An inexact augmented Lagrangian method for nonsmooth optimization on Riemannian manifold *

Kang-Kang Deng †, Zheng Peng †‡

Abstract: We consider a nonsmooth optimization problem on Riemannian manifold, whose objective function is the sum of a differentiable component and a nonsmooth convex function. We propose a manifold inexact augmented Lagrangian method (MIALM) for the considered problem. The problem is reformulated to a separable form. By utilizing the Moreau envelope, we get a smoothing subproblem at each iteration of the proposed method. Theoretically, under suitable assumptions, the convergence to critical point of the proposed method is established. In particular, under the condition of that the approximate global minimizer of the iteration subproblem could be obtained, we prove the convergence to global minimizer of the origin problem. Numerical experiments show that, the MIALM is a competitive method compared to some existing methods.

Keywords: Manifold optimization; Nonsmooth optimization; Augmented Lagrangian method; Moreau envelope.

Mathematics Subject Classification: 90C30, 90C26

1 Introduction

Riemannian manifold optimization is a class of constrained optimization problems, in which the constraint set is a subset of Riemannian manifold $\mathcal{M}$. It has recently aroused considerable research interests due to the wide applications in different fields such as computer vision, signal processing, etc [3]. In these applications, manifold $\mathcal{M}$ could be Stiefel manifold, Grassmann manifold, or symmetric positive definite manifold. Analogy to classical optimization methods in Euclidean space, some Riemannian optimization methods have been explored, e.g., gradient-

*This research was supported by the Natural Science Foundation of China (Grant. 11571074)

†College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China.

‡School of Mathematics and Computational Science, Xiangtan Univeristy, Xiangtan 411105, China. Corresponding author, E-mail: pzheng@fzu.edu.cn.

type methods[3, 10, 37], Newton-type methods[28, 36, 8] and trust region methods[1, 9, 27].

In this paper, we consider a nonsmooth nonconvex Riemannian optimization problem as follows

$$\begin{cases} \min\limits_{X\in\mathbb{R}^{n\times r}} & F(X) := f(X) + g(AX) \\ \text{s.t.} & X \in \mathcal{M}, \end{cases} \tag{1.1}$$

where $f : \mathcal{M} \to \mathbb{R}$ is a smooth but possibly nonconvex function, g is convex but nonsmooth, $\mathcal{M}$ is a Riemannian manifold embedded in Euclidean space $\mathbb{E}$. Many convex or non-convex problems in machine learning applications have the form of problem (1.1), e.g., sparse principle component analysis [40], sparse canonical correlation analysis [34], robust low-rank matrix completion [13, 26] and multi-antenna channel communications [39, 19], etc.

In [2], Absil and Hosseini gave many examples of manifold optimization with nonsmooth objective. We list three representative examples in the following.

Example 1.1 (Sparse principle component analysis (SPCA))**.**

$$\begin{cases} \min\limits_{X\in\mathbb{R}^{n\times r}} & -X^T A^T AX + \lambda\|X\|_1, \\ \text{s.t.} & X^T X = I_r. \end{cases} \tag{1.2}$$

Example 1.2 (Compressed modes in physics (CMs))**.**

$$\begin{cases} \min\limits_{\Psi\in\mathbb{R}^{n\times r}} tr(\Psi^T \Delta\Psi) + \mu\|\Psi\|_1, \\ \text{s.t.} \quad \Psi^T \Psi = I_r. \end{cases} \tag{1.3}$$

Example 1.3 (Robust low-rank matrix completion)**.**

$$\begin{cases} \min\limits_{X\in\mathbb{R}^{n\times n}} & \|P_\Omega(X - M)\|_1, \\ \text{s.t.} & X \in \mathcal{M}_r := \{X|\ \text{rank } (X) = r\}. \end{cases} \tag{1.4}$$

Problem (1.1) is reformulated to a separable form in this paper, and then a manifold inexact augmented Lagrangian method (MIALM) is proposed for the separable form of problem (1.1). The iteration subproblem of the MIALM is formulated to a smooth optimization problem by utilizing the Moreau envelope, it could be solved by some classical Riemannian optimization methods such as Riemannian gradient/Newton/Quasi-Newton method. This algorithmic framework is adapted from [31, 32, 17] for classical nonsmooth composite problem in Euclidean space, which has drawn significant research attentions. The convergence to critical point of the proposed MIALM method is established under some suitable conditions. In particular, if an approximate global minimizer of the iteration subproblem could be obtained, the convergence to global minimizer of the original problem could be proved. Numerical experiments show that, the MIALM is competitive compared to some existing methods.

The rest of this paper is organized as follows. Some related works on nonsmooth manifold optimization problem are summarized in Section 2, and some preliminaries on manifold are given in Section 3. In Section 4, a manifold inexact augmented Lagrangian method is proposed and the iteration subproblem solver is presented. The convergence of the proposed method is established in Section 5. Numerical results on compressed modes problems in physics and sparse PCA are reported in Section 6. Finally, Section 7 concludes this paper by some final remarks.

2 Related works

We summarize some related works for nonsmooth optimization problem on manifold in this section. The existing results mainly focused on two classes of nonsmooth manifold optimization problem: nonsmooth optimization problem with locally Lipschitz objective function, and structured optimization problem having the form of problem (1.1).

Grohs and Hosseini [21] proposed the ϵ-subgradient algorithm for minimizing a locally Lipschitz function on Riemannian manifold. By utilizing ϵ-subgradient-oriented descent directions and the generalized Wolfe line-search on Riemannian manifold, Hosseini, Huang and Yousefpour [24] presented a nonsmooth Riemannian line search algorithm and established the convergence to a stationary point. Grohs [20] presented a nonsmooth trust region algorithm for minimizing locally Lipschitz objective function on Riemannian manifold. The iteration complexity of these subgradient algorithms was also investigated in [5] and [18]. In [25] and [12], the authors proposed the Riemannian gradient sampling algorithms. At each iteration of these methods, the subdifferential of the objective function is approximated by the convex hull of transported gradients of nearby points, and the nearby points are randomly generated in the tangent space of the current iterate.

Some proximal point algorithms on Riemannian manifold were investigated in the recent. Bento, Ferreira and Melo [5] analyzed the iteration complexity of a proximal point algorithm on Hadamard manifold having non-positive sectional curvature. Bento, et al [16] gave the full convergence for any bounded sequence generated by the proximal point method, without assumption on the sign of the sectional curvature on manifold. The Kurdyka-Łojasiewicz inequality on Riemannian manifold is a powerful tool for convergence analysis of optimization methods on manifold. Bento, et al [6] analyzed the full convergence of a steepest descent method and a proximal point method via Kurdyka-Łojasiewicz inequality. Seyedehsomayeh [23] proposed a subgradient-oriented descent method and proved that, if the objective function has the Kurdyka-Łojasiewicz property, the iteration sequence generated by the subgradient-oriented descent method converges to a singular critical point.

By a separable reformulation of problem (1.1), the variable involving Riemannian manifold constraint and that one involving nonsmooth term could be handled separately. To do so, it results in two tractable subproblems. Based on this idea, Lai, et al [30] proposed a splitting of orthogonality constraints (SOC) method for a special case of problem (1.1), in which $f \equiv 0$ and $A = I$. That is

$$\begin{cases} \min\limits_{X} \ g(X), \\ \text{s.t. } X \in \mathcal{M}. \end{cases} \tag{2.1}$$

To solve problem (2.1), the SOC method considered the following separable reformulation:

$$\begin{cases} \min\limits_{X,Y} \ g(Y), \\ \text{s.t. } X \in \mathcal{M}, \ X = Y. \end{cases} \tag{2.2}$$

The associated partial augmented Lagrangian function is

$$\mathcal{L}_\beta := g(Y) - \langle \Lambda, X - Y \rangle + \frac{\beta}{2}\|X - Y\|_F^2 \tag{2.3}$$

where Λ is the Lagrangian multiplier, and β is a penalty parameter. The SOC method updates iterate via

$$\begin{cases} X^{k+1} = \arg\min\limits_{X \in \mathcal{M}} \frac{\beta}{2}\|X - Y^k - \frac{1}{\beta}\Lambda^k\|_F^2, \\ Y^{k+1} = \arg\min g(Y) + \frac{\beta}{2}\|X^{k+1} - Y - \frac{1}{\beta}\Lambda^k\|_F^2, \\ \Lambda^{k+1} = \Lambda^k - \beta(X^{k+1} - Y^{k+1}). \end{cases} \tag{2.4}$$

The X-subproblem is “easy” via projection on $\mathcal{M}$, and the Y -subproblem is often structured in real applications.

