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

A Semi-smooth 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 inexact 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 manifold, 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. Finally, numerical experiments on compressed modes and (constrained) sparse PCA 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 [52, 14], principal component analysis [41, 43], matrix completion [12, 51], medical image analysis [38], and deep learning [19]. 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 [27, 2, 53, 57, 32]. However, nonsmooth manifold optimization is less explored but has drawn increasing attention in recent years [37, 17, 35, 16]. 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, $\psi : \mathbb{R}^m \to \mathbb{R}$ is convex, proper and lower semicontinuous. Besides the nonsmooth function ψ in (1.1), the inequality constraint can be seen as the nonsmooth function $\mathcal{I}_{\mathcal{C}}(x)$, where $\mathcal{I}_{\mathcal{C}}$ is the indicator function of the set $\mathcal{C}$ and $\mathcal{C} := \{x : h_2(x) \leq 0\}$. From the perspective of algorithm design, the

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 by the National Natural Science Foundation of China (12071464, 11671387, 11531014, 11688101) 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
Institue 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.

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, 16, 32] for more examples and details.

1. **Compressed modes (CM) [44].** The CM task 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, it is formulated as the following optimization problem:

$$\min_{Q\in \mathrm{St}(n,r)} \ \mathrm{tr}(Q^\top H Q) + \mu \left\|Q\right\|_1, \tag{1.2}$$

   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) [61].** Let $A \in \mathbb{R}^{p\times n}$ be the data matrix where n and p are the number of variables and the number of observations, respectively. Set $f(Q) = -\mathrm{tr}(Q^\top A^\top A Q)$, $\psi(Q) = \mu \left\|Q\right\|_1$, $h_1(Q) = Q$, $h_2(Q) \equiv 0$ and $\mathcal{M} = \mathrm{St}(n,r)$. Then, the SPCA problem has the form (1.1), i.e. it aims at solving

$$\min_{Q\in \mathrm{St}(n,r)} \ -\mathrm{tr}(Q^\top A^\top A Q) + \mu \left\|Q\right\|_1. \tag{1.3}$$

3. **Constrained SPCA [41].** To further enforce the orthogonality among principle components, the constrained SPCA problem imposes additional constraints on each column of Q, which has the form

$$\begin{aligned} \min_{Q\in \mathrm{St}(n,r)} \ & -\mathrm{tr}(Q^\top A^\top A Q) + \mu \left\|Q\right\|_1, \\ \text{s.t.} \quad & |Q_i^\top A^\top A Q_j| \le \Delta_{ij}, \ \forall\, i \ne j, \end{aligned} \tag{1.4}$$

   where Q_i denotes the i-th column of Q and $\Delta_{ij} \ge 0$ for $i \ne j$ 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 [25, 11]. Riemannian proximal point algorithms are investigated by [26, 35, 16]. Operator splitting methods like the alternating direction methods of multipliers (ADMM) and the augmented Lagrangian methods (ALM) are also promising on manifolds [36, 24], where 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 [16] or the Riemannian proximal gradient method [35] for solving (1.1) requires solving a subproblem with two nonsmooth terms, which is difficult to solve. Moreover, by introducing more auxiliary variables, the multiblock ADMM method is applicable for solving (1.1), 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 an 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 [41, 37, 17, 60], which has been widely studied for many years [3, 50, 13, 18]. 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.

Notations	Descriptions
$[u]_i$	The i-th component of $u \in \mathbb{R}^d$
$[M]$	The set $\{1, 2, \ldots, M\}$, 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 Hseesian 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 geodestic 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 $T_p\mathcal{M}$.

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 [16] and the Riemannian proximal gradient method [35], 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. 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.

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.

2.1 Preliminaries on Manifolds

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 for functions defined on manifolds.

Definition 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 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 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 $g \in \mathcal{D}(\mathcal{N})$.

Definition 4 Let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function on $\mathcal{M}$. Given $p \in \mathcal{M}$, 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 $X(p) \in T_p\mathcal{M}$, $\forall p \in \mathcal{M}$.

Definition 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 [15] and can be used to define the parallel transport of a vector field.

Definition 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 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).

Definition 7 Let X be a vector field. The directional derivative at $p \in \mathcal{M}$ along $v \in T_p\mathcal{M}$ is

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

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 $\nabla X(p; v) = \nabla_v X(p)$.

Definition 8 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 [15, 39, 2] for more details about manifolds.

[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 Chpater 5 in [2].

2.2 Nonsmooth Manifold Optimization

As suggested in [16], 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 [25, 11] 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 [6, 25, 30], 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. The update rule of the subgradient method in the Riemannian setting [25, 11] 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 [16], it is also observed that subgradient based methods are slower than proximal point algorithms and operator splitting methods in the Riemannian setting.

2.2.2 Proximal Point Methods

The extension of proximal point algorithms on manifolds is proposed in [26] and the subgradient methods are suggested in [7] to solve the subproblem. In [26], manifolds with nonpositive 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 on the Stiefel manifold [16] with proved convergence. More specifically, it aims at solving the following problem:

$$\min_{Q \in \mathcal{M}} \ f(Q) + \psi(Q), \tag{2.1}$$

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

$$V_k := \operatorname*{arg\,min}_{V \in T_{Q_k}\mathcal{M}} \langle \mathrm{grad}\, f(Q_k), V\rangle + \frac{1}{2t}\|V\|_F^2 + \psi(Q_k + V) \tag{2.2}$$

via the regularized semismooth Newton method [54]. Besides, Huang et al. proposed a Riemannian proximal gradient method [35] to solve (2.1) for general manifolds by replacing the term $\psi(Q_k + V)$ with $\psi(\mathcal{R}_{Q_k}(V))$, where $\mathcal{R}_{Q_k}$ is a retraction and analyzed the iteration complexity for convex objectives under some assumptions. However, the direct application application of the above two methods for solving (1.1) has to solve the subproblem:

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

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

[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).

2.2.3 Operator Splitting Methods

Operator splitting methods on manifolds split (2.1) into several terms, each of which is easier to solve. For example, the manifold ADMM proposed in [36] rewrites (2.1) to

$$\min_{Q,Z} f(Q) + \psi(Z) \quad \text{s.t.} \quad Q = Z,\ Q \in \mathcal{M}. \tag{2.3}$$

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

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

The X-update requires smooth manifold optimization algorithms and the Z-update is the proximal mapping of ψ. Besides, an inexact ALM framework to solve (2.3) with some convergence results is considered in [24].

When a manifold is embeded in an Eucledean space, classical nonsmooth nonconvex constrained optimization algorithms can also be explored. Lai et al. proposed a splitting method for orthogonality constrained problems (SOC) [37], which reformulates (2.1) into

$$\min_{Q,P,R} f(P) + \psi(R) \quad \text{s.t.} \quad P = R,\ Q = P,\ Q^\top Q = I_r. \tag{2.4}$$

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

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

where the R-update can be solved by a proximal map, 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 (2.4).

Unlike ADMM, ALM usually has theoretical guarantees. In [17], 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 [5]. In [60], the so-called EPALMAL is proposed, where the PALM [10] is used for solving the subproblem. Although both PAMAL and EPALMAL have certain convergence analysis, 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. It has been studied for many decades [9]. Consider the following problem:

$$\min_{x\in\mathbf{R}^n} f(x) + \Phi(x) \quad \text{s.t.} \quad g(x) = 0,\ h(x) \le 0, \tag{2.5}$$

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

$$\mathcal{M} = \{x : g_1(x) = 0,\ h_1(x) \le 0\}, \tag{2.6}$$

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 [3] to solve (2.5). It is shown that any feasible limit point generated by the algorithm is a KKT point under the constant positive linear dependence (CPLD) condition [46], 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 is unbounded. This infeasiblity phenomenon also exists in other literature such as [20].

To alleviate this issue, another AL method is proposed in [41], where two nonmonotone proximal methods are applied to solve the subproblem. They consider the problem (2.5) with an 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 [41] requires that the magnitude of the penalty parameter outgrows that of multipliers. 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. [18] proposed an AL method to solve (2.5) with Φ possibly being a nonconvex non-Lipschitz function. Under a weak constraint qualification called the relaxed constant positive linear dependence (RCPLD) condition [4], they provided a global convergence result.

2.4 Semismooth Newton Methods

The subproblem in ALM generally requires solving a nonsmooth equation, which usually can be efficiently solved by the semismooth Newton method [42, 45, 49]. Under suitable assumptions, the semismooth Newton method has the local superlinear convergence rate. Recently, it is generalized to solving nonsmooth equations [22] on manifolds based on the Clarke generalized covariant derivatives [47, 28]. Below we introduce several definitions for locally Lipschitz vector fields on manifolds.

Definition 9 ([22]) We say a vector field $X : \mathcal{M} \to T\mathcal{M}$ on a manifold $\mathcal{M}$ is locally Lipschitz if for each $p \in \mathcal{M}$ there exist a neighborhood $U \ni p$ and a constant $L_p > 0$ such that for each $x, y \in U$, geodesic γ joining x, y, it has

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

Since a locally Lipschitz vector field X on manifolds is differentiable almost everywhere [22], we denote $\mathcal{D}_X$ as the set of its differentiable points and define the Clarke generalized covariant derivative as follows:

Definition 10 ([28, 22]) Let X be a locally Lipschitz vector field on $\mathcal{M}$. The B-derivative is a set-valued map $\partial_B X : \mathcal{M} \quad \mathcal{L}(T\mathcal{M})$ with

$$\partial_B X(p) := \left\{ H \in \mathcal{L}(T_p\mathcal{M}) :\ \exists \{p_k\} \subset \mathcal{D}_X,\ \lim_{k\to+\infty} p_k = p,\ H = \lim_{k\to+\infty} \nabla X(p_k) \right\}, \tag{2.7}$$

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 $\partial X : \mathcal{M} \quad \mathcal{L}(T\mathcal{M})$ such that $\partial X(p)$ is the convex hull of $\partial_B X(p)$.

The above definitions are consistent with the definitions 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, and thus 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 nonempty compact sets and the maps $\partial_B X, \partial X$ are locally bounded and upper semi-continuous [22]. Having introduced these notions, the Newton method for a locally Lipschitz vector field X on $\mathcal{M}$ is [22]:

$$p_{k+1} := \exp_{p_k}(-H_k^{-1} X(p_k)), \quad \text{where } H_k \in \partial X(p_k). \tag{2.8}$$

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

Definition 11 ([22]) 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

$$\left\| X(p) - P_{qp}[X(q) + H_q \exp_q^{-1} p] \right\| \le \varepsilon d(p,q)^{1+\mu},\ \forall\, q \in B_\delta(p), H_q \in \partial X(q), \tag{2.9}$$

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].

