MATHEMATICS OF OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0364-765X | EISSN 1526-5471 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000
© 0000 INFORMS

A Riemannian smoothing steepest descent method for non-Lipschitz optimization on embedded submanifolds of $\mathbb{R}^n$

Chao Zhang
School of Mathematics and Statistics, Beijing Jiaotong University, Beijing 100044, China. (zc.njtu@163.com)

Xiaojun Chen
Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China. (xiaojun.chen@polyu.edu.hk)

Shiqian Ma
Computational Applied Mathematics and Operations Research, Rice University. (sqma@rice.edu)

In this paper, we study the generalized subdifferentials and the Riemannian gradient sub-consistency that are basis for non-Lipschitz optimization on embedded submanifolds of $\mathbb{R}^n$. We then propose a Riemannian smoothing steepest descent method for non-Lipschitz optimization on complete embedded submanifolds of $\mathbb{R}^n$. We prove that any accumulation point of the sequence generated by the Riemannian smoothing steepest descent method is a stationary point associated with the smoothing function employed in the method, which is necessary for the local optimality of the original non-Lipschitz problem. We also prove that any accumulation point of the sequence generated by our method that satisfies the Riemannian gradient sub-consistency is a limiting stationary point of the original non-Lipschitz problem. Numerical experiments are conducted to demonstrate the advantages of Riemannian ℓ_p $(0<p<1)$ optimization over Riemannian ℓ_1 optimization for finding sparse solutions and the effectiveness of the proposed method.

Key words: Embedded submanifolds of $\mathbb{R}^n$; Non-Lipschitz; Generalized subdifferentials; Smoothing steepest descent method; Riemannian gradient sub-consistency
MSC2000 subject classification: 49M05; 90C26; 57-08

1. Introduction

We consider the Riemannian optimization problem

$$\min \ f(x), \quad x \in \mathcal{M}, \tag{1}$$

where $\mathcal{M}$ is a complete embedded submanifold of $\mathbb{R}^n$ and $f:\mathbb{R}^n \to \mathbb{R}$ is a proper lower semicontinuous function and possibly non-Lipschitz. It is worth mentioning that the results developed in this paper also work for matrix-variable problems, i.e., $f:\mathbb{R}^{m\times n} \to \mathbb{R}$. Such problems arise in a variety of applications in signal processing, computer vision, and data mining [3, 6, 35, 50].

Many classical algorithms for unconstrained and smooth optimization have been extended from Euclidean space to Riemannian manifolds, such as the gradient descent algorithm, the conjugate gradient algorithm, the quasi-Newton algorithm and the trust region method [1, 2, 11, 33]. Recently, Riemannian optimization with a nonsmooth but locally Lipschitz continuous objective function has been considered in the literature. Here the smoothness and locally Lipschitz continuity are interpreted when the function in question is considered in the ambient Euclidean space. The Riemannian Clarke subdifferential of functions over manifolds has been defined and its properties have been discussed in [31]. Several algorithms have been proposed based on the notion of Riemannian Clarke subdifferential. For example, Hosseini and Uschmajew [32] proposed the Riemannian gradient sampling algorithm. This algorithm approximates the subdifferential using the convex hull of transported gradients from tangent spaces of randomly generated nearby points to the tangent space of the current space. The ϵ-subgradient algorithm [29] is a steepest descent method where the descent directions are obtained by a computable approximation of the ϵ-subdifferential. The line search algorithms [30] include the nonsmooth Riemannian BFGS algorithm as a special case. For both the ϵ-subgradient algorithm and the line search algorithms, either the algorithms terminate after a finite number of iterations with the ϵ-subgradient-oriented descent direction being 0, or any accumulation point is a Riemannian Clarke stationary point. Other methods for nonsmooth optimization over Riemannian manifolds include the Riemannian subgradient method [41], the Riemannian ADMM [36, 37, 40], the manifold proximal gradient method [16, 17, 34, 52], manifold proximal point method [15], manifold proximal linear method [53], manifold augmented Lagrangian method [18, 60, 61] and zeroth-order algorithms over Riemannian manifold [39].

The Riemannian generalized subdifferentials have been studied in [4, 38] and are expected to be useful for analyzing non-Lipschitz optimization. To the best of our of knowledge, however, there do not exist optimization algorithms for solving Riemannain non-Lipschitz optimization problems with rigorous convergence results. Consequently, the Riemannian generalized subdifferentials developed in [4, 38] have not been used to show the convergence results for non-Lipschitz optimization yet. Non-Lipschitz optimization in Euclidean space finds many important applications, including but not limited to, finding sparse solutions in signal processing and data mining [21, 24, 42, 43, 48], and neat edge in image restoration [9, 22, 57]. Smoothing methods with a proper updating scheme for the smoothing parameter are efficient for solving large-scale nonsmooth optimization in Euclidean space [19, 22, 23, 25, 58, 59]. With a fixed smoothing parameter, one solves the smoothed problem to update the iterate. Certain strategy is then applied to decide whether and how the smoothing parameter needs to be changed. Under the so-called gradient consistency property, it can be shown that any accumulation point of the smoothing method is a limiting stationary point of the original nonsmooth optimization problem; see for example the definition in [59, pp. 14]. The gradient consistency naturally holds for smoothing functions arising in various real applications with nonsmooth and locally Lipschitz objective functions [12, 13, 19, 55, 58]. Smoothing methods have been widely used to solve unconstrained non-Lipschitz optimization problems [23], and constrained non-Lipschitz optimization with convex feasible sets [59]. However, minimizing a non-Lipschitz function on a nonconvex set has not been widely considered in the literature. In [21], an augmented Lagrangian method for non-Lipschitz nonconvex programming was proposed where the constraint set is nonconvex.

In [58], a smoothing projected gradient method for minimizing a nonsmooth but locally Lipschitz function on a convex feasible set in $\mathbb{R}^n$ was proposed (Algorithm 3.1 of [58]). In [59], a smoothing active set method for linearly constrained non-Lipschitz nonconvex optimization in $\mathbb{R}^n$ (Algorithm 3.1 of [59]) was proposed. As mentioned in Remark 3.2 of [59], a unified framework of smoothing methods can be obtained by slightly modifying Algorithm 3.1 of [59] with the same convergence result developed in [59], including the smoothing steepest descent method if the feasible set is $\mathbb{R}^n$. In this paper the objective function is not necessarily locally Lipschitz. The Riemannian smoothing steepest descent (RSSD) method as well as the convergence analysis that will be developed in this paper extend those from [59]. The RSSD method can be considered as an extension of the smoothing steepest descent method on $\mathbb{R}^n$ from [59] to embedded submanifolds of $\mathbb{R}^n$; see Remark 4 for details.

Main Contributions. Our contributions of this paper are as follows.

(i) We characterize the Riemannian generalized subdifferentials for proper lower semicontinuous functions. We define the notion of limiting stationary point of (1) whose objective function is allowed to be not locally Lipschitz. When the objective function of (1) is locally Lipschitz, a limiting stationary point is a Clarke stationary point, but a Clarke stationary point is not necessarily a limiting stationary point of (1). Compared with the results in [38], Proposition 2 in this paper has not been considered, and Example 2 of this paper has not been given there.

(ii) We define the Riemannian subdifferential of f associated with a smoothing function $\tilde{f}$. We define a stationary point x^* of (1) associated with $\tilde{f}$, and show that x^* being a stationary point of (1) associated with $\tilde{f}$ is a necessary optimality condition for x^* being a local minimizer of (1).

(iii) To build the relationship between the above two notions of stationary points of (1), associated with or without $\tilde{f}$, we define the Riemannian gradient sub-consistency of $\tilde{f}$ at x on $\mathcal{M}$. Under the Riemannian gradient sub-consistency of $\tilde{f}$, any stationary point of (1) associated with $\tilde{f}$ is a limiting stationary point of (1). These concepts and results in (ii) and (iii) are extensions of the corresponding counterparts from [59] for optimization in $\mathbb{R}^n$ to Riemannian optimization on $\mathcal{M}$. We show that the Riemannian gradient sub-consistency holds if the gradient sub-consistency of $\tilde{f}$ holds at x on $\mathbb{R}^n$, provided that f is locally Lipschitz near x on $\mathbb{R}^n$. We also show that for a class of non-Lipschitz functions on $\mathbb{R}^n$, the Riemannian gradient sub-consistency of their smoothing functions holds on $\mathcal{M}$. These two results have not been considered in the existing literature before.

(iv) We design a Riemannian smoothing steepest descent method (RSSD) for solving (1). It is an extension of the smoothing steepest descent method in $\mathbb{R}^n$ from [59] to embedded submanifolds of $\mathbb{R}^n$; see Remark 4 for details. The proposed RSSD method is easy to implement and converges to a stationary point x^* of (1) associated with $\tilde{f}$ where the objective function is nonsmooth, possibly not even locally Lipschitz. Under Riemannian gradient sub-consistency of $\tilde{f}$, x^* is also a limiting stationary point of (1).

(v) When the objective function is locally Lipschitz, the convergence result of our RSSD method is stronger than that of the aforementioned existing methods for Riemannian nonsmooth optimization with locally Lipschitz objective functions. This is because these existing methods can only guarantee that any accumulation point of the sequence is a Clarke stationary point, but our result guarantees that any accumulation point of the sequence is a limiting stationary point.

The rest of this paper is organized as follows. In Section 2, we give a brief review on some basic concepts and properties related to Riemannian manifolds, the generalized subdifferentials and smoothing functions. We define the generalized subdifferentials for non-Lipschitz functions on embedded submanifolds of $\mathbb{R}^n$ that are motivated by [4, 38]. In Section 3, we discuss the properties of the generalized subdifferentials for non-Lipschitz functions on embedded submanifolds of $\mathbb{R}^n$. We also define and discuss the Riemannian gradient sub-consistency that is essential to the convergence analysis of our method. In Section 4, we propose our RSSD method and analyze its convergence behavior. In Section 5, we conduct numerical experiments on two important applications: finding the sparsest vectors in a subspace, and the sparsely-used orthogonal complete dictionary learning. Finally, we draw some concluding remarks in Section 6.

2. Preliminaries

We define some notation first. Throughout this paper, without specification, $\mathcal{M}$ denotes a complete embedded submanifold of $\mathbb{R}^n$. Let $x \in \mathcal{M}$ and $\mathrm{T}_x\mathcal{M}$ be the tangent space of $\mathcal{M}$ at x. The cotangent space at x via the Riemannian metric is denoted as $\mathrm{T}^*_x\mathcal{M}$. We use $\mathrm{T}\mathcal{M}$ to denote the tangent bundle, i.e., the disjoint union of the tangent spaces of $\mathcal{M}$: $\mathrm{T}\mathcal{M} := \{(x,v) \mid x \in \mathcal{M} \text{ and } v \in \mathrm{T}_x\mathcal{M}\}$. We consider the Riemannian metric on $\mathcal{M}$ that is induced from the Euclidean inner product; i.e., for any $\xi, \eta \in \mathrm{T}_x\mathcal{M}$, we have $\langle \xi, \eta \rangle_x = \xi^\top \eta$ if ξ and η are two column vectors of the same dimension, and $\langle \xi, \eta \rangle_x = \mathrm{Tr}(\xi^\top \eta)$ if ξ and η are two matrices of the same dimension, where $\mathrm{Tr}(Z)$ denotes the trace of matrix Z. We use $\|x\|$ to denote the Euclidean norm when x is a vector, and the Frobenius norm when x is a matrix. We use $B_{x,\delta} = \{y \mid \|y - x\| \le \delta\}$ to represent a neighborhood of x with radius $\delta > 0$. For a subset $D \subseteq \mathbb{R}^n$ with nonempty interior, a function $h \in C^1(D)$ means that h is smooth, i.e., continuously differentiable on D. For each $x \in \mathcal{M}$, the Riemannian metric induces an isomorphism between $\mathrm{T}_x\mathcal{M}$ and $\mathrm{T}^*_x\mathcal{M}$ through the mapping $\mathrm{T}_x\mathcal{M} \ni v \mapsto v^* = \langle v, \cdot\rangle_x \in \mathrm{T}_x^*\mathcal{M}$. We define the norm on $\mathrm{T}_x^*\mathcal{M}$ by $\|v^*\|_x^2 = \|v\|_x^2 = \langle v, v\rangle_x$. The subscript x in $\langle\cdot,\cdot\rangle_x$ and $\|\cdot\|_x$ may be omitted when there is no ambiguity.

We now give the definition of the retraction operation.

Definition 1. (Retraction, see [2]). A retraction on a manifold $\mathcal{M}$ is a smooth mapping $R : \mathrm{T}\mathcal{M} \to \mathcal{M}$ with the following two properties, where R_x denotes the restriction of R to the tangent space $\mathrm{T}_x\mathcal{M}$.
(i) $R_x(0_x) = x$, where 0_x denotes the zero element of $\mathrm{T}_x\mathcal{M}$.
(ii) It holds that

$$dR_x(0_x) = id_{\mathrm{T}_x\mathcal{M}},$$

where dR_x is the differential of R_x, and $id_{\mathrm{T}_x\mathcal{M}}$ denotes the identity map on $\mathrm{T}_x\mathcal{M}$.

By the inverse function theorem, we know that R_x is a local diffeomorphism (see, e.g., [30]).

Locally Lipschitz functions on $\mathcal{M}$. We adopt the definition of locally Lipschitz functions on $\mathcal{M}$ in [32]. Let $r : [0,1] \to \mathcal{M}$ be a C^1 curve. The length of r is defined as $l(r) = \int_0^1 \|r'(s)\|ds$. Let $x, y \in \mathcal{M}$. Denote the collection of C^1 curves joining x and y by $C(x,y)$. Then the Riemannian distance between x and y is defined by $\mathrm{dist}(x,y) := \inf\{l(r) : r \in C(x,y)\}$.

Let $\mathcal{M}$ be an embedded submanifold of $\mathbb{R}^n$ with the Riemannian distance, and U be an open subset of $\mathcal{M}$. According to [32], $f : \mathcal{M} \to \mathbb{R}$ is said to satisfy a Lipschitz condition of constant J on U if for any $x, y \in U$ it holds that

$$|f(x) - f(y)| \le J\mathrm{dist}(x,y).$$

A function f is said to be Lipschitz near $x \in \mathcal{M}$ if it satisfies the Lipschitz condition of some constant on an open neighborhood of x. A function f is said to be locally Lipschitz on $\mathcal{M}$ if f is Lipschitz near x for every $x \in \mathcal{M}$.

Generalized subdifferentials on $\mathbb{R}^n$.

In the case that f is nonsmooth but locally Lipschitz continuous near x, the Clarke subdifferential $\partial^\circ f(x)$ of f at $x \in \mathbb{R}^n$ is often used. Let

$$\Omega_f := \{x \in \mathbb{R}^n \mid f \text{ is differentiable at } x\}.$$

According to [26, Theorem 2.5.1, pp. 63], for nonsmooth but locally Lipschitz continuous function f, we have,

$$\partial^\circ f(x) = \mathrm{co}\left\{\lim_{\nu\to\infty} \nabla f(x_\nu) \mid x_\nu \to x,\ x_\nu \in \Omega_f\right\}, \tag{2}$$

where "co" denotes the convex hull.

We now review some important concepts and properties related to generalized subdifferentials of non-Lipschitz functions in Euclidean space $\mathbb{R}^n$ that are often used in nonsmooth analysis [10, 47].

