methods on product manifolds are presented in [10].

Here, the SVRG algorithm [50] is taken as an example. At the current point $X^{s,k}$, we first calculate the full gradient $\mathcal{G}(X^{s,k})$, then randomly sample a subscript from 1 to m and use this to construct a stochastic gradient with reduced variance as $\mathcal{G}(X^{s,k}, \xi_{s,k}) = \nabla f(X^{s,0}) + (\nabla f_{i_{s,k}}(X^{s,k}) - \nabla f_{i_{s,k}}(X^{s,0}))$, finally move along this direction with a given step size to next iteration point

$$X^{s,k+1} = X^{s,k} - \tau_s \xi_{s,k}.$$

For Riemannian SVRG [50], after obtaining the stochastic gradient with reduced Euclidean variance, it first projects this gradient to the tangent space

$$\xi_{s,k} = \mathbf{P}_{T_{X^{s,k}} \mathcal{M}}(\mathcal{G}(X^{s,k})).$$

Then, the following retraction step

$$X^{s,k+1} = R_{X^{s,k}}(-\tau_s \xi_{s,k}),$$

Algorithm 7: Riemannian SVRG [50]

```
1 for  $s = 0, \dots, S - 1$  do
2   calculates the full gradient  $\nabla f(X^{s,0})$  and sets the step size  $\tau_s > 0$ .
3   for  $k = 0, \dots, K - 1$  do
4     Randomly substitute samples get the subscript  $i_{s,k} \subseteq \{1, \dots, m\}$ .
      Calculate a random Euclidean gradient  $\mathcal{G}(X^{s,k})$ 
      
$$\mathcal{G}(X^{s,k}, \xi_{s,k}) = \nabla f(X^{s,0}) + (\nabla f_{i_{s,k}}(X^{s,k}) - \nabla f_{i_{s,k}}(X^{s,0})).$$

      Calculate a random Riemann gradient
      
$$\xi_{s,k} = \mathbf{P}_{T_{X^{s,k}}\mathcal{M}}(\mathcal{G}(X^{s,k})).$$

      Update  $X^{s,k+1}$  in the following format
      
$$X^{s,k+1} = R_{X^{s,k}}(-\tau_s \xi_{s,k}).$$

5   Take  $X^{s+1,0} \leftarrow X^{s,K}$ .
```

968 is executed to get the next feasible point. The detailed version is outlined in Algorithm
969 7.

970 **3.6. Algorithms for Riemannian non-smooth optimization.** As shown in
971 subsections 2.11 to 2.15, many practical problems are with non-smooth objective
972 function and manifold constraints, i.e.,

$$973 \quad \min_{x \in \mathcal{M}} f(x) := g(x) + h(x),$$

974 where g is smooth and h is non-smooth. Riemannian subgradient methods [29, 15]
975 are firstly investigated to solve this kind of problems and their convergence analysis is
976 exhibited in [36] with the help of Kurdyka-Łojasiewicz (KL) inequalities. For locally
977 Lipschitz functions on Riemannian manifolds, a gradient sampling method and a non-
978 smooth Riemannian trust-region method are proposed in [35, 37]. Proximal gradient
979 methods on manifold are presented in [5, 28], where the inner subproblem is solved
980 inexactly by subgradient type methods. The corresponding complexity analysis is
981 given in [11, 12]. Different from the constructions of the subproblem in [5, 28], a more
982 tractable subproblem without manifold constraints is investigated in [25] for convex
983 $h(x)$. By utilizing the semi-smooth Newton method [98], the proposed proximal gra-
984 dient method on manifold enjoys a faster convergence. Another class of methods is
985 based on operator-splitting techniques. Some variants of the alternating direction
986 method of multipliers (ADMM) are studied in [56, 53, 90, 107, 13, 60].

987 We briefly introduce the proximal gradient method on manifold [25] here. Assume
988 that the convex function h is Lipschitz continuous. At each iteration x_k , the following

989 subproblem is constructed

$$\begin{aligned}
990 \quad (3.22) \quad & \min_d \quad \langle \text{grad} g(x_k), d \rangle + \frac{1}{2t} \|d\|_F^2 + h(x_k + d) \\
& \text{s.t.} \quad d \in T_{x_k} \mathcal{M},
\end{aligned}$$

991 where $t > 0$ is a step size. Given a retraction R , problem (3.22) can be seen as a first-
992 order approximation of $f(R_{x_k}(d))$ near the zero element 0_{x_k} on $T_{x_k} \mathcal{M}$. Specifically, it
993 follows from the definition of the Riemannian gradient that $\text{grad} g(x_k) = \nabla g(R_{x_k}(0))$.
994 From the Lipschitz continuous property of h and the definition of R , we have

$$995 \quad |h(R_{x_k}(d)) - h(x_k + d)| \leq L_h \|R_{x_k}(d) - (x_k + d)\|_F = O(\|d\|_F^2),$$

996 where L_h is the Lipschitz constant of h . Therefore, we conclude

$$997 \quad f(R_{x_k}(d)) = \langle \text{grad} g(x_k), d \rangle + h(x_k + d) + O(\|d\|_F^2), \quad d \rightarrow 0.$$

998 Then the next step is to solve (3.22). Since (3.22) is convex and with linear constraints,
999 the KKT conditions are sufficient and necessary for the global optimality. Specifically,
1000 we have

$$1001 \quad d(\lambda) = \text{prox}_{th}(b(\lambda)) - x_k, \quad b(\lambda) = x_k - t(\text{grad} f(x_k) - \mathcal{A}_k^*(\lambda)), \quad \mathcal{A}_k(d(\lambda)) = 0,$$

1002 where $d \in T_{x_k} \mathcal{M}$ is represented by $\mathcal{A}_k(d) = 0$ with a linear operator \mathcal{A}_k , \mathcal{A}_k^* is
1003 the adjoint operator of \mathcal{A}_k . Define $E(\lambda) := \mathcal{A}_k(d(\lambda))$, it is proved in [25] that E is
1004 monotone and then the semi-smooth Newton method in [98] is utilized to solve the
1005 nonlinear equation $E(\lambda) = 0$ to obtain a direction d_k . Combining with a curvilinear
1006 search along d_k with R_{x_k} , the decrease on f is guaranteed and the global convergence
1007 is established.