[4] This is well-defined in a small neighborhood of q [15, Proposition 3.2.9].

In [22], 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 (2.8) has the local convergence rate $1+\mu$. This result is similar to that in Euclidean spaces [42, 45, 49].

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.

3.1 Algorithm

Recall that we consider the following optimization problem:

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

Throughout this paper, we always make the following assumptions:

Assumption 1 *$\mathcal{M}$ is a complete smooth Riemannian manifold.*

Assumption 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 proper, convex and lower semi-continuous 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:

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

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

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

Note that the proximal map of a convex and lower semi-continuous function f is defined as

$$\operatorname{prox}_f(x) := \arg\min_y f(y) + \frac{1}{2}\|x - y\|^2,$$

and simultaneously minimizing L_σ with respect to x, y, z is equivalent to

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

where ψ^σ, $\delta^\sigma_{\mathsf{R}^q}$ are the Moreau-Yosida regularization of ψ, δ_{R^q}. More specifically,

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

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

where δ_{R^q} is the indicator function of $\mathbb{R}^q_-$. The minimization problems (3.5) and (3.6) are related to finding the proximal maps of ψ/σ and δ_{R^q}, which can be easily solved in many cases. In addition, since $\delta^\sigma_{\mathsf{R}^q}$ 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 [41, 18].

Algorithm 3 (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 $y_0 = \operatorname{prox}_{\psi/\sigma_0}(h_1(x_0) + \lambda_0/\sigma_0)$, $z_0 = \Pi_{\mathsf{R}^q}(h_2(x_0) + \gamma_0/\sigma_0)$, where Π_{R^q} is the projection onto $\mathbb{R}^q_-$. Choose a feasible point x_{feas} and a constant Φ such that*

$$\Phi \geq \max\{f(x_{\text{feas}}) + \psi(h_1(x_{\text{feas}})), \ L_{\sigma_0}(x_0, y_0, z_0, \lambda_0, \gamma_0)\}. \tag{3.7}$$

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

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

$$x_k \approx \arg\min_{x\in\mathcal{M}} \; L_k(x) := f(x) + \psi^{\sigma_k}\left(h_1(x) + \frac{\lambda_k}{\sigma_k}\right) + \delta_{\mathsf{R}^q}^{\sigma_k}\left(h_2(x) + \frac{\gamma_k}{\sigma_k}\right). \tag{3.8}$$

Speci cally, we need to nd $x_k \in \mathcal{M}$ that satis es

$$\|\mathrm{grad}\, L_k(x_k)\| < \varepsilon_k, \quad L_k(x_k) \le \Phi, \tag{3.9}$$

(ii). Update y and z using

$$y_k = \mathrm{prox}_{\psi/\sigma_k}\left(h_1(x_k) + \frac{\lambda_k}{\sigma_k}\right), \tag{3.10}$$

$$z_k = \Pi_{\mathsf{R}^q}\left(h_2(x_k) + \frac{\gamma_k}{\sigma_k}\right). \tag{3.11}$$

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

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

3.2 Convergence Analysis

In this part, motivated by the proof in [18], we first give the feasibility result of Algorithm 3.

Theorem 4 *Let $\{(x_k, y_k, z_k)\}$ be the sequence generated by Algorithm 3. Then, we have*

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

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

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

$$\sigma_k = \max\left\{\rho\sigma_{k_i\ 1}, \|\lambda_k\|_2^{1+\alpha}, \|\gamma_k\|_2^{1+\alpha}\right\},\ \forall\, k_i \le k < k_{i+1}.$$

From (3.9) 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

$$\left\|h_1(x_k) - y_k + \frac{\lambda_k}{\sigma_k}\right\|_2^2 + \left\|h_2(x_k) - z_k + \frac{\gamma_k}{\sigma_k}\right\|_2^2 \le 2\frac{\Phi - f(x_k) - \psi(y_k)}{\sigma_k} + \frac{\|\lambda_k\|_2^2 + \|\gamma_k\|_2^2}{\sigma_k^2}. \tag{3.12}$$

Since $f + \psi$ is bounded below and $\sigma_k \to \infty$, we know the first term of (3.12) 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 (3.12) 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 $k_i < k \le k_{i+1}$, we have

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

Also note $\delta_k \leq \tau\delta_{k-1}$ for all $k_i \leq k < k_{i+1} - 1$. By induction, we know

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

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 (3.12), we conclude $\lim_{k\to\infty} \delta_k = 0$.

Next, we consider the convergence result of Algorithm 3 by introducing the extension of the constraint qualifications on manifolds [56]. Consider 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 [56]:

Definition 12 (LICQ) We say that a feasible point $x \in \mathcal{M}$ of (3.1) satisfies the linear independence constraint qualification (LICQ) if $\{\operatorname{grad}[h_2(x)]_i : i \in \mathcal{A}(x)\}$ are linearly independent in $T_x\mathcal{M}$.

It is easy to see that this 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 13 (CPLD) Let $x \in \mathcal{M}$ be a feasible point of (3.1) and define $S(x) := \{\operatorname{grad}[h_2(x)]_i : i \in \mathcal{A}(x)\}$. 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 [56, Theorem 4.1].

Corollary 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 LICQ holds at x_*, then there exists a multiplier γ_* such that the following KKT conditions hold:*

$$0 \in \partial_x \mathcal{L}(x_*, \gamma_*), \tag{3.13a}$$

$$h_2(x_*) \in \mathbb{R}^q_-,\ \gamma_* \in \mathbb{R}^q_+,\ \gamma_*^\top h_2(x) = 0. \tag{3.13b}$$

Remark 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 [56]:

$$y = h_1(x),\ z = h_2(x), \tag{3.14a}$$

$$z \in \mathbb{R}^q_-,\ \gamma \in \mathbb{R}^q_+,\ \gamma^\top z = 0, \tag{3.14b}$$

$$0 \in \partial_y \tilde{\mathcal{L}}(x, y, z, \lambda, \gamma), \tag{3.14c}$$

$$\operatorname{grad}_x \tilde{\mathcal{L}}(x, y, z, \lambda, \gamma) = 0. \tag{3.14d}$$

We have the following relationship between these optimality conditions:

Proposition 1 *When* (3.13) *holds, then* (3.14) *also holds. Consequently, when LICQ holds at a local minimum x_* of* (3.1)*, then* (3.14) *holds.*

Proof From (3.13), we can choose $g \in \partial_x \psi(h_1(x_*))$ such that $\operatorname{grad} f(x_*) + g + \sum_{i=1}^q \gamma_i \operatorname{grad}[h_2(x_*)]_i = 0$. Then, there exist $\{g_i\}_{i\in[s]} \subset \partial_B \psi(h_1(x_*))$ and $\{\alpha_i\}_{i\in[s]} \subset \mathbb{R}_+$ such that $\sum_{i=1}^s \alpha_i = 1$ and $g = \sum_{i=1}^s \alpha_i g_i$. We can find $\{x_k^{(i)}\}_{k\in\mathbb{N}, i\in[s]} \subset \mathcal{M}$ such that $\lim_{k\to\infty} x_k^{(i)} = x_*$, and $\psi \circ h_1$ is differentiable at $x_k^{(i)}$, and $\lim_{k\to\infty} \operatorname{grad} \psi(h_1(x_k^{(i)})) = g_i$. Note that $\operatorname{grad} \psi(h_1(x_k^{(i)})) = \sum_{j=1}^m [\operatorname{grad} \psi(y_k^{(i)})]_j \operatorname{grad}[h_1(x_k^{(i)})]_j$, where $y_k^{(i)} = h_1(x_k^{(i)})$. Since $y_k^{(i)} \to y := h_1(x_*)$, by passing to subsequences if necessary, the limit $\lambda_*^{(i)} := \lim \operatorname{grad} \psi(y_k^{(i)})$ exists. Thus, $\lambda_*^{(i)} \in \partial\psi(y)$. By setting $z = h_2(x_*)$, $\gamma = \gamma_*$ and $\lambda = \sum_{i=1}^s \alpha_i \lambda_*^{(i)}$, we know (3.14) holds.

Finally, we show that Algorithm 3 converges to a KKT point.

Theorem 5 *Suppose there exist* $K \subseteq \mathbb{N}$ *and* $(x_*, y_*, z_*) \in \mathcal{M} \times \mathbb{R}^m \times \mathbb{R}^q_-$ *such that*

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

If CPLD holds at x_**, then there exist* $K_0 \subseteq K$ *and* $\lambda_* \in \mathbb{R}^m, \gamma_* \in \mathbb{R}^q_+$ *such that* $\lim_{k\in K_0} \lambda_{k+1} = \lambda_*$*, and the KKT conditions* (3.14) *hold at* $(x_*, y_*, z_*, \lambda_*, \gamma_*)$*. Moreover, when LICQ holds at* x_**, we can choose* γ_* *such that* $\lim_{k\in K_0} \gamma_{k+1} = \gamma_*$*.*

Proof First, from Theorem 4, we obtain the feasibility condition (3.14a). From (3.10), (3.11) and the property of Moreau-Yosida regularizations, 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

$$\operatorname{grad} L_k(x_k) = \operatorname{grad} f(x_k) + \sum_{i=1}^m [\lambda_{k+1}]_i \operatorname{grad}[h_1(x_k)]_i + \sum_{i=1}^q [\gamma_{k+1}]_i \operatorname{grad}[h_2(x_k)]_i. \tag{3.15}$$

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

$$0 \in \partial\psi(y_k) - (\sigma_k(h_1(x_k) - y_k) + \lambda_k) = \partial\psi(y_k) - \lambda_{k+1}. \tag{3.16}$$

Note that $\lim_{k\in K} y_k = y_*$, then $\{y_k\}_{k\in K}$ is bounded. By the locally boundedness of the subgradient [48, Corollary 24.5.1], $\bigcup_{k\in K} \partial\psi(y_k)$ is also bounded. Then, $\{\lambda_{k+1}\}_{k\in K}$ is bounded, so we can choose $K_1 \subseteq K$, $\lambda_* \in \mathbb{R}^m$ such that $\lim_{k\in K_1} \lambda_{k+1} = \lambda_*$, which implies $\lambda_* \in \partial\psi(y_*)$.