Definition 2. (Subdifferentials). We consider a lower semicontinuous function $f : \mathbb{R}^n \to \mathbb{R}$.
(i) The regular (or Fréchet) subdifferential of f at $x \in \mathbb{R}^n$ is defined as

$$\hat{\partial} f(x) := \{v \mid f(y) \ge f(x) + \langle v, y - x\rangle + o(\|y - x\|)\}.$$

(ii) The limiting subdifferential of f at x is defined as

$$\partial f(x) := \{\lim_{\nu\to\infty} v_\nu \mid \exists\ (x_\nu, f(x_\nu)) \to (x, f(x)), v_\nu \in \hat{\partial} f(x_\nu)\}.$$

For a lower semicontinuous function $f:\mathbb{R}^n\to\mathbb{R}$, according to [47, Definition 8.3, pp. 301], the horizontal subdifferential of f at x is defined as

$$\partial^\infty f(x) := \{\lim_{\nu\to\infty} t_\nu v_\nu \mid \exists\ (x_\nu, f(x_\nu))\to(x,f(x)), t_\nu \downarrow 0,\ v_\nu\in\hat{\partial} f(x_\nu)\}, \tag{3}$$

and according to [10, Definition 1], the Clarke subdifferential of a non-Lipschitz function f at x is defined as

$$\partial^\circ f(x) := \bar{\text{co}}\{\partial f(x)+\partial^\infty f(x)\}, \tag{4}$$

where "$\bar{\text{co}}$" denotes the closure of convex hull.

It is known that

$$\hat{\partial} f(\bar{x}) \subseteq \partial f(\bar{x}) \subseteq \partial^\circ f(\bar{x}). \tag{5}$$

We have the equivalent characterization for the regular subdifferential in the following lemma from [47, Proposition 8.5, pp. 302].

Lemma 1. *A vector $v\in\mathbb{R}^n$ belongs to $\hat{\partial} f(\bar{x})$ if and only if in some neighborhood of $\bar{x}$, there is a function $h\le f$ with $h(\bar{x})=f(\bar{x})$ such that h is differentiable at $\bar{x}$ with $\nabla h(\bar{x})=v$. Moreover, h can be smooth with $h(x)<f(x)$ for all $x\neq\bar{x}$ near $\bar{x}$.*

Generalized subdifferentials on $\mathcal{M}$. Let $h\in C^1(\mathcal{M})$. According to [11, Definition 3.34, pp. 35], the differential of h at x, $dh(x)\in \mathrm{T}_x^*\mathcal{M}$, is a linear operator defined by

$$dh(x)[v] = \frac{d}{dt}h(c(t))\bigg|_{t=0}, \tag{6}$$

where c is a smooth curve on $\mathcal{M}$ passing through x at $t=0$ with velocity v. By [11, Definition 3.58, pp. 42], the Riemannian gradient of h is the vector field $\operatorname{grad} h$ on $\mathcal{M}$ uniquely defined by these identities:

$$\forall (x,v)\in \mathrm{T}\mathcal{M}, \qquad dh(x)[v] = \langle v, \operatorname{grad} h(x)\rangle. \tag{7}$$

In the case that $f:\mathcal{M}\to\mathbb{R}$ is a nonsmooth but locally Lipschitz continuous function, the Riemannian Clarke subdifferential has been studied and used in analyzing the convergence of algorithms [29, 30, 31, 32, 56]. Let

$$\Omega_{f,\mathcal{R}} := \{x\in\mathcal{M} \mid f \text{ is differentiable at } x\}.$$

The Riemannian Clarke subdifferential, denoted as $\partial_{\mathcal{R}}^\circ f(x)$, is defined as [32]

$$\partial_{\mathcal{R}}^\circ f(x) := \text{co}\left\{\lim_{\nu\to\infty} \operatorname{grad} f(x_\nu) \mid x_\nu\to x,\ x_\nu\in\Omega_{f,\mathcal{R}}\right\}. \tag{8}$$

Recall that $\lim_{\nu\to\infty}\operatorname{grad} f(x_\nu)$ in (8) can be explained as follows (see [32]). Let $\{(x_\nu,\xi_\nu)\}\subseteq \mathrm{T}\mathcal{M}$ where $\xi_\nu\in \mathrm{T}_{x_\nu}\mathcal{M}$. We say ξ_ν converges to ξ, denoted as $\lim_{\nu\to\infty}\xi_\nu=\xi$, if $x_\nu\to x$ and if for any smooth vector field ζ on $\mathcal{M}$ it holds that $\langle\xi_\nu,\zeta(x_\nu)\rangle_{x_\nu}\to\langle\xi,\zeta(x)\rangle_x$. An equivalent definition of $\partial_{\mathcal{R}}^\circ f(x)$ [32] relying on the definition of Clarke subdifferential on linear spaces is

$$\partial_{\mathcal{R}}^\circ f(x) = \partial^\circ(f\circ R_x)(0_x) \tag{9}$$

for any retraction R.

The Riemannian regular (or Fréchet) subdifferential for lower semicontinuous functions on Riemannian manifolds has been developed in [4]. Later on, the Riemannian regular, limiting and horizontal subdifferentials have been well studied in [38]. The Riemannian generalized subdifferentials [4, 38] can be considered as natural extensions of the generalized subdifferentials of lower semicontinuous functions on $\mathbb{R}^n$. We will use the following definition for Riemannian subdifferentials throughout the paper.

Definition 3. (Riemannian subdifferentials) Let $f:\mathbb{R}^n\to\mathbb{R}$ be a lower semicontinuous function.
(i) The Riemannian regular (or Fréchet) subdifferential of f at $x\in\mathcal{M}$ is defined as

$$\begin{aligned}\hat{\partial}_{\mathcal{R}} f(x) := \{\operatorname{grad} h(x) \mid\ \exists\ \delta=\delta(h)>0 \text{ such that } h\in C^1(B_{x,\delta})\\ \text{and } f-h \text{ attains a local minimum at } x \text{ on } \mathcal{M}\}.\end{aligned} \tag{10}$$

(ii) The Riemannian limiting subdifferential of f at $x\in\mathcal{M}$ is defined as

$$\partial_{\mathcal{R}} f(x) := \{\lim_{\nu\to\infty} v_\nu \mid \exists\ (x_\nu, f(x_\nu))\to(x,f(x)), v_\nu\in\hat{\partial}_{\mathcal{R}} f(x_\nu)\}. \tag{11}$$

Remark 1. The Riemannian regular (or Fréchet) subdifferential of f at $x\in\mathcal{M}$ in Definition 3.1 of [38] is defined as

$$\partial_F f(x) := \{dh(x) \mid h\in C^1(\mathcal{M}) \text{ and } f-h \text{ attains a local minimum at } x\}. \tag{12}$$

The Riemannian regular subdifferential $\hat{\partial}_{\mathcal{R}} f(x)$ in this paper is essentially only related to the local property of h. By Whitney extension theorem [54], any smooth function on $B_{x,\delta}\cap\mathcal{M}$ can be extended to the whole Euclidean space $\mathbb{R}^n$. Therefore,

$$\begin{aligned}\hat{\partial}_{\mathcal{R}} f(x) &= \{\operatorname{grad} h(x) \mid\ h\in C^1(\mathcal{M}) \text{ and } f-h \text{ attains local minimum at } x\}\\ &= \{\operatorname{grad} h(x) \mid\ dh(x)\in\partial_F f(x)\}.\end{aligned} \tag{13}$$

Hence $\partial_F f(x)$ in (12) and $\hat{\partial}_{\mathcal{R}} f(x)$ in (10) are essentially the same, through the one-to-one correspondence between $\operatorname{grad} h(x)$ in tangent space and $dh(x)$ in cotangent space.

In the next section, we will show that a vector in the Riemannian regular subdifferential $\hat{\partial}_{\mathcal{R}} f(x)$ can be computed via the projection of an arbitrary vector of the regular subdifferential $\hat{\partial} f(x)$ onto $\mathrm{T}_x\mathcal{M}$, if $\mathcal{M}$ is a Riemannian submanifold. We thus prefer to express the condition for h in (10), since such $h\in C^1(B_{x,\delta})$ is also suitable for defining $\hat{\partial} f(x)$. When $\mathcal{M}=\mathbb{R}^n$ and $f:\mathcal{M}\to\mathbb{R}$ is a nonsmooth but locally Lipschitz continuous function, the Riemannian Clarke subdifferential coincides with the Clarke subdifferential in $\mathbb{R}^n$. When $\mathcal{M}=\mathbb{R}^n$, the Riemannian regular and limiting subdifferentials coincide with the usual regular and limiting subdifferentials in $\mathbb{R}^n$.

In this paper, we consider Riemannian optimization with non-Lipschitz objective function f. We will explain later that f may not be locally Lipschitz at some points on $\mathcal{M}$. For this purpose, we give the characterizations of locally Lipschitz functions on $\mathcal{M}$ that are easily checkable in [38], which need the concept of convexity on $\mathcal{M}$. According to Definition 2.5 of [4], a subset U of $\mathcal{M}$ is convex if for any given two points $x,y\in U$, there exists a unique geodesic in U joining x and y such that the length of the geodesic is $\operatorname{dist}(x,y)$. According to Theorem 2.6 in [4], we know that for every $x\in\mathcal{M}$ there exists an open convex set U of $\mathcal{M}$ such that $x\in U$. Then according to Theorem 5.3 in [38] and the relations in (7) and (13), we have the following characterizations for a function on $\mathcal{M}$ to be locally Lipschitz.

Lemma 2. *Let $\mathcal{M}$ be an embedded submanifold of $\mathbb{R}^n$ with the Riemannian distance. Let $f:\mathbb{R}^n\to\mathbb{R}$ be a lower semicontinuous function. Then the following statements are equivalent:*
(i) f is locally Lipschitz near x on $\mathcal{M}$;
(ii) $\hat{\partial}_{\mathcal{R}} f$ is bounded in a neighborhood of x on $\mathcal{M}$.

Smoothing function. We use the following definition of a smoothing function on $\mathbb{R}^n$ as in [59].

Definition 4. (Smoothing function). A function $\tilde{f}(\cdot,\cdot):\mathbb{R}^n\times\mathbb{R}_+\to\mathbb{R}$ is called a smoothing function of $f:\mathbb{R}^n\to\mathbb{R}$, if $\tilde{f}(\cdot,\mu)$ is continuously differentiable in $\mathbb{R}^n$ for any $\mu\in\mathbb{R}_{++}$,

$$\lim_{z\to x,\ \mu\downarrow 0} \tilde{f}(z,\mu) = f(x), \tag{14}$$

and there exist a constant $\kappa>0$ and a function $\omega:\mathbb{R}_{++}\to\mathbb{R}_{++}$ such that

$$|\tilde{f}(x,\mu)-f(x)| \le \kappa\omega(\mu) \quad \text{with} \quad \lim_{\mu\downarrow 0}\omega(\mu)=0. \tag{15}$$

In order to emphasize that μ is a smoothing parameter, we sometimes also write $\tilde f(\cdot,\mu)$ as $\tilde f_\mu(\cdot)$ in this paper.

Example 1. We use the absolute value function $|t|, t\in\mathbb{R}$ as an example to illustrate the smoothing function. We can use the so-called uniform smoothing function

$$s_\mu(t)=\begin{cases}|t|, & \text{if } |t|\ge\frac{\mu}{2}\\ \frac{t^2}{\mu}+\frac{\mu}{4}, & \text{if } |t|<\frac{\mu}{2},\end{cases} \tag{16}$$

with $\kappa=\frac14$ and $\omega(\mu)=\mu$ in (15).

We refer to [19] for more examples of smoothing functions. For the non-Lipschitz term $|t|^p$ where $0<p<1$, its smoothing function can be defined as $(s_\mu(t))^p$, with $\kappa=(\frac14)^p$ and $\omega(\mu)=\mu^p$ in (15).

3. Riemannian generalized subdifferentials and Riemannian gradient sub-consistency

In this section, we first discuss properties of several generalized subdifferentials. We then define and discuss properties of Riemannian gradient sub-consistency of proper lower semicontinuous functions, and related stationary points of (1). These concepts and properties play important roles in the convergence analysis of our RSSD method in the next section. They also provide some basics for minimizing a non-Lipschitz function on an embedded submanifold of $\mathbb{R}^n$.

3.1. Riemannian generalized subdifferentials

Proposition 1. *Let $\mathcal{M}$ be an embedded submanifold of $\mathbb{R}^n$, $x\in\mathcal{M}$, and $f:\mathbb{R}^n\to\mathbb{R}$ be a lower semicontinuous function. Suppose $R:\mathrm{T}\mathcal{M}\to\mathcal{M}$ is a retraction defined in Definition 1. Then*
(i) $\hat\partial_{\mathcal{R}}f(x)=\hat\partial(f\circ R_x)(0_x)$ and $\partial_{\mathcal{R}}f(x)=\partial(f\circ R_x)(0_x)$.
(ii) $v\in\hat\partial_{\mathcal{R}}f(x)$ if and only if $v\in\mathrm{T}_x\mathcal{M}$ and the following holds

$$f\circ R_x(\eta_x)\ge f\circ R_x(0_x)+\langle v,\eta_x\rangle+o(\|\eta_x\|),\quad \forall\eta_x\in\mathrm{T}_x\mathcal{M}. \tag{17}$$

Proof. Statement (i) for $\hat\partial_{\mathcal{R}}f(x)$ holds, according to Theorem 4.3 of [4], Corollary 4.2 of [38], Definition 1 for retraction $R:\mathrm{T}\mathcal{M}\to\mathcal{M}$, and Remark 1 that $\hat\partial_{\mathcal{R}}f(x)$ in Definition 3 of this paper and $\partial_F(x)$ given in Definition 3.1 of [38] are essentially the same. The equivalent characterization of $\hat\partial_{\mathcal{R}}f(x)$ in statement (ii) can be easily obtained from (i). □

Now we give the following proposition about Riemannian regular subdifferential that is useful for computation and theoretical analysis; see the employment of Proposition 2 in equation (37) of Example 4, in equation (39) of Remark 3, as well as in equations (34) and (52) in the proofs of Theorem 2 and Theorem 3, respectively.

Proposition 2. *Let $\mathcal{M}$ be an embedded submanifold of $\mathbb{R}^n$, $x\in\mathcal{M}$, and $f:\mathbb{R}^n\to\mathbb{R}$ be a lower semicontinuous function. Then*

$$\left\{\mathrm{Proj}_{\mathrm{T}_x\mathcal{M}}v\mid v\in\hat\partial f(x)\right\}\subseteq\hat\partial_{\mathcal{R}}f(x). \tag{18}$$

Proof. Using (10) in Definition 3, and the facts that $\mathcal{M}$ is a submanifold embedded in $\mathbb{R}^n$ and $h\in C^1(B_{x,\delta})$, we have

$$\operatorname{grad}h(x)=\mathrm{Proj}_{\mathrm{T}_x\mathcal{M}}\nabla h(x),$$

where $\mathrm{Proj}_{\mathrm{T}_x\mathcal{M}}y$ denotes the projection of $y\in\mathbb{R}^n$ onto $\mathrm{T}_x\mathcal{M}$. Consequently,

$$\begin{aligned}\hat\partial_{\mathcal{R}}f(x)\;=\;&\{\mathrm{Proj}_{\mathrm{T}_x\mathcal{M}}\nabla h(x)\mid\ \exists\,\delta>0\text{ such that }h\in C^1(B_{x,\delta})\text{ and}\\&\quad f-h\text{ attains a local minimum at }x\text{ on }\mathcal{M}\}.\end{aligned} \tag{19}$$

Note that for any $v\in\hat\partial f(x)$, according to Lemma 1, there exists $h\in C^1$, such that $f-h$ attains a local minimum at x on $\mathbb{R}^n$, which is sure to attain a local minimum at x on $\mathcal{M}\subseteq\mathbb{R}^n$. This, combining with (19), indicates that (18) holds. □

Definition 5. A point $x \in \mathcal{M}$ is called a limiting stationary point of (1), if $0 \in \partial_{\mathcal{R}} f(x)$.

According to Proposition 1 (ii), we know that if $\bar{x}$ is a local minimizer of f on $\mathcal{M}$, then $0 \in \hat{\partial}_{\mathcal{R}} f(\bar{x})$. By Definition 3, we have $\hat{\partial}_{\mathcal{R}} f(\bar{x}) \subseteq \partial_{\mathcal{R}} f(\bar{x})$. Hence $\bar{x}$ being a limiting stationary point of (1) is a necessary condition of f achieving a local minimum at $\bar{x}$ on $\mathcal{M}$.

