A RIEMANNIAN SMOOTHING STEEPEST DESCENT METHOD FOR NON-LIPSCHITZ OPTIMIZATION ON SUBMANIFOLDS *

CHAO ZHANG†, XIAOJUN CHEN‡, AND SHIQIAN MA§

Abstract. In this paper, we propose a Riemannian smoothing steepest descent method to minimize a nonconvex and non-Lipschitz function on submanifolds. The generalized subdifferentials on Riemannian manifold and the Riemannian gradient sub-consistency are defined and discussed. 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. Under the Riemannian gradient sub-consistency condition, we also prove that any accumulation point is a Riemannian limiting stationary point of the original non-Lipschitz problem. Numerical experiments are conducted to demonstrate the efficiency of the proposed method.

Key words. Riemannian submanifold, Non-Lipschitz, Smoothing steepest descent method, Riemannian generalized subdifferentials, Riemannian gradient sub-consistency

AMS subject classifications. 65K10, 90C26, 90C46

1. Introduction.

We consider the Riemannian optimization problem

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

where $\mathcal{M}$ is a complete Riemannian submanifold of $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ is a proper lower semi-continuous function which may be nonsmooth and non-Lipschitzian. Such problems arise in a variety of applications in signal processing, computer vision, and data mining [3, 5, 39].

Many classical algorithms for unconstrained and smooth optimization have been extended from the Euclidean space to the Riemannian manifolds, such as the gradient descent algorithm, the conjugate gradient algorithm, the quasi-Newton algorithm and the trust region method [1, 2, 26]. Recently, Riemannian optimization with a nonsmooth but Lipschitz continuous objective function has been considered in the literature. Here the smoothness and Lipschitz continuity are interpreted when the function in question is considered in the ambitent Euclidean space. The Clarke subdifferential of functions over manifolds has been defined and its properties have been discussed in [24]. Several algorithms have been proposed based on the notion of Clarke subdifferential. For example, Hosseini and Uschmajew [25] proposed the Riemannian gradient sampling algorithm. This algorithm approximats 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 [22] is a steepest descent method where the descent directions are obtained by

*Submitted to the editors on April 8, 2021.

Funding: C. Zhang was supported in part by 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-1953210 and CCF-2007797, and UC Davis CeDAR (Center for Data Science and Artificial Intelligence Research) Innovative Data Science Seed Funding Program.

†Department of Applied Mathematics, Beijing Jiaotong University, Beijing 100044, China. (zc.njtu@163.com).

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

§Department of Mathematics, University of California, Davis, CA 95616, USA. (Corresponding author. sqma@ucdavis.edu).

a computable approximation of the ϵ-subdifferential. The line search algorithms [23] 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 Clarke stationary point. Other methods for nonsmooth optimization over Riemannian manifolds include the Riemannian subgradient method [31], the Riemannian ADMM [29, 28], the manifold proximal gradient method [11, 27, 12, 41], manifold proximal point method [10], manifold proximal linear method [42], and manifold augmented Lagrangian method [13, 50, 49].

To the best of our knowledge, there do not exist optimization algorithms for solving Riemannian optimization problems with general non-Lipschitz objective functions, although the Riemannian generalized subdifferentials have been studied for nonsmooth and non-Lipschitz optimization [30]. Non-Lipschitz optimization in Euclidean space finds many important applications, including but not limited to, finding sparse solutions in singal processing and data mining [16, 19, 32, 33, 37], and neat edge in image restoration [6, 17, 46]. Smoothing methods with a proper updating scheme for the smoothing parameter are efficient for solving large-scale nonsmooth optimization in Euclidean space [14, 17, 18, 20, 47, 48]. With a fixed smoothing parameter, one solves the smoothed problem to update the iterate. Certain strategy is then applied to decide weather 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. The gradient consistency naturally holds for smoothing functions arising in various real applications with nonsmooth and Lipschitz objective functions. Smoothing methods have been widely used to solve unconstrained non-Lipschitz optimization problems [18, 20], and constrained non-Lipschitz optimization with feasible region being convex sets [47, 48]. However, minimizing a non-Lipschitz function on a nonconvex set has not been widely considered in the literature. In [16], an augmented Lagrangian method for non-Lipschitz nonconvex programming was proposed where the constraint set is nonconvex.

In this paper, we extend the smoothing steepest descent method in Euclidean space to Riemannian submanifolds. The smoothing steepest descent method is a special case of the smoothing projected gradient method for unconstrained nonsmooth optimization [47]. Our Riemannian smoothing steepest descent method (RSSD) uses the Riemannian gradient of the smoothing function in each iteration. Therefore, we do not need to sample points around the current point to get (sub)gradient information of the current point. This avoids the vector transport comparing with existing gradient-type algorithms such as the Riemannian gradient sampling algorithm [25] and the Riemannian ϵ-subgradient algorithm [22]. Our RSSD is easy to implement and can be shown to converge to a stationary point of the Riemannian optimization with non-Lipschitz objective.

The rest of this paper is organized as follows. In Section 2, we give a brief review on some basic concepts and properties relating to Riemannian manifold, and the generalized subdifferentials for non-Lipschitz functions in Euclidean space. In Section 3, we define the generalized subdifferentials for non-Lipschitz functions on Riemannian submanifolds and discuss their properties. We also define and discuss the Riemannian gradient sub-consistency that is essential to show that any accumulation point of our proposed RSSD method is a Riemannian limiting stationary point. In Section 4, we propose our RSSD method and analyze its convergence behavior. In Section 5, we conduct numerical experiments on two imporant applications: finding a sparse vector

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, $\mathcal{M}$ denotes a finite dimensional complete Riemannian submanifold embeded in an Euclidean space. 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 $\langle \xi, \eta \rangle_x = \mathrm{Tr}(\xi^\top \eta)$, where $\mathrm{T}_x\mathcal{M}$ denotes the tangent space of $\mathcal{M}$ at x, and $\mathrm{Tr}(Z)$ denotes the trace of matrix Z. 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 set of all tangent vectors: $\mathrm{T}\mathcal{M} := \bigcup_{x\in\mathcal{M}} \mathrm{T}_x\mathcal{M}$. 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 δ. For subset $D \subseteq \mathbb{R}^n$, a function $h \in C^1(D)$ means that h is smooth on D.

An important concept in Riemannian optimization is the retraction operation and it is defined below.

DEFINITION 2.1. *(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 properties. Here R_x denotes the restriction of R to the tangent space $T_x\mathcal{M}$.*

(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., [23]). We now review some important concepts and properties related to generalized subgradients, subdifferentials and subderivatives of non-Lipschitz functions in Euclidean space $\mathbb{R}^n$. They are specializations of [36, Definitions 8.3, 8.1] to our setting (note that in our case the function f is finite-valued.)

DEFINITION 2.2. *(Subgradients). We consider a proper lower semi-continuous function $f : \mathbb{R}^n \to \mathbb{R}$. For a vector $v \in \mathbb{R}^n$, we say that*

(i) v is a regular subgradient of f at $\bar{x}$, written as $v \in \hat{\partial} f(\bar{x})$, if

$$f(x) \ge f(\bar{x}) + \langle v, x - \bar{x} \rangle + o(\|x - \bar{x}\|),$$

or equivalently

$$\liminf_{x \to \bar{x},\ x \neq \bar{x}} \frac{f(x) - f(\bar{x}) - \langle v, x - \bar{x} \rangle}{\|x - \bar{x}\|} \ge 0;$$

(ii) v is a (general) limiting subgradient of f at $\bar{x}$, written as $v \in \partial f(\bar{x})$, if there exist $(x^\nu, f(x^\nu)) \to (\bar{x}, f(\bar{x}))$ and $v^\nu \in \hat{\partial} f(x^\nu)$ with $v^\nu \to v$;

(iii) v is a horizontal subgradient of f at $\bar{x}$, written as $v \in \partial^\infty f(\bar{x})$, if the same conditions in (ii) hold, except that instead of $v^\nu \to v$ one has $\lambda^\nu v^\nu \to v$ for some sequence $\lambda^\nu \downarrow 0$.

Here $\hat{\partial} f(\bar{x})$, $\partial f(\bar{x})$, and $\partial^\infty f(\bar{x})$ are called the regular (Fréchet), limiting, and horizontal subdifferentials of f at $\bar{x}$, respectively. According to [36], $\partial^\circ f(\bar{x})$ is called the Clarke subdifferential if

$$\partial^\circ f(\bar{x}) = \mathrm{conv}\{\partial f(\bar{x}) + \partial^\infty f(\bar{x})\},$$

where conv is the convex hull.

DEFINITION 2.3. *(Subderivative). For a proper lower semi-continuous function $f : \mathbb{R}^n \to \mathbb{R}$, the subderivative function $df(\bar{x}) : \mathbb{R}^n \to \bar{\mathbb{R}}$ is defined as*

$$df(\bar{x})(\bar{w}) := \liminf_{\tau \downarrow 0,\ w \to \bar{w}} \frac{f(\bar{x} + \tau w) - f(\bar{x})}{\tau}.$$

We have the two equivalent characterizations for the regular subdifferential in the following two propositions, coming from [36, Exercise 8.4, pp. 301; Proposition 8.5, pp. 302].

PROPOSITION 2.4. *(Regular subgradient from subderivative). It holds that*

$$\hat{\partial} f(\bar{x}) = \{v \mid \langle v, w \rangle \leq df(\bar{x})(w) \quad \text{for all } w\}.$$

PROPOSITION 2.5. *(Variational description of regular subgradients). A vector v belongs to $\hat{\partial} f(\bar{x}) \iff$ in some neighborhood of $\bar{x}$, there is a function $h \leq f$ 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 $\bar{x}$.*

In the case that $f : \mathcal{M} \to \mathbb{R}$ is a nonsmooth but locally Lipschitz continuous function, the Clarke subdifferential has also been studied and used in analyzing the convergence of algorithms, see e.g., [25]. Let

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

The Riemannian Clarke subdifferential, denoted by $\partial^\circ_{\mathcal{R}} f(x)$, is defined as [25]

$$\partial^\circ_{\mathcal{R}} f(x) := \operatorname{conv}\left\{\lim_{\ell \to \infty} \operatorname{grad} f(x_\ell) \mid x_\ell \to x,\ x_\ell \in \Omega_f\right\}, \tag{2.1}$$

where grad dentoes the Riemannian gradient. An alternative definition of $\partial^\circ_{\mathcal{R}} f(x)$ [25] relying on the definition of subdifferential on linear spaces is

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

for any retraction R.