On the other hand, since

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

we know from the optimal condition of the above problem that

$$0 = [\sigma_k(h_2(x_k) - z_k) + \gamma_k]^\top z_k = \gamma_{k+1}^\top z_k \quad \text{and} \quad \gamma_{k+1} \ge 0. \tag{3.17}$$

Let $\mathcal{A}_k = \{i \in [q] : [z_k]_i = 0\}$, from (3.17), 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 know that for sufficiently large k and $i \notin \mathcal{A}_*$, $[\gamma_{k+1}]_i = 0$. Then, when k is large enough (3.15) can be written as

$$\operatorname{grad} L_k(x_k) = \operatorname{grad} f(x_k) + \sum_{i=1}^m [\lambda_{k+1}]_i \operatorname{grad}[h_1(x_k)]_i + \sum_{i\in\mathcal{A}} [\gamma_{k+1}]_i \operatorname{grad}[h_2(x_k)]_i. \tag{3.18}$$

Using the Carathédory's theorem of cones [8], there exist $J_k \subseteq \mathcal{A}_*$ and $[\hat\gamma_k]_j \ge 0$, where $j \in J_k$, such that $\{\operatorname{grad}[h_2(x_k)]_j : j \in J_k\}$ are linearly independent and

$$\operatorname{grad} L_k(x_k) = \operatorname{grad} f(x_k) + \sum_{i=1}^m [\lambda_{k+1}]_i \operatorname{grad}[h_1(x_k)]_i + \sum_{j\in J_k} [\hat\gamma_k]_j \operatorname{grad}[h_2(x_k)]_j, \tag{3.19}$$

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 $\lim_{k\in K_3}[\hat\gamma_k]_i = [\gamma_*]_i$ for $i \in J_*$. Using the fact that $\|\operatorname{grad} L_k(x_k)\| \to 0$, (3.15) implies (3.14d). Since $J_* \subseteq \mathcal{A}_*$ and $[\gamma_*]_i = 0$ for $i \notin J_*$, then (3.14b) 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 $\lim_{k\in k_4}[\hat\gamma_k]_i/M_k = [\hat\gamma]_i$ for $i \in J_*$ and $[\hat\gamma]_i = 0$ otherwise. By the definition of M_k, we have $\hat\gamma \ne 0$ and $\|\hat\gamma\|_\infty = 1$. Dividing (3.19)

by M_k, using the boundedness of $\operatorname{grad} L_k(x_k)$, $\operatorname{grad} f(x_k)$ and $\operatorname{grad}[h_1(x_k)]_i$, letting $k \in K_4 \to \infty$ and noticing that $\lambda_{k+1} \to \lambda_*$, $x_k \to x_*$, we can get

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

Note that $\|\hat{\gamma}\|_\infty = 1$ and $\hat{\gamma} \geq 0$, we know $\{\operatorname{grad}[h_2(x_*)]_j : j \in J_*\}$ 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 LICQ holds at x_* but $\{\gamma_{k+1}\}$ is unbounded, we can divide (3.18) 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 Semismoth Newton Method

Recall that at every step we have to solve

$$\min_{x \in \mathcal{M}} \; L_k(x) = f(x) + \psi^{\sigma_k}\big(h_1(x) + \frac{\lambda_k}{\sigma_k}\big) + \delta^{\sigma_k}_{\mathsf{R}^q}\big(h_2(x) + \frac{\gamma_k}{\sigma_k}\big). \tag{4.1}$$

It is known that L_k is continuously differentiable and its gradient is

$$\operatorname{grad} L_k = \operatorname{grad} f + \operatorname{grad}\Big(\psi^{\sigma_k}\big(h_1 + \frac{\lambda_k}{\sigma_k}\big)\Big) + \operatorname{grad}\Big(\delta^{\sigma_k}_{\mathsf{R}^q}\big(h_2 + \frac{\gamma_k}{\sigma_k}\big)\Big), \tag{4.2}$$

where last two terms in (4.2) 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 $\varphi : \mathcal{M} \mapsto \mathbb{R}$ is continuously differentiable. We make the following assumption.

Assumption 6 *The vector field $X := \operatorname{grad}\varphi$ 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 definition of the semismoothness in (2.9) to a general set-valued map.

Definition 14 Let X be a vector field on $\mathcal{M}$ and $\mathcal{K} : \mathcal{M} \quad \mathcal{L}(T\mathcal{M})$ be a upper-semicontinuous[5] 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 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 $H_q \in \mathcal{K}(q)$,

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

Remark 2 Definition 11 and Definition 14 coincide when $\mathcal{K} = \partial X$ [22, 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 2 *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 $\mathcal{K}_X + \mathcal{K}_Y$.*

[5] We say the map K is upper-semicontinuous if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all q 2 $B_\delta(p)$ we have $P_{qp}K(q)$ $K(p) + \hat{B}_\varepsilon(0)$, where $\hat{B}_\varepsilon(0) := fV$ 2 $L(T_p M) : kVk < \varepsilon g$.

Proof By Definition 14, 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

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

which is consequence of the linearity of P_{qp} and triangular inequality of the norm.

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

Algorithm 7 *Choose $p_0 \in \mathcal{M}$, $\bar{\nu} \in (0, 1]$ and let $\{\eta_k\}$ be a sequence converging to 0. Set $\mu \in (0, 1/2), \delta \in (0, 1), m_{\max} \in \mathbb{N}$ and $\varepsilon_0 \in (0, 1)$.*

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

(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

$$\|(H_k + \omega_k I)V_k + X(p_k)\| \le \tilde{\eta}_k, \tag{4.5}$$

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

$$\langle -X(p_k), V_k\rangle \ge \varepsilon_0 \|V_k\|^2, \tag{4.6}$$

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

Next, find the minimum non-negative integer $m_k \le m_{\max}$ such that

$$\varphi(\exp_{p_k}(\delta^{m_k} V_k)) \le \varphi(p_k) + \mu\delta^{m_k}\langle X(p_k), V_k\rangle. \tag{4.7}$$

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

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

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

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

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

We make several remarks for the above numerical algorithm.

Remark 3 LS-I is a standard way to globalize the semismooth Newton method for minimizing 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 7 reduces to the first-order method. In Theorem 10, we show that the LS-I globalization method has a superlinear convergence under a "convexity assumption" and some regularity conditions.

Remark 4 LS-II is another way to globalize the Newton method [42, 25, 49]. In the Riemannian setting, the convergence result is established under the assumption that $\|X\|^2$ is differentiable [23]. 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 5 We can use the method in [18] to choose an initial point of Algorithm 7 to guarantee the condition (3.9). For LS-I, since it is a descent method, it suffices to choose p_0 such that $L_k(p_0) \le \Phi$. This can be done by

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

where x_{feas} and x_{k-1} is defined in Algorithm 3. 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.

The next theorem establishes the global convergence of Algorithm 7 with LS-I.

Theorem 8 *Let $\{p_k\}$ be the sequence generated by Algorithm 7 with LS-I. Suppose there exists $\delta > 0$ such that $\Omega := \{p \in \mathcal{M} : \varphi(p) \leq \varphi(p_0) + \delta\}$ is compact. Then, the following properties hold:*

- *There exists a constant $m_{\max} \in \mathbb{N}$ such that the line search condition (4.7) holds in $m_{\max}$ steps.*
- *The sequence $\{\varphi(p_k)\}_{k=1}^{\infty}$ is strictly decreasing, i.e. there exists $c > 0$ such that*

$$\varphi(p_k) - \varphi(p_{k+1}) \geq c\|V_k\|^2. \tag{4.8}$$

- *The sequence $\{V_k\}$ and $\{X(p_k)\}$ satisfy $\sum_{k=1}^{\infty} \|V_k\|^2 < \infty$ and $\lim_{k\to\infty} X(p_k) = 0$.*
- *Let $\mathcal{P}$ be the set of accumulation points of $\{p_k\}$, then $\mathcal{P}$ is a non-empty set and for any $p_* \in \mathcal{P}$, we have $X(p_*) = 0$.*

Proof Since X is locally Lipschitz, for every $p \in \Omega$ we can find $R_p, L_p > 0$ such that X is L_p-Lipschitz in $B_{2R_p}(p)$. Since Ω is compact, we can find a finite set $\{q_1, \ldots, q_T\} \in \Omega$ such that $\bigcup_j B_{R_{q_j}}(q_j) \supset \Omega$. Define $R := \min_j R_{q_j}$, $L := \max_j L_{q_j}$. Since there exists $B_{R_{q_j}}(q_j) \ni p$ for every $p \in \Omega$, we find $B_R(p) \subset B_{2R_{q_j}}(q_j)$. Thus, X is L-Lipschitz in $B_R(p)$ for every $p \in \Omega$. Fix $p \in \Omega^\circ$ 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 $0 \leq s, t < R/\|V\|$

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

where the last inequality follows from the Lipschitzness of X and $\ell(\gamma|_{[s,t]}) = |t - s|\|V\|$ and $\|\dot{\gamma}(t)\| = \|\dot{\gamma}(0)\| = \|V\|$. Set $p = p_k$, $V = V_k$ to be the quantities defined in Algorithm 7. From the sufficient descent condition (4.6) in LS-I, we know when $t < R/\|V_k\|$

$$\varphi(p_k) - \varphi(\exp_{p_k}(tV_k)) \geq \langle -X(p_k), tV_k\rangle - \frac{Lt^2}{2}\|V_k\|^2 \begin{cases} \geq t\left(1 - \frac{Lt}{2\varepsilon_0}\right)\langle -X(p_k), V_k\rangle, & \text{if } V_k \text{ is from LS-I,} \\ = t\left(1 - \frac{Lt}{2}\right)\langle -X(p_k), V_k\rangle, & \text{if } V_k = -X(p_k). \end{cases} \tag{4.9}$$

Choosing $m_{\max} = \inf\{m \in \mathbb{N} \mid \delta^m \leq \min\{2\epsilon_0(1-\mu)/L, R/\sup_k \|V_k\|\}\}$. From (4.6), we know $\sup_k \|V_k\| \leq \sup_k \|X(p_k)\|/\varepsilon_0 < \infty$. Thus, $m_{\max} < \infty$ and

$$\delta^{m_{\max}}\left(1 - \frac{L\delta^{m_{\max}}}{2}\right) \geq \delta^{m_{\max}}\left(1 - \frac{L\delta^{m_{\max}}}{2\varepsilon_0}\right) \geq \mu\delta^{m_{\max}}. \tag{4.10}$$