For a locally Lipschitz function f, $x \in \mathcal{M}$ is a Clarke stationary point of (1) if $0 \in \partial^\circ_{\mathcal{R}} f(x)$. The Clarke stationary point of (1) is widely used in the nonsmooth but locally Lipschitz Riemannian optimization literature [29, 30, 31, 32]. We show in the next proposition that a limiting stationary point is a Clarke stationary point.

Proposition 3. *Let $\mathcal{M}$ be an embedded submanifold of $\mathbb{R}^n$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function near $x \in \mathcal{M}$. Then $\partial_{\mathcal{R}} f(x) \subseteq \partial^\circ_{\mathcal{R}} f(x)$.*

Proof. The inclusion holds because

$$\partial^\circ_{\mathcal{R}} f(x) = \partial^\circ (f \circ R_x)(0_x) \supseteq \partial (f \circ R_x)(0_x) = \partial_{\mathcal{R}} f(x).$$

The first equality is due to (9), which transforms the Riemannian Clarke subdifferential of f at x to be the Clarke subdifferential of $f \circ R_x$ at 0_x on the Euclidean space $\mathrm{T}_x\mathcal{M}$. The inclusion comes from (5). The last equality is obtained from Proposition 1 (i). □

We use the following example to show that for f being a locally Lipschitz function on $\mathbb{R}^n$ and $\mathcal{M}$ being an embedded submanifold of $\mathbb{R}^n$, a Clarke stationary point is not necessarily a limiting stationary point.

Example 2. Let us consider the Riemannian optimization problem

$$\min f(x_1, x_2) := \frac{1}{2} x_1^2 - x_1 - |x_2|, \quad x \in \mathcal{M} \tag{20}$$

where $\mathcal{M} = S^1 := \{x \in \mathbb{R}^2 \mid x^\top x = 1\}$ is the unit circle, and f is locally Lipschitz in $\mathbb{R}^2$. Let $\bar{x} = (1, 0)^\top$, and $\bar{x}_\epsilon = (\sqrt{1-\epsilon^2}, \epsilon)^\top$. It is clear that $\bar{x}_\epsilon \to \bar{x}$ when $\epsilon \to 0$, $\|\bar{x}\| = \|\bar{x}_\epsilon\| = 1$, and for any $\epsilon \in (0,1)$,

$$\begin{aligned} f(\bar{x}_\epsilon) &= \frac{1}{2}(1-\epsilon^2) - \sqrt{1-\epsilon^2} - \epsilon \\ &< \frac{1}{2} - (\sqrt{1-\epsilon^2} + \epsilon) \\ &< \frac{1}{2} - 1 = -\frac{1}{2} = f(\bar{x}). \end{aligned}$$

Hence $\bar{x}$ is not a local minimizer of f on S^1.

For $\mathcal{M} = S^1$, we know from [2] that

$$\mathrm{Proj}_{\mathrm{T}_x\mathcal{M}}\, \xi = (I - xx^\top)\xi, \quad \mathrm{T}_x\mathcal{M} = \{z \mid x^\top z = 0\}, \tag{21}$$

and for any $x_\ell \in \Omega_{f,\mathcal{R}}$,

$$\mathrm{grad}\, f(x_\ell) = \mathrm{Proj}_{\mathrm{T}_{x_\ell}\mathcal{M}} \nabla f(x_\ell) = (I - x_\ell x_\ell^\top) \nabla f(x_\ell).$$

By (8), we can calculate that

$$0 \in \partial^\circ_{\mathcal{R}} f(\bar{x}) = \{(0,t)^\top \mid \forall t \in [-1,1]\},$$

which indicates that $\bar{x}$ is a Clarke stationary point of (20).

Using Proposition 1 (ii), we know that $\hat{\partial}_{\mathcal{R}} f(\bar{x}) = \emptyset$. Using Definition 3, and noting that there exists a neighborhood $B_{\bar{x},\delta}$ for some $\delta > 0$, such that f is continuously differentiable at any $\bar{x}_\ell \neq \bar{x}$ in $B_{\bar{x},\delta} \cap S^1$, we have

$$0 \notin \partial_{\mathcal{R}} f(\bar{x}) := \{(0,1)^\top, (0,-1)^\top\}.$$

Hence $\bar{x}$ is not a limiting stationary point of (20).

The algorithm proposed in this paper is related to the smoothing function $\tilde{f}$ that is employed. It is natural that the convergence result also relates to $\tilde{f}$. According to (3.7) of [59], given $x \in \mathbb{R}^n$, the subdifferential of f associated with $\tilde{f}$ at x on $\mathbb{R}^n$ is

$$G_{\tilde{f}}(x) = \{u \in \mathbb{R}^n \mid \nabla_x \tilde{f}(z_k, \mu_k) \to u \text{ for some } z_k \to x,\ \mu_k \downarrow 0\}. \tag{22}$$

We give the following definition for Riemannian subdifferential of f associated with $\tilde{f}$ at x on $\mathcal{M}$.

Definition 6. Given $x \in \mathcal{M}$, the Riemannian subdifferential of f associated with $\tilde{f}$ at x on $\mathcal{M}$ is

$$G_{\tilde{f},\mathcal{R}}(x) = \{v \in \mathrm{T}_x\mathcal{M} \mid \operatorname{grad} \tilde{f}(z_k, \mu_k) \to v \text{ for some } z_k \in \mathcal{M},\ z_k \to x,\ \mu_k \downarrow 0\}. \tag{23}$$

Remark 2. We require here $u \in G_{\tilde{f}}(x)$ and $v \in G_{\tilde{f},\mathcal{R}}(x)$ are vectors in the Euclidean space that f is defined on, and their entries are finite, i.e., they are not ∞ or $-\infty$. It is clear that if $\mathcal{M}$ is the Euclidean space that f is defined on, then $G_{\tilde{f},\mathcal{R}}(x) = G_{\tilde{f}}(x)$.

Example 3. For the smoothing function $\tilde{f}_\mu(t) = (s_\mu(t))^p$ of $f(t) = |t|^p$ with $0 < p < 1$, where $s_\mu(t)$ is the uniform smoothing function of $|t|$ defined in (16), we have

$$s'_\mu(t) = \begin{cases} \operatorname{sign}(t) & \text{if } |t| \geq \frac{\mu}{2} \\ \frac{2t}{\mu} & \text{if } |t| < \frac{\mu}{2} \end{cases} \quad \text{and} \quad [(s_\mu(t))^p]' = p(s_\mu(t))^{p-1} s'_\mu(t).$$

Here $\operatorname{sign}(t) = 1$ if $t > 0$, $\operatorname{sign}(t) = -1$ if $t < 0$, and $\operatorname{sign}(t) = 0$ otherwise. For an arbitrary real number $v \in \mathbb{R}$, and an arbitrarily chosen sequence $\mu_k \downarrow 0$, let $t_k = a{\mu_k}^{2-p}$ with $a = \frac{4^{p-1}v}{2p}$. It is easy to see that

$$\lim_{\mu_k \downarrow 0} [(s_{\mu_k}(t_k))^p]' = 2p4^{1-p}a = v.$$

Hence $G_{\tilde{f}}(0) = (-\infty, \infty)$. For any point $t \neq 0$, we know that $G_{\tilde{f}}(t) = p|t|^{p-1}\operatorname{sign}(t)$.

Definition 7. A point $x \in \mathcal{M}$ is called a stationary point of (1) associated with $\tilde{f}$, if $0 \in G_{\tilde{f},\mathcal{R}}(x)$, i.e.,

$$\liminf_{z \to x,\ z \in \mathcal{M},\ \mu \downarrow 0} \|\operatorname{grad} \tilde{f}(z, \mu)\| = 0. \tag{24}$$

The following result is an extension of Proposition 3.4 of [59] from $\mathbb{R}^n$ to an embedded submanifold of $\mathbb{R}^n$. The key ingredient for the generalization to Riemannian manifold is to show that the sequence of the Riemannian gradients for the smoothing function has 0 as one of its accumulation points.

Theorem 1. *Let $\mathcal{M}$ be an embedded submanifold of $\mathbb{R}^n$. For any smoothing function $\tilde{f}$ of f as defined in Definition 4, if $x^* \in \mathcal{M}$ is a local minimizer of f on $\mathcal{M}$, then x^* is a stationary point of (1) associated with $\tilde{f}$.*

Proof. Since $x^* \in \mathcal{M}$ is a local minimizer of f on $\mathcal{M}$, minima are preserved by composition with diffeomorphisms (see, e.g., the proof of (2) ⇒ (1) in Proposition 2.2 of [5]), we then know that 0_{x^*} is a local minimizer of $\hat{f} = f \circ R_{x^*}$ on the tangent space $\mathrm{T}_{x^*}\mathcal{M}$. Hence there exists a neighborhood $B_{0_{x^*},\delta}$ of 0_{x^*} such that for any $\eta \in \mathrm{T}_{x^*}\mathcal{M} \cap B_{0_{x^*},\delta}$, it holds that $\hat{f}(0_{x^*}) \leq \hat{f}(\eta)$.

Let us denote $\hat{f}_\mu = \tilde{f}_\mu \circ R_{x^*}$ for any fixed $\mu > 0$. We have

$$\begin{aligned} \hat{f}_\mu(0_{x^*}) = \tilde{f}(x^*, \mu) &\leq f(x^*) + \kappa\omega(\mu) \\ &= \hat{f}(0_{x^*}) + \kappa\omega(\mu) \\ &\leq \hat{f}(\eta) + \kappa\omega(\mu) \quad \text{for any } \eta \in B_{0_{x^*},\delta} \\ &= f(x) + \kappa\omega(\mu) \quad \text{for } x = R_{x^*}(\eta) \\ &\leq \tilde{f}(x, \mu) + 2\kappa\omega(\mu) \\ &= \hat{f}_\mu(\eta) + 2\kappa\omega(\mu). \end{aligned}$$

Thus,

$$\hat{f}_\mu(0_{x^*}) \le \hat{f}_\mu(\eta) + 2\kappa\omega(\mu), \quad \text{for any } \eta \in B_{0_{x^*},\delta}. \tag{25}$$

For any $\eta_z \in \mathrm{T}_{x^*}\mathcal{M} \cap B_{0_{x^*},\delta}$, we define $\eta_\mu = 0_{x^*} + \sqrt{\omega(\mu)}\eta_z \in \mathrm{T}_{x^*}\mathcal{M} \cap B_{0_{x^*},\delta}$ for all μ sufficiently small, and $\eta_\mu \to 0_{x^*}$ as $\mu \downarrow 0$. Since $\hat{f}_\mu$ is continuously differentiable on $\mathrm{T}_{x^*}\mathcal{M}$, by Taylor's expansion we have

$$\hat{f}_\mu(0_{x^*}) = \hat{f}_\mu(\eta_\mu) + \langle \operatorname{grad} \hat{f}_\mu(\eta_\mu), -\sqrt{\omega(\mu)}\eta_z \rangle_{x^*} + o(\sqrt{\omega(\mu)}\|\eta_z\|). \tag{26}$$

Substituting (26) into the left hand side of (25), and replacing η by η_μ with μ that is sufficiently small, we get

$$\sqrt{\omega(\mu)}\langle \operatorname{grad} \hat{f}_\mu(\eta_\mu), -\eta_z \rangle_{x^*} + o(\sqrt{\omega(\mu)}\|\eta_z\|) \le 2\kappa\omega(\mu).$$

Dividing both sides of the above inequality by $\sqrt{\omega(\mu)}$, and taking the limit as $\mu \downarrow 0$, we get

$$\limsup_{\mu\downarrow 0} \langle \operatorname{grad} \hat{f}_\mu(\eta_\mu), -\eta_z \rangle_{x^*} \le 0,$$

which implies that

$$\liminf_{\eta \to 0_{x^*},\ \eta \in \mathrm{T}_{x^*}\mathcal{M},\ \mu\downarrow 0} \langle \operatorname{grad} \hat{f}_\mu(\eta), -\eta_z \rangle_{x^*} \le 0. \tag{27}$$

Note that $\eta_z \in \mathrm{T}_{x^*}\mathcal{M} \cap B_{0_{x^*},\delta}$ can be chosen arbitrarily. Let $\mathcal{M}$ be a d-dimensional embedded submanifold of $\mathbb{R}^n$. We can choose $E : \mathbb{R}^n \to \mathrm{T}_{x^*}\mathcal{M}$ to be a linear bijection such that $\{E(e_i)\}_{i=1}^d$ is an orthonormal basis of $\mathrm{T}_{x^*}\mathcal{M}$, where e_i is the i-th unit vector (see, e.g., Section 2 of [56]). Then

$$\operatorname{grad} \hat{f}_\mu(\eta) = \sum_{i=1}^d \lambda_i^\mu E(e_i), \tag{28}$$

for some $\lambda_i^\mu \in \mathbb{R}$. Let us choose

$$\eta_z^{(i,1)} = \epsilon_i E(e_i),\ \ \eta_z^{(i,2)} = -\epsilon_i E(e_i), \quad \text{for } i = 1, 2, \ldots, d,$$

where $\epsilon_i > 0$ is a sufficiently small constant such that $\eta_z^{(i,1)}, \eta_z^{(i,2)} \in B_{0_{x^*},\delta}$. Substituting $\operatorname{grad}\hat{f}_\mu(\eta)$ in (27) by (28), and substituting η_z in (27) by $\eta_z^{(i,1)}$ and $\eta_z^{(i,2)}$, respectively, we obtain

$$\liminf_{\mu\downarrow 0} -\epsilon_i \lambda_i^\mu \ge 0, \quad \text{and} \quad \liminf_{\mu\downarrow 0} \epsilon_i \lambda_i^\mu \ge 0.$$

The above two inequalities indicate

$$\lim_{\mu\downarrow 0} \lambda_i^\mu = 0.$$

Since $i = 1, 2, \ldots, d$ can be chosen arbitrarily, the above equality holds for each i. Hence, we get

$$\liminf_{\eta \to 0_{x^*},\ \eta \in \mathrm{T}_{x^*}\mathcal{M},\ \mu\downarrow 0} \|\operatorname{grad} \hat{f}_\mu(\eta)\| = \lim_{\mu\downarrow 0} \|\sum_{i=1}^d \lambda_i^\mu E(e_i)\| = 0. \tag{29}$$

According to [2, Lemma 7.4.9, pp. 153], we know that for any constant $\tau > 1$, there exist constants $\bar{\delta} > 0$ and $\bar{d} > 0$ such that for all $\|\eta\| \le \bar{d}$ and $x = R_{x^*}(\eta) \in B_{x^*,\bar{\delta}} \cap \mathcal{M}$,

$$\|\operatorname{grad} \tilde{f}(x,\mu)\| = \|\operatorname{grad} \tilde{f}_\mu(R_{x^*}(\eta))\| \le \tau \|\operatorname{grad} \hat{f}_\mu(\eta)\|.$$

Here $\bar{\delta}$ and $\bar{d}$ relate only to τ and the definitions of the retraction R and the Riemannian metric g of $\mathcal{M}$, which can be deduced from the proof of Lemma 7.4.9 of [2]. Taking the limit $\eta \to 0_{x^*}, \eta \in \mathrm{T}_{x^*}\mathcal{M}$, $\mu \downarrow 0$ to both sides of the above inequality and using (29), we get

$$\liminf_{x \to x^*,\ x \in \mathcal{M},\ \mu\downarrow 0} \|\operatorname{grad} \tilde{f}(x,\mu)\| = 0,$$

and hence x^* is a stationary point of (1) associated with $\tilde{f}$ as desired. □

We will show in Section 4 that any accumulation point of the proposed RSSD method is a stationary point of (1) associated with $\tilde{f}$. We will also show in Section 4 that any accumulation point of the proposed RSSD method is also a limiting stationary point of (1), provided $\tilde{f}$ satisfies the Riemannian gradient sub-consistency (to be defined in the next subsection) at the accumulation point.