1008 **3.7. Complexity Analysis.** The complexity analysis of the Riemannian gradi-
1009 ent method and the Riemannian trust region method has been studied in [16]. Similar
1010 to the Euclidean unconstrained optimization, the Riemannian gradient method (us-
1011 ing a fixed step size or Armijo curvilinear search) converges to $\|\text{grad} f(x)\| \leq \varepsilon$ up to
1012 $O(1/\varepsilon^2)$ steps. Under mild assumptions, a modified Riemannian trust region method
1013 converges to $\|\text{grad} f(x)\| \leq \varepsilon$, $\text{Hess} f(x) \succeq -\sqrt{\varepsilon}I$ at most $O(\max\{1/\varepsilon^{1.5}, 1/\varepsilon^{2.5}\})$
1014 iterations. For objective functions with multi-block convex but non-smooth terms, an
1015 ADMM of complexity of $O(1/\varepsilon^4)$ is proposed in [107]. For the cubic regularization
1016 methods on the Riemannian manifold, recent studies [109, 4] show a convergence to
1017 $\|\text{grad} f(x)\| \leq \varepsilon$, $\text{Hess} f(x) \succeq -\sqrt{\varepsilon}I$ with complexity of $O(1/\varepsilon^{1.5})$.

1018 4. Analysis for manifold optimization.

1019 **4.1. Geodesic convexity.** For a convex function in the Euclidean space, any
1020 local minima is also a global minima. An interesting extension is the geodesic convex-
1021 ity of functions. Specifically, a function defined on manifold is said to be geodesically
1022 convex if it is convex along any geodesic. Similarly, a local minima of a geodesically
1023 convex function on manifold is also a global minima. Naturally, a question is how to
1024 distinguish the geodesically convex function.

DEFINITION 11. Given a Riemannian manifold (\mathcal{M}, g) , a set $\mathcal{K} \subset \mathcal{M}$ is called g -fully geodesic, if for any $p, q \in \mathcal{K}$, any geodesic γ_{pq} is located entirely in \mathcal{K} .

For example, $D_c := \{P \in \mathbb{S}_{++}^n \mid \det(P) = c\}$ with a positive constant c is not a convex set in $\mathbb{R}^{n \times n}$, but is a fully geodesic set of Riemannian manifolds (\mathbb{S}_{++}^n, g) , where the Riemannian metric g at P is $g_P(U, V) := \text{tr}(P^{-1}UP^{-1}V)$. Now we present the definition of the g -geodesically convex function.

DEFINITION 12. Given a Riemannian manifold (\mathcal{M}, g) and a g -fully geodesic set $\mathcal{K} \subset \mathcal{M}$, a function $f : \mathcal{K} \rightarrow \mathbb{R}$ is g -geodesically convex if for any $p, q \in \mathcal{K}$ and any geodesic $\gamma_{pq} : [0, 1] \rightarrow \mathcal{K}$ connecting p, q , it holds:

$$f(\gamma_{pq}(t)) \leq (1-t)f(p) + tf(q), \forall t \in [0, 1].$$

A g -fully geodesically convex function may not be convex. For example, $f(x) := (\log x)^2$, $x \in \mathbb{R}_+$ is not convex in the Euclidean space, but is convex with respect to the manifold (\mathbb{R}_+, g) , where $g_x(u, v) := ux^{-1}v$.

Therefore, for a specific function, it is of significant importance to define a proper Riemannian metric to recognize the geodesic convexity. A natural problem is, given a manifold \mathcal{M} and a smooth function $f : \mathcal{M} \rightarrow \mathbb{R}$, whether there is a metric g such that f is geodesic convex with respect to g ? It is generally not easy to prove the existence of such a metric. From the definition of the geodesic convexity, we know that if a function has a non-global local minimum, then this function is not geodesically convex for any metric. For more information on geodesic convexity, we refer to [88].

4.2. Convergence of self-consistent field iterations. In [62, 63], several classical theoretical problems from KSDFT are studied. Under certain conditions, the equivalence between KS energy minimization problems and KS equations are established. In addition, a lower bound of non-zero elements of the charge density is also analyzed. By treating the KS equation as a fixed point equation with respect to a potential function, the Jacobian matrix is explicitly derived using the spectral operator theory and the theoretical properties of the SCF method are analyzed. It is proved that the second-order derivatives of the exchange-correlation energy are uniformly bounded if the Hamiltonian has a sufficiently large eigenvalue gap. Moreover, SCF converges from any initial point and enjoys a local linear convergence rate. Related results can be found in [27, 114, 22, 113, 6, 110].

Specifically, for the KS equation (2.11), we define the potential function

$$(4.1) \quad V := \mathbb{V}(\rho) = L^\dagger \rho + \mu_{xc}(\rho)^\top e$$

and

$$(4.2) \quad H(V) := \frac{1}{2}L + \sum_l \zeta_l w_l w_l^* + V_{ion} + \text{diag}(V).$$

Then, we have $H_{\text{ks}}(\rho) = H(V(\rho))$. From (2.11), X are the eigenvectors corresponding to the p -smallest eigenvalues of $H(V)$, which is dependent on V . Then, a fixed point mapping for V can be written as

$$(4.3) \quad V = \mathbb{V}(F_\phi(V)),$$

1064 where $F_\phi(V) = \text{diag}(X(V)X(V)^\top)$. Therefore, each iteration of SCF is to update V_k
 1065 as

$$1066 \quad (4.4) \quad V_{k+1} = \mathbb{V}(F_\phi(V_k)).$$

1067 For SCF with a simple charge-mixing strategy, the update scheme can be written as

$$1068 \quad (4.5) \quad V_{k+1} = V_k - \alpha(V_k - \mathbb{V}(F_\phi(V_k))),$$

1069 where α is an appropriate step size. Under some mild assumptions, SCF converges
 1070 with a local linear convergence rate.

