Noname manuscript No.
(will be inserted by the editor)

A Semismooth Newton based Augmented Lagrangian Method for Nonsmooth Optimization on Matrix Manifolds

Yuhao Zhou · Chenglong Bao · Chao Ding · Jun Zhu

Received: date / Accepted: date

Abstract This paper is devoted to studying an augmented Lagrangian method for solving a class of manifold optimization problems, which have nonsmooth objective functions and non-negative constraints. Under the constant positive linear dependence condition on manifolds, we show that the proposed method converges to a stationary point of the nonsmooth manifold optimization problem. Moreover, we propose a globalized semismooth Newton method to solve the augmented Lagrangian subproblem on manifolds efficiently. The local superlinear convergence of the manifold semismooth Newton method is also established under some suitable conditions. We also prove that the semismoothness on submanifolds can be inherited from that in the ambient manifold. Finally, numerical experiments on compressed modes and (constrained) sparse principal component analysis illustrate the advantages of the proposed method.

Keywords Nonsmooth manifold optimization · Semismooth Newton method · Augmented Lagrangian method · Riemannian manifold

1 Introduction

Manifold optimization is recently growing in popularity as it naturally arises from various applications in many fields, including phase retrieval [17, 61], principal component analysis [49, 51], matrix completion [15, 60], medical image analysis [45], and deep learning [23]. It is concerned with optimization problems with a manifold constraint, and has been extensively studied when the objective function is smooth during the past decades [2, 33, 38, 39, 62, 66]. However, nonsmooth manifold optimization is less explored but has drawn increasing attention in recent years [19, 21, 42, 44]. In this paper, we consider the following nonsmooth and nonconvex manifold optimization problem:

$$\min_{x} \; \{f(x) + \psi(h_1(x))\}\,, \text{ s.t. } x \in \mathcal{M},\; h_2(x) \leq 0, \tag{1.1}$$

where $\mathcal{M}$ is a Riemannian manifold, $f : \mathcal{M} \to \mathbb{R}$, $h_1 : \mathcal{M} \to \mathbb{R}^m$, $h_2 : \mathcal{M} \to \mathbb{R}^q$ are continuously differentiable, and $\psi : \mathbb{R}^m \to \mathbb{R}$ is convex. Besides the nonsmooth function ψ in (1.1), the inequality constraint can be seen as the nonsmooth function $\delta_{\mathcal{C}}(x)$, where $\delta_{\mathcal{C}}$ is the indicator function of the set

The work was supported in part by National Natural Science Foundation of China (11901338, 61620106010). The research of the third author was supported in part by the National Natural Science Foundation of China (12071464) and the Beijing Natural Science Foundation (Z190002).

Yuhao Zhou
Department of Computer Science and Technology, Tsinghua University, China.

Chenglong Bao
Yau Mathematical Sciences Center, Tsinghua University, China and Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, China.

Chao Ding
Institute of Applied Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, China.

Jun Zhu
Department of Computer Science and Technology, Tsinghua University, China.

$\mathcal{C} := \{x : h_2(x) \leq 0\}$. From the perspective of algorithm design, the two nonsmooth terms in (1.1) are difficult to handle in general. On the other hand, the model (1.1) has many important applications in machine learning and scientific computing. We list some typical examples as follows and refer the readers to [1, 19, 38] for more examples and details.

1. **Compressed modes (CM) [52].** The CM problem seeks for sparse eigenfunctions to a class of Hamiltonian operators, as these localized spatial bases play an important role in representing the rapid varying functions in physics and quantum chemistry. Numerically, let H be a discretization of the Hamiltonian operator, then the CM problem is formulated as:
$$\min_{Q \in \mathrm{St}(n,r)} \left\{ \mathrm{tr}(Q^\top H Q) + \mu \|Q\|_1 \right\}, \tag{1.2}$$
where $\mathrm{St}(n,r) := \{Q \in \mathbb{R}^{n\times r} : Q^\top Q = I_r\}$ is the Stiefel manifold.
2. **Sparse principal component analysis (SPCA) [71].** Let $A \in \mathbb{R}^{p\times n}$ be a data matrix where n and p are the number of variables and the number of observations, respectively. Setting $f(Q) = -\mathrm{tr}(Q^\top A^\top A Q)$, $\psi(Q) = \mu \|Q\|_1$, $h_1(Q) = Q$, $h_2(Q) \equiv 0$ and $\mathcal{M} = \mathrm{St}(n,r)$, the SPCA problem has the form (1.1), i.e. it aims at solving
$$\min_{Q \in \mathrm{St}(n,r)} \left\{ -\mathrm{tr}(Q^\top A^\top A Q) + \mu \|Q\|_1 \right\}. \tag{1.3}$$
3. **Constrained SPCA [49].** To further enforce the orthogonality among principal components, the constrained SPCA problem imposes additional constraints on each column of Q, which has the form
$$\begin{aligned} \min_{Q \in \mathrm{St}(n,r)} \quad & \left\{ -\mathrm{tr}(Q^\top A^\top A Q) + \mu \|Q\|_1 \right\}, \\ \text{s.t.} \quad & |Q_i^\top A^\top A Q_j| \leq \mathit{\Delta}_{ij}, \ \forall\, i \neq j, \end{aligned} \tag{1.4}$$
where Q_i denotes the i-th column of Q and $\mathit{\Delta}_{ij} \geq 0$ are the predefined tolerances.

It is worth mentioning that (1.1) can be regarded as an unconstrained nonsmooth optimization problem on $\mathcal{M}$ when the inequality constraint is dropped, i.e., $h_2(x) \equiv 0$. Various methods are designed to solve such problems. The subgradient methods in the Riemannian setting are studied in [14, 30]. Riemannian proximal point algorithms are investigated by [19, 32, 42]. Operator splitting methods like the alternating direction methods of multipliers (ADMM) and the augmented Lagrangian methods (ALM) are also promising on manifolds [29, 43], in which unconstrained manifold optimization algorithms are used to solve the subproblems. However, due to the existence of two nonsmooth terms in (1.1), the existing manifold-based algorithms may not be directly applicable for solving (1.1) in general cases. For example, the direct application of ManPG [19] or the Riemannian proximal gradient method [42] requires dealing with a subproblem with two nonsmooth terms, which is generally difficult to solve. Moreover, by introducing more auxiliary variables, the multiblock ADMM method can be exploited, but there is no convergence guarantee of such algorithm, to the best of our knowledge.

On the other hand, in many situations $\mathcal{M}$ is embedded in a Euclidean space and can be specified by equality constraints, e.g., the Stiefel manifold or the Oblique manifold. In these cases, (1.1) can be viewed as a constrained optimization problem in Euclidean spaces [21, 44, 49, 69], which has been widely studied for many years [5, 16, 22, 59]. However, due to the complex constraints induced by embedded manifolds, the constraint qualifications may not be satisfied, and the nonlinear methods are not applicable for solving optimization problems over abstract manifolds. Motivated by the above analysis, we aim at designing numerical algorithms for solving (1.1) by exploiting the intrinsic structure of manifolds.

The Main Contributions

In this paper, we propose a manifold-based augmented Lagrangian method to solve (1.1), which consists of two nonsmooth terms. Compared to the existing methods, the proposed algorithm satisfies the manifold constraint automatically at each step and exploits the second-order geometric property of manifolds. The main idea of the proposed method is to introduce auxiliary variables that split (1.1) into a smooth manifold constrained term, a nonsmooth term and an inequality constrained term. Then, we apply the augmented Lagrangian method to solve the equivalent version of (1.1) and show the global

Table 1: Notations used in this work.

Notations	Descriptions
$[u]_i$	The i-th component of $u \in \mathbb{R}^d$
$[M]$	The set $\{1, 2, \cdots, M\}$, where M is a positive integer
$\mathcal{M}$	A complete n-dimensional smooth Riemannian manifold
$T_p\mathcal{M}$	The tangent space at $p \in \mathcal{M}$
$T\mathcal{M}$	The tangent bundle of $\mathcal{M}$
$\mathcal{D}(\mathcal{M})$	The set of smooth functions on $\mathcal{M}$ with compact support
$\mathrm{d}\varphi\|_p$	The differential of the smooth map φ at $p \in \mathcal{M}$
$\operatorname{grad}\varphi$	The gradient of the function φ on manifolds
$\partial\varphi$	The Clarke subgradient of the function φ on manifolds
$\operatorname{Hess}\varphi$	The Hessian of the function φ on manifolds
X, Y	Vector fields on manifolds
$\mathcal{X}(\mathcal{M})$	The set of all smooth vector fields on $\mathcal{M}$
∂X	The Clarke generalized covariant derivative of the vector field X
$\nabla_X Y$	The Levi-Civita connection of two vector fields X and Y
$\nabla X(p; v)$	The directional derivative of a vector field X at p along v
$\nabla X(p)$	The operator $v \mapsto \nabla_v X(p)$ from $T_p\mathcal{M}$ to $T_p\mathcal{M}$
$P_\gamma^{s\to t}$	The parallel transport along a curve γ from $\gamma(s)$ to $\gamma(t)$
P_{pq}	The parallel transport along the geodesic from p to q
$\exp_p$	The exponential map at $p \in \mathcal{M}$
$\mathcal{L}(T_p\mathcal{M})$	The linear space of all linear operators from $T_p\mathcal{M}$ to $T_p\mathcal{M}$
$\mathbf{P}_p V$	The projection of a vector $V \in \mathbb{R}^d$ into $T_p\mathcal{M}$
$B_p(r)$	The open ball $\{q \in \mathcal{M} : d(p, q) < r\}$
$\\|Q\\|_1$	The ℓ_1-norm of $Q \in \mathbb{R}^{n\times r}$, namely $\sum_{i,j} \|Q_{ij}\|$
$\\|Q\\|_F$	The Frobenius norm of the matrix $Q \in \mathbb{R}^{n\times r}$
$\\|Q\\|$	The operator norm of the matrix $Q \in \mathbb{R}^{n\times r}$
$\\|x\\|_2$, $\\|x\\|_{\mathbb{R}^n}$	The ℓ_2-norm of the vector $x \in \mathbb{R}^n$
$\\|x\\|_\infty$	The ℓ_∞-norm of the vector $x \in \mathbb{R}^n$
$\langle\eta, \xi\rangle_p$,$\langle\eta, \xi\rangle$	The Riemannian inner product of $\eta, \xi \in T_p\mathcal{M}$
$\\|\eta\\|_p$, $\\|\eta\\|$	The norm of $\eta \in T_p\mathcal{M}$

convergence property under some constraint qualifications on manifolds. Using the Moreau-Yosida identity, the augmented Lagrangian subproblem can be converted to a continuous and differentiable manifold optimization problem, but it is not second-order differentiable. This subproblem is inherently different from the subproblems in ManPG [19] and the Riemannian proximal gradient method [42], which are nonsmooth problems on the tangent space. To solve the augmented Lagrangian subproblem, we propose a globalized version of the semismooth Newton method on manifolds and prove its local superlinear convergence under some reasonable assumptions. We also provide a theorem showing that when the manifold $\mathcal{M}$ is a compact submanifold of another Riemannian manifold $\bar{\mathcal{M}}$, the semismoothness of a vector field X on $\mathcal{M}$ can be inherited from the semismoothness of its extension to $\bar{\mathcal{M}}$. Numerical results in compressed modes (1.2), sparse PCA (1.3), and the constrained sparse PCA (1.4) show the advantages of the proposed method comparing with existing approaches.

The remaining parts of this article are organized as follows: Preliminaries on manifolds and backgrounds about previously mentioned optimization methods are presented in Sec. 2. The augmented Lagrangian method together with its convergence analysis are given in Sec. 3. The globalized semismooth Newton method for dealing with the subproblem, and its global convergence together with the transition to local superlinear convergence are shown in Sec. 4. The semismooth property on submanifolds and the method for calculating the Clarke generalized covariant derivative are also explored in Sec. 4. Numerical experiments are reported in Sec. 5. Finally, this article is concluded in Sec. 6.

2 Background

In this section, we review some concepts of manifolds and briefly discuss some related literature of nonsmooth manifold optimization, nonsmooth nonconvex ALM in Euclidean spaces, and the semismooth Newton method.

A Hausdorff topological space $\mathcal{M}$ is said to be an n-dimensional manifold if it has a countable basis and for each $p \in \mathcal{M}$ there exist a neighborhood U of p, an open subset $\hat{U} \subset \mathbb{R}^n$ and a map $\varphi : U \to \hat{U}$ such that φ is a homeomorphism. The pair (U, φ) is called a chart.

Notations used in the remaining part of this article are listed in Table 1. As the Euclidean spaces can be interpreted as the linear manifolds [2], our notations for manifolds are consistent with those used in Euclidean spaces when the function is defined on $\mathbb{R}^n$, e.g. $\operatorname{grad}\varphi = \nabla\varphi$ if φ is defined on $\mathbb{R}^n$. Now, we briefly review basic definitions and properties of functions defined on manifolds. Most of these definitions can be found in, e.g., [18, Chapter 1-3] and [2, Chapter 3].

Definition 2.1 A tangent vector $\xi_p : \mathcal{D}(\mathcal{M}) \to \mathbb{R}$ to a manifold $\mathcal{M}$ at a point p is a linear operator such that for every $f \in \mathcal{D}(\mathcal{M})$, $\xi_p f = \dot{\gamma}(0) f := \left.\frac{\mathrm{d}(f(\gamma(t)))}{\mathrm{d}t}\right|_{t=0}$, where $\gamma : (-1, 1) \to \mathcal{M}$ is a smooth curve on $\mathcal{M}$ with $\gamma(0) = p$.

The tangent space $T_p\mathcal{M}$ is the space containing all tangent vectors at p, which is an n-dimensional $\mathbb{R}$-linear space. A Riemannian metric $\langle \cdot, \cdot \rangle_p$ gives an inner product on $T_p\mathcal{M}$, which smoothly depends on p.[1] Moreover, a Riemannian metric gives a metric on $\mathcal{M}$ and the gradient of functions defined on $\mathcal{M}$. Below we assume that $\mathcal{M}$ is equipped with a Riemannian metric.

Definition 2.2 Given $p, q \in \mathcal{M}$, the distance between p and q is defined as

$$d(p, q) = \inf \left\{ \ell(\gamma) : \ell(\gamma) := \int_0^1 \sqrt{\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle} \mathrm{d}t \right\},$$

where inf is taken over all piecewise smooth curves $\gamma : [0, 1] \to \mathcal{M}$ with $\gamma(0) = p$ and $\gamma(1) = q$.

Definition 2.3 Let $f : \mathcal{M} \to \mathcal{N}$ be a smooth map between smooth manifolds $\mathcal{M}, \mathcal{N}$, the differential of f at $p \in \mathcal{M}$, denoted by $\mathrm{d}f|_p$, is a map from $T_p\mathcal{M}$ to $T_{f(p)}\mathcal{N}$ such that $(\mathrm{d}f|_p\eta)g := \eta(g \circ f)$ for all $g \in \mathcal{D}(\mathcal{N})$ and $\eta \in T_p\mathcal{M}$.

Definition 2.4 Let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function and $p \in \mathcal{M}$, the gradient of f at p is defined as the unique tangent vector $\operatorname{grad} f(p) \in T_p\mathcal{M}$ that satisfies

$$\xi_p f = \langle \xi_p, \operatorname{grad} f(p) \rangle, \ \forall \xi_p \in T_p\mathcal{M}.$$

The uniqueness of $\operatorname{grad} f(p)$ follows from the Riesz representation theorem. Define $T\mathcal{M} := \bigcup_{p\in\mathcal{M}} T_p\mathcal{M}$ to be the tangent bundle of $\mathcal{M}$ and a map $X : \mathcal{M} \to T\mathcal{M}$ to be a vector field on $\mathcal{M}$ if $X(p) \in T_p\mathcal{M}$ for all $p \in \mathcal{M}$.

Definition 2.5 For any $X, Y \in \mathcal{X}(\mathcal{M})$, a map $\nabla_X Y \in \mathcal{X}(\mathcal{M})$ is called the Levi-Civita connection if it is an affine connection[2] and satisfies

$$X \langle Y, Z \rangle = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla_X Z \rangle \quad \text{and} \quad \nabla_X Y - \nabla_Y X = XY - YX,$$

for all $X, Y, Z \in \mathcal{X}(\mathcal{M})$.

The Levi-Civita connection is unique [18] and can define the parallel transport of a vector field.

Definition 2.6 A vector field X is parallel along a smooth curve γ if $\nabla_{\dot{\gamma}} X = 0$.

Given a smooth curve γ and $\eta \in T_{\dot{\gamma}(0)}\mathcal{M}$, there exists a unique parallel vector field X_η along γ such that $X_\eta(0) = \eta$. We define the parallel transport along γ to be $P_\gamma^{0\to t}\eta := X_\eta(t)$. A curve γ is called a geodesic if it is parallel to itself, i.e. $\nabla_{\dot{\gamma}}\dot{\gamma} = 0$, which implies $P_\gamma^{0\to t}\dot{\gamma}(0) = \dot{\gamma}(t)$. For given initial conditions $\gamma(0) = p \in \mathcal{M}$, $\dot{\gamma}(0) = \eta \in T_p\mathcal{M}$, the geodesic equation $\nabla_{\dot{\gamma}}\dot{\gamma} = 0$ has a solution locally. Let V_p be the set of $\eta \in T_p\mathcal{M}$ such that γ is a geodesic, $\gamma(0) = p$, $\dot{\gamma}(0) = \eta$ and $\gamma(1)$ exists. The exponential map $\exp_p : V_p \to \mathcal{M}$ is defined as $\eta \mapsto \gamma(1)$. When the geodesic from p to q is unique, denoted by γ_{pq}, we

[1] The subscript p in $\langle \cdot, \cdot \rangle_p$ is usually omitted for simplicity.

[2] An affine connection is $\mathcal{D}(\mathcal{M})$-linear w.r.t. X, $\mathbb{R}$-linear w.r.t. X, Y, and satisfies the product rule, see Chapter 5 in [2].

define $P_{pq} := P_{\gamma_{pq}}^{0\to 1}$. We highlight that the parallel transport $P_\gamma^{0\to t}$ is a linear isometry, i.e., $P_\gamma^{0\to t}$ is linear and $\langle \xi, \zeta \rangle = \left\langle P_\gamma^{0\to t}\xi, P_\gamma^{0\to t}\zeta \right\rangle$ for $\xi, \zeta \in T_p\mathcal{M}$ (see Sec. 5.4 in [2] for details).

The exponential map is not always tractable, e.g., it may be expansive to compute or does not have a closed-form solution (since we need to solve a differential equation). However, it is possible that the convergence properties of an optimization algorithm remains the same when the exponential map is replaced with its first-order approximation [2, 4]. Such an approximation is called a retraction and defined as follows.

Definition 2.7 A C^2 map $R : T\mathcal{M} \to \mathcal{M}$ is a retraction if $R_p(0) = p$ and $\frac{\mathrm{d}}{\mathrm{d}t} R_p(t\eta)|_{t=0} = \eta$ for all $\eta \in T_p\mathcal{M}$ and $p \in \mathcal{M}$, where we denote $R_p := R(p, \cdot)$.

Generally, we only require that for every $p \in \mathcal{M}$, R is defined on a neighborhood of $(p, 0) \in T\mathcal{M}$.

Definition 2.8 ([27, 28]) Let X be a vector field on $\mathcal{M}$. The directional derivative at $p \in \mathcal{M}$ along $v \in T_p\mathcal{M}$ is defined as

$$\nabla X(p; v) := \lim_{t\to 0^+} \frac{1}{t}\left[P_{\exp_p(tv),p} X(\exp_p(tv)) - X(p)\right] \in T_p\mathcal{M}. \tag{2.1}$$

We say X is directionally differentiable at p if $\nabla X(p; v)$ exists for all $v \in T_p\mathcal{M}$. When X is smooth at p, we know $\nabla X(p; v) = \nabla_v X(p)$ (see [57, p. 234]).

Definition 2.9 Let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function. The Hessian of f at $p \in \mathcal{M}$, denoted by $\operatorname{Hess} f(p)$, is defined as a linear operator on $T_p\mathcal{M}$ such that $\operatorname{Hess} f(p)[v] := \nabla_v \operatorname{grad} f(p)$ for all $v \in T_p(\mathcal{M})$.

We refer readers to [2, 18, 46, 47] for more details about manifolds.

2.2 Nonsmooth Manifold Optimization

As suggested in [19], most nonsmooth manifold optimization algorithms can be classified into three categories: subgradient methods, proximal point algorithms and operator splitting methods.

2.2.1 Subgradient Methods

The subgradient methods on manifolds [14, 30] naturally generalize their Euclidean space counterparts. Suppose $f : \mathcal{M} \to \mathbb{R}$ is a locally Lipschitz function[3] on a manifold $\mathcal{M}$ and (U, φ) is a chart containing $p \in \mathcal{M}$. From [8, 30, 37], the Clarke generalized directional derivative of f at p, denoted by $f^\circ(p; v)$, is defined by

$$f^\circ(p; v) := \limsup_{y\to p, t\downarrow 0} \frac{\hat{f}(\varphi(y) + t\mathrm{d}\varphi|_p v) - \hat{f}(\varphi(y))}{t},$$

where $\hat{f} := f \circ \varphi^{-1}$ and $\mathrm{d}\varphi|_p$ is the differential of φ at p. The Clarke subgradient is

$$\partial f(p) := \{\xi \in T_p\mathcal{M} : \langle \xi, v\rangle \le f^\circ(p; v),\ \forall v \in T_p\mathcal{M}\}.$$

Indeed, $f^\circ(p; v)$ is the Clarke directional derivative of $\hat{f}$ in Euclidean spaces and is independent of the choice of φ. Moreover, the Clarke subgradient $\partial f(p)$ can be obtained from the Euclidean version as shown in Proposition 3.1 of [65]:

$$\partial f(p) = (\mathrm{d}\varphi|_p)^{-1}[G_{\varphi(p)}^{-1} \partial \hat{f}(\varphi(p))], \tag{2.2}$$

where $G_{\varphi(p)} \in \mathbb{R}^{n\times n}$ is the metric matrix such that its (i, j)-th element is $g_{ij} := \langle (\mathrm{d}\varphi|_p)^{-1} e_i, (\mathrm{d}\varphi|_p)^{-1} e_j\rangle$, where $\{e_i\}_{i\in[n]}$ is the standard basis of $\mathbb{R}^n$, i.e., the j-th component of e_i is δ_{ij}.

The update rule of the subgradient method in the Riemannian setting [14, 30] is $p_{k+1} = \exp_{p_k}(t_k v_k)$, where $v_k \in \partial f(p_k)$ and t_k is the stepsize. These methods are known to be slow in the Euclidean setting. From the experiments in [19], it is also observed that subgradient based methods are slower than proximal point algorithms and operator splitting methods in the Riemannian setting.

[3] We say a function f on a manifold is locally Lipschitz if $f \circ \varphi^{-1}$ is locally Lipschitz in U for every chart (U, φ).

2.2.2 Proximal Point Methods

The extension of proximal point algorithms on manifolds is proposed in [32] and the subgradient methods are suggested in [9] to solve the subproblem. In [32], manifolds with non-positive sectional curvature are considered, which exclude many important applications such as optimization problems on the Stiefel manifold. Very recently, Chen et al. proposed the proximal gradient method (ManPG) on the Stiefel manifold [19] with proved convergence. More specifically, it aims at solving the following problem:

$$\min_{Q \in \mathcal{M}} \; \{f(Q) + \psi(Q)\}, \tag{2.3}$$

where $\mathcal{M} = \mathrm{St}(n,r)$ is the Stiefel manifold, f is smooth with Lipschitz gradient, and ψ is convex and Lipschitz. In each step, the descent direction is determined by solving the subproblem

$$V_k := \operatorname*{arg\,min}_{V \in T_{Q_k}\mathcal{M}} \left\{ \langle \operatorname{grad} f(Q_k), V \rangle + \frac{1}{2t} \|V\|_F^2 + \psi(Q_k + V) \right\} \tag{2.4}$$

via the regularized semismooth Newton method [63]. Besides, Huang and Wei extended an accelerated version of the proximal gradient method to manifolds [41]. They also proposed a Riemannian proximal gradient method [42] to solve (2.3) for general manifolds by replacing the term $\psi(Q_k+V)$ with $\psi(R_{Q_k}(V))$, where R_{Q_k} is a retraction, and analyzed the iteration complexity for convex objectives under some assumptions. However, the direct application of the above three methods for solving (1.1) has to deal with the subproblem:

$$V_k := \operatorname*{arg\,min}_{V \in T_{Q_k}\mathcal{M}} \left\{ \langle \operatorname{grad} f(Q_k), V \rangle + \frac{1}{2t} \|V\|_F^2 + \underbrace{\psi(Q_k + V) + \delta_{\mathcal{C}}(Q_k + V)}_{(\text{ or } \psi(R_{Q_k}(V)) + \delta_{\mathcal{C}}(R_{Q_k}(V)))} \right\},$$

where $\delta_{\mathcal{C}}$ is the indicator function of the feasible set corresponding to the inequality constraints in (1.1). In general, the above problem is difficult due to the existence of two nonsmooth terms.

It is worth mentioning that (2.3) may be solved in a more efficient way when $\mathcal{M}$ has specific structures, e.g. $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$ where $\mathcal{M}_1$ and $\mathcal{M}_2$ are two manifolds. Chen et al. [20] proposed an alternating manifold proximal gradient algorithm to solve (2.3), which alternatively updates the variables in Gauss-Seidel fashion based on the linearized formulation (2.4). It is empirically observed that this effective alternating strategy leads to better performance than ManPG [20]. Exploring the special structure of the manifold is a promising direction, and we leave it as our future work.

2.2.3 Operator Splitting Methods

Operator splitting methods on manifolds split (2.3) into several terms, each of which is easier to solve. For example, the manifold ADMM proposed in [43] rewrites (2.3) to

$$\min_{Q,Z} \; \{f(Q) + \psi(Z)\} \quad \text{s.t.} \quad Q = Z, \; Q \in \mathcal{M}. \tag{2.5}$$

Then, a two-block ADMM is used to solve it, which has the following update rules:

$$\begin{aligned}
Q_{k+1} &:= \operatorname*{arg\,min}_{Q \in \mathcal{M}} \left\{ f(Q) + \frac{\rho}{2} \|Q - Z_k + U_k\|_F^2 \right\}, \\
Z_{k+1} &:= \operatorname*{arg\,min}_{Z} \left\{ \psi(Z) + \frac{\rho}{2} \|Q_{k+1} - Z + U_k\|_F^2 \right\}, \\
U_{k+1} &:= U_k + Q_{k+1} - Z_{k+1}.
\end{aligned}$$

The Q-update requires smooth manifold optimization algorithms and the Z-update is the proximal mapping of ψ. Besides, an inexact ALM framework to solve (2.5) with some convergence results is considered in [29].

When the manifold $\mathcal{M}$ can be embedded in a Euclidean space, classical nonsmooth nonconvex constrained optimization algorithms can also be explored. Lai et al. proposed a splitting method for orthogonality constrained problems (SOC) [44], which reformulates (2.3) into

$$\min_{Q,P,R} \{f(P) + \psi(R)\} \quad \text{s.t.} \quad P = R, \; Q = P, \; Q^\top Q = I_r. \tag{2.6}$$

A three-block ADMM is then used to solve the above problem:

$$\begin{aligned} P_{k+1} &:= \underset{P\in\mathbb{R}^{n\times r}}{\arg\min} \left\{ f(P) + \frac{\rho}{2}\left\|P - R_k + \Lambda_k\right\|_F^2 + \frac{\rho}{2}\left\|P - Q_k + \Gamma_k\right\|_F^2 \right\}, \\ R_{k+1} &:= \underset{R\in\mathbb{R}^{n\times r}}{\arg\min} \left\{ \psi(R) + \frac{\rho}{2}\left\|P_{k+1} - R + \Lambda_k\right\|_F^2 \right\}, \\ Q_{k+1} &:= \underset{Q\in\mathbb{R}^{n\times r}}{\arg\min}\ \frac{\rho}{2}\left\|P_{k+1} - Q + \Gamma_k\right\|_F^2 \quad \text{s.t.} \quad Q^\top Q = I_r, \\ \Lambda_{k+1} &:= \Lambda_k + P_{k+1} - R_{k+1}, \quad \Gamma_{k+1} := \Gamma_k + P_{k+1} - Q_{k+1}, \end{aligned}$$

where the R-update can be solved by a proximal map, and the Q-update has the closed form solution, and the P-update can be solved using gradient based methods. Although these ADMM-type methods are simple, to the best of our knowledge, it is unclear whether they converge to a KKT point of (2.6).

Unlike ADMM, ALM usually has theoretical guarantees. In [21], Chen et al. proposed the proximal alternating minimized augmented Lagrangian method (PAMAL), which solves the augmented Lagrangian subproblem by the proximal alternating minimization (PAM) scheme [7]. In [69], the so-called EPALMAL is proposed, where the PALM [12] is used for solving the subproblem. Although both PAMAL and EPALMAL have certain convergence guarantee, the analysis is not complete for solving (1.1).