Combining (4.9) with (4.10), the line search condition (4.7) holds at most in $m_{\max}$ iterations. Setting $c = \mu\delta^{m_{\max}}\varepsilon_0 > 0$, the conditions (4.7) and (4.21) imply

$$\varphi(p_k) - \varphi(p_{k+1}) \geq \mu\delta^{m_k}\langle -X(p_k), V_k\rangle \geq c\|V_k\|^2, \tag{4.11}$$

which is (4.8). As φ is continuous and Ω is compact, φ is bounded below. Together with the monotonicity of $\varphi_(p_k)$, there exists some φ_* such that $\varphi(p_k) \to \varphi_*$ as $k \to \infty$. Moreover, the telescoping sum of (4.11) and $k \to \infty$ implies that

$$c\sum_{k=1}^{\infty}\|V_k\|^2 \leq \varphi(p_0) - \varphi_* < \infty \Rightarrow \|V_k\| \to 0, k \to \infty.$$

By the upper-semicontinuity of $\mathcal{K}$ and the compactness of Ω, there exist $r > 0$ and $q_1, \ldots, q_N \in \Omega$ such that $\bigcup_{i=1}^N B_r(q_i) \supset \Omega$ and $P_{qq_j}\mathcal{K}(q) \subset \mathcal{K}(q_j) + B_\varepsilon(0)$ for every $q \in B_r(q_j)$. Since the parallel transport is an isometry, then $\{H_k\}$ is bounded. Together with the boundedness o f $X(p_k)$ and (4.5), it has

$$\|X(p_k)\| \leq \tilde{\eta}_k + \|H_k + \omega_k I\|\|V_k\| \leq \eta_k + \|H_k\|\|V_k\| + \|X(p_k)\|^{\bar{\nu}}\|V_k\| \to 0, k \to \infty.$$

The compactness of Ω and the continuity of X can easily derive that $\mathcal{P} \neq \emptyset$ and $X(p_*) = 0$ for all $p_* \in \mathcal{P}$.

4.1 The Analysis of Convergence Rate

To obtain the convergence rate of Algorithm 7, we first introduce some basic results of Riemannian manifolds and useful lemmas. Given $p \in \mathcal{M}$ and $r > 0$, define $B_r(p)$ to be the open ball on $\mathcal{M}$ with center p and radius r.

Lemma 1 (Lemma 2.3 and 2.4 in [21]) *Fix $p \in \mathcal{M}$, the following properties hold.*

- *There exists $r > 0$ such that for every $q \in B_r(p)$, the exponential map $\exp_q$ is a di eomorphism from $\{v \in T_q\mathcal{M} : \|v\| < 2r\}$ to $B_{2r}(q)$.*
- *There exist $K > 0$ and $r > 0$ such that for every $v, w \in T_q\mathcal{M}$ with $\|v\|, \|w\| < 2r$, it has $|d(\exp_q v, \exp_q w)^2 - \|v - w\|^2| \le K\|v\|^2\|w\|^2$.*

Lemma 2 *Suppose X is a locally Lipschitz vector eld on $\mathcal{M}$, $p_* \in \mathcal{M}$, $\mathcal{K}$ is de ned in De nition 14 and all elements in $\mathcal{K}(p_*)$ are nonsingular and there exists $\lambda > 0$ such that $\lambda \ge \max\{\|H^{-1}\| : H \in \mathcal{K}(p_*)\}$. 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*

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

Proof The proof is the same as the proof in [22, 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, which is a generalization of Theorem 2.3 in [29].

Lemma 3 *Suppose φ is a continuously di erentiable function with Lipschitz gradient, $p, q \in \mathcal{M}$. Let $\gamma : [0,1] \to \mathcal{M}$ be a geodesic joining p, q. Then, there exist $\xi \in (0,1), M_\xi \in \partial \operatorname{grad}\varphi(\gamma(\xi))$ 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 $\hat\varphi(t) = \varphi \circ \gamma$. We know that $\hat\varphi$ is continuously differentiable with Lipschitz gradient in $[0,1]$. By Theorem 2.3 in [29], 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 $\hat\varphi'(t) = \langle \operatorname{grad}\varphi(\gamma(t)), \dot\gamma(t)\rangle$ 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 where $\hat M_\xi \in \partial_B^2\hat\varphi(\xi)$, then 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 3 and Lemma 2 provides the following theorem that characterizes the second-order optimality conditions for (4.3).

Theorem 9 *Suppose $\varphi : \mathcal{M} \to \mathbb{R}$ is continuously di erentiable with locally Lipschitz gradient, then we have*

- *(Second-order necessary condition). Suppose $x_* \in \mathcal{M}$ is a local minima of φ, then for any $v \in T_x\ \mathcal{M}$, there is $H_v \in \partial \operatorname{grad}\varphi(x_*)$ such that $\langle H_v v, v\rangle \ge 0$.*
- *(Second-order su cient condition). Suppose $x_* \in \mathcal{M}$ such that $\operatorname{grad}\varphi(x_*) = 0$ and all elements in $\partial \operatorname{grad}\varphi(x_*)$ are positive de nite, then x_* is a strict local minima of φ.*

Remark 6 Using Lemma 3, we can see that the second-order sufficient condition implies the strongly geodesic convexity of φ near p_* as defined in [58].

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

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 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 $\bar{v}_k := \frac{rv_k}{\|v_k\|} \to v \in T_p\mathcal{M}$ with $\|v\| = r$. Define $t_k = \|v_k\|/r$, $q_k = \exp_p v_k$, $r_k = \exp_p \bar{v}_k$ and $s_k = \exp_p t_k v$, it has $\{q_k, r_k, s_k\} \subset B_{r_0}(p)$. Thus, we have

$$\begin{aligned}\|P_{q_k,p} X(q_k) - \nabla X(p; v_k)\| &\leq \|P_{q_k,p} X(q_k) - P_{s_k,p} X(s_k)\| + \|P_{s_k,p} X(s_k) - \nabla X(p; v_k)\| \\ &\leq \underbrace{\|P_{p,s_k} P_{q_k,p} X(q_k) - P_{q_k,s_k} X(q_k))\|}_{\text{(I)}} + \underbrace{\|P_{q_k,s_k} X(q_k) - X(s_k)\|}_{\text{(II)}} \\ &\quad + \underbrace{\|P_{s_k,p} X(s_k) - t_k \nabla X(p; v)\|}_{\text{(III)}} + \underbrace{\|\nabla X(p; v_k) - t_k \nabla X(p; v)\|}_{\text{(IV)}}.\end{aligned} \tag{4.12}$$

From Lemma 10 in [33], there exists $C > 0$ such that

$$\text{(I)} \leq \|P_{s_k,q_k} P_{p,s_k} P_{q_k,p} - \mathrm{id}\| \|X(q_k)\| \leq C \max(d(p, q_k), d(q_k, s_k)) \|X(q_k)\| = o(t_k), \tag{4.13}$$

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 1 we have

$$\text{(II)} \leq L d(q_k, s_k) = o(t_k), \quad \text{(III)} = \|P_{s_k,p} X(s_k) - X(p) - t_k \nabla X(p; v)\| = o(t_k). \tag{4.14}$$

Let $q_k^t = \exp_p t\bar{v}_k$ and $q_v^t = \exp_p tv$, it has

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

from Lemma 10 in [33] and Lemma 1. An intermediate result of the above estimation is

$$\|\nabla X(p; \bar{v}_k) - \nabla X(p; v)\| \leq \|\bar{v}_k - \bar{v}\| \to 0, \text{ if } k \to \infty. \tag{4.15}$$

Thus, the last term in (4.12) is

$$\text{(IV)} = t_k \|\nabla X(p; \bar{v}_k) - \nabla X(p; v)\| = o(t_k) \tag{4.16}$$

Combining (4.13), (4.14) and (4.16), it has $\|P_{q_k,p} X(q_k) - \nabla X(p; v_k)\| = o(t_k) = o(\|v_k\|)$, which contradicts with our assumption.

Lemma 5 *Suppose* $\{p_k\} \subset \mathcal{M}$ *and* $V_k \in T_{p_k}\mathcal{M}$ *satisfy* $\lim_{k\to\infty} p_k = p \in \mathcal{M}$ *and* $d(\exp_{p_k} V_k, p) = o(d(p_k, p))$. *Then,* $\lim_{k\to\infty} \|V_k\|/d(p_k, p) = 1$ *and* $d(\exp_{p_k} V_k, p) = o(\|V_k\|)$.

Proof Define $q_{k+1} := \exp_{p_k} V_k$. Let $B_r(p), K$ be the quantities in Lemma 1. Without the loss of generality, we may assume $p_k \in B_r(p)$ for every k. Using Lemma 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

$$\limsup_{k\to\infty} \frac{d(p_k, q_{k+1})^2}{d(p_k, p)^2} \leq K \limsup_{k\to\infty} d(q_{k+1}, p)^2 + \limsup_{k\to\infty} \left(1 + \frac{d(q_{k+1}, p)}{d(p_k, p)}\right)^2 = 1. \tag{4.17}$$

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 has $d(q_{k+1}, p)^2 \leq \varepsilon d(p_k, p)^2 \leq 2\varepsilon d(q_{k+1}, p)^2 + 2\varepsilon d(p_k, q_{k+1})^2 + \varepsilon K d(q_{k+1}, p)^2 d(p_k, q_{k+1})^2$. Thus, we know

$$d(p_{k+1}, p)^2 \leq \frac{2\varepsilon}{1 - 2\varepsilon - \varepsilon K d(p_k, p_{k+1})^2} d(p_k, p_{k+1})^2 = o\left(d(p_k, p_{k+1})^2\right).$$

On the other hand,

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

Combining with (4.17), we have $\lim_{k\to\infty}\frac{d(p_k,p)}{d(p_k,q_{k+1})}=1$.

Theorem 10 *Under the same assumptions as in Theorem 8, let $\mathcal{K}$ be the set-valued map used in Algorithm 7. 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 de nite, then we have $p_k\to p_*$ as $k\to\infty$ and for su ciently large k, it has*

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

where $\bar\nu\in(0,1]$ is the parameter de ned in Algorithm 7.