Chen, et al [15] proposed a proximal alternating minimization augmented Lagrangian (PAMAL) method of multipliers for problem (1.1) with $A = I$ and $\mathcal{M} = St_n$. Specifically, the PAMAL method first reformulates the problem to:

$$\begin{cases} \min\limits_{X,Y,Q} \ f(Y) + h(Q), \\ \text{s.t. } X = Y, X = Q, X \in \mathcal{M}. \end{cases} \tag{2.5}$$

Then it considers the augmented Lagrangian method of multipliers framework aiming to obtain the solution for the jointed variable (X, Y, Q) at each iteration. The iterate is produced by

$$\begin{cases} (X^{k+1}, Y^{k+1}, Q^{k+1}) = \arg\min_{X,Y,Q} \mathcal{L}_\beta(X, Y, Q; \Lambda_1^k, \Lambda_2^k), \\ \Lambda_1^{k+1} = \Lambda_1^k - \beta(X^{k+1} - Y^{k+1}), \\ \Lambda_2^{k+1} = \Lambda_2^k - \beta(X^{k+1} - Q^{k+1}), \end{cases} \tag{2.6}$$

where $\mathcal{L}_\beta$ is the augmented Lagrangian function associated to (2.5). The subproblem on the jointed variable (X, Y, Q) is intractable, hence the authors proposed a proximal alternating minimization method to handle it. Hong, et al [22] considered a more general form where $\mathcal{M}$ is the generalized orthogonal constraint, and proposed a PAMAL-type algorithm in which a proximal alternating linearized minimization method was used for iteration subproblem.

Kovnatsky, et al [29] proposed a manifold ADMM (MADMM) for a general manifold optimization problem as follows

$$\begin{cases} \min_{X,Y} \ f(X) + g(Y) \\ \text{s.t. } AX = Y, X \in \mathcal{M} \end{cases} \tag{2.7}$$

The associated partial augmented Lagrangian function is

$$\mathcal{L}_\beta(X, Y; \Lambda) := f(X) + g(Y) - \langle \Lambda, AX - Y \rangle + \frac{\beta}{2}\|AX - Y\|_F^2.$$

The MADMM has the iterate as follows

$$\begin{cases} X^{k+1} = \arg\min_{X \in \mathcal{M}} \mathcal{L}_\beta(X, Y^k, \Lambda^k) \\ Y^{k+1} = \arg\min_{Y} \mathcal{L}_\beta(X^{k+1}, Y, \Lambda^k) \\ \Lambda^{k+1} = \Lambda^k - \beta(AX^{k+1} - Y^{k+1}) \end{cases} \tag{2.8}$$

More recently, Chen, et al [14] proposed a manifold proximal gradient method (ManPG) for problem (1.1) with $A = I$, i.e.

$$\min_X f(X) + g(X), \quad \text{s.t. } X \in \mathcal{M} \tag{2.9}$$

At the k-th iteration, the search direction D^k of ManPG is obtained by

$$\begin{cases} \min_D \ \langle D, \text{grad} f(X^k) \rangle + \frac{\beta}{2}\|D\|_F^2 + g(X^k + D), \\ \text{s.t. } D \in T_{X^k}\mathcal{M}, \end{cases} \tag{2.10}$$

where $D \in T_{X^k}\mathcal{M}$ can be represented by a linear system $\mathcal{A}_k(D) = 0$. The subproblem (2.10) is solved by applying the semi-smooth Newton method to the KKT system. The next iterate X^{k+1} is then obtained by

$$X^{k+1} = R_{X^k}(\alpha_k D^k).$$

3 Preliminaries

3.1 Riemannian optimization

Let $\mathcal{M}$ be a smooth manifold, and $\mathbb{E}$ be the Euclidean space. The tangent space of $\mathcal{M}$ at $x \in \mathcal{M}$ is denoted by $T_x\mathcal{M}$. A Riemannian manifold $(\mathcal{M}, \langle\cdot,\cdot\rangle)$ is a smooth manifold equipped with inner product $\langle\cdot,\cdot\rangle_x$ on each point $x \in \mathcal{M}$. Let (U, φ) be a chart, where U is an open set with $x \in U \subset \mathcal{M}$ and φ is a homeomorphism between U and open set $\varphi(U)$ in Euclidean space. Given a smooth Riemannian manifold $\mathcal{M}$, the chart exists at each point $x \in \mathcal{M}$.

Definition 3.1 (Riemannian Gradient)**.** Riemannian gradient, denoted by $\text{grad}f(x) \in T_x\mathcal{M}$, is the unique tangent vector satisfying

$$\langle \text{grad}f(x), \xi\rangle_x = df(x)[\xi],\ \forall\ \xi \in T_x\mathcal{M}. \tag{3.1}$$

If $\mathcal{M}$ is an embedded manifold of Euclidean space, the Riemannian metric between $u, v \in T_x\mathcal{M}$ could be introduced by an inner product of Euclidean space, i.e. $\langle u, v\rangle_x = \langle u, v\rangle$, where the later is classical inner product of Euclidean space. In the sense, we have

$$\text{grad}f(x) = \text{Proj}_{T_x\mathcal{M}}(\nabla f(x)) \tag{3.2}$$

where $\nabla f(x)$ is the gradient in Euclidean space, $\text{Proj}_{T_x\mathcal{M}}$ is a projection on tangent space $T_x\mathcal{M}$.

Definition 3.2 (Riemannian Hessian)**.** Given a smooth function $f : \mathcal{M} \to \mathbb{R}$, the Riemannian Hessian of f at x in $\mathcal{M}$ is linear mapping $\text{Hess}f(x)$ of $T_x\mathcal{M}$ into itself defined by

$$\text{Hess}f(x)[\xi_x] = \nabla_{\xi_x}\text{grad}f(x) \tag{3.3}$$

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

Definition 3.3 (Retraction)**.** A retraction on manifold $\mathcal{M}$ is a smooth mapping $R : T\mathcal{M} \to \mathcal{M}$ having the following properties. Let $R_x : T_x\mathcal{M} \to \mathcal{M}$ be the restriction of R to $T_x\mathcal{M}$, then

- $R_x(0_x) = x$, where 0_x is zero element of $T_x\mathcal{M}$
- $dR_x(0_x) = id_{T_x\mathcal{M}}$, where $id_{T_x\mathcal{M}}$ is the identity mapping on $T_x\mathcal{M}$

Definition 3.4 (Vector Transport)**.** The vector transport $\mathcal{T}$ is a smooth mapping with

$$T\mathcal{M} \oplus T\mathcal{M} \to T\mathcal{M} : (\eta_x, \xi_x) \mapsto \mathcal{T}_{\eta_x}(\xi_x) \in T\mathcal{M}, \forall\ x \in \mathcal{M}, \tag{3.4}$$

where $\mathcal{T}$ satisfies that

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

Definition 3.5 (The Clarke subdifferential on Riemannian manifold)**.** For a locally Lipschitz continuous function f on $\mathcal{M}$, the Riemannian generalized directional derivative of f at $x \in \mathcal{M}$ in direction $v \in T_x\mathcal{M}$ is given by

$$f^\circ(x; v) = \lim_{y\to x} \sup_{t\downarrow 0} \frac{f \circ \varphi^{-1}(\varphi(y) + t\mathrm{D}\varphi(y)[v]) - f \circ \varphi^{-1}(\varphi(y))}{t}, \tag{3.5}$$

where (φ, U) is coordinate chart at x. The generalized gradient or the Clarke subdifferential of f at $x \in \mathcal{M}$ is

$$\partial f(x) = \{\xi \in T_x\mathcal{M} : \langle \xi, v\rangle_x \le f^\circ(x; v), \forall v \in T_x\mathcal{M}\}. \tag{3.6}$$

Consider a manifold minimization problem

$$\begin{cases} \min\limits_x & f(x) \\ \text{s.t.} & c_i(x) = 0, i = 1, \cdots, m, \\ & x \in \mathcal{M}. \end{cases} \tag{3.7}$$

Let $\Omega := \{x \in \mathcal{M} : c_i(x) = 0, i = 1 \cdots, m\}$. Given $x^* \in \Omega$, assume that the Linear Independent Constraint Qualification (LICQ) holds at x^*, then normal cone $\mathcal{N}_\Omega(x^*)$ is [35]:

$$\mathcal{N}_\Omega(x^*) = \left\{ \sum_{i=1}^m \lambda_i \mathrm{grad} c_i(x^*) \,\middle|\, \lambda \in \mathbb{R}^m \right\} \tag{3.8}$$

Hence, for the first-oder optimality condition of problem (3.7), we have

Lemma 3.1 ([38],Proposition 2.7)**.** *Suppose $x^* \in \mathcal{M}$ and $c_i(x^*) = 0, i = 1 \cdots, m$. if*

$$\partial f(x^*) \cap (-\mathcal{N}_\Omega(x^*)) \neq \emptyset, \tag{3.9}$$

then x^* is a stationary solution of problem (3.7).

