Journal of Machine Learning Research 25 (2024) 1-32 Submitted 3/23; Revised 1/24; Published 3/24

A projected semismooth Newton method for a class of nonconvex composite programs with strong prox-regularity

Jiang Hu HUJIANGOPT@GMAIL.COM
Massachusetts General Hospital and Harvard Medical School
Harvard University
Boston, MA 02114, USA

Kangkang Deng* FREEDENG1208@GMAIL.COM
Beijing International Center for Mathematical Research
Peking University
Beijing, 100871, China

Jiayuan Wu 1901110043@PKU.EDU.CN
College of Engineering
Peking University
Beijing, 100871, China

Quanzheng Li LI.QUANZHENG@MGH.HARVARD.EDU
Massachusetts General Hospital and Harvard Medical School
Harvard University
Boston, MA 02114, USA

Editor: Animashree Anandkumar

Abstract

This paper aims to develop a Newton-type method to solve a class of nonconvex composite programs. In particular, the nonsmooth part is possibly nonconvex. To tackle the nonconvexity, we develop a notion of strong prox-regularity which is related to the singleton property and Lipschitz continuity of the associated proximal operator, and we verify it in various classes of functions, including weakly convex functions, indicator functions of proximally smooth sets, and two specific sphere-related nonconvex nonsmooth functions. In this case, the problem class we are concerned with covers smooth optimization problems on manifold and certain composite optimization problems on manifold. For the latter, the proposed algorithm is the first second-order type method. Combining with the semismoothness of the proximal operator, we design a projected semismooth Newton method to find a root of the natural residual induced by the proximal gradient method. Due to the possible nonconvexity of the feasible domain, an extra projection is added to the usual semismooth Newton step and new criteria are proposed for the switching between the projected semismooth Newton step and the proximal step. The global convergence is then established under the strong prox-regularity. Based on the BD regularity condition, we establish local superlinear convergence. Numerical experiments demonstrate the effectiveness of our proposed method compared with state-of-the-art ones.

Keywords: nonconvex composite optimization, strong prox-regularity, projected semismooth Newton method, superlinear convergence

*. Corresponding author.

1. Introduction

The nonconvex composite minimization problem has attracted lots of attention in signal processing, statistics, and machine learning. The formulation we are concerned with is:

$$\min_{x\in\mathbb{R}^n} \quad \varphi(x) := f(x) + h(x), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable and possibly nonconvex, $h : \mathbb{R}^n \to (-\infty, \infty]$ is a proper, closed, and extended real-valued function. Note that h can be nonsmooth and nonconvex. In this paper, we consider a class of nonsmooth and nonconvex functions h satisfying the following strong prox-regularity.

Definition 1 (strong prox-regularity) *We call a proper, closed, and extended real-valued function $h : \mathbb{R}^n \to \bar{\mathbb{R}}$ is strongly prox-regular with respect to a closed set $\mathcal{C} \supset \mathrm{dom}(h)$, a positive constant γ, and a norm function $\|\cdot\|$, if the proximal operator $\mathrm{prox}_{th}(\cdot) := \arg\min_u th(u) + \frac{1}{2}\|\cdot - u\|_2^2$ is single-valued and Lipschitz continuous over the closed γ-neighborhood of $\mathcal{C}$, denoted as $\{x + tv : x \in \mathcal{C} \subset \mathbb{R}^n, v \in \mathbb{R}^n \text{ with } \|v\| = 1, 0 \le t \le \gamma\}$.*

We call the above definition strong prox-regularity due to the uniform γ for all $x \in \mathcal{C}$, which can be seen as an enhanced version of the prox-regularity (Rockafellar and Wets, 2009, Definition 13.27, Proposition 13.37). Note that the strong prox-regularity holds for any closed $\mathcal{C} \subset \mathbb{R}^n$ and $\gamma > 0$ if h is convex (Moreau, 1965). Here, we present some classes of nonconvex functions satisfying Definition 1.

(i) h is weakly convex. A function is called weakly convex with modulus $\rho > 0$ if $h(x) + \frac{\rho}{2}\|x\|^2$ is convex. By using the same idea for the convex functions, one can verify that prox_{th} is single-valued and Lipschitz continuous when $t < \frac{1}{\rho}$. Thus, h is strongly prox-regular with $\mathcal{C} = \mathbb{R}^n$, $\gamma = t$, and the ℓ_2-norm $\|\cdot\|_2$ for any $t < \frac{1}{\rho}$. Optimization with weakly convex objective functions has been considered in (Davis and Drusvyatskiy, 2019).

(ii) h is the indicator function of a proximally smooth set (Clarke et al., 1995). For a set $\mathcal{X} \subset \mathbb{R}^n$, define its closed r-neighborhoods

$$\mathcal{X}(r) := \{u \in \mathbb{R}^n : d_{\mathcal{X}}(u) \le r\}, \text{ with } d_{\mathcal{X}}(u) := \inf\{\|u - x\| : x \in \mathcal{X}\}. \tag{2}$$

We say that $\mathcal{X}$ is r-proximally smooth if the nearest-point projection $\mathrm{proj}_{\mathcal{X}}$ is single-valued on $\mathcal{X}(r)$. In addition, the proximal operator (which is the same as the projection operator onto $\mathcal{X}$) is Lipschitz continuous (Clarke et al., 1995, Theorem 4.8) on $\mathcal{X}(r)$. Thus, the indicator function $\delta_{\mathcal{X}}(\cdot)$ is strongly prox-regular with $\mathcal{C} = \mathcal{X}$, $\gamma = r$, and $\|\cdot\|$. Note that the projection operator onto a smooth and compact manifold embedded in Euclidean space is a smooth mapping on a neighborhood of the manifold (Foote, 1984). It is also worth mentioning that the Stiefel manifold is 1-proximally smooth (Balashov and Tremba, 2022, Proposition 1).

As shown above, optimization with weakly convex regularizers or constraints of the proximally smooth set can be fitted into (1). The strong prox-regularity serves as a general concept to put different problem classes together and allows us to derive a uniform

algorithmic design and theoretic analysis. Since the proximal operator is single-valued and Lipschitz continuous on a closed set, one can further explore the differentiability and design second-order type algorithms to obtain the algorithmic speedup and fast convergence rate guarantee.

It has been shown in (Böhm and Wright, 2021) that two popular nonsmooth nonconvex regularizers, the minimax concave penalty (Zhang, 2010) and the smoothly clipped absolute deviation (Fan, 1997), are weakly convex. Since any smooth manifold is proximally smooth, the manifold optimization problems (Absil et al., 2009; Hu et al., 2020; Boumal, 2023) take the form (1). Besides, we are also motivated by the following applications, where h is from the oblique manifold and a simple ℓ_1 norm or the constraint of nonnegativity. Let us note that such h is not weakly convex or the indicator function of a smooth manifold.

1.1 Motivating examples

Example 1. Sparse PCA on oblique manifold

In (Huang and Wei, 2021), the authors consider the following formulation of sparse PCA:

$$\min_{X\in \mathrm{Ob}(n,p)} \quad \|X^\top A^\top AX - D^2\|_F^2 + \lambda\|X\|_1, \tag{3}$$

where $\mathrm{Ob}(n,p) = \{X \in \mathbb{R}^{n\times p} : \mathrm{diag}(X^\top X) = 1_p\}$ with $\mathrm{diag}(B)$ being a vector consisting of the diagonal entries of B and $1_p \in \mathbb{R}^n$ of all elements 1, D is a diagonal matrix whose diagonal entries are the first p largest singular values of A, $\|\cdot\|_F$ denotes the matrix Frobenius norm, $\|X\|_1 := \sum_{i=1}^n \sum_{j=1}^p |X_{ij}|$, and $\lambda > 0$ is a parameter to control the sparsity. Problem (3) takes the form (1) by letting

$$h(X) = \lambda\|X\|_1 + \delta_{\mathrm{Ob(n,p)}}(X), \tag{4}$$

where $\delta_{\mathcal{C}}(\cdot)$ denotes the indicator function of the set $\mathcal{C}$, which takes the value zero on $\mathcal{C}$ and $+\infty$ otherwise. Utilizing the separable structure and the results by (Xiao and Bai, 2021), the i-th column of $\mathrm{prox}_{th}(X)$, denoted by $(\mathrm{prox}_{th}(X))_i$, is

$$(\mathrm{prox}_{th}(X))_i = \begin{cases} (\underbrace{0,\ldots,0}_{j-1}, \mathrm{sign}(X_{ij}), \underbrace{0,\ldots,0}_{n-j})^\top, & \text{if } \ w \geq 0, \\ -w_i^-/\|w_i^-\|_2 \cdot \mathrm{sign}(X_i), & \text{otherwise,} \end{cases}$$

where $w_i = \lambda t - |X_i|$, X_i is the i-th column of X, $w_i^- = \min(w_i, 0)$, $\mathrm{sign}(a)$ returns 1 if $a \geq 0$ and -1 otherwise, and $j = \arg\min_{1\leq k\leq n} w_i(k)$. Note that prox_{th} is not unique for all $X \in \mathbb{R}^{n\times p}$ and $t > 0$. We will give the specific $\mathcal{C}$, γ, and $\|\cdot\|$ such that prox_{th} is strongly prox-regular later in Section 3.