A definition of generalized subdifferentials for nonsmooth non-Lipschitz function on manifold is given as follows by [30, Definition 3.1].

DEFINITION 2.6. *Let $f : \mathcal{M} \to \mathbb{R}$ be any lower semicontinuous function. The Riemannian Fréchet subdifferential of f at $x \in \mathcal{M}$ is defined as*

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

where $dh(x)$ is the differential of h at $x \in \mathcal{M}$. The Riemannian limiting subdifferential of f at $x \in \mathcal{M}$ is defined as

$$\partial f(x) := \{\lim_{\ell \to \infty} v_\ell^* \mid v_\ell^* \in \partial_F f(x_\ell),\ (x_\ell, f(x_\ell)) \to (x, f(x))\}.$$

The Riemannian horizontal subdifferential of f at $x \in \mathcal{M}$ is defined as

$$\partial^\infty f(x) := \{\lim_{\ell \to \infty} t_\ell v_\ell^* \mid v_\ell^* \in \partial_F f(x),\ (x_\ell, f(x_\ell)) \to (x, f(x)) \text{ and } t_\ell \downarrow 0\}.$$

Let h be a $C^1(\mathcal{M})$ function at x. The differential of h at x, $dh(x) \in \mathrm{T}_x\mathcal{M}^*$, is an element of $\mathrm{T}_x\mathcal{M}^*$, which is defined as

$$dh(x)(v) = \langle \mathrm{grad} h(x), v\rangle, \quad \forall v \in \mathrm{T}_x\mathcal{M},$$

where $\mathrm{grad} h(x)$ is the Riemannian gradient of h at $x \in \mathcal{M}$.

We use the following definition of a smoothing function on $\mathbb{R}^n$ as in [48].

DEFINITION 2.7. *(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 $\mu \in \mathbb{R}_{++}$,*

$$\lim_{z\to x,\ \mu\downarrow 0} \tilde{f}(z,\mu) = f(x), \tag{2.2}$$

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

$$|\tilde{f}(x,\mu) - f(x)| \le \kappa\omega(\mu) \quad \textit{with} \quad \lim_{\mu\downarrow 0} \omega(\mu) = 0. \tag{2.3}$$

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 2.8. 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

$$s_\mu(t) = \begin{cases} |t| & \text{if } |t| \ge \frac{\mu}{2} \\ \frac{t^2}{\mu} + \frac{\mu}{4} & \text{if } |t| < \frac{\mu}{2}, \end{cases} \tag{2.4}$$

with $\kappa = \frac{1}{4}$ and $\omega(\mu) = \mu$ in (2.3).

We refer to [14] for more examples of smoothing functions. For non-Lipschitz term $|t|^p$ where $0 < p < 1$, its smoothing function can be defined as $(s_\mu(t))^p$, with $\kappa = (\frac{1}{4})^p$ and $\omega(\mu) = \mu^p$ in (2.3).

3. Riemannian generalized subdifferentials and Riemannian gradient sub-consistency. In this section, we define and discuss several genearlized subdifferentials, Riemannian gradient sub-consistency of proper lower semicontinuous functions, and related stationary points of (1.1). These concepts play important roles in the convergence analysis of our proposed method in the next section.

3.1. Riemanian generalized subdifferentials. Motivated by the generalized Clarke subdifferential on Riemannian manifold in (2.1), and the generalized subdifferentials for a lower semicontiunouos function on Riemannian manifold in Definition 2.6 given by Ledyaev and Zhu [30], we define the generalized subdifferentials for lower semicontinuous functions. Similar as [45] for the nonsmooth but Lipschitz case, we define the generalized subdifferentials on the tangent space, not on the cotangent space as in Definition 2.6 by [30]. Since the Riemannian gradient of a differentiable function is defined on the tangent space, from the computational point of view, we find that it is more reasonable to define the generalized subdifferential of a nonsmooth function on the tangent space.

DEFINITION 3.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a lower semicontinuous function. The Riemannian Fréchet subdifferential of f at $x \in \mathcal{M}$ is defined as*

$$\hat{\partial}_{\mathcal{R}} f(x) := \{\mathrm{grad}\, h(x) \mid \exists\ \delta > 0 \textit{ such that } h \in C^1(B_{x,\delta}) \textit{ and } f - h \textit{ attains a local minimum at } x \textit{ on } \mathcal{M}\}. \tag{3.1}$$

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

$$\partial_{\mathcal{R}} f(x) := \{\lim_{\ell\to\infty} v_\ell^* \mid v_\ell^* \in \hat{\partial}_{\mathcal{R}} f(x_\ell), (x_\ell, f(x_\ell)) \to (x, f(x))\}. \tag{3.2}$$

The Riemannian horizontal subdifferential of f at $x \in \mathcal{M}$ is defined as

$$\partial^\infty_{\mathcal{R}} f(x) := \{\lim_{\ell\to\infty} t_\ell v_\ell^* \mid v_\ell^* \in \hat{\partial}_{\mathcal{R}} f(x_\ell), (x_\ell, f(x_\ell)) \to (x, f(x)) \text{ and } t_\ell \downarrow 0\}. \tag{3.3}$$

The Riemannian regular subdifferential is essentially only related to the local property of h. By Whitney extension theorem [43], any smooth function on $B_{x,\delta} \cap \mathcal{M}$ can be extended on the whole Euclidean space $\mathbb{R}^n$. When $\mathcal{M} = \mathbb{R}^n$, the Riemannian Fréchet, limiting, and horizontal subdifferentials coincide with the usual Fréchet, limiting, and horizontal subdifferentials in $\mathbb{R}^n$. When f is Lipschitz continuous, we know that $\partial^\infty_{\mathcal{R}} f(x) = \{0\}$, and then the Riemannian Clarke subdifferential is

$$\partial^\circ_{\mathcal{R}} f(x) = \mathrm{conv}\{\partial_{\mathcal{R}} f(x)\},$$

which is widely used in the Riemannian optimization literature [22, 24, 25, 23].

We make a brief comparison and build up the relation between Definition 3.1 and Definition 2.6. First we note that $\hat{\partial}_{\mathcal{R}} f(x) \subseteq \mathrm{T}_x\mathcal{M}$, and $\partial_F f(x) \subseteq \mathrm{T}_x\mathcal{M}^*$. There is a one-to-one correspondence between element $\mathrm{grad} h(x) \in \hat{\partial}_{\mathcal{R}} f(x)$ and $dh(x) \in \partial_F f(x)$. That is, for any $dh(x) \in \partial_F f(x)$, there is a unique $\mathrm{grad}\, h(x) \in \hat{\partial}_{\mathcal{R}} f(x)$ that corresponds to it. Moreover, we have

$$dh(x)(\cdot) := \langle \mathrm{grad}\, h(x), \cdot \rangle, \tag{3.4}$$

because for any $x \in \mathcal{M}$ and $\xi \in \mathrm{T}_x\mathcal{M}$,

$$\langle \mathrm{grad}\, h(x), \xi \rangle = dh(x)(\xi) = \left.\frac{dh(\gamma(t))}{dt}\right|_{t=0},$$

where γ is a curve on $\mathcal{M}$ with $\gamma(0) = x$ and $\dot\gamma(0) = \xi$.

Using Definition 3.1, and the facts that $\mathcal{M}$ is a submanifold embedded in $\mathbb{R}^n$ and $h \in C^1(B_{x,\delta})$, we have

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

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,

$$\begin{aligned}\partial_{\mathcal{R}} f(x) = \{\mathrm{Proj}_{\mathrm{T}_x\mathcal{M}} \nabla h(x) \mid \exists\ \delta > 0 \text{ such that } h \in C^1(B_{x,\delta}) \text{ and} \\ f - h \text{ attains a local minimum at } x \text{ on } \mathcal{M}\}.\end{aligned} \tag{3.5}$$

Note that for any $v \in \hat{\partial} f(x)$, according to Proposition 2.5, 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 (3.5), indicates that

$$\hat{\partial}_{\mathcal{R}} f(x) \supseteq \{\mathrm{Proj}_{\mathrm{T}_x\mathcal{M}}\, v \mid v \in \hat{\partial} f(x)\}. \tag{3.6}$$

By using Definition 3.1, we have $\hat{\partial}_{\mathcal{R}} f(x) \subseteq \partial_{\mathcal{R}} f(x)$.

We provide an equivalent characterization of $\hat{\partial}_{\mathcal{R}} f(x)$ below.

PROPOSITION 3.2. *Let R be any given retraction as defined in Definition 2.1. Then $v \in \hat{\partial}_{\mathcal{R}} f(x)$ if and only if $v \in \mathrm{T}_x\mathcal{M}$ and the following holds*

$$f \circ R(\eta_x) \geq f \circ R(0_x) + \langle v, \eta_x \rangle + o(\|\eta_x\|), \quad \forall \eta_x \in \mathrm{T}_x\mathcal{M}. \tag{3.7}$$

Proof. By Definition 3.1, $v \in \hat{\partial}_{\mathcal{R}} f(x)$ if and only if there exists $h \in C^1(B_{x,\delta})$ for some $\delta > 0$ such that $f - h$ attains local minimum at x on $\mathcal{M}$, and $\operatorname{grad} h(x) = v$. The latter statement is equivalent to the fact that $f \circ R_x - h \circ R_x$ obtains local minimum at 0_x in $\mathrm{T}_x\mathcal{M}$. By Definition 2.1 and the fact that $\mathcal{M}$ is endowed with a Riemannian metric, we have

$$\operatorname{grad}(h \circ R_x)(0_x) = \operatorname{grad} h(x) = v.$$

This implies that $v \in \hat{\partial}(f \circ R_x)(0_x)$, when considering $\mathrm{T}_x\mathcal{M}$ is an Euclidean space itself. By Definition 2.2, we know that $v \in \mathrm{T}_x\mathcal{M}$ satisfies (3.7). ∎

According to Proposition 3.2, we easily find that if $\bar{x}$ is a local minimizer of f on $\mathcal{M}$, then $0 \in \hat{\partial}_{\mathcal{R}} f(\bar{x})$.