Proof Since $\mathcal{K}(p_*)\cup\partial X(p_*)$ is positive definite, by Lemma 2 and the upper-semicontinuity of $\mathcal{K}$ and ∂X, we can find a neighborhood $U\ni p_*$ and constants $\omega,M>0$ such that $M\|V\|^2\ge\langle HV,V\rangle\ge2\omega\|V\|^2$ for every $p\in U,H\in\mathcal{K}(p)\cup\partial X(p),V\in T_p\mathcal{M}$. By Lemma 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

$$\big\|X(q)+H_q\exp_q^{-1}p_*\big\|\le d(p_*,q)^{1+\nu},\ \forall q\in B_{r_0}(p_*),\tag{4.18}$$

and X is L-Lipshitz in $B_{r_0}(p_*)$.

Given arbitrary $0<r<r_0$, by Theorem 8, there exists some $K_0>0$ such that $\omega_k=\|X(p_k)\|^{\bar\nu}<1/2$, $\|V_k\|<r/2$ and $(2M+1)^{1+\bar\nu}\|V_k\|^{\bar\nu}\le\omega$ for $k\ge K_0$. From Lemma 3, we know $\varphi(q)-\varphi(p_*)\ge\omega d(q,p_*)^2$ whenever $q\in B_r(p_*)$. Since p_* is an accumulation point, we can find $K_1\ge K_0>0$ such that $\varphi(p_{K_1})-\varphi(p_*)<\omega r^2/4$ and $p_{K_1}\in B_r(p_*)$. Note that $d(p_{K_1},p_{K_1+1})=\delta^{m_{K_1}}\|V_{K_1}\|<r/2$. Then, $d(p_{K_1+1},p_*)\le d(p_{K_1},p_*)+d(p_{K_1},p_{K_1+1})<r$, 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 $\{p_k\}_{k\ge K_1}\subset B_r(p_*)$ which implies that $p_k\to p_*$ as $k\to\infty$. Below we assume that $k\ge K_1$.

By the positive definiteness of $\mathcal{K}$, the CG method in Algorithm 7 is able to find a direction V_k satisfying (4.5). Thus, we know $\|X(p_k)\|\le\|H_k+\omega_kI\|\,\|V_k\|/(1-\|X(p_k)\|^{\bar\nu})\le(2M+1)\,\|V_k\|$. Then, we obtain

$$\begin{aligned}\langle -X(p_k),V_k\rangle&=\langle(H_k+\omega_kI)V_k,V_k\rangle-\langle(H_k+\omega_kI)V_k+X(p_k),V_k\rangle\\&\ge2\omega\,\|V_k\|^2-\tilde\eta_k\,\|V_k\|\ge2\omega\,\|V_k\|^2-(2M+1)^{1+\bar\nu}\,\|V_k\|^{2+\bar\nu}\ge\omega\|V_k\|^2.\end{aligned}\tag{4.19}$$

Thus, the condition (4.6) holds. Note that $\|(H_k+\omega_k)^{-1}\|\le(2\omega)^{-1}$, and $\|X(p_k)\|\le Ld(p_k,p_*)$. 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

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

Using (4.19) we find $\|V_k\|\le\|X_k\|/\omega\le Ld(p_k,p_*)/\omega$. From Lemma 1, we know

$$d(\exp_{p_k}V_k,p_*)^2\le\|V_k-\exp_{p_k}^{-1}p_*\|^2+K\|V_k\|^2d(p_k,p_*)^2\le O\big(d(p_k,p_*)^{2+2\min\{\nu,\bar\nu\}}\big).\tag{4.20}$$

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

$$\varphi(\exp_{p_k}V_k)\le\varphi(p_k)+\mu\,\langle X(p_k),V_k\rangle\,,\tag{4.21}$$

holds with $m_k=0$. Define $U_k:=\exp_p^{-1}p_k$, $W_k:=\exp_p^{-1}\exp_{p_k}V_k$. Applying Lemma 3 to p_*, p_k and $\exp_{p_k}V_k$, we have

$$\varphi(p_k)=\varphi(p_*)+\frac{1}{2}\langle\tilde R_kU_k,U_k\rangle,\ \text{and}\ \varphi(\exp_{p_k}V_k)=\varphi(p_*)+\frac{1}{2}\langle\tilde M_kW_k,W_k\rangle,$$

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 $\xi \in (0,1)$. Substracting these two equations and using Lemma 5, we know

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

Thus, we have

$$\begin{aligned}\varphi(\exp_{p_k} V_k) - \varphi(p_k) - \frac{1}{2}\langle X(p_k), V_k\rangle &= -\frac{1}{2}\langle \tilde{R}_k U_k, U_k\rangle - \frac{1}{2}\langle X(p_k), V_k\rangle + o\left(\|V_k\|^2\right)\\ &= -\frac{1}{2}\langle V_k, X(p_k) - P_{p\ p_k}\tilde{R}_k U_k\rangle - \frac{1}{2}\langle P_{p_k p}\ V_k + U_k, \tilde{R}_k U_k\rangle + o\left(\|V_k\|^2\right).\end{aligned} \tag{4.22}$$

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, 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

$$P_{p_k p}\ X(p_k) = \tilde{R}_k U_k + o(\|U_k\|). \tag{4.23}$$

Moreover, applying $\lim_{k\to\infty} p_k = p_*$ and (4.20) to Lemma 5, we have

$$\frac{\|V_k\|}{\|U_k\|} = \frac{\|V_k\|}{d(p_k, p_*)} \to 1 \text{ and } \|W_k\| = d(\exp_{p_k} V_k, p_*) = o(\|V_k\|) \text{ as } k \to \infty. \tag{4.24}$$

Then, (4.23), (4.24) and Lemma 1 imply

$$\begin{aligned}&|\langle V_k, X(p_k) - P_{p\ p_k}\tilde{R}_k U_k\rangle| \le \|V_k\|\,\|X(p_k) - P_{p\ p_k}\tilde{R}_k U_k\| \le o(\theta_k \|V_k\|\,\|U_k\|) = o\big(\theta_k \|V_k\|^2\big),\\ &|\langle P_{p_k p}\ V_k + U_k, \tilde{R}_k U_k\rangle| \le \|V_k - \exp_{p_k}^{-1} p_*\|\|\tilde{R}_k U_k\| \le O\left(\|W_k\|\|U_k\|\right) = o\left(\|V_k\|^2\right).\end{aligned} \tag{4.25}$$

Combing (4.6), (4.22) and (4.25), we have

$$\varphi(\exp_{p_k} V_k) - \varphi(p_k) = \frac{1}{2}\langle X(p_k), V_k\rangle + o\left(\|V_k\|^2\right) \le \mu\langle X(p_k), V_k\rangle - \varepsilon_0\left(\frac{1}{2} - \mu\right)\|V_k\|^2 + o\left(\|V_k\|^2\right).$$

Thus, (4.21) holds for sufficiently large k and $m_k = 0$ in LS-I.

4.2 Calculations of Clarke Generalized Covariant Derivatives

In this section, we discuss the approach that calculates the Clarke generalized covariant derivative of a locally Lipschitz vector field X, which is difficult in general manifolds. Here, we consider an embedded manifold $\mathcal{M}$ with the following assumption:

Assumption 11 *The manifold $\mathcal{M}$ is embedded in $\mathbb{R}^d$ and there exist an open set $U \subset \mathbb{R}^d$ containing $\mathcal{M}$ and a vector-valued map $\bar{X} : U \subset \mathbb{R}^d \to \mathbb{R}^d$ such that $\bar{X}|_{\mathcal{M}} = X$.*

If X is differentiable at $p \in \mathcal{M}$, it is known from [31] that

$$\nabla X(p)[V] = \mathbf{P}_p(\nabla \bar{X}(p)[V]), \tag{4.26}$$

where $\mathbf{P}_p$ is the projection onto $T_p\mathcal{M}$. If X is not differentiable at $p \in \mathcal{M}$, by a direct calculation from Definition 10, it is easy to see that

$$\partial X(p) \subseteq \mathbf{P}_p(\partial \bar{X}(p))^6.$$

The following example shows that the above inclusion can be strict, and thus we need further discussion for finding an element in $\partial X(p)$.

[6] ∂X is the Clarke generalized covariant derivative on $\mathcal{M}$ or the Clarke generalized Jocobian in $\mathbb{R}^d$. The exact meaning depends on the domain of X.

Example 1 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 $Y(x, y) = \left(2|x - 1/\sqrt{2}|,\ 4|y - 1/\sqrt{2}|\right)^\top$. 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 (4.26) that

$$\nabla X(p) = (I - pp^\top)\nabla\bar{X}(p) = (I - pp^\top)(\nabla Y(p) - p^\top Y(p)I). \tag{4.27}$$

Moreover, by a direct calculation, it is known that

$$\mathbf{P}_{p_0}(\partial\bar{X}(p_0)) = (I - p_0p_0^\top)\partial\bar{X}(p_0) = \mathrm{co}\left\{\pm\begin{pmatrix}1 & -2\\ -1 & 2\end{pmatrix},\ \pm\begin{pmatrix}1 & 2\\ -1 & -2\end{pmatrix}\right\}, \tag{4.28}$$

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, 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 (4.27). The second term in (4.27) converges to 0 since $Y(p_0) = 0$. Observe that $\nabla Y(p_k) = \mathrm{diag}(2, -4)$ when $\theta_k > \pi/4$ and $\nabla Y(p_k) = \mathrm{diag}(-2, 4)$ 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 $[-1, 1] \subset [-3, 3]$. Thus, $\partial X(p_0) \subset \mathbf{P}_p(\partial\bar{X}(p))$. 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 12 *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*

- *F is continuously differentiable;*
- *$f_j : \mathbb{R} \to \mathbb{R}^{n_j}$, $j = 1, \ldots, d$ 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 $(q - \delta, q + \delta)\backslash\{q\}$;*
- *The left and right derivatives of f_j exist.*

Based on the above assumptions, the next lemma finds an element in ∂X by choosing a proper path that converges to the non-differentiable point.