3.2 Proximal operator and retraction-smooth

For a proper, convex and low semicontinuous function $g : \mathbb{E} \to \mathbb{R}$, the proximal operator with parameter $\mu \ge 0$, denoted by $\mathrm{prox}_{\mu g}$, is defined by

$$\mathrm{prox}_{\mu g}(v) := \arg\min_x \{g(x) + \frac{1}{2\mu}\|x - v\|^2\}. \tag{3.10}$$

The associated Moreau envelope is a function $M : \mathbb{E} \to \mathbb{R}$ given by

$$\begin{aligned} M_{\mu g}(v) :&= \min_x \{g(x) + \frac{1}{2\mu}\|x - v\|^2\} \\ &= g(\text{prox}_{\mu g}(v)) + \frac{1}{2\mu}\|\text{prox}_{\mu g}(v) - v\|^2. \end{aligned} \tag{3.11}$$

The Moreau envelope is a continuously differentiable function, even when g is not. This is:

Lemma 3.2 (Theorem 6.60 in [4])**.** *Let $g : \mathbb{E} \to \mathbb{R}$ be a proper closed and convex function, and $\mu \geq 0$. Then $M_{\mu g}$ is $\frac{1}{\mu}$-smooth in $\mathbb{E}$, and for $\forall\ v \in \mathbb{E}$ one has*

$$\nabla M_{\mu g}(v) = \frac{1}{\mu}(v - prox_{\mu g}(v)). \tag{3.12}$$

Lemma 3.2 states that, the Moreau envelope is continuously differentiable in Euclidean space $\mathbb{E}$. Next we will show the relationship between Retraction smoothness in submanifold of Euclidean space and smoothness in Euclidean space.

Definition 3.6 (Retraction-Smooth)**.** A function $f : \mathcal{M} \to \mathbb{R}$ is said to be retraction L-smooth if for $\forall\ x, y \in \mathcal{M}$, it holds that

$$f(y) \leq f(x) + \langle \text{grad} f(x), \xi \rangle_x + \frac{L}{2}\|\xi\|_x^2, \tag{3.13}$$

where $\xi \in T_x\mathcal{M}$ and $R_x(\xi) = y$.

Let $\mathcal{M}$ be a Riemannian submanifold of $\mathbb{E}$. The following lemma states that, if $f : \mathbb{R}^n \to \mathbb{R}$ has Lipschitz continuous gradient, then f is also retraction smooth on $\mathcal{M}$.

Lemma 3.3. *[Lemma 4 in [10]] Let $\mathbb{E}$ be a Euclidean space (for example, $\mathbb{E} = \mathbb{R}^n$) and $\mathcal{M}$ be a compact Riemannian submanifold of $\mathbb{E}$. If $f : \mathbb{E} \to \mathbb{R}$ has Lipschitz continuous gradient in the convex hull of $\mathcal{M}$, then there exists a positive constant L_g such that*

$$f(R_{x_k}(\eta)) \leq f(x_k) + \langle \eta, gradf(x_k) \rangle + \frac{L_g}{2}\|\eta\|^2 \tag{3.14}$$

holds at $\forall\ \eta \in T_{x_k}\mathcal{M}$.

Lemma 3.3 was proved in [10]. For the sake of completeness, we give a proof as follows.

Proof. By Lipschitz continuity, ∇f is Lipschitz along any line segment in $\mathbb{E}$ jointing x and y. Hence, there exists $L > 0$ such that

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2,\ \forall x, y \in \mathcal{M}. \tag{3.15}$$

It also holds if $y = R_x(\eta),\ \forall \eta \in T_x\mathcal{M}$. Since $\mathrm{grad} f(x)$ is a orthogonal projection of $\nabla f(x)$ on $T_x\mathcal{M}$, we have

$$\begin{aligned} \langle \nabla f(x), R_x(\eta) - x \rangle &= \langle \nabla f(x), \eta + R_x(\eta) - x - \eta \rangle \\ &= \langle \mathrm{grad} f(x), \eta \rangle + \langle \nabla f(x), R_x(\eta) - x - \eta \rangle . \end{aligned} \tag{3.16}$$

It is easy to deduce from (3.15) and (3.16) that

$$f(R_x(\eta)) \leq f(x) + \langle \mathrm{grad} f(x), \eta \rangle + \frac{L}{2}\|R_x(\eta) - x\|^2 + \|\nabla f(x)\|\|R_x(\eta) - x - \eta\|.$$

Since $\nabla f(x)$ is continuous on compact set $\mathcal{M}$, there exists $G > 0$ such that $\|\nabla f(x)\| \leq G,\ \forall\, x \in \mathcal{M}$. By Definition 3.3 and the compactness of manifold, there exists $\alpha, \beta \geq 0$ such that, for all $x \in \mathcal{M}$ and all $\eta \in T_x\mathcal{M}$, we have

$$\|R_x(\eta) - x\| \leq \alpha\|\eta\|^2, \text{and } \|R_x(\eta) - x - \eta\| \leq \beta\|\eta\|^2.$$

Hence

$$f(R_x(\eta)) \leq f(x) + \langle \mathrm{grad} f(x), \eta \rangle + \left(\frac{L}{2}\alpha^2 + G\beta\right)\|\eta\|^2.$$

Let $L_g = \left(\frac{L}{2}\alpha^2 + G\beta\right)$, we have (3.14) and complete the proof. □

4 The proposed method

4.1 Problem reformulation

For regularity, we need the following assumptions.

Assumption 4.1.

A. $\mathcal{M}$ is a compact Riemannian submanifold embedded in Euclidean space $\mathbb{E}$;

B. f is smooth but not necessary convex, g is a nonsmooth convex function on $\mathbb{E}$, and $\partial g(Y)$ is uniformly bounded for all $Y \in \mathbb{R}^{d\times r}$, where $\partial g(Y)$ is a subgradient of g at Y in usual sense.

By introducing auxiliary variable $Y = AX$, problem (1.1) can be reformulated to

$$\begin{cases} X^* = \arg \min\limits_{X\in\mathbb{R}^{n\times r}} f(X) + g(Y) \\ \text{s.t. } AX = Y,\ X \in \mathcal{M}. \end{cases} \tag{4.1}$$

where $A \in \mathbb{R}^{d\times n}$. The partial Lagrangian function associated to problem (4.1) is

$$L(X, Y; Z) := f(X) + g(Y) - \langle Z, AX - Y \rangle \tag{4.2}$$

By Lemma (3.1), we obtain the KKT system of problem (4.1) as follows:

Proposition 4.1. *Suppose in problem* (4.1) *that* f *is smooth with Lipschitz gradient and* g *is convex and locally Lipschitz continuous. Then,* (X^*, Y^*) *satisfies the KKT conditions if there exists a Lagrange multiplier* Z^* *such that*

$$\begin{cases} 0 \in & Proj_{T_{X^*}\mathcal{M}}(\nabla f(X^*) - A^T Z^*), \\ 0 \in & \partial g(Y^*) + Z^*, \\ & AX^* = Y^*. \end{cases} \tag{4.3}$$

4.2 Manifold inexact augmented Lagrangian method

The augmented Lagrangian associated with (4.1) is

$$\begin{aligned} \mathcal{L}_\rho(X, Y; Z) &= L(X, Y; Z) + \frac{\rho}{2}\|AX - Y\|_F^2 \\ &= f(X) + g(Y) - \langle Z, AX - Y\rangle + \frac{\rho}{2}\|AX - Y\|_F^2. \end{aligned} \tag{4.4}$$

For a given (X^k, Y^k, Z^k), the next iterate generated by the manifold inexact augmented Lagrangian method (MIALM) is

$$\begin{cases} (X^{k+1}, Y^{k+1}) & = \arg\min_{X\in\mathcal{M}, Y} \mathcal{L}_\rho(X, Y; Z^k), \\ Z^{k+1} & = Z^k - \rho(AX^{k+1} - Y^{k+1}). \end{cases} \tag{4.5}$$

The (X, Y)- subproblem is intractable due to the nonsmoothess and joint minimization. Hence, an efficient Riemannian optimization method should be proposed for this subproblem in MIALM (4.5). Notice that, for fixed $\rho > 0$ and Z we aim to solve

$$\min_{X\in\mathcal{M}, Y\in\mathbb{R}^{d\times r}} \Psi(X, Y) := \mathcal{L}_\rho(X, Y; Z) \tag{4.6}$$