Definition 3.3. *A point $x \in \mathcal{M}$ is called a limiting stationary point of the Riemannian optimization problem (1.1), if $0 \in \partial_{\mathcal{R}} f(x)$.*

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

Definition 3.4. *The subdifferential of f associated with $\tilde{f}$ at $x \in \mathbb{R}^n$ is*

$$G_{\tilde{f}}(x) = \{u \in \mathbb{R}^n \ : \ \nabla_x \tilde{f}(z_k, \mu_k) \to u \quad \text{for some } z_k \to x,\ \mu_k \downarrow 0\}, \tag{3.8}$$

and the Riemannian subdifferential of f associated with $\tilde{f}$ at $x \in \mathcal{M}$ is

$$G_{\tilde{f},\mathcal{R}}(x) = \{v \in \mathbb{R}^n : \operatorname{grad} \tilde{f}(z_k, \mu_k) \to v \text{ for some } z_k \in \mathcal{M},\ z_k \to x,\ \mu_k \downarrow 0\}. \tag{3.9}$$

Remark 3.5. Here $u \in G_{\tilde{f}}(x)$ and $v \in G_{\tilde{f},\mathcal{R}}(x)$ are vectors in $\mathbb{R}^n$ whose entries are finite, i.e., they are not ∞ or $-\infty$.

Example 3.6. 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 (2.4), we have

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

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 3.7. *We say that $x^* \in \mathcal{M}$ is a stationary point of f associated with $\tilde{f}$ on the submanifold $\mathcal{M}$, if*

$$\liminf_{x \to x^*,\ x \in \mathcal{M},\ \mu \downarrow 0} \| \operatorname{grad} \tilde{f}(x, \mu)\| = 0. \tag{3.10}$$

The following result is an extension of Proposition 3.4 of [48] from $\mathbb{R}^n$ to the submanifold $\mathcal{M}$.

Proposition 3.8. *For any smoothing function $\tilde{f}$ of f as defined in Definition 2.7, if $x^* \in \mathcal{M}$ is a local minimizer of f on the submanifold $\mathcal{M}$, then x^* is a stationary point of f associated with $\tilde{f}$ on the submanifold $\mathcal{M}$.*

Proof. Note that $x^* \in \mathcal{M}$ is a local minimizer of f on the submanifold $\mathcal{M}$. Since minima are preserved by composition with diffeomorphisms (see, e.g., the proof of (2) $\Rightarrow$ (1) in Proposition 2.2 of [4]), 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,

$$\hat{f}_\mu(0_{x^*}) \leq \hat{f}_\mu(\eta) + 2\kappa\omega(\mu), \quad \text{for any } \eta \in B_{0_{x^*},\delta}. \tag{3.11}$$

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

$$\hat{f}_\mu(0_{x^*}) = \hat{f}_\mu(\eta_\mu) + \langle \operatorname{grad} \hat{f}_\mu(\eta_\mu), -\sqrt{\omega(\mu)}\eta_z \rangle + o(\sqrt{\omega(\mu)}\|\eta_z\|). \tag{3.12}$$

Substituting (3.12) into the left hand side of (3.11), and replacing η by η_μ with μ that is sufficiently small, we get

$$\sqrt{\omega(\mu)}\langle \operatorname{grad} \hat{f}_\mu(\eta_\mu), -\eta_z \rangle + o(\sqrt{\omega(\mu)}\|\eta_z\|) \leq 2\kappa\omega(\mu).$$

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

$$\limsup_{\mu \downarrow 0} \langle \operatorname{grad} \hat{f}_\mu(\eta_\mu), -\eta_z \rangle \leq 0,$$

which implies that

$$\liminf_{\eta \to 0_{x^*},\ \eta \in \mathrm{T}_{x^*}\mathcal{M},\ \mu \downarrow 0} \langle \operatorname{grad} \hat{f}_\mu(\eta), -\eta_z \rangle \leq 0. \tag{3.13}$$

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 submanifold. 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 [45]). Then

$$\operatorname{grad} \hat{f}_\mu(\eta) = \sum_{i=1}^{d} \lambda_i^\mu E(e_i), \tag{3.14}$$

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 (3.13) by (3.14), and substituting η_z in (3.13) by $\eta_z^{(i,1)}$ and $\eta_z^{(i,2)}$, respectively, we obtain

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

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

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

That is,

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

and hence x^* is a stationary point of f associated with $\tilde{f}$ on $\mathcal{M}$ as desired. ∎

We will show later that any accumulation point of our proposed RSSD method is a stationary point of f associated with $\tilde{f}$ on $\mathcal{M}$.

3.2. Riemannian gradient sub-consistency. Now we define the Riemannian gradient sub-consistency of $\tilde{f}$ at $x \in \mathcal{M}$, which makes 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 will be essential to show that any accumulation point of the RSSD method developed in this paper is also a Riemannian limiting stationary point.

DEFINITION 3.9. *A smoothing function $\tilde{f}$ of the function f is said to satisfy the gradient sub-consistency at $x \in \mathbb{R}^n$ if*

$$G_{\tilde{f}}(x) \subseteq \partial f(x), \tag{3.15}$$

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

$$G_{\tilde{f},\mathcal{R}}(x) \subseteq \partial_{\mathcal{R}} f(x). \tag{3.16}$$

We say that $\tilde{f}$ satisfies the gradient sub-consistency on $\mathbb{R}^n$ if (3.15) holds for any $x \in \mathbb{R}^n$, and that $\tilde{f}$ satisfies the Riemannian gradient sub-consistency on $\mathcal{M}$ if (3.16) holds for any $x \in \mathcal{M}$.

If the inclusion is substituted by the equality in (3.15) for any $x \in \mathbb{R}^n$, then $\tilde{f}$ satisfies the gradient consistency on $\mathbb{R}^n$. Clearly, the gradient consistency indicates the gradient sub-consistency. The gradient consistency on $\mathbb{R}^n$ has been well studied in smoothing methods for nonsmooth optimization. For nonsmooth but Lipschitz function f, it has been shown that the gradient consistency property holds for various smoothing functions in many real applications [7, 8, 14, 44, 47].

The following proposition demonstrates that if the gradient sub-consistency of $\tilde{f}$ in $\mathbb{R}^n$ holds, then the Riemannian gradient sub-consistency of $\tilde{f}$ holds on $\mathcal{M}$, provided that f is locally Lipschitz.

PROPOSITION 3.10. *Let f be a locally Lipschitz function with $\tilde{f}$ being a smoothing function of f. If the gradient sub-consistency of $\tilde{f}$ holds on $\mathbb{R}^n$, then the Riemannian gradient sub-consistency on $\mathcal{M}$ holds.*

Proof. For any $x \in \mathcal{M}$, 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. Then there exist subsequences $x_{\mu_k} \in \mathcal{M}$, $x_{\mu_k} \to x$, $\mu_k \downarrow 0$ as $k \to \infty$, and a vector $u \in G_{\tilde{f}}(x)$ such that

$$u = \lim_{x_{\mu_k} \to x,\ x_{\mu_k} \in \mathcal{M},\ \mu_k \downarrow 0} \nabla_x \tilde{f}(x_{\mu_k}, \mu_k), \tag{3.17}$$

and

$$\begin{aligned} v &= \lim_{x_{\mu_k} \to x,\ x_{\mu_k} \in \mathcal{M},\ \mu_k \downarrow 0} \operatorname{grad} \tilde{f}(x_{\mu_k}, \mu_k) \\ &= \lim_{x_{\mu_k} \to x,\ x_{\mu_k} \in \mathcal{M},\ \mu_k \downarrow 0} \operatorname{Proj}_{\mathrm{T}_{x_{\mu_k}}\mathcal{M}} \nabla_x \tilde{f}(x_{\mu_k}, \mu_k), \\ &= \operatorname{Proj}_{\mathrm{T}_x\mathcal{M}} u. \end{aligned} \tag{3.18}$$

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 (3.17), and $\| \operatorname{Proj}_{\mathrm{T}_{x_{\mu_k}}\mathcal{M}} u - \operatorname{Proj}_{\mathrm{T}_x\mathcal{M}} u \| \to 0$ because $S(x) := \mathrm{T}_x\mathcal{M}$ is continuous and convex-valued (i.e., $S(x)$ is a convex set for each fixed x), and $\operatorname{Proj}_{S(x)}$ is continuous according to Example 5.57 of [36].

Since the gradient sub-consistency $G_{\tilde{f}} \subseteq \partial f(x)$ holds, we know that $u \in \partial f(x)$. By the definition of the 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 the Riemannian Fréchet subdifferential in (3.6), we have

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

and using the same arguments of proving (3.18), 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 Riemannian gradient sub-consistency holds. □

For non-Lipschitz functions, we first use the smoothing function $s_\mu(t)$ of $|t|^p$ to illustrate that the gradient consistency on $\mathbb{R}^n$ holds. It is known that $\partial f(0) = (-\infty, \infty)$, and $\partial f(t) = p|t|^{p-1}\operatorname{sign}(t)$. This, combined with Example 3.6, yields that

$$G_{\tilde{f}}(0) = (-\infty, \infty) = \partial f(0),$$

and for any point $t \neq 0$,

$$G_{\tilde{f}}(t) = p|t|^{p-1}\text{sign}(t) = \partial f(t).$$

Thus the smoothing function $\tilde{f}$ of the non-Lipschitz function $f = |t|^p$ satisfies the gradient consistency on $\mathbb{R}^n$.

Furthermore, we consider a class of non-Lipschitz optimization on submanifold $\mathcal{M}$ as follows

$$\min_{x\in\mathcal{M}} \ f(x) := \hat{f}(x) + \lambda\|Bx\|_p^p, \tag{3.19}$$

where $\hat{f}$ is a smooth function, $\mathcal{M}$ is a submanifold, $B \in \mathbb{R}^{m\times n}$ is a given matrix of full column rank, and $p \in (0,1)$, and $\lambda > 0$ are given constants. Many applications can be formulated in the form of (3.19), such as finding the sparsest vector 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 $|t|$ satisfying Definition 2.7. Then the function

$$\tilde{f}(x,\mu) = \hat{f}(x) + \lambda\sum_{i=1}^{m}[\tilde{s}_\mu((Bx)_i)]^p \tag{3.20}$$

is a smoothing function of f defined in (3.19). We then have the following proposition.