Lemma 6 *Suppose Assumption 11 and Assumption 12 hold. For any $q \in \mathcal{M}$, let $\{q^{(n)}\} \subset \mathcal{M}$ be a sequence that converges to q. If the sequence $\{q^{(n)}\}$ satisfies that for each $j \in [d]$, it has either $q_j^{(n)} > q_j\ \forall n \in \mathbb{N}$ or $q_j^{(n)} < q_j\ \forall n \in \mathbb{N}$, 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, $r_j : T_q\mathcal{M} \to \mathbb{R}^{n_j}$ is a linear operator such that*

$$r_j(v) = \begin{cases} v_j \lim\limits_{t\downarrow q_j} f_j'(t), & \text{if } q_j^{(n)} > q_j, \\ v_j \lim\limits_{t\uparrow q_j} f_j'(t), & \text{if } q_j^{(n)} < q_j. \end{cases} \tag{4.29}$$

Proof This is a direct consequence of Definition 10 and (4.26).

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.

Assumption 13 *Let $Q \in \mathrm{St}(n, r)$, $Z \in \mathbb{R}^{n\times r}$ and $V = \mathbf{P}_Q Z \in T_Q\mathrm{St}(n, r)$. For all i, j, it has $Q_{ij} \in \{\pm 1\}$ if $V_{ij} = 0$.*

Remark 7 Let $\mathcal{I} = \{(i, j) : V_{ij} = 0 \text{ for all } V \in T_Q\mathrm{St}(n, r)\}$. It is easy to know that when $(i, j) \notin \mathcal{I}$ the set $\mathcal{Z}_{ij} := \{Z \in \mathbb{R}^{n\times r} : (\mathbf{P}_Q Z)_{ij} = 0\}$ has zero Lebesgue measure as it is a linear space and $\dim \mathcal{Z}_{ij} < nr$, which implies that the probability of $V_{ij} = 0$ is zero when Z is a random matrix in $\mathbb{R}^{n\times r}$ for $(i, j) \notin \mathcal{I}$.

Remark 8 When $Q_{ij} \in \{\pm 1\}$, by the direct computation, we know $(\mathbf{P}_Q Z)_{ij} = 0$ since $Q^\top Q = I_r$. Conversely, we prove that $Q_{ij} \in \{\pm 1\}$ for all $(i,j) \in \mathcal{I}$ where $\mathcal{I}$ is defined in Remark 7. If $(i_0, j_0) \in \mathcal{I}$, then $N \in (T_Q \mathrm{St}(n,r))^\perp$, where $N_{ij} = 1$ if $i = i_0, j = j_0$ and $N_{ij} = 0$ otherwise. From [2, Example 3.6.2], we know $(T_Q \mathrm{St}(n,r))^\perp = \{QS : S \in \mathbb{R}^{r\times r}, S^\top = S\}$. Without the loss of generality, we can assume $i = j = 1$ and there exists $S = S^\top$ such that $QS = N$. We can rewrite $QS = N, Q^\top Q = I_r$ 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 $a = c$. Similarly, from $x(az^\top + x^\top B) = 0$ and $A^\top(yz^\top + AB) = 0$, note that $xa + A^\top y = 0$ and $xx^\top + A^\top A = I_{r-1}$, we know $B = 0$. Note $yz^\top + AB = 0$ and $B = 0$, then either $y = 0$ or $z = 0$. When $y = 0$, then $1 = a^2 + y^\top y = a^2$. When $z = 0$, then $1 = ab + x^\top z = a^2$. Thus, $Q_{11} = a = \pm 1$.

Theorem 14 *Let $X(P) = F(P, f_{11}(P_{11}), f_{12}(P_{12}), \ldots, f_{nr}(P_{nr}))$ be a locally Lipschitz vector field on $\mathrm{St}(n,r)$ such that the Assumption 12 holds. Given $Q \in \mathrm{St}(n,r)$ and $V = \mathbf{P}_Q Z \in T_Q\mathrm{St}(n,r)$ and the Assumption 13 holds. Then we know $\mathbf{P}_Q \circ dF|_Q \circ (\mathrm{id}_{\mathbb{R}^{n\ r}}, H_{11}, \ldots, H_{nr}) \in \partial X(Q)$, where $H_{ij} : T_Q\mathrm{St}(n,r) \to \mathbb{R}^{n_{ij}}$ is a linear map such that $H_{ij}(W) = W_{ij}D_{ij}$ and*

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

Proof Consider a sequence $Q^{(n)} = \exp_Q(t_n V)$, where $t_n \downarrow 0$. If $V_{ij} > 0$, 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 $Q_{ij} = 1$, note that $Q^\top Q = I_r$, we know $Q^{(n)}_{ij} < 1$ for all n. Other cases are similar, and we can use Lemma 6 to find the derivative.

Thus, we derive the following algorithm:

Algorithm 15 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 14.

It is noted that the Algorithm 15 find an element of the Clarke generalized covariant derivative of vector fields on $\mathrm{St}(n,r)$ almost surely.

5 Numerical Experiments

In this section, we evaluate our algorithm on three problems mentioned before: compressed modes (CM) [44], sparse PCA and the constrained sparse PCA [41]. In CM and SPCA, we compare our algorithm with SOC [37], ManPG [16, ManPG-Ada (Algorithm 2)], accelerated ManPG (AManPG) [34], and accelerated Riemannian proximal gradient (ARPG) [35]. In the constrained SPCA, we compare our algorithm with ALSPCA [41]. Codes of SOC and ManPG are provided by [16], codes of AManPG and ARPG are provided by [35], and the code of ALSPCA is provided by [41]. All codes are implemented by 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 [16] to solve the Schrödinger equation of 1D free-electron model with periodic boundary condition:

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

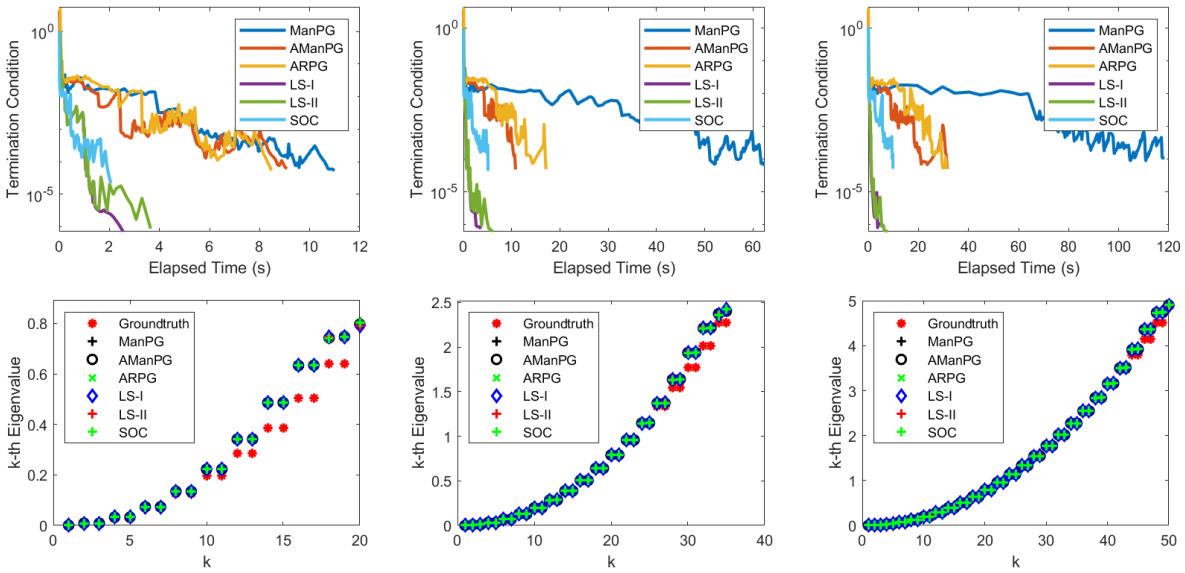

Fig. 1: Comparisons on CM with $(n, \mu) = (500, 0.05)$ and $r = 20, 35, 50$. The top row plots the termination condition, which is the sum of the left hand parts in (5.3) and (5.4) for SOC, and similar for other algorithms. The bottom row plots the eigenvalues of $Q^\top HQ$, where Q is the solution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

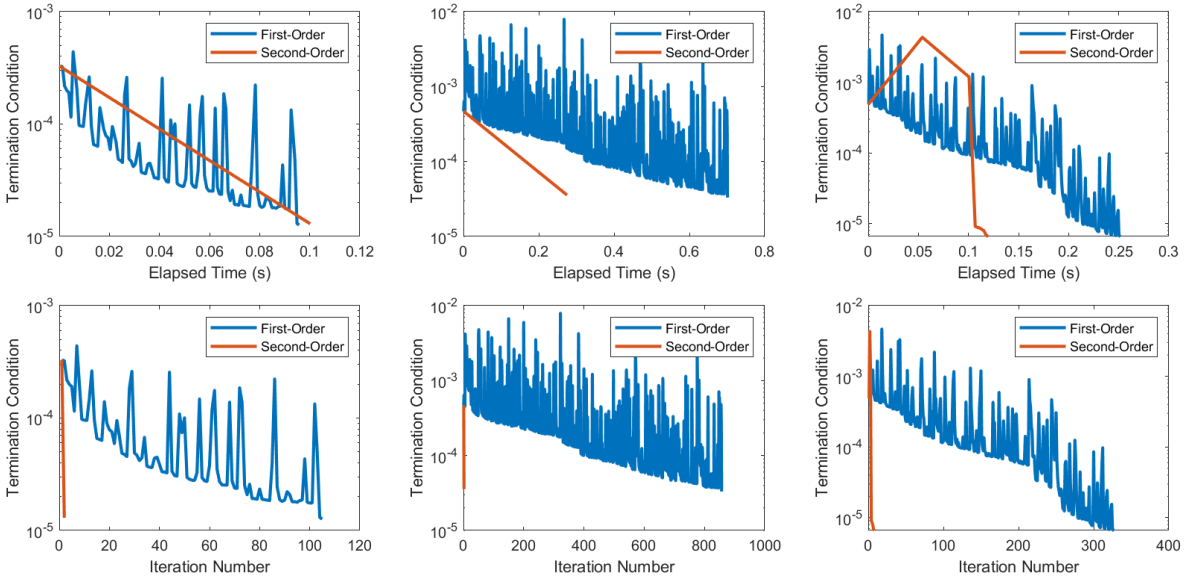

Fig. 2: Comparisons on the first-order method [53] and the second-order method in solving the CM subproblem (3.8). Columns represent the behaviour of these methods in solving different subproblems in Algorithm 3.