Example 2. Nonnegative PCA on oblique manifold

If the nonnegativity of the principal components is required, we have the following nonnegative PCA model

$$\min_{X\in \mathrm{Ob}^+(n,p)} \quad \|X^\top A^\top AX - D^2\|_F^2, \tag{5}$$

where $\mathrm{Ob}^+(n,p) := \mathrm{Ob}(n,p) \cap \{X \in \mathbb{R}^{n\times p} : X_{ij} \geq 0\}$ and D is defined as in (3). Note that a more general formulation with smooth objective function over $\mathrm{Ob}^+(n,p)$ has been considered in (Jiang et al., 2022). Problem (5) falls into (1) by letting

$$h(X) = \delta_{\mathrm{Ob}^+(n,p)}(X) \tag{6}$$

is the indicator function of $\mathrm{Ob}^+(n,p)$. Due to the separable structure, the i-th column of $\mathrm{prox}_{th}(X)$, denoted by $(\mathrm{prox}_{th}(X))_i$, is

$$(\mathrm{prox}_{th}(X))_i = \begin{cases} (\underbrace{0,\ldots,0}_{j-1}, 1, \underbrace{0,\ldots,0}_{n-j}), & \text{if } \max(X_i) \leq 0, \\ X_i^+/\|X_i^+\|_2, & \text{otherwise,} \end{cases}$$

where $j = \arg\min_{1\leq k\leq n} X_{ik}$ in the first case, $X_i^+ = \max(X_i, 0)$, and X_i is the i-th column of X. Note that this projection is not unique for all $X \in \mathbb{R}^{n\times p}$, e.g., $X = 0$. We will show its strong prox-regularity later in Section 3.

Example 3. Sparse least square regression with probabilistic simplex constraint

The authors of (Xiao and Bai, 2021; Li et al., 2021) consider the spherical constrained formulation of the following optimization problems:

$$\min_{y\in\mathbb{R}^n} \ \frac{1}{2}\|Ay - b\|_2^2, \ \text{ s.t. } \ y \in \Delta_n, \tag{7}$$

where $\Delta_n = \{y \in \mathbb{R}^n : y \geq 0, 1_n^\top y = 1\}$, $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$. By decomposing $y = x \odot x$ with the Hadamard product $\odot$ (i.e., $y_i = x_i^2, \ i = 1,\ldots,n$), it holds that

$$y \in \Delta_n \iff x \in \mathrm{Ob}(n,1).$$

Adding a sparsity constraint on x leads to the following optimization problem

$$\min_{x\in\mathbb{R}^n} \ \frac{1}{2}\|A(x\odot x) - b\|_2^2 + \lambda\|x\|_1, \ \text{ s.t. } \ x \in \mathrm{Ob}(n,1). \tag{8}$$

By taking $h(x) = \lambda\|x\|_1 + \delta_{\mathrm{Ob}(n,1)}$, problem (8) has the form (1). Due to the separable structure of the proximal operator of (4), the strong prox-regularity of h here is similar to that of (4).

1.2 Literature review

The composite optimization problem arises from various applications, such as signal processing, statistics, and machine learning. When h is convex, extensive first-order methods are designed, such as the proximal gradients and its Nesterov's accelerated versions, the alternating direction methods of multipliers, etc. We refer to (Boyd et al., 2011; Beck, 2017) for details. For faster convergence, second-order methods, such as proximal Newton methods (Lee et al., 2014; Kanzow and Lechner, 2021) and semismooth Newton methods (Mifflin, 1977; Qi and Sun, 1993, 1999; Byrd et al., 2016; Milzarek and Ulbrich, 2014; Zhao

et al., 2010; Xiao et al., 2018; Li et al., 2018a,b) are also developed for the nonsmooth problem (1). If h is nonconvex, the proximal gradient methods are developed for $\ell_{1/2}$ norm in (Xu et al., 2012) and more nonconvex regularizers (Gong et al., 2013; Yang, 2017). The global convergence is established by utilizing the smoothness of f and the explicit solution of the proximal subproblem.

In the case of h being weakly convex, subgradient-type methods (Davis and Drusvyatskiy, 2019; Davis et al., 2018) and proximal point-type method (Drusvyatskiy, 2018) yield lower complexity bound. Optimization with prox-regular functions has recently attracted much attention. The authors (Themelis et al., 2018) propose a gradient-type method to solve the forward-backward envelope of φ. This can be seen as a variable-metric first-order method. Since the Moreau envelope of a prox-regular function is continuously differentiable, a nonsmooth Newton method is designed to solve the gradient system of the Moreau envelope in (Khanh et al., 2020, 2021). Note that the indicator function of a proximally smooth set is prox-regular (Clarke et al., 1995), the authors of (Balashov and Tremba, 2022) developed a generalized Newton method to fixed point equation induced by the projected gradient method.

In the case of h being the indicator function of a Riemannian manifold, the efficient Riemannian algorithms have been extensively studied in the last decades (Absil et al., 2009; Wen and Yin, 2013; Hu et al., 2020; Boumal, 2023). When h takes the form (4), the manifold proximal gradient methods (Chen et al., 2020; Huang and Wei, 2021) are designed. These approaches only use first-order information and do not have superlinear convergence. In addition, manifold augmented Lagrangian methods are also proposed in works (Deng and Peng, 2022; Zhou et al., 2021), in which the subproblem is solved by the first-order method or second-order method. When it comes to the case of (6), a second-order type method is proposed in the recent work (Jiang et al., 2022). While in their subproblems, only the second-order information of the smooth part is explored.

1.3 Our contributions

In this paper, we propose a projected semismooth Newton method to deal with a class of nonsmooth and nonconvex composite programs. In particular, the nonsmooth part is nonconvex but satisfies the proposed strong prox-regularity properties. Our main contributions are as follows:

- We introduce the concept of strong prox-regularity. Different from the classic prox-regularity, the strong prox-regularity enjoys some kind of uniform proximal regularity around a closed region containing all feasible points. A crucial property is that the proximal operator of a strongly prox-regular function locally behaves like that of convex functions. With the strong prox-regularity, the stationary condition can be reformulated as a single-valued residual mapping which is Lipschitz continuous on the closed region. We present several classes of functions satisfying both the strong prox-regularity condition, including weakly convex functions and indicator functions of proximally smooth sets (including manifold constraints). In particular, two specific sphere-related nonsmooth and nonconvex functions, which are not weakly convex or indicator functions of a smooth manifold, are verified to satisfy the strong prox-regularity.

- As shown in Section 1.1, two sphere-related nonsmooth and nonconvex functions result in composite optimization problems on manifolds. In this paper, we propose the first second-order type method to solve this kind of problem, which outperforms state-of-the-art first-order methods (Chen et al., 2020; Huang and Wei, 2021). It is worth mentioning that first-order methods (Chen et al., 2020; Huang and Wei, 2021) fail in solving the nonnegative PCA on the oblique manifold due to their dependence on the Lipschitz continuity of the nonsmooth part.

- By introducing the strong prox-regularity condition and semismoothness, we design a residual-based projected semismooth Newton method to solve the nonconvex composite optimization problem (1). To tackle the nonconvexity, we add an extra projection on the usual semismooth Newton step and switch to the proximal gradient step if two proposed inexact conditions are not satisfied. Compared with the Moreau-envelope based approaches (Khanh et al., 2020, 2021), we decouple the composite structures and design a second-order method by utilizing the second-order derivative of the smooth part and the generalized Jacobian of the proximal operator of h.

- The global convergence of the proposed projected semismooth Newton method is presented. Other than the strong prox-regularity condition and the semismoothness, the assumptions are standard and can be achieved by various applications including our motivating examples. We prove the switching conditions are locally satisfied, which allows the local transition to the projected semismooth Newton step. By assuming the BD-regularity condition, we show the local superlinear convergence. Numerical experiments on various applications demonstrate the efficiency over state-of-the-art ones.

1.4 Notation

Given a matrix A, we use $\|A\|_F$ to denote its Frobenius norm, $\|A\|_1 := \sum_{ij} |A_{ij}|$ to denote its ℓ_1 norm, and $\|A\|_2$ to denote its spectral norm. For a vector x, we use $\|x\|_2$ and $\|x\|_1$ to denote its Euclidean norm and ℓ_1 norm, respectively. The symbol $\mathbb{B}$ will denote the closed unit ball in $\mathbb{R}^n$, while $\mathbb{B}(x, \epsilon)$ will stand for the closed ball of the radius of $\epsilon > 0$ centered at x.

1.5 Organization

The outline of this paper is as follows. In Section 2, we present the preliminaries on the subdifferential, concepts of stationarity, and semismoothness. Various nonconvex and nonsmooth functions satisfying the strong prox-regularity and semismoothness are demonstrated in Section 3. Then, we propose a projected semismooth Newton method in Section 4. The corresponding convergence analysis of the proposed method is provided in Section 5. We illustrate the efficiency of our proposed method by several numerical experiments in Section 6. Finally, a brief conclusion is given in Section 7.

2. Preliminaries

In this section, we first review some basic notations of subdifferential and give the definition of the prox-regular function. We also introduce several concepts of stationarity and present the definition of semismoothness.