Let

$$\begin{aligned} \psi_Z(X) := & \inf_Y \Psi(X, Y) \\ = & f(X) + g(\text{Prox}_{g/\rho}(AX - \mu Z)) \\ & + \frac{\rho}{2}\|AX - \frac{1}{\rho}Z - \text{Prox}_{\mu g}(AX - \frac{1}{\rho}Z)\|_F^2 - \frac{1}{2\rho}\|Z\|_F^2. \end{aligned} \tag{4.7}$$

The new iterate $(\bar{X}, \bar{Y})$ is produced sequentially by

$$\bar{X} = \arg\min_{X\in\mathcal{M}} \psi_Z(X), \quad \bar{Y} = \text{Prox}_{g/\rho}(A\bar{X} - \frac{1}{\rho}Z). \tag{4.8}$$

In the sense, the MIALM iterate could be rewritten to

$$\begin{cases} X^{k+1} = \arg\min_{X\in\mathcal{M}} \psi_{Z^k}(X), \\ Y^{k+1} = \text{Prox}_{g/\rho}(AX^{k+1} - \frac{1}{\rho}Z^k), \\ Z^{k+1} = Z^k - \rho(AX^{k+1} - Y^{k+1}). \end{cases} \tag{4.9}$$

By (3.12), we have

$$\begin{aligned}\nabla\psi_Z(X) &= \nabla f(X) + \rho A^T\left(AX - \frac{1}{\rho}Z - \mathrm{Prox}_{\mu g}(AX - \frac{1}{\rho}Z)\right)\\ &= \nabla f(X) + \rho A^T\left(\mathrm{Prox}_{\rho g^*}(AX - \frac{1}{\rho}Z)\right)\end{aligned} \tag{4.10}$$

where g^* is the conjugate operator of g and defined by $g^*(x) = \sup_v\{\langle x, v\rangle - g(v)\}$. By Lemma 3.3, $\psi_Z(\cdot)$ is retraction smooth over Riemannian manifold $\mathcal{M}$, and its Riemannian gradient is

$$\mathrm{grad}\psi_Z(X) = \mathrm{Proj}_{T_X\mathcal{M}}(\nabla\psi_Z(X)).$$

Thus, at the k-th iteration, the X-subproblem is identical to find X^{k+1} such that

$$\mathrm{grad}\psi_{Z^k}(X) = 0.$$

Algorithm 1 summarizes the proposed manifold inexact augmented Lagrangian method in detail.

Remark 4.1. 1) The proposed method is an ALM-type method. The complexity of X-subproblem is as same as that of MADMM. However, our method obtains a joint optimal solution which guarantees the convergence, while the MADMM does not.

2) All efficient Riemannian optimization methods are applicable for the X-subproblem, e.g., Riemannian gradient method, Riemannian Newton method, etc.

3) The proposed method is utilizable for smooth Riemannian optimization problem under set-constrained, in which $g(X) = \delta_\Omega(X)$, the indictor function of constraint set Ω.

4.3 Riemannian optimization subproblem

The main computational cost of Algorithm 1 is to solve the X-subproblem. It is a smooth optimization problem on Riemannian manifold. The X-subproblem could be restated as follows

$$\min_X \psi(X), \quad \text{s.t. } X \in \mathcal{M}. \tag{4.15}$$

where $\psi = \psi_{\bar{Z}}$ is given by (4.7). It is a retraction smooth function, so problem (4.6) can be solved by some Riemannian gradient methods including Riemannian gradient descent (RGD), Riemannian conjugate gradient (RCG) and Riemannian trust region (RTR) method, etc. In this paper, we adopt a RGD method to problem (4.15), see Algorithm 2 for details.

Algorithm 1 Manifold Inexact augmented Lagrangian method for problem (1.1)

Input: Let $Z_{\min} < Z_{\max}$, $X_0 \in \mathcal{M}$, $\bar{Z}_0 \in \mathbb{R}^{d\times r}$, tolerance $\epsilon_{\min} \geq 0$, $\epsilon_0 > 0$, $\rho_0 > 1$, $\sigma > 1, 0 < \tau < 1$.

for $k = 0, 1, \cdots$ **do**

Compute (X^{k+1}, Y^{k+1}) by solving the following problem within a tolerance ϵ_k,

$$\min_{X\in\mathcal{M}} \psi_{\bar{Z}^k}(X), \tag{4.11}$$

where $\{\epsilon_k\}_{k\in\mathbb{N}} \downarrow 0$, and

$$Y^{k+1} = \text{Prox}_{g/\rho_k}(AX^{k+1} - \bar{Z}^k). \tag{4.12}$$

Update Lagrangian multiplier Z^{k+1} by

$$Z^{k+1} = \bar{Z}^k - \rho_k(AX^{k+1} - Y^{k+1}) \tag{4.13}$$

Project Z^{k+1} onto $\{Z : Z_{\min} \leq Z \leq Z_{\max}\}$ and denoted by $\bar{Z}^{k+1}$.

Update penalty parameter by

$$\rho_{k+1} = \begin{cases} \rho_k, & \text{if } \|AX^{k+1} - Y^{k+1}\|_\infty \leq \tau\|AX^k - Y^k\|_\infty \\ \sigma\rho_k, & \text{otherwise} \end{cases} \tag{4.14}$$

end for

5 Convergence analysis

For convenience of notation, we rewrite original problem (4.1) to a standard constraint optimization problem on manifold. Specifically, let $W = [X; Y] \in \mathbb{R}^{(n+d)\times r}$, and $\mathcal{N} = \mathcal{M}\times\mathbb{R}^{d\times r}$ be a product manifold. Then, problem (4.1) can be rewritten to

$$\min_{W} \ \theta(W), \ \text{s.t.} \ \ h(W) = 0, \ W \in \mathcal{N}. \tag{5.1}$$

where $\theta(W) = f(X) + g(Y)$, and $h(W) = [A, -I]W \in \mathbb{R}^{d\times r}$. The augmented Lagrangian function associated to problem (5.1) is

$$\mathcal{L}_\rho(W; Z) = \theta(W) + \sum_{i=1}^{d}\sum_{j=1}^{r} Z_{ij}[h(W)]_{ij} + \frac{\rho}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}[h(W)]_{ij}^2 \tag{5.2}$$

The KKT condition can be given by

$$0 \in \partial\theta(W^*) + \sum_{i=1}^{d}\sum_{j=1}^{r} Z_{ij}\text{grad}[h(W^*)]_{ij}, \quad h(W^*) = 0, \quad W^* \in \mathcal{N}, \tag{5.3}$$

where $\partial\theta(W^*)$ is Riemannian subdifferential of θ at W^*. The KKT condition (5.3) is identical to (4.3) because that $\mathcal{M}$ is a Riemannian submanifold embedded in Euclidean space. Inspired

Algorithm 2 Riemannian gradient method for subproblem (4.15)

1: **Given:** $X^0 \in \mathcal{M}$, tolerance $\epsilon > 0$. Let $\eta^0 = -\text{grad}\psi(X^0)$.
2: **Initialize:** k = 0.
3: **while** $\|\eta^k\| \geq \epsilon$ **do**
4: Pick $\eta^k = -\text{grad}\psi(X^k)$ and a step size α_k, compute

$$X^{k+1} = \mathrm{R}_{X^k}(\alpha_k \eta^k). \tag{4.16}$$

5: **end while**

by Zhang, Yang and Song [35], the constraint qualifications of problem (5.1) is given by:

Definition 5.1 (LICQ)**.** Linear independence constraint qualifications (LICQ) are said to hold at $W^* \in \mathcal{N}$ for problem (5.1) if

$$\{\text{grad}[h(W^*)]_{ij} | i = 1, \cdots, d; j = 1, \cdots, r\} \text{ are linearly independent in } T_{W^*}\mathcal{N}. \tag{5.4}$$

We will analyze the convergence of Algorithm 1 in the following two cases:

1) The iterate (X^{k+1}, Y^{k+1}) is an ϵ_k-stationary point of iteration subproblem, i.e.,

$$\|\text{grad}\psi_{\bar{Z}^k}(X^{k+1})\| \leq \epsilon_k. \tag{5.5}$$

2) The iterate (X^{k+1}, Y^{k+1}) is an ϵ_k-global minimizer of iteration subproblem, i.e.,

$$\mathcal{L}_{\rho_k}(W^{k+1}; \bar{Z}^k) \leq \mathcal{L}_{\rho_k}(W; \bar{Z}^k) + \epsilon_k, \ \forall \ W \in \mathcal{N}. \tag{5.6}$$

Remark 5.1. In the case 1), (5.5) is indeed to find W^{k+1} such that

$$\delta^k \in \partial \mathcal{L}_{\rho_k}(W^{k+1}; \bar{Z}^k), \ \|\delta^k\| \leq \epsilon_k.$$