1071 **THEOREM 13.** *Suppose that $\lambda_{p+1}(H(V)) - \lambda_p(H(V)) > \delta$, $\forall V$, the second or-*
 1072 *der derivatives of ϵ_{xc} are upper bounded and there is a constant θ such that $\|L^\dagger +$*
 1073 *$\frac{\partial \mu_{\text{xc}}(\rho)}{\partial \rho} e\|_2 \leq \theta$, $\forall \rho \in \mathbb{R}^n$. Let $b_1 := 1 - \frac{\theta}{\delta} > 0$, $\{V_k\}$ be a sequence generated by (4.5)*
 1074 *with a step size of α satisfying*

$$1075 \quad 0 < \alpha < \frac{2}{2 - b_1}.$$

1076 *Then, $\{V^i\}$ converges to a solution of the KS equation (2.11), and its convergence*
 1077 *rate is not worse than $|1 - \alpha| + \alpha(1 - b_1)$.*

1078 **4.3. Pursuing global optimality.** In the Euclidean space, a common way to
 1079 escape the local minimum is to add white noise to the gradient flow, which leads to a
 1080 stochastic differential equation

$$1081 \quad dX(t) = -\nabla f(X(t))dt + \sigma(t)dB(t),$$

1082 where $B(t)$ is the standard n -by- p Brownian motion. A generalized noisy gradient
 1083 flow on the Stiefel manifold is investigated in [101]

$$1084 \quad dX(t) = -\text{grad } f(X(t))dt + \sigma(t) \circ dB_{\mathcal{M}}(t),$$

1085 where $B_{\mathcal{M}}(t)$ is the Brownian motion on the manifold $\mathcal{M} := \text{St}(n, p)$. The construc-
 1086 tion of a Brownian motion is then given in an extrinsic form. Theoretically, it can
 1087 converge to the global minima by assuming second-order continuity.

4.4. Community detection. For community detection problems, a commonly
 used model is called the degree-correlated stochastic block model (DCSBM). It as-
 sumes that there are no overlaps between nodes in different communities. Specifically,
 the hypothesis node set $[n] = \{1, \dots, n\}$ contains k communities, $\{C_1^*, \dots, C_k^*\}$ sat-
 isfying

$$C_a^* \cap C_b^* = \emptyset, \forall a \neq b \text{ and } \cup_{a=1}^k C_a^* = [n].$$

In DCSBM, the network is a random graph, which can be represented by a matrix
 with all elements 0 to 1 represented by $B \in \mathbb{S}^k$. Let $A \in \{0, 1\}^{n \times n}$ be the adjacency
 matrix of this network and $A_{ii} = 0, \forall i \in [n]$. Then for $i \in C_a^*, j \in C_b^*, i \neq j$,

$$A_{ij} = \begin{cases} 1, & \text{with probability } B_{ab}\theta_i\theta_j, \\ 0, & \text{with probability } 1 - B_{ab}\theta_i\theta_j, \end{cases}$$

1088 where the heterogeneity of nodes is characterized by the vector θ . More specifically,
 1089 larger θ_i corresponds to i with more edges connecting other nodes. For DCSBM, the
 1090 aforementioned relaxation model (2.15) is proposed in [106]. By solving (2.15), an
 1091 approximation of the global optimal solution can be obtained with high probability.

THEOREM 14. Define $G_a = \sum_{i \in C_a^*} \theta_i$, $H_a = \sum_{b=1}^k B_{ab} G_b$, $F_i = H_a \theta_i$. Let U^* and Φ^* be global optimal solutions for (2.15) and (2.14), respectively and define $\Delta = U^*(U^*)^\top - \Phi^*(\Phi^*)^\top$. Suppose that $\max_{1 \leq a < b \leq k} \frac{B_{ab} + \delta}{H_a H_b} < \lambda < \min_{1 \leq a \leq k} \frac{B_{aa} - \delta}{H_a^2}$ (where $\delta > 0$). Then, with high probability, we have

$$\|\Delta\|_{1,\theta} \leq \frac{C_0}{\delta} \left(1 + \left(\max_{1 \leq a \leq k} \frac{B_{Aa}}{H_a^2} \|f\|_1 \right) \right) (\sqrt{n\|f\|_1} + n).$$

1092 **4.5. Max cut.** Consider the SDP relaxation (2.2) and the non-convex relaxation
 1093 problem with low rank constraints (2.3). If $p \geq \sqrt{2n}$, the composition of a solution
 1094 V^* of (2.3), i.e., $V^*(V^*)^\top$, is always an optimal solution of SDP (2.2) [9, 71, 19].
 1095 If $p \geq \sqrt{2n}$, for almost all matrices C , problem (2.3) has a unique local minimum
 1096 and this minimum is also a global minimum of the original problem (2.1) [17]. The
 1097 relationship between solutions of the two problems (2.2) and (2.3) is presented in [67].
 1098 Define $\text{SDP}(C) = \max\{\langle C, X \rangle : X \succeq 0, X_{ii} = 1, i \in [n]\}$. A point $V \in \text{Ob}(p, n)$ is
 1099 called an ε -approximate concave point of (2.3), if

$$\langle U, \text{Hess } f(V)[U] \rangle \leq \varepsilon \|u\|^2, \quad \forall U \in T_V \text{Ob}(p, n).$$

1101 The following theorem tells the approximation quality of an ε -approximate concave
 1102 point of (2.3).

1103 THEOREM 15. For any ε -approximate concave point V of (2.3), we have

$$1104 \quad (4.6) \quad \text{tr}(CV^\top V) \geq \text{SDP}(C) - \frac{1}{p-1}(\text{SDP}(C) + \text{SDP}(-C)) - \frac{n}{2}\varepsilon.$$

1106 Another problem with similar applications is the \mathbb{Z}_2 synchronization problem [7].
 1107 Specifically, given noisy observations $Y_{ij} = z_i z_j + \sigma W_{ij}$, where $W_{i>j} \sim \mathcal{N}(0, 1)$ and
 1108 $W_{ij} = W_{ji}$, $W_{ii} = 0$, we want to estimate the unknown labels $z_i \in \{\pm 1\}$. It can be
 1109 seen as a special case of the max cut problem with $p = 2$. The following results are
 1110 presented in [7].