2.1 Subdifferential and prox-regular functions

Let $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper, lower semicontinuous, and extended real-valued function. The domain of φ is defined as $\mathrm{dom}(\varphi) = \{x \in \mathbb{R}^n : \varphi(x) < +\infty\}$. A vector $v \in \mathbb{R}^n$ is said to be a Fréchet subgradient of φ at $x \in \mathrm{dom}(\varphi)$ if

$$\liminf_{\substack{y \to x, \\ y \neq x}} \frac{\varphi(y) - \varphi(x) - \langle v, y - x\rangle}{\|y - x\|_2} \geq 0. \tag{9}$$

The set of vectors $v \in \mathbb{R}^p$ satisfying (9) is called the Fréchet subdifferential of φ at $x \in \mathrm{dom}(\varphi)$ and denoted by $\widehat{\partial}\varphi(x)$. The limiting subdifferential, or simply the subdifferential, of φ at $x \in \mathrm{dom}(\varphi)$ is defined as

$$\partial\varphi(x) = \left\{v \in \mathbb{R}^n : \exists x^k \to x, v^k \to v \text{ with } \varphi\left(x^k\right) \to \varphi(x), v^k \in \widehat{\partial}\varphi\left(x^k\right)\right\}.$$

By convention, if $x \notin \mathrm{dom}(\varphi)$, then $\partial\varphi(x) = \emptyset$. The domain of $\partial\varphi$ is defined as $\mathrm{dom}(\partial\varphi) = \{x \in \mathbb{R}^n : \partial\varphi(x) \neq \emptyset\}$. For the indicator function $\delta_{\mathcal{S}} : \mathbb{R}^n \to \{0, +\infty\}$ associated with the non-empty closed set $\mathcal{S} \subseteq \mathbb{R}^n$, we have

$$\widehat{\partial}\delta_{\mathcal{S}}(x) = \left\{v \in \mathbb{R}^n : \limsup_{y \to x, y \in \mathcal{S}} \frac{\langle v, y - x\rangle}{\|y - x\|_2} \leq 0\right\} \quad \text{and} \quad \partial\delta_{\mathcal{S}}(x) = \mathcal{N}_{\mathcal{S}}(x)$$

for any $x \in \mathcal{S}$, where $\mathcal{N}_{\mathcal{S}}(x)$ is the normal cone to $\mathcal{S}$ at x.

The function φ is prox-bounded (Rockafellar and Wets, 2009, Definition 1.23) if there exists $\lambda > 0$ such that $e_\lambda\varphi(x) := \inf_y\{\varphi(y) + \frac{1}{2\lambda}\|y - x\|_2^2\} > -\infty$ for some $x \in \mathbb{R}^n$. The supremum of the set of all such λ is the threshold λ_φ of prox-boundedness for φ. The function φ is prox-regular (Rockafellar and Wets, 2009, Definition 13.27) at $\bar{x}$ for $\bar{v}$ if φ is finite and locally lower semicontinuous at $\bar{x}$ with $\bar{v} \in \partial\varphi(\bar{x})$, and there exist $\varepsilon > 0$ and $\rho \geq 0$ such that

$$\varphi(x') \geq \varphi(x) + \left\langle v, x' - x\right\rangle - \frac{\rho}{2}\|x' - x\|_2^2, \ \forall x' \in \mathbb{B}(\bar{x}, \varepsilon), \tag{10}$$

when $v \in \partial\varphi(x)$, $\|v - \bar{v}\|_2 < \varepsilon, \|x - \bar{x}\|_2 < \varepsilon, \varphi(x) < \varphi(\bar{x}) + \varepsilon$. If the above inequality holds for all $\bar{v} \in \partial\varphi(\bar{x})$, φ is said to be prox-regular at $\bar{x}$. Note that the inequality (10) holds for all $x' \in \mathrm{dom}(\varphi)$ and $v \in \partial\varphi(x)$ with a uniform ρ if φ is weakly convex. It follows from (Rockafellar and Wets, 2009, Exercise 13.35) that the summation of a smooth function and a prox-regular function is prox-regular as well.

For prox-regular functions, we have the following fact.

Proposition 2 *((Rockafellar and Wets, 2009, Proposition 13.37), (Khanh et al., 2020, Lemma 6.3)) Let $\varphi : \mathbb{R}^n \to \bar{\mathbb{R}}$ be proper, lower semicontinuous, and prox-bounded with threshold λ_φ. Suppose φ is finite and prox-regular at $\bar{x}$ for $\bar{v} \in \partial\varphi(\bar{x})$. Then for any sufficiently small $\gamma \in (0, \lambda_\varphi)$, the proximal mapping $\mathrm{prox}_{\gamma\varphi}$ is single-valued and Lipschitz continuous around $\bar{x} + \gamma\bar{v}$ and satisfies the condition $\mathrm{prox}_{\lambda\varphi}(\bar{x} + \gamma\bar{v}) = \bar{x}$.*

Our proposed prox-regularity condition is a stronger version of the well-known prox-regularity condition in optimization theory. Specifically, our condition requires the proximal operator to be single-valued and Lipschitz continuous for a closed region $\mathcal{C}$ with a uniform γ. As shown later, the uniformity of γ plays a critical role in determining the lower bound of step sizes in algorithmic design.

2.2 Concepts of stationarity and their relationship

There are two definitions of stationarities based on the subdifferential and the proximal gradient iteration.

- Critical point: x is a critical point if

$$0 \in \partial\varphi(x) = \nabla f(x) + \partial h(x). \tag{11}$$

- Fixed point of the proximal mapping:

$$x \in \operatorname{prox}_{th}(x - t\nabla f(x)), \tag{12}$$

  where $t > 0$.

It follows from the definition of prox_{th} that any point x satisfying (12) yields $0 \in \nabla f(x) + \partial h(x)$, which implies x is also a critical point. Inversely, a critical point may not satisfy (12) due to the nonconvexity of h. Therefore, equation (12) defines a stronger stationary point than (11).

2.3 Semismoothness

By the Rademacher's theorem, a locally Lipschitz operator is almost everywhere differentiable. For a locally Lipschitz F, denote by D_F the set of the differential points of F. The B-subdifferential at x is defined as

$$\partial_B F(x) := \left\{ \lim_{k\to\infty} J\left(x^k\right) \mid x^k \in D_F, x^k \to x \right\},$$

where $J(x)$ represents the Jacobian of F at the differentiable point x. Obviously, $\partial_B F(x)$ may not be a singleton. The Clarke subdifferential $\partial_C F(x)$ is defined as

$$\partial_C F(x) = \operatorname{conv}\left(\partial_B F(x)\right),$$

where $\operatorname{conv}(A)$ represents the closed convex hull of A. A locally Lipschitz continuous operator F is called semismooth at x with respect to $\partial_B F$ ($\partial_C F$) if

- F is directionally differentiable at x, i.e., for any direction d, the limit $\lim_{t\downarrow 0} \frac{F(x+td)-F(x)}{t}$ exists.
- For all d and $J \in \partial_B F(x+d)$ ($\partial_C F(x+d)$), it holds that

$$\|F(x+d) - F(x) - Jd\|_2 = o(\|d\|_2), \quad d \to 0.$$

We say F is semismooth with respect to $\partial_B F$ ($\partial_C F$) if F is semismooth for any $x \in \mathbb{R}^n$ with respect to $\partial_B F$ ($\partial_C F$). If f is twice continuously differentiable and prox_{th} is single-valued, Lipschitz continuous, and semismooth with respect to its B-subdifferential $D(x)$, one can follow (Chan and Sun, 2008, Lemma 1) to verify that if $I - t\nabla^2 f(x)$ is nonsingular, the operator $F(x) := \mathrm{prox}_{th}(x - t\nabla f(x)) - x$ is semismooth with respect to

$$M(x) := \{I - D(I - t\nabla^2 f(x)) : D \in D(x)\} \tag{13}$$

by using the definition of semismoothness.

3. Semismooth and strongly prox-regular functions

Let us verify the semismoothness and the strongly prox-regularity condition for some typical nonconvex nonsmooth functions h.

3.1 Weakly convex function

Following (Moreau, 1965), one can verify that the strong prox-regularity holds for ρ-weakly convex functions if $t \leq 1/\rho$. The semismoothness of the proximal operator of a weakly convex function generally does not hold, which happens in the convex case as well. While two popular nonconvex regularizers for reducing bias are the minimax concave penalty (MCP) (Zhang, 2010) and the smoothly clipped absolute deviation (Fan, 1997), the semismoothness is satisfied. Specifically, the MCP is defined as

$$h_{\lambda,\theta}(x) := \begin{cases} \lambda|x| - \frac{x^2}{2\theta}, & |x| \leq \theta\lambda, \\ \frac{\theta\lambda^2}{2}, & \text{otherwise}, \end{cases}$$

where λ and θ are two positive parameters. It is weakly convex with modulus $\rho = \theta^{-1}$. If $t < \theta$, the closed-form expression of the proximal operator is

$$\mathrm{prox}_{th}(x) = \begin{cases} 0, & |x| < t\lambda, \\ \frac{x - \lambda t\,\mathrm{sign}(x)}{1-(t/\theta)}, & t\lambda \leq |x| \leq \theta\lambda, \\ x, & |x| > \theta\lambda. \end{cases}$$

The semismoothness property of the MCP regularizer is presented in (Shi et al., 2019). Analogously, one can also verify the weak convexity of the SCAD regularizer and the semismoothness of its proximal operator. We refer to (Böhm and Wright, 2021) and (Shi et al., 2019) for the details. Numerical results in (Shi et al., 2019) exhibit the efficiency of semismooth Newton methods.