Theorem 5.1. *Suppose $\{W^k\}_{k\in\mathbb{N}}$ is a sequence generated by Algorithm 1, Assumption 4.1 holds, and* (5.5) *holds for all $k \geq 0$.Then, sequence $\{W^k\}_{k\in\mathbb{N}}$ has at least one cluster point. Furthermore, if W^* is a cluster point, and LICQ holds at W^*, then W^* is a KKT point of problem* (5.1)*.*

Proof. To prove the first part of Theorem 5.1, we need to show that sequence $\{W^k\}_{k\in\mathbb{N}}$ is bounded. By Assumption 4.1, $\mathcal{M}$ is a compact manifold, hence $\{X^k\}$ is bounded. By

$$Y^{k+1} = \text{Prox}_{g/\rho}(AX^{k+1} - \frac{1}{\rho}\bar{Z}^k),$$

there exists $\nu^k \in \partial g(Y^{k+1})$ such that

$$0 = \nu^k - \rho_k(AX^{k+1} - \frac{1}{\rho}\bar{Z}^k - Y^{k+1}).$$

Again using Assumption 4.1, $\partial g(Y^{k+1})$ is bounded, and hence ν^k is also bounded. It is obvious that $\bar{Z}^k \in [Z_{min}, Z_{max}]$ is bounded. Since sequence $\{\rho_k\}_{k\in\mathbb{N}}$ is nondescreasing, we get $\rho_k \geq \rho_0$ ($\forall k \in \mathbb{N}$). Hence $\{Y^k\}_{k\in\mathbb{N}}$ is bounded. In summary, sequence $\{W^k\}_{k\in\mathbb{N}}$ is bounded.

Next, we will show that W^* is a feasible point of (5.1). By the updating rule of W in Algorithm 1, we have $W^k \in \mathcal{N}$.

If $\{\rho_k\}_{k\in\mathbb{N}}$ is bounded, by the updating rule of ρ_k, there exists a $k_0 \in \mathbb{N}$ such that

$$\|h(W^k)\|_\infty \leq \tau \|h(W^{k-1})\|_\infty, \quad \forall k \geq k_0,$$

where $\tau \in (0,1)$. Hence, $h(W^*) = 0$.

If $\{\rho_k\}$ is unbounded, by Remark 5.1 we have

$$\delta^k \in \partial \mathcal{L}_{\rho_k}(W^{k+1}; \bar{Z}^k), \ \|\delta^k\| \leq \epsilon_k.$$

where $\epsilon^k \downarrow 0$ as $k \to \infty$. Thus there exists $U^k \in \partial\theta(W^k)$ such that

$$U^k + \sum_{i=1}^{d}\sum_{j=1}^{r} \left(\bar{Z}^k_{ij} + \rho_k[h(W^k)]_{ij}\right) \mathrm{grad}[h(W^k)]_{ij} = \delta^k. \tag{5.7}$$

Dividing both sides of (5.7) by ρ_k, we have

$$\sum_{i=1}^{d}\sum_{j=1}^{r} \left(\bar{Z}^k_{ij}/\rho_k + [h(W^k)]_{ij}\right) \mathrm{grad}[h(W^k)]_{ij} = (\delta^k - U^k)/\rho_k \tag{5.8}$$

where $\{\bar{Z}^k\}$ is bounded, and $\delta^k \downarrow 0$. Since $\theta(W) = f(X) + g(Y)$, where g is a convex function on $\mathbb{E}$, and

$$\partial\theta(W) = \begin{pmatrix} \mathrm{grad} f(X) \\ \partial g(Y) \end{pmatrix},$$

where $\partial g(Y)$ is a subdifferential (set) in usual sense. Invoked by Proposition B.24(b) in [7], the set $\bigcup_{k\in\mathcal{K}} \partial g(Y^k)$ is bounded because that $\{Y^k\}_{k\in\mathcal{K}}$ is a bounded set. In addition, $f(X)$ is a retraction smooth function, hence the Riemannian gradient sequence $\{\mathrm{grad} f(X^k)\}_{k\in\mathcal{K}}$ is bounded. Thus, we have that $\bigcup_{k\in\mathcal{K}} \partial\theta(W^k)$ is bounded. This means that $\{U^k\}$ is bounded. Taking limits as $k \in \mathcal{K}$ going to infinity on both sides of (5.7), and using the continuity and differentiability of h, we have,

$$\sum_{i=1}^{d}\sum_{j=1}^{r} ([h(W^*)]_{ij}) \,\mathrm{grad}[h(W^*)]_{ij} = 0 \tag{5.9}$$

Note that LICQ holds at W^*, we conclude that $[h(W^*)]_{ij} = 0$ for all i, j.

Since $\{U^k\}_{k\in\mathcal{K}}$ is bounded, there exists a subsequence $\mathcal{K}_1 \subset \mathcal{K}$ such that $\lim_{k\to\infty, k\in\mathcal{K}_1} U^k = U^*$. Recall that $\lim_{k\to\infty, k\in\mathcal{K}_1} W^k = W^*$. We get

$$U^* \in \partial\theta(W^*)$$

by the closedness property of the limiting subdifferential. Together with $Z_{ij}^{k+1} = \bar{Z}^k + \rho_k [h(W^k)]_{ij}$ for all i, j, one can get from Algorithm 1 that, for all $k \in \mathcal{K}_1$,

$$U^k + \sum_{i=1}^{d}\sum_{j=1}^{r} \left(Z_{ij}^{k+1}\right) \operatorname{grad}[h(W^k)]_{ij} = \delta^k \tag{5.10}$$

where δ^k satisfying $\|\delta^k\| \le \epsilon^k$, and $U^k \in \partial\theta(W^k)$.

We claim that $\{Z^k\}$ is bounded. Otherwise, assume $\{Z^k\}$ is unbounded, we have

$$\frac{U^k}{\|Z^{k+1}\|_\infty} + \sum_{i=1}^{d}\sum_{j=1}^{r} \left(\frac{Z_{ij}^{k+1}}{\|Z^{k+1}\|_\infty}\right) \operatorname{grad}[h(W^k)]_{ij} = \frac{\delta^k}{\|Z^{k+1}\|_\infty}$$

Since $\frac{Z^{k+1}}{\|Z^{k+1}\|_\infty} \in [-1, 1]$ is bounded, there exists a subsequence $\mathcal{K}_2 \subset \mathcal{K}_1$ such that $\lim_{k\to\infty, k\in\mathcal{K}_2} \frac{Z^{k+1}}{\|Z^{k+1}\|_\infty} = \bar{Z}$, where $\bar{Z}$ is a nonzero matrix. Taking the limit on $k \in \mathcal{K}_2$ going to infinity, we obtain

$$\sum_{i=1}^{d}\sum_{j=1}^{r} \bar{Z}_{ij} \operatorname{grad}[h(W^*)]_{ij} = 0, \tag{5.11}$$

which contradicts the LICQ condition at W^*.

Since $\{U^k\}$ is bounded and $\{\delta^k\} \downarrow 0$, there exists a subsequence $\mathcal{K}_3 \subset \mathcal{K}_2$ such that $\lim_{k\to\infty, k\in\mathcal{K}_3} U^k = U^*$ and $\lim_{k\to\infty, k\in\mathcal{K}_3} Z^k = Z^*$. By the continuity of mapping grad h, and taking limits on $k \in \mathcal{K}_3$ going to infinity on both sides of (5.10), we have

$$U^* + \sum_{i=1}^{d}\sum_{j=1}^{r} \left(Z_{ij}^*\right) \operatorname{grad}[h(W^*)]_{ij} = 0. \tag{5.12}$$

□

Lemma 5.1. *Suppose that $W \in \mathcal{N} = \mathcal{M} \times R^{d\times r}$ where $\mathcal{M}$ is a stiefel manifold, denoted by $St(n, r)$). Then the LICQ always holds at $\forall\ W \in \mathcal{N}$.*

Proof. Let $e_i \in \mathbb{R}^d$ be a m-dimensional vector in which the i-th entry is 1 and 0 for others, and $\bar{e}_j \in \mathbb{R}^r$ be a r-dimensional vector in which the j-th entry is 1 and 0 for others. Then

$$\nabla[h(W)]_{ij} = \begin{pmatrix} A^T e_i \bar{e}_j^T \\ -e_i \bar{e}_j^T \end{pmatrix}, \quad i = 1, \cdots, d; j = 1, \cdots, r.$$

A basis of the normal cone of $St(n,r)$ at X, denoted by $N_X St(n,r)$, is given by

$$\left\{X(\bar{e}_i\bar{e}_j^T + \bar{e}_j\bar{e}_i^T) : i = 1,\cdots,r, j = 1,\cdots,r\right\}.$$

It is easy to show that, for $\forall\ W \in \mathcal{N}$, all the vectors in the set

$$\left\{\begin{pmatrix} A^T e_i\bar{e}_j^T \\ -e_i\bar{e}_j^T \end{pmatrix}, i = 1,\cdots,d; j = 1,\cdots,r.\right\}\bigcup\left\{\begin{pmatrix} X(\bar{e}_i\bar{e}_j^T + \bar{e}_j\bar{e}_i^T) \\ 0 \end{pmatrix}, i = 1,\cdots,r; j = 1,\cdots,r.\right\}$$

are linearly independent. Hence, if there exists Z such that

$$\sum_{i=1}^{d}\sum_{j=1}^{r} Z_{ij}\nabla[h(W)]_{ij} \in N_W\mathcal{N}, \tag{5.13}$$

then we have $Z = 0$. By assumptions in this lemma, $\mathcal{N}$ is a submanifold of Euclidean space. So, it derives immediately from (5.13) that

$$\sum_{i=1}^{d}\sum_{j=1}^{r} Z_{ij}\mathrm{grad}[h(W)]_{ij} = 0.$$

Which implies LICQ holds at W and completes the proof. □

We consider the case that a ϵ_k-global minimizer of the iteration subproblem could be obtained at each iteration of Algorithm 1.