1111 THEOREM 16. If $\sigma < \frac{1}{8}\sqrt{n}$, then, with a high probability, all second-order stable
 1112 points Q of problem (2.3) ($p = 2$) have the following non-trivial relationship with the
 1113 true label z , i.e., for each such σ , there is ε such that

$$1114 \quad \frac{1}{n} \|Q^\top z\|_2 \geq \varepsilon.$$

1115 **4.6. Burer-Monteiro factorizations of smooth semidefinite programs.**

1116 Consider the following SDP

$$1117 \quad (4.7) \quad \min_{X \in \mathbb{S}^{n \times n}} \text{tr}(CX) \quad \text{s.t. } \mathcal{A}(X) = b, X \succeq 0,$$

1118 where $C \in \mathbb{S}^{n \times n}$ is a cost matrix, $\mathcal{A} : \mathbb{S}^{n \times n} \rightarrow \mathbb{R}^m$ is a linear operator and $\mathcal{A}(X) = b$
 1119 leads to m equality constraints on X , i.e., $\text{tr}(A_i X) = b_i$ with $A_i \in \mathbb{S}^{n \times n}$, $b \in \mathbb{R}^m$, $i =$

1120 $1, \dots, m$. Define \mathcal{C} as the constraint set

$$1121 \quad \mathcal{C} = \{X \in \mathbb{S}^{n \times n} : \mathcal{A}(X) = b, X \succeq 0\}.$$

1122 If \mathcal{C} is compact, it is proved in [9, 71] that (4.7) has a global minimum of rank r with
 1123 $\frac{r(r+1)}{2} \leq m$. This allows to use the Burer-Monteiro factorizations [19] (i.e., let $X =$
 1124 YY^\top with $Y \in \mathbb{R}^{n \times p}$, $\frac{p(p+1)}{2} \geq m$) to solve the following non-convex optimization
 1125 problem

$$1126 \quad (4.8) \quad \min_{Y \in \mathbb{R}^{n \times p}} \text{tr}(CYY^\top) \quad \text{s.t. } \mathcal{A}(YY^\top) = b.$$

1127 Here, we define the constraint set

$$1128 \quad (4.9) \quad \mathcal{M} = \mathcal{M}_p = \{Y \in \mathbb{R}^{n \times p} : \mathcal{A}(YY^\top) = b\}.$$

1129 Since \mathcal{M} is non-convex, there may exist many non-global local minima of (4.8). It
 1130 is claimed in [20] that each local minimum of (4.8) maps to a global minimum of
 1131 (4.7) if $\frac{p(p+1)}{2} > m$. By utilizing the optimality theory of manifold optimization, any
 1132 second-order stationary point can be mapped to a global minimum of (4.7) under mild
 1133 assumptions [18]. Note that (4.9) is generally not a manifold. When the dimension
 1134 of the space spanned by $\{A_1 Y, \dots, A_m Y\}$, denoted by $\text{rank } \mathcal{A}$, is fixed for all Y , \mathcal{M}_p
 1135 defines a Riemannian manifold. Hence, we need the following assumptions.

1136 **ASSUMPTION 17.** *For a given p such that \mathcal{M}_p is not empty, assume at least one*
 1137 *of the following conditions are satisfied.*

1138 (SDP.1) $\{A_1 Y, \dots, A_m Y\}$ are linearly independent in $\mathbb{R}^{n \times p}$ for all $Y \in \mathcal{M}_p$

1139 (SDP.2) $\{A_1 Y, \dots, A_m Y\}$ span a subspace of constant dimension in $\mathbb{R}^{n \times p}$ for all Y in
 1140 an open neighborhood of $\mathcal{M}_p \in \mathbb{R}^{n \times p}$.

1141 By comparing the optimality conditions of (4.8) and the KKT conditions of (4.7), the
 1142 following equivalence between (4.7) and (4.8) is established in [18, Theorem 1.4].

1143 **THEOREM 18.** *Let p be such that $\frac{p(p+1)}{2} > \text{rank } \mathcal{A}$. Suppose that Assumption*
 1144 *17 holds. For almost any cost matrix $C \in \mathbb{S}^{n \times n}$, if $Y \in \mathcal{M}_p$ satisfies first- and*
 1145 *second-order necessary optimality conditions for (4.8), then Y is globally optimal and*
 1146 *$X = YY^\top$ is globally optimal for (4.7).*

1147 **4.7. Little Grothendieck problem with orthogonality constraints.** Given
 1148 a positive semidefinite matrix $C \in \mathbb{R}^{dn \times dn}$, the little Grothendieck problem with
 1149 orthogonality constraints can be expressed as

$$1150 \quad (4.10) \quad \max_{O_1, \dots, O_d \in \mathcal{O}_d} \sum_{i=1}^n \sum_{j=1}^n \text{tr}(C_{ij}^\top O_i O_j^\top),$$

1151 where C_{ij} represents the (i, j) -th $d \times d$ block of C , \mathcal{O}_d is a group of $d \times d$ orthogonal
 1152 matrices (i.e., $O \in \mathcal{O}_d$ if and only if $O^\top O = O O^\top = I$.) A SDP relaxation of (4.10)
 1153 is as follows [8]

$$1154 \quad (4.11) \quad \max_{\substack{G \in \mathbb{R}^{dn \times dn} \\ G_{ii} = I_{d \times d}, G \succeq 0}} \text{tr}(CG).$$

1155 For the original problem (4.10), a randomized approximation algorithm is presented
 1156 in [8]. Specifically, it consists of the following two procedures.

- Let G be a solution to problem (4.11). Denote by the Cholesky decomposition $G = LL^\top$. Let X_i be a $d \times (nd)$ matrix such that $L = (X_1^\top, X_2^\top, \dots, X_n^\top)^\top$.
- Let $R \in \mathbb{R}^{(nd) \times d}$ be a real-valued Gaussian random matrix whose entries are i.i.d. $\mathcal{N}(0, \frac{1}{d})$. The approximate solution of the problem (4.10) can be calculated as follows

$$V_i = \mathcal{P}(X_i R),$$

where $\mathcal{P}(Y) = \arg \min_{Z \in \mathcal{O}_d} \|Z - Y\|_F$ with $Y \in \mathbb{R}^{d \times d}$.

For the solution obtained in the above way, a constant approximation ratio on the objective function value is shown, which recovers the known $\frac{2}{\pi}$ approximation guarantee for the classical little Grothendieck problem.