PROPOSITION 3.11. *The smoothing function $\tilde{f}$ that is constructed in (3.20) for the non-Lipschitz objective function f in (3.19) satisfies the gradient sub-consistency on $\mathbb{R}^n$, and the Riemannian gradient sub-consistency on the submanifold $\mathcal{M}$.*

Proof. For any $x \in \mathbb{R}^n$, let us denote the index sets

$$I_1 := \{i \mid (Bx)_i \neq 0\}, \quad \text{and} \quad I_2 := \{i \mid (Bx)_i = 0\},$$

and correspondingly for any $z \in \mathbb{R}^n$, define

$$f_1(z) := \lambda\sum_{i\in I_1}|(Bz)_i|^p, \quad \text{and} \quad f_2(z) := \lambda\sum_{i\in I_2}|(Bz)_i|^p,$$
$$\tilde{f}_1(z,\mu) := \lambda\sum_{i\in I_1}[\tilde{s}_\mu((Bz)_i)], \quad \text{and} \quad \tilde{f}_2(z,\mu) := \lambda\sum_{i\in I_2}[\tilde{s}_\mu((Bz)_i)].$$

Clearly

$$\lambda\|Bz\|_p^p = f_1(z) + f_2(z), \quad \text{and} \quad \tilde{f}(z,\mu) = \hat{f}(z) + \tilde{f}_1(z,\mu) + \tilde{f}_2(z,\mu).$$

For any $u \in G_{\tilde{f}}(x)$, we know that there exist sequence $z_k \to x$, and $\mu_k \downarrow 0$ as $k \to \infty$ such that

$$u = \lim_{z_k\to x,\ \mu_k\downarrow 0} \nabla_x\tilde{f}(z_k,\mu_k). \tag{3.21}$$

It is clear that

$$\nabla_x\tilde{f}(z_k,\mu_k) = \nabla\hat{f}(z_k) + \nabla_x\tilde{f}_1(z_k,\mu_k) + \nabla_x\tilde{f}_2(z_k,\mu_k), \tag{3.22}$$

and

$$\lim_{k\to\infty}\nabla\hat{f}(z_k) = \nabla\hat{f}(x) \quad \text{and} \quad \lim_{k\to\infty}\nabla_x\tilde{f}_1(z_k,\mu_k) = \nabla f_1(x). \tag{3.23}$$

By direct computation,

$$\nabla_x \tilde{f}_2(z_k,\mu_k) = \sum_{i\in I_2} \lambda p\left(\tilde{s}_{\mu_k}((Bz_k)_i)\right)^{p-1}\left[\tilde{s}_{\mu_k}((Bz_k)_i)\right]' B_{i\cdot}^\top = B_{I_2}^\top Y_k. \tag{3.24}$$

Here $B_{i\cdot}$ is the i-th row of B, B_{I_2} is the submatrix of B defined by $B_{I_2} = (B_{i\cdot})_{i\in I_2}$, and

$$Y_k := Y_k(z_k,\mu_k) = \lambda p\left(\left(\tilde{s}_{\mu_k}((Bz_k)_i)\right)^{p-1}[\tilde{s}_{\mu_k}((Bz_k)_i)]'\right)_{i\in I_2} \in \mathbb{R}^{|I_2|},$$

with $|I_2|$ being the cardinality of the index set I_2. Let $N(C)$ be the null space of the matrix C and $N(C)^\perp$ be its orthogonal complement. It is known that Y_k can be uniquely written as

$$Y_k = Y_k^1 + Y_k^2, \quad \text{where } Y_k^1 \in N(B_{I_2}^\top),\ Y_k^2 \in N(B_{I_2}^\top)^\perp. \tag{3.25}$$

We claim that $\{Y_k^2\}$ is bounded along with $z_k \to x$, $\mu_k \downarrow 0$ as $k\to\infty$. Otherwise, there exists an infinite subsequence $K_1 \subseteq \{1,2,\ldots\}$ such that

$$\lim_{k\to\infty,\ k\in K_1} \|Y_k^2\| = \infty.$$

Because

$$\frac{Y_k^2}{\|Y_k^2\|} \in N(B_{I_2}^\top)^\perp, \quad \text{and } \left\{\frac{Y_k^2}{\|Y_k^2\|}\right\} \text{ is bounded,}$$

there exists an infinite subsequence $K_2 \subseteq K_1$ such that

$$\lim_{k\to\infty,\ k\in K_2} \frac{Y_k^2}{\|Y_k^2\|} = \bar{Y} \in N(B_{I_2}^\top)^\perp, \quad \text{with } \|\bar{Y}\| = 1. \tag{3.26}$$

This, together with (3.24) implies that $B_{I_2}^\top \bar{Y} \neq 0$, and

$$\begin{aligned}\lim_{k\to\infty,\ k\in K_2} \|\nabla_x \tilde{f}_2(z_k,\mu_k)\| &= \lim_{k\to\infty,\ k\in K_2} \|B_{I_2}^\top\left(Y_k^1 + Y_k^2\right)\| \\ &= \lim_{k\to\infty,\ k\in K_2} \left\|B_{I_2}^\top\left(\|Y_k^2\|\frac{Y_k^2}{\|Y_k^2\|}\right)\right\| = \infty.\end{aligned}$$

Hence, by using (3.22) and (3.23), we find $\|\nabla_x \tilde{f}(z_k,\mu_k)\| \to \infty$ as $k\to\infty, k\in K_2$, which contradicts to (3.21) that $u\in\mathbb{R}^n$ cannot have components tending to infinity.

From the boundedness of $\{Y_k^2\}$, we know that there exists an infinite subsequence $K_3 \subseteq \{1,2,\ldots\}$ such that

$$\lim_{k\to\infty,\ k\in K_3} Y_k^2 = \hat{Y} \in \mathbb{R}^{I_2}.$$

Hence

$$u = \lim_{z_k\to x,\ \mu_k\downarrow 0} \nabla_x \tilde{f}(z_k,\mu_k) = \nabla \hat{f}(x) + \nabla f_1(x) + B_{I_2}^\top \hat{Y}. \tag{3.27}$$

Let us define the function

$$h(z) = \hat{f}(z) + f_1(z) + \sum_{i\in I_2} \hat{Y}_i (Bz)_i.$$

Note that for any $\nu \in R$, there exists some $\delta > 0$ such that

$$|t|^p > \nu t \quad \text{for any } |t| \le \delta.$$

We can easily find that there exists a neighborhood $B_{x,\bar{\delta}}$ of x such that for any $z \in B_{x,\bar{\delta}}$, $h(z) \leq f(z)$, and $h(x) = f(x)$. Thus by Proposition 2.5, we have

$$\nabla h(x) = \nabla \hat{f}(x) + \nabla f_1(x) + B_{I_2}^{\top} \hat{Y} \in \hat{\partial} f(x) \subseteq \partial f(x).$$

This, combining with (3.27) yields $u \in \partial f(x)$. Since both $x \in \mathbb{R}^n$ and $u \in G_{\tilde{f}}(x)$ are arbitrary, we get that $\tilde{f}$ defined in (3.20) satisfies the gradient sub-consistency on $\mathbb{R}^n$.

Below we show that $\tilde{f}$ also satisfies the Riemannian gradient sub-consistency on the submanifold $\mathcal{M}$.

For any $x \in \mathcal{M}$, let $v \in G_{\tilde{f},\mathcal{R}}(x)$. Then there exists infinite sequence $z_k \in \mathcal{M}$, $z_k \to x$, $\mu_k \downarrow 0$ as $k \to \infty$ such that

$$\begin{aligned} v &= \lim_{z_k \to x,\ z_k \in \mathcal{M},\ \mu_k \downarrow 0} \operatorname{grad} \tilde{f}(z_k, \mu_k) \\ &= \lim_{z_k \to x,\ z_k \in \mathcal{M},\ \mu_k \downarrow 0} \operatorname{Proj}_{\mathrm{T}_{z_k}\mathcal{M}} \nabla_x \tilde{f}(z_k, \mu_k). \end{aligned} \tag{3.28}$$

If $\{\nabla_x \tilde{f}(z_k, \mu_k)\}$ is bounded, noting that $\tilde{f}$ satisfies the gradient sub-consistency on $\mathbb{R}^n$, and following the similar arguments in the proof of Proposition 3.10, we can show that $v \in \partial_{\mathcal{R}} f(x)$.

Otherwise, there exists an infinite subsequence $K \subseteq \{1, 2, \ldots\}$ such that

$$\{\|\nabla_x \tilde{f}(z_k, \mu_k)\|\}_{k \in K} \to \infty,$$

which indicates that $\{\|\nabla_x \tilde{f}_2(z_k, \mu_k)\|\}_{k \in K} \to \infty$ by noting (3.22) and (3.23). By (3.24) and (3.25), we know

$$\nabla_x \tilde{f}_2(z_k, \mu_k) = B_{I_2}^{\top} Y_k^2, \quad \text{where } Y_k^2 \in N(B_{I_2}^{\top})^{\perp}. \tag{3.29}$$

Hence

$$\{\|B_{I_2}^{\top} Y_k^2\|\}_{k \in K} \to \infty, \quad \text{and } \{\|Y_k^2\|\}_{k \in K} \to \infty. \tag{3.30}$$

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

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

which implies

$$\left\|\operatorname{Proj}_{\mathrm{T}_{z_k}\mathcal{M}} g_k^2\right\| \leq \left\|\operatorname{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}(g_k^1 + g_k^2)\right\| + \left\|g_k^1\right\|. \tag{3.31}$$

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 (3.31), we have

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

The two terms on the right-hand side of the above inequality are bounded by noting (3.28) and (3.23). Thus

$$\left\{\left\|\operatorname{Proj}_{\mathrm{T}_{z_k}\mathcal{M}} \nabla_x \tilde{f}_2(z_k, \mu_k)\right\|\right\} \text{ is bounded.} \tag{3.32}$$

Using (3.30), we may assume without loss of generality that

$$\lim_{k\to\infty,\ k\in K} \frac{Y_k^2}{\|Y_k^2\|} = \bar{Y} \neq 0.$$

We can write

$$B_{I_2}^\top \frac{Y_k^2}{\|Y_k^2\|} = d_k^1 + d_k^2, \quad \text{where } d_k^1 \in \mathrm{T}_{z_k}\mathcal{M}, d_k^2 \in (\mathrm{T}_{z_k}\mathcal{M})^\perp; \tag{3.33}$$

$$\nabla \hat{f}(z_k) + \nabla_x \tilde{f}_1(z_k, \mu_k) = a_k^1 + a_k^2, \quad \text{where } a_k^1 \in \mathrm{T}_{z_k}\mathcal{M}, a_k^2 \in (\mathrm{T}_{z_k}\mathcal{M})^\perp. \tag{3.34}$$

Here $(\mathrm{T}_{z_k}\mathcal{M})^\perp$ is the orthogonal complement of $\mathrm{T}_{z_k}\mathcal{M}$.