Theorem 5.2. *Assume that $\{W^k\}_{k\in\mathbb{N}}$ is a sequence generated by Algorithm 1, Assumption 4.1 holds, and (5.6) is satisfied at each iteration of Algorithm 1. Let W^* be a limit point of $\{W^k\}_{k\in\mathbb{N}}$. Then one has*

$$\sum_{i=1}^{d}\sum_{j=1}^{r}[h(W^*)]_{ij}^2 \leq \sum_{i=1}^{d}\sum_{j=1}^{r}[h(W)]_{ij}^2,\ \forall\ W \in \mathcal{N}. \tag{5.14}$$

Proof. Consider the following two cases: $\{\rho_k\}$ bounded and $\rho_k \to \infty$.

If $\{\rho_k\}$ is bounded, then there exists k_0 such that $\rho_k = \rho_{k_0}$ for all $k \geq k_0$. Hence

$$\sum_{i=1}^{d}\sum_{j=1}^{r}[h(W^{k+1})]_{ij}^2 \leq \tau\sum_{i=1}^{d}\sum_{j=1}^{r}[h(W^k)]_{ij}^2, i = 1,\cdots,m; j = 1,\cdots,r.$$

Which implies that $h(W^k) \to 0$ as $k \to \infty$. We have $h(W^*) = 0$, and (5.14) holds.

Then to the case $\rho_k \to \infty$. Since W^* is a limit point of $\{W^k\}$, there exists a subsequence $\mathcal{K} \subset \mathbb{N}$ such that

$$\lim_{k\to\infty,\ k\in\mathcal{K}} W^k = W^*.$$

Assume by contradiction there exists $W \in \mathcal{N}$ such that

$$\sum_{i=1}^{d}\sum_{j=1}^{r}[h(W^*)]_{ij}^2 \geq \sum_{i=1}^{d}\sum_{j=1}^{r}[h(W)]_{ij}^2,$$

By the boundedness of $\{\bar{Z}^k\}$ and $\rho_k \to \infty$, there exist $c > 0$ and $k_0 \in \mathbb{N}$ such that, for all $k \in \mathcal{K}$ and $k \geq k_0$ we have

$$\sum_{i=1}^{d}\sum_{j=1}^{r}([h(W^{k+1})]_{ij} + \frac{1}{\rho_k}\bar{Z}_{ij}^k)^2 \geq \sum_{i=1}^{d}\sum_{j=1}^{r}([h(W)]_{ij} + \frac{1}{\rho_k}\bar{Z}_{ij}^k)^2 + c.$$

Therefore

$$\theta(W^{k+1}) + \frac{\rho_k}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}([h(W^{k+1})]_{ij} + \frac{1}{\rho_k}\bar{Z}_{ij}^k)^2 \geq \theta(W) + \frac{\rho_k}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}([h(W)]_{ij} + \frac{1}{\rho_k}\bar{Z}_{ij}^k)^2 + \frac{\rho_k c}{2} + \theta(W^{k+1}) - \theta(W)$$

Since $\lim\limits_{k\to\infty, k\in\mathcal{K}} W^k = W^*$, and $\{\epsilon_k\}$ is bounded, there exists $k_1 > k_0$ such that, for all $k \in \mathcal{K}, k \geq k_1$ we have

$$\frac{\rho_k c}{2} + \theta(W^{k+1}) - \theta(W) > \epsilon_k.$$

Therefore,

$$\theta(W^{k+1}) + \frac{\rho_k}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}([h(W^{k+1})]_{ij} + \frac{1}{\rho_k}\bar{Z}_{ij}^k)^2 \geq \theta(W) + \frac{\rho_k}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}([h(W)]_{ij} + \frac{1}{\rho_k}\bar{Z}_{ij}^k)^2 + \epsilon_k.$$

This contradicts (5.6), hence we have (5.14) and complete the proof. □

Theorem 5.3. *In Algorithm 1, let $\epsilon_{\min} = 0$ and W^* be a limit point of sequence $\{W^k\}_{k\in\mathbb{N}}$. If iterate W^{k+1} is a ϵ_k- global minimizer satisfying* (5.6)*, then W^* is a global minimizer of problem* (4.1)*. Meanwhile, X^* is a global minimizer of problem* (1.1)*.*

Proof. By (5.6), we have

$$\theta(W^{k+1}) + \frac{\rho_k}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}([h(W^{k+1})]_{ij} + \frac{1}{\rho_k}\bar{Z}_{ij}^k)^2 \leq \theta(W) + \frac{\rho_k}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}([h(W)]_{ij} + \frac{1}{\rho_k}\bar{Z}_{ij}^k)^2 + \epsilon_k$$

for all $W \in \mathcal{N}$. Since $h(W) = 0$, we arrive that

$$\theta(W^{k+1}) + \frac{\rho_k}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}([h(W^{k+1})]_{ij} + \frac{1}{\rho_k}\bar{Z}_{ij}^k)^2 \leq \theta(W) + \frac{\rho_k}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}(\frac{1}{\rho_k}\bar{Z}_{ij}^k)^2 + \epsilon_k.$$

Which implies that

$$\theta(W^{k+1}) \leq \theta(W) + \frac{\rho_k}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}(\frac{1}{\rho_k}\bar{Z}_{ij}^k)^2 + \epsilon_k. \tag{5.15}$$

If $\rho_k \to \infty$, by taking limits on both sides of (5.15) as $k \in \mathcal{K}, k \to \infty$, and using $\lim\limits_{k\to\infty, k\in\mathcal{K}} \epsilon_k = 0$, we get

$$\theta(W^*) \leq \theta(W),\ \forall\ W \in \mathcal{N}.$$

In case of that $\{\rho_k\}$ is bounded, there exists $k_0 \in \mathbb{N}$ such that $\rho_k = \rho_{k_0}$ for all $k > k_0$. By (5.6) we have

$$\theta(W^{k+1}) + \frac{\rho_{k_0}}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}([h(W^{k+1})]_{ij} + \frac{1}{\rho^{k_0}}\bar{Z}_{ij}^k)^2 \leq \theta(W) + \frac{\rho_{k_0}}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}([h(W)]_{ij} + \frac{1}{\rho_{k_0}}\bar{Z}_{ij}^k)^2 + \epsilon_k$$

for $W \in \mathcal{N}$. Since $h(W) = 0$, we get

$$\theta(W^{k+1}) + \frac{\rho_{k_0}}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}([h(W^{k+1})]_{ij} + \frac{1}{\rho^{k_0}}\bar{Z}_{ij}^k)^2 \leq \theta(W) + \frac{\rho_{k_0}}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}(\frac{1}{\rho_{k_0}}\bar{Z}_{ij}^k)^2 + \epsilon_k \tag{5.16}$$

for all $k \geq k_0$. Let $\mathcal{K}_1 \subset \mathcal{K}$, and

$$\lim_{k\to\infty, k\in\mathcal{K}_1} \bar{Z}^k = Z^*.$$

Taking limits on both sides of (5.16) as $k \to \infty, k \in \mathcal{K}_1$, and noting that $h(W^*) = 0$, we get

$$\theta(W^*) + \frac{\rho_{k_0}}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}(\frac{1}{\rho_{k_0}}Z_{ij}^*)^2 \leq \theta(W) + \frac{\rho_{k_0}}{2}\sum_{i=1}^{d}\sum_{j=1}^{r}(\frac{1}{\rho_{k_0}}Z_{ij}^*)^2.$$

Hence

$$\theta(W^*) \leq \theta(W),\ \forall\ W \in \mathcal{N},$$

and the proof is completed. □

6 Experiments

Numerical experiments for testing the performance of the proposed MIALM method, with compared to some existing methods including SOC [30], PAMAL [22], MADMM [29] and ManPG [14], are presented in the current section. All the methods are used to solve the compressed modes and sparse PCA problem. In the MIALM and MADMM, the Riemannian manifold optimization subproblem is handled by "Manopt", a Matlab toolbox for optimization on manifolds [11]. In the SOC, PAMAL and ManPG methods, the code provided by Chen [14] are used (all codes are available in online). All experiments are run on a personal computer having 4.0GHz Intel Core i7 CPU and 16 GB RAM.