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:

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

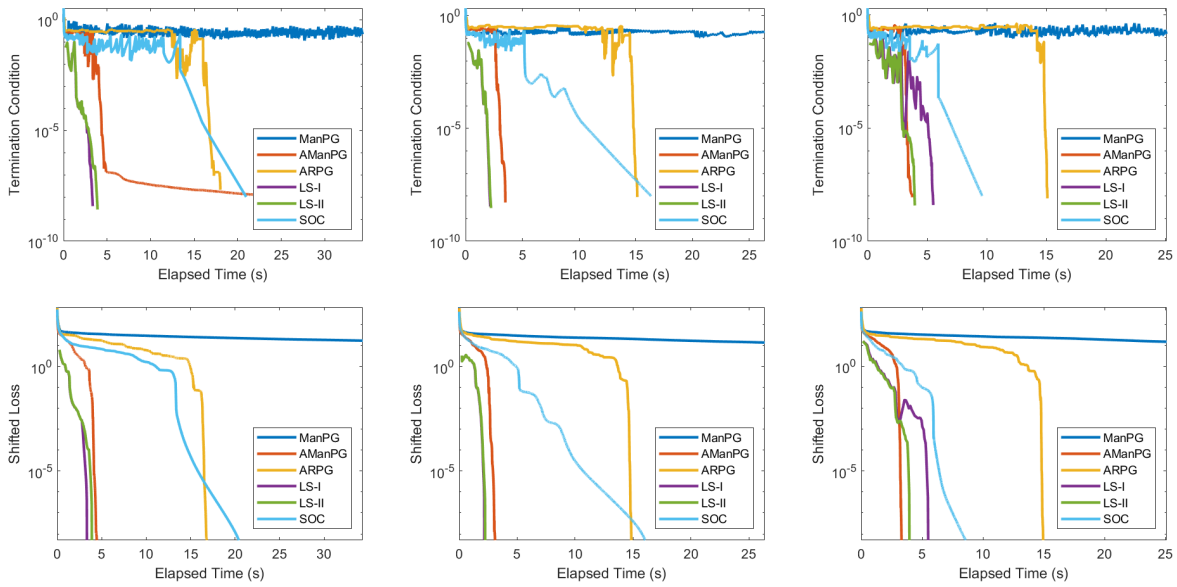

Fig. 3: Comparisons on SPCA with $(n, r, \mu) = (500, 20, 1.00)$ and three different $A^\top A$. The shifted loss is the loss substracted by the minimal loss among all iterations, i.e., $F(x_k) - \min_j F(x_j)$.

The solution of the above problem satisfies

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

As suggested by [16, 34], ManPG and AManPG use $t^{-1}\|V_*\|_\infty/(\|Q\|_F+1) \le 5\times 10^{-5}$ 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:

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

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

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

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

Similarly, our algorithm rewrites the problem into

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

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

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

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

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

$$\operatorname{grad} \hat{L}_k(Q) = -2HQ + (\sigma Q + \Lambda) - \sigma F, \text{ and } \operatorname{Hess} \hat{L}_k(Q)[Z] = -2HZ + \sigma Z \odot E,$$

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

$$\operatorname{grad} L_k(Q) = \mathbf{P}_Q \operatorname{grad} \hat{L}_k(Q), \text{ and } \operatorname{Hess} L_k(Q)[Z] = \mathbf{P}_Q(\operatorname{Hess} \hat{L}_k(Q)[Z] - Z \operatorname{sym}(Q^\top \operatorname{grad} \hat{L}_k(Q))),$$

where $\operatorname{sym} Q := (Q + Q^\top)/2$. It is easy to check that $\operatorname{grad} L_k$ fulfills Assumption 12 with $f_{ij}(Q_{ij}) = F_{ij}$. Thus, Algorithm 15 can be applied to find an element of $\partial \operatorname{grad} L_k$ when $\hat{L}_k$ is non-differentiable at Q.

The parameters of ManPG and SOC are the same as those in [16]. The codes and parameters of AManPG and ARPG are adopted from the SPCA implementation of Huang et al. [35]. All of the four methods are terminated when their termination conditions are satisfied or the number of iterations exceeds 30000. In our algorithm, we set $\tau = 0.97$, $\rho = 1.25$, $\alpha = 1.01$. We set $\varepsilon_k = 0.95^k$ and the initial value of σ is 1.0. To solve our subproblem, Therefore, we use the first-order method to find a good initial point, and start the second-order method when $\|\operatorname{grad} L_k\| < 5 \times 10^{-4}$, where L_k is defined in (3.8). The maximum number of iterations in CG and the first-order method is 1000. In Algorithm 7, the linesearch parameter is $\mu = 0.1$. We set $\lambda_k = 0.7^k$ and the minimal stepsize of the linesearch is 10^{-4}. When CG finds a negative direction p_k, we set $\lambda_k = -2\operatorname{Hess} L_k(Q_k)[p_k, p_k] / \|p_k\|^2$ and restart the iteration.

We report the results in Table. 2 and Fig. 1. 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. We note that when $r = 10$ and 15, LS-I looks extremely slow because the second-order method starts too early such that most evaluations of second-order directions are wasted. This could be solved by a careful tuning of parameters. From Fig. 1, we also find ManPG has difficulty attaining the desired termination condition when r is large. The performance of LS-I and LS-II are similar, except for the aforementioned cases in which the second-order method does not starts 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. We also try to run the first-order method [53] 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 Fig. 2 and reported in Table. 3. From Table. 3 and the bottom row of Fig. 2, we see that to attain 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. 3. 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 Fig. 2.

We also note that in the Euclidean setting the sparsity may be exploited to accelerate the conjugate gradient in the second-order method [40]. 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 7 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 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

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 [16].

		CPU (s)						Loss					
		ManPG	AManPG	ARPG	LS-I	LS-II	SOC	ManPG	AManPG	ARPG	LS-I	LS-II	SOC
n	200	11.54	3.86	6.43	**1.16**	1.60	1.78	14.10	14.10	14.10	14.10	14.10	14.10
	500	21.02	8.32	9.15	**4.00**	5.87	5.41	18.60	18.60	18.60	18.60	18.60	18.60
	1000	66.30	**8.60**	11.30	11.45	9.66	14.99	23.30	23.30	23.30	23.30	23.30	23.30
	1500	44.73	40.18	42.85	**23.24**	26.72	39.69	26.90	26.80	26.80	26.80	26.80	26.80
	2000	42.48	46.86	51.47	**27.59**	27.74	46.33	29.80	29.70	29.70	29.70	29.70	29.70
r	10	15.35	15.63	15.71	88.89	**11.67**	17.28	10.70	10.70	10.70	10.70	10.70	10.70
	15	35.35	**9.55**	11.17	37.23	10.09	15.47	16.50	16.40	16.40	16.40	16.40	16.40
	25	87.64	22.73	23.93	**17.19**	20.55	27.11	32.00	32.00	32.00	32.00	32.00	32.00
	30	83.13	27.41	29.35	**11.28**	14.84	18.10	42.90	42.90	42.90	42.90	42.90	42.90
μ	0.05	102.63	14.86	13.98	6.89	7.87	**6.34**	15.10	15.10	15.10	15.10	15.10	15.10
	0.15	65.08	17.09	21.48	**13.73**	15.11	32.51	31.00	31.00	31.00	31.00	31.00	31.00
	0.20	51.46	**12.96**	24.12	17.14	14.46	33.71	38.30	38.20	38.20	38.20	38.20	38.20
	0.25	42.74	**13.41**	25.43	53.06	20.10	34.80	45.30	45.20	45.20	45.20	45.20	45.30

Table 3: 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 and CPU time of the second-order method. The column First-Order is the total number of iterations and CPU time of the first-order method [53] 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)		Iteration		Time (s)	
		LS-I	LS-II	LS-I	LS-II	LS-I	LS-II	LS-I	LS-II
n	200	273	367	2.30	4.40	27233	24887	7.00	7.00
	500	286	362	4.60	9.80	33403	28714	15.00	13.00
	1000	124	81	2.90	2.00	1754	3932	1.00	2.00
	1500	232	358	20.20	61.90	66829	59074	68.00	60.00
	2000	130	206	12.90	32.80	23439	21273	31.00	27.00
r	10	551	300	82.70	57.40	56994	24357	27.00	11.00
	15	330	256	35.20	54.30	28721	18450	18.00	11.00
	25	227	342	12.40	27.70	28154	23752	26.00	21.00
	30	162	422	15.60	74.20	24970	19603	26.00	21.00
μ	0.05	165	285	5.10	13.20	9075	8209	6.00	5.00
	0.15	138	219	5.50	12.60	20983	19213	16.00	16.00
	0.20	101	145	5.20	12.10	12352	10388	9.00	7.00
	0.25	271	299	37.80	64.70	73177	35754	55.00	26.00

that since ManPG generally cannot achieve our requirement on the accuracy, we terminate it when the computational time exceeds 300 seconds.

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 $\{w_i\}$ are sampled from the standard Gaussian distribution. Finally, columns of A are normalized to the zero mean and the unit length. Results are reported in in Table. 4 and Fig. 3. We find that the losses of the solutions obtained by these methods are comparable, and our algorithm 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 [41]. In this section, we compare our algorithm with [41], which will be referred to as ALSPCA. In our implementation, we apply the first-order feasible method on Stiefel manifold

Table 4: Comparison on SPCA. $A \in \mathbb{R}^{50\times n}$, $(n, r, \mu) = (2000, 20, 1.0)$ and one of them varies.