3.2. Riemannian gradient sub-consistency Now we define the Riemannian gradient sub-consistency of $\tilde{f}$ at $x \in \mathcal{M}$, which makes a connection between the Riemannian subdifferential $G_{\tilde{f},\mathcal{R}}(x)$ associated with $\tilde{f}$ and the Riemannian limiting subdifferential $\partial_{\mathcal{R}} f(x)$. The Riemannian gradient sub-consistency is essential to show that any accumulation point of the RSSD method is a limiting stationary point of (1). Hence when minimizing a nonsmooth but locally Lipschitz function on $\mathcal{M}$, the RSSD method has stronger convergence result than the existing methods that guarantee any accumulation point is a Clarke stationary point of (1), e.g., ϵ-subgradient algorithm [29], line search algorithms [30], Riemannian gradient sampling algorithm [32], and Riemannian proximal gradient methods [34].

Definition 8. Given $x \in \mathbb{R}^n$, a smoothing function $\tilde{f}$ of the function f is said to satisfy the gradient sub-consistency at x on $\mathbb{R}^n$ if

$$G_{\tilde{f}}(x) \subseteq \partial f(x). \tag{30}$$

Given $x \in \mathcal{M}$, $\tilde{f}$ is said to satisfy the Riemannian gradient sub-consistency at x on $\mathcal{M}$ if

$$G_{\tilde{f},\mathcal{R}}(x) \subseteq \partial_{\mathcal{R}} f(x). \tag{31}$$

We say that $\tilde{f}$ satisfies the gradient sub-consistency on $\mathbb{R}^n$ if (30) holds for any $x \in \mathbb{R}^n$, and that $\tilde{f}$ satisfies the Riemannian gradient sub-consistency on $\mathcal{M}$ if (31) holds for any $x \in \mathcal{M}$.

Later we will show that if f is nonsmooth but locally Lipschitz near x on $\mathbb{R}^n$, $\tilde{f}$ is a smoothing function of f, and the gradient sub-consistency of the smoothing function $\tilde{f}$ at x on $\mathbb{R}^n$ holds, then the Riemannian gradient sub-consistency of $\tilde{f}$ on $\mathcal{M}$ holds. Furthermore, we also provide in (35) a Riemannian optimization problem that minimizing a non-Lipschitz function f on $\mathcal{M}$. We show that its smoothing function $\tilde{f}$ defined in (42) satisfies the Riemannian gradient sub-consistency on $\mathcal{M}$. It is worth mentioning that in numerical experiments of Section 5, both the problem (59) of finding the sparsest vectors in a subspace, and the problem in (65) for sparsely-used orthogonal complete dictionary learning using ℓ_p $(0<p<1)$ regularization are examples of the model (35).

If the inclusion is substituted by the equality in (30), then we say $\tilde{f}$ satisfies the gradient consistency at x on $\mathbb{R}^n$. If the inclusion is substituted by the equality in (31), then we say $\tilde{f}$ satisfies the Riemannian gradient consistency at x on $\mathcal{M}$. Clearly, the gradient consistency indicates the gradient sub-consistency. The gradient consistency of $\tilde{f}$ on $\mathbb{R}^n$ has been well studied in smoothing methods for nonsmooth optimization. For nonsmooth but locally Lipschitz function f, it has been shown that the gradient consistency property on $\mathbb{R}^n$ holds for various smoothing functions in many real applications [12, 13, 19, 55, 58].

The following theorem demonstrates that given an embedded submanifold $\mathcal{M}$ of $\mathbb{R}^n$, $x \in \mathcal{M}$, if the gradient sub-consistency of $\tilde{f}$ at x on $\mathbb{R}^n$ holds, then the Riemannian gradient sub-consistency of $\tilde{f}$ holds at x on $\mathcal{M}$, provided that f is locally Lipschitz near x on $\mathbb{R}^n$.

Theorem 2. *Given an embedded submanifold $\mathcal{M}$ of $\mathbb{R}^n$ and a vector $x \in \mathcal{M}$, let f be a locally Lipschitz function near x on $\mathbb{R}^n$, with $\tilde{f}$ being a smoothing function of f. If the gradient sub-consistency of $\tilde{f}$ at x on $\mathbb{R}^n$ holds, then the Riemannian gradient sub-consistency of $\tilde{f}$ at x on $\mathcal{M}$ holds.*

Proof. Let $v \in G_{\tilde f,\mathcal{R}}(x)$. Note that $G_{\tilde f}(x) \subseteq \partial f(x)$ is bounded if f is a locally Lipschitz function near x on $\mathbb{R}^n$. Then there exist subsequences $\{x_{\mu_k}\} \subset \mathcal{M}$, $x_{\mu_k} \to x$, and $\{\mu_k\}$, $\mu_k \downarrow 0$ as $k \to \infty$, and a vector $u \in G_{\tilde f}(x)$ such that

$$u = \lim_{x_{\mu_k}\to x,\ x_{\mu_k}\in\mathcal{M},\ \mu_k\downarrow 0} \nabla_x \tilde f(x_{\mu_k},\mu_k), \tag{32}$$

and

$$\begin{aligned} v &= \lim_{x_{\mu_k}\to x,\ x_{\mu_k}\in\mathcal{M},\ \mu_k\downarrow 0} \operatorname{grad} \tilde f(x_{\mu_k},\mu_k) \\ &= \lim_{x_{\mu_k}\to x,\ x_{\mu_k}\in\mathcal{M},\ \mu_k\downarrow 0} \operatorname{Proj}_{\mathrm{T}_{x_{\mu_k}}\mathcal{M}} \nabla_x \tilde f(x_{\mu_k},\mu_k), \\ &= \operatorname{Proj}_{\mathrm{T}_x\mathcal{M}} u. \end{aligned} \tag{33}$$

The last equality holds because

$$\begin{aligned} &\|\operatorname{Proj}_{\mathrm{T}_{x_{\mu_k}}\mathcal{M}} \nabla_x \tilde f(x_{\mu_k},\mu_k) - \operatorname{Proj}_{\mathrm{T}_x\mathcal{M}} u\| \\ \le\ &\|\operatorname{Proj}_{\mathrm{T}_{x_{\mu_k}}\mathcal{M}} \nabla_x \tilde f(x_{\mu_k},\mu_k) - \operatorname{Proj}_{\mathrm{T}_{x_{\mu_k}}\mathcal{M}} u\| + \|\operatorname{Proj}_{\mathrm{T}_{x_{\mu_k}}\mathcal{M}} u - \operatorname{Proj}_{\mathrm{T}_x\mathcal{M}} u\| \\ \le\ &\|\nabla_x \tilde f(x_{\mu_k},\mu_k) - u\| + \|\operatorname{Proj}_{\mathrm{T}_{x_{\mu_k}}\mathcal{M}} u - \operatorname{Proj}_{\mathrm{T}_x\mathcal{M}} u\| \\ \to\ &0, \end{aligned}$$

as $x_{\mu_k} \to x,\ x_{\mu_k} \in \mathcal{M},\ \mu_k \downarrow 0$. Here the second inequality comes from the fact that $\operatorname{Proj}_{\mathrm{T}_{x_{\mu_k}}\mathcal{M}}$ is nonexpansive. Moreover, $\|\nabla_x \tilde f(x_{\mu_k},\mu_k) - u\| \to 0$ by (32), and $\|\operatorname{Proj}_{\mathrm{T}_{x_{\mu_k}}\mathcal{M}} u - \operatorname{Proj}_{\mathrm{T}_x\mathcal{M}} u\| \to 0$ because $\operatorname{Proj} : x \to \operatorname{Proj}_{\mathrm{T}_x\mathcal{M}}$ is continuously differentiable according to [11, Exercise 3.66, pp. 59].

Since the gradient sub-consistency at x on $\mathbb{R}^n$ holds, i.e., $G_{\tilde f}(x) \subseteq \partial f(x)$, we know that $u \in \partial f(x)$. By the definition of limiting subdifferential of f on $\mathbb{R}^n$,

$$\exists\ u_\ell \in \hat\partial f(x_\ell),\ (x_\ell, f(x_\ell)) \to (x, f(x)) \text{ such that } \lim_{\ell\to\infty} u_\ell = u.$$

By the characterization of Riemannian regular subdifferential in (18), we have

$$v_\ell = \operatorname{Proj}_{\mathrm{T}_{x_\ell}\mathcal{M}} u_\ell \in \hat\partial_{\mathcal{R}} f(x_\ell), \tag{34}$$

and using the same arguments of proving (33), we have

$$\lim_{\ell\to\infty} v_\ell = \lim_{\ell\to\infty} \operatorname{Proj}_{\mathrm{T}_{x_\ell}\mathcal{M}} u_\ell = \operatorname{Proj}_{\mathrm{T}_x\mathcal{M}} u = v.$$

This implies $v \in \partial_{\mathcal{R}} f(x)$, and hence the smoothing function $\tilde f$ of f satisfies the Riemannian gradient sub-consistency at x on $\mathcal{M}$. □

Furthermore, we consider the following non-Lipschitz Riemannian minimization problem:

$$\min_{x\in\mathcal{M}}\ f(x) := \hat f(x) + \lambda \sum_{i=1}^{m} \varphi(|d_i^\top x|), \tag{35}$$

where $\hat f$ is a continuously differentiable function, $\mathcal{M}$ is an embedded submanifold of $\mathbb{R}^n$, $0 \neq d_i \in \mathbb{R}^n$, $i = 1,\dots,m$, are nonzero vectors, $\lambda > 0$ is a given constant, and $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a nonsmooth penalty function. The problem with $\mathcal{M} = \mathbb{R}^n$ has been well investigated in [23], which includes many widely used nonsmooth penalty functions φ in variable selection, image restoration, and signal reconstruction.

If φ is nonsmooth but locally Lipschitz, it is easy to see that f is nonsmooth but locally Lipschitz. In this case, its Riemannian gradient sub-consistency has been investigated in Theorem 2. Below we only focus on φ that is not locally Lipschitz. Motivated by Assumption 1.1 in [23], we require φ to satisfy the following assumption.

Assumption 1. *The function $\varphi:\mathbb{R}_+\to\mathbb{R}_+$ is continuous at 0 with $\varphi(0)=0$, $\varphi'(0^+)=\infty$, and φ is nonsmooth but locally Lipschitz in $(0,\infty)$.*

For instance, the bridge penalty φ_1 used in [20, 22, 23, 24], the log penalty φ_2 [23], and the penalty φ_3 used in [44]

$$\varphi_1(t)=t^p,\ \varphi_2(t)=\log(\alpha t^p+1),\ \varphi_3(t)=\min\{t^p,1\},\quad \text{for some } 0<p<1,\alpha>0, \tag{36}$$

are not locally Lipschitz functions on $\mathbb{R}_+:=\{t\in\mathbb{R}\mid t\ge 0\}$ that satisfy Assumption 1. If the objective function f is not locally Lipschitz on $\mathbb{R}^n$, it may also be not locally Lipschitz on $\mathcal{M}$ as well (see the following example).

Example 4. Let us consider $\mathcal{M}=S^1$ which is the unit circle as in Example 2, $\bar{x}=(\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2})^\top$, and

$$f(x)=|x_1-x_2|^{\frac{1}{2}}+|x_1+x_2|^{\frac{1}{2}}.$$

For each $\gamma>0$, let us define

$$h_\gamma(x):=|x_1-x_2|^{\frac{1}{2}}+\gamma(x_1+x_2),\quad \text{and}\quad \delta_\gamma:=\frac{1}{2\gamma^2}>0.$$

Then it is clear that $f(\bar{x})-h_\gamma(\bar{x})=0$, and we claim that

$$f(x)-h_\gamma(x)\ =\ |x_1+x_2|^{\frac{1}{2}}-\gamma(x_1+x_2)\ge 0,\quad \text{for any } x\in B_{\bar{x},\delta_\gamma}.$$

To see this, note that for any $x\in B_{\bar{x},\delta_\gamma}$, if $x_1+x_2\le 0$, then it is obvious that $f(x)-h_\gamma(x)\ge 0$. We then only need to consider $x\in B_{\bar{x},\delta_\gamma}$ and $x_1+x_2>0$. In this case, $f(x)-h_\gamma(x)\ge 0$ is equivalent to

$$(x_1+x_2)^{\frac{1}{2}}\ge\gamma(x_1+x_2),$$

that is, $(x_1+x_2)^{\frac{1}{2}}\le\frac{1}{\gamma}$. In view of $x\in B_{\bar{x},\delta_\gamma}$, we know that

$$\max\{|x_1-\bar{x}_1|,|x_2-\bar{x}_2|\}\le\sqrt{(x_1-\bar{x}_1)^2+(x_2-\bar{x}_2)^2}\le\frac{1}{2\gamma^2},$$

which indicates

$$x_1+x_2=(x_1-\bar{x}_1)+(x_2-\bar{x}_2)\le\frac{1}{\gamma^2},$$

and consequently $(x_1+x_2)^{\frac{1}{2}}\le\frac{1}{\gamma}$. Thus $f(x)-h_\gamma(x)\ge 0$ also holds in this case.

Therefore, $f-h_\gamma$ attains minimum at $\bar{x}$ in a neighborhood $B_{\bar{x},\delta_\gamma}$ of $\bar{x}$. According to Lemma 1, we have for any $\gamma>0$,

$$v_\gamma=\nabla h_\gamma(\bar{x})=\frac{1}{2\sqrt[4]{2}}\begin{pmatrix}1\\-1\end{pmatrix}+\gamma\begin{pmatrix}1\\1\end{pmatrix}\in\hat{\partial}f(\bar{x}).$$

Thus according to (21) and Proposition 2, we find for any $\gamma>0$,

$$\begin{aligned}u_\gamma &= \mathrm{Proj}_{\mathrm{T}_{\bar{x}}\mathcal{M}}v_\gamma=(I-\bar{x}\bar{x}^\top)v_\gamma\\ &=\frac{1}{2}\begin{pmatrix}1&1\\1&1\end{pmatrix}v_\gamma=\gamma\begin{pmatrix}1\\1\end{pmatrix}\in\hat{\partial}_{\mathcal{R}}f(\bar{x}).\end{aligned}\tag{37}$$

It is easy to see that $\|u_\gamma\|\to\infty$ as $\gamma\to\infty$. In view of Lemma 2, we know that f is not locally Lipschitz on S^1.

Remark 3. We consider a general embedded submanifold $\mathcal{M}$ of $\mathbb{R}^n$. Let $\bar{x} \in \mathcal{M}$, and

$$I_{\bar{x}} = \{i \in \{1, \ldots, m\} \mid d_i^\top \bar{x} \neq 0\} \text{ and } J_{\bar{x}} = \{i \in \{1, \ldots, m\} \mid d_i^\top \bar{x} = 0\}. \tag{38}$$

Assume $J_{\bar{x}} \neq \emptyset$. Now we consider the model (35) with $\varphi = \varphi_1$. Using arguments similar to that in the above simple example, we choose an arbitrary $i_0 \in J_{\bar{x}}$ and let

$$h_\gamma(x) = \hat{f}(x) + h_1(x) + h_{2,\gamma}(x),$$

where

$$h_1(x) = \lambda \sum_{i \in I_{\bar{x}}} |d_i^\top x|^p, \quad h_{2,\gamma}(x) = \lambda \gamma d_{i_0}^\top x \quad \text{for any } \gamma > 0.$$

It is easy to see that $f - h_\gamma$ attains local minimum in a neighborhood $B_{\bar{x},\delta_\gamma}$ for a positive constant δ_γ, and $f(\bar{x}) = h_\gamma(\bar{x})$. Hence by Lemma 1 we have $\nabla h_\gamma(\bar{x}) \in \hat{\partial} f(\bar{x})$, and consequently by Proposition 2, we find

$$u_\gamma = \mathrm{Proj}_{\mathrm{T}_{\bar{x}}\mathcal{M}} \nabla h_\gamma(\bar{x}) \in \hat{\partial}_{\mathcal{R}} f(\bar{x}). \tag{39}$$

As long as there exists a point $\bar{x} \in \mathcal{M}$ such that $\mathrm{Proj}_{\mathrm{T}_{\bar{x}}\mathcal{M}} \nabla h_{2,\gamma}(\bar{x}) \neq 0$, since $\gamma > 0$ can be chosen arbitrarily large, we can conclude that f is not locally Lipschitz on $\mathcal{M}$ according to Lemma 2.