6.1 Compressed modes in Physics

In physics, the compressed modes problem (CMs) seeks spatially localized solutions of the independent-particle Schrödinger equation:

$$\hat{H}\phi(x) = \lambda\phi(x), \quad x \in \Omega, \tag{6.1}$$

where $\hat{H} = -\frac{1}{2}\Delta$ and Δ is a Laplacian operator. Consider the $1D$ free-electron (FE) model, where $\hat{H} = -\frac{1}{2}\partial_{x^2}$. By a proper discretization, the compressed modes problem can be reformulated to

$$\begin{cases} \min\limits_{X\in\mathbb{R}^{n\times k}} tr(X^T HX) + \mu\|X\|_1, \\ \text{s.t.} \quad X^T X = I_d, \end{cases} \tag{6.2}$$

where H denotes the discretized Schrödinger operator, μ is a regularization parameter. The interesting readers are referred to [33] for more details. For problem (6.2), both SOC and PAMAL consider the identical form as follows:

$$\begin{cases} \min\limits_{\Psi,Q,P\in\mathbb{R}^{n\times r}} tr(X^T HX) + \mu\|Q\|_1, \\ \text{s.t.} \quad Q = X, P = X, P^T P = I_r. \end{cases} \tag{6.3}$$

The MADMM handles the separable reformulation of the form

$$\begin{cases} \min\limits_{\Psi,Q\in\mathbb{R}^{n\times r}} tr(X^T HX) + \mu\|Q\|_1, \\ \text{s.t.} \quad Q = X, X^T X = I_r. \end{cases} \tag{6.4}$$

In our experiments, the domain $\Omega := [0, 50]$ is discretized with n equally spaced nodes. The parameters of MIALM are set to : $\tau = 0.99, \sigma = 1.05, \rho_0 = \lambda_{\max}(H)/2, Z_{\min} = -100\cdot 1_{d\times r}, Z_{\max} = 100\cdot 1_{d\times r}, Z^0 = 0_{d\times r}$ and $\epsilon_k = \max(10^{-5}, 0.9^k)$, where $k \in \mathbb{N}$ is the iteration counter. We terminated MIALM if $\|X^k - Y^k\|_F^2 \le 10^{-9}$ or $k \ge 500$. For the inner iteration of the MIALM, a Barzilai-Borwein stepsize is used to accelerate, and it is terminated if $\|\text{grad}\Psi_{\bar{Z}^k}(X)\|_X \le \epsilon_k$ or the inner iteration number exceeds 20. The final objective value obtained by the MIALM method is denoted by F_M.

For the MADMM, the penalty parameter is set to $\rho = \lambda_{\max}(H)/2$. We terminated MADMM if $\|X^k - Y^k\|_F^2 \le 10^{-9}$ or $F(X^k) \le F_M + 10^{-7}$, or $k \ge 500$. The inner iteration of the MADMM terminates if the norm of Riemannian gradient of X-subproblem is less than 10^{-5} or the inner iteration number exceeds 20. For the SOC, PAMAL and ManPG, the parameters are set to as same as in [14], except that the penalty parameter $\rho = 2\lambda_{\max}(H)$ in SOC and PAMAL. The ManPG terminates if stopping criterion described in [14] is met or $F(X^k) \le F_M + 10^{-7}$. For easy comparison, Table 1 lists the objective function value, sparsity of solution and cpu time. One can find from Table 1 that, our MILAM method outperforms to the other methods.

Table 1: Comparisons of MIALM and ManPG, MADMM, PAMAL, SOC on CMs problem

	μ	MIALM			ManPG			MADMM		
		time	F_M	sp	time	F	sp	time	F	sp
$r = 2, n = 128$	0.1	0.021	0.943	0.835	0.036	0.943	0.836	0.112	0.943	0.836
	0.2	0.016	1.639	0.881	0.024	1.639	0.882	0.024	1.639	0.882
	0.3	0.020	2.265	0.901	0.029	2.265	0.900	0.167	2.265	0.903
	μ	PAMAL			SOC					
		time	F	sp	time	F	sp			
	0.1	0.049	0.943	0.837	0.024	0.943	0.837			
	0.2	0.038	1.639	0.882	0.017	1.639	0.882			
	0.3	0.088	2.265	0.901	0.026	2.265	0.901			

	r	MIALM			ManPG			MADMM		
		time	F_M	sp	time	F	sp	time	F	sp
$\mu = 0.2, n = 256$	2	0.021	2.167	0.892	0.071	2.167	0.892	0.153	2.167	0.892
	4	0.063	4.334	0.887	0.233	4.334	0.886	0.311	4.338	0.884
	6	0.345	6.500	0.889	0.722	6.500	0.884	0.531	6.509	0.881
	r	PAMAL			SOC					
		time	F	sp	time	F	sp			
	2	0.127	2.167	0.892	0.057	2.167	0.892			
	4	0.709	4.334	0.888	0.273	4.334	0.888			
	6	3.036	6.500	0.887	0.980	6.500	0.887			

	n	MIALM			ManPG			MADMM		
		time	F_M	sp	time	F	sp	time	F	sp
$\mu = 0.6, r = 2$	200	0.018	2.265	0.901	0.028	2.265	0.901	0.167	2.265	0.903
	300	0.017	2.996	0.910	0.051	2.996	0.910	0.128	3.005	0.909
	500	0.026	3.956	0.920	0.132	3.956	0.920	0.282	4.048	0.916
	n	PAMAL			SOC					
		time	F	sp	time	F	sp			
	200	0.045	2.265	0.902	0.028	2.265	0.901			
	300	0.085	2.996	0.910	0.041	2.996	0.910			
	500	0.253	3.956	0.920	0.137	3.956	0.920			

6.2 Sparse principle component analysis

Given a data set $\{b_1, \cdots, b_m\}$ where $b_i \in \mathbb{R}^{n\times 1}$. The sparse PCA problem is

$$\begin{cases} \min\limits_{X\in\mathbb{R}^{n\times r}} \sum\limits_{i=1}^{m} \|b_i - XX^T b_i\|_2^2 + \mu\|X\|_1, \\ \text{s.t.} \quad X^T X = I_r, \end{cases} \tag{6.5}$$

where μ is a regularization parameter. Let $B = [b_1, \cdots, b_m]^T \in \mathbb{R}^{m\times n}$, problem (6.5) has the form:

$$\begin{cases} \min\limits_{X\in\mathbb{R}^{n\times r}} -tr(X^T B^T BX) + \mu\|X\|_1, \\ \text{s.t.} \quad X^T X = I_r. \end{cases} \tag{6.6}$$

In our experiments, the random data matrix $B \in \mathbb{R}^{m\times n}$ is generated by MATLAB function $randn(m, n)$, in which the entries of B follow the standard Gaussian distribution. We shift the columns of B such that thier mean equal to 0, and finally the column-vectors are normalized. The parameters of MIALM are set to the same as that used the CMs problem, except that the stopping criterion is modified to $\|X^k - Y^k\|_F^2 \leq 10^{-8}$, the penalty parameter $\rho_0 = \lambda_{\max}^2(B^T B)/2$. Similarly, the parameters of the MADMM are also set as the same as that used the CMs problem, except that the penalty parameter $\rho_0 = \lambda_{max}^2(B^T B)/2$. For the SOC, PAMAL and ManPG methods, the stopping criterion and parameter settings provided in [14] are copied. The interesting readers are referred to [14] for details. Table 2 lists performance of all methods on the sparse PCA problem for comparisions.

7 Conclusions

We propose a manifold inexact augmented Lagrangian method for nonsmooth composite minimization problem on Riemannian manifold. In each iteration of the proposed mehod, a smooth Riemannian manifold minimization subproblem is obtained by utilizing the Moreau envelope. The convergence of the proposed method is established under some suitable assumptions. Numerical experiments show that, the proposed method is competitive compared to some existing state-of-the-art methods.