For any scalar $\alpha > 0$ and $g_k \in \mathbb{R}^n$, it is not difficult to show that

$$\mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}\, \alpha g_k = \alpha\, \mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}\, g_k. \tag{3.35}$$

Thus

$$\begin{aligned}
&\mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}\, \nabla_x \tilde{f}_2(z_k, \mu_k) \\
=&\,\mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}\, B_{I_2}^\top Y_k^2 = \mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}\, \|Y_k^2\| B_{I_2}^\top \frac{Y_k^2}{\|Y_k^2\|} \\
=&\|Y_k^2\|\, \mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}\, B_{I_2}^\top \frac{Y_k^2}{\|Y_k^2\|} = \|Y_k^2\|\, \mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}(d_k^1 + d_k^2) = \|Y_k^2\| d_k^1,
\end{aligned}$$

where the third equality employs (3.35). In view of (3.30) and (3.32), we get

$$\lim_{k\to\infty,\ k\in K} d_k^1 = 0.$$

By using (3.33) and (3.34), we get

$$\begin{aligned}
\mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}\, \nabla_x \tilde{f}(z_k, \mu_k) &= \mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}(\nabla \hat{f}(z_k) + \nabla_x \tilde{f}_1(z_k, \mu_k) + B_{I_2}^\top Y_k^2) \\
&= \mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}(a_k^1 + a_k^2 + d_k^1 + d_k^2) \\
&= \mathrm{Proj}_{\mathrm{T}_{z_k}\mathcal{M}}(a_k^1 + d_k^1) = a_k^1 + d_k^1.
\end{aligned}$$

Consequently,

$$\begin{aligned}
v &= \lim_{z_k\to x,\ z_k\in\mathcal{M},\ \mu_k\downarrow 0} \mathrm{Proj}_{T_{z_k}\mathcal{M}}\, \nabla_x \tilde{f}(z_k, \mu_k) \\
&= \lim_{k\to\infty,\ k\in K} (a_k^1 + d_k^1) = \lim_{k\to\infty,\ k\in K} a_k^1 \\
&= \lim_{z_k\to x,\ z_k\in\mathcal{M},\ \mu_k\downarrow 0} \mathrm{Proj}_{T_{z_k}\mathcal{M}}(\nabla \hat{f}(x) + \nabla f_1(x)).
\end{aligned}$$

We now define function $\bar{h}(z) = \hat{f}(z) + f_1(z)$. It is then easy to check that there exists a neighborhood $B_{x,\delta}$ for some $\delta > 0$ such that $\bar{h}(z) \leq f(z)$ with $\bar{h}(x) = f(x)$, and $\nabla \bar{h}(x) = \nabla \hat{f}(x) + \nabla f_1(x)$. Then by Proposition 2.5, $\nabla \bar{h}(x) \in \hat{\partial} f(x)$. Hence

$$v = \mathrm{Proj}_{T_{z_k}\mathcal{M}}(\nabla \hat{f}(x) + \nabla f_1(x)) \in \hat{\partial}_{\mathcal{R}} f(x) \subseteq \partial_{\mathcal{R}} f(x).$$

Therefore, $\tilde{f}$ satisfies the Riemannian gradient sub-consistency as desired. ∎

Algorithm 4.1 Riemannian smoothing steepest descent method (RSSD) for solving (1.1)

Input: $x_0 \in \mathcal{M}$, $\delta_{opt} \geq 0$, $\delta_0 > 0$, $\mu_{opt} \geq 0$, $\mu_0 > 0$, $\beta \in (0,1)$, $\bar{\alpha} > 0$, $\theta_\delta \in (0,1)$, $\theta_\mu \in (0,1)$.
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 $t_\ell := \beta^m \bar{\alpha}$ where m is the smallest integer such that

$$\tilde{f}(R_{x_\ell}(\beta^m \bar{\alpha} \eta_\ell), \mu_\ell) \leq \tilde{f}(x_\ell, \mu_\ell) - \sigma \beta^m \bar{\alpha} \|\operatorname{grad} \tilde{f}(x_\ell, \mu_\ell)\|^2. \tag{4.1}$$

    Set $x_{\ell+1} := R_{x_\ell}(t_\ell \eta_\ell)$.
  end if
end for

4. Riemannian smoothing steepest descent method. In this section, we present our RSSD method, which is detailed in Algorithm 4.1.

A few remarks for Algorithm 4.1 are in demand. First, the line search (4.1) is well defined and t_ℓ can be found in finite trials. 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 $\eta_\ell = -\operatorname{grad} \tilde{f}(x_l, \mu_l)$. 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 (4.1) is well defined.

The convergence result of Algorithm 4.1 requires the following assumption.

Assumption 4.1. For any fixed $\bar{\mu} > 0$ and any given vector $\bar{x} \in \mathcal{M}$, the level set $\mathcal{L}_{\bar{x},\bar{\mu}} = \{x \in \mathcal{M} \mid \tilde{f}(x, \bar{\mu}) \leq \tilde{f}(\bar{x}, \bar{\mu})\}$ is compact.

It is easy to see that this assumption holds if $\mathcal{M}$ is a sphere or the Stiefel manifold.

PROPOSITION 4.2. *Assume Assumption 4.1 holds. Let* $K = \{\ell \mid \|\eta_\ell\| \leq \delta_\ell\}$ *and* $\{x_\ell\}$ *be an infinite sequence generated by Algorithm 4.1 with* $\delta_{opt} = \mu_{opt} = 0$. *Then* K *is an infinite set and*

$$\lim_{\ell \to \infty,\ \ell \in K} \delta_\ell = 0 \quad \textit{and} \quad \lim_{\ell \to \infty,\ \ell \in K} \mu_\ell = 0. \tag{4.2}$$

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

$$\delta_\ell \equiv \delta_{\bar{\ell}}, \quad \mu_\ell \equiv \mu_{\bar{\ell}}, \quad \text{and} \quad \|\eta_\ell\| > \delta_{\bar{\ell}}.$$

Therefore, for $\ell \geq \bar{\ell}$, we have $x_{\ell+1} = R_{x_\ell}(t_\ell \eta_\ell)$, where t_ℓ is obtained by using the line search (4.1) with fixed $\mu_{\bar{\ell}}$. Using Assumption 4.1 and Corollary 4.3.2 of [2], we obtain

$$\lim_{\ell\to\infty} \|\eta_\ell\| = \lim_{\ell\to\infty} \|\operatorname{grad} \tilde{f}(x_\ell, \mu_\ell)\| = \lim_{\ell\to\infty} \|\operatorname{grad} \tilde{f}(x_\ell, \mu_{\bar{\ell}})\| = 0,$$

which contradicts to $\|\eta_\ell\| > \delta_{\bar{\ell}}$ for all $\ell \geq \bar{\ell}$. 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 (4.2) as desired. ∎

THEOREM 4.3. *Assume Assumption 4.1 holds. Let $K = \{\ell \mid \|\eta_\ell\| \leq \delta_\ell\}$ and $\{x_\ell\}$ be an infinite sequence generated by Algorithm 4.1 with $\delta_{opt} = \mu_{opt} = 0$. Then the following statements hold.*

(i) Any accumulation point of $\{x_\ell\}_{\ell\in K}$ is a stationary point of (1.1) associated with $\tilde{f}$ on the submanifold $\mathcal{M}$.

(ii) In addition, if $\tilde{f}$ satisfies the Riemannian gradient sub-consistency, then any accumulation point of $\{x_\ell\}_{\ell\in K}$ is a Riemannian limiting stationary point of (1.1).

Proof. By Algorithm 4.1 and Proposition 4.2, we have

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

Let x^* be any accumulation point of $\{x_\ell\}_{\ell\in K}$ with $\check{K}$ being a subsequence of K such that $\lim_{\ell\to\infty,\ \ell\in\check{K}} x_\ell = x^*$. 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 (1.1) associated with $\tilde{f}$ on the submanifold $\mathcal{M}$. That is, statement (i) holds.

In addition, if $\tilde{f}$ satisfies the Riemannian gradient sub-consistency, 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 Riemannian limiting stationary point of (1.1). Consequently statement (ii) holds. ∎

5. Numerical experiments. In this section, we apply our RSSD method (Algorithm 4.1) to solve two problems: finding a sparse vector in a subspace (FSV), and the sparsely-used orthogonal complete dictionary learning problem (ODL).

5.1. Finding a sparse vector in a subspace. The FSV problem seeks the sparsest vector 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 [34, 35]. This problem is also known as dual principal component pursuit and finds applications in robust subspace recovery [40, 52]. 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

$$\min \ \|Qx\|_0, \quad \text{s.t. } x \in S^{n-1}, \tag{5.1}$$

where $S^{n-1} = \{x \in R^n \mid \|x\| = 1\}$ is the unit sphere, and $\|z\|_0$ counts the number of nonzero entries of z. Because of the combinatorial nature of the cardinality function

$\|\cdot\|_0$, (5.1) is very difficult to solve in practice. In the literature, people have been focusing on its ℓ_1 norm relaxation given below [35, 34, 40, 52]:

$$\min \ \|Qx\|_1, \quad \text{s.t. } x \in S^{n-1}, \tag{5.2}$$

where $\|z\|_1 := \sum_i |z_i|$ is the ℓ_1 norm of vector z. Many algorithms have been proposed for solving (5.2), including the Riemannian gradient sampling algorithm [25], projected subgradient method [51], Riemannian subgradient method [31], manifold proximal point algorithm [10] and so on.

Moreover, for the compressive sensing problems that have the same objective functions as (5.1) and (5.2), people have found that using the ℓ_p quasi-norm $\|z\|_p$ $(0 < p < 1)$ to replace $\|z\|_1$ can help promote the sparsity of z [9, 21, 15, 19, 32, 33]. Motivated by this, we propose to consider the following ℓ_p $(0 < p < 1)$ minimization model for the FSV problem:

$$\min \ f(x) := \|Qx\|_p^p, \quad \text{s.t. } x \in S^{n-1}, \tag{5.3}$$

where $\|z\|_p^p := \sum_i |z_i|^p$. Note that algorithms proposed in [25, 51, 31, 10] for solving (5.2) do not apply to (5.3), because the objective function in (5.3) is non-Lipschitz. We propose to solve (5.3) using our RSSD algorithm, and we now show the details.