THEOREM 19. *Given a symmetric matrix $C \succeq 0$. Let $V_1, \dots, V_n \in \mathcal{O}_d$ be obtained as above. Then*

$$\mathbf{E} \left[\sum_{i=1}^n \sum_{j=1}^n \text{tr} (C_{ij}^T V_i V_j^T) \right] \geq \alpha(d)^2 \max_{O_1, \dots, O_n \in \mathcal{O}_d} \sum_{i=1}^n \sum_{j=1}^n \text{tr} (C_{ij}^T O_i O_j^T),$$

where

$$\alpha(d) := \mathbf{E} \left[\frac{1}{d} \sum_{j=1}^d \sigma_j(Z) \right],$$

$Z \in \mathbb{R}^{d \times d}$ is a Gaussian random matrix whose components i.i.d. $\mathcal{N}(0, \frac{1}{d})$ and $\sigma_j(Z)$ is the j -th singular value of Z .

5. Conclusions. Manifold optimization has been extensively studied in literatures. We review the definition of manifold optimization, a few related applications, algorithms and analysis. However, there are still many issues and challenges. Many manifold optimization problems that can be effectively solved are still limited to relatively simple structures such as orthogonal constraints, rank constraints and etc. For other manifolds with complicated structures, what are the most efficient choices of Riemannian metrics and retraction operators are not obvious. Another interesting topic is to combine the manifold structure with the characteristics of specific problems and applications, such as graph-based data analysis, real-time data flow analysis, biomedical image analysis, etc. Nonsmooth problems appear to be more and more attractive.

REFERENCES

- [1] P.-A. ABSIL, C. G. BAKER, AND K. A. GALLIVAN, *Trust-region methods on Riemannian manifolds*, Found. Comput. Math., 7 (2007), pp. 303–330.
- [2] P.-A. ABSIL, R. MAHONY, AND R. SEPULCHRE, *Optimization algorithms on matrix manifolds*, Princeton University Press, Princeton, NJ, 2008.
- [3] P.-A. ABSIL AND J. MALICK, *Projection-like retractions on matrix manifolds*, SIAM Journal on Optimization, 22 (2012), pp. 135–158.
- [4] N. AGARWAL, N. BOUMAL, B. BULLINS, AND C. CARTIS, *Adaptive regularization with cubics on manifolds with a first-order analysis*, arXiv preprint arXiv:1806.00065, (2018).
- [5] M. BACÁK, R. BERGMANN, G. STEIDL, AND A. WEINMANN, *A second order nonsmooth variational model for restoring manifold-valued images*, SIAM Journal on Scientific Computing, 38 (2016), pp. A567–A597.

- [6] Z. BAI, D. LU, AND B. VANDEREYCKEN, *Robust Rayleigh quotient minimization and nonlinear eigenvalue problems*, SIAM J. Sci. Comput., 40 (2018), pp. A3495–A3522.
- [7] A. S. BANDEIRA, N. BOUMAL, AND V. VORONINSKI, *On the low-rank approach for semidefinite programs arising in synchronization and community detection*, in Conference on Learning Theory, 2016, pp. 361–382.
- [8] A. S. BANDEIRA, C. KENNEDY, AND A. SINGER, *Approximating the little grothendieck problem over the orthogonal and unitary groups*, Mathematical programming, 160 (2016), pp. 433–475.
- [9] A. I. BARVINOK, *Problems of distance geometry and convex properties of quadratic maps*, Discrete & Computational Geometry, 13 (1995), pp. 189–202.
- [10] G. BÉCIGNEUL AND O.-E. GANEA, *Riemannian adaptive optimization methods*, arXiv preprint arXiv:1810.00760, (2018).
- [11] G. BENTO, J. NETO, AND P. OLIVEIRA, *Convergence of inexact descent methods for nonconvex optimization on Riemannian manifolds*, arXiv preprint arXiv:1103.4828, (2011).
- [12] G. C. BENTO, O. P. FERREIRA, AND J. G. MELO, *Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds*, Journal of Optimization Theory and Applications, 173 (2017), pp. 548–562.
- [13] E. G. BIRGIN, G. HAESER, AND A. RAMOS, *Augmented lagrangians with constrained subproblems and convergence to second-order stationary points*, Computational Optimization and Applications, 69 (2018), pp. 51–75.
- [14] S. BONNABEL, *Stochastic gradient descent on Riemannian manifolds*, IEEE Transactions on Automatic Control, 58 (2013), pp. 2217–2229.
- [15] P. B. BORCKMANS, S. E. SELVAN, N. BOUMAL, AND P.-A. ABSIL, *A Riemannian subgradient algorithm for economic dispatch with valve-point effect*, Journal of Computational and Applied Mathematics, 255 (2014), pp. 848–866.
- [16] N. BOUMAL, P.-A. ABSIL, AND C. CARTIS, *Global rates of convergence for nonconvex optimization on manifolds*, IMA J. Numer. Anal., (2016).
- [17] N. BOUMAL, V. VORONINSKI, AND A. BANDEIRA, *The non-convex Burer-Monteiro approach works on smooth semidefinite programs*, in Advances in Neural Information Processing Systems, 2016, pp. 2757–2765.
- [18] N. BOUMAL, V. VORONINSKI, AND A. S. BANDEIRA, *Deterministic guarantees for Burer-Monteiro factorizations of smooth semidefinite programs*, arXiv preprint arXiv:1804.02008, (2018).
- [19] S. BURER AND R. D. MONTEIRO, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Mathematical Programming, 95 (2003), pp. 329–357.
- [20] S. BURER AND R. D. MONTEIRO, *Local minima and convergence in low-rank semidefinite programming*, Mathematical Programming, 103 (2005), pp. 427–444.
- [21] J.-F. CAI, H. LIU, AND Y. WANG, *Fast rank-one alternating minimization algorithm for phase retrieval*, Journal of Scientific Computing, 79 (2019), pp. 128–147.
- [22] Y. CAI, L.-H. ZHANG, Z. BAI, AND R.-C. LI, *On an eigenvector-dependent nonlinear eigenvalue problem*, 2017.
- [23] L. CAMBIER AND P.-A. ABSIL, *Robust low-rank matrix completion by Riemannian optimization*, SIAM Journal on Scientific Computing, 38 (2016), pp. S440–S460.
- [24] T. CARSON, D. G. MIXON, AND S. VILLAR, *Manifold optimization for K-means clustering*, in 2017 International Conference on Sampling Theory and Applications (SampTA), IEEE, 2017, pp. 73–77.
- [25] S. CHEN, S. MA, A. M.-C. SO, AND T. ZHANG, *Proximal gradient method for manifold optimization*, arXiv preprint arXiv:1811.00980, (2018).
- [26] M. CHO AND J. LEE, *Riemannian approach to batch normalization*, in Advances in Neural Information Processing Systems, 2017, pp. 5225–5235.
- [27] X. DAI, Z. LIU, L. ZHANG, AND A. ZHOU, *A conjugate gradient method for electronic structure calculations*, SIAM Journal on Scientific Computing, 39 (2017), pp. A2702–A2740.
- [28] G. DE CARVALHO BENTO, J. X. DA CRUZ NETO, AND P. R. OLIVEIRA, *A new approach to the proximal point method: convergence on general Riemannian manifolds*, Journal of Optimization Theory and Applications, 168 (2016), pp. 743–755.
- [29] G. DIRR, U. HELMKE, AND C. LAGEMAN, *Nonsmooth Riemannian optimization with applications to sphere packing and grasping*, in Lagrangian and Hamiltonian Methods for