		CPU (s)						Loss					
		ManPG	AManPG	ARPG	LS-I	LS-II	SOC	ManPG	AManPG	ARPG	LS-I	LS-II	SOC
n	500	130.04	6.77	48.26	5.25	**4.99**	18.86	-337.50	-337.70	-337.70	**-337.80**	**-337.80**	-337.30
	1000	216.74	12.05	79.78	12.46	**11.98**	34.94	-749.70	-749.70	**-749.90**	-749.70	-749.70	-749.60
	1500	286.69	34.72	87.92	25.50	**22.71**	57.74	-1206.80	-1207.60	**-1207.80**	-1207.60	-1207.60	-1206.90
	2000	284.57	34.64	112.64	40.48	**29.77**	94.48	-1621.10	-1621.50	-1621.80	**-1622.60**	**-1622.60**	-1621.30
	2500	295.38	57.50	116.51	68.77	**26.41**	124.66	-2073.30	-2077.00	**-2077.30**	-2077.10	-2077.10	-2076.20
	3000	297.92	81.80	134.70	78.58	**41.24**	158.95	-2535.30	-2542.00	-2541.80	**-2542.30**	**-2542.30**	-2541.10
r	5	46.15	**7.37**	25.00	14.96	15.79	55.96	-1497.50	-1497.50	-1497.50	**-1497.60**	**-1497.60**	-1497.50
	10	164.57	**18.52**	44.10	28.68	18.90	75.23	-1639.70	-1640.10	-1640.10	**-1640.50**	**-1640.50**	-1639.70
	15	247.58	38.15	67.22	43.32	**24.76**	86.39	-1641.00	-1641.50	**-1641.60**	-1640.80	-1640.80	-1640.90
	25	299.09	62.88	146.32	30.75	**25.99**	105.45	-1627.10	-1636.00	-1636.40	-1636.40	-1636.40	**-1636.50**
μ	0.25	252.48	**70.98**	80.56	80.91	83.91	93.19	-1871.50	-1874.10	-1874.10	**-1874.20**	**-1874.20**	-1871.40
	0.50	277.07	57.01	86.69	48.89	**36.06**	93.67	-1777.10	**-1780.80**	**-1780.80**	-1780.60	-1780.60	-1778.60
	0.75	280.89	51.82	98.02	44.07	**26.06**	94.44	-1695.10	-1698.00	-1697.70	**-1698.10**	**-1698.10**	-1695.80
	1.25	276.42	37.36	112.67	39.56	**23.36**	93.86	-1551.90	-1552.00	-1552.20	**-1553.30**	**-1553.30**	-1552.00

Table 5: 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	**7.58**	10.19	**37.75**	34.35	11.86	**11.98**	**-15.31**	-8.84	**-8.58**	-8.29
	10	**12.68**	19.61	**37.86**	34.24	22.80	**23.16**	**-15.30**	-8.82	**-8.12**	-7.99
	15	**17.58**	25.37	38.58	**39.36**	**32.94**	32.78	**-15.24**	-8.70	**-8.11**	-7.86
	20	**20.74**	27.57	**39.86**	37.63	42.59	**43.03**	**-15.09**	-9.17	-7.94	**-8.53**
	25	**27.66**	32.02	**39.98**	38.86	51.59	**52.05**	**-15.13**	-9.02	-8.04	**-8.13**
n	500	**8.99**	15.46	**72.45**	68.53	**35.73**	35.68	**-15.20**	-8.95	-8.28	**-8.38**
	1000	**13.30**	22.82	**56.15**	49.97	40.47	**41.67**	**-15.16**	-8.87	**-8.09**	-7.94
	2000	**20.74**	27.57	**39.86**	37.63	42.59	**43.03**	**-15.09**	-9.17	-7.94	**-8.53**
	4000	46.53	**38.75**	**23.84**	21.61	43.16	**44.08**	**-15.10**	-9.01	**-8.30**	-7.96
	6000	**59.52**	61.42	**18.33**	16.90	43.10	**43.76**	**-15.25**	-9.18	**-8.72**	-7.98

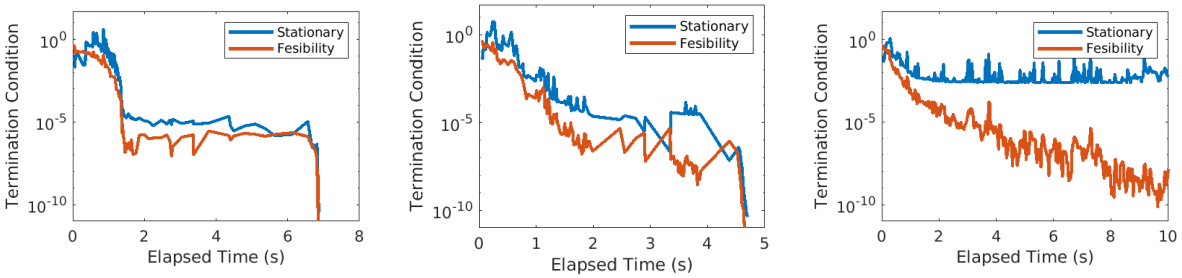

Fig. 4: Results of Algorithm 3 and 7 on constrained SPCA with $(n, \mu) = (600, 1.0)$ and $r = 3, 5, 10$, respectively. We use the first-order method to find an initial point for Algorithm 7 and start the second-order method when $\|\operatorname{grad} L_k\| \le 5\times10^{-4}$. "Feasibility" refers 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 $r = 10$.

[53] to solve the subproblem when the iterates do not meet the start conditions of the semismooth Newton method. Following [41], the termination condition of subproblem is based on the feasiblity 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 [41]. 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 the zero mean and the unit length.

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 [41]. It is defined as follows:

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

This quantity measures how well the data is explained by the found principal components. We set $\Delta_{ij} = 10^{-8}$ and report the results in Table. 5. We find solutions obtained from these algorithms are comparable in terms of CPAV. 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 Fig. 4, the semismooth Newton method can start when $r = 3$ and $r = 5$. When $r = 10$, the behaviour of our algorithm is similar to the first order methods as it is difficult to meet 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 10 is likely violated for the case when r is large. A similar phenomenon has also been observed in SDP [59, 55]. 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.

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. 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.
4. 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.
5. 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.
6. 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.

7. 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.
8. D. P. Bertsekas, Nonlinear programming, Journal of the Operational Research Society, 48 (1997), pp. 334–334.
9. D. P. Bertsekas, Constrained optimization and Lagrange multiplier methods, Academic press, 2014.
10. J. Bolte, S. Sabach, and M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Mathematical Programming, 146 (2014), pp. 459–494.
11. 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.
12. N. Boumal and P.-a. Absil, RTRMC: A Riemannian trust-region method for low-rank matrix completion, in Advances in neural information processing systems, 2011, pp. 406–414.
13. S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al., Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and Trends® in Machine learning, 3 (2011), pp. 1–122.
14. 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.
15. M. P. d. Carmo, Riemannian geometry, Birkhäuser, 1992.
16. S. Chen, S. Ma, A. M.-C. So, and T. Zhang, Proximal gradient method for nonsmooth optimization over the stiefel manifold, SIAM Journal on Optimization, 30 (2020), pp. 210–239.
17. 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.
18. 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.
19. M. Cho and J. Lee, Riemannian approach to batch normalization, in Advances in Neural Information Processing Systems, 2017, pp. 5225–5235.
20. 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.
21. 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.
22. F. R. de Oliveira, O. P. Ferreira, et al., Newton method for finding a singularity of a special class of locally lipschitz continuous vector fields on riemannian manifolds., J. Optim. Theory Appl., 185 (2020), pp. 522–539.
23. F. R. de Oliveira and F. R. Oliveira, A global version of the Newton method for finding a singularity of the nonsmooth vector fields on Riemannian manifolds, arXiv preprint arXiv:2006.01559, (2020).
24. K. Deng and Z. Peng, An inexact augmented lagrangian method for nonsmooth optimization on Riemannian manifold, arXiv preprint arXiv:1911.09900, (2019).
25. 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.
26. O. Ferreira and P. Oliveira, Proximal point algorithm on Riemannian manifolds, Optimization, 51 (2002), pp. 257–270.
27. D. Gabay, Minimizing a differentiable function over a differential manifold, Journal of Optimization Theory and Applications, 37 (1982), pp. 177–219.
28. E. Ghahraei, S. Hosseini, and M. R. Pouryayevali, Pseudo-Jacobian and characterization of monotone vector fields on Riemannian manifolds, J. Convex Anal, 24 (2017), pp. 149–168.
29. 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.
30. 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.
31. J. Hu, X. Liu, Z. Wen, and Y. Yuan, A brief introduction to manifold optimization, arXiv preprint arXiv:1906.05450, (2019).
32. 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.
33. W. Huang, P.-A. Absil, and K. A. Gallivan, A riemannian symmetric rank-one trust-region method, Mathematical Programming, 150 (2015), pp. 179–216.
34. W. Huang and K. Wei, Extending fista to riemannian optimization for sparse pca, arXiv preprint arXiv:1909.05485, (2019).
35. W. Huang and K. Wei, Riemannian proximal gradient methods, arXiv preprint arXiv:1909.06065, (2019).
36. 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.
37. R. Lai and S. Osher, A splitting method for orthogonality constrained problems, Journal of Scientific Computing, 58 (2014), pp. 431–449.
38. 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.
39. J. M. Lee, Introduction to smooth manifolds, Springer, 2012.
40. 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.
41. Z. Lu and Y. Zhang, An augmented lagrangian approach for sparse principal component analysis, Mathematical Programming, 135 (2012), pp. 149–193.
42. R. Mifflin, Semismooth and semiconvex functions in constrained optimization, SIAM Journal on Control and Optimization, 15 (1977), pp. 959–972.
43. 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.
44. 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.
45. L. Qi and J. Sun, A nonsmooth version of Newton's method, Mathematical programming, 58 (1993), pp. 353–367.
46. 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.
47. F. Rampazzo and H. J. Sussmann, Commutators of flow maps of nonsmooth vector fields, Journal of Differential Equations, 232 (2007), pp. 134–175.
48. R. T. Rockafellar, Convex analysis, 28, Princeton university press, 1970.
49. D. Sun and J. Sun, Semismooth matrix-valued functions, Mathematics of Operations Research, 27 (2002), pp. 150–169.
50. 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.
51. B. Vandereycken, Low-rank matrix completion by Riemannian optimization, SIAM Journal on Optimization, 23 (2013), pp. 1214–1236.
52. I. Waldspurger, A. d'Aspremont, and S. Mallat, Phase recovery, maxcut and complex semidefinite programming, Mathematical Programming, 149 (2015), pp. 47–81.
53. Z. Wen and W. Yin, A feasible method for optimization with orthogonality constraints, Mathematical Programming, 142 (2013), pp. 397–434.
54. 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.
55. 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.

56. 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.
57. 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.
58. H. Zhang and S. Sra, First-order methods for geodesically convex optimization, in Conference on Learning Theory, 2016, pp. 1617–1638.
59. 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.
60. 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.
61. H. Zou, T. Hastie, and R. Tibshirani, Sparse principal component analysis, Journal of computational and graphical statistics, 15 (2006), pp. 265–286.