For instance, if $\mathcal{M}$ is the unit sphere in $\mathbb{R}^n$:

$$S^{n-1} = \{x \in \mathbb{R}^n \mid \|x\| = 1\}, \tag{40}$$

then as $d_{i_0}^\top \bar{x} = 0$, we have

$$\mathrm{Proj}_{\mathrm{T}_{\bar{x}}\mathcal{M}} \nabla h_{2,\gamma}(\bar{x}) = (I - \bar{x}\bar{x}^\top) \lambda \gamma d_{i_0} = \lambda \gamma d_{i_0} \neq 0. \tag{41}$$

Hence f is not locally Lipschitz on $\mathcal{M}$.

Many applications can be formulated in the form of (35), such as finding the sparsest vectors in a subspace, and the sparsely-used orthogonal complete dictionary learning that will be discussed later in Section 5.

Let $\tilde{s}_\mu(t)$ be a smoothing function of $|t|$, $\tilde{\varphi}$ be a smoothing function of φ satisfying Definition 4, and the function

$$\tilde{f}(x, \mu) = \hat{f}(x) + \lambda \sum_{i=1}^m \tilde{\varphi}(\tilde{s}_\mu(d_i^\top x), \mu) \tag{42}$$

be a smoothing function of f defined in (35). For instance, for $\varphi = \varphi_1$ and $\varphi = \varphi_2$ in (36), we can choose

$$\tilde{\varphi}(t, \mu) = \varphi(t),$$

and for φ_3, we can use

$$\tilde{\varphi}_3(t, \mu) = \begin{cases} t^p - (t^p - 1)_+ & \text{if } |t^p - 1| \geq \frac{\mu}{2} \\ t^p - \left(\frac{(t^p-1)^2}{2\mu} + \frac{t^p - 1}{2} + \frac{\mu}{8}\right) & \text{if } |t^p - 1| < \frac{\mu}{2}. \end{cases}$$

Theorem 3. *The smoothing function $\tilde{f}$ that is constructed in (42) for the non-Lipschitz objective function f in (35) satisfies the Riemannian gradient sub-consistency on $\mathcal{M}$.*

Proof. For an arbitrary $x \in \mathbb{R}^n$, let the index sets I_x and J_x be defined in (38) with $\bar{x}$ being replaced by x. Let D_{J_x} be the matrix whose columns are d_i, $i \in J_x$, i.e.,

$$D_{J_x} = (d_i)_{i \in J_x} \in \mathbb{R}^{n \times |J_x|}, \tag{43}$$

with $|J_x|$ being the cardinality of the index set J_x.

If $J_x = \emptyset$, then f is locally Lipschitz near x on $\mathbb{R}^n$. It is clear that $\tilde{f}$ satisfies the gradient sub-consistency at x on $\mathbb{R}^n$. Thus $\tilde{f}$ satisfies the Riemannian gradient sub-consistency at x on $\mathcal{M}$ as shown in Theorem 2.

Otherwise $J_x \neq \emptyset$. Define

$$f_1(z) := \lambda \sum_{i \in I_x} \varphi(|d_i^\top z|), \quad \text{and} \quad f_2(z) := \lambda \sum_{i \in J_x} \varphi(|d_i^\top z|),$$
$$\tilde{f}_1(z,\mu) := \lambda \sum_{i \in I_x} \tilde{\varphi}(\tilde{s}_\mu(d_i^\top z), \mu), \quad \text{and} \quad \tilde{f}_2(z,\mu) := \lambda \sum_{i \in J_x} \tilde{\varphi}(\tilde{s}_\mu(d_i^\top z), \mu).$$

Clearly

$$\lambda \sum_{i=1}^m \varphi(|d_i^\top z|) = f_1(z) + f_2(z), \quad \text{and} \quad \tilde{f}(z,\mu) = \hat{f}(z) + \tilde{f}_1(z,\mu) + \tilde{f}_2(z,\mu).$$

It is clear that

$$\nabla_x \tilde{f}(z_k,\mu_k) = \nabla \hat{f}(z_k) + \nabla_x \tilde{f}_1(z_k,\mu_k) + \nabla_x \tilde{f}_2(z_k,\mu_k), \tag{44}$$

and

$$\lim_{k\to\infty} \nabla \hat{f}(z_k) = \nabla \hat{f}(x) \quad \text{and} \quad \lim_{k\to\infty} \nabla_x \tilde{f}_1(z_k,\mu_k) = \nabla f_1(x). \tag{45}$$

By direct computation,

$$\nabla_x \tilde{f}_2(z_k,\mu_k) = \sum_{i \in J_x} \tilde{\varphi}'(s_{\mu_k}(d_i^\top z_k), \mu_k) s'_{\mu_k}(d_i^\top z_k) d_i = D_{J_x} u_k, \tag{46}$$

where

$$u_k := u_k(z_k,\mu_k) = \left(\tilde{\varphi}'(s_{\mu_k}(d_i^\top z_k), \mu_k) s'_{\mu_k}(d_i^\top z_k)\right)_{i \in J_x} \in \mathbb{R}^{|J_x|}.$$

Let $v \in G_{\tilde{f},\mathcal{R}}(x)$. Then there exist infinite sequences $\{z_k\} \subset \mathcal{M}$, $z_k \to x$, and $\{\mu_k\}$, $\mu_k \downarrow 0$ as $k \to \infty$ such that

$$\begin{aligned} v &= \lim_{z_k\to x,\ z_k\in\mathcal{M},\ \mu_k\downarrow 0} \operatorname{grad} \tilde{f}(z_k,\mu_k) \\ &= \lim_{z_k\to x,\ z_k\in\mathcal{M},\ \mu_k\downarrow 0} \operatorname{Proj}_{\mathrm{T}_{z_k}\mathcal{M}} \nabla_x \tilde{f}(z_k,\mu_k). \end{aligned} \tag{47}$$

For any $g_k^1, g_k^2 \in \mathbb{R}^n$, it is easy to see that

$$\begin{aligned} &\left| \left\| \operatorname{Proj}_{\mathrm{T}_{z_k}\mathcal{M}} g_k^2 \right\| - \left\| \operatorname{Proj}_{\mathrm{T}_{z_k}\mathcal{M}} (g_k^1 + g_k^2) \right\| \right| \\ \le &\left\| \operatorname{Proj}_{\mathrm{T}_{z_k}\mathcal{M}} (g_k^1 + g_k^2) - \operatorname{Proj}_{\mathrm{T}_{z_k}\mathcal{M}} g_k^2 \right\| \le \left\| g_k^1 \right\|, \end{aligned}$$

which implies

$$\left\| \operatorname{Proj}_{\mathrm{T}_{z_k}\mathcal{M}} g_k^2 \right\| \le \left\| \operatorname{Proj}_{\mathrm{T}_{z_k}\mathcal{M}} (g_k^1 + g_k^2) \right\| + \left\| g_k^1 \right\|. \tag{48}$$

By substituting $g_k^1 = \nabla \hat{f}(z_k) + \nabla_x \tilde{f}_1(z_k,\mu_k)$ and $g_k^2 = \nabla_x \tilde{f}_2(z_k,\mu_k)$ into (48), we have

$$\left\|\mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}\, \nabla_x \tilde{f}_2(z_k,\mu_k)\right\| \le \left\|\mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}\, \nabla_x \tilde{f}(z_k,\mu_k)\right\| + \left\|\nabla \hat{f}(z_k) + \nabla_x \tilde{f}_1(z_k,\mu_k)\right\|.$$

The two terms on the right-hand side of the above inequality are bounded by noting (47) and (45). Thus

$$\left\{\left\|\mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}\, \nabla_x \tilde{f}_2(z_k,\mu_k)\right\|\right\} \text{ is bounded.} \tag{49}$$

We can write

$$\begin{aligned} D_{J_x} u_k &= b_k^1 + b_k^2, \ \text{ where } b_k^1 \in \mathrm{T}_{z_k}\mathcal{M}, b_k^2 \in (\mathrm{T}_{z_k}\mathcal{M})^\perp; &(50)\\ \nabla \hat{f}(z_k) + \nabla_x \tilde{f}_1(z_k,\mu_k) &= a_k^1 + a_k^2, \ \text{ where } a_k^1 \in \mathrm{T}_{z_k}\mathcal{M}, a_k^2 \in (\mathrm{T}_{z_k}\mathcal{M})^\perp. &(51)\end{aligned}$$

Here $(\mathrm{T}_{z_k}\mathcal{M})^\perp$ is the orthogonal complement of $\mathrm{T}_{z_k}\mathcal{M}$. By (49) and (46), we know that $\{b_k^1\}$ is bounded.

Let $r = \mathrm{rank}(D_{J_x})$ be the rank of D_{J_x} and $\mathrm{Range}(D_{J_x})$ be the range of D_{J_x}. Let $\{j_1, j_2, \ldots, j_r\} \subseteq J_x$ such that $\{d_{j_i}, i = 1,2\ldots,r\}$ constitutes a basis for $\mathrm{Range}(D_{J_x})$. We define $\xi_i = d_{j_i}$, $i = 1,2\ldots,r$. If $r < n$, we can find $\xi_i \in \mathbb{R}^n$, $i = r+1,\ldots,n$, such that $\{\xi_1,\xi_2,\ldots,\xi_n\}$ constitutes a basis for $\mathbb{R}^n$. Let us define the matrix $\Xi = \begin{pmatrix}\xi_1,\xi_2,\ldots,\xi_n\end{pmatrix} \in \mathbb{R}^{n\times n}$ that is invertible. Then the linear system with unknown vector w

$$\Xi w = b_k^1$$

is consistent, and has a unique solution $w_k = \Xi^{-1} b_k^1$. It is clear that $\{w_k\}$ is bounded.

Let $\bar{K} \subseteq K$ be an infinite sequence such that $\lim_{k\to\infty,\ k\in\bar{K}} w_k = \bar{w}$. By using (50) and (51), we get

$$\begin{aligned}\mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}\, \nabla_x \tilde{f}(z_k,\mu_k) &= \mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}(\nabla \hat{f}(z_k) + \nabla_x \tilde{f}_1(z_k,\mu_k) + D_{J_x} u_k)\\ &= \mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}(a_k^1 + a_k^2 + b_k^1 + b_k^2)\\ &= \mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}(a_k^1 + b_k^1) = a_k^1 + b_k^1.\end{aligned}$$

Consequently,

$$\begin{aligned} v &= \lim_{z_k\to x,\ z_k\in\mathcal{M},\ \mu_k\downarrow 0} \mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}\, \nabla_x \tilde{f}(z_k,\mu_k)\\ &= \lim_{k\to\infty,\ k\in\bar{K}} (a_k^1 + b_k^1)\\ &= \lim_{z_k\to x,\ z_k\in\mathcal{M},\ \mu_k\downarrow 0,\ k\in\bar{K}} \mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}\left(\nabla \hat{f}(z_k) + \nabla_x \tilde{f}_1(z_k,\mu_k) + \Xi w_k\right)\\ &= \mathrm{Proj}_{\mathrm{T}_x\mathcal{M}}\left(\nabla \hat{f}(x) + \nabla f_1(x) + \Xi\bar{w}\right),\end{aligned}$$

where the last equality can be obtained using the arguments for (33).

We now define the function

$$\bar{h}(z) = \hat{f}(z) + f_1(z) + \sum_{i=1}^{r} \bar{w}_i d_{j_i}^T z + \sum_{i=r+1}^{n} \bar{w}_i \xi_i^T (z - x).$$

It is then easy to check that there exists a neighborhood $B_{x,\delta}$ for some $\delta > 0$ such that $\bar{h}(z) \le f(z)$ with $\bar{h}(x) = f(x)$, and $\nabla\bar{h}(x) = \nabla\hat{f}(x) + \nabla f_1(x) + \Xi\bar{w}$. Then by Lemma 1, $\nabla\bar{h}(x) \in \hat{\partial} f(x)$. Hence

$$v = \mathrm{Proj}_{\mathrm{T}_x\mathcal{M}}\left(\nabla \hat{f}(x) + \nabla f_1(x) + \Xi\bar{w}\right) \in \hat{\partial}_{\mathcal{R}} f(x) \subseteq \partial_{\mathcal{R}} f(x), \tag{52}$$

which indicates that $\tilde{f}$ satisfies the Riemannian gradient sub-consistency at x on $\mathcal{M}$ in this case.

Since $x \in \mathcal{M}$ is arbitrary, we have that $\tilde{f}$ satisfies the Riemannian gradient sub-consistency on $\mathcal{M}$ as desired. □

At the end of this section, we make clear the relation between the set of limiting stationary points defined in Definition 5

$$S_l = \{x^* \in \mathcal{M} \mid 0 \in \partial_{\mathcal{R}} f(x^*)\},$$

and that of stationary points associated with $\tilde{f}$ defined in Definition 7

$$S_{\tilde{f}} = \{x^* \in \mathcal{M} \mid 0 \in G_{\tilde{f},\mathcal{R}}(x^*)\}.$$

Any local minimizer of the Riemannian optimization problem (1) lies in both sets. If the Riemannian gradient sub-consistency holds on $\mathcal{M}$, i.e., $G_{\tilde{f},\mathcal{R}}(x^*) \subseteq \partial_{\mathcal{R}} f(x^*)$ for any $x^* \in \mathcal{M}$, then $S_{\tilde{f}} \subseteq S_l$.

4. Riemannian smoothing steepest descent method

In this section, we present our RSSD method for solving the Riemannian optimization problem (1), which is detailed in Algorithm 1. The objective function in (1) is allowed to be non-Lipschitz on $\mathcal{M}$. We always assume there exists at least one global optimal solution of (1).