2.3 Augmented Lagrangian Methods for Nonsmooth and Nonconvex Problems

Here, we review some augmented Lagrangian methods in constrained optimization, which has been studied for many decades [10]. We consider the following problem:

$$\min_{x\in\mathbb{R}^n} \{f(x) + \Phi(x)\} \quad \text{s.t.} \quad g(x) = 0,\ h(x) \le 0, \tag{2.7}$$

where f, g, h are smooth and Φ is lower semicontinuous. It is noted that the original problem (1.1) has the above form when the manifold $\mathcal{M}$ can be written as

$$\mathcal{M} = \{x : g_1(x) = 0,\ h_1(x) \le 0\}, \tag{2.8}$$

where g_1 and h_1 are parts of g and h, respectively. When $\Phi \equiv 0$ and the constraints can be divided into $g_1(x) = 0$, $g_2(x) = 0$, $h_1(x) \le 0$ and $h_2(x) \le 0$ such that the minimization problem is easier on $\{x : g_2(x) = 0,\ h_2(x) \le 0\}$, Anderani et al. proposed an Augmented Lagrangian (AL) method [5] to solve (2.7). It is shown that any feasible limit point generated by the algorithm is a KKT point under the constant positive linear dependence (CPLD) condition [54], which is weaker than the linear independence constraint qualification (LICQ) condition. However, this method cannot guarantee that any limit point is feasible when the penalty parameter, i.e., the coefficient of the quadratic term in the augmented Lagrangian function, tends to infinity as the iteration proceeds (see Theorem 4.1(i) in [5]). This infeasiblity phenomenon also exists in other literature such as [25].

To alleviate this issue, another AL method is proposed in [49], where two nonmonotone proximal methods are applied to solve the subproblem. They consider the problem (2.7) with the additional assumption that Φ is a convex function. A feasible point is assumed to be known, and is used to guarantee that the augmented Lagrangian function is uniformly bounded from above at points generated in subproblems. Besides, the method in [49] modifies the update rule of the penalty parameter to ensure that the penalty grows faster than Lagrangian multipliers (see, e.g., (3.14) in Algorithm 3.1). Using these two properties, the convergence result that any limit point is a KKT point under Robinson's constraint qualification is established. Recently, Chen et al. [22] proposed an AL method to solve (2.7) with Φ possibly being a nonconvex non-Lipschitz function. Under a weak constraint qualification called the relaxed constant positive linear dependence (RCPLD) condition [6], they provided a global convergence result.

2.4 Semismooth Newton Methods

The subproblem in ALM generally requires tackling a nonsmooth equation, which usually can be efficiently solved by the semismooth Newton method [50, 53, 58]. Under suitable assumptions, the semismooth Newton method has the local superlinear convergence rate. Recently, this method is generalized to solving nonsmooth equations [27] on manifolds based on the Clarke generalized covariant derivatives [34, 55]. Below we introduce several definitions for locally Lipschitz vector fields on manifolds.

Definition 2.10 ([27]) Let $L > 0$ and $\Omega \subset \mathcal{M}$. We say a vector field $X : \mathcal{M} \to T\mathcal{M}$ on a manifold $\mathcal{M}$ is L-Lipschitz in Ω if for each $x, y \in \Omega$, and each geodesic γ joining x, y, it holds

$$\left\| P_{\gamma}^{0\to 1} X(x) - X(y) \right\| \le L\ell(\gamma),$$

and we say that X is locally Lipschitz at $p \in \mathcal{M}$ if there exist a neighborhood $U_p \ni p$ and a constant $L_p > 0$ such that X is L_p-Lipschitz in U_p. If X is locally Lipschitz at every $p \in \mathcal{M}$, we say that X is locally Lipschitz on $\mathcal{M}$.

Since a locally Lipschitz vector field X on manifolds is differentiable almost everywhere [27], we denote $\mathcal{D}_X$ as the set of its differentiable points and define the Clarke generalized covariant derivative as follows:

Definition 2.11 ([27, 34]) Let X be a locally Lipschitz vector field on $\mathcal{M}$. The B-derivative is a set-valued map $\partial_B X : \mathcal{M} \rightrightarrows \mathcal{L}(T\mathcal{M})$ with

$$\partial_B X(p) := \left\{ H \in \mathcal{L}(T_p\mathcal{M}) : \ \exists \{p_k\} \subset \mathcal{D}_X, \ \lim_{k\to+\infty} p_k = p, \ H = \lim_{k\to+\infty} \nabla X(p_k) \right\}, \tag{2.9}$$

where the last limit means that $\|\nabla X(p_k)[P_{pp_k} v] - P_{pp_k} H v\| \to 0$ for all $v \in T_p\mathcal{M}$. The Clarke generalized covariant derivative is a set-valued map $\partial X : \mathcal{M} \rightrightarrows \mathcal{L}(T\mathcal{M})$ such that $\partial X(p)$ is the convex hull of $\partial_B X(p)$.

The above definitions are consistent with those in $\mathbb{R}^n$ as the tangent spaces can be identified to $\mathbb{R}^n$ so $P_{p_k p}$ and P_{pp_k} are the identical mappings. The properties of the Clarke generalized covariant derivative are similar to those in Euclidean spaces. For example, $\partial_B X(p)$ and $\partial X(p)$ are non-empty compact sets and the maps $\partial_B X, \partial X$ are locally bounded and upper semicontinuous [27, Proposition 3.1]. Having introduced these notions, the Newton method for a locally Lipschitz vector field X on $\mathcal{M}$ is [27]:

$$p_{k+1} := \exp_{p_k}(-H_k^{-1} X(p_k)), \quad \text{where } H_k \in \partial X(p_k). \tag{2.10}$$

To obtain the convergence rate, we have to impose the semismooth property of the vector field X.

Definition 2.12 ([27]) Let X be a locally Lipschitz vector field on $\mathcal{M}$. We say X is semismooth with order μ at $p \in \mathcal{M}$ if it is directionally differentiable in a neighborhood U of p, and for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\left\| X(p) - P_{qp}[X(q) + H_q \exp_q^{-1} p] \right\| \le \varepsilon d(p,q)^{1+\mu}, \ \forall\, q \in B_\delta(p), H_q \in \partial X(q), \tag{2.11}$$

where $B_\delta(p) := \{q \in \mathcal{M} : d(p,q) < \delta\}$ and $\exp_q^{-1} p$ is the inverse of the exponential map[4].

In [27], it is shown that if X is locally Lipschitz, $X(p_*) = 0$, all elements in $\partial X(p_*)$ are nonsingular and X is semismooth at p_* with order μ, then the Newton iteration (2.10) has the local convergence rate $1 + \mu$. This result is similar to that in Euclidean spaces [50, 53, 58].

3 An Augmented Lagrangian Framework

In this section, we present an augmented Lagrangian method to solve (1.1) and establish its convergence result. The method for solving the subproblem is deferred to the next section.

[4] This is well-defined in a small neighborhood of q [18, Proposition 3.2.9].

3.1 Algorithm

Recall that we consider the following optimization problem:

$$\min_x \{f(x)+\psi(h_1(x))\}, \quad \text{s.t. } x\in\mathcal{M},\ h_2(x)\le 0. \tag{3.1}$$

Throughout this paper, we always make the following assumptions:

Assumption 3.1 *$\mathcal{M}$ is a complete smooth Riemannian manifold.*

Assumption 3.2 *$f:\mathcal{M}\to\mathbb{R}$, $h_1:\mathcal{M}\to\mathbb{R}^m$, $h_2:\mathcal{M}\to\mathbb{R}^q$ are continuously differentiable, and $\psi:\mathbb{R}^m\to\mathbb{R}$ is a convex function. $f(x)+\psi(y)$ is bounded below for $(x,y)\in\mathcal{M}\times\mathbb{R}^m$.*

Note that we can reformulate (3.1) to the following problem:

$$\min_{x,y,z} \{f(x)+\psi(y)\}, \quad \text{s.t. } x\in\mathcal{M},\ y=h_1(x),\ z=h_2(x),\ z\le 0. \tag{3.2}$$

The augmented Lagrangian function $L_\sigma:\mathcal{M}\times\mathbb{R}^m\times\mathbb{R}^q_-\times\mathbb{R}^m\times\mathbb{R}^q\to\mathbb{R}$ of (3.2) is given by

$$L_\sigma(x,y,z,\lambda,\gamma)=f(x)+\psi(y)+\frac{\sigma}{2}\left\|h_1(x)-y+\frac{\lambda}{\sigma}\right\|_2^2+\frac{\sigma}{2}\left\|h_2(x)-z+\frac{\gamma}{\sigma}\right\|_2^2-\frac{\|\lambda\|_2^2+\|\gamma\|_2^2}{2\sigma}. \tag{3.3}$$

We note that simultaneously minimizing L_σ with respect to x,y,z is equivalent to

$$\min_{x\in\mathcal{M}}\left\{f(x)+\psi^\sigma\left(h_1(x)+\frac{\lambda}{\sigma}\right)+\delta^\sigma_{\mathbb{R}^q_-}\left(h_2(x)+\frac{\gamma}{\sigma}\right)\right\}, \tag{3.4}$$

where ψ^σ, $\delta^\sigma_{\mathbb{R}^q_-}$ are the Moreau-Yosida regularization of ψ, $\delta_{\mathbb{R}^q_-}$, respectively. More specifically, it holds that

$$\psi^\sigma(x):=\min_{y\in\mathbb{R}^m}\left\{\psi(y)+\frac{\sigma}{2}\|x-y\|_2^2\right\}, \tag{3.5}$$

$$\delta^\sigma_{\mathbb{R}^q_-}(x):=\min_{z\in\mathbb{R}^q}\left\{\delta_{\mathbb{R}^q_-}(z)+\frac{\sigma}{2}\|x-z\|_2^2\right\}, \tag{3.6}$$

where $\delta_{\mathbb{R}^q_-}$ is the indicator function of $\mathbb{R}^q_-$. Since the above two functions are crucial in developing our algorithm, we present some important properties.

Proposition 3.1 (Theorem 4.1.4 in [35]) *Let $\sigma>0$ and $f:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ be a closed proper convex function, then the Moreau-Yosida regularization f^σ is continuously differentiable, and its gradient is*

$$\nabla f^\sigma(x)=\frac{1}{\sigma}(x-\mathrm{prox}_{f/\sigma}(x)), \tag{3.7}$$

where $\mathrm{prox}_{f/\sigma}$ is the proximal map of the function f/σ:

$$\mathrm{prox}_{f/\sigma}(x):=\arg\min_y\left\{f(y)+\frac{\sigma}{2}\|x-y\|_2^2\right\}. \tag{3.8}$$

The minimization problems (3.5) and (3.6) are related to finding the proximal maps $\mathrm{prox}_{\psi/\sigma}$ and $\mathrm{prox}_{\delta_{\mathbb{R}^q_-}}$, which can be easily solved in many cases. In addition, since $\delta^\sigma_{\mathbb{R}^q_-}$ and ψ^σ are continuously differentiable, (3.4) is a smooth optimization problem on manifolds. These observations suggest the following augmented Lagrangian method, whose framework is similar to [22, 49].

Algorithm 3.1 (An augmented Lagrangian method of solving (3.2)) *Choose initial values $x_0\in\mathcal{M}$, $\gamma_0\in\mathbb{R}^q_+$, $\lambda_0\in\mathbb{R}^m$, $\sigma_0>0$, $\alpha,\tau\in(0,1)$, $\rho>1$ and a sequence $\{\varepsilon_k\}\subseteq\mathbb{R}_+$ converging to 0. Let $y_0=\mathrm{prox}_{\psi/\sigma_0}(h_1(x_0)+\lambda_0/\sigma_0)$, $z_0=\Pi_{\mathbb{R}^q_-}(h_2(x_0)+\gamma_0/\sigma_0)$, where $\Pi_{\mathbb{R}^q_-}$ is the projection onto $\mathbb{R}^q_-$. Choose a feasible point x_{feas} and a constant Φ such that*

$$\Phi\ge\max\{f(x_{\mathrm{feas}})+\psi(h_1(x_{\mathrm{feas}})),\ L_{\sigma_0}(x_0,y_0,z_0,\lambda_0,\gamma_0)\}. \tag{3.9}$$

Our algorithm repeats the following steps for $k=1,2,\ldots$

(i) Find $x_k \in \mathcal{M}$ such that

$$\|\mathrm{grad}\, L_k(x_k)\| < \varepsilon_k, \quad L_k(x_k) \le \Phi, \tag{3.10}$$

where

$$L_k(x) := f(x) + \psi^{\sigma_k}\left(h_1(x) + \frac{\lambda_k}{\sigma_k}\right) + \delta_{\mathbb{R}^q_-}^{\sigma_k}\left(h_2(x) + \frac{\gamma_k}{\sigma_k}\right). \tag{3.11}$$

(ii) Update y and z using

$$y_k = \mathrm{prox}_{\psi/\sigma_k}\left(h_1(x_k) + \frac{\lambda_k}{\sigma_k}\right), \tag{3.12}$$

$$z_k = \Pi_{\mathbb{R}^q_-}\left(h_2(x_k) + \frac{\gamma_k}{\sigma_k}\right). \tag{3.13}$$

(iii) Update the multipliers:

$$\lambda_{k+1} = \lambda_k + \sigma_k(h_1(x_k) - y_k), \quad \gamma_{k+1} = \gamma_k + \sigma_k(h_2(x_k) - z_k).$$

(iv) Let $\delta_k = \max\{\|h_1(x_k) - y_k\|_2, \ \|h_2(x_k) - z_k\|_2\}$.
If $\delta_k \le \tau\delta_{k-1}$, then $\sigma_{k+1} = \sigma_k$. Otherwise, set

$$\sigma_{k+1} = \max\left\{\rho\sigma_k, \ \|\lambda_{k+1}\|_2^{1+\alpha}, \ \|\gamma_{k+1}\|_2^{1+\alpha}\right\}. \tag{3.14}$$

3.2 Convergence Analysis

In this part, motivated by the proof in [22], we first give the feasibility result of Algorithm 3.1.

Theorem 3.1 *Let $\{(x_k, y_k, z_k)\}$ be the sequence generated by Algorithm 3.1. Then, we have*

$$\lim_{k\to\infty} (\|h_1(x_k) - y_k\|_2 + \|h_2(x_k) - z_k\|_2) = 0.$$

Consequently, if $x_* \in \mathcal{M}$ is an accumulation point of $\{x_k\}$, then $(x_*, h_1(x_*), h_2(x_*))$ is a feasible accumulation point of $\{(x_k, y_k, z_k)\}$. Moreover, $\{x_k\}$ always contains an accumulation point if $\mathcal{M}$ is compact.

Proof First, consider the case where $\{\sigma_k\}$ is bounded. There exists $k_1 \in \mathbb{N}$ such that $\delta_{k+1} \le \tau\delta_k$ for any $k \ge k_1$. Then, $\lim_{k\to\infty} \delta_k = 0$ since $\tau < 1$. By the definition of δ_k, we know $\lim_{k\to\infty} \|h_1(x_k) - y_k\|_2 = 0$ and $\lim_{k\to\infty} \|h_2(x_k) - z_k\|_2 = 0$.

In the case where $\{\sigma_k\}$ is unbounded. By the update rule, σ_k is updated for infinitely many times. Then, we can find $k_1 < k_2 < \cdots$ such that

$$\sigma_k = \sigma_{k_i} = \max\{\rho\sigma_{k_{i-1}}, \|\lambda_{k_i}\|_2^{1+\alpha}, \|\gamma_{k_i}\|_2^{1+\alpha}\}, \ \forall\, k_i \le k < k_{i+1}.$$

From (3.10) and the definition of $L_k(x)$, we know that $L_{\sigma_k}(x_k, y_k, z_k, \lambda_k, \gamma_k) \le \Phi$, where L_{σ_k} is defined in (3.3). Therefore, we have

$$\left\|h_1(x_k) - y_k + \frac{\lambda_k}{\sigma_k}\right\|_2^2 + \left\|h_2(x_k) - z_k + \frac{\gamma_k}{\sigma_k}\right\|_2^2 \le 2\frac{\Phi - f(x_k) - \psi(y_k)}{\sigma_k} + \frac{\|\lambda_k\|_2^2 + \|\gamma_k\|_2^2}{\sigma_k^2}. \tag{3.15}$$

Since $f + \psi$ is bounded below and $\sigma_k \to \infty$, then the first term of (3.15) converges to 0. Next, we show that the second term also converges to 0. Notice that $\|\lambda_{k_i}\|_2^{1+\alpha} \le \sigma_{k_i}$ and $\sigma_{k_i} \to \infty$, we have $\|\lambda_{k_i}\|_2 / \sigma_{k_i} \le \sigma_{k_i}^{-\frac{\alpha}{1+\alpha}} \to 0$ as $i \to \infty$. A similar result also holds for $\|\gamma_{k_i}\|_2 / \sigma_{k_i}$. Then by (3.15) and the definition of δ_{k_i}, we have $\lim_{i\to\infty} \delta_{k_i} = 0$. By the update rule of λ_k, for $k_i < k < k_{i+1}$, it holds that

$$\frac{\|\lambda_k\|_2}{\sigma_k} \le \frac{\|\lambda_{k-1}\|_2}{\sigma_{k_i}} + \|h_1(x_{k-1}) - y_{k-1}\|_2 \le \frac{\|\lambda_{k-1}\|_2}{\sigma_{k_i}} + \delta_{k-1}.$$

Also note that $\delta_k \le \tau\delta_{k-1}$ for all $k_i \le k < k_{i+1} - 1$. Then, for $k_i < k < k_{i+1}$ the following inequality holds by induction

$$\frac{\|\lambda_k\|_2}{\sigma_k} \le \frac{\|\lambda_{k_i}\|_2}{\sigma_{k_i}} + \sum_{j=k_i}^{k-1} \delta_j \le \frac{\|\lambda_{k_i}\|_2}{\sigma_{k_i}} + \delta_{k_i} \sum_{j=0}^{k-k_i-1} \tau^j \le \frac{\|\lambda_{k_i}\|_2}{\sigma_{k_i}} + \frac{\delta_{k_i}}{1-\tau}.$$

Thus, we have $\lim_{k\to\infty} \|\lambda_k\|_2 / \sigma_k = 0$. Similarly, $\lim_{k\to\infty} \|\gamma_k\|_2 / \sigma_k = 0$. Therefore, from (3.15), we conclude that $\lim_{k\to\infty} \delta_k = 0$. □

Next, we consider the convergence result of Algorithm 3.1 by introducing an extension of the constraint qualifications on manifolds [65]. Consider the problem (3.1), we define the active set of a feasible point x to be $\mathcal{A}(x) := \{i \in [q] : [h_2(x)]_i = 0\}$. The following constraint qualification can be introduced [65]:

Definition 3.1 (LICQ) We say that a feasible point $x \in \mathcal{M}$ of (3.1) satisfies the linear independence constraint qualification (LICQ) if $\{\mathrm{grad}\,[h_2(x)]_i : i \in \mathcal{A}(x)\}$ are linearly independent in $T_x\mathcal{M}$.

The above definition is the same as that in the Euclidean case except that Euclidean gradients are replaced by Riemannian gradients. Indeed, there is a weaker constraint qualification:

Definition 3.2 (CPLD) Let $x \in \mathcal{M}$ be a feasible point of (3.1) and define $S(x) := \{\mathrm{grad}\,[h_2(x)]_i : i \in \mathcal{A}(x)\}$. We say that x satisfies the constant positive linear dependence constraint qualification (CPLD) if for each subset $S_0(x) \subseteq S(x)$ whose elements are linearly dependent with non-negative coefficients, S_0 remains linearly dependent in a neighborhood of x.

The first-order optimality condition of (3.1) can be stated as follows using the LICQ condition, which is a direct consequence of [65, Theorem 4.1].

Corollary 3.1 *Define the Lagrangian of* (3.1) *as*

$$\mathcal{L}(x,\gamma) := f(x) + \psi(h_1(x)) + \gamma^\top h_2(x), \quad x \in \mathcal{M},\ \gamma \geq 0.$$

Suppose x_* is a local minimum of (3.1) *and the LICQ holds at x_*, then there exists a multiplier γ_* such that the following KKT conditions hold:*

$$0 \in \partial_x \mathcal{L}(x_*, \gamma_*), \tag{3.16a}$$

$$h_2(x_*) \in \mathbb{R}^q_-,\ \gamma_* \in \mathbb{R}^q_+,\ \gamma_*^\top h_2(x_*) = 0. \tag{3.16b}$$

Remark 3.1 The Lagrangian of the equivalent problem (3.2) is

$$\tilde{\mathcal{L}}(x,y,z,\lambda,\gamma) = f(x) + \psi(y) + \lambda^\top (h_1(x) - y) + \gamma^\top (h_2(x) - z).$$

Under the LICQ condition, a necessary optimality condition is the following KKT system [65]:

$$y = h_1(x),\ z = h_2(x), \tag{3.17a}$$

$$z \in \mathbb{R}^q_-,\ \gamma \in \mathbb{R}^q_+,\ \gamma^\top z = 0, \tag{3.17b}$$

$$0 \in \partial_y \tilde{\mathcal{L}}(x,y,z,\lambda,\gamma), \tag{3.17c}$$

$$\mathrm{grad}_x\, \tilde{\mathcal{L}}(x,y,z,\lambda,\gamma) = 0. \tag{3.17d}$$

Before presenting the relationship of these two optimality conditions, we give the following chain rule.

Lemma 3.1 *Let $p \in \mathcal{M}$, and $h : \mathcal{M} \to \mathbb{R}^m$ be a smooth map, $\psi : \mathbb{R}^m \to \mathbb{R}$ be a convex function. Then,*

$$\partial(\psi \circ h)(p) = \left\{ \sum_{i=1}^m \alpha_i\, \mathrm{grad}\,[h(p)]_i : \alpha \in \partial\psi(h(p)) \right\}. \tag{3.18}$$

Proof For any point $p \in \mathcal{M}$ and a fixed chart (U_p, φ) at p, the map $h \circ \varphi^{-1} : \varphi(U_p) \subset \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at p, and hence the chain rule in Euclidean spaces (e.g., Theorem 2.3.9 in [24]) implies

$$\partial(\psi \circ \hat{h})(\varphi(p)) = \mathrm{co} \left\{ \sum_{i=1}^m \alpha_i \xi_i : \alpha \in \partial\psi(y) \right\} = \left\{ \sum_{i=1}^m \alpha_i \xi_i : \alpha \in \partial\psi(y) \right\},$$

where $\hat{h} := h \circ \varphi^{-1}$, $y := h(p)$, $\xi_i := \nabla \hat{h}_i(\varphi(p))$ and "co" denotes the convex hull, and the last equality is from the fact that $\partial\psi(y)$ is a compact convex set. Note that from [2, p. 46], $(\mathrm{d}\varphi|_p)^{-1} G^{-1} \xi_i = \mathrm{grad}\,[h(p)]_i$, where G is the metric matrix defined around (2.2). Combining with (2.2), we know (3.18) holds. □

Using the chain rule, the following proposition shows that the two optimality conditions are equivalent.

Proposition 3.2 *Conditions* (3.16) *and* (3.17) *are equivalent:*

(i) If (x,γ) satisfies (3.16), then there exist y, z, λ such that $(x, y, z, \lambda, \gamma)$ satisfies (3.17);
(ii) If $(x, y, z, \lambda, \gamma)$ satisfies (3.17), then (x,γ) satisfies (3.16).

Proof When the conditions in (3.16) hold, we set $y = h_1(x)$, $z = h_2(x)$, and choose $g \in \partial(\psi \circ h_1)(x)$ such that $\operatorname{grad} f(x) + g + \sum_{i=1}^{q}[\gamma]_i \operatorname{grad}[h_2(x)]_i = 0$. By (3.18), there exists $\alpha \in \partial\psi(y)$ such that $g = \sum_{i=1}^{m} \alpha_i \operatorname{grad}[h_1(x)]_i$. Therefore, (3.17d) holds by setting $\lambda = \alpha$. As (3.17c) is equivalent to $\lambda \in \partial\psi(y)$, it also holds by the choice of λ. Other conditions directly follow from (3.16).

Conversely, when the conditions in (3.17) hold, (3.17c) and (3.18) imply that $\lambda \in \partial\psi(y)$ and $\sum_{i=1}^{m} \lambda_i \operatorname{grad}[h_1(x)]_i \in \partial(\psi \circ h_1)(x)$. Combining with (3.17d), the condition (3.16a) holds. $\square$

Finally, we show that Algorithm 3.1 converges to a KKT point.

Theorem 3.2 *Suppose there exist $K \subseteq \mathbb{N}$ and $(x_*, y_*, z_*) \in \mathcal{M} \times \mathbb{R}^m \times \mathbb{R}^q_-$ such that*

$$\lim_{K \ni k \to \infty} \left(d(x_k, x_*) + \|y_k - y_*\|_2 + \|z_k - z_*\|_2\right) = 0.$$

If the CPLD condition holds at x_*, then there exist $K_0 \subseteq K$ and $\lambda_* \in \mathbb{R}^m, \gamma_* \in \mathbb{R}^q_+$ such that $\lim_{K_0 \ni k \to \infty} \lambda_{k+1} = \lambda_*$, and the KKT conditions (3.17) hold at $(x_*, y_*, z_*, \lambda_*, \gamma_*)$. Moreover, when the LICQ holds at x_*, we can choose γ_* such that $\lim_{K_0 \ni k \to \infty} \gamma_{k+1} = \gamma_*$.

Proof First, from Theorem 3.1, we obtain the feasibility condition (3.17a). From (3.12), (3.13) and the property of Moreau-Yosida regularizations in Proposition 3.1, we know that

$$y_k = h_1(x_k) + \frac{\lambda_k}{\sigma_k} - \frac{1}{\sigma_k}\nabla\psi^{\sigma_k}\left(h_1(x_k) + \frac{\lambda_k}{\sigma_k}\right), \text{ and } z_k = h_2(x_k) + \frac{\gamma_k}{\sigma_k} - \frac{1}{\sigma_k}\nabla\delta^{\sigma_k}_{\mathbb{R}^q_-}\left(h_2(x_k) + \frac{\gamma_k}{\sigma_k}\right).$$

From the definition of λ_{k+1} and γ_{k+1}, combining the above two equations, we have

$$\operatorname{grad} L_k(x_k) = \operatorname{grad} f(x_k) + \sum_{i=1}^{m}[\lambda_{k+1}]_i \operatorname{grad}[h_1(x_k)]_i + \sum_{i=1}^{q}[\gamma_{k+1}]_i \operatorname{grad}[h_2(x_k)]_i, \tag{3.19}$$

where the chain rule for gradients on manifolds is used[5].

Since $y_k = \operatorname{prox}_{\psi/\sigma_k}(h_1(x_k) + \lambda_k/\sigma_k)$, then

$$0 \in \partial\psi(y_k) - (\sigma_k(h_1(x_k) - y_k) + \lambda_k) = \partial\psi(y_k) - \lambda_{k+1}. \tag{3.20}$$

Note that $\lim_{K \ni k \to \infty} y_k = y_*$, then $\{y_k\}_{k \in K}$ is bounded. By the locally boundedness of the subgradient [56, Corollary 24.5.1], $\bigcup_{k \in K} \partial\psi(y_k)$ is also bounded. Therefore, $\{\lambda_{k+1}\}_{k \in K}$ is bounded, so we can choose $K_1 \subseteq K$, $\lambda_* \in \mathbb{R}^m$ such that $\lim_{K_1 \ni k \to \infty} \lambda_{k+1} = \lambda_*$, which implies $\lambda_* \in \partial\psi(y_*)$.

On the other hand, since

$$z_k = \Pi_{\mathbb{R}^q_-}\left(h_2(x_k) + \frac{\gamma_k}{\sigma_k}\right) = \arg\min_{z \le 0} \left\| h_2(x_k) + \frac{\gamma_k}{\sigma_k} - z \right\|_2^2,$$

then from the optimal condition of the above problem, it holds that

$$0 = [\sigma_k(h_2(x_k) - z_k) + \gamma_k]^\top z_k = \gamma_{k+1}^\top z_k \quad \text{and} \quad \gamma_{k+1} \ge 0. \tag{3.21}$$

Let $\mathcal{A}_k = \{i \in [q] : [z_k]_i = 0\}$, from (3.21), we know $[\gamma_{k+1}]_i = 0$ for $i \notin \mathcal{A}_k$. Define $\mathcal{A}_* := \{i \in [q] : [z_*]_i = 0\}$. Since $z_k \to z_*$, we also find that for sufficiently large k and $i \notin \mathcal{A}_*$, $[\gamma_{k+1}]_i = 0$. Then, when k is large enough (3.19) can be written as

$$\operatorname{grad} L_k(x_k) = \operatorname{grad} f(x_k) + \sum_{i=1}^{m}[\lambda_{k+1}]_i \operatorname{grad}[h_1(x_k)]_i + \sum_{i \in \mathcal{A}_*}[\gamma_{k+1}]_i \operatorname{grad}[h_2(x_k)]_i. \tag{3.22}$$

---

[5] Let $g : \mathcal{M} \to \mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}$ be two smooth maps, we can use Definition 2.4 to compute the gradient of $f \circ g$. For $p \in \mathcal{M}$, $\xi_p \in T_p\mathcal{M}$, choosing any curve $\gamma : (-1, 1) \to \mathcal{M}$ such that $\gamma(0) = p$ and $\dot\gamma(0) = \xi_p$, we have $\langle \xi_p, \operatorname{grad}(f \circ g)(p)\rangle = \xi_p(f \circ g) = \frac{\mathrm{d}(f \circ g \circ \gamma)(t)}{\mathrm{d}t}\Big|_{t=0} = \sum_{i=1}^{m} \alpha_i \frac{\mathrm{d}(g_i \circ \gamma)(t)}{\mathrm{d}t}\Big|_{t=0} = \sum_{i=1}^{m} \alpha_i \cdot (\xi_p g_i) = \sum_{i=1}^{m} \alpha_i \langle \xi_p, \operatorname{grad}[g(p)]_i\rangle$, where $\alpha := \nabla f(q)$, $q := g(\gamma(t))$ and g_i is the i-th component of g. Thus, $\operatorname{grad}(f \circ g)(p) = \sum_{i=1}^{m} \alpha_i \operatorname{grad}[g(p)]_i$.