3.2 Smooth and compact embedded manifold

Since any smooth manifold is a proximally smooth set, there exists a neighborhood $\mathcal{X}(r)$ of the form (2) such that the projection is single-valued and Lipschitz continuous (Clarke et al., 1995, Theorem 4.8). On the other hand, the projection onto smooth and compact embedded manifold is also a smooth mapping (Foote, 1984) on $\mathcal{X}(r)$. Putting them together, we conclude that the indicator function is strongly prox-regular and the

corresponding projection operator is smooth over $\mathcal{X}(r)$. For a special sphere-constrained smooth optimization problem, the Bose-Einstein condensates, we will show the numerical superiority of our proposed method using strong prox-regularity and semismoothness. For general smooth optimization problems with orthogonal constraints, we refer the reader to (Gawlik and Leok, 2017) for the calculations of the generalized Jacobian of the polar decomposition.

3.3 Two specific oblique manifold related nonconvex functions

We shall show that the nonconvex and nonsmooth functions (4) and (6) satisfy the strong prox-regularity and semismoothness.

Lemma 3 *The functions h defined in both* (4) *and* (6) *are strongly prox-regular and their proximal operators are semismooth with respect to their B-subdifferentials. Specifically,*

(i) *Let $\mathcal{C}_1 = \mathrm{Ob}(n,p)$, $\|V\|_{2,\infty} := \max_{i=1,2,\ldots,p} \|V_i\|_2$, and $\gamma_1 = \frac{1}{(\lambda+1)n}$. The function $h(X) = \lambda\|X\|_1 + \delta_{\mathrm{Ob}(n,p)}(X)$ is strongly prox-regular with respect to $\mathcal{C}_1$, γ_1, and $\|\cdot\|_{2,\infty}$. Moreover, the proximal mapping prox_{th} is semismooth over the set $\mathcal{D}_1 = \{X + tV : X \in \mathcal{C}_1, \|V\|_{2,\infty} = 1, 0 \le t \le \gamma_1\}$ with repsect to $\partial_B \mathrm{prox}_{th}$.*

(ii) *Let $\mathcal{C}_2 = \mathrm{Ob}^+(n,p)$ and $0 < \gamma_2 < 1$. The function $h(X) = \delta_{\mathrm{Ob}^+(n,p)}(X)$ is strongly prox-regular with respect to $\mathcal{C}_2$, γ_2, and $\|\cdot\|_{2,\infty}$. Moreover, the proximal mapping prox_{th} is semismooth over the set $\mathcal{D}_2 = \{X + tV : X \in \mathcal{C}_2, \|V\|_{2,\infty} = 1, 0 \le t \le \gamma_2\}$ with repsect to $\partial_B \mathrm{prox}_{th}$.*

Proof Let us prove (i) and (ii), respectively.

(i) Note that for any vector $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$, $\|x\|_\infty \ge 1/\sqrt{n}$. Following from the definition of the proximal mapping (4), we have for $t \le \gamma_1$, the proximal mapping prox_{th} is single-valued and Lipschitz continuous over $\mathcal{D}_1$.

Since the proximal mapping (4) is separable with respect to the columns in X, its semismoothness property can be reduced to the case of $p = 1$. Note that the non-differential points of prox_{th} are in the set $\mathcal{A} := \{x \in \mathbb{R}^n : \exists i,\ |x_i| = \lambda t\}$. At a nondifferentiable point $x \in \mathcal{A}$, let $d \in \mathbb{R}^n$ be a direction. Without loss of generality, assume $x_i = t\lambda$ and $|x_j| \ne t\lambda$ for all $j \ne i$. If $d_i > 0$, we have $\partial_B \mathrm{prox}_{th}(x+d) = \frac{\mathrm{diag}(1_{\tilde{w}<0})}{\|\tilde{w}^-\|_2} - \frac{\tilde{w}^-(\tilde{w}^-)^\top}{\|\tilde{w}^-\|_2^3} =: J(x+d)$, with $\tilde{w}^- = \min(\lambda t - |x+d|, 0) \odot \mathrm{sign}(x)$. Define $\tilde{d}_j = d_j$ if $j \ne i$ and 0 otherwise. Note that $\mathrm{prox}_{th}(x+d) = \mathrm{prox}_{th}(x+\tilde{d})$, $J(x+d) = J(x+\tilde{d})$ and $J(x+d)d = J(x+\tilde{d})\tilde{d}$. Thus,

$$\begin{aligned} &\|\mathrm{prox}_{th}(x+d) - \mathrm{prox}_{th}(x) - J(x+d)d\|_2 \\ =&\|\mathrm{prox}_{th}(x+\tilde{d}) - \mathrm{prox}_{th}(x) - J(x+\tilde{d})\tilde{d}\|_2 \\ =&o(\|\tilde{d}\|_2) = o(\|d\|_2). \end{aligned}$$

One can draw a similar conclusion for the case $d_i < 0$. Combining them together, we conclude that prox_{th} is semismooth.

(ii) It follows from the definition of the proximal mapping (6) that $\mathrm{prox}_{th}(X)$ is single-valued and Lipschitz continuous over $\mathcal{D}_2$. Analogous to the case above, one can prove the semismooth property of prox_{th}.

■

The strong prox-regularity and semismoothness established in the above lemma allow us to design efficient second-order methods for solving the applications in Subsection 1.1. Corresponding numerical experiments will be conducted in Section 6.

4. A projected Semismooth Newton method

To solve (1), the proximal gradient method is

$$x^{k+1} \in \arg\min_x \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2t_k}\|x - x^k\|_2^2 + h(x) = \mathrm{prox}_{t_k h}(x^k - t_k \nabla f(x^k)), \quad (14)$$

where $t_k > 0$ is the step size depending on the Lipschitz constant of ∇f. Since h is nonconvex, $\mathrm{prox}_{t_k h}$ is usually a set-valued mapping. To accelerate (14), the author (Yang, 2017) investigates the techniques of extrapolation and nonmonotone line search.

If h is strongly prox-regular with respect to $\mathcal{C} \supset \mathrm{dom}(h)$, γ, and $\|\cdot\|$, then $\mathrm{prox}_{th}(x^k - t\nabla f(x^k))$ is single-valued and Lipschitz continuous (SL) whenever $\|t\nabla f(x^k)\| \le \gamma$ and x^k belongs to the closed set $\mathcal{C}$. To ensure the compactness of the sequence $\{x^k\}$, one usually investigates the coercive property and the descent property of φ. Specifically, any level set $\{x : \varphi(x) \le \alpha\}$ with $\alpha \in \mathbb{R}$ is compact for a coercive φ. If the sequence $\{\varphi(x^k)\}$ is decreasing, $\{x^k\} \subset \{x : \varphi(x) \le \varphi(x^0)\}$ is a compact set. Moreover, the norm $\|\nabla f(x)\|$ is upper bounded by a finite constant $L > 0$ over $\{x : \varphi(x) \le \varphi(x^0)\}$ due to the smoothness. The proximal operator prox_{th} is SL if $t \le \frac{\gamma}{L}$. For this choice of t, we are able to design a second-order method to solve the fixed point equation:

$$0 = F(x) := x - \mathrm{prox}_{th}(x - t\nabla f(x)), \quad (15)$$

where t is set as $\min\{\gamma, 1\}/L$. It follows the SL property of prox_{th} and twice continuous differentiability of f that F is single-valued, Lipschitz continuous, and semismooth.

In what follows, we assume that prox_{th} is semismooth with respect to its B-subdifferential. Then, F is semismooth with respect to $M(x)$. This allows us to design a semismooth Newton method for solving (1). One typical benefit of second-order methods is the superlinear or faster local convergence rate. Specifically, we first solve the linear system

$$(M_k + \mu_k I)d^k = -F(x^k), \quad (16)$$

where $M_k \in M(x^k)$ defined by (13) is a generalized Jacobian and $\mu_k = \kappa\|F(x^k)\|_2$ with a positive constant κ. Note that the shift term $\mu_k I$ can be used to promote the positive definiteness of the coefficient matrix of (16), particularly in the convex setting (Xiao et al., 2018; Li et al., 2018b). The semismooth Newton step is then defined as

$$z^k = \mathcal{P}_{\mathrm{dom}(h)}(x^k + d^k), \quad (17)$$

where the projection onto $\mathrm{dom}(h)$ is necessary for the globalization due to the nonconvexity of h. We remark that the strong prox-regularity in Definition 1 is crucial for the design

of semismooth Newton methods. For a general prox-regular function h, we know from Proposition 2 that for $v \in \partial h(x)$, the proximal operator prox_{th} is a singleton and Lipschitz continuous around $x+tv$ for sufficiently small t. Since $\nabla f(x)$ could be far away from $\partial h(x)$, the proximal operator $\text{prox}_{th}(x - t\nabla f(x))$ may not be a singleton. On the other hand, a uniform t for all iterates may not exist. This non-singleton property causes difficulty in designing second-order methods.