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 (5.3):

$$\tilde{f}(x, \mu) = \sum_{i=1}^{m} [s_\mu((Qx)_i)]^p, \tag{5.4}$$

where $s_\mu(t)$ is the uniform smoothing function for $|t|$ defined in (2.4).

Note that our RSSD can also solve the ℓ_1 norm minimization problem (5.2). Therefore, we compare our RSSD with two existing algorithms: Riemannian gradient sampling (RGS) method [24] and Riemannian nonsmooth BFGS (RBFGS) method [23] on (5.2). For the ℓ_p quisi-norm minimization problem (5.3), since no existing method is available for solving it, we only use our RSSD method to solve it, and we test RSSD with different $p \in (0, 1)$ to see the effect of p to the problem (5.3).

The FSV problems are generated as follows. We choose $n \in \{5, 10, 15, 20\}$ and $m \in \{4n, 6n, 8n, 10n, 12n, 14n\}$. The subspace W is generated following the same way as [24]. More specifically, we first generate the vector $e = (1, \ldots, 1, 0, \ldots, 0)^\top$ whose first n components are 1 and the remaining $m - n$ components are 0. We then generate $n - 1$ random vectors in $\mathbb{R}^m$. The subspace W is the span of e and these $n - 1$ random vectors. We use $Q \in \mathbb{R}^{m \times n}$ to denote the matrix whose columns form an orthonormal basis of W. The minimum value for $\|Qx\|_0$ on the sphere is likely to equal to n in this case.

We terminate our RSSD method if $\mu_l < 10^{-6}$ and $\delta_l < 10^{-4}$. As suggested in [23], we terminate the RGS and the RBFGS methods if one of the following two conditions is satisfied:

(i) the step size is less than the machine precision $2.22 \cdot 10^{-16}$;
(ii) $\epsilon_k \leq 10^{-6}$ and $\delta_k \leq 10^{-12}$.

Moreover, the maximum number of iterations is set to $n_{\max} = 1000$ for all three methods. Our RSSD, as well as the RGS method, are implemented in MATLAB. The RBFGS codes were provided to us by Wen Huang, one of the authors of [23], and they were written in C++ with a MATLAB interface. The parameters in the RGS and the RBFGS methods are set following the suggestions in [25] and [23]. The parameters of our RSSD method are set as

$$\mu_0 = 1,\ \delta_0 = 0.1, \theta_\mu = 0.5,\ \theta_\delta = 0.5. \tag{5.5}$$

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

$$x^0 = \text{randn}(n, 1); x^0 = x^0/\text{norm}(x^0).$$

For each (m, n), we generate 50 random instances with 50 random initial points. We claim that an algorithm successfully finds the sparsest vector if $\|Q\hat{x}\|_0 = n$ where $\hat{x}$ is the computed solution. Here, when we count the number of nonzeros of $Q\hat{x}$, we truncate the entries as

$$(Q\hat{x})_i = 0, \quad \text{if } |(Q\hat{x})_i| < \tau,$$

where $\tau > 0$ is a pregiven tolerance. We report the number of successful cases out of 50 cases in Tables 1 and 2. For RGS, we run the algorithm for 50 runs for each initial point and we also report the standard deviation of the average number of the successful cases.

Tables 1 and 2 record the number of success for the three methods for the ℓ_1 model (5.2) with different parameters (m, n). The bold numbers in the tables highlight the largest number of success for the corresponding (m, n). Comparing RGS, RBFGS and RSSD, we see that our RSSD method can provide a solution with the best accuracy, because when $\tau = 10^{-8}$, both the RGS and the RBFGS fail to recover the groundtruth, but our RSSD method can still recover the groudtruth in many instances. From Tables 1 and 2 we see that, in the total 64 cases, RSSD performed the best in 39 cases. For other cases that RSSD is not the best, it is still comparable in most cases.

In Tables 1 and 2 we also report the results for RSSD-g, which incorporates a global technique to RSSD by selecting the best parameters $(\theta_\mu, \theta_\delta)$ from a subset of choices. More specifically, it is worth mentioning that Example 3.6 indicates that the different relations of the sequence of unknowns and the sequence of the smoothing parameters may yield different accumulation points. The parameter θ_μ in our RSSD method controls the speed of the smoothing function that approaches to the original function, and the parameter θ_δ determines the requirement of accuracy for the approximated solution along with the iterations. The different relations of the two sequences can be obtained by using different choices of $(\theta_\mu, \theta_\delta)$. The number of successful instances can be improved if we tune the parameters $(\theta_\mu, \theta_\delta)$ for different settings of (m, n). We record in the last column of Tables 1 and 2 the numbers of successful instances of our RSSD method by selecting the best result using the different choices of

$$(\theta_\mu, \theta_\delta) = (0.5, 0.5), (0.1, 0.5), (0.5, 0.1), (0.8, 0.2), (0.2, 0.8). \tag{5.6}$$

We see from Tables 1 and 2 that by selecting the best parameters in (5.6), the performance of RSSD is clearly significantly improved.

TABLE 1
Number of success from 50 random initial points for the ℓ_1 minimization model (5.2) and $n = 5, 10$ (result for RGS is "average number of success $\pm$ standard deviation").

(n, m)	τ	RGS	RBFGS	RSSD	RSSD-g
$(5, 20)$	10^{-5}	$\mathbf{19.96 \pm 1.442}$	16	16	22
	10^{-6}	$\mathbf{19.96 \pm 1.442}$	0	16	22
	10^{-7}	0.36 ± 0.598	0	**16**	22
	10^{-8}	0 ± 0	0	**16**	22
$(5, 30)$	10^{-5}	$\mathbf{26.12 \pm 2.537}$	22	21	30
	10^{-6}	$\mathbf{26.12 \pm 2.537}$	0	21	30
	10^{-7}	0.62 ± 0.667	0	**21**	30
	10^{-8}	0 ± 0	0	**2**	30
$(5, 40)$	10^{-5}	$\mathbf{45.54 \pm 1.555}$	31	28	43
	10^{-6}	$\mathbf{45.54 \pm 1.555}$	1	28	43
	10^{-7}	0.78 ± 0.932	0	**28**	43
	10^{-8}	0 ± 0	0	**28**	43
$(5, 50)$	10^{-5}	$\mathbf{50 \pm 0}$	46	49	50
	10^{-6}	$\mathbf{50 \pm 0}$	44	49	50
	10^{-7}	0.7 ± 0.814	26	**49**	50
	10^{-8}	0 ± 0	0	**49**	50
$(10, 60)$	10^{-5}	24.16 ± 2.17	**26**	25	38
	10^{-6}	24 ± 2.231	**26**	25	38
	10^{-7}	0 ± 0	18	**25**	38
	10^{-8}	0 ± 0	0	**25**	38
$(10, 80)$	10^{-5}	$\mathbf{32.5 \pm 2.013}$	31	29	44
	10^{-6}	$\mathbf{32.42 \pm 1.960}$	31	29	44
	10^{-7}	0.002 ± 0.141	18	**29**	44
	10^{-8}	0 ± 0	0	**29**	44
$(10, 100)$	10^{-5}	$\mathbf{44.68 \pm 2.035}$	40	44	48
	10^{-6}	$\mathbf{44.56 \pm 1.971}$	40	44	48
	10^{-7}	0.02 ± 0.141	24	**44**	48
	10^{-8}	0 ± 0	0	**44**	48
$(10, 120)$	10^{-5}	$\mathbf{46.44 \pm 1.756}$	41	36	46
	10^{-6}	$\mathbf{46.22 \pm 1.753}$	41	36	46
	10^{-7}	0.1 ± 0.303	31	**36**	46
	10^{-8}	0 ± 0	0	**36**	46

We now report the results of solving the ℓ_p minimization model (5.3) using our RSSD method. In Tables 3 and 4 we again report the number of successes from 50 random instances. Here we only report the results for $\tau = 10^{-8}$. We also include the results for the ℓ_1 minimization model (5.2) for the purpose of comparison. Note that Table 4 correponds to the RSSD-g, i.e., RSSD with parameters $(\theta_\mu, \theta_\delta)$ chosen as the best one in (5.6). From Tables 3 and 4 we see that the ℓ_p minimization model (5.3) can indeed be better than the ℓ_1 minimization model (5.2), as long as an appropriate p is used.

5.2. Sparsely-used orthogonal complete dictionary learning. Given a set of data $Y = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m] \in \mathbb{R}^{n\times m}$, the sparsely-used orthogonal complete dictionary learning (ODL) seeks a dictionary that can sparsely represent Y. More specifically, ODL seeks an orthogonal matrix $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{n\times n}$ 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 [39] for more details of this model. This problem can be modeled

TABLE 2
Number of success from 50 random initial points for the ℓ_1 minimization model (5.2) and $n = 15, 20$ (result for RGS is "average number of success $\pm$ standard deviation").