Using the Carathéodory's theorem of cones [11], there exist $J_k \subseteq \mathcal{A}_*$ and $[\hat{\gamma}_k]_j \geq 0$, where $j \in J_k$, such that $\{\mathrm{grad}\,[h_2(x_k)]_j : j \in J_k\}$ are linearly independent and

$$\mathrm{grad}\, L_k(x_k) = \mathrm{grad}\, f(x_k) + \sum_{i=1}^{m} [\lambda_{k+1}]_i \,\mathrm{grad}\,[h_1(x_k)]_i + \sum_{j \in J_k} [\hat{\gamma}_k]_j \,\mathrm{grad}\,[h_2(x_k)]_j, \tag{3.23}$$

Since $J_k \subseteq [q]$ is a finite set, we can choose J_* and $K_2 \subseteq K_1$ such that $J_k = J_*$ for all $k \in K_2$ and $|K_2| = \infty$. Define $M_k = \max\{[\hat{\gamma}_k]_i : i \in J_*\}$ for $k \in K_2$. When $\{M_k\}_{k \in K_2}$ is bounded, then we can find $K_3 \subseteq K_2$ and $\gamma_* \in \mathbb{R}^p_+$ such that $[\gamma_*]_i = 0$ for $i \notin J_*$ and $\lim_{K_3 \ni k \to \infty} [\hat{\gamma}_k]_i = [\gamma_*]_i$ for $i \in J_*$. Using the fact that $\|\mathrm{grad}\, L_k(x_k)\| \to 0$, (3.19) implies (3.17d). Since $J_* \subseteq \mathcal{A}_*$ and $[\gamma_*]_i = 0$ for $i \notin J_*$, then (3.17b) follows.

When $\{M_k\}_{k \in K_2}$ is unbounded, we can find $K_4 \subseteq K_3$ and $\hat{\gamma} \in \mathbb{R}^q_+$ such that $\lim_{K_4 \ni k \to \infty} [\hat{\gamma}_k]_i / M_k = [\hat{\gamma}]_i$ for $i \in J_*$ and $[\hat{\gamma}]_i = 0$ otherwise. By the definition of M_k, we have $\hat{\gamma} \neq 0$ and $\|\hat{\gamma}\|_\infty = 1$. Dividing (3.23) by M_k, using the boundedness of $\mathrm{grad}\, L_k(x_k)$, $\mathrm{grad}\, f(x_k)$ and $\mathrm{grad}\,[h_1(x_k)]_i$, letting $k \in K_4 \to \infty$ and noticing that $\lambda_{k+1} \to \lambda_*$, $x_k \to x_*$, we can get

$$\sum_{j \in J_*} [\hat{\gamma}]_j \,\mathrm{grad}\,[h_2(x_*)]_j = 0.$$

Note that $\|\hat{\gamma}\|_\infty = 1$ and $\hat{\gamma} \geq 0$, we know $\{\mathrm{grad}\,[h_2(x_*)]_j : j \in J_*\}$ are linearly dependent with non-negative coefficients. However, they are linearly independent near x_*, which contradicts to the CPLD assumption. Thus, M_k is always bounded. Moreover, when the LICQ holds at x_* but $\{\gamma_{k+1}\}$ is unbounded, we can divide (3.22) by $\|\gamma_{k+1}\|_2$ and yield a contradiction to the LICQ condition. Therefore, $\{\gamma_{k+1}\}$ is bounded and contains a convergent subsequence. □

4 A Globalized Semismooth Newton Method

Recall that at every step we have to find an approximated stationary point of the following problem.

$$\min_{x \in \mathcal{M}} \; L_k(x) = f(x) + \psi^{\sigma_k}\left(h_1(x) + \frac{\lambda_k}{\sigma_k}\right) + \delta^{\sigma_k}_{\mathbb{R}^q_-}\left(h_2(x) + \frac{\gamma_k}{\sigma_k}\right). \tag{4.1}$$

It is known that L_k is continuously differentiable, and from the calculation around (3.19), the gradient of L_k is

$$\mathrm{grad}\, L_k(x) = \mathrm{grad}\, f(x) + \sum_{i=1}^{m} \alpha_i(x) \,\mathrm{grad}\,[h_1(x)]_i + \sum_{i=1}^{q} \beta_i(x) \,\mathrm{grad}\,[h_2(x)]_i, \tag{4.2}$$

where $\alpha(x)$ and $\beta(x)$ are gradients of the Moreau-Yosida regularization ψ^{σ_k} and $\delta^{\sigma_k}_{\mathbb{R}^q_-}$ at $h_1(x) + \lambda_k/\sigma_k$ and $h_2(x) + \gamma_k/\sigma_k$, respectively. Note that $\alpha(x)$ and $\beta(x)$ are continuous but not differentiable, so the Newton method cannot be applied, and we need the semismooth Newton method. For simplicity, we consider the following abstract problem

$$\min \;\varphi(x), \;\text{ s.t. } x \in \mathcal{M}, \tag{4.3}$$

where $\varphi : \mathcal{M} \to \mathbb{R}$ is continuously differentiable. We make the following assumption.

Assumption 4.1 *The vector field $X := \mathrm{grad}\,\varphi$ is locally Lipschitz and can be factorized into $X = X_1 + X_2 + X_3$ such that X_1 is smooth and X_2, X_3 are locally Lipschitz.*

Due to the existence of two nonsmooth terms in X, we need to extend the Clarke generalized covariant derivatives in the definition of the semismoothness in (2.11) to a general set-valued map.

Definition 4.1 Let X be a vector field on $\mathcal{M}$ and $\mathcal{K} : \mathcal{M} \rightrightarrows \mathcal{L}(T\mathcal{M})$ be an upper semicontinuous[6] set-valued map such that $\mathcal{K}(p)$ is a non-empty compact subset of $\mathcal{L}(T_p\mathcal{M})$. Suppose X is Lipschitz and

[6] We say the map $\mathcal{K}$ is upper semicontinuous if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all $q \in B_\delta(p)$ we have $P_{qp}\mathcal{K}(q) \subset \mathcal{K}(p) + \hat{B}_\varepsilon(0)$, where $\hat{B}_\varepsilon(0) := \{V \in \mathcal{L}(T_p\mathcal{M}) : \|V\| < \varepsilon\}$.

directionally differentiable in a neighborhood U of $p \in \mathcal{M}$. We say that X is semismooth at p with order μ with respect to $\mathcal{K}$ if for every $\varepsilon > 0$, there exists $\delta > 0$ such that for every $q \in B_\delta(p)$ and $H_q \in \mathcal{K}(q)$,

$$\left\|X(p) - P_{qp}[X(q) + H_q \exp_q^{-1} p]\right\| \leq \varepsilon d(p,q)^{1+\mu}. \tag{4.4}$$

Moreover, we say X is strongly semismooth at p with respect to $\mathcal{K}$ if there exist $C > 0, \delta > 0$ such that for every $q \in B_\delta(p)$ and $H_q \in \mathcal{K}(q)$,

$$\left\|X(p) - P_{qp}[X(q) + H_q \exp_q^{-1} p]\right\| \leq C d(p,q)^2. \tag{4.5}$$

Remark 4.1 Definition 2.12 and Definition 4.1 coincide when $\mathcal{K} = \partial X$ [27, Proposition 3.1].

When one of the nonsmooth terms vanishes, i.e. $X_2 = 0$ or $X_3 = 0$, we can choose $\mathcal{K} = \partial_B X_3$ or $\mathcal{K} = \partial_B X_2$. When both terms X_2 and X_3 are non-trivial, it is difficult to compute $\partial_B(X_2 + X_3)$ in general as we only know $\partial_B(X_2 + X_3) \subset \partial_B X_2 + \partial_B X_3$. In this case, we choose $\mathcal{K} = \partial_B X_2 + \partial_B X_3$. The next proposition guarantees that such a choice does not affect the semismoothness of $X_2 + X_3$.

Proposition 4.1 *Let X, Y be vector fields on $\mathcal{M}$ and $p \in \mathcal{M}$. Suppose X is semismooth at p with order μ with respect to $\mathcal{K}_X$, and Y is semismooth at p with order μ with respect to $\mathcal{K}_Y$. Then, $X + Y$ is semismooth at p with order μ with respect to $\mathcal{K}_X + \mathcal{K}_Y$.*

Proof By Definition 4.1, for every $\varepsilon > 0$, there exists $\delta > 0$ such that (4.4) holds for both X and Y. Then, for every $q \in B_\delta(p)$, $H_X \in \mathcal{K}_X(q)$, $H_Y \in \mathcal{K}_Y(q)$, we have

$$\begin{aligned} &\left\|(X+Y)(p) - P_{qp}[(X+Y)(q) + (H_X + H_Y)\exp_q^{-1} p]\right\| \\ &\leq \left\|X(p) - P_{qp}[X(q) + H_X \exp_q^{-1} p]\right\| + \left\|Y(p) - P_{qp}[Y(q) + H_Y \exp_q^{-1} p]\right\| < 2\varepsilon d(p,q)^{1+\mu}, \end{aligned}$$

where the first inequality follows from the linearity of P_{qp} and the triangle inequality of the norm. Therefore, the vector field $X + Y$ is semismooth with order μ with respect to $\mathcal{K}_X + \mathcal{K}_Y$. □

From the numerical perspective, we impose the assumption on $\mathcal{K}$ in our following analysis.

Assumption 4.2 *For every $p \in \mathcal{M}$, every element in $\mathcal{K}(p)$ is self-adjoint (symmetric).*

Since the Riemannian Hessian is self-adjoint and an element in $\partial X(p)$ is the limit of a sequence of self-adjoint operators, then the Clarke generalized covariant derivatives fulfill the above assumption.

Now, we are ready to present the semismooth Newton method for finding $p \in \mathcal{M}$ such that $X(p) = 0$.

Algorithm 4.1 *Choose $p_0 \in \mathcal{M}$, $\bar{\nu} \in (0,1]$ and let $\{\eta_k\} \subset \mathbb{R}_+$ be a sequence converging to 0. Set $\mu \in (0, 1/2), \delta \in (0,1), m_{\max} \in \mathbb{N}$, and $p, \beta_0, \beta_1 > 0$,*

Our algorithm repeats the following steps for $k = 0, 1, 2, \dots$

(i) Choose $H_k \in \mathcal{K}(p_k)$ and use the conjugate gradient (CG) method to find $V_k \in T_{p_k}\mathcal{M}$ such that

$$\|(H_k + \omega_k I)V_k + X(p_k)\| \leq \tilde{\eta}_k, \tag{4.6}$$

where $\omega_k := \|X(p_k)\|^{\bar{\nu}}$, $\tilde{\eta}_k := \min\left\{\eta_k, \|X(p_k)\|^{1+\bar{\nu}}\right\}$. Note that CG may fail when H_k is not positive definite, we choose the first-order direction $V_k = -X(p_k)$ in this case.

(ii) Choose the stepsize by one of the following linesearch methods:

(LS-I) If V_k is not a sufficient descent direction of φ, i.e. it does not satisfy

$$\langle -X(p_k), V_k\rangle \geq \min\{\beta_0, \beta_1 \|V_k\|^p\} \|V_k\|^2, \tag{4.7}$$

then, we set V_k to be $-X(p_k)$.

Next, find the minimum non-negative integer m_k such that

$$\varphi(R_{p_k}(\delta^{m_k} V_k)) \leq \varphi(p_k) + \mu\delta^{m_k}\langle X(p_k), V_k\rangle. \tag{4.8}$$

(LS-II) Find the minimum non-negative integer $m_k \leq m_{\max}$ such that

$$\|X(R_{p_k}(\delta^{m_k} V_k))\| \leq (1 - 2\mu\delta^{m_k})\|X(p_k)\|.$$

If m_k cannot be found, then we set $m_k = m_{\max}$.

(iii) *Set* $p_{k+1} = R_{p_k}(\delta^{m_k} V_k)$.

We make several remarks for the above numerical algorithm.

Remark 4.2 LS-I is a standard way to globalize the semismooth Newton method for minimizing smooth functions. Note that the CG method may fail, or V_k may not be a descent direction as H_k may not be positive definite. In both cases, Algorithm 4.1 reduces to the first-order method. It is noted that the condition (4.8) always holds for finite m_k as shown in Theorem 4.1, so the requirement $m_k \leq m_{\max}$ in LS-I is not needed. In Theorem 4.3, we show that the LS-I globalization method has a superlinear convergence under a "convexity assumption" and some regularity conditions.

Remark 4.3 LS-II is another way to globalize the Newton method [30, 50, 58]. In the Riemannian setting, the convergence result is established under the assumption that $\|X\|^2$ is differentiable [28]. However, $\|X\|^2 = \|X_1 + X_2 + X_3\|^2$ is not differentiable in general. Thus, LS-II does not have a convergence guarantee. In our experiments, we find both LS-I and LS-II have a similar performance for "convex problems" like (1.2) and LS-II is suitable for "nonconvex problems" like (1.3) and (1.4).

Remark 4.4 We can use the method in [22] to choose an initial point of Algorithm 4.1 to guarantee the condition (3.10). For LS-I, since it is a descent method, it suffices to choose p_0 such that $L_k(p_0) \leq \Phi$. This can be done by

$$p_0 = \begin{cases} x_{\text{feas}}, & L_k(x_{k-1}) > \Phi, \\ x_{k-1}, & L_k(x_{k-1}) \leq \Phi, \end{cases}$$

where x_{feas} and x_{k-1} is defined in Algorithm 3.1. The same initialization method is used for the LS-II method. However, the LS-II method has no convergence guarantee as it is not a descent method.

4.1 The Analysis of Global Convergence

We collect two technical lemmas for the subsequent analysis of Algorithm 4.1. The second part of the first one can be regarded as an approximated cosine law on manifolds.

Lemma 4.1 (Lemma 2.3 and 2.4 in [26]) *Fix* $p \in \mathcal{M}$*, the following properties hold.*

(i) *There exists* $r > 0$ *such that for every* $q \in B_r(p)$*, the exponential map* $\exp_q$ *is a diffeomorphism from* $\{v \in T_q\mathcal{M} : \|v\| < 2r\}$ *to* $B_{2r}(q)$.

(ii) *There exist* $K > 0$ *and* $r > 0$ *such that for every* $v, w \in T_q\mathcal{M}$ *with* $\|v\|, \|w\| < 2r$*, it holds*

$$|d(\exp_q v, \exp_q w)^2 - \|v - w\|^2| \leq K\|v\|^2\|w\|^2. \tag{4.9}$$

The second lemma is an extension of Proposition 2 in [3], which follows from the compactness of U and Taylor's theorem. Since it is the key lemma for extending the analysis of Algorithm 4.1 from the exponential map to general retractions, we give its proof for completeness.

Lemma 4.2 *Let* $U \subset \mathcal{M}$ *be a compact subset and* R *be a retraction, then there exist* $C, r > 0$ *such that for any* $q \in U$ *and* $v \in T_q\mathcal{M}$ *with* $\|v\| < r$*, the following inequalities hold*

$$d(R_q v, q) \leq C\|v\| \quad \text{and} \quad d(R_q v, \exp_q v) \leq C\|v\|^2. \tag{4.10}$$

Proof From the compactness of U, there exist $C, K, r_0 > 0$ such that Lemma 4.1 holds for every $q \in U$. By shrinking the constant r_0 if necessary, we could also assume that for every $q \in U$, $R_q v$ is defined for every $v \in T_q\mathcal{M}$ with $\|v\| < r_0$. For a fixed point $q \in \mathcal{M}$, let $\{e_i\}_{i\in[n]}$ be an orthonormal basis of $T_q\mathcal{M}$, we define the isomorphism $\psi_q : \mathbb{R}^n \to T_q\mathcal{M}$ as $\psi_q(w_1, \ldots, w_n) = \sum_{i=1}^n w_i e_i$, and the chart $\varphi_q : B_{r_0}(q) \to \mathbb{R}^n$ as $\varphi_q\left(\exp_q \psi_q(w)\right) = w$, where $\|w\|_{\mathbb{R}^n} < r_0$,[7] which implies that $(\psi_q \circ \varphi_q)(p) = \exp_q^{-1} p$ for every $p \in B_q(r_0)$.

For $\|v\| < r_0$, we define the curve $\gamma(t) := R_q(tv)$, then

$$d(R_q v, q) \leq \ell(\gamma|_{[0,1]}) = \int_0^1 \|\dot{\gamma}(t)\| \mathrm{d}t = \int_0^1 \|\mathrm{d}R_q|_{tv}[v]\| \mathrm{d}t \leq \tilde{C}_q \|v\|, \tag{4.11}$$

[7] Indeed, this is called the normal coordinate (see, e.g., [18, p. 86] and [47, p. 131]).

where $\tilde{C}_q := \sup_{\|v\|<r_0} \|\mathrm{d}R_q|_v\|$. Since $R_q v$ is C^2 with respect to q and v and U is compact, we know $\tilde{C} := \sup_{q\in U} \tilde{C}_q < \infty$. Thus, $\exp_q^{-1} R_q v$ is defined for $q \in U$ and $\|v\| < r := r_0/\tilde{C}$.

Let $\hat{B}_r := \{w \in \mathbb{R}^n : \|w\|_{\mathbb{R}^n} < r\}$, and $\hat{R}_q, \hat{E}_q : \hat{B}_r \to \mathbb{R}^n$ be the maps $\varphi_q \circ R_q \circ \psi_q|_{\hat{B}_r}$, $\varphi_q \circ \exp_q \circ \psi_q|_{\hat{B}_r}$, respectively. Since the differentials of $\exp_q$ and R_q are the same at 0, then the differentials of $\hat{E}_q$ and $\hat{R}_q$ also coincide at 0. Let $\hat{D}_q := \hat{E}_q - \hat{R}_q$, then Taylor's theorem states that for any $w \in \hat{B}_r$, there exists $t_w \in (0,1)$ such that $\|\hat{D}_q(w)\|_{\mathbb{R}^n} = \frac{1}{2}\|\nabla^2 \hat{D}_q(t_w w)[w,w]\|_{\mathbb{R}^n} \le \hat{C}_q \|w\|^2_{\mathbb{R}^n}$, where $\hat{C}_q := \frac{1}{2}\sup_{\|w\|_{\mathbb{R}^n}<r} \|\nabla^2 \hat{D}_q(w)\|$.

Since ψ_q is a linear isomorphism, then $(\psi_q \circ \hat{D}_q \circ \psi_q^{-1})(v) = (\psi_q \circ (\varphi_q \circ \exp_q - \varphi_q \circ R_q))(v) = (\psi_q \circ \varphi_q \circ \exp_q)(v) - (\psi_q \circ \varphi_q \circ R_q)(v) = v - \exp_q^{-1} R_q v$. Therefore, by Lemma 4.1, we have

$$d(R_q v, \exp_q v)^2 \le \|v - \exp_q^{-1} R_q v\|^2 + K\|\exp_q^{-1} R_q v\|^2 \|v\|^2 \le (\hat{C}_q^2 + \tilde{C}^2 K)\|v\|^4$$

for every $v \in T_q\mathcal{M}$ with $\|v\| < r$, where the last inequality follows from (4.11) and $\|\exp_q^{-1} R_q v\| = d(R_q v, q)$. Finally, since $R_q v$ and $\exp_q v$ are both C^2 with respect to q and v, $\hat{C}_q$ is uniformly bounded on the compact set U. Thus, the required constant C exists and is finite. □

The next theorem establishes the global convergence of Algorithm 4.1 with LS-I.

Theorem 4.1 *Let $\{p_k\}$ be the sequence generated by Algorithm 4.1 with LS-I. Suppose there exists $\delta > 0$ such that $\Omega := \{p \in \mathcal{M} : \varphi(p) \le \varphi(p_0) + \delta\}$ is compact. Then, the following properties hold:*

(i) For every $k \in \mathbb{N}$, there exists $m_k < \infty$ such that (4.8) *holds. Moreover, if there exist $k_0 \in \mathbb{N}$ and $\varepsilon_0 > 0$ such that the following holds for every $k > k_0$,*

$$\langle -X(p_k), V_k\rangle \ge \varepsilon_0 \|V_k\|^2, \tag{4.12}$$

then there exists an $m_{\max} \in \mathbb{N}$ such that $m_k \le m_{\max}$ for all $k \in \mathbb{N}$.

(ii) Let $\mathcal{P}$ be the set of accumulation points of $\{p_k\}$, then $\mathcal{P}$ is a non-empty set and every $p_* \in \mathcal{P}$ is a stationary point of φ, i.e., $X(p_*) = 0$.

(iii) $\lim_{k\to\infty} \|X(p_k)\| = 0$ *and* $\lim_{k\to\infty} \|V_k\| = 0$.

Proof (i) Let r, C be the constants such that Lemma 4.2 holds for $U = \Omega$. Since X is locally Lipschitz, for every $p \in \Omega$ we can find $S_p, L_p > 0$ such that X is L_p-Lipschitz in $B_{2S_p}(p)$. Since Ω is compact, we can find a finite set $\{q_1, \ldots, q_T\} \in \Omega$ such that $\bigcup_j B_{S_{q_j}}(q_j) \supset \Omega$. Define $R := \min\{r, \min_j S_{q_j}\}$, $\tilde{L} := \max_j L_{q_j}$. Since there exists $B_{S_{q_j}}(q_j) \ni p$ for every $p \in \Omega$, we find $B_R(p) \subset B_{2S_{q_j}}(q_j)$. Thus, X is $\tilde{L}$-Lipschitz in $B_R(p)$ for every $p \in \Omega$. Fix $p \in \Omega$ with $\varphi(p) < \varphi(p_0) + \delta$ and $V \in T_p\mathcal{M}$, and define $\gamma(t) = \exp_p(tV)$. Consider the function $\hat{\varphi} = \varphi \circ \gamma$, we know that $\hat{\varphi}'(t) = \langle X(\gamma(t)), \dot{\gamma}(t)\rangle$. Note that for $0 \le s, t < R/\|V\|$,

$$\begin{aligned}|\hat{\varphi}'(t) - \hat{\varphi}'(s)| &= |\langle X(\gamma(t)), \dot{\gamma}(t)\rangle - \langle X(\gamma(s)), \dot{\gamma}(s)\rangle| = \left|\left\langle X(\gamma(t)) - P_\gamma^{s\to t} X(\gamma(s)), \dot{\gamma}(t)\right\rangle\right| \\ &\le \left\|X(\gamma(t)) - P_\gamma^{s\to t} X(\gamma(s))\right\| \|\dot{\gamma}(t)\| \le \tilde{L}\ell(\gamma|_{[s,t]}) \|\dot{\gamma}(t)\| = \tilde{L}\|V\|^2 |t-s|,\end{aligned}$$

where the last inequality follows from the Lipschitzness of X and $\ell(\gamma|_{[s,t]}) = |t-s|\,\|V\|$ and $\|\dot{\gamma}(t)\| = \|\dot{\gamma}(0)\| = \|V\|$.[8] Thus, $\hat{\varphi}$ is $\tilde{L}\|V\|^2$-smooth. Besides, we know φ is Lipschitz on the compact set Ω since $\operatorname{grad}\varphi = X$ is bounded. Setting $p = p_k$, $V = V_k$ to be the quantities defined in Algorithm 4.1, $\hat{L}$ to be the Lipschitz constant of φ and $L := \tilde{L} + 2\hat{L}C$, we know when $0 \le t < R/\|V_k\|$, it holds that $|\varphi(R_{p_k}(tV_k)) - \varphi(\exp_{p_k}(tV_k))| \le \hat{L}d(R_{p_k}(tV_k), \exp_{p_k}(tV_k))$, and

$$\begin{aligned}\varphi(p_k) - \varphi(R_{p_k}(tV_k)) &\overset{(4.10)}{\ge} \varphi(p_k) - \varphi(\exp_{p_k}(tV_k)) - \hat{L}C\|tV_k\|^2 \\ &\ge \langle -X(p_k), tV_k\rangle - \frac{Lt^2}{2}\|V_k\|^2.\end{aligned} \tag{4.13}$$

[8] Since γ is a geodesic which is parallel to itself, i.e., $\nabla_{\dot{\gamma}}\dot{\gamma} = 0$, then $\frac{\mathrm{d}}{\mathrm{d}t}\|\dot{\gamma}(t)\|^2 = 2\langle \nabla_{\dot{\gamma}}\dot{\gamma}, \dot{\gamma}\rangle \equiv 0$. Therefore, we have $\|\dot{\gamma}(t)\| = \|\dot{\gamma}(0)\| = \|V\|$ and $\ell(\gamma|_{[s,t]}) = \int_s^t \|\dot{\gamma}(u)\|\mathrm{d}u = |t-s|\|\dot{\gamma}(0)\|$.

In view of (4.13), when $V_k \neq 0$, we have for $0 \leq t < R/\|V_k\|$,

$$\varphi(p_k) - \varphi(R_{p_k}(tV_k)) \begin{cases} \geq t\Big(1 - \frac{Lt}{2\min\{\beta_0, \beta_1\|V_k\|^p\}}\Big)\langle -X(p_k), V_k\rangle, & \text{if (4.7) holds,} \\ = t\Big(1 - \frac{Lt}{2}\Big)\langle -X(p_k), V_k\rangle, & \text{if } V_k = -X(p_k), \end{cases}$$

where β_0, β_1 and p are constants defined in Algorithm 4.1. Thus, for a fixed $k \in \mathbb{N}$, (4.8) holds whenever $\delta^{m_k} \leq \min\{R/\|V_k\|, 2\min\{1, \beta_0, \beta_1\|V_k\|^p\}(1-\mu)/L\}$. On the other hand, (4.8) automatically holds for $m_k = 0$ when $V_k = 0$. Thus, $\{\varphi(p_k)\}$ is a non-increasing sequence, which implies $\{p_k\} \subset \Omega$.

Next, we show that $\sup_k \|V_k\| < \infty$. We assume on the contrary that there exists a subsequence $\{k_j\}$ such that $\|V_{k_j}\| \to \infty$. The compactness of Ω and the continuity of X imply that $\sup_k \|X(p_k)\| < \infty$. Since V_{k_j} satisfies (4.7) or $V_{k_j} = -X(p_{k_j})$, it holds that $\|V_{k_j}\| \leq \|X(p_{k_j})\|/\min\{1, \beta_0, \beta_1\|V_{k_j}\|^p\}$, which is bounded and hence contradicts to the assumption that $\|V_{k_j}\| \to \infty$. Therefore, $\|V_k\|$ is uniformly bounded and (4.13) holds for all $k \in \mathbb{N}$ and $0 \leq t < R/\sup_k \|V_k\| =: \bar{t}$. Moreover, when (4.12) holds, we can choose m_k such that $\delta^{m_k} \leq \min\{\bar{t}, 2\min\{1, \varepsilon_0\}(1-\mu)/L\}$. Since the term on the right-hand side is independent of k and positive, $\{m_k\}$ is uniformly bounded.

(ii) Since Ω is compact and $\{p_k\} \subset \Omega$, the set $\mathcal{P}$ is non-empty. Let $p_* \in \mathcal{P}$, then there exists a subsequence $\{k_j\}$ such that $p_{k_j} \to p_*$ as $j \to \infty$. Now, we prove that $X(p_*) = 0$. Assume on the contrary that $X(p_*) \neq 0$, then we claim that $\inf_j \|V_{k_j}\| > 0$. Otherwise, when $\inf_j \|V_{k_j}\| = 0$, in view of (4.6), we have either $\|X(p_{k_j})\| = \|V_{k_j}\|$ or $\|X(p_{k_j})\| \leq \eta_{k_j} + \|(H_{k_j} + \omega_{k_j} I)V_{k_j}\|$. The upper-semicontinuity of $\mathcal{K}$ and ∂X shows that they are locally bounded and the compactness of Ω gives that $\mathcal{K} \cup \partial X$ is uniformly bounded in Ω. Therefore, $\sup_j \|H_{k_j}\| < \infty$. Since $\eta_{k_j} \to 0$, $\omega_{k_j} = \|X(p_{k_j})\|^{\bar{\nu}}$ is bounded and $\inf_j \|V_{k_j}\| = 0$, we conclude that $\liminf_{j\to\infty} \|X(p_{k_j})\| = 0$, which contradicts to $X(p_*) \neq 0$. Thus, $\|V_{k_j}\|$ is bounded away from zero.