Algorithm 1 Riemannian smoothing steepest descent (RSSD) method for solving (1)

```
Input: $x_0 \in \mathcal{M}$, $\delta_{opt} \geq 0$, $\delta_0 > 0$, $\mu_{opt} \geq 0$, $\mu_0 > 0$, $\sigma \in (0,1), \beta \in (0,1)$, $\bar{\alpha} > 0$, $\theta_\delta \in (0,1)$, $\theta_\mu \in (0,1)$.
for $\ell = 0,1,2,\ldots$ do
    Compute $\eta_\ell = -\operatorname{grad} \tilde{f}(x_\ell, \mu_\ell)$.
    if $\|\eta_\ell\| \leq \delta_{opt}$ and $\mu_\ell \leq \mu_{opt}$ then
        return
    else if $\|\eta_\ell\| \leq \delta_\ell$ then
        $\mu_{\ell+1} := \theta_\mu \mu_\ell, \delta_{\ell+1} := \theta_\delta \delta_\ell$,
        $x_{\ell+1} := x_\ell$.
    else
        $\mu_{\ell+1} = \mu_\ell, \delta_{\ell+1} = \delta_\ell$.
        Find $t_\ell := \beta^{m_\ell} \bar{\alpha}$ where $m_\ell$ is the smallest integer such that
            $\tilde{f}(R_{x_\ell}(\beta^{m_\ell} \bar{\alpha} \eta_\ell), \mu_\ell) \leq \tilde{f}(x_\ell, \mu_\ell) - \sigma \beta^{m_\ell} \bar{\alpha} \|\operatorname{grad} \tilde{f}(x_\ell, \mu_\ell)\|^2$.    (53)
        Set $x_{\ell+1} := R_{x_\ell}(t_\ell \eta_\ell)$.
    end if
end for
```

A few remarks for Algorithm 1 are in order. First, the line search (53) is well defined and t_ℓ can be found in finite trials. To see this, note that for fixed μ_ℓ, $\tilde{f}(\cdot, \mu_\ell)$ is continuously differentiable. Clearly, we have

$$\lim_{t \downarrow 0} \frac{\tilde{f}_{\mu_\ell} \circ R_{x_\ell}(t\eta_\ell) - \tilde{f}_{\mu_\ell} \circ R_{x_\ell}(0_{x_\ell})}{t} = (\tilde{f}_{\mu_\ell} \circ R_{x_\ell})'(0_{x_\ell}, \eta_\ell) = \langle \operatorname{grad} \tilde{f}(x_\ell, \mu_\ell), \eta_\ell \rangle.$$

Note that $\eta_\ell = -\operatorname{grad} \tilde{f}(x_\ell, \mu_\ell)$. Thus there exists $\alpha > 0$ such that for all $t \in (0, \alpha)$,

$$\tilde{f}_{\mu_\ell} \circ R_{x_\ell}(t\eta_\ell) \leq \tilde{f}_{\mu_\ell} \circ R_{x_\ell}(0_{x_\ell}) - t\sigma \|\operatorname{grad} \tilde{f}(x_\ell, \mu_\ell)\|^2.$$

This guarantees that the line search step (53) is well defined.

The following result is an extension of Theorem 3.5 together with Remark 3.2 of [59]. The key ingredient for the extension is to show that the index set K related to the Riemannian gradient of the smoothing function defined in Theorem 4 is an infinite set.

THEOREM 4. *Let $K=\{\ell \mid \|\eta_\ell\|\le \delta_\ell\}$ and $\{x_\ell\}$ be an infinite sequence generated by Algorithm 1 with $\delta_{opt}=\mu_{opt}=0$. Then the following statements hold.*
(i) Any accumulation point x^* of $\{x_\ell\}_{\ell\in K}$ is a stationary point of (1) associated with $\tilde f$.
(ii) In addition, if $\tilde f$ satisfies the Riemannian gradient sub-consistency at x^* on $\mathcal{M}$, then x^* is a limiting stationary point of (1).

Proof. We first claim that if there exists an accumulation point $x^*\in\mathcal{M}$, then K is an infinite set, and

$$\lim_{\ell\to\infty,\ \ell\in K}\delta_\ell=0 \quad \text{and} \quad \lim_{\ell\to\infty,\ \ell\in K}\mu_\ell=0. \tag{54}$$

Suppose on the contrary that K is a finite set. This means there exists $\bar\ell$ such that for all $\ell\ge\bar\ell$,

$$\delta_\ell\equiv\delta_{\bar\ell},\quad \mu_\ell\equiv\mu_{\bar\ell},$$

and

$$\eta_\ell=-\operatorname{grad}\tilde f(x_\ell,\mu_{\bar\ell}),\quad \|\eta_\ell\|>\delta_{\bar\ell}>0. \tag{55}$$

Therefore, for $\ell\ge\bar\ell$, we have $x_{\ell+1}=R_{x_\ell}(t_\ell\eta_\ell)$, where t_ℓ is obtained by using the line search (53) with fixed $\mu_{\bar\ell}$. Then Algorithm 1 becomes a Riemannian steepest descent method for minimizing a smooth function $\tilde f(\cdot,\mu_{\bar\ell})$ on $\mathcal{M}$. According to Theorem 4.3.1 of [2], we have $\operatorname{grad}\tilde f(x^*,\mu_{\bar\ell})=0$, which contradicts (55). Therefore, K is an infinite set. Note that for each $\ell\in K$, we have

$$\mu_{\ell+1}=\theta_\mu\mu_\ell \quad \text{and} \quad \delta_{\ell+1}=\theta_\delta\delta_\ell$$

with decaying factors $\theta_\mu\in(0,1)$ and $\theta_\delta\in(0,1)$. This, together with K being an infinite set, yields (54) as desired.

By Algorithm 1, we have

$$\lim_{\ell\to\infty,\ \ell\in K}\|\operatorname{grad}\tilde f(x_\ell,\mu_\ell)\|=\lim_{\ell\to\infty,\ \ell\in K}\|\eta_\ell\|\le\lim_{\ell\to\infty,\ \ell\in K}\delta_\ell=0.$$

Let $\check K$ be a subsequence of K such that $\lim_{\ell\to\infty,\ \ell\in\check K}x_\ell=x^*$. The completeness of $\mathcal{M}$ guarantees that $x^*\in\mathcal{M}$. Thus

$$\liminf_{x\to x^*,\ x\in\mathcal{M},\ \mu\downarrow 0}\|\operatorname{grad}\tilde f(x,\mu)\|=0,\ \text{and}\ 0\in G_{\tilde f,\mathcal{R}}(x^*).$$

Hence x^* is a stationary point of (1) associated with $\tilde f$. That is, statement (i) holds.

In addition, if $\tilde f$ satisfies the Riemannian gradient sub-consistency at x^* on $\mathcal{M}$, then we know $G_{\tilde f,\mathcal{R}}(x^*)\subseteq\partial_{\mathcal{R}}f(x^*)$. Thus we find $0\in\partial_{\mathcal{R}}f(x^*)$. Hence x^* is a limiting stationary point of (1). Consequently statement (ii) holds. □

The sequence $\{x_\ell\}$ generated by Algorithm 1 is guaranteed to have an accumulation point, if the following assumption holds.

ASSUMPTION 2. *For any $\bar\mu\in(0,\mu_0]$ and any given vector $\bar x\in\mathcal{M}$, the level set $\mathcal{L}_{\bar x,\bar\mu}=\{x\in\mathcal{M}\mid \tilde f(x,\bar\mu)\le\tilde f(\bar x,\bar\mu)\}$ is compact.*

Assumption 2 holds if $\mathcal{M}$ is compact. Assumption 2 also holds if f is coercive in $\mathbb{R}^n$, i.e., $|f(x)|\to\infty$ if $\|x\|\to\infty$, because for an arbitrary $\bar\mu\in(0,\mu_0]$ and an arbitrary given vector $\bar x\in\mathcal{M}$, by using Definition 4 for smoothing function, $x\in\mathcal{L}_{\bar x,\bar\mu}$ implies that

$$f(x)\le f(\bar x)+2\kappa\omega(\bar\mu),$$

which, together with the coercivity of f, yields that $\mathcal{L}_{\bar{x},\bar{\mu}}$ is compact.

Next, we explain how the RSSD method can be considered as an extension of the smoothing steepest descent method from $\mathbb{R}^n$ to an embedded submanifold of $\mathbb{R}^n$ in the following remark. Here the smoothing steepest descent method comes from Algorithm 3.1 and Remark 3.2 of [59]. To be specific, we set $\Omega = \mathbb{R}^n$ in Algorithm 3.1 of [59], and substitute "the active set method in Algorithm 2.1" in Algorithm 3.1 [59] by the well-known "steepest descent method with Armijo line search"; see, e.g., subsection 1.2 of [7]. Then we get the smoothing steepest descent method on $\mathbb{R}^n$.

Remark 4. When $\mathcal{M} = \mathbb{R}^n$, let us set the parameters in our RSSD method to be

$$\theta_\mu = \theta_\delta = \zeta, \quad \delta_0 = \hat{\gamma}\mu_0,$$

where ζ and $\hat{\gamma}$ are the parameters in the smoothing steepest descent method on $\mathbb{R}^n$. On the other hand, for the smoothing steepest descent method on $\mathbb{R}^n$, let us set $n_1 = 1$, choose the steepset descent method with Armijo line search to be the same as (53) in our RSSD method, and

$$\mu_{k+1} = \zeta \mu_k.$$

Then from the same initial point x_0, the sequence $\{x_0, x_\ell\}_{\ell \in K}$ where $K = \{\ell \mid \|\eta_\ell\| \le \delta_\ell\}$ generated by our RSSD method coincides with the sequence generated by the smoothing steepest descent method.

In computation, we do not set $\mu_{opt} = \delta_{opt} = 0$, but instead set them to be small positive real numbers. For instance, we can set $\mu_{opt} = \delta_{opt} = \epsilon$ for a given small positive real number ϵ. Then we expect to get an ϵ-approximate stationary point $\hat{x}$ of (1) associated with $\tilde{f}$ defined as follows, by implementing Algorithm 1.

Definition 9. We say that $\hat{x} \in \mathcal{M}$ is an ϵ-approximate stationary point of (1) associated with $\tilde{f}$, if

$$\mu \le \epsilon \quad \text{and} \quad \|\operatorname{grad} \tilde{f}(\hat{x}, \mu)\| \le \epsilon. \tag{56}$$

This definition of an ϵ-approximate stationary point of (1) associated with $\tilde{f}$ is motivated by [28], where smoothing direct-search methods in nonsmooth optimization on $\mathbb{R}^n$ have been developed to obtain an ϵ-approximate solution.

Theorem 5 below is novel even when $\mathcal{M} = \mathbb{R}^n$. It has not been considered before for the smoothing steepest descent method in $\mathbb{R}^n$.

Theorem 5. *Under Assumption 2, after finite iterations, Algorithm 1 with $\mu_{opt} = \delta_{opt} = \epsilon$ will reach an iterate point that is an ϵ-approximate stationary point of (1) with respect to $\tilde{f}$.*

Proof. Let $\theta = \max\{\theta_\mu, \theta_\delta\}$, and

$$n_K := \left\lceil \max\left\{ \log_\theta \frac{\epsilon}{\mu_0}, \log_\theta \frac{\epsilon}{\|\eta_0\|}, 1 \right\} \right\rceil.$$

Here $\lceil r \rceil$ refers to the smallest integer that is no less than the real number r.

We then have

$$\theta^{n_K}\mu_0 \le \epsilon \quad \text{and} \quad \theta^{n_K}\|\eta_0\| \le \epsilon.$$

Let us denote $K_\epsilon = \{k_1, k_2, \ldots, k_{n_K}\}$ where $k_i < k_j$ for $1 \le i < j \le n_K$ such that K_ϵ contains the first n_K elements that satisfy $\|\eta_{k_i}\| \le \delta_{k_i}$. Thus

$$\mu_{k_{n_K}} \le \theta^{n_K}\mu_0 \le \epsilon \quad \text{and} \quad \theta_{k_{n_K}} \le \theta^{n_K}\|\eta_0\| \le \epsilon,$$

and we get the iterate point x_{n_K} as an ϵ-approximate stationary point of (1) associated with $\tilde{f}$.

From iterations x_{k_i} to $x_{k_{i+1}}$ for $1 \leq i \leq n_K - 1$, the smoothing parameter keeps the same as μ_{k_i} and Algorithm 1 performs the iterates of the Riemannian steepest descent method for minimizing the smooth function $\tilde{f}(x, \mu_{k_i})$ on $\mathcal{M}$. According to Corollary 4.3.2 of [2],

$$\lim_{\ell \to \infty} \| \operatorname{grad} \tilde{f}(x^\ell, \mu_{k_i}) \| = 0.$$

This implies that after a finite number of steps (say $\widehat{\ell}_i$), we will get $\|\eta_{k_i + \widehat{\ell}_i}\| = \| \operatorname{grad} \tilde{f}(x^{k_i + \widehat{\ell}_i}, \mu_{k_i})\| \leq \delta_{k_i}$. Hence after

$$k_{n_K} = \sum_{i=1}^{n_K - 1} \widehat{\ell}_i$$

steps, we will get an ϵ-approximate stationary point of (1) associated with $\tilde{f}$ as desired. □

Remark 5. Complexity for non-Lipschitz optimization in $\mathbb{R}^n$ has been investigated in [8, 9, 19, 20]. In [20], it is shown that solving a non-Lipschitz optimization problem is strongly NP-hard. In [8, 9], a smoothing sequential quadratic regularization (SSQR) algorithm was proposed for solving non-Lipschitz optimization. The worst-case iteration complexity of the SSQR algorithm for finding an ϵ affine-scaled stationary point is $\mathcal{O}(\epsilon^{-2})$. It is worth mentioning that the construction of a special strongly convex quadratic minimization problem, and a special rule for updating the smoothing parameter at each iteration are essential to show the worst-case complexity in [8, 9]. New techniques need to be developed to obtain the iteration complexity of our RSSD method for Riemannian non-Lipschitz optimization and we leave it as a future work.

5. Numerical experiments

In this section, we apply our RSSD method (Algorithm 1) to solve two problems: finding the sparsest vectors in a subspace (FSV), and the sparsely-used orthogonal complete dictionary learning problem (ODL). A notebook with 1.80GHz CPU and 16GB of RAM is used for the numerical experiments. We implement Algorithm 1 in MATLAB (version R2018b).

5.1. Finding the sparsest vectors in a subspace

The FSV problem seeks the sparsest vectors in an n-dimensional linear subspace $W \subset \mathbb{R}^m$ $(m > n)$. This problem has been studied recently and it finds interesting applications and connection with sparse dictionary learning, sparse PCA, and many other problems in signal processing and machine learning [45, 46]. This problem is also known as dual principal component pursuit and finds applications in robust subspace recovery [51, 63]. Let $Q \in \mathbb{R}^{m \times n}$ denote a matrix whose columns form an orthonormal basis of W. The FSV problem can be formulated as

$$\min \ \|Qx\|_0, \quad \text{s.t. } x \in S^{n-1}, \tag{57}$$

where S^{n-1} is the unit sphere, and $\|z\|_0$ counts the number of nonzero entries of z. Because of the combinatorial nature of the cardinality function $\|\cdot\|_0$, (57) is very difficult to solve in practice. In the literature, people have been focusing on its ℓ_1 norm relaxation given below [46, 45, 51, 63]:

$$\min \ \|Qx\|_1, \quad \text{s.t. } x \in S^{n-1}, \tag{58}$$

where $\|z\|_1 := \sum_i |z_i|$ is the ℓ_1 norm of vector z. Many algorithms have been proposed for solving (58), including the Riemannian gradient sampling algorithm [32], projected subgradient method [62], Riemannian subgradient method [41], manifold proximal point algorithm [15] and so on.

Moreover, for the compressive sensing problems that have the same objective functions as (57) and (58), people have found that using the ℓ_p quasi-norm $\|z\|_p^p := \sum_i |z_i|^p$ $(0 < p < 1)$ to replace

$\|z\|_1$ can help to promote the sparsity of z [14, 27, 20, 24, 42, 43]. Motivated by this, we propose the following ℓ_p $(0<p<1)$ minimization model for the FSV problem:

$$\min\ f(x) := \|Qx\|_p^p, \quad \text{s.t. } x \in S^{n-1}. \tag{59}$$

We will illustrate that comparing with (58), (59) with proper choices of $0<p<1$ is a better approximation to (57). To this end, we construct a simple example below for which the global minimizers of the Riemannian ℓ_0 model are known ahead of time.

Let $V = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_5] \in \mathbb{R}^{35\times 5}$ be a matrix, whose columns have $5, 6, 7, 8$, and 9 nonzero entries sequentially, and each nonzero entry of the column is the only nonzero entry in its row. Specifically, the nonzero entries of V are

$$\begin{aligned}
V(1:5,1) &= (\,1^0\ 2^0\ 3^0\ 4^0\ 5^0\,)^T;\\
V(6:11,2) &= (\,1^1\ 2^1\ 3^1\ 4^1\ 5^1\ 6^1\,)^T;\\
V(12:18,3) &= (\,1^2\ 2^2\ 3^2\ 4^2\ 5^2\ 6^2\ 7^2\,)^T;\\
V(19:26,4) &= (\,1^3\ 2^3\ 3^3\ 4^3\ 5^3\ 6^3\ 7^3\ 8^3\,)^T;\\
V(27:35,5) &= (\,1^4\ 2^4\ 3^4\ 4^4\ 5^4\ 6^4\ 7^4\ 8^4\ 9^4\,)^T.
\end{aligned}$$

Let the linear space W be the span of column vectors of V, and let $Q = [\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_5] \in \mathbb{R}^{35\times 5}$ be a matrix where $\mathbf{q}_j = \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|}$ for $j = 1, \dots, 5$. Clearly the columns of the matrix Q form an orthonormal basis of W.