Note that the pure semismooth Newton step is generally not guaranteed to converge from arbitrary starting points. For globalization, we switch to the proximal gradient step when the semismooth Newton step does not decrease the norm of the residual (15) or increases the objective function value to a certain amount. To be specific, the Newton step z^k is accepted if the following conditions are simultaneously satisfied:

$$\|F(z^k)\|_2 \leq \nu\rho_k, \tag{18}$$

$$\varphi(z^k) \leq \varphi(x^k) + \eta\rho_k^{1-q}\|F(z^k)\|_2^q, \tag{19}$$

where ρ_k is the norm of the residual of the last accepted Newton iterate until k with an initialization $\rho_0 > 0$, $\eta > 0$, and $\nu, q \in (0,1)$. Otherwise, the semismooth Newton step z^k fails, and we do a proximal gradient step, i.e.,

$$x^{k+1} = \text{prox}_{th}(x^k - t\nabla f(x^k)) = x^k - F(x^k). \tag{20}$$

Due to the choice of $t = \min\{\gamma, 1\}/L$, we will show in the next section that there is a sufficient decrease in the objective function value $\varphi(x^{k+1})$. Under the BD-regularity condition (Any element of $\partial_B F(x^*)$ at the stationary point x^* is nonsingular (Qi, 1993; Pang and Qi, 1993)), we show in the next section that the semismooth Newton steps will always be accepted when the iterates are close to the optimal solution. The proposed switching between the Newton step and the proximal gradient step ensures that its theoretical convergence is independent of the specific value chosen for $\kappa > 0$ in (16). However, selecting an appropriate κ is beneficial for achieving satisfactory numerical performance. The detailed algorithm is presented in Algorithm 1.

Algorithm 1 A projected semismooth Newton method for solving (1)

Input: The constants $L > 0$, $\gamma > 0$, $\nu \in (0,1), q \in (0,1), \eta > 0, \rho_0 > 0$, $\kappa > 0$, and an initial point $x^0 \in \mathbb{R}^n$, set $k = 0$.

1: **while** the condition is not met **do**

2: Calculate the semismooth Newton direction d^k by solving the linear equation

$$(M(x^k) + \mu_k I)d^k = -F(x^k).$$

3: Set $z^k = \mathcal{P}_{\text{dom}(h)}(x^k + d^k)$. If the conditions (18) and (19) are satisfied, set $x^{k+1} = z^k$. Otherwise, set $x^{k+1} = x^k - F(x^k)$.

4: Set $k = k + 1$.

5: **end while**

5. Convergence analysis

In this section, we will present the convergence properties of the proposed projected semismooth Newton method, i.e., Algorithm 1. It consists of two parts, the global convergence to a stationary point from any starting point and the local superlinear convergence.

5.1 Global convergence

First of all, we introduce the following assumptions.

Assumption 4 *For problem* (1)*, we assume*

- *the function f is twice continuously differentiable, its gradient ∇f is Lipschitz continuous with modulus $L > 0$.*
- *the function h is strongly prox-regular with respect to $\mathcal{C}$ and γ.*
- *the function φ is bounded from below and coercive.*

With the above assumption, the proximal gradient step (20) leads to a sufficient decrease on φ.

Lemma 5 *Suppose that Assumption 4 holds. Then for any $t_k \in (0, \frac{1}{L}]$ we have*

$$\varphi(x^k) - \varphi(x^{k+1}) \geq \left(\frac{1}{2t_k} - \frac{L}{2}\right) \|x^{k+1} - x^k\|_2^2. \tag{21}$$

Proof It follows from the optimality of x^{k+1} that

$$\left\langle \nabla f(x^k), x^{k+1} - x^k \right\rangle + \frac{1}{2t_k}\|x^{k+1} - x^k\|_2^2 + h(x^{k+1}) \leq h(x^k).$$

By Assumption 4 and $t_k \in (0, \frac{1}{L})$, we have

$$\begin{aligned} f(x^{k+1}) + h(x^{k+1}) &\leq f(x^k) + \left\langle \nabla f(x^k), x^{k+1} - x^k \right\rangle + \frac{L}{2}\|x^{k+1} - x^k\|_2^2 + h(x^{k+1}) \\ &\leq f(x^k) + h(x^k) + \left(\frac{L}{2} - \frac{1}{2t_k}\right) \|x^{k+1} - x^k\|_2^2. \end{aligned}$$

The proof is completed. ■

From the above lemma, the convergence of the proximal gradient method for solving (1) can be obtained by the coercive property of φ. When the projected semismooth Newton update z^k is accepted, the function value $\varphi(z^k)$ may increase while the residual decreases as guaranteed by (18) and (19). This allows us to show global convergence.

Theorem 6 *Let $\{x^k\}$ be the iterates generated by Algorithm 1. Suppose that Assumption 4 holds. Let $t_k \equiv t \in (0, \min(\gamma, 1)/L]$, Then we have*

$$\lim_{k\to\infty} \|F(x^k)\|_2 = 0.$$

Proof If x^{k+1} is obtained by the proximal gradient update, it holds from Lemma 5 that

$$\varphi(x^k) - \varphi(x^{k+1}) \geq \left(\frac{1}{2t} - \frac{L}{2}\right) \|F(x^k)\|_2^2. \tag{22}$$

It follows the Lipschitz properties of prox_{th} and $\nabla f(x)$ that F is Lipschitz continuous. Let L_F be the Lipschitz constant of F. From the triangle inequality, we have

$$\|F(x^{k+1})\|_2 \leq \|F(x^k)\|_2 + \|F(x^{k+1}) - F(x^k)\|_2 \leq (L_F + 1)\|F(x^k)\|_2.$$

Plugging the above inequality into (22) leads to

$$\varphi(x^k) - \varphi(x^{k+1}) \geq c_1 \|F(x^{k+1})\|_2^2, \tag{23}$$

where $c_1 := \left(\frac{1}{2t} - \frac{L}{2}\right) \frac{1}{(L_F+1)^2} > 0$.

If the Newton update z^k is accepted, the conditions (18) and (19) imply that

$$\begin{aligned} \varphi(x^k) - \varphi(x^{k+1}) &\geq -\eta \rho_k^{1-q} \|F(x^{k+1})\|_2^q \\ &= c_1 \|F(x^{k+1})\|_2^2 - (c_1 \|F(x^{k+1})\|_2^{2-q} + \eta \rho_k^{1-q}) \|F(x^{k+1})\|_2^q \end{aligned}$$

and $\rho_{k+1} = \|F(x^{k+1})\|_2 \leq \nu \rho_k$. Since $\rho_k \in (0, \rho_0)$ for all k, $c_1 \|F(x^{k+1})\|_2^{2-q} + \eta \rho_k^{1-q}$ is bounded by a constant, denoted by c_2. Hence, for the projected semismooth Newton step, it holds

$$\varphi(x^k) - \varphi(x^{k+1}) \geq c_1 \|F(x^{k+1})\|_2^2 - c_2 \rho_{k+1}^q. \tag{24}$$

Combining (23) and (24), we have

$$\varphi(x^0) - \varphi(x^{K+1}) = \sum_{i=1}^{K} (\varphi(x^k) - \varphi(x^{k+1})) \geq c_1 \sum_{k=0}^{K} \|F(x^{k+1})\|_2^2 - c_2 \sum_{k \in \mathcal{K}_N} \rho_{k+1}^q,$$

where $\mathcal{K}_N \subset \{1, 2, \ldots, K+1\}$ consists of the indices where the projected semismooth Newton updates are accepted. It is easy to see that $\sum_{k \in \mathcal{K}_N} \rho_{k+1}^q \leq \rho_0^q \sum_{k=1}^{K+1} \nu^{qk} = \frac{\rho_0^q (1 - \nu^{q(K+1)})}{1-\nu} \leq \frac{\rho_0^q}{1-\nu^q}$. Therefore,

$$c_1 \sum_{k=0}^{K} \|F(x^{k+1})\|_2^2 \leq \varphi(x^0) - \varphi(x^{K+1}) + \frac{c_2 \rho_0^p}{1 - \nu^p}.$$

Since φ is bounded from below, we have

$$\sum_{k=0}^{\infty} \|F(x^k)\|_2^2 < \infty,$$

which implies that $\lim_{k \to \infty} \|F(x^k)\|_2 = 0$. We complete the proof. ■

5.2 Local superlinear convergence

The local superlinear convergence of the semismooth Newton update has been studied in (Qi and Sun, 1993, 1999; Xiao et al., 2018). The difficulties in our case lie in the extra nonconvex projection operator $\mathcal{P}_{\mathrm{dom}(h)}$ and the switching conditions (18) and (19). We make the following assumptions.

Assumption 7 *Let $\{x^k\}$ be the iterates generated by Algorithm 1.*

(A1) *The iterate x^k converges to x^* with $F(x^*)=0$, as $k\to\infty$.*

(A2) *The Hessian $\nabla^2 f$ is continuous around x^*.*

(A3) *The mapping F is semismooth at x^* with respect to $M(x)$. In addition, there exists $C>0$ such that each element $M\in M(x^*)$ defined by (13) is nonsingular with $\|M^{-1}\|_2\le C$.*

(A4) *The function φ is Lipschitz continuous over $\mathrm{dom}(h)$ with modulus L_φ, i.e., for all $x,y\in\mathrm{dom}(h)$,*

$$|\varphi(x)-\varphi(y)|\le L_\varphi\|x-y\|_2.$$

Since the convergence of $\{\|F(x^k)\|_2\}$ is proved in Theorem 6, any accumulation point of $\{x^k\}$ has zero residual. The Assumption (A1) reads that the full sequence $\{x^k\}$ is convergent. The Assumption (A2) holds for any twice continuously differentiable f. The Assumption (A3) is the standard BD-regularity condition used in (Qi, 1993; Pang and Qi, 1993; Milzarek and Ulbrich, 2014; Xiao et al., 2018).