From (4.8) and the fact that $\{\varphi(p_k)\}$ is non-increasing, we know

$$\begin{aligned} \varphi(p_{k_{j+1}}) - \varphi(p_{k_j}) \leq \varphi(p_{k_j+1}) - \varphi(p_{k_j}) &\leq \mu\delta^{m_{k_j}}\langle X(p_{k_j}), V_{k_j}\rangle \\ &\leq -\mu\delta^{m_{k_j}}\min\{\beta_0, \beta_1\|V_{k_j}\|^p\}\|V_{k_j}\|^2 < 0. \end{aligned}$$

Taking $j \to \infty$, the term on the left vanishes. Since $\|V_{k_j}\|$ is bounded away from both sides, $\delta^{m_{k_j}} \to 0$ as $j \to \infty$, and hence $m_{k_j} \to \infty$. Thus, for sufficiently large j, we know $t_j := \delta^{m_{k_j}-1} < \bar{t}$. As m_k is the smallest integer such that (4.8) holds, when replacing m_k with $m_k - 1$, we have

$$\varphi(q_j) > \varphi(p_{k_j}) + \mu t_j\langle X(p_{k_j}), V_{k_j}\rangle,$$

where $q_j := R_{p_{k_j}}(t_j V_{k_j})$. Together with (4.13), it holds

$$\mu t_j\langle X(p_{k_j}), V_{k_j}\rangle < \varphi(q_j) - \varphi(p_{k_j}) \leq t_j\langle X(p_{k_j}), V_{k_j}\rangle + \frac{Lt_j^2}{2}\|V_{k_j}\|^2,$$

and thus,

$$0 \leq (1-\mu)\min\{\beta_0, \beta_1\|V_{k_j}\|^p\}\|V_{k_j}\|^2 \leq (1-\mu)\langle -X(p_{k_j}), V_{k_j}\rangle < \frac{Lt_j}{2}\|V_{k_j}\|^2.$$

As $\{V_{k_j}\}$ is bounded, by passing to a subsequence if necessary, we assume that $\lim_{j\to\infty} V_{k_j} = V_* \neq 0$. Dividing the above inequality by $\|V_{k_j}\|^2$ and letting $j \to \infty$, we obtain $(1-\mu)\min\{\beta_0, \beta_1\|V_*\|^p\} = 0$, which leads to a contradiction since $\mu \in (0, 1/2)$. Thus, we conclude that $X(p_*) = 0$. From the choice of V_{k_j} in LS-I, we can also conclude that $\lim_{j\to\infty} V_{k_j} = 0$.

(iii) If the whole sequence $\{\|X(p_k)\|\}$ contains an accumulation point other than 0, then there exists a subsequence $\{\|X(p_{k_j})\|\}$ such that $\|X(p_{k_j})\| \to \alpha > 0$ as $j \to \infty$. By passing to a subsequence if necessary, we could assume $p_{k_j} \to p_*$ as $j \to \infty$. Thus, $p_* \in \mathcal{P}$ and the above discussion gives that $X(p_*) = 0$, which leads to a contradiction. Similarly, it can be shown that the only accumulation point of $\{\|V_k\|\}$ is 0. $\square$

To obtain the locally superlinear convergence rate of Algorithm 4.1, we first introduce some basic results of Riemannian manifolds and technical lemmas.

Lemma 4.3 *Suppose X is a locally Lipschitz vector field on $\mathcal{M}$, $p_* \in \mathcal{M}$, $\mathcal{K}$ is defined in Definition 4.1 and all elements in $\mathcal{K}(p_*)$ are nonsingular and there exists $\lambda > 0$ such that $\lambda \geq \max\left\{\|H^{-1}\| : H \in \mathcal{K}(p_*)\right\}$. Then, for every $\varepsilon > 0$ and $\varepsilon\lambda < 1$, there exists a neighborhood U of p_* such that all elements in $\mathcal{K}(p)$ are nonsingular for $p \in U$ and*

$$\left\|H^{-1}\right\| \leq \frac{\lambda}{1-\varepsilon\lambda}, \ \forall\, p \in U, \ H \in \mathcal{K}(p).$$

Proof The proof is the same as the proof in [27, Lemma 4.2] in which $\mathcal{K} = \partial X$ and only the upper-semicontinuity of ∂X is used. □

Below is a second-order Taylor theorem on manifolds, a generalization of Theorem 2.3 in [36].

Lemma 4.4 *Suppose $\varphi : \mathcal{M} \to \mathbb{R}$ is a continuously differentiable function with Lipschitz gradient, and $p, q \in \mathcal{M}$. Let $\gamma : [0,1] \to \mathcal{M}$ be a geodesic joining p and q. Then, there exist $\xi \in (0,1)$ and $M_\xi \in \partial \operatorname{grad} \varphi(\gamma(\xi))$ such that*

$$\varphi(q) - \varphi(p) = \langle \operatorname{grad} \varphi(p), \dot\gamma(0)\rangle + \frac{1}{2}\left\langle P_\gamma^{\xi\to 0} M_\xi P_\gamma^{0\to\xi}\dot\gamma(0), \dot\gamma(0)\right\rangle.$$

Proof Define $\hat\varphi = \varphi \circ \gamma$. We know that $\hat\varphi$ is continuously differentiable with Lipschitz gradient in $[0,1]$. By Theorem 2.3 in [36], there exist $\xi \in (0,1), \hat M_\xi \in \partial^2\hat\varphi(\xi)$ such that $\hat\varphi(1) - \hat\varphi(0) = \hat\varphi'(0) + \frac{\hat M_\xi}{2}$. Note that due to $\hat\varphi'(t) = \langle \operatorname{grad}\varphi(\gamma(t)), \dot\gamma(t)\rangle$ and $\nabla_{\dot\gamma}\dot\gamma = 0$, we find $\hat\varphi''(t) = \left\langle \nabla_{\dot\gamma(t)} \operatorname{grad}\varphi(\gamma(t)), \dot\gamma(t)\right\rangle$, whenever $\hat\varphi'$ is differentiable at t. From the linearity of the parallel transport, it suffices to consider the case $\hat M_\xi \in \partial_B^2\hat\varphi(\xi)$, i.e. there exists $\{t_k\} \to \xi$ such that $\hat\varphi''$ exists at t_k and $\hat\varphi''(t_k) \to \hat M_\xi$. Since $\hat\varphi''(t_k) = \langle \nabla_{\dot\gamma(t_k)} \operatorname{grad}\varphi(\gamma(t_k)), \dot\gamma(t_k)\rangle = \langle P_\gamma^{t_k\mapsto\xi}\nabla_{P_\gamma^{\xi\mapsto t_k}\dot\gamma(\xi)} \operatorname{grad}\varphi(\gamma(t_k)), \dot\gamma(\xi)\rangle$ and note that the parallel transport is an isometry, then there is $M_\xi \in \partial \operatorname{grad}\varphi(\gamma(\xi))$ such that $\hat M_\xi = \langle M_\xi\dot\gamma(\xi), \dot\gamma(\xi)\rangle$. Note $\dot\gamma(\xi) = P_\gamma^{0\to\xi}\dot\gamma(0)$, then the conclusion holds. □

A direct consequence of Lemma 4.3 and 4.4 provides the following theorem that characterizes the second-order optimality conditions of (4.3).

Theorem 4.2 *Suppose $\varphi : \mathcal{M} \to \mathbb{R}$ is continuously differentiable with locally Lipschitz gradient, then the following properties hold*

– *(Second-order necessary condition). Suppose $x_* \in \mathcal{M}$ is a local minimum of φ, then for any $v \in T_{x_*}\mathcal{M}$, there is $H_v \in \partial \operatorname{grad}\varphi(x_*)$ such that $\langle H_v v, v\rangle \geq 0$.*
– *(Second-order sufficient condition). Suppose $x_* \in \mathcal{M}$ such that $\operatorname{grad}\varphi(x_*) = 0$ and all elements in $\partial \operatorname{grad}\varphi(x_*)$ are positive definite, then x_* is a strict local minimum of φ.*

Remark 4.5 Using Lemma 4.4, we can see that the second-order sufficient condition implies the strongly geodesic convexity of φ near p_* as defined in [67].

The next two lemmas give the local analysis around the critical point of X. The following lemma analyzes the B-differentiability on manifolds.

Lemma 4.5 *Let X be a locally Lipschitz vector field on $\mathcal{M}$ and $p \in \mathcal{M}$. If X is directionally differentiable at p and $X(p) = 0$, then $\|P_{\exp_p v, p}X(\exp_p v) - \nabla X(p; v)\| = o(\|v\|)$ as $T_p\mathcal{M} \ni v \to 0$.*

Proof Suppose the conclusion does not hold, then there exist $\delta > 0$ and $v_k \in T_p\mathcal{M} \to 0$ such that

$$\|P_{\exp_p v_k, p}X(\exp_p v_k) - \nabla X(p; v_k)\| \geq \delta\,\|v_k\|.$$

By Lemma 4.1 and the locally Lipschitz condition on X, there exists $r_0 > 0$ such that for every $p_1, p_2 \in B_{r_0}(p)$, there exists a unique shortest geodesic joining p_1 and p_2 and X is L-Lipschitz in $B_{r_0}(p)$. Let $0 < r < r_0$, by taking the subsequence of $\{v_k\}$, we assume that $\bar v_k := \frac{r v_k}{\|v_k\|} \to v \in T_p\mathcal{M}$ with $\|v\| = r$. We define $t_k = \|v_k\|/r$, $q_k = \exp_p v_k$, $r_k = \exp_p \bar{v}_k$ and $s_k = \exp_p t_k v$, then it holds that $\{q_k, r_k, s_k\} \subset B_{r_0}(p)$, and hence we have

$$\begin{aligned}\|P_{q_k,p}X(q_k) - \nabla X(p; v_k)\| &\leq \|P_{q_k,p}X(q_k) - P_{s_k,p}X(s_k)\| + \|P_{s_k,p}X(s_k) - \nabla X(p; v_k)\| \\ &\leq \underbrace{\|P_{p,s_k}P_{q_k,p}X(q_k) - P_{q_k,s_k}X(q_k))\|}_{\text{(A)}} + \underbrace{\|P_{q_k,s_k}X(q_k) - X(s_k)\|}_{\text{(B)}} \\ &\quad + \underbrace{\|P_{s_k,p}X(s_k) - t_k\nabla X(p; v)\|}_{\text{(C)}} + \underbrace{\|\nabla X(p; v_k) - t_k\nabla X(p; v)\|}_{\text{(D)}}.\end{aligned} \tag{4.14}$$

From Lemma 10 in [40], there exists $C > 0$ such that

$$\text{(A)} \leq \|P_{s_k,q_k}P_{p,s_k}P_{q_k,p} - \mathrm{id}\|\|X(q_k)\| \leq C\max(d(p, q_k), d(q_k, s_k))\|X(q_k)\| = o(t_k), \tag{4.15}$$

since $\|X(q_k)\| \leq L\|v_k\| \to 0$ as $v_k \to 0$. Moreover, since X is locally Lipschitz and directional differentiable at p, using Lemma 4.1, as $k \to \infty$, we have

$$\text{(B)} \leq Ld(q_k, s_k) = o(t_k), \quad \text{(C)} = \|P_{s_k,p}X(s_k) - X(p) - t_k\nabla X(p; v)\| = o(t_k), \tag{4.16}$$

where the first estimate follows from Definition 2.10 and $\ell(\gamma) = d(q_k, s_k)$. Let $q_k^t = \exp_p t\bar{v}_k$ and $q_v^t = \exp_p tv$, from Lemma 10 in [40] and Lemma 4.1, we have when $t \to 0$,

$$\begin{aligned}\|P_{q_k^t,p}X(q_k^t) - P_{q_v^t,p}X(q_v^t)\| &\leq \|P_{p,q_v^t}P_{q_k^t,p}X(q_k^t) - P_{q_k^t,q_v^t}X(q_k^t)\| + \|P_{q_k^t,q_v^t}X(q_k^t) - X(q_v^t)\| \\ &\leq \|P_{q_v^t,q_k^t}P_{p,q_v^t}P_{q_k^t,p} - \mathrm{id}\|\|X(q_k^t)\| + Ld(q_k^t, q_v^t) = O(t^2) + O(t)\|\bar{v}_k - v\|,\end{aligned}$$

where the constants in the big-O notation depend on p, v and $\mathcal{M}$. An intermediate result of the above estimate is

$$\|\nabla X(p; v_k) - \nabla X(p; v)\| \leq O(\|\bar{v}_k - v\|) = o(1). \tag{4.17}$$

Thus, the last term in (4.14) is

$$\text{(D)} = t_k\|\nabla X(p; \bar{v}_k) - \nabla X(p; v)\| = o(t_k) \tag{4.18}$$

Combining (4.15), (4.16) and (4.18), we find that $\|P_{q_k,p}X(q_k) - \nabla X(p; v_k)\| = o(t_k) = o(\|v_k\|)$, which contradicts with our assumption. □

Lemma 4.6 *Suppose $\{p_k\} \subset \mathcal{M}$ and $V_k \in T_{p_k}\mathcal{M}$ satisfy $\lim_{k\to\infty} p_k = p \in \mathcal{M}$, $\lim\limits_{k\to\infty} \|V_k\| = 0$ and $d(R_{p_k}V_k, p) = o(d(p_k, p))$. Then, $\lim\limits_{k\to\infty} \|V_k\|/d(p_k, p) = 1$ and $d(R_{p_k}V_k, p) = o(\|V_k\|)$.*

Proof Define $q_{k+1} := R_{p_k}V_k$. Let $B_r(p), K$ be the quantities in Lemma 4.1. Without the loss of generality, we may assume $p_k \in B_r(p)$ and $\|v\| < r$ for every k, and shrink the ball $B_r(p)$ such that (4.10) holds in $B_r(p)$. Using Lemma 4.1, we know the following inequality holds for every geodesic triangle in $B_r(p)$ whose edges are a, b, c:

$$a^2 \leq b^2 + c^2 - 2bc\cos A + Kb^2c^2 \leq (b + c)^2 + Kb^2c^2,$$

where A is the angle between b and c. Let $a = d(p_k, q_{k+1})$, $b = d(p_k, p)$, $c = d(q_{k+1}, p)$, then we have

$$\limsup_{k\to\infty} \frac{d(p_k, q_{k+1})^2}{d(p_k, p)^2} \leq K\limsup_{k\to\infty} d(q_{k+1}, p)^2 + \limsup_{k\to\infty}\left(1 + \frac{d(q_{k+1}, p)}{d(p_k, p)}\right)^2 = 1. \tag{4.19}$$

Define $a = d(p_k, p)$, $b = d(p_k, q_{k+1})$, $c = d(q_{k+1}, p)$, note that $d(q_{k+1}, p) = o(d(p_k, p))$, then for any $\varepsilon > 0$ there exists $k_\varepsilon > 0$ such that for any $k \geq k_\varepsilon$, it holds that $d(q_{k+1}, p)^2 \leq \varepsilon d(p_k, p)^2 \leq 2\varepsilon d(q_{k+1}, p)^2 + 2\varepsilon d(p_k, q_{k+1})^2 + \varepsilon Kd(q_{k+1}, p)^2 d(p_k, q_{k+1})^2$. Thus, we know

$$d(q_{k+1}, p)^2 \leq \frac{2\varepsilon}{1 - 2\varepsilon - \varepsilon Kd(p_k, q_{k+1})^2} d(p_k, q_{k+1})^2.$$

From (4.19) and the fact that $p_k \to p$ as $k \to \infty$, there exists $K_0 > 0$ such that $Kd(p_k, q_{k+1})^2 \leq 2Kd(p_k, p)^2 \leq 2$ for $k \geq K_0$. Thus, $d(q_{k+1}, p)^2 \leq 4\varepsilon d(p_k, q_{k+1})^2$ for all $\varepsilon \in (0, 1/8)$ and $k \geq \max(k_\varepsilon, K_0)$, i.e., $d(q_{k+1}, p)^2 = o(d(p_k, q_{k+1})^2)$. From (4.10), we have

$$d(p_k, q_{k+1}) \leq d(p_k, \exp_{p_k} V_k) + d(\exp_{p_k} V_k, q_{k+1}) \leq \|V_k\| + C\|V_k\|^2,$$

and hence $d(q_{k+1}, p) = o(d(p_k, q_{k+1})) = o(\|V_k\|)$.

On the other hand,

$$\limsup_{k\to\infty} \frac{d(p_k, p)^2}{d(p_k, q_{k+1})^2} \leq K \limsup_{k\to\infty} d(q_{k+1}, p)^2 + \limsup_{k\to\infty} \left(1 + \frac{d(q_{k+1}, p)}{d(p_k, q_{k+1})}\right)^2 = 1.$$

Combining with (4.19), we have $\lim_{k\to\infty} \frac{d(p_k,p)}{d(p_k,q_{k+1})} = 1$. $\square$

Theorem 4.3 *Under the same assumptions as in Theorem 4.1, let $\mathcal{K}$ be the set-valued map used in Algorithm 4.1. Denote p_* be any accumulation point of $\{p_k\}$. If X is semismooth at p_* with order ν with respect to $\mathcal{K} \cup \partial X$, and all elements of $\mathcal{K}(p_*) \cup \partial X(p_*)$ are positive definite, then we have $p_k \to p_*$ as $k \to \infty$ and for sufficiently large k, it holds*

$$d(p_{k+1}, p_*) \leq O\big(d(p_k, p_*)^{1+\min\{\nu,\bar{\nu}\}}\big),$$

where $\bar{\nu} \in (0, 1]$ is the parameter defined in Algorithm 4.1.

Proof Since $\mathcal{K}(p_*) \cup \partial X(p_*)$ is positive definite, by Lemma 4.3 and the upper-semicontinuity of $\mathcal{K}$ and ∂X, we can find a neighborhood $U \ni p_*$ and constants $\omega, M > 0$ such that $M\|V\|^2 \geq \langle HV, V\rangle \geq 2\omega\|V\|^2$ for every $p \in U, H \in \mathcal{K}(p) \cup \partial X(p), V \in T_p\mathcal{M}$. By Lemma 4.1 and the semismooth condition of X, there exists $r_0 \in (0, 1/2]$ such that $B_{r_0}(p_*) \subset U$, the unique shortest geodesic joining points in $B_{r_0}(p_*)$ exists, and

$$\big\|X(q) + H_q \exp_q^{-1} p_*\big\| \leq d(p_*, q)^{1+\nu},\ \forall q \in B_{r_0}(p_*), \tag{4.20}$$

and X is L-Lipschitz in $B_{r_0}(p_*)$. From Lemma 4.2 and the compactness of Ω (the set defined in Theorem 4.1), we can further assume that (4.10) holds for $q \in \Omega$ and $\|v\| < r_0$, and denote the constant in (4.10) as $C > 0$.

From Theorem 4.1, both $\|V_k\|$ and $\|X(p_k)\|$ converge to zero as $k \to \infty$. Given arbitrary $0 < r < r_0$, there exists some $K_0 > 0$ such that $\omega_k = \|X(p_k)\|^{\bar{\nu}} < 1/2$, $\|V_k\| < r/(2C)$, $\beta_1\|V_k\|^p \leq \omega$ and $(2M + 1)^{1+\bar{\nu}}\|V_k\|^{\bar{\nu}} \leq \omega$ for $k \geq K_0$. From Lemma 4.4, we know $\varphi(q) - \varphi(p_*) \geq \omega d(q, p_*)^2$ whenever $q \in B_r(p_*)$. Since p_* is an accumulation point, we can find $K_1 \geq K_0 > 0$ such that $\varphi(p_{K_1}) - \varphi(p_*) < \omega r^2/4$ and $p_{K_1} \in B_r(p_*)$. Note that from (4.10), $d(p_{K_1}, p_{K_1+1}) \leq C\delta^{m_{K_1}}\|V_{K_1}\| < r/2$. Then, $d(p_{K_1+1}, p_*) \leq d(p_{K_1}, p_*) + d(p_{K_1}, p_{K_1+1}) < r$, i.e., $p_{K_1+1} \in B_r(p_*)$. Since $\varphi(p_k)$ is non-increasing, we still have $\varphi(p_{K_1+1}) - \varphi(p_*) < \omega r^2/4$. By induction, we know $\{p_k\}_{k\geq K_1} \subset B_r(p_*)$ which implies that $p_k \to p_*$ as $k \to \infty$. Below we assume that $k \geq K_1$.

By the positive definiteness of $\mathcal{K}\cup\partial X$ near p_* and the self-adjoint assumption of $\mathcal{K}$ (Assumption 4.2), the CG method in Algorithm 4.1 is able to find a direction V_k satisfying (4.6). Thus, we know $\|X(p_k)\| \leq \|H_k + \omega_k I\| \, \|V_k\| / (1 - \|X(p_k)\|^{\bar{\nu}}) \leq (2M + 1)\, \|V_k\|$. Then, we obtain

$$\begin{aligned} \langle -X(p_k), V_k\rangle &= \langle (H_k + \omega_k I)V_k, V_k\rangle - \langle (H_k + \omega_k I)V_k + X(p_k), V_k\rangle \\ &\geq 2\omega\, \|V_k\|^2 - \tilde{\eta}_k\, \|V_k\| \geq 2\omega\, \|V_k\|^2 - (2M + 1)^{1+\bar{\nu}}\, \|V_k\|^{2+\bar{\nu}} \geq \omega\|V_k\|^2. \end{aligned} \tag{4.21}$$

Thus, the condition (4.7) holds as $\omega \geq \beta_1\|V_k\|^p$. Note that $\|(H_k + \omega_k I)^{-1}\| \leq (2\omega)^{-1}$, and $\|X(p_k)\| \leq Ld(p_k, p_*)$. Define $C := 2\max\{(2\omega)^{-1}, L^{1+\bar{\nu}} + (2\omega)^{-1}L\}$, by the definition of $\omega_k, \tilde{\eta}_k$, we have

$$\begin{aligned} \|V_k - \exp_{p_k}^{-1} p_*\| &\leq \|(H_k + \omega_k I)^{-1}X(p_k) + \exp_{p_k}^{-1} p_*\| + (2\omega)^{-1}\tilde{\eta}_k \\ &\leq (2\omega)^{-1}\|X(p_k) + (H_k + \omega_k I)\exp_{p_k}^{-1} p_*\| + (2\omega)^{-1}\tilde{\eta}_k \\ &\leq (2\omega)^{-1}\left(\|X(p_k) + H_k \exp_{p_k}^{-1} p_*\| + \omega_k\|\exp_{p_k}^{-1} p_*\|\right) + (2\omega)^{-1}\tilde{\eta}_k \\ &\leq (2\omega)^{-1}\left(d(p_k, p_*)^{1+\nu} + \omega_k d(p_k, p_*)\right) + (2\omega)^{-1}\tilde{\eta}_k \leq Cd(p_k, p_*)^{1+\min\{\nu,\bar{\nu}\}}. \end{aligned}$$

Using (4.21) we find $\|V_k\| \leq \|X(p_k)\|/\omega \leq Ld(p_k, p_*)/\omega$. From Lemma 4.1, we know

$$\begin{aligned} d(R_{p_k}V_k, p_*)^2 &\overset{(4.10)}{=} O\big(d(\exp_{p_k} V_k, p_*)^2\big) \\ &\leq O\big(\|V_k - \exp_{p_k}^{-1} p_*\|^2 + K\|V_k\|^2 d(p_k, p_*)^2\big) \leq O\big(d(p_k, p_*)^{2+2\min\{\nu,\bar{\nu}\}}\big), \end{aligned} \tag{4.22}$$

where K is the constant defined in Lemma 4.1. To complete the proof, we need to show that for $\mu \in (0, 1/2)$ and sufficiently large k, the linesearch condition

$$\varphi(R_{p_k}V_k) \leq \varphi(p_k) + \mu\, \langle X(p_k), V_k\rangle\,, \tag{4.23}$$

holds with $m_k = 0$. Define $U_k := \exp_{p_*}^{-1} p_k$, $W_k := \exp_{p_*}^{-1} R_{p_k} V_k$. Applying Lemma 4.4 to p_*, p_k and $R_{p_k} V_k$, we have

$$\varphi(p_k) = \varphi(p_*) + \frac{1}{2}\langle \tilde{R}_k U_k, U_k\rangle, \text{ and } \varphi(R_{p_k} V_k) = \varphi(p_*) + \frac{1}{2}\langle \tilde{M}_k W_k, W_k\rangle,$$

where $\tilde{R}_k = P_{p_r p_*} R_k P_{p_* p_r}$, $\tilde{M}_k = P_{p_m p_*} M_k P_{p_* p_m}$, R_k is a Clarke generalized covariant derivative at $p_r := \exp_{p_*}(\theta_k U_k)$ for some $\theta_k \in (0,1)$, and M_k is a Clarke generalized covariant derivative at $p_m := \exp_{p_*}(\xi_k W_k)$ for some $\xi_k \in (0,1)$. Subtracting the above two equations, we know

$$\varphi(R_{p_k} V_k) - \varphi(p_k) = -\frac{1}{2}\langle \tilde{R}_k U_k, U_k\rangle + O\left(\|W_k\|^2\right).$$

Thus, we have

$$\begin{aligned}\varphi(R_{p_k} V_k) - \varphi(p_k) - \frac{1}{2}\left\langle X(p_k), V_k\right\rangle &= -\frac{1}{2}\langle \tilde{R}_k U_k, U_k\rangle - \frac{1}{2}\left\langle X(p_k), V_k\right\rangle + O\left(\|W_k\|^2\right) \\ &= -\frac{1}{2}\langle V_k, X(p_k) - P_{p_* p_k}\tilde{R}_k U_k\rangle - \frac{1}{2}\langle P_{p_k p_*} V_k + U_k, \tilde{R}_k U_k\rangle + O\left(\|W_k\|^2\right).\end{aligned} \tag{4.24}$$

Since $P_{p_r p_*} \exp_{p_r}^{-1} p_* = -\exp_{p_*}^{-1} p_r = -\theta_k U_k$ and $X(p_*) = 0$, then by the semismoothness $P_{p_r p_*} X(p_r) - \theta_k \tilde{R}_k U_k = P_{p_r p_*}(X(p_r) + R_k \exp_{p_r}^{-1} p_*) - X(p_*) = o(\|\theta_k U_k\|)$. From Lemma 4.5, we know $P_{p_r p_*} X(p_r) = \theta_k \nabla X(p_*; U_k) + o(\|\theta_k U_k\|)$ and $P_{p_k p_*} X(p_k) = \nabla X(p_*; U_k) + o(\|U_k\|)$, which implies

$$P_{p_k p_*} X(p_k) = \tilde{R}_k U_k + o(\|U_k\|). \tag{4.25}$$

Moreover, applying $\lim\limits_{k\to\infty} p_k = p_*$ and (4.22) to Lemma 4.6, we have

$$\frac{\|V_k\|}{\|U_k\|} = \frac{\|V_k\|}{d(p_k, p_*)} \to 1 \text{ and } \|W_k\| = d(R_{p_k} V_k, p_*) = o(\|V_k\|) \text{ as } k \to \infty. \tag{4.26}$$

From Lemma 4.1, it holds

$$\begin{aligned}\|V_k - \exp_{p_k}^{-1} p_*\|^2 &\le d(\exp_{p_k} V_k, p_*)^2 + K\|V_k\|^2 \|\exp_{p_k}^{-1} p_*\|^2 \\ &\overset{(4.10)}{\le} \|W_k\|^2 + o(\|V_k\|^2) \overset{(4.26)}{=} o(\|V_k\|^2).\end{aligned} \tag{4.27}$$

Then, the above three displays imply that

$$\begin{aligned}|\langle V_k, X(p_k) - P_{p_* p_k}\tilde{R}_k U_k\rangle| &\le \|V_k\|\,\|X(p_k) - P_{p_* p_k}\tilde{R}_k U_k\| \le o(\|V_k\|\,\|U_k\|) = o\left(\|V_k\|^2\right), \\ |\langle P_{p_k p_*} V_k + U_k, \tilde{R}_k U_k\rangle| &\le \|V_k - \exp_{p_k}^{-1} p_*\|\|\tilde{R}_k U_k\| \le o\left(\|V_k\|^2\right).\end{aligned} \tag{4.28}$$

Combining (4.21), (4.24) and (4.28), we have

$$\begin{aligned}\varphi(R_{p_k} V_k) - \varphi(p_k) &= \frac{1}{2}\left\langle X(p_k), V_k\right\rangle + o\left(\|V_k\|^2\right) \\ &\le \mu\left\langle X(p_k), V_k\right\rangle - \omega\left(\frac{1}{2} - \mu\right)\|V_k\|^2 + o\left(\|V_k\|^2\right).\end{aligned}$$

Thus, (4.23) holds for sufficiently large k and $m_k = 0$ in LS-I. $\square$

4.3 The Semismooth Condition on Submanifolds

In this section, we assume that the manifold $\mathcal{M}$ is embedded in an ambient space $\bar{\mathcal{M}}$ and the vector field X on $\mathcal{M}$ is the restriction of a vector field $\bar{X}$ defined on $\bar{\mathcal{M}}$. Our goal is to answer a natural question: can the Lipschitzness, the directional differentiability and the semismoothness of X be inherited from those of $\bar{X}$? Throughout this section, we make the following assumptions.

Assumption 4.3 *The n-dimensional Riemannian manifold $\mathcal{M}$ is a compact embedded submanifold of the d-dimensional Riemannian manifold $\bar{\mathcal{M}}$, i.e., $\mathcal{M} \subset \bar{\mathcal{M}}$ and the inclusion map $\iota : \mathcal{M} \hookrightarrow \bar{\mathcal{M}}$ is a smooth injection, and the differential of ι is injective, $\mathcal{M}$ and $\iota(\mathcal{M})$ are homeomorphism, and the Riemannian metric of $\mathcal{M}$ is inherited from that of $\bar{\mathcal{M}}$ (see, e.g., [47, Chapter 8]).*

Assumption 4.4 *The vector field X is the restriction of $\bar{X}$ on $\mathcal{M}$, where $\bar{X}$ is a vector field on $\bar{\mathcal{M}}$.*

It is noted that the compactness of $\mathcal{M}$ is not essential as the semismoothness involves only local properties. For simplicity, we impose the assumption that $\mathcal{M}$ is compact. To answer the aforementioned question, we need to introduce some concepts about submanifolds. Notations in Table 1 are defined for $\mathcal{M}$, and we add a line over them to represent corresponding notations on $\bar{\mathcal{M}}$, e.g., $\overline{\exp}_p$, $\bar{P}_\gamma^{0\to t}$ and $\bar{\nabla}_X Y$ are the exponential map, the parallel transport and the Levi-Civita connection on $\bar{\mathcal{M}}$, respectively. We use $d_{\mathcal{M}}$ and $d_{\bar{\mathcal{M}}}$ to denote the distance on $\mathcal{M}$ and $\bar{\mathcal{M}}$, respectively.