Denote by $\mathbf{e}_i \in \mathbb{R}^5$ the i-th column of the identity matrix for $i = 1, \dots, 5$. It is then easy to see that for the Riemannian ℓ_0 model, the sparsest vector in the linear space W has five nonzero entries and $\pm\mathbf{e}_1$ are the only two global minimizers, and $\pm\mathbf{e}_i$, $i = 2, 3, 4, 5$ are local minimizers corresponding to vectors $\pm Q\mathbf{e}_i \in \mathbb{R}^{35}$ in the linear space W with $6, 7, 8, 9$ nonzero entries, respectively. By direct computation, we list in Table 1 the objective values at $\pm\mathbf{e}_i$ for $i = 1, \dots, 5$ for the three models, respectively.

Table 1. Objective values of the three models at $\pm\mathbf{e}_i$, $i = 1, \dots, 5$

	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
$\lVert Q(\pm\mathbf{e}_i)\rVert_0$	5	6	7	8	9
$\lVert Q(\pm\mathbf{e}_i)\rVert_1$	2.2361	2.2014	2.0473	1.9385	1.8631
$\lVert Q(\pm\mathbf{e}_i)\rVert_p^p$, $p=0.1$	4.6134	5.3531	5.8989	6.3284	6.6766
$\lVert Q(\pm\mathbf{e}_i)\rVert_p^p$, $p=0.01$	4.9599	5.9310	6.8764	7.8018	8.7088
$\lVert Q(\pm\mathbf{e}_i)\rVert_p^p$, $p=0.001$	4.9960	5.9931	6.9875	7.9798	8.9702

It is obvious that $\pm\mathbf{e}_1$ are not global minimizers of the Riemannian ℓ_1 model, and $\|Q(\pm\mathbf{e}_5)\|_1$ achieves the lowest values among the five objective values. For $p = 0.1, 0.01, 0.001$, we find that $\pm\mathbf{e}_1$ achieve the lowest objective values among $\pm\mathbf{e}_i, i = 1, \dots, 5$. In view of the facts that each entry of Q is nonnegative and each nonzero entry of the column of Q is the only nonzero entry in its row, we know that $Q\mathbf{e}_i \geq 0$, $i = 1, \dots, 5$, and there is no index $\bar{j}$ such that

$$(Q\mathbf{e}_{i_1})_{\bar{j}} > 0 \text{ and } (Q\mathbf{e}_{i_2})_{\bar{j}} > 0, \quad \forall\ i_1 \neq i_2, i_1, i_2 \in \{1, \dots, 5\}.$$

For any $x \in S^4$, we have $x = \sum_{i=1}^5 x_i \mathbf{e}_i$, $\sum_{i=1}^5 x_i^2 = 1$, and $|x_i| \geq x_i^2$ for $i = 1, \dots, 5$. It is easy to see that

$$|Qx| = |\sum_{i=1}^5 x_i Q\mathbf{e}_i| = \sum_{i=1}^5 |x_i| Q\mathbf{e}_i \geq \sum_{i=1}^5 x_i^2 Q\mathbf{e}_i,$$

and consequently

$$\begin{aligned}\|Qx\|_p^p = \||Qx|\|_p^p &\geq \|\sum_{i=1}^5 x_i^2 Q\mathbf{e}_i\|_p^p \\ &\geq \sum_{i=1}^5 x_i^2 \|Q\mathbf{e}_i\|_p^p \geq \sum_{i=1}^5 x_i^2 \|Q\mathbf{e}_1\|_p^p = \|Q\mathbf{e}_1\|_p^p,\end{aligned}$$

where the second inequality is obtained from the concavity of $\|\cdot\|_p^p$ for $0<p<1$. Therefore, $\pm\mathbf{e}_1$ are the global minimizers of the Riemannian ℓ_p model for $p=0.1, 0.01, 0.001$, which coincide to the global minimizers of the original Riemannian ℓ_0 model. Moreover, the objective values at $\pm\mathbf{e}_i$, $i=1,\ldots,5$ for the Riemannian ℓ_p model keep increasing for i as that for the Riemannian ℓ_0 model.

There exists at least one nonzero row of Q, say $\hat{q}_{i_0}^\top$, and there exists a vector $\hat{x}\in S^{n-1}$ such that $\hat{q}_{i_0}^\top \hat{x}=0$. Then by the same arguments as in Remark 2, the objective function in (59) is not locally Lipschitz on S^{n-1}. Hence algorithms proposed in [32, 62, 41, 15] for solving (58) do not apply to (59). We propose to solve (59) using our RSSD method. We also use the proposed RSSD method to solve the Riemannian ℓ_1 norm minimization problem (58). Now we show the details below.

According to [2], the tangent space at $x\in S^{n-1}$ is

$$\mathrm{T}_x S^{n-1} := \{z\in\mathbb{R}^n \mid x^\top z = 0\},$$

and the projection of $\xi\in\mathbb{R}^n$ onto the tangent space $\mathrm{T}_x S^{n-1}$ is

$$\mathrm{Proj}_{\mathrm{T}_x S^{n-1}}\,\xi = (I - xx^\top)\xi.$$

In our RSSD algorithm, we use $R_x(\xi) = (x+\xi)/\|x+\xi\|$ as the retraction function. We use the following smoothing function for (59):

$$\tilde{f}(x,\mu) = \sum_{i=1}^m [s_\mu((Qx)_i)]^p, \tag{60}$$

where $s_\mu(t)$ is the uniform smoothing function for $|t|$ defined in (16).

The parameters of our RSSD method are set as

$$\mu_0 = 1,\;\; \delta_0 = 0.1, \theta_\mu = 0.5,\;\; \theta_\delta = 0.5. \tag{61}$$

We choose 50 initial points x_0 from normally distributed random vectors, using MATLAB code

$$\mathrm{randn}('\mathrm{state}', j); x_0 = \mathrm{randn}(n,1); x_0 = x_0/\mathrm{norm}(x_0),$$

for $j=1,\ldots,50$.

For each instance, we terminate it when the CPU time reaches 50 seconds and find that all the 50 computed solutions fall in $\{\pm\mathbf{e}_i, i=1,\ldots,5\}$ corresponding to $p=0.1, 0.01, 0.001$. The CPU time is measured with MATLAB command "cputime". Here we say $\hat{x}$ fall in $\{\pm\mathbf{e}_i\}$ for $i=1,\ldots,5$, if

$$\mathrm{gap}(\hat{x}, \pm\mathbf{e}_i) = \min\{\|\hat{x}-\mathbf{e}_i\|, \|\hat{x}+\mathbf{e}_i\|\} \leq 10^{-8}. \tag{62}$$

We record in Table 2 the frequencies of the computed solutions that fall in $\{\pm\mathbf{e}_i, i=1,\ldots,5\}$, respectively. We can conclude that the Riemannian ℓ_p model with $p=0.001$ succeeds to find the true global minimizers of the original Riemannian ℓ_0 model 10 times from the 50 initial points. In contrast, the Riemannian ℓ_1 minimization model does not find the true global minimizers from the 50 initial points.

This example demonstrates that there indeed exists a problem for which the Riemannian ℓ_1 model fails to find the sparsest vector in a subspace, while the Riemannian ℓ_p model with suitable $0<p<1$ can find the sparsest vector in a subspace. Hence it is useful to develop an algorithm for solving Riemannian non-Lipschitz optimization with rigorous convergence result. This is the main motivation of this paper.

TABLE 2. Frequencies of the computed solutions that fall in $\pm\mathbf{e}_i$, $i=1,\dots,5$ from the 50 initial points, using Riemannian ℓ_1 model and Riemannian ℓ_p model, respectively.

	$\pm\mathbf{e}_1$	$\pm\mathbf{e}_2$	$\pm\mathbf{e}_3$	$\pm\mathbf{e}_4$	$\pm\mathbf{e}_5$
ℓ_1 model	0	4	15	10	21
ℓ_p model, $p=0.1$	1	4	16	9	20
ℓ_p model, $p=0.01$	6	5	14	8	17
ℓ_p model, $p=0.001$	10	9	10	5	16

5.2. Sparsely-used orthogonal complete dictionary learning Given a set of data $Y=[\mathbf{y}_1,\mathbf{y}_2,\dots,\mathbf{y}_m]\in\mathbb{R}^{n\times m}$, the sparsely-used orthogonal complete dictionary learning (ODL) seeks a dictionary that can sparsely represent Y. More specifically, ODL seeks an orthogonal matrix $X=[\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_n]\in\mathbb{R}^{n\times n}$ and a sparse matrix $S\in\mathbb{R}^{n\times m}$ such that $Y\approx XS$. The matrix X is called an orthogonal dictionary. We refer to [50] for more details of this model. This problem can be modeled as the Riemannian ℓ_0 minimization problem [49]:

$$\min\ \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{y}_i^\top X\|_0,\quad \text{s.t. } X\in\mathrm{St}(n,n), \tag{63}$$

where $\mathrm{St}(n,n)=\{X\in\mathbb{R}^{n\times n}\mid X^\top X=I_n\}$ is the orthogonal group, which is a special case of the Steifel manifold. To overcome the computational difficulty of the Riemannian ℓ_0 minimization model, the ℓ_0 term is usually replaced by the ℓ_1 norm in the literature, which leads to the following Riemannian ℓ_1 minimization problem for ODL [49, 50]:

$$\min\ \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{y}_i^\top X\|_1,\quad \text{s.t. } X\in\mathrm{St}(n,n). \tag{64}$$

Here we again consider the Riemannian ℓ_p $(0<p<1)$ quasi-norm minimization model

$$\min\ \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{y}_i^\top X\|_p^p,\quad \text{s.t. } X\in\mathrm{St}(n,n), \tag{65}$$

and apply the RSSD method to solve it. We now specify the details. The tangent space of the Stiefel manifold $\mathrm{St}(n,n)$ is

$$\mathrm{T}_X\mathrm{St}:=\{\xi\in\mathbb{R}^{n\times n}\ :\ \xi^\top X+X^\top\xi=0\}.$$

The projection of $Z\in\mathbb{R}^{n\times n}$ onto the tangent space $\mathrm{T}_X\mathrm{St}(n,n)$ is

$$\mathrm{Proj}_{\mathrm{T}_X\mathrm{St}}Z=Z-\frac{1}{2}X(X^\top Z+Z^\top X). \tag{66}$$

We use the QR factorization as the retraction on the Stiefel manifold, which is given by $R_X(\xi)=\mathrm{qf}(X+\xi)$. Here $\mathrm{qf}(A)$ denotes the Q factor of the QR decomposition of A.

In [41], Li et al. proposed a Riemannian subgradient method and its variants – Riemannian incremental subgradient method and Riemannian stochastic subgradient method – for solving the Riemannian ℓ_1 minimization problem (64). In this section, we use our RSSD to solve the Riemannian ℓ_p minimization model (65) with $p=0.001$, and compare its performance with the algorithms proposed in [41] for solving (64). We thus generate the synthetic data for ODL in a similar manner as [41], which is detailed below. We first generate the underlying orthogonal dictionary $X^*\in\mathrm{St}(n,n)$ with $n=30$ whose entries are drawn according to standard Gaussian distribution. The number of samples $m=\lfloor 10\cdot n^{1.5}\rfloor=1643$. The sparse matrix $S^*\in\mathbb{R}^{n\times m}$ is generated such that the entries follow the Bernoulli-Gaussian distribution with parameter 0.5. Finally, we set $Y = X^*S^*$. We generate 50 instances using this procedure. For each instance, we generate two different initial points: one is a standard Gaussian random vector denoted as x_0^{Gauss}, and the other one is a uniform random vector denoted as x_0^{uniform}. For the ease of presentation, we denote the three algorithms in [41] – Riemannian subgradient method, Riemannian incremental subgradient method, and Riemannian stochastic subgradient method – as R-Full, R-Inc and R-Sto, respectively. We use the same parameters in (61) for RSSD. The codes for R-Full, R-Inc and R-Sto were downloaded from the author's webpage[1].

For each instance, we terminate each method if the CPU time reaches 50 seconds. We truncate the entries of $Y^\top \hat{X}$ as

$$(Y^\top \hat{X})_{ij} = 0, \quad \text{if } |(Y^\top \hat{X})_{ij}| < \tau,$$

where $\tau > 0$ is a pregiven tolerance, and $\hat{X}$ is the computed solution. We report the average of the sparsity level of $Y^\top \hat{X}$ over 50 instances in Table 3, where the sparsity level is computed by

$$\textbf{sparsity level} = \frac{\text{number of zero entries of } Y^\top \hat{X}}{mn}.$$

Note that the desired sparsity level of $Y^\top \hat{X}$ is 0.5 because of the way that S^* was generated. We see from Table 3 that the ℓ_p minimization model with $p = 0.001$ solved by our RSSD method provides the best results in terms of the sparsity level.

For each instance, we compute the sparsity level at the latest iterate point obtained by each method when the CPU time reaches $t = 1, 2, \dots, 50$ seconds, respectively. We then compute the corresponding average sparsity level of the 50 instances and plot the trajectory of the sparsity level at $t = 1, 2, \dots, 50$ seconds in Figures 1 and 2. We use a log-scale on the x-axis in Figures 1 and 2. From these figures, it is clear that the ℓ_p minimization model (65) with $p = 0.001$ solved by the RSSD method provides the best results in terms of sparsity level. More specifically, the RSSD method can improve the sparsity to the desired level 0.5 after about 15 seconds, while the other three algorithms stopped making progress after about one second and none of the three algorithms achieves the sparsity level higher than 0.4. This example also demonstrates the necessity of developing the RSSD method for solving the Riemannian non-Lipschitz optimization.

Table 3. Average of sparsity levels of computed solutions from 50 instances

Initial points	ℓ_1 minimization model			ℓ_p model, $p = 0.001$
	R-Full	R-Inc	R-Sto	RSSD
x_0^{Gauss}, $\tau = 10^{-4}$	0.3727	0.3857	0.3456	**0.5000**
x_0^{Gauss}, $\tau = 10^{-5}$	0.3697	0.3852	0.3450	**0.4895**
x_0^{Uniform}, $\tau = 10^{-4}$	0.3727	0.3784	0.3234	**0.5000**
x_0^{Uniform}, $\tau = 10^{-5}$	0.3675	0.3773	0.3222	**0.4915**

We admit that from our numerical experience, the computation cost of our RSSD method for each iteration is higher than the Riemannian subgradient-type methods including R-Full, R-Inc, and R-Sto methods in [41]. The reason is that the Armijo line search is used in the RSSD method, while no line search strategy is adopted in the Riemannian subgradient-type methods in [41]. It is also possible to develop more efficient smoothing algorithms than the RSSD method for solving Riemannian non-Lipschitz optimization, by making use of the theoretical analysis of Riemannian

[1] https://github.com/lixiao0982/Riemannian-subgradient-methods.