- Nonlinear Control 2006, Springer, 2007, pp. 29–45.
- [30] J. DUCHI, E. HAZAN, AND Y. SINGER, *Adaptive subgradient methods for online learning and stochastic optimization*, Journal of Machine Learning Research, 12 (2011), pp. 2121–2159.
- [31] A. EDELMAN, T. A. ARIAS, AND S. T. SMITH, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 303–353.
- [32] D. GABAY, *Minimizing a differentiable function over a differential manifold*, J. Optim. Theory Appl., 37 (1982), pp. 177–219.
- [33] B. GAO, X. LIU, X. CHEN, AND Y. YUAN, *A new first-order algorithmic framework for optimization problems with orthogonality constraints*, SIAM Journal on Optimization, 28 (2018), pp. 302–332.
- [34] Y. GAO AND D. SUN, *A majorized penalty approach for calibrating rank constrained correlation matrix problems*, tech. report, National University of Singapore, 2010.
- [35] P. GROHS AND S. HOSSEINI, *Nonsmooth trust region algorithms for locally Lipschitz functions on Riemannian manifolds*, IMA Journal of Numerical Analysis, 36 (2015), pp. 1167–1192.
- [36] S. HOSSEINI, *Convergence of nonsmooth descent methods via Kurdyka–Łojasiewicz inequality on Riemannian manifolds*, Hausdorff Center for Mathematics and Institute for Numerical Simulation, University of Bonn (2015,(INS Preprint No. 1523)), (2015).
- [37] S. HOSSEINI AND A. USCHMAJEV, *A Riemannian gradient sampling algorithm for nonsmooth optimization on manifolds*, SIAM Journal on Optimization, 27 (2017), pp. 173–189.
- [38] J. HU, B. JIANG, L. LIN, Z. WEN, AND Y. YUAN, *Structured quasi-Newton methods for optimization with orthogonality constraints*, arXiv preprint arXiv:1809.00452, (2018).
- [39] J. HU, B. JIANG, X. LIU, AND Z. WEN, *A note on semidefinite programming relaxations for polynomial optimization over a single sphere*, Science China Mathematics, 59 (2016), pp. 1543–1560.
- [40] J. HU, A. MILZAREK, Z. WEN, AND Y. YUAN, *Adaptive regularized newton method for Riemannian optimization*, arXiv preprint arXiv:1708.02016, (2017).
- [41] J. HU, A. MILZAREK, Z. WEN, AND Y. YUAN, *Adaptive quadratically regularized Newton method for Riemannian optimization*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 1181–1207.
- [42] W. HUANG, *Optimization algorithms on Riemannian manifolds with applications*, PhD thesis, The Florida State University, 2013.
- [43] W. HUANG, P.-A. ABSIL, AND K. GALLIVAN, *A Riemannian BFGS method without differentiated retraction for nonconvex optimization problems*, SIAM J. Optim., 28 (2018), pp. 470–495.
- [44] W. HUANG, P.-A. ABSIL, AND K. A. GALLIVAN, *A Riemannian symmetric rank-one trust-region method*, Math. Program., 150 (2015), pp. 179–216.
- [45] W. HUANG, P.-A. ABSIL, AND K. A. GALLIVAN, *Intrinsic representation of tangent vectors and vector transports on matrix manifolds*, Numerische Mathematik, 136 (2017), pp. 523–543.
- [46] W. HUANG, K. A. GALLIVAN, AND P.-A. ABSIL, *A Broyden class of quasi-Newton methods for Riemannian optimization*, SIAM J. Optim., 25 (2015), pp. 1660–1685.
- [47] B. IANNAZZO AND M. PORCELLI, *The Riemannian Barzilai–Borwein method with nonmonotone line search and the matrix geometric mean computation*, IMA Journal of Numerical Analysis, 00 (2017), pp. 1–23.
- [48] B. JIANG AND Y.-H. DAI, *A framework of constraint preserving update schemes for optimization on Stiefel manifold*, Mathematical Programming, 153 (2015), pp. 535–575.
- [49] B. JIANG, Y.-F. LIU, AND Z. WEN, *L_p -norm regularization algorithms for optimization over permutation matrices*, SIAM Journal on Optimization, 26 (2016), pp. 2284–2313.
- [50] B. JIANG, S. MA, A. M.-C. SO, AND S. ZHANG, *Vector transport-free svrg with general retraction for Riemannian optimization: Complexity analysis and practical implementation*, arXiv preprint arXiv:1705.09059, (2017).
- [51] I. T. JOLLIFFE, N. T. TREDAFILOV, AND M. UDDIN, *A modified principal component technique based on the lasso*, Journal of computational and Graphical Statistics, 12 (2003), pp. 531–547.
- [52] R. E. KASS, *Nonlinear regression analysis and its applications*, J. Am. Stat. Assoc., 85 (1990), pp. 594–596.
- [53] A. KOVNATSKY, K. GLASHOFF, AND M. M. BRONSTEIN, *Madmm: a generic algorithm for non-smooth optimization on manifolds*, in European Conference on Computer Vision,