References

[1] P.-A. Absil, C. G. Baker, and K. A. Gallivan. Trust-region methods on riemannian manifolds. *Foundations of Computational Mathematics*, 7(3):303–330, 2007.

Table 2: Comparisons of MIALM and ManPG, MADMM, PAMAL, SOC on SPCA ($m = 50$)

	μ	MIALM			ManPG			MADMM		
		time	F_M	sp	time	F	sp	time	F	sp
$r = 2, n = 200$	0.5	0.038	-6.839	0.461	0.035	-6.819	0.458	0.193	-6.767	0.454
	0.6	0.038	-5.304	0.543	0.042	-5.248	0.545	0.201	-5.147	0.539
	0.8	0.043	-2.439	0.722	0.047	-2.369	0.732	0.199	-2.285	0.732
	μ	PAMAL			SOC					
		time	F	sp	time	F	sp			
	0.5	1.919	-6.847	0.460	0.251	-6.826	0.458			
	0.6	2.123	-5.267	0.545	0.302	-5.262	0.544			
	0.8	2.247	-2.387	0.733	0.281	-2.371	0.732			

	r	MIALM			ManPG			MADMM		
		time	F_M	sp	time	F	sp	time	F	sp
$\mu = 0.6, n = 200$	2	0.040	-5.308	0.548	0.039	-5.290	0.547	0.199	-5.209	0.538
	3	0.047	-7.563	0.562	0.058	-7.530	0.561	0.223	-7.369	0.552
	5	0.091	-11.625	0.594	0.117	-11.571	0.591	0.291	-11.304	0.582
	r	PAMAL			SOC					
		time	F	sp	time	F	sp			
	2	0.040	-5.308	0.548	0.251	-5.329	0.544			
	3	3.322	-7.597	0.562	0.442	-7.552	0.561			
	5	6.828	-11.687	0.592	0.674	-11.727	0.588			

	n	MIALM			ManPG			MADMM		
		time	F_M	sp	time	F	sp	time	F	sp
$\mu = 0.6, r = 2$	200	0.039	-5.323	0.539	0.040	-5.283	0.541	0.203	-5.166	0.538
	300	0.048	-8.128	0.473	0.043	-8.112	0.473	0.227	-7.955	0.467
	500	0.085	-14.139	0.399	0.054	-14.134	0.399	0.303	-13.698	0.385
	n	PAMAL			SOC					
		time	F	sp	time	F	sp			
	200	2.187	-5.282	0.545	0.288	-5.290	0.542			
	300	3.037	-8.106	0.477	0.443	-8.108	0.474			
	500	9.618	-14.106	0.400	1.283	-14.109	0.398			

[2] P.-A. Absil and S. Hosseini. A collection of nonsmooth riemannian optimization problems. In *Nonsmooth Optimization and Its Applications*, pages 1–15. Springer, 2019.

[3] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[4] A. Beck. *First-order methods in optimization*, volume 25. SIAM, 2017.

[5] 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(2):548–562, 2017.

[6] G. d. C. Bento, J. C. d. L. Neto, and P. R. Oliveira. Convergence of inexact descent methods for nonconvex optimization on riemannian manifolds. *arXiv preprint arXiv:1103.4828*, 2011.

[7] D. P. Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

[8] M. Bortoloti, T. A. Fernandes, O. P. Ferreira, and J. Yuan. Damped newton's method on riemannian manifolds. *arXiv preprint arXiv:1803.05126*, 2018.

[9] N. Boumal. Riemannian trust regions with finite-difference hessian approximations are globally convergent. In *International Conference on Geometric Science of Information*, pages 467–475, 2015.

[10] N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2018.

[11] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research*, 15(1):1455–1459, 2014.

[12] J. V. Burke, F. E. Curtis, A. S. Lewis, M. L. Overton, and L. E. Simões. Gradient sampling methods for nonsmooth optimization. *arXiv preprint arXiv:1804.11003*, 2018.

[13] L. Cambier and P.-A. Absil. Robust low-rank matrix completion by riemannian optimization. *SIAM Journal on Scientific Computing*, 38(5):S440–S460, 2016.

[14] S. Chen, S. Ma, A. M.-C. So, and T. Zhang. Proximal gradient method for manifold optimization. *arXiv preprint arXiv:1811.00980*, 2018.

[15] W. Chen, H. Ji, and Y. You. An augmented lagrangian method for 1-regularized optimization problems with orthogonality constraints. *SIAM Journal on Scientific Computing*, 38(4):B570–B592, 2016.

[16] 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(3):743–755, 2016.

[17] N. K. Dhingra, S. Z. Khong, and M. R. Jovanovic. The proximal augmented lagrangian method for nonsmooth composite optimization. *IEEE Transactions on Automatic Control*, 2018.

[18] O. P. Ferreira, M. S. Louzeiro, and L. Prudente. Iteration-complexity of the subgradient method on riemannian manifolds with lower bounded curvature. *Optimization*, 68(4):713–729, 2019.

[19] R. H. Gohary and T. N. Davidson. Noncoherent mimo communication: Grassmannian constellations and efficient detection. *IEEE Transactions on Information Theory*, 55(3):1176–1205, 2009.

[20] P. Grohs and S. Hosseini. Nonsmooth trust region algorithms for locally lipschitz functions on riemannian manifolds. *IMA Journal of Numerical Analysis*, 36(3):1167–1192, 2015.

[21] P. Grohs and S. Hosseini. ε-subgradient algorithms for locally lipschitz functions on riemannian manifolds. *Advances in Computational Mathematics*, 42(2):333–360, 2016.

[22] Z. Hong, X. Zhang, D. Chu, and L. Liao. Nonconvex and nonsmooth optimization with generalized orthogonality constraints: An approximate augmented lagrangian method. *Journal of Scientific Computing*, 72(1):1–42, 2017.

[23] S. Hosseini. Convergence of nonsmooth descent methods via kurdyka–lojasiewicz inequality on riemannian manifolds. *Hausdorff Center for Mathematics and Institute for Numerical Simulation, University of Bonn (2015,(INS Preprint No. 1523))*, 2015.

[24] S. Hosseini, W. Huang, and R. Yousefpour. Line search algorithms for locally lipschitz functions on riemannian manifolds. *SIAM Journal on Optimization*, 28(1):596–619, 2018.

[25] S. Hosseini and A. Uschmajew. A riemannian gradient sampling algorithm for nonsmooth optimization on manifolds. *SIAM Journal on Optimization*, 27(1):173–189, 2017.

[26] W. Huang. Optimization algorithms on riemannian manifolds with applications. 2013.

[27] W. Huang, P.-A. Absil, and K. A. Gallivan. A riemannian symmetric rank-one trust-region method. *Mathematical Programming*, 150(2):179–216, 2015.

[28] W. Huang, K. A. Gallivan, and P.-A. Absil. A broyden class of quasi-newton methods for riemannian optimization. *SIAM Journal on Optimization*, 25(3):1660–1685, 2015.

[29] A. Kovnatsky, K. Glashoff, and M. M. Bronstein. Madmm: a generic algorithm for non-smooth optimization on manifolds. In *European Conference on Computer Vision*, pages 680–696. Springer, 2016.

[30] R. Lai and S. Osher. A splitting method for orthogonality constrained problems. *Journal of Scientific Computing*, 58(2):431–449, 2014.

[31] X. Li, D. Sun, and K.-C. Toh. A highly efficient semismooth newton augmented lagrangian method for solving lasso problems. *SIAM Journal on Optimization*, 28(1):433–458, 2018.

[32] M. Lin, Y.-J. Liu, D. Sun, and K.-C. Toh. Efficient sparse semismooth newton methods for the clustered lasso problem. *SIAM Journal on Optimization*, 29(3):2026–2052, 2019.

[33] V. Ozoliņš, R. Lai, R. Caflisch, and S. Osher. Compressed modes for variational problems in mathematics and physics. *Proceedings of the National Academy of Sciences*, 110(46):18368–18373, 2013.

[34] D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[35] W. H. Yang, L.-H. Zhang, and R. Song. Optimality conditions for the nonlinear programming problems on riemannian manifolds. *Pacific Journal of Optimization*, 10(2):415–434, 2014.

[36] X. Yuan, W. Huang, P.-A. Absil, and K. Gallivan. A riemannian quasi-newton method for computing the karcher mean of symmetric positive definite matrices. *Florida State University,(FSU17-02)*, 2017.

[37] H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638, 2016.

[38] J. Zhang, S. Ma, and S. Zhang. Primal-dual optimization algorithms over riemannian manifolds: an iteration complexity analysis. *arXiv preprint arXiv:1710.02236*, 2017.

[39] L. Zheng and D. N. C. Tse. Communication on the grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel. *IEEE Transactions on Information Theory*, 48(2):359–383, 2002.

[40] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.