For the projection operator $\mathcal{P}_{\mathrm{dom}(h)}$ in Algorithm 1, we prove the following bounded property, which has also been used in the convergence rate analysis for the generalized power method for the group synchronization problems (Liu et al., 2017b, Lemma 1) (Liu et al., 2017a, Proposition 3.3) (Liu et al., 2020, Lemma 2).

Proposition 8 *For all $x\in\mathbb{R}^n$ and $y\in\mathrm{dom}(h)$, it holds $\|\mathcal{P}_{\mathrm{dom}(h)}(x)-y\|_2\le 2\|x-y\|_2$.*

Proof Following the definition of $\mathcal{P}_{\mathrm{dom}(h)}$, we have

$$\|\mathcal{P}_{\mathrm{dom}(h)}(x)-y\|_2\le\|\mathcal{P}_{\mathrm{dom}(h)}(x)-x\|_2+\|x-y\|_2\le 2\|x-y\|_2.$$

■

The following lemma shows that the switching conditions (18) and (19) are satisfied by the projected semismooth Newton update when k is large enough.

Lemma 9 *Let $\{x^k\}$ be the iterates generated by Algorithm 1. Suppose that Assumptions 4 and 7 hold. Then for sufficiently large k, the Newton update z^k is always accepted.*

Proof Let us first define a constant $\gamma_F\in\left(0,\min\left\{\frac{1}{8C},\frac{\nu}{32C^2L_F},\frac{\eta^{\frac{1}{1-q}}}{32C^2(L_\varphi 3^q C^q)^{\frac{1}{1-q}}}\right\}\right)$, where $C,\nu,\eta,q,L_F,L_\varphi$ are defined previously. It follows from (Qi, 1993, Lemma 2.6) and (A3) that there exists $\varepsilon>0$ such that for any $x\in\mathbb{B}(x^*,\varepsilon)$ and $M\in M(x)$,

$$\|F(x)-F(x^*)-(M+\kappa\|F(x)\|_2 I)(x-x^*)\|_2\le\gamma_F\|x-x^*\|_2,\ \|(M+\kappa\|F(x)\|I)^{-1}\|_2\le 2C. \tag{25}$$

For the projected semismooth Newton update $z^k = \mathcal{P}_{\mathrm{dom}(h)}(x^k - (M_k + \mu_k I)^{-1}F(x^k))$, it hold that

$$\begin{aligned}\|z^k - x^*\|_2 &= \|\mathcal{P}_{\mathrm{dom}(h)}(x^k - (M_k + \mu_k I)^{-1}F(x^k)) - x^*\|_2 \\ &\leq 2\|(M_k + \mu_k I)^{-1}(F(x^k) - F(x^*) - (M_k + \mu_k I)(x^k - x^*))\|_2 \\ &\leq 4\gamma_F C\|x^k - x^*\|_2,\end{aligned} \tag{26}$$

where we assume $x^k \in \mathbb{B}(x^*, \varepsilon)$. Due to the choice of γ_F, we have $z^k \in \mathbb{B}(x^*, \varepsilon)$. Note that

$$\|x^k - x^*\|_2 \leq \|z^k - x^*\|_2 + \|z^k - x^k\|_2 \leq 4\gamma_F C\|x^k - x^*\|_2 + 4C\|F(x^k)\|_2. \tag{27}$$

Then

$$\|x^k - x^*\|_2 \leq \frac{4C}{1 - 4\gamma_F C}\|F(x^k)\|_2. \tag{28}$$

Combining (26) and (28) implies

$$\|z^k - x^*\|_2 \leq \frac{16\gamma_F C^2}{1 - 4\gamma_F C}\|F(x^k)\|_2. \tag{29}$$

Hence,

$$\|F(z^k)\|_2 = \|F(z^k) - F(x^*)\|_2 \leq L_F\|z^k - x^*\|_2 \leq \frac{16\gamma_F C^2 L_F}{1 - 4\gamma_F C}\|F(x^k)\|_2 \leq \nu\|F(x^k)\|_2. \tag{30}$$

In addition, note that

$$\begin{aligned}\|z^k - x^*\|_2 &= \|(M_k + \mu_k I)^{-1}\left(F(z^k) - F(x^*) - (M_k + \mu_k I)(z^k - x^*) - F(z^k)\right)\|_2 \\ &\leq 2\gamma_F C\|z^k - x^*\|_2 + 2C\|F(z^k)\|_2.\end{aligned}$$

This gives

$$\|z^k - x^*\|_2 \leq \frac{2C}{1 - 2\gamma_F C}\|F(z^k)\|_2. \tag{31}$$

The changes between $\varphi(z^k)$ and $\varphi(x^k)$ can be estimated by

$$\begin{aligned}\varphi(z^k) - \varphi(x^k) &\leq \varphi(z^k) - \varphi(x^*) \leq L_\varphi\|z^k - x^*\|_2 \\ &= L_\varphi\|z^k - x^*\|_2^{1-q}\|z^k - x^*\|_2^q \\ &\leq L_\varphi\left(\frac{16\gamma_F C^2}{1 - 4\gamma_F C}\right)^{1-q}\left(\frac{2C}{1 - 2\gamma_F C}\right)^q\|F(x^k)\|_2^{1-q}\|F(z^k)\|_2^q \\ &\leq \eta\|F(x^k)\|_2^{1-q}\|F(z^k)\|_2^q.\end{aligned} \tag{32}$$

Due to the convergence of residual, for any proximal gradient step index k_0, there always exists a $k > k_0$ such that $\|F(x^k)\|_2 \leq \rho_k$. Then all followed iterates are projected semismooth Newton steps because of (30) and (32). This completes the proof. ■

The above lemma establishes the local transition to the projected semismooth Newton step. Utilizing the semismoothness, we have the locally superlinear convergence on the iterates generated by Algorithm 1.

Theorem 10 *Let $\{x^k\}$ be the iterates generated by Algorithm 1. Suppose that Assumptions 4 and 7 hold. Then there exists a finite $K > 0$, such that for all $k \geq K$, $\{x^k\}$ converges to x^* Q-superlinearly.*

Proof From Lemma 9, there exists a K such that the projected semismooth Newton update is accepted for $k \geq K$. It follows from the semismoothness of F that

$$\begin{aligned}\|x^{k+1} - x^k\|_2 &= \|\mathcal{P}_{\mathrm{dom}(h)}(x^k - (M_k + \mu_k I)^{-1}F(x^k)) - x^*\|_2 \\ &\leq 4C\|F(x^k) - F(x^*) - (M_k + \mu_k I)(x^k - x^*)\|_2 \\ &= o(\|x^k - x^*\|_2),\end{aligned}$$

where we use $\mu_k = \kappa\|F(x^k)\|_2$ and $F(x^k) \to 0$ (i.e., (A1)) for the last equality. This means $\{x^k\}$ converges to x^* Q-superlinearly. ■

6. Numerical experiments

In this section, some numerical experiments are presented to evaluate the performance of our proposed Algorithm 1, denoted by ProxSSN. We compare ProxSSN with the existing methods including AManPG and ARPG (Huang and Wei, 2021). We also test the proximal gradient descent method (ProxGD for short) as in (14). Here, a nonmonotone line search with Barzilai–Borwein (BB) step size (Barzilai and Borwein, 1988) is used for acceleration. Let $s^k = x^k - x^{k-1}$ and $y^k = \nabla f(x^k) - \nabla f(x^{k-1})$. The BB step sizes are defined as

$$\beta_k^1 = \frac{\langle s^k, s^k\rangle}{|\langle s^k, y^k\rangle|}, \text{ and } \beta_k^2 = \frac{|\langle s^k, y^k\rangle|}{\langle y^k, y^k\rangle}. \tag{33}$$

Given $\varrho, \delta \in (0,1)$, the nonmonotone Armijo line search is to find the smallest nonnegative integer ℓ satisfying

$$\varphi(\mathrm{prox}_{t_k(\ell)h}(x^k - t_k(\ell)\nabla f(x^k))) \leq C_k + \frac{\varrho}{2t_k(\ell)}\|\mathrm{prox}_{t_k(\ell)h}(x^k - t_k(\ell)\nabla f(x^k)) - x_k\|_2^2. \tag{34}$$

Here, $t_k(\ell) := \beta_k\delta^\ell$, β_k is set to β_k^1 and β_k^2 alternatively, and the reference value C_k is calculated via $C_k = (\varpi Q_{k-1}C_{k-1} + \varphi(x^k))/Q_k$ where $\varpi \in [0,1]$, $C_0 = \varphi(x^0)$, $Q_k = \varpi Q_{k-1} + 1$ and $Q_0 = 1$. Once ℓ is obtained, we set $t_k = \beta_k\delta^\ell$ and the next iterate is then given by $x^{k+1} = \mathrm{prox}_{t_k h}(x^k - t_k\nabla f(x^k))$.