- Springer, 2016, pp. 680–696.
- [54] D. KRESSNER, M. STEINLECHNER, AND B. VANDEREYCKEN, *Low-rank tensor completion by Riemannian optimization*, BIT Numer. Math., 54 (2014), pp. 447–468.
 - [55] R. LAI AND J. LU, *Localized density matrix minimization and linear-scaling algorithms*, Journal of Computational Physics, 315 (2016), pp. 194–210.
 - [56] R. LAI AND S. OSHER, *A splitting method for orthogonality constrained problems*, Journal of Scientific Computing, 58 (2014), pp. 431–449.
 - [57] R. LAI, Z. WEN, W. YIN, X. GU, AND L. M. LUI, *Folding-free global conformal mapping for genus-0 surfaces by harmonic energy minimization*, Journal of Scientific Computing, 58 (2014), pp. 705–725.
 - [58] Y. LECUN, Y. BENGIO, AND G. HINTON, *Deep learning*, nature, 521 (2015), p. 436.
 - [59] C. J. LI, M. WANG, H. LIU, AND T. ZHANG, *Near-optimal stochastic approximation for online principal component estimation*, Mathematical Programming, 167 (2018), pp. 75–97.
 - [60] C. LIU AND N. BOUMAL, *Simple algorithms for optimization on Riemannian manifolds with constraints*, arXiv preprint arXiv:1901.10000, (2019).
 - [61] H. LIU, J.-F. CAI, AND Y. WANG, *Subspace clustering by (k, k) -sparse matrix factorization*, Inverse Problems & Imaging, 11 (2017), pp. 539–551.
 - [62] X. LIU, X. WANG, Z. WEN, AND Y. YUAN, *On the convergence of the self-consistent field iteration in Kohn–Sham density functional theory*, SIAM Journal on Matrix Analysis and Applications, 35 (2014), pp. 546–558.
 - [63] X. LIU, Z. WEN, X. WANG, M. ULBRICH, AND Y. YUAN, *On the analysis of the discretized Kohn–Sham density functional theory*, SIAM Journal on Numerical Analysis, 53 (2015), pp. 1758–1785.
 - [64] X. LIU, Z. WEN, AND Y. ZHANG, *Limited memory block Krylov subspace optimization for computing dominant singular value decompositions*, SIAM J. Sci. Comput., 35 (2013), pp. A1641–A1668.
 - [65] X. LIU, Z. WEN, AND Y. ZHANG, *An efficient Gauss–Newton algorithm for symmetric low-rank product matrix approximations*, SIAM Journal on Optimization, 25 (2015), pp. 1571–1608.
 - [66] Y. LIU, F. SHANG, J. CHENG, H. CHENG, AND L. JIAO, *Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds*, in Advances in Neural Information Processing Systems, 2017, pp. 4868–4877.
 - [67] S. MEI, T. MISIAKIEWICZ, A. MONTANARI, AND R. I. OLIVEIRA, *Solving SDPs for synchronization and maxcut problems via the Grothendieck inequality*, arXiv preprint arXiv:1703.08729, (2017).
 - [68] A. MONTANARI AND E. RICHARD, *Non-negative principal component analysis: Message passing algorithms and sharp asymptotics*, IEEE Transactions on Information Theory, 62 (2016), pp. 1458–1484.
 - [69] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer, New York, second ed., 2006.
 - [70] E. OJA AND J. KARHUNEN, *On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix*, Journal of mathematical analysis and applications, 106 (1985), pp. 69–84.
 - [71] G. PATAKI, *On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues*, Mathematics of operations research, 23 (1998), pp. 339–358.
 - [72] P. PULAY, *Convergence acceleration of iterative sequences. the case of SCF iteration*, Chemical Physics Letters, 73 (1980), pp. 393–398.
 - [73] P. PULAY, *Improved SCF convergence acceleration*, Journal of Computational Chemistry, 3 (1982), pp. 556–560.
 - [74] C. QI, *Numerical optimization methods on Riemannian manifolds*, PhD thesis, Florida State University, 2011.
 - [75] W. RING AND B. WIRTH, *Optimization methods on Riemannian manifolds and their application to shape space*, SIAM J. Optim., 22 (2012), pp. 596–627.
 - [76] H. SATO, H. KASAI, AND B. MISHRA, *Riemannian stochastic variance reduced gradient*, arXiv preprint arXiv:1702.05594, (2017).
 - [77] R. M. SCHOEN AND S.-T. YAU, *Lectures on harmonic maps*, vol. 2, Amer Mathematical Society, 1997.