According to [47, p. 226], for every $p \in \mathcal{M}$, the tangent space $T_p\bar{\mathcal{M}}$ can be decomposed into $T_p\mathcal{M} \oplus (T_p\mathcal{M})^\perp$, i.e., $v \in T_p\bar{\mathcal{M}}$ can be uniquely written as $v_\top + v_\perp$, where $v_\top \in T_p\mathcal{M}$ and $v_\perp \in (T_p\mathcal{M})^\perp$. Let Y, Z be smooth vector fields on $\mathcal{M}$, and $\bar{Y}, \bar{Z}$ be smooth vector fields on $\bar{\mathcal{M}}$ such that $Y = \bar{Y}|_{\mathcal{M}}$ and $Z = \bar{Z}|_{\mathcal{M}}$, then the <u>second fundamental form of $\mathcal{M}$</u> is defined as $\mathrm{II}(Y, Z) := \left(\bar{\nabla}_{\bar{Y}}\bar{Z}\right)_\perp$. Proposition 8.1 in [47] shows that $\mathrm{II}(Y, Z)$ is independent of the extensions $\bar{Y}, \bar{Z}$; $\mathrm{II}(Y, Z)$ is $C^\infty(\mathcal{M})$-bilinear; and the value of $\mathrm{II}(Y, Z)$ at $p \in \mathcal{M}$ depends only on $Y(p), Z(p)$. Thus, we could safely write $\mathrm{II}(v, w)$ for $v, w \in T_p\mathcal{M}$, and $\mathrm{II}(Y, Z)$ is still valid even for non-smooth vector fields Y, Z. Moreover, the Gauss formula [47, Theorem 8.2, Corollary 8.3] relates Levi-Civita connections on $\mathcal{M}$ and $\bar{\mathcal{M}}$ as follows:

$$\bar{\nabla}_Y Z = \nabla_Y Z + \mathrm{II}(Y, Z) \quad \text{and} \quad \bar{\nabla}_{\dot{\gamma}} Z = \nabla_{\dot{\gamma}} Z + \mathrm{II}(\dot{\gamma}, Z), \tag{4.29}$$

where $\gamma : (-1, 1) \to \mathcal{M}$ is a smooth curve. Note that $\mathrm{II}(Y, Z)$ is orthogonal to the tangent space of $\mathcal{M}$, the above equations imply that $\nabla_Y Z$ is the projection of $\bar{\nabla}_Y Z$ onto $T\mathcal{M}$, i.e.,

$$\nabla_Y Z = (\bar{\nabla}_Y Z)_\top. \tag{4.30}$$

Before presenting the main theorem of this section, we need to introduce the directional differentiability in the Hadamard sense whose Euclidean counterpart can be found in [13, Definition 2.45].

Definition 4.2 A vector field $X : \mathcal{M} \to T\mathcal{M}$ is <u>directionally differentiable at $p \in \mathcal{M}$ in the Hadamard sense</u> if it is directionally differentiable at p and

$$\nabla X(p; v) = \lim_{\substack{t \downarrow 0 \\ T_p\mathcal{M} \ni v' \to v}} \frac{1}{t}\left[P_{\exp_p(tv'), p} X(\exp_p(tv')) - X(p)\right]. \tag{4.31}$$

It is clear that the above definition is a stronger version than the directional differentiability in Definition 2.8. In Euclidean spaces, the Lipschitzness and the directional differentiability can imply the directional differentiability in the Hadamard sense [13, Proposition 2.49]. However, it is unclear whether this result can be generalized to Riemannian manifolds.

Theorem 4.4 *Let $\bar{\mathcal{K}} : \bar{\mathcal{M}} \rightrightarrows \mathcal{L}(T\bar{\mathcal{M}})$ be an upper semicontinuous set-valued map such that $\bar{\mathcal{K}}(p)$ is a non-empty compact set for every $p \in \bar{\mathcal{M}}$, and $\mathcal{K} : \mathcal{M} \rightrightarrows \mathcal{L}(T\mathcal{M})$ be a set-valued map satisfying the conditions in Definition 4.1. Suppose X is continuous on $\mathcal{M}$, then the following statements hold.*

(i) If $\bar{X}$ is locally Lipschitz at $p \in \mathcal{M}$, then X is locally Lipschitz at p (see Definition 2.10).
(ii) If $\bar{X}$ is directionally differentiable at $p \in \mathcal{M}$ in the Hadamard sense, then X is directionally differentiable at p in the Hadamard sense, and the directional derivative is

$$\nabla X(p; v) = \bar{\nabla} X(p; v) - \mathrm{II}(v, X(p)) = (\nabla \bar{X}(p; v))_\top. \tag{4.32}$$

(iii) Suppose there exists a neighborhood $V_p \subset \mathcal{M}$ of p such that one of the following assumptions holds:

(a) $X(p) = 0$, and for every $q \in V_p, H_q \in \mathcal{K}(q), v \in T_q\mathcal{M}$, there exists $\bar{H}_q \in \bar{\mathcal{K}}(q)$ such that $(\bar{H}_q v)_\top = H_q v$.

(b) X is locally Lipschitz at p, and for every $q \in V_p, H_q \in \mathcal{K}(q), v \in T_q\mathcal{M}$, there exists $\bar{H}_q \in \bar{\mathcal{K}}(q)$ such that $(\bar{H}_q v)_\top = H_q v$ and $(\bar{H}_q v)_\perp = \mathrm{II}(v, X(q))$.

Fix $\mu \in [0,1)$, if for every $\varepsilon > 0$, there exists $\delta > 0$ such that for every $q \in \mathcal{M}$ with $d_{\bar{\mathcal{M}}}(p,q) < \delta$ and every $\bar{H}_q \in \bar{\mathcal{K}}(q)$, the inequality (4.4) holds, i.e.,

$$\|\bar{X}(p) - \bar{P}_{qp}[\bar{X}(q) + \bar{H}_q \overline{\exp}_q^{-1} p]\| \le \varepsilon d_{\bar{\mathcal{M}}}(p,q)^{1+\mu}. \tag{4.33}$$

Then, there exists $\tilde{\delta} > 0$ such that for every $q \in \mathcal{M}$ with $d_{\mathcal{M}}(p,q) < \tilde{\delta}$ and every $H_q \in \mathcal{K}(q)$, the following inequality holds.

$$\|X(p) - P_{qp}[X(q) + H_q \exp_q^{-1} p]\| \le 2\varepsilon d_{\mathcal{M}}(p,q)^{1+\mu}. \tag{4.34}$$

(iv) Suppose either the assumption (iii.a) or (iii.b) is fulfilled, and $\bar{X}$ is Lipschitz and directionally differentiable in the Hadamard sense in a neighborhood $\bar{U}_p \subset \bar{\mathcal{M}}$ of $p \in \mathcal{M}$. If $\bar{X}$ is semismooth at p with order $\mu \in [0,1)$ with respect to $\bar{\mathcal{K}}$, then X is semismooth at p with order μ with respect to $\mathcal{K}$. Moreover, if $\bar{X}$ is strongly semismooth at p with respect to $\bar{\mathcal{K}}$, then X is also strongly semismooth at p with respect to $\mathcal{K}$.

Before proving the above theorem, we show that the Clarke generalized covariant derivative satisfies the assumption (iii.b), and illustrate how to verify the semismoothness condition in Theorem 4.3 (see Proposition 4.2, Corollary 4.1 and Example 4.1).

First, we briefly review how to represent quantities on manifolds under a local coordinate. Let (U,φ) be a chart near $p \in \mathcal{M}$, and $\{e_i\}$ be the standard basis of $\mathbb{R}^n$, i.e., the j-th component of e_i is δ_{ij}, and let $E_i := (\mathrm{d}\varphi)^{-1}[e_i]$, then E_i is a smooth vector field near p and $\{E_i(q)\}$ is a basis of $T_q\mathcal{M}$ for every $q \in U$. Let $g_{ij} := \langle E_i, E_j\rangle$, $G := (g_{ij})_{i,j\in[n]}$ and Γ_{ij}^k be the Christoffel symbols such that $\nabla_{E_i} E_j = \sum_{k=1}^n \Gamma_{ij}^k E_k$, and g^{ij} be such that $(g^{ij})_{i,j\in[n]} = G^{-1}$, and we call G the metric matrix. The gradient of a smooth function $f : \mathcal{M} \to \mathbb{R}$ is $\operatorname{grad} f = \sum_{i,j=1}^n (g^{ij} E_i f) E_j$ (see [47, p. 27]). Two smooth vector fields X, Y defined near p can be represented as $X = \sum_{i=1}^n X^i E_i$ and $Y = \sum_{i=1}^n Y^i E_i$, and their inner product is $\langle X, Y\rangle = \sum_{i,j=1}^n g_{ij} X^i Y^j$. The covariant derivative $\nabla_X Y$ can be written as (see [47, p. 92])

$$\nabla_X Y = \sum_{i=1}^n (XY^i) E_i + \sum_{i,j,k=1}^n X^i Y^j \Gamma_{ij}^k E_k. \tag{4.35}$$

When (U,φ) is the chart constructed in the proof of Lemma 4.2, φ^{-1} is the normal coordinate at $p \in \mathcal{M}$ (see [47, p. 132]). Proposition 5.24 in [47] shows that $g_{ij}(p) = \delta_{ij}$, $\Gamma_{ij}^k(p) = 0$, and all partial derivatives of g_{ij} vanish at p. Thus, $\operatorname{grad} f(p) = \sum_{i=1}^n (E_i f)(p) E_i(p)$, $\nabla_X Y(p) = \sum_{i=1}^n (XY^i)(p) E_i(p)$ and $\langle X(p), Y(p)\rangle = \sum_{i=1}^n X^i(p) Y^i(p)$.

The following two lemmas are some chain rules on manifolds, which are proved by pulling functions on manifolds back to Euclidean spaces and then applying the Euclidean chain rules.

Lemma 4.7 *Let $\mathcal{M}, \mathcal{N}$ be two Riemannian manifolds, and $F : \mathcal{M} \to \mathcal{N}$ be a smooth map, $g : \mathcal{N} \to \mathbb{R}$ be a map that is locally Lipschitz at $F(p) \in \mathcal{N}$, then the following chain rule holds at $p \in \mathcal{M}$.*

$$\partial(g \circ F)(p) \subseteq \partial g(F(p)) \mathrm{d}F(p), \tag{4.36}$$

where the inclusion means that for every $\xi \in \partial(g \circ F)(p)$, there exists $\zeta \in \partial g(F(p))$ such that $\langle \xi, v\rangle = \langle \zeta, \mathrm{d}F|_p[v]\rangle$ for every $v \in T_p\mathcal{M}$.

Proof Let (U,φ) and (V,ψ) be charts near $p \in \mathcal{M}$ and $F(p) \in \mathcal{N}$ with $\varphi(0) = p$ and $\psi(0) = F(p)$, respectively. We denote $G_{\mathcal{N}}, G_{\mathcal{M}}$ as the metric matrices on $\mathcal{N}, \mathcal{M}$ defined above, respectively. Define $\hat{g} := g \circ \psi^{-1}$ and $\hat{F} := \psi \circ F \circ \varphi^{-1}$. Then, Theorem 2.3.10 in [24] yields that

$$\partial(\hat{g} \circ \hat{F})(0) \subseteq \partial \hat{g}(\hat{F}(0)) \mathrm{d}\hat{F}(0).$$

Fix $\xi \in \partial(g \circ F)(p)$, then from (2.2) we know $\hat{\xi} := G_{\mathcal{M}} \mathrm{d}\varphi|_p[\xi] \in \partial(\hat{g} \circ \hat{F})(0)$. The above display shows that there exists $\hat{\zeta} \in \partial \hat{g}(\hat{F}(0))$ such that for every $\hat{v} \in \mathbb{R}^n$, it holds that $\hat{\xi}^\top \hat{v} = \hat{\zeta}^\top (\mathrm{d}\hat{F}(0)[\hat{v}])$. Let $v :=$

$(\mathrm{d}\varphi|_p)^{-1}[\hat{v}]$, then $\hat{\xi}^\top \hat{v} = (\mathrm{d}\varphi|_p[\xi])^\top G_{\mathcal{M}}(\mathrm{d}\varphi|_p[v]) = \langle \xi, v\rangle$. Since $\zeta := (\mathrm{d}\psi|_{F(p)})^{-1} G_{\mathcal{N}}^{-1}\hat{\zeta} \in \partial g(F(p))$ and $\mathrm{d}\hat{F} = \mathrm{d}\psi \circ \mathrm{d}F \circ (\mathrm{d}\varphi)^{-1}$, we find that $\hat{\zeta}^\top(\mathrm{d}\hat{F}(0)[\hat{v}]) = (\mathrm{d}\psi|_{F(p)}[\zeta])^\top G_{\mathcal{N}}(\mathrm{d}\psi|_{F(p)}[\mathrm{d}F|_p[v]]) = \langle \zeta, \mathrm{d}F|_p[v]\rangle$. Therefore, $\langle \xi, v\rangle = \langle \zeta, \mathrm{d}F|_p[v]\rangle$ for every $v \in T_p\mathcal{M}$, and hence $\xi \in \partial g(F(p))\mathrm{d}F(p)$. □

Lemma 4.8 *Let $\mathcal{M}$ be a Riemannian manifold, X be a vector field on $\mathcal{M}$ that is locally Lipschitz at $p \in \mathcal{M}$, and Y be a smooth vector field on $\mathcal{M}$. Define $f := \langle X, Y\rangle$, then the following chain rule holds.*

$$\partial f(p) = \langle \partial X(p), Y(p)\rangle + \langle X(p), \nabla Y(p)\rangle\,, \tag{4.37}$$

where $\xi \in \partial f(p)$ is regarded as the operator $\zeta \mapsto \langle \xi, \zeta\rangle$ for $\zeta \in T_p\mathcal{M}$, and the right-hand side is understood as the set of operators $\zeta \mapsto \langle H\zeta, Y(p)\rangle + \langle X(p), \nabla_\zeta Y(p)\rangle$ for $\zeta \in T_p\mathcal{M}$ and $H \in \partial X(p)$.

Proof Let (U_p, φ) be a chart such that φ^{-1} is the normal coordinate. Suppose that $X = \sum_{i=1}^n X^i E_i$ and $Y = \sum_{i=1}^n Y^i E_i$, we define $\hat{f} := f \circ \varphi^{-1}$, $\hat{X}^j := X^j \circ \varphi^{-1}$, $\hat{X} := (\hat{X}^1, \ldots, \hat{X}^n)$, and $\hat{Y}$ likewise. Thus, $\hat{f}(x) = \hat{X}(x)^\top G(x)\hat{Y}(x)$ for $x \in \hat{U}_p := \varphi(U_p)$, and can be regarded as the composition $\mathcal{A} \circ \tilde{X}$, where $\mathcal{A}(x, v) := v^\top G(x)\hat{Y}(x)$ and $\tilde{X}(x) := (x, \hat{X}(x))$ for $x \in U_p, v \in \mathbb{R}^n$. Since $\mathcal{A}$ is smooth and $G(0) = I_n, \nabla G(0) = 0$, we know $\nabla \mathcal{A}(0, v)[w, \xi] = v^\top \nabla \hat{Y}(0)[w] + \xi^\top \hat{Y}(0)$ for $w, \xi \in \mathbb{R}^n$. Note that $\partial \tilde{X} = (\mathrm{Id}, \partial \hat{X})$, then the chain rule in Euclidean spaces [24, Theorem 2.6.6] implies that

$$\partial \hat{f}(0) = \nabla \mathcal{A}(0, \hat{X}(0))[\mathrm{Id}, \partial \hat{X}(0)] = \hat{X}(0)^\top \nabla \hat{Y}(0) + \hat{Y}(0)^\top \partial \hat{X}(0). \tag{4.38}$$

We define the linear bijection $J : \mathcal{L}(\mathbb{R}^n) \to \mathcal{L}(T_p\mathcal{M})$ as

$$(J\hat{H})v := \sum_{i=1}^n [\hat{H}\hat{v}]_i E_i(p),$$

where $\hat{H} \in \mathcal{L}(\mathbb{R}^n)$, $v \in T_p\mathcal{M}$ and $\hat{v} := \mathrm{d}\varphi|_p[v]$. Below we show that $J(\partial \hat{X}(0)) = \partial X(p)$.

Suppose $\hat{H} \in \partial_B \hat{X}(0)$, then there exists $x_k \to 0$ such that $\hat{X}$ is differentiable at x_k and $\nabla \hat{X}(x_k) \to \hat{H}$ as $k \to \infty$. Let $H := J\hat{H}$, $\hat{v} \in \mathbb{R}^n$, $v := (\mathrm{d}\varphi|_p)^{-1}[\hat{v}]$, $p_k := \varphi^{-1}(x_k)$ and $v_k := (\mathrm{d}\varphi|_{p_k})^{-1}[\hat{v}]$, then $\nabla \hat{X}^i(x_k)[\hat{v}] = \mathrm{d}\hat{X}^i|_{x_k}[\hat{v}] = \mathrm{d}X^i|_{p_k}[v_k] = v_k X^i(p_k)$, and therefore

$$\nabla_{v_k} X(p_k) = \sum_{i=1}^n \nabla \hat{X}^i(x_k)[\hat{v}]E_i + \sum_{i,j,k=1}^n X^i Y^j \Gamma_{ij}^k E_k.$$

Let $k \to \infty$ and note $\Gamma_{ij}^k(p) = 0$, we find $\nabla_{v_k} X(p_k) \to \sum_{i=1}^n [\hat{H}\hat{v}]_i E_i(p) = Hv$. Since $\hat{v}$ is arbitrary, we conclude that $\nabla X(p_k) \to H \in \partial_B X(p)$. By reversing this argument, we can show that $J^{-1}H \in \partial \hat{X}(0)$ for every $H \in \partial X(p)$. Then, $J(\partial_B \hat{X}(0)) = \partial_B X(p)$, and by the linearity of J, we know $J(\partial \hat{X}(0)) = \partial X(p)$.

Thus, $\hat{Y}(0)^\top(\hat{H}\hat{v}) = \sum_{i=1}^n Y^i(p)\langle J\hat{H}v, E_i(p)\rangle = \langle J\hat{H}v, Y(p)\rangle$ for every $\hat{H} \in \partial \hat{X}(0)$, $v \in T_p\mathcal{M}$ and $\hat{v} := \mathrm{d}\varphi|_p[v]$. Similarly, it holds that $\hat{X}(0)^\top \nabla \hat{Y}(0)[\hat{v}] = \langle X(p), \nabla Y(p)[v]\rangle$. Besides, (2.2) yields $\partial \hat{f}(0)[\hat{v}] = \partial f(p)[v]$. Therefore, the proof is completed by plugging these equations into (4.38). □

The following proposition relates the Clarke generalized covariant derivatives on $\mathcal{M}$ and its ambient space $\bar{\mathcal{M}}$. As a special case when $\mathcal{M} := \mathbb{R}^n \subset \mathbb{R}^{n+k} =: \bar{\mathcal{M}}$, the second fundamental form $\mathrm{II}(u, v) \equiv 0$ and (4.39) can be obtained from the corollary in [24, p. 75].

Proposition 4.2 *Let $p \in \mathcal{M}$ and $v \in T_p\mathcal{M}$, if $\bar{X}$ is locally Lipschitz at p, then*

$$\partial X(p)[v] \subseteq \bar{\partial}\bar{X}(p)[v] - \mathrm{II}(v, X(p)), \tag{4.39}$$

where $\partial X(p)[v] := \{Hv : H \in \partial X(p)\}$ and $\bar{\partial}\bar{X}(p)[v] = \{\bar{H}v : \bar{H} \in \partial \bar{X}(p)\}$.

Proof Since Theorem 4.4 shows that X is also locally Lipschitz at p, then $\partial X(p)$ is well-defined. Let $U_p \subset \mathcal{M}$ be a neighborhood of p, $\xi \in T_p\mathcal{M}$ and V be a smooth vector field on $\bar{\mathcal{M}}$ such that $V_\top(q) = P_{pq}\xi$ for $q \in U_p$, and $\iota : \mathcal{M} \hookrightarrow \bar{\mathcal{M}}$ be the embedding map. Define $\bar{f}_V := \langle \bar{X}, V\rangle$ and $f_V := \bar{f}_V \circ \iota$, then Lemma 4.7 shows that $\partial f_V(p) \subseteq \bar{\partial}\bar{f}_V(p)\mathrm{d}\iota(p) = \bar{\partial}\bar{f}_V(p)|_{T_p\mathcal{M}}$.

Let $v \in T_p\mathcal{M}$, then Lemma 4.8 gives that $\bar{\partial}\bar{f}_V(p)[v] = \langle \bar{\partial}\bar{X}(p)[v], V(p)\rangle + \langle \bar{X}(p), \bar{\nabla}_v V(p)\rangle$. Since $v, \bar{X}(p) \in T_p\mathcal{M}$, Proposition 8.4 in [47] shows that

$$\langle \bar{X}(p), \bar{\nabla}_v V_\perp(p)\rangle = \langle \bar{X}(p), (\bar{\nabla}_v V_\perp(p))_\top\rangle = \langle X(p), \nabla_v V_\perp(p)\rangle = -\langle \mathrm{II}(v, X(p)), V_\perp(p)\rangle\,.$$

Note that $\nabla_v V_\top(p) = \nabla_v(P_{pq}\xi)(p) = 0$, then $\langle \bar{X}(p), \bar{\nabla}_v V_\top(p) \rangle = \langle X(p), \nabla_v V_\top(p) \rangle = 0$, and thus,

$$\langle \bar{X}(p), \bar{\nabla}_v V(p) \rangle = \langle \bar{X}(p), \bar{\nabla}_v V_\perp(p) + \bar{\nabla}_v V_\top(p) \rangle = -\langle \mathrm{II}(v, X(p)), V_\perp(p) \rangle = -\langle \mathrm{II}(v, X(p)), V(p) \rangle,$$

where the last equation follows from $\langle \mathrm{II}(v, X(p)), V_\top(p) \rangle = 0$.

Combining these arguments, we know $\partial f_V(p)[v] \subseteq \langle \bar{\partial}\bar{X}(p)[v] - \mathrm{II}(v, X(p)), V(p) \rangle$. Note that $\bar{X}(q) \in T_q\mathcal{M}$ for $q \in U_p$, then $f_V(q) = \langle X(q), V_\top(q) \rangle$ for $q \in U_p$. Since $\nabla_v V_\top(p) = 0$ and $\partial X(p)[v] \subset T_p\mathcal{M}$, then Lemma 4.8 yields $\partial f_V(p)[v] = \langle \partial X(p)[v], V_\top(p) \rangle = \langle \partial X(p)[v], V(p) \rangle$. Therefore, we know

$$\langle \partial X(p)[v], V(p) \rangle \subseteq \langle \bar{\partial}\bar{X}(p)[v] - \mathrm{II}(v, X(p)), V(p) \rangle.$$

Note that $V(p)_\top = \xi$ and $V(p)_\perp$ are arbitrary and $\partial X(p)[v], \bar{\partial}\bar{X}(p)[v]$ are compact convex sets, the inclusion relationship (4.39) follows from [56, Corollary 13.1.1]. ☐

Corollary 4.1 *Suppose that $\bar{X}$ is Lipschitz and directionally differentiable in the Hadamard sense in a neighborhood $\bar{U}_p \subset \bar{\mathcal{M}}$ of $p \in \mathcal{M}$. If $\bar{X}$ is semismooth at p with order $\mu \in [0, 1)$ with respect to $\bar{\partial}\bar{X}(p)$, then X is also semismooth at p with order $\mu \in [0, 1)$ with respect to $\partial X(p)$. Moreover, if $\bar{X}$ is strongly semismooth at p with respect to $\bar{\partial}\bar{X}(p)$, then X is also strongly semismooth at p with respect to $\partial X(p)$.*

Proof Proposition 4.2 shows that the assumption (iii.b) in Theorem 4.4 is satisfied, then the conclusion follows from Theorem 4.4. ☐

Remark 4.6 When $X = X_1 + X_2 + X_3$ as in Assumption 4.1, where X_2, X_3 are locally Lipschitz and X_1 is smooth. Suppose X_2, X_3 can be extended to $\bar{X}_2, \bar{X}_3$ in a neighborhood of $\mathcal{M}$ in $\bar{\mathcal{M}}$. If we could apply Corollary 4.1 to show that X_2, X_3 are semismooth with respect to the Clarke generalized covariant derivatives, then Proposition 4.1 would imply that $X_2 + X_3$ is semismooth with respect to $\partial X_2 + \partial X_3$.

Example 4.1 In the CM problem (1.2), the function L_k used in Algorithm 4.1 (see (5.8) for its gradient) can be naturally extended to the Euclidean space $\mathbb{R}^{n\times r}$. We denote this extension by $\bar{L}_k$. It is the sum of the smooth function $\mathrm{tr}(X^\top HX)$ and the Moreau-Yosida regularization of the $\|\cdot\|_1$-norm whose Euclidean gradient is a piecewise linear map. It is well-known that a piecewise linear map (in Euclidean spaces) is strongly semismooth (see, e.g., [31, Proposition 7.4.7]), and hence $\nabla\bar{L}_k$ is also strongly semismooth (with respect to the Clarke generalized derivative). Since $X(Q) := \operatorname{grad} L_k(Q) = \mathbf{P}_Q\bar{\nabla}\bar{L}_k$ for $Q \in \mathrm{St}(n, r)$, and $\mathbf{P}_Q(V) = V - Q(Q^\top V + V^\top Q)/2$ (see, e.g., [38]), then X can also be extended to a vector-valued map $\bar{X}$ in $\mathbb{R}^{n\times r}$. Note that $\bar{X}$ is the composition of a smooth map and a strongly semismooth map, so $\bar{X}$ is also strongly semismooth. By Corollary 4.1, we know that $\operatorname{grad} L_k = X$ is strongly semismooth with respect to $\partial \operatorname{grad} L_k(p_*)$. Indeed, the SPCA problem (1.3) and the CM problem have the same non-smooth term, so $\operatorname{grad} L_k$ for (1.3) is also strongly semismooth. The constrained SPCA problem (1.4) has an additional non-smooth term, i.e., the indicator function of a convex set. The Euclidean gradient of the Moreau-Yosida regularization of this term is also strongly semismooth, and therefore $\operatorname{grad} L_k$ for (1.4) is strongly semismooth.

We prove Theorem 4.4 in the remaining part of this section. First, we give a simple result that will be frequently used in our proof: For $p \in \bar{\mathcal{M}}$, $\xi \in T_p\bar{\mathcal{M}}$, and a curve $\gamma : [0, 1] \to \bar{\mathcal{M}}$ with $\gamma(0) = p$, the vector field $Z(t) := \bar{P}_\gamma^{0\to t}\xi$ is parallel to γ, and hence by the definition of parallel transport we have

$$\bar{\nabla}_{\dot{\gamma}}\bar{P}_\gamma^{0\to t}\xi = 0. \tag{4.40}$$

Similarly, for $p \in \mathcal{M}, \zeta \in T_p\mathcal{M}$ and $\gamma : [0, 1] \to \mathcal{M}$, it holds that $\nabla_{\dot{\gamma}}P_\gamma^{0\to t}\zeta = 0$.

The following lemma compares the parallel transports of the manifold $\mathcal{M}$ and its ambient space $\bar{\mathcal{M}}$.

Lemma 4.9 *There exists $C > 0$ such that for every $p \in \mathcal{M}$, and every smooth curve $\gamma : [-1, 1] \to \mathcal{M}$ with $\gamma(0) = p$, and every $v \in T_p\mathcal{M}$, $t \in [0, 1]$, it holds that*

$$\|P_\gamma^{0\to t}v - \bar{P}_\gamma^{0\to t}v\| \le C\ell(\gamma|_{[0,t]})\|v\|. \tag{4.41}$$

Moreover, there exists $C' > 0$ such that for every $p \in \mathcal{M}$, $w \in T_p\mathcal{M}$ with $\|w\| \le 2$, and every $t \in [0, 1]$, the following inequality holds for $\gamma(t) := \exp_p(tw)$.

$$\|P_\gamma^{0\to t}v - \bar{P}_\gamma^{0\to t}v - t\mathrm{II}(\dot{\gamma}(t), P_\gamma^{0\to t}v)\| \le C't^2. \tag{4.42}$$

Proof Let $Y : [0,1] \to T\mathcal{M}$ be a smooth vector field along γ. For any fixed $t \in [0,1]$ and $\xi \in T_{\gamma(t)}\bar{\mathcal{M}}$, it holds that

$$\begin{aligned}\left\langle \frac{\mathrm{d}}{\mathrm{d}s}\bar{P}_\gamma^{s\to t}Y(s),\xi\right\rangle &= \frac{\mathrm{d}}{\mathrm{d}s}\left\langle \bar{P}_\gamma^{s\to t}Y(s),\xi\right\rangle = \frac{\mathrm{d}}{\mathrm{d}s}\left\langle Y(s),\bar{P}_\gamma^{t\to s}\xi\right\rangle \\ &= \left\langle \bar{\nabla}_{\dot\gamma}Y(s),\bar{P}_\gamma^{t\to s}\xi\right\rangle + \left\langle Y(s),\bar{\nabla}_{\dot\gamma}\bar{P}_\gamma^{t\to s}\xi\right\rangle \overset{(4.40)}{=} \left\langle \bar{P}_\gamma^{s\to t}\bar{\nabla}_{\dot\gamma}Y(s),\xi\right\rangle,\end{aligned}$$

Therefore, $\frac{\mathrm{d}}{\mathrm{d}s}\bar{P}_\gamma^{s\to t}Y(s) = \bar{P}_\gamma^{s\to t}\bar{\nabla}_{\dot\gamma}Y(s)$.