The reasons of not using ManPG (Chen et al., 2020), RPG (Huang and Wei, 2021) or the algorithms proposed in (Lai and Osher, 2014; Kovnatsky et al., 2016) is that their performance can not measure up with AManPG or ARPG in tests of (Huang and Wei, 2021). For ARPG and AManPG, we use the code provided by (Huang and Wei, 2021). The codes were written in MATLAB and run on a standard PC with 3.00 GHz AMD R5 microprocessor and 16GB of memory. The reported time is wall-clock time in seconds.

. all codes are available at `https://www.math.fsu.edu/~whuang2/files/RPG_v0.2.zip`

6.1 Sparse principal component analysis

In this subsection, we consider the sparse PCA problem (3), which can be regarded as a nonsmooth problem on the oblique manifold. Let $f(X) := \|X^T A^T A X - D^2\|_F^2$. AManPG solves the following subproblem in each iteration:

$$\eta_{X^k} = \underset{\eta \in T_{X^k}\mathrm{Ob}(n,p)}{\arg\min} \left\langle \mathrm{grad} f(X^k), \eta \right\rangle + \frac{\tilde{L}}{2}\|\eta\|_F^2 + \lambda \|X^k + \eta\|_1,$$

where $\tilde{L} > L$ with L being the Lipschitz constant of f, $\mathrm{grad} f(X^k)$ denotes the Riemannian gradient of f at X^k, and $T_X\mathrm{Ob}(n,p)$ is the tangent space to $\mathrm{Ob}(n,p)$ at X. We refer to (Chen et al., 2020) for more details. In the k-th iteration of ARPG, one needs to solve the subproblem:

$$\eta_{X^k} = \underset{\eta \in T_{X^k}\mathrm{Ob}(n,p)}{\arg\min} \left\langle \mathrm{grad} f(X^k), \eta \right\rangle + \frac{\tilde{L}}{2}\|\eta\|_F^2 + \lambda \|\mathcal{R}_{X^k}(\eta)\|_1,$$

where $\mathcal{R}$ denotes a retraction operator on $\mathrm{Ob}(n,p)$. The termination condition of both AManPG and ARPG is as follows:

$$\|\tilde{L}\eta_{X^k}\|_F^2 \le \mathrm{tol}, \tag{35}$$

where $\mathrm{tol} > 0$ is a given tolorance. The ProxGD and ProxSSN methods are applied to solve problem (3) by setting $f(X) := \|X^T A^T A X - D^2\|_F^2$, $h(X) = \lambda\|X\|_1 + \delta_{\mathrm{Ob(n,p)}}(X)$. ProxGD has the following update rule

$$X^{k+1} = \mathrm{prox}_{t_k h}(X^k - t_k \nabla f(X^k)).$$

The following relative KKT condition is set as a stopping criterion for our algorithm and ProxGD:

$$\mathrm{err} := \frac{\left\|X^k - \mathrm{prox}_{t_k h}(X^k - t_k \nabla f(X^k))\right\|_F}{t_k(1 + \|X^k\|_F)} \le \mathrm{tol}. \tag{36}$$

Note that t_k is fixed in ProxSSN. Based on Lemma 3, we can calculate the proximal mapping and its generalized Jacobian in our ProxSSN at a low cost.

Implementation details The parameters of AManPG and ARPG are set the same as in (Huang and Wei, 2021). For ProxSSN, we set $q = 20, \nu = 0.9999, \eta = 10^{-6}, t = 1/\lambda_{\max}(A^T A)$, and the initial value $\kappa = 1$. The maximum number of iterations is 10000. The starting point of all algorithms is the leading p right singular vectors of the matrix A. Due to the evaluation criterion being different for different algorithms, we first run ARPG when (35) is satisfied with $\mathrm{tol} = 10^{-10} \times n \times p$ or the number of iterations exceeds 10000, and denote F_{ARPG} as the obtained objective value. The other algorithms are terminated when the objective value satisfies $F(X^k) \le F_{\mathrm{ARPG}} + 10^{-6}$ or (35) (or (36)) is satisfied with $\mathrm{tol} = 10^{-10} \times n \times p$, or the number of iterations exceeds 10000.

In our experiments, the data matrix $A \in \mathbb{R}^{m \times n}$ is produced by MATLAB function $randn(m, n)$, in which all entries of A follow the standard Gaussian distribution. Next, we shift the columns of A such that they have zero-mean, and normalize the resulting matrix by its spectral norm.

6.1.1 Numerical results

In Figure 1, we present the trajectories of the objective function values with respect to the wall-clock time for the cases of $n = 300$ and $n = 400$, where $\varphi_{\min}$ is the minimum objective value of all algorithms in the iterative process. It can be seen that our proposed ProxSSN converges fastest among all algorithms. AManPG and ARPG have comparable performances. Figures 2 and 3 shows the performance of all algorithms under different n, p. We see that all algorithms have similar objective values, but the consuming time of ProxSSN is the least. We present the wall-clock time in the column "time" and the objective function value in the column "obj" in Table 1 for different combinations of m, n, p, where similar conclusions can be drawn. It should be noted that computational time for larger values of n or p may decrease as the stopping criterion, defined by $\text{tol} = 10^{-10} \times n \times p$, varies with n and p.

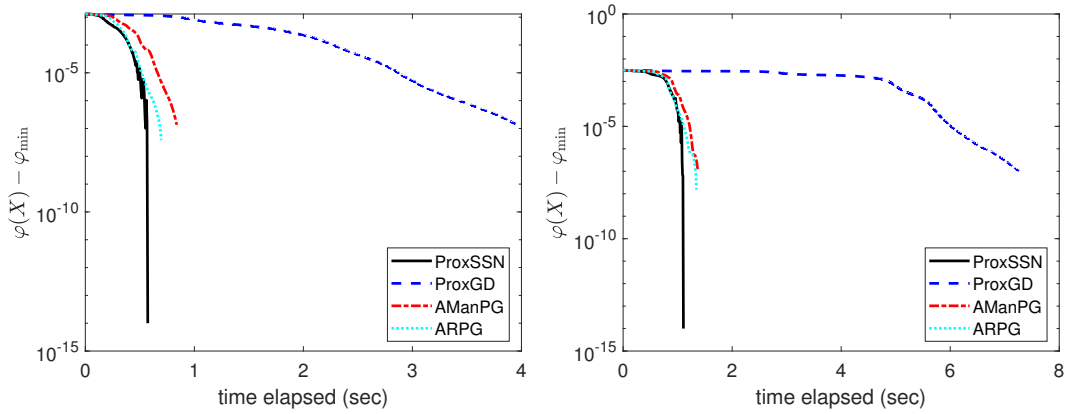

Figure 1: The trajectories of the objective function values with respect to the wall-clock time on the sparse PCA problem (3) with $p = 10, \lambda = 0.01$. Left: $n = 300$; right: $n = 400$

We also compare the accuracy and efficiency of ProxSSN with other algorithms using the performance profiling method proposed in (Dolan and Moré, 2002). Let $t_{i,s}$ be some performance quantity (e.g. the wall-clock time or the gap between the obtained objective function value and $\varphi_{\min}$, lower is better) associated with the s-th solver on problem i. Then, one computes the ratio $r_{i,s}$ as $t_{i,s}$ over the smallest value obtained by n_s solvers on problem i, i.e., $r_{i,s} := \frac{t_{i,s}}{\min\{t_{i,s}: 1 \le s \le n_s\}}$. For $\tau > 0$, the value

$$\pi_s(\tau) := \frac{\text{number of problems where } \log_2(r_{i,s}) \le \tau}{\text{total number of problems}}$$

indicates that solver s is within a factor $2^\tau \ge 1$ of the performance obtained by the best solver. Then the performance plot is a curve $\pi_s(\tau)$ for each solver s as a function of τ. In Figure 4, we show the performance profiles of the criterion, the wall-clock time and the gap in the objective function values. In particular, the intercept point of the axis "ratio of problems" and the curve in each subfigure is the percentage of the faster one among the

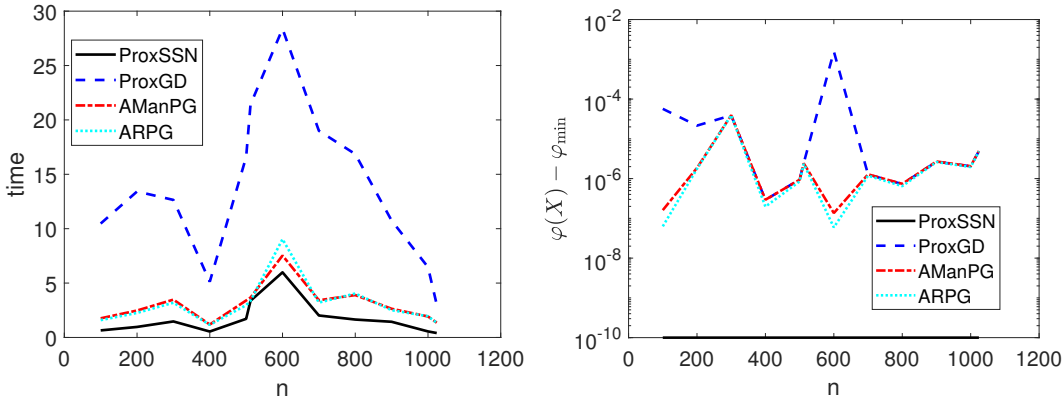

Figure 2: Comparisons of wall-clock time and the objective function values on the sparse PCA problem (3) with $p = 20, \lambda = 0.01$ for different n.

four solvers. These figures show that both the wall-clock time and the gap in the objective function values of ProxSSN are much better than other algorithms on most problems.