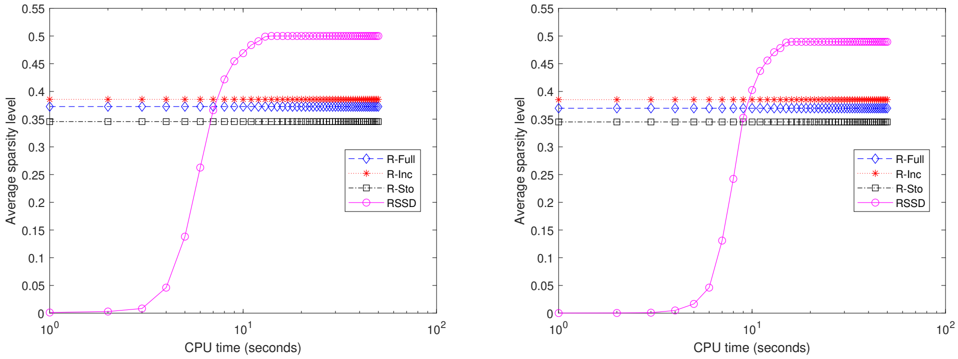

FIGURE 1. Average sparsity level versus CPU time of 50 instances using Gaussian initial points. Left: $\tau = 10^{-4}$; Right: $\tau = 10^{-5}$. Log-scale on the x-axis.

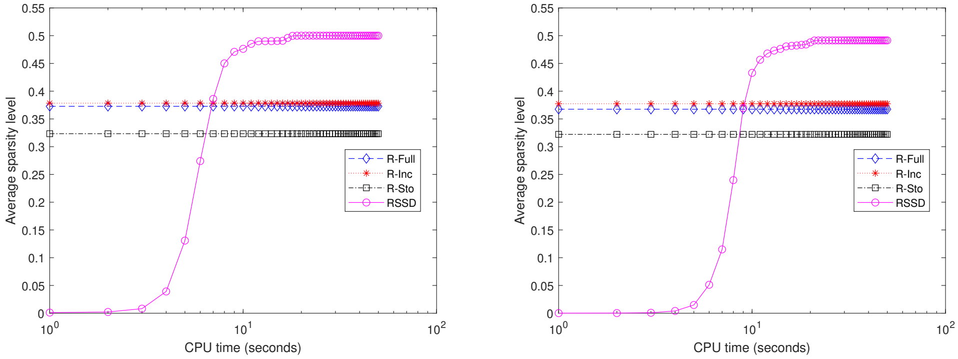

FIGURE 2. Average sparsity level versus CPU time of 50 instances using uniform initial points. Left: $\tau = 10^{-4}$; Right: $\tau = 10^{-5}$. Log-scale on the x-axis.

generalized subdifferentials and Riemannian gradient sub-consistency property developed in this paper.

Recall that we use the CPU time budget (50 seconds) to terminate each method in numerical experiments. For our RSSD method, we record μ_ℓ and $\|\eta_\ell\| = \| -\operatorname{grad}\tilde{f}(x_\ell, \mu_\ell)\|$ when the sparsity level becomes stable, and try the values around them. We find that $\mu_{opt} \in [2\times 10^{-3}, 3\times 10^{-3}]$ and $\delta_{opt} \in [4\times 10^{-3}, 10^{-2}]$ are suitable as stopping criterion of the RSSD method for the sparsely-used ODL problem, and in this case the sparsity level keeps almost the same but the number of iterations are around 1/3 of that given by the CPU budget (50 seconds). If smaller values of μ_{opt} and δ_{opt} are also used as a stopping criteria, together with the CPU budget (50 seconds), the RSSD method will often reach the CPU time budget, but the sparsity level keeps almost the same.

6. Concluding remarks

In this paper, we study the Riemannian generalized subdifferentials, and Riemannian gradient sub-consistency relating to non-Lipschitz optimization on embedded submanifolds of $\mathbb{R}^n$. We then develop RSSD, a novel Riemannian smoothing steepest descent method, for minimizing a non-Lipschitz function over embedded submanifolds of $\mathbb{R}^n$. We prove that any accumulation point generated by our RSSD method is a stationary point of (1) associated with the smoothing function employed in the method, which is necessary for local optimality of (1). Moreover, we also prove that any accumulation point is a limiting stationary point of (1), if the Riemannian gradient sub-consistency property holds at the accumulation point. We show that smoothing functions satisfy the Riemannian gradient sub-consistency under mild conditions. Numerical results on finding the sparsest vectors in a subspace and the sparsely-used orthogonal complete dictionary learning demonstrate the necessity of studying non-Lipschitz optimization on embedded submanifolds of $\mathbb{R}^n$ and the effectiveness of our RSSD method for solving non-Lipschitz optimization on embedded submanifolds of $\mathbb{R}^n$.

Acknowledgments. The authors are grateful to two anonymous referees for their constructive comments that have greatly improved the presentation of this paper. C. Zhang was supported in part by National Natural Science Foundation of China (No. 12171027) and Natural Science Foundation of Beijing (No. 1202021). X. Chen was supported in part by Hong Kong Research Council Grant PolyU15300219. S. Ma was supported in part by NSF grants DMS-2243650, CCF-2308597, UC Davis CeDAR (Center for Data Science and Artificial Intelligence Research) Innovative Data Science Seed Funding Program, and a startup fund from Rice University.

References

[1] Absil PA, Gallivan KA (2009) Accelerated line-search and trust-region methods. *SIAM J. Numer. Anal.* 47:997–1018.

[2] Absil PA, Mahony R, Sepulchre R (2008) *Optimization Algorithms on Matrix Manifolds* (Princeton University Press, Princeton, NJ).

[3] Adler RL, Dedieu JP, Margulies JY, Martens M, Shub M (2002) Newton's method on Riemannian manifolds and a geometric model for the human spine. *IMA J. Numer. Anal.* 22:359–390.

[4] Azagra D, Ferrera J, Løpez-Mesas F (2005) Nonsmooth analysis and Hamilton-Jacobi equations on Riemannian manifolds. *J. Funct. Anal.* 91:304–361.

[5] Azagra D, Ferrera J, Sanz B (2012) Viscosity solutions to second order partial differential equations on Riemannian manifolds. *J. Differ. Equations* 245:307–336.

[6] Bačák M, Bergmann R, Steidl G, Weinmann A (2016) A second order non-smooth variational model for restoring manifold-valued images. *SIAM J. Sci. Comput.* 38:A567–A597.

[7] Bertsekas DP (1999) *Nonlinear Programming* (Athena Scientific, Belmont, Massachusetts).

[8] Bian W, Chen X (2013) Worst-case complexity of smoothing quadratic regularization methods for non-Lipschitzian optimization. *SIAM J. Optim.* 23:1718–1741.

[9] Bian W, Chen X (2015) Linearly constrained non-Lipschitz optimization for image restoration. *SIAM J. Imaging Sci.* 8:2294–2322.

[10] Bolte J, Danilids D, Lewis A, Shiota M (2007) Clarke subgradients of stratifiable functions. *SIAM J. Optim.* 18:556–572.

[11] Boumal N (2023) *An introduction to optimization on smooth manifolds* (Cambridge University Press), URL http://dx.doi.org/10.1017/9781009166164.

[12] Burke JV, Hoheisel T (2013) Epi-convergent smoothing with applications to convex composite functions. *SIAM J. Optim.* 23:1457–1479.

[13] Burke JV, Hoheisel T, Kanzow C (2013) Gradient consistency for integral-convolution smoothing functions. *Set-Valued Var. Anal.* 21:359–376.

[14] Chartrand R, Yin W (2008) Iteratively reweighted algorithms for compressive sensing. *ICASSP*.

[15] Chen S, Deng Z, Ma S, So AMC (2021) Manifold proximal point algorithms for dual principal component pursuit and orthogonal dictionary learning. *IEEE Trans. Signal Process.* 69:4759–4773.

[16] Chen S, Ma S, So AMC, Zhang T (2020) Proximal gradient method for nonsmooth optimization over the Stiefel manifold. *SIAM J. Optim.* 30:210–239.

[17] Chen S, Ma S, Xue L, Zou H (2020) An alternating manifold proximal gradient method for sparse principal component analysis and sparse canonical correlation analysis. *INFORMS J. Optim.* 2:192–208.

[18] Chen W, Ji H, You Y (2016) An augmented Lagrangian method for ℓ_1-regularized optimization problems with orthogonality constraints. *SIAM J. Sci. Comput.* 38:B570–B592.

[19] Chen X (2012) Smoothing methods for nonsmooth, nonconvex minimization. *Math. Program.* 134:71–99.

[20] Chen X, Ge D, Wang Z, Ye Y (2014) Complexity of unconstrained L_2-L_p minimization. *Math. Program.* 143:371–383.

[21] Chen X, Guo L, Luo Z, Ye JJ (2017) An augmented Lagrangian method for non-Lipschitz nonconvex programming. *SIAM J. Numer. Anal.* 55:168–193.

[22] Chen X, Ng MK, Zhang C (2012) Non-Lipschitz ℓ_p regularization and box constrained model for image restoration. *IEEE Trans. Image Process.* 21:4709–4721.

[23] Chen X, Niu L, Yuan Y (2013) Optimality conditions and smoothing trust region Newton method for non-Lipschitz optimization. *SIAM J. Optim.* 23:1528–1552.

[24] Chen X, Xu F, Ye Y (2010) Lower bound theory of nonzero entries in solutions of $\ell_2 - \ell_p$ minimization. *SIAM J. Sci. Comput.* 32:2832–2852.

[25] Chen X, Zhou W (2010) Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization. *SIAM J. Imaging Sci.* 3:765–790.

[26] Clarke FH (1990) *Optimization and Nonsmooth Analysis* (Journal Wiley & Sons).

[27] Foucart S, Lai MJ (2009) Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. *Appl. Comput. Harmon. Anal.* 26:395–407.

[28] Garmanjani R, Vicente LN (2013) Smoothing and worst-case complexity for direct-search methods in nonsmooth optimization. *IMA J. Numer. Anal.* 3:1008–1028.

[29] Grohs P, Hosseini S (2016) ϵ-subgradient algorithms for locally Lipschitz functions on Riemannian manifolds. *Adv. Comput. Math.* 42:333–366.

[30] Hosseini S, Huang W, Yousefpour R (2018) Line search algorithms for locally Lipschitz functions on Riemannian manifolds. *SIAM J. Optim.* 28:596–619.

[31] Hosseini S, Pouryayevali MR (2001) Generalized gradients and characterization of epi-Lipschitz sets in Riemannian manifolds. *Nonlinear Anal.-Theor.* 74:3884–3895.

[32] Hosseini S, Uschmajew A (2017) A Riemannian gradient sampling algorithm for nonsmooth optimization on manifolds. *SIAM J. Optim.* 27:173–189.

[33] Huang W, Absil PA, Gallivan KA (2018) A Riemannian BFGS method without differentiated retraction for nonconvex optimization problems. *SIAM J. Optim.* 28:470–495.

[34] Huang W, Wei K (2022) Riemannian proximal gradient methods. *Math. Program.* 194:371–413.

[35] Jiang B, Meng X, Wen Z, Chen X (2023) An exact penalty approach for optimization with nonnegative orthogonality constraints. *Math. Program.* 198:855–897.

[36] Kovnatsky A, Glashoff K, Bronstein MM (2016) MADMM: a generic algorithm for non-smooth optimization on manifolds. *European Conference on Computer Vision*, 680–696 (Springer).

[37] Lai R, Osher S (2014) A splitting method for orthogonality constrained problems. *J. Sci. Comput.* 58:431–449.

[38] Ledyaev YS, Zhu QJ (2007) Nonsmooth analysis on smooth manifolds. *Trans. Amer. Math. Soc.* 359:3687–3732.

[39] Li J, Balasubramanian K, Ma S (2022) Stochastic zeroth-order Riemannian derivative estimation and optimization. *Math. Oper. Res.* 48(2):1183–1211.

[40] Li J, Ma S, Srivastava T (2022) A Riemannian ADMM. *https://arxiv.org/abs/2211.02163* .

[41] Li X, Chen S, Deng Z, Qu Q, Zhu Z, So AMC (2021) Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods. *SIAM J. Optim.* 31:1605–1634.

[42] Liu YF, Dai YH, Ma S (2015) Joint power and admission control: Non-convex L_q approximation and an effective polynomial time deflation approach. *IEEE Trans. Signal Process.* 63:3641–3656.

[43] Liu YF, Ma S, Dai YH, Zhang S (2016) A smoothing SQP framework for a class of composite L_q minimization over polyhedron. *Math. Program.* 158:467–500.

[44] Nguyen CT, Alcantara JH, Okuno T, Takeda A, Chen JS (2021) Unified smoothing approach for best hyperparameter selection problem using a bilevel optimization strategy. *arXiv: 2110.12630v1* .

[45] Qu Q, Sun J, Wright J (2016) Finding a sparse vector in a subspace: Linear sparsity using alternating directions. *IEEE Trans. Inf. Theory* 62:5855–5880.

[46] Qu Q, Zhu Z, Li X, Tsakiris MC, Wright J, Vidal R (2020) Finding the sparsest vectors in a subspace: Theory, algorithms, and applications. *https://arxiv.org/abs/2001.06970* .

[47] Rockafellar RT, Wets RJB (1998) *Variational Analysis* (Springer, New York).

[48] Shang F, Cheng J, Liu Y, Luo ZQ, Lin Z (2018) Bilinear factor matrix norm minimization for robust PCA: algoirthms and applications. *IEEE Trans. Pattern Anal.* 40:2066–2080.

[49] Spielman DA, Wang H, Wright J (2012) Exact recovery of sparsely-used dictionaries. *Conference on Learning Theory.*

[50] Sun J, Qu Q, Wright J (2017) Complete dictionary recovery over the sphere I: overview and the geometric picture. *IEEE Trans. Inform. Theory* 63:853–884.

[51] Tsakiris MC, Vidal R (2018) Dual principal component pursuit. *J. Mach. Learn. Res.* 19:1–49.

[52] Wang B, Ma S, Xue L (2022) Riemannian stochastic proximal gradient methods for nonsmooth optimization over the Stiefel manifold. *J. Mach. Learn. Res.* 23:1–33.

[53] Wang Z, Liu B, Chen S, Ma S, Xue L, Zhao H (2021) A manifold proximal linear method for sparse spectral clustering with application to single-cell RNA sequencing data analysis. *INFORMS J. Optim.* 4(2):200–214.

[54] Whitney H (1934) Analytic extensions of differentiable functions defined in closed sets. *Trans. Amer. Math. Soc.* 36:63–89.

[55] Xu M, Ye JJ, Zhang L (2015) Smoothing SQP mehtods for solving degenerate nonsmooth constrained optimization problems with applications to bilevel programs. *SIAM J. Optim.* 25:1388–1410.

[56] Yang W, Zhang LH, Song R (2014) Optimality conditions for the nonlinear programming problems on Riemannian manifolds. *Pacific J. Optim.* 10:415–434.

[57] Zeng C, Wu C, Jia R (2019) Non-Lipschitz models for image restoration with impulse noise removel. *SIAM J. Imaging Sci.* 12:420–458.

[58] Zhang C, Chen X (2009) Smoothing projected gradient method and its application to stochastic linear complementarity problem. *SIAM J. Optim.* 20:627–649.

[59] Zhang C, Chen X (2020) A smoothing active set method for linearly constrained non-Lipschitz nonconvex optimization. *SIAM J. Optim.* 30:1–30.

[60] Zhou Y, Bao C, Ding C, Zhu J (2022) A semi-smooth Newton based augmented Lagrangian method for nonsmooth optimization on matrix manifolds. *to appear in Math. Program., https://doi.org/10.1007/s10107-022-01898-1* .

[61] Zhu H, Zhang X, Chu D, Liao L (2017) Nonconvex and nonsmooth optimization with generalized orthogonality constraints: An approximate augmented Lagrangian method. *J. Sci. Comput.* 72:331–372.

[62] Zhu Z, Ding T, Robinson DP, Tsakiris MC, Vidal R (2019) A linearly convergent method for non-smooth non-convex optimization on the Grassmannian with applications to robust subspace and dictionary learning. *NeurIPS.*

[63] Zhu Z, Wang Y, Robinson DP, Naiman D, Vidal R, Tsakiris MC (2018) Dual principal component pursuit: Improved analysis and efficient algorithms. *NeurIPS.*