Setting $Y(t) = P_\gamma^{0\to t}v$ for $v \in T_p\mathcal{M}$, we know

$$\bar{\nabla}_{\dot\gamma}Y \overset{(4.29)}{=} \nabla_{\dot\gamma}Y + \mathrm{II}(\dot\gamma, Y) \overset{(4.40)}{=} \mathrm{II}(\dot\gamma, Y),$$

and thus,

$$Y(t) - \bar{P}_\gamma^{0\to t}Y(0) = \int_0^t \frac{\mathrm{d}}{\mathrm{d}s}\bar{P}_\gamma^{s\to t}Y(s)\mathrm{d}s = \int_0^t \bar{P}_\gamma^{s\to t}\mathrm{II}(\dot\gamma(s),Y(s))\mathrm{d}s. \tag{4.43}$$

Since II smoothly depends on the Riemannian metric of $\bar{\mathcal{M}}$ and $\mathcal{M}$ is compact, the constant $C := \sup\{\|\mathrm{II}(u,v)\| : p \in \mathcal{M}, u,v \in T_p\mathcal{M}, \|u\| = \|v\| = 1\}$ is finite. As the parallel transport is an isometry, it holds that $\|Y(t)\| = \|Y(0)\| = \|v\|$ and

$$\|Y(t) - \bar{P}_\gamma^{0\to t}Y(0)\| \le \int_0^t C\|v\|\|\dot\gamma(s)\|\mathrm{d}s = C\|v\|\ell(\gamma|_{[0,t]}).$$

Moreover, we have

$$\begin{aligned}t\mathrm{II}(\dot\gamma(t),Y(t)) - [Y(t) - \bar{P}_\gamma^{0\to t}Y(0)] &\overset{(4.43)}{=} \int_0^t \mathrm{II}(\dot\gamma(t),Y(t)) - \bar{P}_\gamma^{s\to t}\mathrm{II}(\dot\gamma(s),Y(s))\mathrm{d}s \\ = \int_0^t\int_s^t \frac{\mathrm{d}}{\mathrm{d}u}\bar{P}_\gamma^{u\to t}\mathrm{II}(\dot\gamma(u),Y(u))\mathrm{d}u\mathrm{d}s &= \int_0^t\int_s^t \bar{P}_\gamma^{u\to t}\bar{\nabla}_{\dot\gamma}\mathrm{II}(\dot\gamma(u),Y(u))\mathrm{d}u\mathrm{d}s.\end{aligned}$$

When $\gamma(t) := \exp_p(tw)$ for $w \in T_p\mathcal{M}$ with $\|w\| \le 2$, the above integrand is continuous with respect to v,w,t,u,p, and thus, the constant $C' := 2\sup\{\|\bar{\nabla}_{\dot\gamma}\mathrm{II}(\dot\gamma, P_\gamma^{0\to t}v)\| : p \in \mathcal{M}$ and $w,v \in T_p\mathcal{M}, \|w\| \le 2, \|v\| \le 2, t \in [0,1]\}$ is finite. Therefore, (4.42) holds. ☐

The following lemma controls the error of parallelly transporting a vector along the shortest geodesic on $\mathcal{M}$ and then moving back along the shortest geodesic on $\bar{\mathcal{M}}$.

Lemma 4.10 *Fix $p \in \mathcal{M}$, then there exists $C > 0$ such that for every $v \in T_p\mathcal{M}$ with $\|v\| \le 2$, $t \in [0,1]$ and $w,\xi \in T_p\bar{\mathcal{M}}$, the following inequality holds*

$$|\left\langle \bar{P}_{\gamma(t),p}\bar{P}_\gamma^{0\to t}w,\xi\right\rangle - \langle w,\xi\rangle| \le Ct^2\|w\|\|\xi\|, \tag{4.44}$$

where $\gamma(t) := \exp_p(tv)$. Consequently,

$$\|\bar{P}_{p,\gamma(t)} - \bar{P}_\gamma^{0\to t}\| \le Ct^2. \tag{4.45}$$

Proof Let $w,\xi \in T_p\bar{\mathcal{M}}$ and $v \in T_p\mathcal{M}$ with $\|v\| \le 2$ and $\gamma(t) := \exp_p(tv)$, then

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \bar{P}_{\gamma(t),p}\bar{P}_\gamma^{0\to t}w,\xi\right\rangle &= \frac{\mathrm{d}}{\mathrm{d}t}\left\langle \bar{P}_\gamma^{0\to t}w,\bar{P}_{p,\gamma(t)}\xi\right\rangle \\ &= \left\langle \bar{\nabla}_{\dot\gamma}\bar{P}_\gamma^{0\to t}w,\bar{P}_{p,\gamma(t)}\xi\right\rangle + \left\langle \bar{P}_\gamma^{0\to t}w,\bar{\nabla}_{\dot\gamma}\bar{P}_{p,\gamma(t)}\xi\right\rangle \\ &\overset{(4.40)}{=} \left\langle \bar{P}_\gamma^{0\to t}w,\bar{\nabla}_{\dot\gamma}\bar{P}_{p,\gamma(t)}\xi\right\rangle, \end{aligned}\tag{4.46}$$

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left\langle \bar{P}_{\gamma(t),p}\bar{P}_\gamma^{0\to t}w,\xi\right\rangle &= \left\langle \bar{\nabla}_{\dot\gamma}\bar{P}_\gamma^{0\to t}w,\bar{\nabla}_{\dot\gamma}\bar{P}_{p,\gamma(t)}\xi\right\rangle + \left\langle \bar{P}_\gamma^{0\to t}w,\bar{\nabla}_{\dot\gamma}\bar{\nabla}_{\dot\gamma}\bar{P}_{p,\gamma(t)}\xi\right\rangle \\ &\overset{(4.40)}{=} \left\langle \bar{P}_\gamma^{0\to t}w,\bar{\nabla}_{\dot\gamma}\bar{\nabla}_{\dot\gamma}\bar{P}_{p,\gamma(t)}\xi\right\rangle. \end{aligned}\tag{4.47}$$

Let $Y(q) := \bar{P}_{pq}\xi$ for $q \in \bar{\mathcal{M}}$, then $Y(\gamma(t)) = \bar{P}_{p,\gamma(t)}\xi$. From [18, Proposition 2.2(c)], we know $\bar{\nabla}_{\dot\gamma}Y = \bar{\nabla}_{\dot\gamma}\bar{P}_{p,\gamma(t)}\xi$. By the definition of Y, $\bar{\nabla}_vY(p) = 0$ for every $v \in T_p\bar{\mathcal{M}}$. Therefore, the right-hand side of

(4.46) vanishes at $t = 0$. Since $\bar\nabla_{\dot\gamma}\bar\nabla_{\dot\gamma}\bar P_{p,\gamma(t)}\xi$ is continuous with respect to t, v and is linear with respect to ξ, then

$$C := \sup\left\{\|\bar\nabla_{\dot\gamma}\bar\nabla_{\dot\gamma}\bar P_{p,\gamma(t)}\xi\| : t \in [0,1], \|v\| \le 2, \|\xi\| \le 1\right\} < \infty.$$

Applying Taylor's theorem at $t = 0$, we know for every $t \in [0,1]$ and $w, \xi \in T_p\bar{\mathcal{M}}$,

$$|\left\langle \bar P_{\gamma(t),p}\bar P_\gamma^{0\to t}w, \xi\right\rangle - \langle w, \xi\rangle| \le Ct^2\|w\|\|\xi\|.$$

Since C is independent of v, then the above holds for every $\|v\| \le 2$. Finally, (4.45) follows from

$$\|\bar P_{p,\gamma(t)} - \bar P_\gamma^{0\to t}\| = \|\bar P_{\gamma(t),p}\bar P_\gamma^{0\to t} - \mathrm{Id}\| \le Ct^2.$$

This completes the proof. □

The lemma below bounds the error of the inverse exponential maps of $\mathcal{M}$ and $\bar{\mathcal{M}}$, which is proved by extending $\exp_p$ to a retraction on $\bar{\mathcal{M}}$ using the Fermi coordinate [47, p. 135] and applying [70, Lemma 3] to bound the difference of two inverse retractions.

Lemma 4.11 *Fix $p_0 \in \mathcal{M}$, there exist $C, r > 0$ such that the following inequality holds for every $p, q \in B_r(p_0)$.*

$$\|\exp_p^{-1} q - \overline{\exp}_p^{-1} q\| \le C d_{\bar{\mathcal{M}}}(p,q)^2. \tag{4.48}$$

Proof From Lemma 4.1, there exists $r > 0$ such that $\exp_p$ is a diffeomorphism from $\{v \in T_p\mathcal{M} : \|v\| < 4r\}$ to $B_{4r}(p)$ for every $p \in B_{4r}(p_0)$. According to Theorem 5.25 in [47], there exists $r' > 0$ such that $\overline{\exp}$ is a diffeomorphism from $V := \{(p,v) \in N\mathcal{M} : p \in B_{2r}(p_0), \|v\| < r'\}$ to its image under V, where $N\mathcal{M} := \bigcup_{p\in\mathcal{M}}(T_p\mathcal{M})^\perp$ is the normal bundle. Let $E_1, \ldots, E_d$ be vector fields on $B_{4r}(p_0)$ such that $\{E_1(p), \ldots, E_n(p)\}$ and $\{E_{n+1}(p), \ldots, E_d(p)\}$ are orthonormal bases of $T_p\mathcal{M}$ and $(T_p\mathcal{M})^\perp$, respectively. For $p \in B_r(p_0)$, $v := \sum_{i=1}^d w_iE_i(p) \in T_p\bar{\mathcal{M}}$ with $\|v_\top\| < r$ and $\|v_\perp\| < r'$, we define $q(p,v) := \exp_p(v_\top)$ and

$$E(p,v) := \overline{\exp}_{q(p,v)}\left(\sum_{i=n+1}^d w_iE_i(q(p,v))\right).$$

The above discussion shows that E is well-defined, and $E(p,v) = \exp_p v$ if $v \in T_p\mathcal{M}$ and $\|v\| < r$, and $E(p,v) = \overline{\exp}_p v$ if $v \in (T_p\mathcal{M})^\perp$ and $\|v\| < r'$. Thus, E can be regarded as a retraction on $\bar{\mathcal{M}}$ defined near p_0, and (4.48) follows from [70, Lemma 3]. □

Proof (Proof of Theorem 4.4)

(i) First, we verify the local Lipschitz property of X. Since $\bar X$ is locally Lipschitz at $p \in \mathcal{M}$, then there exist $L_p > 0$ and a neighborhood $\bar U_p \subset \bar{\mathcal{M}}$ at p such that $\bar X$ is L_p-Lipschitz in $\bar U_p$. By Lemma 4.9, we know for every $p, q \in \mathcal{M} \cap \bar U_p$ and every geodesic γ joining p and q, it holds that

$$\begin{aligned}\|P_\gamma^{0\to1}X(p) - X(q)\| &\le \|P_\gamma^{0\to1}X(p) - \bar P_\gamma^{0\to1}X(p)\| + \|\bar P_\gamma^{0\to1}X(p) - X(q)\| \\ &\le \left(C\sup_{p\in\mathcal{M}}\|X(p)\| + L_p\right)\ell(\gamma),\end{aligned}$$

where C is the constant in Lemma 4.9. Since $\mathcal{M}$ is compact and X is continuous, $M := \sup_{p\in\mathcal{M}}\|X(p)\| < \infty$, then X is Lipschitz on $\mathcal{M} \cap \bar U_p$, i.e., X is locally Lipschitz at p.

(ii) Next, we show that X is directionally differentiable at p in the Hadamard sense. Fix $v \in T_p\mathcal{M}$ with $\|v\| = 1$ and let $\gamma(t) := \exp_p(tv')$ for $t \in (-1,1)$ and $v' \in T_p\mathcal{M}$ with $\|v'\| \le 2$, and decompose $P_\gamma^{t\to0}X(\exp_p(tv')) - X(p)$ into

$$\underbrace{(P_\gamma^{t\to0} - \bar P_\gamma^{t\to0})X(\gamma(t)) - tS(t)}_{\text{(A)}} + \underbrace{(\bar P_\gamma^{t\to0} - \bar P_{\gamma(t),p})X(\gamma(t))}_{\text{(B)}} + \underbrace{\bar P_{\gamma(t),p}X(\gamma(t)) - X(p) + tS(t)}_{\text{(C)}}, \tag{4.49}$$

where $S(t) := \mathrm{II}(-\dot\gamma(0), P_\gamma^{t\to0}X(\gamma(t)))$.

From Lemma 4.9 and 4.10, there exists $C > 0$ (independent of v') such that

$$|\text{(A)}| \overset{(4.42)}{\le} Ct^2 \quad \text{and} \quad |\text{(B)}| \le \|\bar P_\gamma^{t\to0} - \bar P_{\gamma(t),p}\|\|X(\gamma(t))\| \overset{(4.45)}{\le} CMt^2.$$

When $t \to 0^+$ and $v' \to v$, we know $|(\mathrm{A})| + |(\mathrm{B})| = O(t^2)$ and $v'_t := t^{-1}\overline{\exp}_p^{-1}\gamma(t) \to v$, and hence the continuity of $S(t)$ and the Hadamard differentiability (4.31) imply that

$$\begin{aligned}
&\lim_{\substack{t\downarrow 0\\ v'\to v}} \frac{1}{t}\left[P_{\exp_p(tv'),p}X(\exp_p(tv')) - X(p)\right] = \lim_{\substack{t\downarrow 0\\ v'\to v}} \frac{1}{t}(\mathrm{C})\\
&= \lim_{t\downarrow 0} \frac{1}{t}\left[P_{\overline{\exp}_p(tv'_t),p}X(\overline{\exp}_p(tv'_t)) - X(p)\right] + \lim_{\substack{t\downarrow 0\\ v'\to v}} S(t) \overset{(4.31)}{=} \bar\nabla X(p;v) - \mathrm{II}(v, X(p)).
\end{aligned}$$

Note that $\mathrm{II}(v, X(p)) \in (T_p\mathcal{M})^\perp$, we find $\nabla X(p;v) = \bar\nabla X(p;v) - \mathrm{II}(v,X(p)) = (\bar\nabla X(p;v))_\top$ for every $\|v\| = 1$. Since $\nabla X(p;tv) = t\nabla X(p;v)$ for all $t > 0$, we know X is directionally differentiable at p in the Hadamard sense.

(iii.b) Finally, we need to verify the inequality (4.4). Suppose that the assumption (iii.b) holds. Since the parallel transport is an isometry, then from (4.33), for any $\varepsilon > 0$, there exists $\delta \in (0,1)$ such that for every $q \in \mathcal{M}$ with $d_{\bar{\mathcal{M}}}(p,q) < \delta$ and $\bar H_q \in \bar{\mathcal{K}}(q)$, it holds that

$$\|\bar P_{pq}X(p) - X(q) - \bar H_q\overline{\exp}_q^{-1}p\| \le \varepsilon d_{\bar{\mathcal{M}}}(p,q)^{1+\mu} \le \varepsilon d_{\mathcal{M}}(p,q)^{1+\mu}. \tag{4.50}$$

Without the loss of generality, we could assume that $B_\delta(p) \subset V_p$, where V_p is the neighborhood of p in the assumption (iii.b). For every $q \in \mathcal{M}$ and $H_q \in \mathcal{K}(q)$ such that $d_{\mathcal{M}}(q,p) < \delta$, the assumption (iii.b) states that there exists $\bar H_q \in \bar{\mathcal{K}}(q)$ satisfying $(\bar H_q \exp_q^{-1} p)_\perp = \mathrm{II}(\exp_q^{-1} p, X(q))$ and $(\bar H_q \exp_q^{-1} p)_\top = H_q \exp_q^{-1} p$. Let $\alpha = d_{\mathcal{M}}(p,q)$, $\xi := \alpha^{-1}\exp_p^{-1} q$ and $\gamma(t) := \exp_p(t\xi)$, it holds that

$$\begin{aligned}
\|P_{pq}X(p) - X(q) - H_q\exp_q^{-1}p\| &\le \underbrace{\|P_{pq}X(p) - \bar P_\gamma^{0\to\alpha}X(p) + \mathrm{II}(\exp_q^{-1}p, X(q))\|}_{(\mathrm{A})}\\
&+ \underbrace{\|(\bar P_\gamma^{0\to\alpha} - \bar P_{pq})X(p)\|}_{(\mathrm{B})} + \underbrace{\|\bar P_{pq}X(p) - X(q) - \bar H_q\overline{\exp}_q^{-1}p\|}_{(\mathrm{C})} + \underbrace{\|\bar H_q(\overline{\exp}_q^{-1}p - \exp_q^{-1}p)\|}_{(\mathrm{D})}.
\end{aligned}$$

Let $r, \tilde C > 0$ be the constants such that Lemma 4.9, 4.10 and 4.11 hold with $p_0 = p$, and $\tilde L_p > 0$ be the Lipschitz constant of X at p, and by the compactness of $\mathcal{M}$ and the continuity of $\mathrm{II}(v,w)$, we further assume that $\sup\{\|\mathrm{II}(v,w)\| : p \in \mathcal{M}, v, w \in T_p\mathcal{M}, \|v\| = \|w\| = 1\} < \tilde C < \infty$. Note that $\dot\gamma(\alpha) = -\alpha^{-1}\exp_q^{-1}p$ and $\mathrm{II}(\exp_q^{-1}p, X(q)) = -\alpha\mathrm{II}(\dot\gamma(\alpha), X(q))$, then

$$\begin{aligned}
|(\mathrm{A})| &\le \|P_{pq}X(p) - \bar P_\gamma^{0\to\alpha}X(p) - \alpha\mathrm{II}(\dot\gamma(\alpha), P_\gamma^{0\to\alpha}X(p))\| + \alpha\|\mathrm{II}(\dot\gamma(\alpha), P_\gamma^{0\to\alpha}X(p) - X(q))\|\\
&\overset{(4.42)}{\le} \tilde C\alpha^2 + \tilde C\alpha\|\dot\gamma(\alpha)\|\|P_\gamma^{0\to\alpha}X(p) - X(q)\| \le \tilde C(1 + \tilde L_p)\alpha^2,
\end{aligned}$$

where the last inequality follows from the Lipschitzness of X at p. From (4.45) we find that $|(\mathrm{B})| \le \tilde C M\alpha^2$. Since $\mathcal{K}$ is locally bounded, there exists $\tilde M > 0$ such that all elements in $\bigcup_{d(p,q)\le\delta}\bar{\mathcal{K}}(q)$ are bounded by $\tilde M$, and thus (4.48) yields $|(\mathrm{D})| \le \tilde C\tilde M\alpha^2$. Combining with (4.50) and letting $\hat C := \tilde C(M + \tilde M + \tilde L_p + 1)$, it holds that

$$\|P_{pq}X(p) - X(q) - H_q\exp_q^{-1}p\| \le \varepsilon d_{\mathcal{M}}(p,q)^{1+\mu} + \hat C d_{\mathcal{M}}(p,q)^2.$$

Thus, the inequality (4.34) holds for $q \in \mathcal{M}$ with $d_{\mathcal{M}}(p,q) < \tilde\delta := \min\{r, \delta, (\varepsilon/\hat C)^{\frac{1}{1-\mu}}\}$.

(iii.a) When $X(p) = 0$, we know

$$\begin{aligned}
\|X(q) + H_q\exp_q^{-1}p\| &= \|(\bar X(q) + \bar H_q\exp_q^{-1}p)_\top\| \le \|\bar X(q) + \bar H_q\exp_q^{-1}p\|\\
&\le \|\bar X(q) + \bar H_q\overline{\exp}_q^{-1}p\| + \|\bar H_q(\overline{\exp}_q^{-1}p - \exp_q^{-1}p)\|.
\end{aligned}$$

The above two terms can be controlled by (4.33) and (4.48), respectively, and therefore the conclusion follows from a similar discussion as in (iii.b).

(iv) The conclusion for the semismoothness follows from (i), (ii) and (iii), and the conclusion for the strong semismoothness can be obtained by a similar discussion. $\square$

In this section, we discuss the approach of calculating the Clarke generalized covariant derivative of a locally Lipschitz vector field X, which is difficult in general manifolds. Here, we follow Assumption 4.3 and 4.4, and further assume that $\bar{\mathcal{M}}$ is the Euclidean space $\mathbb{R}^d$.

If X is differentiable at $p \in \mathcal{M}$, it is known from (4.30) (see also [38]) that for every $v \in T_p\mathcal{M}$,

$$\nabla X(p)[v] = \mathbf{P}_p(\bar{\nabla}\bar{X}(p)[v]), \tag{4.51}$$

where $\mathbf{P}_p$ is the projection onto $T_p\mathcal{M}$. If X is not differentiable at $p \in \mathcal{M}$, Proposition 4.2 shows that $\partial X(p)[v] \subseteq \mathbf{P}_p(\bar{\partial}\bar{X}(p)[v])$. As illustrated by the following example, the inclusion relationship can be strict, and thus we need further discussion for finding an element in $\partial X(p)$.

Example 4.2 Consider the manifold $\mathbb{S}^1 = \{(x, y) \in \mathbb{R}^2 : x^2+y^2 = 1\} \subset \mathbb{R}^2$ and the vector-valued function $Y : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $Y(x, y) = \big(2|x - 1/\sqrt{2}|,\ 4|y - 1/\sqrt{2}|\big)^\top$. Let $p_0 = (1/\sqrt{2}, 1/\sqrt{2})^\top \in \mathbb{S}^1$. Note that the projection onto the tangent space $T_p\mathbb{S}^1$ is $I - pp^\top$, we can define the vector field $X : \mathbb{S}^1 \to T\mathbb{S}^1$ by $X(p) = (I - pp^\top)Y(p) \in T_p\mathbb{S}^1$ for $p \in \mathbb{S}^1$. In this case, $\bar{X} : \mathbb{R}^2 \to \mathbb{R}^2$ is also $\bar{X}(p) = (I - pp^\top)Y(p)$ for $p \in \mathbb{R}^2$. If $\bar{X}$ is differentiable at $p \in \mathbb{S}^1$, we know from (4.51) that

$$\nabla X(p) = (I - pp^\top)\bar{\nabla}\bar{X}(p) = (I - pp^\top)(\bar{\nabla}Y(p) - p^\top Y(p)I). \tag{4.52}$$

Moreover, by a direct calculation, it is known that

$$\mathbf{P}_{p_0}(\bar{\partial}\bar{X}(p_0)) = (I - p_0p_0^\top)\bar{\partial}\bar{X}(p_0) = \mathrm{co}\left\{\pm\begin{pmatrix} 1 & -2 \\ -1 & 2 \end{pmatrix},\ \pm\begin{pmatrix} 1 & 2 \\ -1 & -2 \end{pmatrix}\right\}, \tag{4.53}$$

where "co" is the convex hull. Since $T_{p_0}\mathbb{S}^1 = \{v \in \mathbb{R}^2 : v^\top p_0 = 0\} = \{(t, -t) \in \mathbb{R}^2 : t \in \mathbb{R}\}$, then a linear operator on $T_{p_0}\mathbb{S}^1$ can be uniquely determined by a real number $a \in \mathbb{R}$, i.e., $(t, -t) \mapsto a(t, -t)$. As $\mathbf{P}_{p_0}(\partial\bar{X}(p_0)) \subset \mathcal{L}(T_{p_0}\mathbb{S}^1)$, we know the set $\mathbf{P}_{p_0}(\partial\bar{X}(p_0))$ is equivalent to $\mathrm{co}\{\pm 3, \pm 1\} = [-3, 3]$.

Next, we consider $\partial X(p_0)$. Let $p_k \in \mathbb{S}^1$ be a sequence converging to p_0. We may assume $p_k = (\cos\theta_k, \sin\theta_k)$, where $\theta_k \to \pi/4$ and $\theta_k \neq \pi/4$. The derivative of $X(p_k)$ can be given by (4.52). The second term in (4.52) converges to 0 since $Y(p_0) = 0$. Observe that $\bar{\nabla}Y(p_k) = \mathrm{diag}(2, -4)$ when $\theta_k > \pi/4$ and $\bar{\nabla}Y(p_k) = \mathrm{diag}(-2, 4)$ when $\theta_k < \pi/4$. Then, we know the Clarke generalized covariant derivative of X at p_0 is

$$\partial X(p_0) = \mathrm{co}\left\{\pm\begin{pmatrix} 1 & 2 \\ -1 & -2 \end{pmatrix}\right\},$$

which is equivalent to $\mathrm{co}\{\pm 1\} = [-1, 1]$. It is clear that $[-1, 1] \subset [-3, 3]$, and thus $\partial X(p_0) \subset \mathbf{P}_p(\bar{\partial}\bar{X}(p))$. This strict inclusion is mainly because we can only find the two tangent directions on $\mathbb{S}^1$ converging to p_0, while the other two normal directions are available only in $\mathbb{R}^2$.

We make the next assumption on the vector field X, which covers the problems in our experiments.

Assumption 4.5 *For any $p = (p_1, \ldots, p_d) \in \mathcal{M}$, the vector field $X(p) := F(p, f_1(p_1), \ldots, f_d(p_d)) \in T_p\mathcal{M}$ is locally Lipschitz, where*

(i) $F : \mathbb{R}^d \times \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d} \to T_p\mathcal{M}$ is continuously differentiable;
(ii) $f_j : \mathbb{R} \to \mathbb{R}^{n_j}$, $j = 1, \ldots, d$ and the non-differentiable points of f_j are isolated, i.e. for any point $q \in \mathbb{R}$, there exists some $\delta > 0$ such that f_j is continuously differentiable on $(q - \delta, q + \delta)\backslash\{q\}$;
(iii) The left and right derivatives of f_j exist.

Based on the above assumptions, the next lemma finds an element in $\partial X(q)$ by choosing a proper path that converges to the point $q \in \mathcal{M}$.

Lemma 4.12 *Fix $q \in \mathcal{M}$. Suppose that Assumption 4.5 holds, and $\{q^{(n)}\} \subset \mathcal{M}$ is a sequence converging to q such that for each $j \in [d]$, it holds that either $q_j^{(n)} > q_j$ for all $n \in \mathbb{N}$, or $q_j^{(n)} < q_j$ for all $n \in \mathbb{N}$. Then, we have $\mathbf{P}_q \circ dF|_q \circ (\mathrm{id}_{\mathbb{R}^d}, r_1, r_2, \ldots, r_d) \in \partial X(q)$, where $dF|_q$ is the differential of $F : \mathbb{R}^{d+\sum_{i=1}^d n_i} \to \mathbb{R}^d$ at $(q, f_1(q_1), \ldots, f_d(q_d))$, and $r_j : T_q\mathcal{M} \to \mathbb{R}^{n_j}$ is a linear operator such that*

$$r_j(v) = \begin{cases} v_j \lim\limits_{t\downarrow q_j} f_j'(t), & \text{if } q_j^{(n)} > q_j, \\ v_j \lim\limits_{t\uparrow q_j} f_j'(t), & \text{if } q_j^{(n)} < q_j. \end{cases} \tag{4.54}$$

Proof This is a direct consequence of Definition 2.11 and (4.51).

It is noted that the existence of the sequence $\{q^{(n)}\}$ depends on the manifold. The next theorem gives a construction of such a sequence on the Stiefel manifold, which relies on the following assumption.

Assumption 4.6 *Let $Q \in \mathrm{St}(n,r)$ and $V \in T_Q\mathrm{St}(n,r)$. Then, for all $i \in [n], j \in [r]$, it holds that $Q_{ij} \in \{\pm 1\}$ whenever $V_{ij} = 0$.*

Theorem 4.5 *Let $Q \in \mathrm{St}(n,r)$, $V \in T_Q\mathrm{St}(n,r)$, and $X(P) = F(P, f_{11}(P_{11}), f_{12}(P_{12}), \ldots, f_{nr}(P_{nr}))$ be a locally Lipschitz vector field on $\mathrm{St}(n,r)$ such that Assumption 4.5 and 4.6 hold. Then, $\mathbf{P}_Q \circ dF|_Q \circ (\mathrm{id}_{\mathbb{R}^{n\times r}}, H_{11}, \ldots, H_{nr}) \in \partial X(Q)$, where $H_{ij} : T_Q\mathrm{St}(n,r) \to \mathbb{R}^{n_{ij}}$ is a linear map satisfying $H_{ij}(W) = W_{ij}D_{ij}$ and*

$$D_{ij} = \begin{cases} \lim_{t\downarrow Q_{ij}} f'_{ij}(t), & V_{ij} > 0, \text{ or } V_{ij} = 0 \text{ but } Q_{ij} = -1, \\ \lim_{t\uparrow Q_{ij}} f'_{ij}(t), & V_{ij} < 0, \text{ or } V_{ij} = 0 \text{ but } Q_{ij} = 1. \end{cases}$$

Proof When $V_{ij} > 0$, we consider the sequence $Q^{(n)} = \exp_Q(t_n V)$ with $t_n \downarrow 0$. Since $\frac{d}{dt}\big|_{t=0} \exp_Q(tV) = V$, then $Q^{(n)}_{ij} > Q_{ij}$ for sufficiently large n. In the case where $V_{ij} = 0$ and $Q_{ij} = -1$, since $P^\top P = I_r$ for every $P \in \mathrm{St}(n,r)$, then there exists a sequence $Q^{(n)}$ converging to Q such that $Q^{(n)}_{ij} > -1 = Q_{ij}$ for all $n \in \mathbb{N}$. The other two cases are similar, and hence we can use Lemma 4.12 to find the derivative.