(n, m)	τ	RGS	RBFGS	RSSD	RSSD-g
$(15, 90)$	10^{-5}	12.16 ± 2.054	9	**16**	31
	10^{-6}	12.16 ± 2.054	9	**16**	31
	10^{-7}	0 ± 0	4	**16**	31
	10^{-8}	0 ± 0	0	**16**	31
$(15, 120)$	10^{-5}	16.78 ± 2.401	17	**20**	36
	10^{-6}	16.74 ± 2.319	17	**20**	36
	10^{-7}	0 ± 0	11	**20**	36
	10^{-8}	0 ± 0	0	**20**	36
$(15, 150)$	10^{-5}	36.84 ± 1.856	40	**41**	48
	10^{-6}	36.8 ± 1.906	40	**41**	48
	10^{-7}	0 ± 0	36	**41**	48
	10^{-8}	0 ± 0	0	**41**	48
$(15, 180)$	10^{-5}	26.76 ± 2.162	**33**	26	40
	10^{-6}	26.66 ± 2.125	**33**	26	40
	10^{-7}	0 ± 0	**31**	26	40
	10^{-8}	0 ± 0	0	**26**	40
$(20, 160)$	10^{-5}	19.64 ± 2.819	28	**41**	43
	10^{-6}	19.62 ± 2.849	28	**41**	43
	10^{-7}	0 ± 0	20	**41**	43
	10^{-8}	0 ± 0	0	**38**	43
$(20, 200)$	10^{-5}	20.74 ± 2.717	25	**29**	46
	10^{-6}	20.74 ± 2.717	24	**29**	46
	10^{-7}	0 ± 0	23	**29**	46
	10^{-8}	0 ± 0	0	**29**	46
$(20, 240)$	10^{-5}	20.60 ± 2.441	**30**	24	35
	10^{-6}	20.58 ± 2.400	**30**	24	35
	10^{-7}	0 ± 0	**30**	24	35
	10^{-8}	0 ± 0	0	**24**	35
$(20, 280)$	10^{-5}	24.62 ± 2.230	**32**	27	37
	10^{-6}	24.60 ± 2.222	**32**	27	37
	10^{-7}	0 ± 0	**32**	27	37
	10^{-8}	0 ± 0	0	**27**	37

as an ℓ_0 minimization problem [38]:

$$\min \ \frac{1}{m} \sum_{i=1}^{m} \|\mathbf{y}_i^\top X\|_0, \quad \text{s.t. } X \in \mathrm{St}(n, n), \tag{5.7}$$

where $\mathrm{St}(n, n) = \{X \in \mathbb{R}^{n \times n} \ : \ X^\top X = I_n\}$ is the Stiefel manifold. To overcome the computational difficulty of the ℓ_0 minimization model, the ℓ_0 term is usually replaced by the ℓ_1 norm in the literture, which leads to the following ℓ_1 minimization problem for ODL [38, 39]:

$$\min \ \frac{1}{m} \sum_{i=1}^{m} \|\mathbf{y}_i^\top X\|_1, \quad \text{s.t. } X \in \mathrm{St}(n, n). \tag{5.8}$$

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

$$\min \ \frac{1}{m} \sum_{i=1}^{m} \|\mathbf{y}_i^\top X\|_p^p, \quad \text{s.t. } X \in \mathrm{St}(n, n), \tag{5.9}$$

TABLE 3

Number of success among runs from 50 random initial points for the ℓ_1 minimization model (5.2) and the ℓ_p minimization model (5.3) with $p = 0.9, \ldots, 0.1$, with $\tau = 10^{-8}$ by our RSSD method with $(\theta_\mu, \theta_\delta) = (0.5, 0.5)$.

(m,n)	ℓ_1	ℓ_p minimization model with $0 < p < 1$								
		0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
$(5,20)$	16	17	17	19	17	19	19	19	**20**	**20**
$(5,30)$	0	21	**22**	**22**	21	0	**22**	**22**	**22**	**22**
$(5,40)$	28	29	**35**	33	28	31	31	30	30	33
$(5,50)$	49	49	**50**	49	49	49	**50**	49	49	48
$(10,60)$	25	27	**28**	26	25	25	25	25	25	25
$(10,80)$	29	27	**31**	30	28	28	28	29	30	28
$(10,100)$	**44**	43	42	**44**	41	43	43	43	**44**	43
$(10,120)$	36	35	35	37	38	35	37	38	36	**41**
$(15,90)$	16	16	18	18	18	**19**	**19**	17	16	16
$(15,120)$	20	21	23	19	21	24	23	24	**25**	21
$(15,150)$	41	43	**44**	39	35	38	38	37	35	38
$(15,180)$	26	26	26	**30**	29	26	27	26	26	26
$(20,160)$	**41**	40	**41**	36	33	34	37	38	39	39
$(20,200)$	29	29	30	**33**	**33**	**33**	30	30	30	27
$(20,240)$	**24**	22	23	20	20	21	21	19	19	21
$(20,280)$	27	26	28	27	**29**	25	24	24	24	26

TABLE 4

Number of success among runs from 50 random initial points for the ℓ_1 minimization model (5.2) and the ℓ_p minimization model (5.3) with $p = 0.9, \ldots, 0.1$, with $\tau = 10^{-8}$ by our RSSD-g method, i.e., RSSD method using multiple choices $(\theta_\mu, \theta_\delta)$ in (5.6).

(m,n)	ℓ_1	ℓ_p minimization model with $0 < p < 1$								
		0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
$(5,20)$	22	21	20	22	22	23	23	**25**	23	21
$(5,30)$	30	30	31	32	32	30	32	31	**35**	27
$(5,40)$	43	40	41	43	43	**44**	42	43	42	42
$(5,50)$	**50**	**50**	**50**	**50**	**50**	**50**	**50**	**50**	**50**	**50**
$(10,60)$	38	39	41	37	39	39	38	**42**	39	38
$(10,80)$	44	42	43	42	42	**45**	43	**45**	43	40
$(10,100)$	48	47	48	48	**49**	48	48	48	48	47
$(10,120)$	46	47	46	46	46	46	44	45	46	**48**
$(15,90)$	31	26	26	31	30	28	32	**33**	30	31
$(15,120)$	35	**39**	36	35	36	35	35	33	34	32
$(15,150)$	**48**	47	**48**	47	**48**	47	47	**48**	**48**	45
$(15,180)$	40	41	40	42	42	42	40	40	**43**	38
$(20,160)$	43	**45**	44	41	42	44	43	43	**45**	43
$(20,200)$	**46**	**46**	45	45	43	45	45	40	41	41
$(20,240)$	35	32	31	33	34	35	34	32	35	**36**
$(20,280)$	35	38	38	38	39	**41**	40	**41**	40	38

and apply our RSSD method to solve it. We now specify the details. The tangent space of the Stiefel manifold is

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

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 [31], Li et.al. proposed a Riemannian subgradient method and its variants – Riemannian incremental subgradient method and Riemannian stochastic subgradient method – for solving the ℓ_1 minimization problem (5.8). In this section, we compare our RSSD for solving the ℓ_p minimization problem (5.9) and compare its performance with the algorithms proposed in [31] for solving (5.8). We thus generate the synthetic data for ODL in a similar manner as [31], 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 [31] – Riemannian subgradient method, Riemannian incremental subgradient method, and Riemannian stochastic subgradient method – as R-Full, R-Inc and R-Sto, respectively. We use our RSSD to solve the ℓ_p minimization model (5.9) with $p = 0.001$. Moreover, we again 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 use the same parameters in (5.5) for RSSD. The codes for R-Full, R-Inc and R-Sto were downloaded from the author's webpage[1].

All the algorithms were run in MATLAB (R2018b) on a notebook with 1.80GHz CPU and 16GB of RAM. For each instance, we terminated the algorithm when the CPU time reaches 50 seconds. We report the average of the sparsity level of $Y^\top \hat{X}$ over 50 instances in Table 5, 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 5 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.

Moreover, we plot the trajectory of the sparisty level in Figures 1 and 2. From these figures, it is clear that the ℓ_p minimization model (5.9) with $p = 0.001$ solved by our RSSD method provides the best results in terms of sparsity level. More specifically, our RSSD method can improve the sparsity to the desired level, while the other three algorithms stopped making progress after about one second.

6. Concluding remarks. In this paper, we developed RSSD, a novel Riemannian smoothing steepest descent method, for minimizing a non-Lipschitz function over Riemannian submanifolds. We studied some useful concepts such as the Riemannian generalized subdifferentials, and Riemannian gradient sub-consistency. We proved that any accumulation point generated by our RSSD method is a stationary point associated with the smoothing function employed in the method, which is necessary for local optimality of (1.1). Moreover, under the Riemannian gradient sub-consistency,

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

TABLE 5
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**

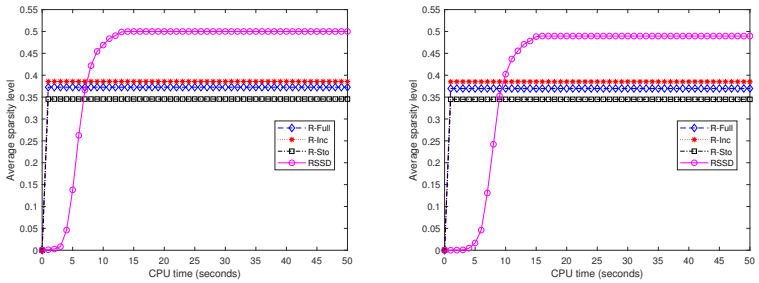

FIG. 1. *Average sparsity level versus CPU time of 50 instances using Guassian initial points. Left: $\tau = 10^{-4}$; Right: $\tau = 10^{-5}$.*

we also proved that any accumulation point is a limiting stationary point of (1.1). Numerical results on finding a sparse vector in a subspace and the sparsely-used orthogonal complete dictionary learning demonstrate the advantage of the non-Lipschitz minimization models and the efficiency of our RSSD method.

Acknowledgements. We are very grateful to Professor Wen Huang of Xiamen University for providing the C++ code for the Riemannian BFGS method, and Hui Shi for helps on the numerical experiments.