- [78] M. SEIBERT, M. KLEINSTEUBER, AND K. HÜPER, *Properties of the BFGS method on Riemannian manifolds*, Mathematical System Theory C Festschrift in Honor of Uwe Helmke on the Occasion of his Sixtieth Birthday, (2013), pp. 395–412.
- [79] O. SHAMIR, *A stochastic PCA and SVD algorithm with an exponential convergence rate*, in International Conference on Machine Learning, 2015, pp. 144–152.
- [80] D. SIMON AND J. ABELL, *A majorization algorithm for constrained correlation matrix approximation*, Linear Algebra Appl., 432 (2010), pp. 1152–1164.
- [81] A. SINGER AND Y. SHKOLNISKY, *Three-dimensional structure determination from common lines in cryo-em by eigenvectors and semidefinite programming*, SIAM journal on imaging sciences, 4 (2011), pp. 543–572.
- [82] S. T. SMITH, *Optimization techniques on Riemannian manifolds*, Fields Institute Communications, 3 (1994).
- [83] W. SUN AND Y. YUAN, *Optimization theory and methods: nonlinear programming*, vol. 1, Springer Science & Business Media, 2006.
- [84] A. TOTH, J. A. ELLIS, T. EVANS, S. HAMILTON, C. KELLEY, R. PAWLOWSKI, AND S. SLATTERY, *Local improvement results for Anderson acceleration with inaccurate function evaluations*, SIAM Journal on Scientific Computing, 39 (2017), pp. S47–S65.
- [85] C. UDRISTE, *Convex functions and optimization methods on Riemannian manifolds*, vol. 297, Springer Science & Business Media, 1994.
- [86] M. ULBRICH, Z. WEN, C. YANG, D. KLOCKNER, AND Z. LU, *A proximal gradient method for ensemble density functional theory*, SIAM Journal on Scientific Computing, 37 (2015), pp. A1975–A2002.
- [87] B. VANDEREYCKEN, *Low-rank matrix completion by Riemannian optimization*, SIAM Journal on Optimization, 23 (2013), pp. 1214–1236.
- [88] N. K. VISHNOI, *Geodesic convex optimization: Differentiation on manifolds, geodesics, and convexity*, arXiv preprint arXiv:1806.06373, (2018).
- [89] I. WALDSPURGER, A. D’ASPREMONT, AND S. MALLAT, *Phase recovery, maxcut and complex semidefinite programming*, Mathematical Programming, 149 (2015), pp. 47–81.
- [90] Y. WANG, W. YIN, AND J. ZENG, *Global convergence of admm in nonconvex nonsmooth optimization*, Journal of Scientific Computing, 78 (2019), pp. 29–63.
- [91] K. WEI, J.-F. CAI, T. F. CHAN, AND S. LEUNG, *Guarantees of Riemannian optimization for low rank matrix recovery*, SIAM Journal on Matrix Analysis and Applications, 37 (2016), pp. 1198–1222.
- [92] Z. WEN, A. MILZAREK, M. ULBRICH, AND H. ZHANG, *Adaptive regularized self-consistent field iteration with exact Hessian for electronic structure calculation*, SIAM J. Sci. Comput., 35 (2013), pp. A1299–A1324.
- [93] Z. WEN, C. YANG, X. LIU, AND Y. ZHANG, *Trace-penalty minimization for large-scale eigenspace computation*, Journal of Scientific Computing, 66 (2016), pp. 1175–1203.
- [94] Z. WEN AND W. YIN, *A feasible method for optimization with orthogonality constraints*, Math. Program., 142 (2013), pp. 397–434.
- [95] Z. WEN, W. YIN, AND Y. ZHANG, *Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm*, Mathematical Programming Computation, 4 (2012), pp. 333–361.
- [96] Z. WEN AND Y. ZHANG, *Accelerating convergence by augmented Rayleigh–Ritz projections for large-scale eigenpair computation*, SIAM Journal on Matrix Analysis and Applications, 38 (2017), pp. 273–296.
- [97] X. WU, Z. WEN, AND W. BAO, *A regularized newton method for computing ground states of Bose-Einstein condensates*, arXiv preprint arXiv:1504.02891, (2015).
- [98] X. XIAO, Y. LI, Z. WEN, AND L. ZHANG, *A regularized semi-smooth newton method with projection steps for composite convex programs*, Journal of Scientific Computing, (2018), pp. 1–26.
- [99] T. XIE AND F. CHEN, *Non-convex clustering via proximal alternating linearized minimization method*, International Journal of Wavelets, Multiresolution and Information Processing, 16 (2018), p. 1840013.
- [100] W. H. YANG, L.-H. ZHANG, AND R. SONG, *Optimality conditions for the nonlinear programming problems on Riemannian manifolds*, Pacific Journal of Optimization, 10 (2014), pp. 415–434.

- [101] H. YUAN, X. GU, R. LAI, AND Z. WEN, *Global optimization with orthogonality constraints via stochastic diffusion on manifold*, arXiv preprint arXiv:1707.02126, (2017).
- [102] R. ZASS AND A. SHASHUA, *Nonnegative sparse pca*, in Advances in neural information processing systems, 2007, pp. 1561–1568.
- [103] H. ZHANG AND W. W. HAGER, *A nonmonotone line search technique and its application to unconstrained optimization*, SIAM J. Optim., 14 (2004), pp. 1043–1056.
- [104] H. ZHANG, S. J. REDDI, AND S. SRA, *Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds*, in Advances in Neural Information Processing Systems, 2016, pp. 4592–4600.
- [105] H. ZHANG AND S. SRA, *First-order methods for geodesically convex optimization*, in Conference on Learning Theory, 2016, pp. 1617–1638.
- [106] J. ZHANG, H. LIU, Z. WEN, AND S. ZHANG, *A sparse completely positive relaxation of the modularity maximization for community detection*, SIAM J. Sci. Comput., 40 (2018), pp. A3091–A3120.
- [107] J. ZHANG, S. MA, AND S. ZHANG, *Primal-dual optimization algorithms over Riemannian manifolds: an iteration complexity analysis*, arXiv preprint arXiv:1710.02236, (2017).
- [108] J. ZHANG, Z. WEN, AND Y. ZHANG, *Subspace methods with local refinements for eigenvalue computation using low-rank tensor-train format*, Journal of Scientific Computing, 70 (2017), pp. 478–499.
- [109] J. ZHANG AND S. ZHANG, *A cubic regularized Newton’s method over Riemannian manifolds*, arXiv preprint arXiv:1805.05565, (2018).
- [110] L. ZHANG AND R. LI, *Maximization of the sum of the trace ratio on the Stiefel manifold, ii: Computation*, Science China Mathematics, 58 (2015), pp. 1549–1566.
- [111] X. ZHANG, J. ZHU, Z. WEN, AND A. ZHOU, *Gradient type optimization methods for electronic structure calculations*, SIAM Journal on Scientific Computing, 36 (2014), pp. C265–C289.
- [112] Y. ZHANG, Y. LAU, H.-W. KUO, S. CHEUNG, A. PASUPATHY, AND J. WRIGHT, *On the global geometry of sphere-constrained sparse blind deconvolution*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4894–4902.
- [113] Z. ZHAO, Z.-J. BAI, AND X.-Q. JIN, *A Riemannian newton algorithm for nonlinear eigenvalue problems*, SIAM Journal on Matrix Analysis and Applications, 36 (2015), pp. 752–774.
- [114] X. ZHU, *A Riemannian conjugate gradient method for optimization on the Stiefel manifold*, Computational Optimization and Applications, 67 (2017), pp. 73–110.