The following proposition shows that for a fixed $Q \in \mathrm{St}(n,r)$, Assumption 4.6 holds for almost every $V \in T_Q\mathrm{St}(n,r)$.

Proposition 4.3 *Let $Q \in \mathrm{St}(n,r)$ and $\mathcal{I} = \{(i,j) : V_{ij} = 0 \text{ for all } V \in T_Q\mathrm{St}(n,r)\}$, then the following properties hold:*

(i) If $(i,j) \in \mathcal{I}$, then $Q_{ij} \in \{\pm 1\}$;
(ii) If $(i,j) \notin \mathcal{I}$, then the set $\mathcal{Z}_{ij} := \{Z \in \mathbb{R}^{n\times r} : (\mathbf{P}_Q Z)_{ij} = 0\}$ has zero Lebesgue measure.

Proof First, the second property holds since $\mathcal{Z}_{ij}$ is a linear subspace in $\mathbb{R}^{n\times r}$ with $\dim \mathcal{Z}_{ij} < nr$.

Next, we prove that $Q_{ij} \in \{\pm 1\}$ for all $(i,j) \in \mathcal{I}$. If $(i_0, j_0) \in \mathcal{I}$, then the matrix $N \in (T_Q\mathrm{St}(n,r))^\perp$, where $N_{ij} = 1$ if $i = i_0, j = j_0$ and $N_{ij} = 0$ otherwise. From [2, Example 3.6.2], the normal space can be written as $(T_Q\mathrm{St}(n,r))^\perp = \{QS : S \in \mathbb{R}^{r\times r}, S^\top = S\}$. Without the loss of generality, we can assume $i = j = 1$ and there exists $S = S^\top$ such that $QS = N$, and then the equations $QS = N, Q^\top Q = I_r$ can be rewritten using block matrices:

$$\begin{pmatrix} a & x^\top \\ y & A \end{pmatrix} \begin{pmatrix} b & z^\top \\ z & B \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} a & y^\top \\ x & A^\top \end{pmatrix} \begin{pmatrix} a & x^\top \\ y & A \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & I_{r-1} \end{pmatrix},$$

where $a, b \in \mathbb{R}$, $y, z \in \mathbb{R}^{r-1}$, $x \in \mathbb{R}^{n-1}$, $A \in \mathbb{R}^{(n-1)\times(r-1)}$ and $B^\top = B \in \mathbb{R}^{(r-1)\times(r-1)}$.

From $a(ab + x^\top z) = a$ and $y^\top(yb + Az) = 0$, note that $a^2 + y^\top y = 1$ and $ax^\top + y^\top A = 0$, we know $a = b$. Similarly, $B = 0$ holds by combining equations $x(az^\top + x^\top B) = 0$, $A^\top(yz^\top + AB) = 0$, $xa + A^\top y = 0$ and $xx^\top + A^\top A = I_{r-1}$. Note $yz^\top + AB = 0$ and $B = 0$, then either $y = 0$ or $z = 0$ holds. Therefore, $1 = a^2 + y^\top y = a^2$ when $y = 0$; and $1 = ab + x^\top z = a^2$ when $z = 0$. Thus, $Q_{11} = a = \pm 1$. □

When $Z \in \mathbb{R}^{n\times r}$ is randomly sampled from a probability measure that is absolutely continuously with respect to the Lebesgue measure in $\mathbb{R}^{n\times r}$, the tangent vector $V := \mathbf{P}_Q Z$ fulfills Assumption 4.6 almost surely. Therefore, we derive the following algorithm, which finds an element of the Clarke generalized covariant derivative almost surely.

Algorithm 4.2 **Input:** $Q \in \mathrm{St}(n,r)$, **Output:** $H \in \partial X(Q)$.

(i) Sample $Z \in \mathbb{R}^{n\times r}$ from the standard Gaussian distribution.
(ii) Calculate the projection $V = \mathbf{P}_Q Z$.
(iii) If there exists (i,j) such that $V_{ij} = 0$ and $Q_{ij} \notin \{\pm 1\}$, re-run (i)-(ii).
(iv) Calculate H by Theorem 4.5.

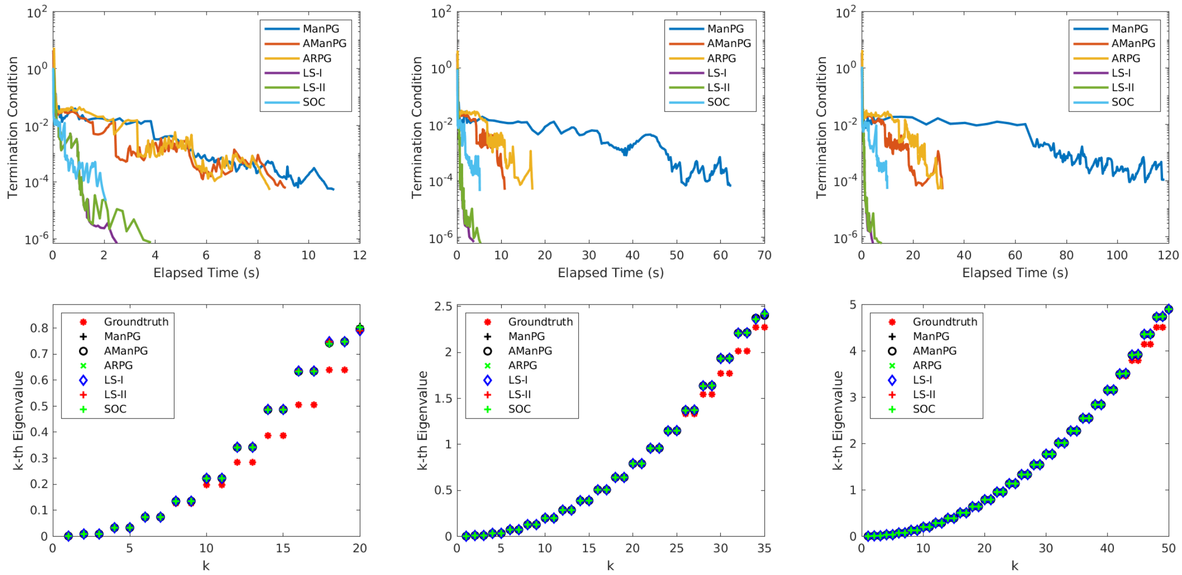

Fig. 1: Comparisons on CM with $(n, \mu) = (500, 0.05)$ and $r = 20, 35, 50$. The top row plots the termination condition, which is the sum of the left-hand parts in (5.3) and (5.4) for SOC, and similar for other algorithms. The bottom row plots the eigenvalues of $Q^\top HQ$, where Q is the solution of the corresponding method. "LS-I" and "LS-II" refer to our algorithms using the exponential map.

5 Numerical Experiments

In this section, we evaluate our algorithm on three problems mentioned before: compressed modes (CM) [52], sparse PCA and the constrained sparse PCA [49]. In CM and SPCA, we compare our algorithm with SOC [44], ManPG [19, ManPG-Ada (Algorithm 2)], accelerated ManPG (AManPG) [41], and accelerated Riemannian proximal gradient (ARPG) [42]. In the constrained SPCA, we compare our algorithm with ALSPCA [49]. Codes of SOC and ManPG are provided in [19], codes of AManPG and ARPG are provided by [42], and the code of ALSPCA is provided by [49]. All codes are implemented in MATLAB and evaluated on Intel i9-9900K CPU. Reported results are averaged over 20 runs with different random initial points.

5.1 Compressed Modes

We consider the CM problem (1.2) and follows the setting of [19] to solve the Schrödinger equation of 1D free-electron model with periodic boundary condition:

$$-\frac{1}{2}\Delta\phi(x) = \lambda\phi(x), \quad x \in [0, 50].$$

We discretize the domain $[0, 50]$ into n nodes, and let H be the discretized version of $-\frac{1}{2}\Delta$. The CM problem needs to solve (1.2). The first-order optimality condition is

$$0 \in 2\mathbf{P}_Q(HQ) + \mu\mathbf{P}_Q(\partial \|Q\|_1). \tag{5.1}$$

It is worth mentioning that this condition is difficult to check in general because of the existence of the projection. Recall that ManPG solves the following subproblem:

$$V_* = \underset{V \in T_Q\mathrm{St}(n,r)}{\arg\min} \left\{ \left\langle \mathrm{grad}_Q \operatorname{tr}(Q^\top HQ), V \right\rangle + \frac{1}{2t}\|V\|_F^2 + \|Q + V\|_1 \right\}.$$

The solution of the above problem satisfies

$$-V_*/t \in 2\mathbf{P}_Q(HQ) + \mu\mathbf{P}_Q\left(\partial \|Q + V_*\|_1\right).$$

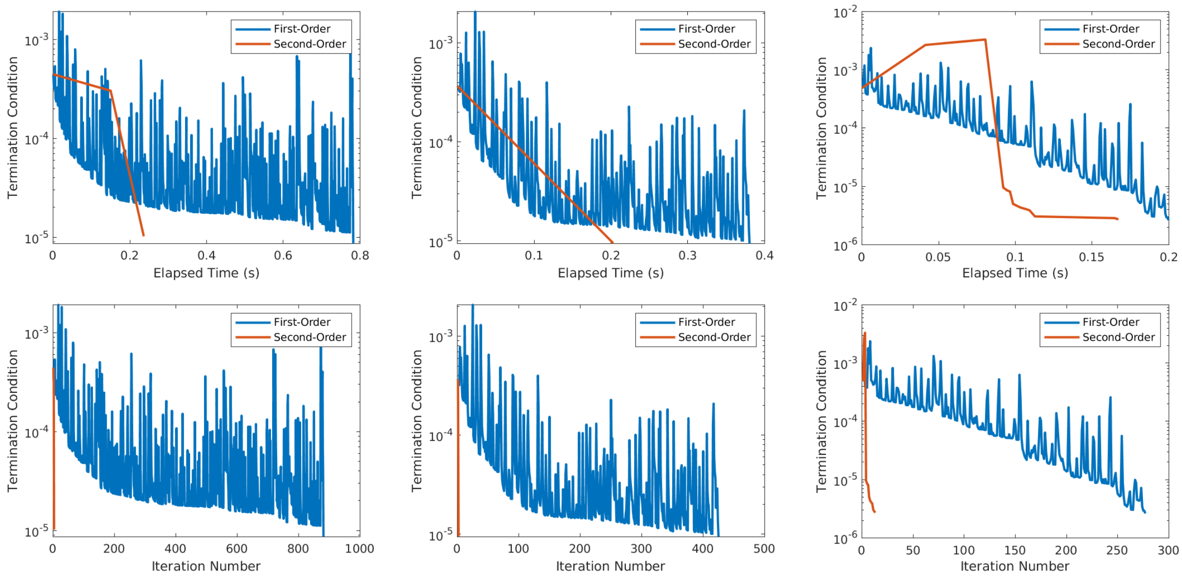

Fig. 2: Comparisons on the first-order method [62] and the second-order method in solving the CM subproblem (3.11). Columns represent the behavior of these methods in solving different subproblems in Algorithm 3.1.

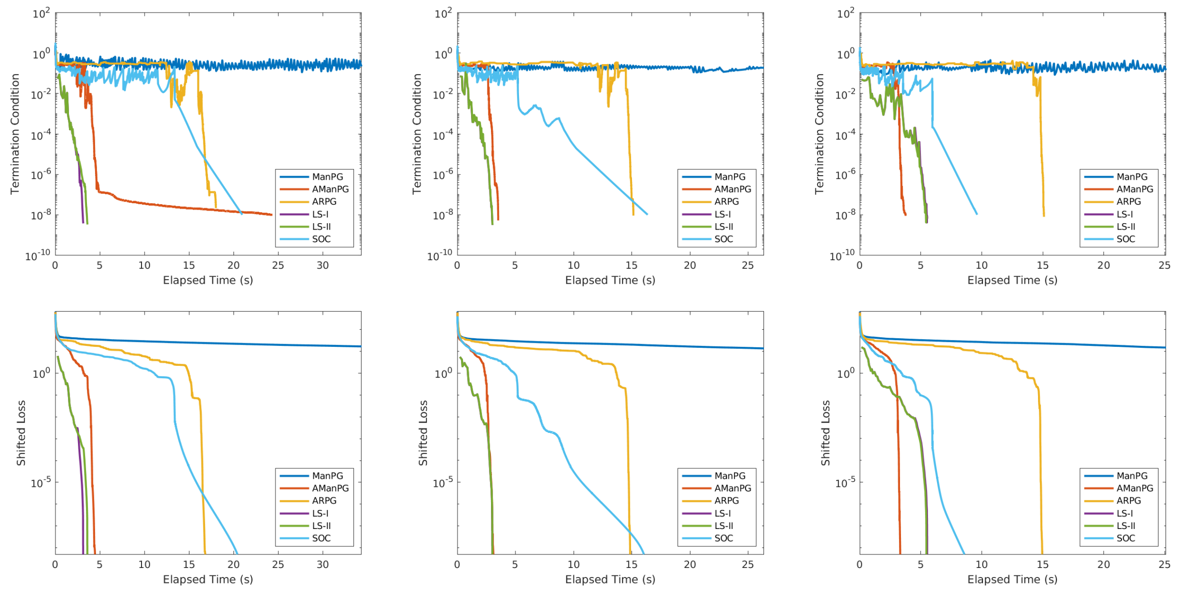

Fig. 3: Comparisons on SPCA with $(n, r, \mu) = (500, 20, 1.00)$ and three different $A^\top A$. The shifted loss is the loss subtracted by the minimal loss among all iterations, i.e., $F(x_k) - \min_j F(x_j)$. "LS-I" and "LS-II" refer to our algorithm using the retraction based on the QR decomposition.

As suggested by [19, 41], ManPG and AManPG use $t^{-1} \|V_*\|_\infty / (\|Q\|_F + 1) \leq 5 \times 10^{-5}$ as the termination condition, which can be regarded as an approximation of the optimality condition (5.1) of the CM problem.

For SOC and our algorithm, the termination rules are their KKT conditions. Specifically, SOC rewrites (1.2) into the following problem:

$$\begin{aligned} &\min_{Q,P,R \in \mathbb{R}^{n \times r}} \left\{ \operatorname{tr}(P^\top H P) + \mu \|R\|_1 \right\}, \\ &\text{s.t. } Q = P, R = P, Q \in \operatorname{St}(n, r). \end{aligned} \tag{5.2}$$

The Lagrangian of (5.2) is $L_S(Q,P,R,\Gamma,\Lambda) = \mathrm{tr}(P^\top HP) + \mu\|R\|_1 + \Gamma^\top(Q-P) + \Lambda^\top(R-P)$, where $P, R \in \mathbb{R}^{n\times r}$ and $Q \in \mathrm{St}(n,r)$. We terminate SOC when both the following conditions are satisfied.

$$\frac{\|Q-P\|_\infty}{\max\{\|Q\|_F, \|P\|_F\}+1} + \frac{\|R-P\|_\infty}{\max\{\|R\|_F, \|P\|_F\}+1} \le 5\times 10^{-7}, \tag{5.3}$$

$$\frac{\|\mathrm{grad}_Q L_S\|_\infty}{\|Q\|_F+1} + \frac{\|\nabla_P L_S\|_\infty}{\|P\|_F+1} + \frac{\min_{G\in\partial_R L_S}\|G\|_\infty}{\|R\|_F+1} \le 5\times 10^{-5}. \tag{5.4}$$

Similarly, our algorithm rewrites the problem into

$$\begin{aligned} &\min_{Q,R\in\mathbb{R}^{n\times r}} \left\{\mathrm{tr}(Q^\top HQ) + \mu\|R\|_1\right\}, \\ &\text{s.t. } Q = R, Q \in \mathrm{St}(n,r). \end{aligned} \tag{5.5}$$

The Lagrangian is $L_N(Q,R,\Lambda) = \mathrm{tr}(Q^\top HQ) + \mu\|R\|_1 + \Lambda^\top(Q-R)$, where $R \in \mathbb{R}^{n\times r}$ and $Q \in \mathrm{St}(n,r)$. The termination conditions are both (5.6) and (5.7):

$$\frac{\|Q-R\|_\infty}{\max\{\|Q\|_F, \|R\|_F\}+1} \le 5\times 10^{-7}, \tag{5.6}$$

$$\frac{\|\mathrm{grad}_Q L_N\|_\infty}{\|Q\|_F+1} + \frac{\min_{G\in\partial_R L_N}\|G\|_\infty}{\|R\|_F+1} \le 5\times 10^{-5}. \tag{5.7}$$

Below we illustrate how to apply our algorithm in this problem. In the subproblem of our algorithm, we need to find an approximated stationary point of $L_k(Q) = \mathrm{tr}(Q^\top HQ) + G_\sigma(Q + \Lambda/\sigma)$, where G_σ is the Moreau-Yosida regularization of $\lambda\|\cdot\|_1$. Let $\bar{L}_k : \mathbb{R}^{n\times r} \to \mathbb{R}$ be the extension of L_k into the Euclidean space, i.e., $\bar{L}_k|_{\mathrm{St}(n,r)} = L_k$. The Euclidean gradient and Hessian of $\bar{L}_k$ at a differentiable point Q can be easily computed as

$$\nabla\bar{L}_k(Q) = 2HQ + (\sigma Q + \Lambda) - \sigma F, \tag{5.8}$$

$$\nabla^2\bar{L}_k(Q)[Z] = 2HZ + \sigma Z \odot E, \tag{5.9}$$

where $\odot$ is the Hadamard product and $E, F \in \mathbb{R}^{n\times r}$ are matrices such that $E_{ij} = \mathbf{1}_{\{|Q_{ij}+\Lambda_{ij}/\sigma|\le\lambda/\sigma\}}$ and $F_{ij} = \mathrm{prox}_{\lambda\|\cdot\|_1/\sigma}(Q_{ij} + \Lambda_{ij}/\sigma)$. When $\bar{L}_k$ is differentiable at Q, the Riemannian gradient and Hessian can be written as [2]

$$\mathrm{grad}\, L_k(Q) = \mathbf{P}_Q\, \mathrm{grad}\, \bar{L}_k(Q), \tag{5.10}$$

$$\mathrm{Hess}\, L_k(Q)[Z] = \mathbf{P}_Q(\mathrm{Hess}\, \bar{L}_k(Q)[Z] - Z\,\mathrm{sym}(Q^\top \mathrm{grad}\, \bar{L}_k(Q))), \tag{5.11}$$

where $\mathrm{sym}\, Q := (Q + Q^\top)/2$. We note that $\mathrm{grad}\, L_k$ fulfills Assumption 4.5 with $f_{ij}(Q_{ij}) = F_{ij}$, and thus Algorithm 4.2 can be applied to find an element of $\partial\, \mathrm{grad}\, L_k$ when $\bar{L}_k$ is non-differentiable at Q.

The parameters of ManPG and SOC are the same as those in [19]. The codes and parameters of AManPG and ARPG are adopted from the SPCA implementation of Huang et al. [42]. All the four methods are terminated when their termination conditions are satisfied or the number of iterations exceeds 30000.

We consider the exponential map and two retractions from [2, Example 4.1.3] in Algorithm 4.1. One of these retractions is based on the polar decomposition:

$$R_X^{\mathrm{polar}}(V) := (X+V)(I + V^\top V)^{-\frac{1}{2}},$$

and the other one is based on the QR decomposition:

$$R_X^{\mathrm{QR}}(V) := \mathrm{qf}(X+V),$$

where $\mathrm{qf}(X)$ is the Q component in the QR decomposition of X.

In our algorithm, let Δ_k, Γ_k be the left-hand sides of (5.6) and (5.7) evaluated at the k-th step, respectively. We set $\tau = 0.97$, $\rho = 1.25$, $\alpha = 1.01$, $\sigma_0 = 1$, and $\varepsilon_k = \min\{0.95^k, 5\Delta_k\}$, in which the 0.95^k component ensures that $\varepsilon_k \to 0$ and the $5\Delta_k$ component suggests that the violation of the stationary condition (5.7) would not be large when the violation of the feasibility condition (5.6) is small. We also update σ_k using (3.14) when $\Delta_k > 2.5\Gamma_k$, i.e., the converse of the aforementioned case.

To solve our subproblem, we use the first-order method to find a good initial point, and start the second-order method when $\|\mathrm{grad}\, L_k\| < 5 \times 10^{-4} =: \Delta_G$, where L_k is defined in (3.11). The maximum numbers of iterations in CG and the first-order method are both 1000. In Algorithm 4.1, the linesearch parameter is $\mu = 0.1$. We set $\omega_k = \min\{0.7^k, 200\|X(p_k)\|\}$ and the minimal stepsize of the linesearch is 10^{-4}. When CG finds a negative direction p_k, we set $\omega_k = -2\mathrm{Hess}\, L_k(Q_k)[p_k, p_k]/\, \|p_k\|^2$ and restart the iteration. Parameters in the condition (4.7) are $p = 0.05, \beta_0 = 1, \beta_1 = 10^{-3}$. The threshold Δ_G will be modified during the iterations when one of the following cases occurs.

- When the linesearch does not find an m_k such that $\delta^{m_k} > 10^{-4}$, we replace Δ_G with $0.95\Delta_G$.
- When the number of iterations of Algorithm 4.1 exceeds 10, we replace Δ_G with $0.9\Delta_G$.
- When $\|V_k\| > 10^4$, i.e., $H_k + \omega_k I$ is nearly singular, we replace Δ_G with $0.8\Delta_G$.

These modifications of Δ_G may be helpful when Algorithm 4.1 starts too early. We also note that from Table 3 the second-order method generally starts (i.e., the condition $\|\,\mathrm{grad}\, L_k\| < \Delta_G$ holds) during nearly the last 1/4 iterations of Algorithm 3.1.

We report the results in Table 2 and Figure 1, from which we see that all methods find solutions with similar objective function values and comparable sparsity levels. It is noted that our method is generally faster than (or comparable to) other methods. The running time for different retractions in our algorithm are close, except that when $r = 10$ and 15, LS-I seems extremely slow because the second-order method starts too early such that most evaluations of second-order directions are wasted. This issue could be addressed by a careful tuning of parameters. From Figure 1, we find ManPG has difficulty attaining the desired termination condition when r is large. The performance of the two linesearch methods are similar, except for the aforementioned cases in which the second-order method does not start appropriately.

Next, we examine whether the second-order method is helpful in this problem. We terminate our method when both (5.6) and (5.7) are less than 5×10^{-8} and report the number of iterations and the computational cost of the second-order method used to solve subproblems.[9] We also try to run the first-order method [62] from the same initial point in each subproblem and terminate it when it attains the same accuracy as the second-order method. Results are illustrated in Figure 2 and reported in Table 4. From Table 4 and the bottom row of Figure 2, we see that to achieve the same accuracy, the number of iterations of the second-order method is significantly smaller than the first-order method. Since the second-order method requires solving a linear equation in each iteration, the first-order method may be faster than the second-order method in terms of the computational cost as can be observed in Table 4. However, by comparing the computational cost of these two methods in LS-I, we see that the second-order method can be faster in some cases due to its fast convergence, which is also illustrated in the top row of Figure 2. The average iteration number of CG and the percentage that CG detects a negative curvature direction are also reported in Table 4, from which we observe that negative curvature directions are frequently detected when $r = 10$ and 15. We suspect that this is possibly because the CM problem is more "singular" in these cases, which may also be supported by Table 5, in which the minimum eigenvalue is found to be significantly smaller.

Finally, although the CM problem appears "convex", the positive definiteness of $\partial\, \mathrm{grad}\, L_k$ required by Theorem 4.3 is neither obvious nor easy to theoretically verify. Indeed, finding all elements in $\partial\, \mathrm{grad}\, L_k$ is also difficult (see the discussions in Sec. 4.4). We partially verify this condition using numerical simulations and report the minimum eigenvalue of one element in $\partial\, \mathrm{grad}\, L_k$. The results in Table 5 suggest that the positive-definite condition may be fulfilled in this problem.

We also note that in the Euclidean setting the sparsity may be exploited to accelerate the conjugate gradient in the second-order method [48]. However, this is not straightforward in the Riemannian setting since the existence of the projection in the Riemannian Hessian may destroy the sparsity. Our current implementation of Algorithm 4.1 does not exploit the sparsity of the solution. It is believed that the second-order could be accelerated if we could exploit the sparsity to design a faster CG method.

5.2 Sparse PCA

In this section, we consider the SPCA problem (1.3). We compare our algorithm with AManPG, ARPG and SOC in high accuracy. The termination conditions are similar to those in CM and the threshold

[9] We only consider our method in this setting since other algorithms mentioned in Table 2 have difficulty in reaching these stopping conditions. Indeed, these observations are also valid for the original termination conditions with a lower significance.

Table 2: Comparison on CM. $(n, r, \mu) = (1000, 20, 0.1)$ and one of them varies. ManPG is the adaptive version (ManPG-Ada) in [19]. LSe, LSq, LSp denote our algorithm with the exponential map, the retraction using QR decomposition the retraction using polar decomposition, respectively.

		ManPG	AManPG	ARPG	SOC	LSe-I	LSe-II	LSq-I	LSq-II	LSp-I	LSp-II
		Running Time (s)									
n	200	11.54	3.86	6.43	1.78	1.20	1.70	1.21	1.76	**1.10**	1.54
	500	21.02	8.32	9.15	5.41	4.00	6.19	**3.96**	5.88	4.15	6.14
	1000	66.30	**8.60**	11.30	14.99	12.06	11.34	14.07	11.11	11.93	9.49
	1500	44.73	40.18	42.85	39.69	**24.00**	27.93	24.42	27.85	25.88	33.46
	2000	42.48	46.86	51.47	46.33	33.50	28.42	31.77	29.05	27.91	**26.22**
r	10	15.35	15.63	15.71	17.28	31.60	12.74	70.55	15.74	80.23	**11.26**
	15	35.35	**9.55**	11.17	15.47	49.14	12.07	53.59	11.74	51.58	12.01
	25	87.64	22.73	23.93	27.11	**17.39**	20.99	17.79	20.61	18.41	20.53
	30	83.13	27.41	29.35	18.10	11.49	15.49	11.40	15.13	**11.02**	14.42
μ	0.05	102.63	14.86	13.98	**6.34**	7.19	8.04	7.17	7.97	7.03	7.60
	0.15	65.08	17.09	21.48	32.51	14.22	16.03	14.13	15.69	**13.98**	15.59
	0.20	51.46	**12.96**	24.12	33.71	17.31	20.66	19.94	19.92	21.92	22.65
	0.25	42.74	**13.41**	25.43	34.80	18.08	20.84	61.23	53.44	35.93	44.30
		Loss Function: $\mathrm{tr}(X^\top HX) + \mu\|X\|_1$									
n	200	14.10	14.10	14.10	14.10	14.10	14.10	14.10	14.10	14.10	14.10
	500	18.60	18.60	18.60	18.60	18.60	18.60	18.60	18.60	18.60	18.60
	1000	23.30	23.30	23.30	23.30	23.30	23.30	23.30	23.30	23.30	23.30
	1500	26.90	26.80	26.80	26.80	26.80	26.80	26.80	26.80	26.80	26.80
	2000	29.80	29.70	29.70	29.70	29.70	29.70	29.70	29.70	29.70	29.70
r	10	10.70	10.70	10.70	10.70	10.70	10.70	10.70	10.70	10.70	10.70
	15	16.50	16.40	16.40	16.40	16.40	16.40	16.40	16.40	16.40	16.40
	25	32.00	32.00	32.00	32.00	32.00	32.00	32.00	32.00	32.00	32.00
	30	42.90	42.90	42.90	42.90	42.90	42.90	42.90	42.90	42.90	42.90
μ	0.05	15.10	15.10	15.10	15.10	15.10	15.10	15.10	15.10	15.10	15.10
	0.15	31.00	31.00	31.00	31.00	31.00	31.00	31.00	31.00	31.00	31.00
	0.20	38.30	38.20	38.20	38.20	38.20	38.20	38.20	38.20	38.20	38.20
	0.25	45.30	45.20	45.20	45.30	45.20	45.20	45.20	45.20	45.20	45.20

Table 3: The average iteration number of Algorithm 3.1 in the CM problem. The number in the bracket is the percentage that the second-order method starts.

		LSe-I	LSe-II	LSq-I	LSq-II	LSp-I	LSp-II
n	200	78 (23%)	78 (23%)	78 (23%)	79 (23%)	74 (25%)	74 (25%)
	500	75 (31%)	76 (32%)	75 (31%)	76 (32%)	74 (31%)	75 (32%)
	1000	46 (25%)	44 (22%)	45 (18%)	44 (21%)	44 (17%)	43 (20%)
	1500	57 (37%)	58 (38%)	57 (37%)	58 (38%)	57 (36%)	58 (38%)
	2000	56 (34%)	55 (34%)	56 (31%)	55 (34%)	55 (33%)	55 (35%)
r	10	103 (38%)	75 (16%)	93 (16%)	75 (16%)	98 (13%)	80 (17%)
	15	67 (33%)	48 (7%)	70 (4%)	48 (6%)	71 (4%)	50 (7%)
	25	61 (36%)	61 (36%)	61 (36%)	61 (37%)	61 (34%)	61 (35%)
	30	56 (27%)	59 (30%)	56 (27%)	59 (31%)	56 (28%)	58 (30%)
μ	0.05	67 (34%)	67 (34%)	67 (34%)	67 (34%)	67 (33%)	68 (34%)
	0.15	54 (32%)	54 (32%)	54 (32%)	54 (32%)	54 (33%)	55 (33%)
	0.20	51 (23%)	51 (24%)	51 (19%)	50 (21%)	53 (22%)	51 (22%)
	0.25	49 (9%)	48 (8%)	59 (8%)	49 (5%)	54 (7%)	49 (5%)

is set to 5×10^{-8}. In our algorithm, we set $\tau = 0.99$, the maximum numbers of iterations in CG and the first-order method are both 300. We set $\varepsilon_k = 0.9^k$ and the initial value of σ is $3\lambda_{\max}(A^\top A)$, where $\lambda_{\max}$ denotes the maximal eigenvalue. Other parameters of our algorithm are the same as those in CM. Note that since ManPG generally cannot achieve our requirement on the accuracy, we omit it in this experiment.