REFERENCES

[1] P.-A. ABSIL AND K. A. GALLIVAN, *Accelerated line-search and trust-region methods*, SIAM J. Numer. Anal., 47 (2009), pp. 997–1018.
[2] P.-A. ABSIL, R. MAHONY, AND R. SEPULCHRE, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2008.
[3] R. L. ADLER, J. P. DEDIEU, J. Y. MARGULIES, M. MARTENS, AND M. SHUB, *Newton's method on Riemannian manifolds and a geometric model for the human spine*, IMA J. Numer. Anal., 22 (2002), pp. 359–390.
[4] D. AZAGRA, J. FERRERA, AND B. SANZ, *Viscosity solutions to second order partial differential equations on Riemannian manifolds*, J. Differ. Equations, 245 (2012), pp. 307–336.
[5] M. BAČÁK, R. BERGMANN, G. STEIDL, AND A. WEINMANN, *A second order non-smooth variational model for restoring manifold-valued images*, SIAM J. Sci. Comput., 38 (2016), pp. A567–A597.
[6] W. BIAN AND X. CHEN, *Linearly constrained non-Lipschitz optimization for image restoration*, SIAM J. Imaging Sci., 8 (2015), pp. 2294–2322.
[7] J. V. BURKE AND T. HOHEISEL, *Epi-convergent smoothing with applications to convex composite functions*, SIAM J. Optim., 23 (2013), pp. 1457–1479.
[8] J. V. BURKE, T. HOHEISEL, AND C. KANZOW, *Gradient consistency for integral-convolution smoothing functions*, Set-Valued Var. Anal., 21 (2013), pp. 359–376.
[9] R. CHARTRAND AND W. YIN, *Iteratively reweighted algorithms for compressive sensing*, in ICASSP, 2008.
[10] S. CHEN, Z. DENG, S. MA, AND A. M.-C. SO, *Manifold proximal point algo-*

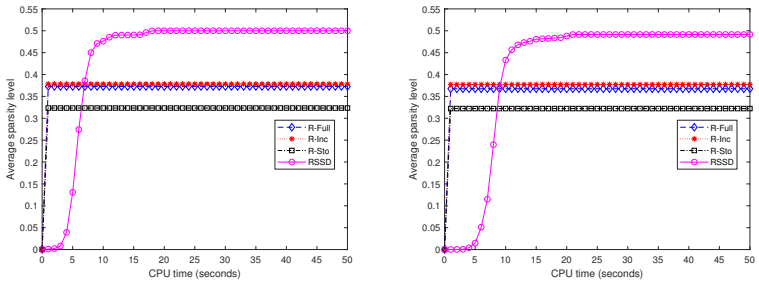

FIG. 2. *Average sparsity level versus CPU time of 50 instances using uniform initial points. Left:* $\tau = 10^{-4}$*; Right:* $\tau = 10^{-5}$.

rithms for dual principal component pursuit and orthogonal dictionary learning, https://arxiv.org/abs/2005.02356, (2020).

[11] S. CHEN, S. MA, A. M.-C. SO, AND T. ZHANG, *Proximal gradient method for nonsmooth optimization over the Stiefel manifold*, SIAM J. Optim., 30 (2020), pp. 210–239.

[12] S. CHEN, S. MA, L. XUE, AND H. H. ZOU, *An alternating manifold proximal gradinet method for sparse principal component analysis and sparse canonical correlation analysis*, INFORMS J. Optimization, 2 (2020), pp. 192–208.

[13] W. CHEN, H. JI, AND Y. YOU, *An augmented Lagrangian method for* ℓ_1*-regularized optimization problems with orthogonality constraints*, SIAM J. Sci. Comput., 38 (2016), pp. B570–B592.

[14] X. CHEN, *Smoothing methods for nonsmooth, nonconvex minimization*, Math. Program., Ser. B, 134 (2012), pp. 71–99.

[15] X. CHEN, D. GE, Z. WANG, AND Y. YE, *Complexity of unconstrained* L_2*-*L_p *minimization*, Math. Program., 143 (2014), pp. 371–383.

[16] X. CHEN, L. GUO, Z. LUO, AND J. J. YE, *An augmented Lagrangian method for non-Lipschitz nonconvex programming*, SIAM J. Numer. Anal., 55 (2017), pp. 168–193.

[17] X. CHEN, M. K. NG, AND C. ZHANG, *Non-Lipschitz* ℓ_p *regularization and box constrained model for image restoration*, IEEE Trans. Image Process., 21 (2012), pp. 4709–4721.

[18] X. CHEN, L. NIU, AND Y. YUAN, *Optimality conditions and smoothing trust region Newton method for non-Lipschitz optimization*, SIAM J. Optim., 23 (2013), pp. 1528–1552.

[19] X. CHEN, F. XU, AND Y. YE, *Lower bound theory of nonzero entries in solutions of* $\ell_2 - \ell_p$ *minimization*, SIAM J. Sci. Comput., 32 (2010), pp. 2832–2852.

[20] X. CHEN AND W. ZHOU, *Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization*, SIAM J. Imaging Sci., 3 (2010), pp. 765–790.

[21] S. FOUCART AND M. J. LAI, *Sparsest solutions of underdetermined linear systems via* ℓ_q*-minimization for* $0 < q \leq 1$, Appl. Comput. Harmon. Anal., 26 (2009), pp. 395–407.

[22] P. GROHS AND S. HOSSEINI, ϵ*-subgradient algorithms for locally Lipschitz functions on Riemannian manifolds*, Adv. Comput. Math., 42 (2016), pp. 333–366.

[23] S. HOSSEINI, W. HUANG, AND R. YOUSEFPOUR, *Line search algorithms for locally Lipschitz functions on Riemannian manifolds*, SIAM J. Optim., 28 (2018), pp. 596–619.

[24] S. HOSSEINI AND M. R. POURYAYEVALI, *Generalized gradients and characterization of epi-Lipschitz sets in Riemannian manifolds*, Nonlinear Anal.-Theor., 74 (2001), pp. 3884–3895.

[25] S. HOSSEINI AND A. USCHMAJEW, *A Riemannian gradient sampling algorithm for nonsmooth optimization on manifolds*, SIAM J. Optim., 27 (2017), pp. 173–189.

[26] W. HUANG, P.-A. ABSIL, AND K. A. GALLIVAN, *A Riemannian BFGS method without differentiated retraction for nonconvex optimization problems*, SIAM J. Optim., 28 (2018), pp. 470–495.

[27] W. HUANG AND K. WEI, *Riemannian proximal gradient methods*, accepted in Math. Program., (2021).

[28] A. KOVNATSKY, K. GLASHOFF, AND M. M. BRONSTEIN, *MADMM: a generic algorithm for non-smooth optimization on manifolds*, in European Conference on Computer Vision, Springer, 2016, pp. 680–696.

[29] R. LAI AND S. OSHER, *A splitting method for orthogonality constrained problems*, J. Sci. Comput., 58 (2014), pp. 431–449.

[30] Y. S. LEDYAEV AND Q. J. ZHU, *Nonsmooth analysis on smooth manifolds*, Trans. Amer. Math.

Soc., 359 (2007), pp. 3687–3732.
[31] X. Li, S. Chen, Z. Deng, Q. Qu, Z. Zhu, and A. M.-C. So, *Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods*, arXiv: 1911.05047v3, accepted in SIAM J. Optim., (2019).
[32] Y.-F. Liu, Y.-H. Dai, and S. Ma, *Joint power and admission control: Non-convex l_q approximation and an effective polynomial time deflation approach*, IEEE Trans. Signal Process., 63 (2015), pp. 3641–3656.
[33] Y.-F. Liu, S. Ma, Y.-H. Dai, and S. Zhang, *A smoothing SQP framework for a class of composite L_q minimization over polyhedron*, Math. Program., Ser. A, 158 (2016), pp. 467–500.
[34] Q. Qu, J. Sun, and J. Wright, *Finding a sparse vector in a subspace: Linear sparsity using alternating directions*, IEEE Trans. Inf. Theory, 62 (2016), pp. 5855–5880.
[35] Q. Qu, Z. Zhu, X. Li, M. C. Tsakiris, J. Wright, and R. Vidal, *Finding the sparsest vectors in a subspace: Theory, algorithms, and applications*, https://arxiv.org/abs/2001.06970, (2020).
[36] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, New York, 1998.
[37] F. Shang, J. Cheng, Y. Liu, Z.-Q. Luo, and Z. Lin, *Bilinear factor matrix norm minimization for robust PCA: algoirthms and applications*, IEEE Trans. Pattern Anal., 40 (2018), pp. 2066–2080.
[38] D. A. Spielman, H. Wang, and J. Wright, *Exact recovery of sparsely-used dictionaries*, in Conference on Learning Theory, 2012.
[39] J. Sun, Q. Qu, and J. Wright, *Complete dictionary recovery over the sphere I: overview and the geometric picture*, IEEE Trans. Inform. Theory, 63 (2017), pp. 853–884.
[40] M. C. Tsakiris and R. Vidal, *Dual principal component pursuit*, J. Mach. Learn. Res., 19 (2018), pp. 1–49.
[41] B. Wang, S. Ma, and L. Xue, *Riemannian stochastic proximal gradient methods for nonsmooth optimization over the Stiefel manifold*, https://arxiv.org/pdf/2005.01209.pdf, (2020).
[42] Z. Wang, B. Liu, S. Chen, S. Ma, L. Xue, and H. Zhao, *A manifold proximal linear method for sparse spectral clustering with application to single-cell RNA sequencing data analysis*, https://arxiv.org/abs/2007.09524, (2020).
[43] H. Whitney, *Analytic extensions of differentiable functions defined in closed sets*, Trans. Amer. Math. Soc., 36 (1934), pp. 63–89.
[44] M. Xu, J. J. Ye, and L. Zhang, *Smoothing SQP mehtods for solving degenerate nonsmooth constrained optimization problems with applications to bilevel programs*, SIAM J. Optim., 25 (2015), pp. 1388–1410.
[45] W. Yang, L.-H. Zhang, and R. Song, *Optimality conditions for the nonlinear programming problems on Riemannian manifolds*, Pacific J. Optim., 10 (2014), pp. 415–434.
[46] C. Zeng, C. Wu, and R. Jia, *Non-Lipschitz models for image restoration with impulse noise removel*, SIAM J. Imaging Sci., 12 (2019), pp. 420–458.
[47] C. Zhang and X. Chen, *Smoothing projected gradient method and its application to stochastic linear complementarity problem*, SIAM J. Optim., 20 (2009), pp. 627–649.
[48] C. Zhang and X. Chen, *A smoothing active set method for linearly constrained non-Lipschitz nonconvex optimization*, SIAM J. Optim., 30 (2020), pp. 1–30.
[49] Y. Zhou, C. Bao, C. Ding, and J. Zhu, *A semi-smooth Newton based augmented Lagrangian method for nonsmooth optimization on matrix manifolds*, https://arxiv.org/abs/2103.02855, (2021).
[50] H. Zhu, X. Zhang, D. Chu, and L. Liao, *Nonconvex and nonsmooth optimization with generalized orthogonality constraints: An approximate augmented Lagrangian method*, J. Sci. Comput., 72 (2017), pp. 331–372.
[51] Z. Zhu, T. Ding, D. P. Robinson, M. C. Tsakiris, and R. Vidal, *A linearly convergent method for non-smooth non-convex optimization on the Grassmannian with applications to robust subspace and dictionary learning*, in NeurIPS, 2019.
[52] Z. Zhu, Y. Wang, D. P. Robinson, D. Naiman, R. Vidal, and M. C. Tsakiris, *Dual principal component pursuit: Improved analysis and efficient algorithms*, in NeurIPS, 2018.