Table 4: Comparison on CM with higher accuracy. $(n, r, \mu) = (1000, 20, 0.1)$ and one of them varies. The termination threshold is 5×10^{-8}. The column Second-Order is the total number of iterations, CPU time, average CG iteration number of the second-order method, together with the percentage that CG detects the negative curvature direction (denoted by CG Rst). The column First-Order is the total number of iterations and CPU time of the first-order method [62] started from the same initial point as the second-order method and terminated when it attains the same optimality condition.

		Second-Order								First-Order			
		Iteration		Time (s)		Avg. CG Iter		CG Rst (%)		Iteration		Time (s)	
		LS-I	LS-II	LS-I	LS-II	LS-I	LS-II	LS-I	LS-II	LS-I	LS-II	LS-I	LS-II
		Exponential Map											
n	200	288	369	3.06	5.12	66.13	82.87	0.00	0.26	28924	23605	9.22	7.52
	500	298	358	5.73	10.44	60.53	88.65	0.00	0.00	39901	27498	18.83	13.17
	1000	244	90	7.22	4.27	45.13	96.49	4.34	24.03	3504	4374	2.65	3.24
	1500	251	363	26.61	65.24	156.09	256.81	1.66	3.97	68778	58475	70.11	60.56
	2000	203	202	21.52	33.61	157.13	249.48	4.72	0.18	37675	21100	53.22	27.54
r	10	267	373	24.93	76.76	290.48	568.20	21.91	18.23	3955	23636	1.94	11.84
	15	240	306	30.79	81.31	313.80	592.89	25.47	19.91	6277	11495	4.20	9.39
	25	237	334	14.51	28.99	104.94	141.85	0.00	0.00	29139	25811	27.10	23.94
	30	239	441	24.02	83.16	139.21	254.22	0.00	0.79	33426	20143	36.68	22.24
μ	0.05	174	292	5.61	13.44	73.71	93.62	0.00	0.00	9940	7906	7.40	5.91
	0.15	182	215	7.28	13.31	83.12	124.71	0.00	0.00	27398	19627	22.43	16.62
	0.20	121	178	6.11	27.31	104.88	297.51	1.13	4.47	13972	16872	10.42	13.02
	0.25	133	378	10.96	142.51	169.43	652.74	3.53	10.39	15413	23251	12.23	19.91
		Retraction using QR Decomposition											
n	200	264	377	2.39	5.12	60.13	83.44	0.00	0.00	25699	25639	8.15	8.16
	500	265	373	4.32	9.96	53.93	84.43	0.00	0.00	30973	29728	14.61	14.04
	1000	110	86	5.12	3.63	89.69	85.90	20.93	18.24	2478	4965	1.85	3.69
	1500	238	360	22.33	65.66	142.90	263.24	2.35	5.87	69516	59054	71.92	61.12
	2000	136	197	14.89	32.68	182.04	249.38	5.51	0.31	23681	21536	29.78	28.15
r	10	603	335	87.72	65.44	415.50	548.50	34.42	14.05	44448	26438	22.12	13.01
	15	393	331	50.38	70.59	286.62	481.59	44.88	23.94	24163	18022	16.03	11.89
	25	225	327	12.83	26.98	103.27	137.76	0.55	0.00	29205	24441	26.70	22.44
	30	177	439	16.66	77.70	145.66	238.25	0.00	0.00	27444	20960	29.86	23.22
μ	0.05	165	278	5.25	13.06	74.65	97.48	0.00	0.00	9431	7887	7.00	5.85
	0.15	136	217	5.64	12.71	98.06	120.97	0.00	0.00	21286	20372	17.11	17.39
	0.20	99	161	6.75	16.89	151.63	211.26	8.47	0.33	18256	13665	13.92	10.27
	0.25	210	354	33.98	89.84	313.85	450.47	20.33	0.36	79478	37584	61.58	29.45
		Retraction using Polar Decomposition											
n	200	274	369	2.62	5.07	61.67	83.38	0.00	0.00	28362	25357	9.02	8.06
	500	295	335	5.19	9.91	55.56	91.45	0.00	0.00	39868	27719	19.18	13.40
	1000	113	71	4.84	3.78	84.84	106.35	18.23	18.42	2783	5549	2.09	4.18
	1500	251	390	23.54	74.01	142.35	267.34	2.79	4.64	73268	62999	75.87	65.86
	2000	137	201	14.82	32.11	184.55	240.82	5.04	0.18	23561	21548	29.89	29.70
r	10	491	264	69.46	50.06	408.00	536.71	32.46	14.51	45323	34803	22.18	17.09
	15	394	274	50.06	56.21	284.55	462.63	44.71	23.48	25270	18112	16.94	12.11
	25	217	328	12.65	28.22	106.78	143.70	0.55	0.00	28576	26171	25.92	23.86
	30	160	432	15.85	76.47	156.98	239.43	0.00	0.00	24631	20308	26.58	22.41
μ	0.05	160	281	5.17	13.37	77.30	98.57	0.00	0.00	9345	8016	6.88	5.97
	0.15	136	214	5.73	12.49	96.72	122.45	0.00	0.00	21143	18790	17.08	14.89
	0.20	93	151	5.74	14.92	143.59	199.05	7.25	0.28	16898	13309	12.59	10.08
	0.25	194	281	26.74	72.03	269.42	465.45	20.43	0.41	75149	36278	57.65	28.19

Table 5: The minimum eigenvalue of the Hessian matrix of (3.11) in the CM problem (1.2). $(n, r, \mu) = (1000, 20, 0.1)$ and one of them varies. We report the results of 5 different runs. Let $\{x_k\}$ be the sequence generated by Algorithm 3.1 and I be the index set at which Algorithm 4.1 is called, then each number of this table is $\min_{k \in I} \lambda_{\min}(H_k)$, where $\lambda_{\min}$ is the minimum eigenvalue and $H_k \in \partial \operatorname{grad} L_k(x_k)$ is from Algorithm 4.2.

		Minimum Eigenvalue ($\times 10$)				
n	200	3.04960	3.05820	3.05160	3.23330	1.75420
	500	0.97663	1.08610	1.14290	1.17090	1.01000
	1000	0.13422	2.06550	0.22452	1.88420	0.05863
	1500	0.65285	0.61208	0.63683	0.86749	0.80422
	2000	0.24362	0.24363	0.24154	0.24101	0.23518
r	10	0.00264	0.00570	0.00287	0.00901	0.00005
	15	0.06096	0.05403	0.01370	0.00334	0.06086
	25	0.60886	0.60154	0.36930	0.76388	0.06192
	30	3.01550	6.18060	2.96390	6.18660	5.35900
μ	0.05	0.58725	1.03500	1.01720	1.12870	1.15590
	0.15	0.57333	0.73569	0.39304	0.57299	0.53624
	0.20	0.34736	0.20896	0.45627	0.34404	0.03558
	0.25	0.01539	0.16556	0.05790	0.01262	0.05763

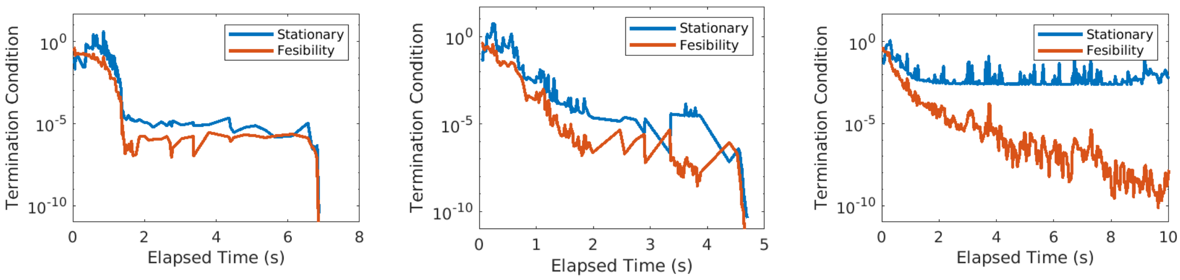

Fig. 4: Results of Algorithm 3.1 and 4.1 on constrained SPCA with $(n, \mu) = (600, 1.0)$ and $r = 3, 5, 10$, respectively. We use the first-order method to find an initial point for Algorithm 4.1 and start the second-order method when $\|\operatorname{grad} L_k\| \le 5 \times 10^{-4}$. "Feasibility" refers to the condition like (5.6) and "Stationary" refers to the condition like (5.7). The second-order method is able to start when $r = 3, 5$ and fails to start when $r = 10$.

The data matrix $A \in \mathbb{R}^{50 \times n}$ is generated as follows: First, we randomly generate A from the standard Gaussian distribution. Then, we modify the singular values of A to $\{w_i^4 + 10^{-5}\}_{i=1}^{50}$ to make A ill-conditioned, where $\{w_i\}$ is sampled from the standard Gaussian distribution. Finally, columns of A are normalized to have zero mean and unit length. Results are reported in Table 6 and Figure 3. We find that the losses of the solutions obtained by these methods are comparable, and our algorithm (LS-II) is faster than other methods when one of μ, r, n is large.

5.3 Constrained Sparse PCA

Since the SPCA problem in Sec. 5.2 has no guarantee on finding eigenvectors, the constrained SPCA problem (1.4) is considered in [49]. In this section, we compare our algorithm with the method in [49], which will be referred to as ALSPCA. In our implementation, we apply the first-order feasible method on Stiefel manifolds [62] to solve the subproblem when the iterates do not meet the start conditions of the semismooth Newton method. Following [49], the termination condition of subproblem is based on the feasibility conditions (5.6). In our algorithm, we set $\tau = 0.25$, $\rho = 10$, $\alpha = 1.01$, the maximum number of iterations of the first-order method is 2000. We set $\varepsilon_k = 0.1^k$ and the initial value of σ is 1. The parameters of ALSPCA are the same as those in [49]. The data matrix $A \in \mathbb{R}^{50 \times n}$ is generated from the standard Gaussian distribution and each column of A is normalized to have zero mean and unit length.

Table 6: Comparison on SPCA. $A \in \mathbb{R}^{50\times n}$, $(n, r, \mu) = (2000, 20, 1.0)$ and one of them varies. LSE, LSQ, LSP denote our algorithm with the exponential map, the retraction using QR decomposition the retraction using polar decomposition, respectively.

		AManPG	ARPG	SOC	LSE-I	LSE-II	LSQ-I	LSQ-II	LSP-I	LSP-II
		Running Time (s)								
n	500	6.77	48.25	18.85	7.69	7.60	5.12	5.19	5.14	**5.03**
	1000	12.05	79.78	34.94	14.66	18.24	13.56	**11.93**	13.19	12.06
	1500	34.72	87.92	57.74	29.49	28.77	24.94	25.12	27.76	**23.55**
	2000	34.64	112.64	94.48	33.30	42.67	47.19	34.15	45.30	**30.55**
	2500	57.50	116.51	124.66	45.11	42.63	46.11	**37.25**	52.52	37.63
	3000	81.80	134.70	158.95	77.56	81.44	99.62	**58.87**	79.36	64.45
r	5	**7.37**	25.00	55.96	20.86	19.32	17.73	17.86	18.37	16.04
	10	18.52	44.10	75.23	22.49	32.83	25.89	**18.12**	29.05	19.12
	15	38.14	67.22	86.39	35.84	31.50	38.00	**18.94**	30.52	19.12
	25	62.88	146.31	105.45	39.49	35.56	27.73	**27.45**	33.83	28.75
μ	0.25	**70.97**	80.56	93.19	116.40	141.50	89.27	96.15	87.74	97.62
	0.50	57.01	86.69	93.66	53.81	50.27	58.37	41.38	57.38	**38.75**
	0.75	51.82	98.02	94.44	37.36	39.85	43.89	32.15	39.06	**30.23**
	1.25	37.36	112.67	93.86	40.34	30.16	53.26	**23.25**	52.10	24.18
		Loss Function: $-\mathrm{tr}(X^\top A^\top AX) + \mu\|X\|_1$								
n	500	-337.79	-337.73	-337.32	**-337.82**	**-337.82**	**-337.82**	**-337.82**	**-337.82**	**-337.82**
	1000	-749.79	**-749.92**	-749.62	-749.61	-749.61	-749.61	-749.61	-749.61	-749.61
	1500	-1207.65	**-1207.83**	-1206.96	-1207.57	-1207.48	-1207.57	-1207.48	-1207.57	-1207.48
	2000	-1621.52	-1621.85	-1621.35	**-1622.69**	**-1622.69**	**-1622.69**	**-1622.69**	**-1622.69**	**-1622.69**
	2500	-2077.05	**-2077.30**	-2076.24	-2077.13	-2077.13	-2077.13	-2077.13	-2077.13	-2077.13
	3000	-2542.05	-2541.80	-2541.15	**-2542.36**	**-2542.36**	**-2542.36**	**-2542.36**	**-2542.36**	**-2542.36**
r	5	-1497.56	-1497.50	-1497.52	-1497.66	**-1497.67**	**-1497.67**	**-1497.67**	**-1497.67**	**-1497.67**
	10	-1640.15	-1640.17	-1639.74	**-1640.63**	**-1640.63**	**-1640.63**	**-1640.63**	**-1640.63**	**-1640.63**
	15	-1641.56	**-1641.60**	-1640.95	-1640.68	-1640.79	-1640.79	-1640.79	-1640.79	-1640.79
	25	-1636.00	-1636.46	**-1636.54**	-1636.48	-1636.48	-1636.48	-1636.48	-1636.48	-1636.48
μ	0.25	-1874.13	-1874.15	-1871.42	**-1874.23**	**-1874.23**	-1874.22	-1874.22	-1874.22	-1874.22
	0.50	**-1780.83**	-1780.81	-1778.62	-1780.71	-1780.69	-1780.69	-1780.69	-1780.69	-1780.69
	0.75	-1698.08	-1697.79	-1695.84	**-1698.17**	**-1698.17**	**-1698.17**	**-1698.17**	**-1698.17**	**-1698.17**
	1.25	-1552.05	-1552.28	-1552.04	**-1553.42**	-1553.41	-1553.41	-1553.41	-1553.41	-1553.41

Table 7: Comparison on constrained SPCA. $(n, r, \mu) = (2000, 20, 1.0)$ and one of them varies. EQUALITY and INEQUALITY are the logarithm of the violation of equality/manifold and inequality constraints, respectively.

		CPU (s)		SPARSITY (%)		CPAV (%)		EQUALITY		INEQUALITY	
		OURS	ALSPCA	OURS	ALSPCA	OURS	ALSPCA	OURS	ALSPCA	OURS	ALSPCA
r	5	**9.01**	12.17	**37.83**	34.43	11.82	**11.93**	**-15.31**	-9.19	-8.21	**-8.46**
	10	**14.67**	20.80	**38.16**	33.74	22.77	**23.19**	**-15.32**	-8.74	**-8.29**	-8.19
	15	**21.10**	31.29	**38.85**	38.62	**33.03**	32.97	**-15.27**	-8.73	**-8.33**	-8.19
	25	**32.51**	33.67	**39.23**	38.68	51.59	**52.09**	**-15.14**	-9.06	**-8.36**	-7.98
	30	**39.79**	39.84	39.65	**40.51**	60.09	**60.31**	**-15.12**	-9.03	**-8.23**	-7.91
n	500	**10.14**	17.13	**72.62**	68.67	**35.71**	35.61	**-15.23**	-8.93	**-8.32**	-8.26
	1000	**15.64**	22.97	**56.07**	49.80	40.57	**41.78**	**-15.20**	-8.99	**-8.23**	-7.96
	2000	**26.23**	28.91	**38.74**	38.01	42.65	**42.94**	**-15.12**	-8.76	-8.11	**-8.22**
	4000	52.48	**44.21**	**23.30**	21.65	43.22	**44.06**	**-15.08**	-9.13	**-8.62**	-7.89
	6000	71.72	**60.98**	**17.68**	16.99	43.13	**43.74**	**-15.22**	-8.89	**-8.79**	-7.94

In our experiments, we observe that our algorithm and ALSPCA find very different solutions for the same penalty parameter μ. To better compare their performance, we set $\mu = 1$ for our algorithm, and tune μ to achieve a comparable sparsity for ALSPCA. We compare the quality of solutions in terms of the cumulative percentage of adjusted variance (CPAV) proposed by [49]. It is defined as follows:

$$\mathrm{CPAV}(V) = \frac{1}{\mathrm{tr}(A^\top A)}\Big(\mathrm{tr}(V^\top A^\top AV) - \sqrt{\sum_{i\neq j}(V_i^\top A^\top AV_j)^2}\Big).$$

This quantity measures how well the data is explained by the found principal components. We set $\varDelta_{ij} = 10^{-8}$ and report the results in Table 7. We find solutions obtained from these algorithms are comparable in terms of CPAV, and we also find there is a trade-off between the sparsity and the explainability (CPAV). The speed of our algorithm is comparable with ALSPCA for large n, r, and is faster for small n, r. Moreover, as shown in the first two columns in Figure 4, the semismooth Newton method can start when $r = 3$ and $r = 5$. When $r = 10$, the behavior of our algorithm is similar to the first order methods as it is difficult to fulfill the stationary conditions like (5.7). One possible reason is that (1.4) has $r(r-1)$ inequality constraints such that the non-degeneracy condition required by Theorem 4.3 is likely violated for the case when r is large. A similar phenomenon has also been observed in SDP [68, 64]. Even in this case, we find that our algorithm is still faster than the ALSPCA method.

6 Conclusion

The paper proposes an augmented Lagrangian method for solving a class of nonsmooth optimization problems on manifolds with proved convergence. Using the Moreau-Yosida regularization, the augmented Lagrangian subproblem can be efficiently solved by the globalized semismooth Newton method, which exploits the second order geometry structure of manifolds. The local superlinear convergence rate is established under certain non-degenerate conditions that are similar to the case in the Euclidean space. Our numerical experiments on various applications show the advantages of the proposed method over the existing methods. The work done in this paper on ALM for solving the nonsmooth optimization problems on manifolds is by no means complete. There are many unanswered questions on both theory and algorithm design. For example, the convergence results of ALM for solving the nonsmooth manifold optimization problem are established under the constraint qualifications such as CPLD and LICQ. A systematical study on the convergence analysis of ALM under weaker assumptions is certainly of paramount necessity for solving the nonsmooth manifold optimization problems. Another direction is to design more efficient algorithms for solving the ALM subproblems in particular for the more challenging problems such as the constrained sparse PCA. It is our firm belief that a better using of the inherent second order geometry structure of manifolds rather than projecting them into the tangent spaces will lead to more efficient optimization methods for solving nonsmooth manifold optimization problems.

Acknowledgement

The authors would like to thank the two anonymous referees and the associate editor for their valuable comments and constructive suggestions, which have greatly improved the quality and the presentation of this paper.

References

1. P.-A. Absil and S. Hosseini, A collection of nonsmooth Riemannian optimization problems, in Nonsmooth Optimization and Its Applications, Springer, 2019, pp. 1–15.
2. P.-A. Absil, R. Mahony, and R. Sepulchre, Optimization algorithms on matrix manifolds, Princeton University Press, 2009.
3. P.-A. Absil and J. Malick, Projection-like retractions on matrix manifolds, SIAM Journal on Optimization, 22 (2012), pp. 135–158.
4. R. L. Adler, J.-P. Dedieu, J. Y. Margulies, M. Martens, and M. Shub, Newton's method on Riemannian manifolds and a geometric model for the human spine, IMA Journal of Numerical Analysis, 22 (2002), pp. 359–390.

5. R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt, On augmented Lagrangian methods with general lower-level constraints, SIAM Journal on Optimization, 18 (2008), pp. 1286–1309.
6. R. Andreani, G. Haeser, M. L. Schuverdt, and P. J. Silva, A relaxed constant positive linear dependence constraint qualification and applications, Mathematical Programming, 135 (2012), pp. 255–273.
7. H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality, Mathematics of Operations Research, 35 (2010), pp. 438–457.
8. D. Azagra, J. Ferrera, and F. López-Mesas, Nonsmooth analysis and Hamilton–Jacobi equations on Riemannian manifolds, Journal of Functional Analysis, 220 (2005), pp. 304–361.
9. 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.
10. D. P. Bertsekas, Constrained optimization and Lagrange multiplier methods, Academic Press, 1982.
11. D. P. Bertsekas, Nonlinear programming, Journal of the Operational Research Society, 48 (1997), pp. 334–334.
12. J. Bolte, S. Sabach, and M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Mathematical Programming, 146 (2014), pp. 459–494.
13. J. F. Bonnans and A. Shapiro, Perturbation analysis of optimization problems, Springer Science & Business Media, 2000.
14. 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.
15. N. Boumal and P.-a. Absil, RTRMC: A Riemannian trust-region method for low-rank matrix completion, in Advances in Neural Information Processing Systems, vol. 24, 2011.
16. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and Trends® in Machine learning, 3 (2011), pp. 1–122.
17. 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.
18. M. P. d. Carmo, Riemannian geometry, Birkhäuser, 1992.
19. S. Chen, S. Ma, A. Man-Cho So, and T. Zhang, Proximal gradient method for nonsmooth optimization over the stiefel manifold, SIAM Journal on Optimization, 30 (2020), pp. 210–239.
20. S. Chen, S. Ma, L. Xue, and H. Zou, An alternating manifold proximal gradient method for sparse principal component analysis and sparse canonical correlation analysis, INFORMS Journal on Optimization, 2 (2020), pp. 192–208.
21. W. Chen, H. Ji, and Y. You, An augmented Lagrangian method for ℓ_1-regularized optimization problems with orthogonality constraints, SIAM Journal on Scientific Computing, 38 (2016), pp. B570–B592.
22. X. Chen, L. Guo, Z. Lu, and J. J. Ye, An augmented Lagrangian method for non-Lipschitz nonconvex programming, SIAM Journal on Numerical Analysis, 55 (2017), pp. 168–193.
23. M. Cho and J. Lee, Riemannian approach to batch normalization, in Advances in Neural Information Processing Systems, vol. 30, 2017.
24. F. H. Clarke, Optimization and nonsmooth analysis, SIAM, 1990.
25. F. E. Curtis, H. Jiang, and D. P. Robinson, An adaptive augmented Lagrangian method for large-scale constrained optimization, Mathematical Programming, 152 (2015), pp. 201–245.
26. A. Daniilidis, R. Deville, E. Durand-Cartagena, and L. Rifford, Self-contracted curves in Riemannian manifolds, Journal of Mathematical Analysis and Applications, 457 (2018), pp. 1333–1352.
27. F. R. de Oliveira and O. P. Ferreira, Newton method for finding a singularity of a special class of locally Lipschitz continuous vector fields on Riemannian manifolds, Journal of Optimization Theory and Applications, 185 (2020), pp. 522–539.
28. F. R. de Oliveira and F. R. Oliveira, A global Newton method for the nonsmooth vector fields on Riemannian manifolds, Journal of Optimization Theory and Applications, (2021), pp. 1–15.

29. K. Deng and Z. Peng, An inexact augmented Lagrangian method for nonsmooth optimization on Riemannian manifold, arXiv preprint arXiv:1911.09900, (2019).
30. 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.
31. F. Facchinei and J.-S. Pang, Finite-dimensional variational inequalities and complementarity problems, Springer Science & Business Media, 2007.
32. O. Ferreira and P. Oliveira, Proximal point algorithm on Riemannian manifolds, Optimization, 51 (2002), pp. 257–270.
33. D. Gabay, Minimizing a differentiable function over a differential manifold, Journal of Optimization Theory and Applications, 37 (1982), pp. 177–219.
34. E. Ghahraei, S. Hosseini, and M. R. Pouryayevali, Pseudo-Jacobian and characterization of monotone vector fields on Riemannian manifolds, Journal of Convex Analysis, 24 (2017), pp. 149–168.
35. J.-B. Hiriart-Urruty and C. Lemaréchal, Convex analysis and minimization algorithms, vol. 2, Springer, 1993.
36. J.-B. Hiriart-Urruty, J.-J. Strodiot, and V. H. Nguyen, Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data, Applied Mathematics and Optimization, 11 (1984), pp. 43–56.
37. S. Hosseini and M. Pouryayevali, Generalized gradients and characterization of epi-Lipschitz sets in Riemannian manifolds, Nonlinear Analysis: Theory, Methods & Applications, 74 (2011), pp. 3884–3895.
38. J. Hu, X. Liu, Z.-W. Wen, and Y.-X. Yuan, A brief introduction to manifold optimization, Journal of the Operations Research Society of China, 8 (2020), pp. 199–248.
39. J. Hu, A. Milzarek, Z. Wen, and Y. Yuan, Adaptive quadratically regularized Newton method for Riemannian optimization, SIAM Journal on Matrix Analysis and Applications, 39 (2018), pp. 1181–1207.
40. W. Huang, P.-A. Absil, and K. A. Gallivan, A Riemannian symmetric rank-one trust-region method, Mathematical Programming, 150 (2015), pp. 179–216.
41. W. Huang and K. Wei, Extending FISTA to Riemannian optimization for sparse PCA, arXiv preprint arXiv:1909.05485, (2019).
42. W. Huang and K. Wei, Riemannian proximal gradient methods, Mathematical Programming, (2021), pp. 1–43.
43. 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.
44. R. Lai and S. Osher, A splitting method for orthogonality constrained problems, Journal of Scientific Computing, 58 (2014), pp. 431–449.
45. 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.
46. J. M. Lee, Introduction to smooth manifolds, Springer, 2012.
47. J. M. Lee, Introduction to Riemannian manifolds, Springer, 2018.
48. 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 (2018), pp. 433–458.
49. Z. Lu and Y. Zhang, An augmented Lagrangian approach for sparse principal component analysis, Mathematical Programming, 135 (2012), pp. 149–193.
50. R. Mifflin, Semismooth and semiconvex functions in constrained optimization, SIAM Journal on Control and Optimization, 15 (1977), pp. 959–972.
51. A. Montanari and E. Richard, Non-negative principal component analysis: Message passing algorithms and sharp asymptotics, IEEE Transactions on Information Theory, 62 (2015), pp. 1458–1484.
52. 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 of the United States of America, 110 (2013), pp. 18368–18373.
53. L. Qi and J. Sun, A nonsmooth version of Newton's method, Mathematical Programming, 58 (1993), pp. 353–367.

54. L. Qi and Z. Wei, On the constant positive linear dependence condition and its application to SQP methods, SIAM Journal on Optimization, 10 (2000), pp. 963–981.
55. F. Rampazzo and H. J. Sussmann, Commutators of flow maps of nonsmooth vector fields, Journal of Differential Equations, 232 (2007), pp. 134–175.
56. R. T. Rockafellar, Convex analysis, 28, Princeton University Press, 1970.
57. M. D. Spivak, A comprehensive introduction to differential geometry, vol. 2, Publish or Perish, 3rd ed., 1999.
58. D. Sun and J. Sun, Semismooth matrix-valued functions, Mathematics of Operations Research, 27 (2002), pp. 150–169.
59. D. Sun, J. Sun, and L. Zhang, The rate of convergence of the augmented Lagrangian method for nonlinear semidefinite programming, Mathematical Programming, 114 (2008), pp. 349–391.
60. B. Vandereycken, Low-rank matrix completion by Riemannian optimization, SIAM Journal on Optimization, 23 (2013), pp. 1214–1236.
61. I. Waldspurger, A. d'Aspremont, and S. Mallat, Phase recovery, maxcut and complex semidefinite programming, Mathematical Programming, 149 (2015), pp. 47–81.
62. Z. Wen and W. Yin, A feasible method for optimization with orthogonality constraints, Mathematical Programming, 142 (2013), pp. 397–434.
63. 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, 76 (2018), pp. 364–389.
64. L. Yang, D. Sun, and K.-C. Toh, SDPNAL+: A majorized semismooth Newton-CG augmented Lagrangian method for semidefinite programming with nonnegative constraints, Mathematical Programming Computation, 7 (2015), pp. 331–366.
65. 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.
66. 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.
67. H. Zhang and S. Sra, First-order methods for geodesically convex optimization, in Conference on Learning Theory, 2016, pp. 1617–1638.
68. X.-Y. Zhao, D. Sun, and K.-C. Toh, A Newton-CG augmented Lagrangian method for semidefinite programming, SIAM Journal on Optimization, 20 (2010), pp. 1737–1765.
69. H. Zhu, X. Zhang, D. Chu, and L.-Z. Liao, Nonconvex and nonsmooth optimization with generalized orthogonality constraints: An approximate augmented Lagrangian method, Journal of Scientific Computing, 72 (2017), pp. 331–372.
70. X. Zhu and H. Sato, Riemannian conjugate gradient methods with inverse retraction, Computational Optimization and Applications, 77 (2020), pp. 779–810.
71. H. Zou, T. Hastie, and R. Tibshirani, Sparse principal component analysis, Journal of Computational and Graphical Statistics, 15 (2006), pp. 265–286.