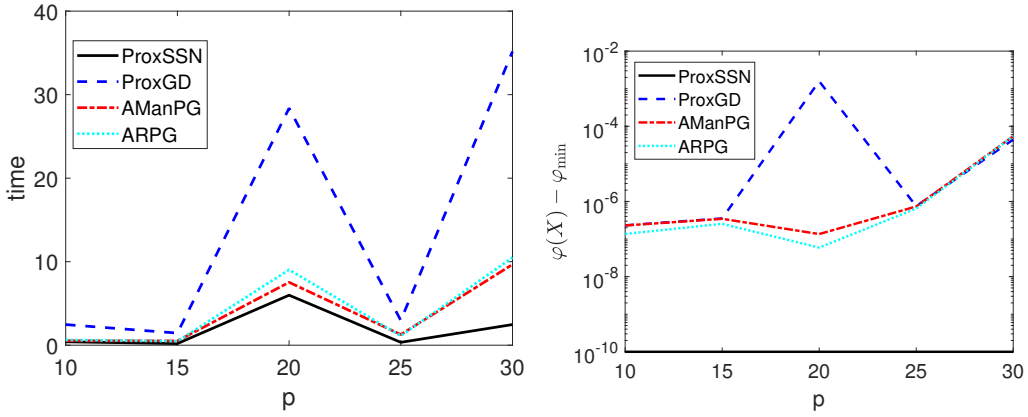

Figure 3: Comparisons of wall-clock time and the objective function values on the sparse PCA problem (3) with $n = 512, \lambda = 0.01$ for different p.

Table 1: Computational results of oblique SPCA

(m, n, p)	ProxSSN		ProxGD		AManPG		ARPG	
	time	obj	time	obj	time	obj	time	obj
100 / 500 / 10	1.58	1.28380	13.59	1.28380	1.99	1.28380	1.75	1.28380
100 / 500 / 15	1.16	1.85986	7.05	1.85986	1.82	1.85986	1.696	1.85986
100 / 500 / 20	1.71	2.44963	16.59	2.44963	3.40	2.44963	2.96	2.44963
100 / 500 / 25	2.36	3.00555	15.66	3.00555	4.05	3.00555	3.97	3.00555
100 / 500 / 30	0.84	3.58139	12.15	3.58139	3.21	3.58139	3.16	3.58139
100 / 600 / 10	0.40	1.39524	2.47	1.39524	0.53	1.39524	0.62	1.39524
100 / 600 / 15	0.19	2.04237	1.45	2.04237	0.51	2.04237	0.48	2.04237

(m,n,p)	ProxSSN		ProxGD		AManPG		ARPG	
	time	obj	time	obj	time	obj	time	obj
100 / 600 / 20	5.98	2.68717	28.37	2.68875	7.53	2.68717	9.03	2.68717
100 / 600 / 25	0.34	3.31583	2.96	3.31583	1.27	3.31583	1.11	3.31583
100 / 600 / 30	2.47	3.93575	35.16	3.93579	9.68	3.93580	10.49	3.93580
100 / 700 / 10	1.50	1.50657	8.51	1.50657	1.65	1.50657	1.68	1.50657
100 / 700 / 15	0.60	2.21769	2.61	2.21769	0.80	2.21769	0.84	2.21769
100 / 700 / 20	2.02	2.92664	19.00	2.92664	3.42	2.92664	3.20	2.92664
100 / 700 / 25	2.22	3.59936	21.64	3.59936	5.04	3.59936	4.55	3.59936
100 / 700 / 30	2.44	4.23529	42.76	4.23540	6.04	4.23529	5.07	4.23529
100 / 800 / 10	0.27	1.60610	1.64	1.60610	0.45	1.60610	0.53	1.60610
100 / 800 / 15	0.46	2.36806	4.67	2.36806	0.87	2.36806	0.91	2.36806
100 / 800 / 20	1.64	3.09902	16.86	3.09902	3.89	3.09902	4.03	3.09902
100 / 800 / 25	1.38	3.82806	19.20	3.82806	4.08	3.82806	3.98	3.82806
100 / 800 / 30	5.77	4.55643	41.61	4.55681	13.49	4.55644	12.97	4.55644
100 / 900 / 10	0.76	1.71069	4.21	1.71069	0.80	1.71069	0.90	1.71069
100 / 900 / 15	0.28	2.51949	2.68	2.51949	0.93	2.51949	0.88	2.51949
100 / 900 / 20	1.44	3.28293	10.74	3.28294	2.63	3.28294	2.50	3.28294
100 / 900 / 25	1.66	4.09218	22.07	4.09218	6.50	4.09218	6.23	4.09218
100 / 900 / 30	3.70	4.81562	38.57	4.81896	13.25	4.81563	13.80	4.81563
100 / 1000 / 10	1.42	1.80718	10.43	1.80718	1.81	1.80718	1.69	1.80718
100 / 1000 / 15	2.38	2.64274	19.57	2.64274	3.65	2.64274	3.45	2.64274
100 / 1000 / 20	0.55	3.47447	6.46	3.47447	1.92	3.47447	1.94	3.47447
100 / 1000 / 25	2.23	4.25629	29.63	4.25629	6.83	4.25629	6.84	4.25629
100 / 1000 / 30	5.37	5.08015	44.92	5.08103	17.01	5.08015	15.79	5.08015

6.2 Sparse least square regression

In this subsection, we consider the sparse least-square problem (8), which can be regarded as a nonsmooth problem on the oblique manifold. We test the same algorithms as in subsection 6.1 for the comparisons. All parameters and strategies follow the setup discussed in the last subsection except tol $= 10^{-10}nm$. The numerical results are presented in Figures 5-7. In general, the overall performance of different methods is similar to the results shown in the last subsection. It is clear that ProxSSN is the fastest method for solving problem (8), both in terms of the objective function value and the wall-clock time. Table 2 shows the detailed results for different combinations of m, n. We see that ProxSSN compares favorably with the other algorithms and outperforms the first-order algorithm ProxGD.

Table 2: Computational results of least square regression

(m,n)	ProxSSN		ProxGD		AManPG		ARPG	
	time	obj	time	obj	time	obj	time	obj
20 / 3000	0.04	3.43796e-02	2.92	3.43839e-02	0.07	3.43799e-02	0.07	3.43798e-02
20 / 3200	0.18	3.39873e-02	1.07	3.39886e-02	0.08	3.39886e-02	0.09	3.39885e-02
20 / 3400	0.11	3.23110e-02	5.61	3.23240e-02	0.26	3.23123e-02	0.29	3.23122e-02
20 / 3600	0.17	3.17896e-02	5.18	3.20365e-02	0.24	3.20365e-02	0.27	3.20364e-02
20 / 3800	0.05	3.43032e-02	5.94	3.43061e-02	0.24	3.43048e-02	0.27	3.43047e-02
20 / 4000	0.09	3.44652e-02	6.07	3.54900e-02	0.36	3.44664e-02	0.42	3.44663e-02
20 / 4200	0.21	3.60764e-02	6.34	3.67852e-02	0.43	3.60786e-02	0.50	3.60785e-02
20 / 4400	0.11	3.36402e-02	6.49	3.68569e-02	0.60	3.36402e-02	0.80	3.36402e-02
20 / 4600	0.13	3.39844e-02	6.59	3.71441e-02	0.92	3.35500e-02	1.32	3.35500e-02
20 / 4800	0.16	3.40047e-02	6.90	3.40144e-02	0.34	3.40059e-02	0.42	3.40058e-02

(m,n)	ProxSSN		ProxGD		AManPG		ARPG	
	time	obj	time	obj	time	obj	time	obj
20 / 5000	0.06	3.32278e-02	6.90	3.54494e-02	0.86	3.32286e-02	1.22	3.32285e-02
30 / 3000	0.05	3.73733e-02	3.28	3.85623e-02	0.18	3.73735e-02	0.18	3.73734e-02
30 / 3200	0.05	3.46184e-02	5.82	3.51461e-02	0.28	3.46273e-02	0.31	3.46273e-02
30 / 3400	0.08	3.57899e-02	2.21	3.59235e-02	0.14	3.59235e-02	0.15	3.59234e-02
30 / 3600	0.17	3.73116e-02	3.56	3.73122e-02	0.17	3.73122e-02	0.20	3.73121e-02
30 / 3800	0.14	3.76258e-02	6.57	3.90207e-02	0.63	3.76263e-02	0.83	3.76263e-02
30 / 4000	0.03	4.06294e-02	6.83	4.08145e-02	0.31	4.08106e-02	0.37	4.08105e-02
30 / 4200	0.08	3.96908e-02	6.98	4.07081e-02	0.40	3.96931e-02	0.48	3.96930e-02
30 / 4400	0.03	3.95462e-02	7.57	3.95534e-02	0.27	3.95500e-02	0.30	3.95500e-02
30 / 4600	0.10	3.55181e-02	5.54	3.55193e-02	0.20	3.55193e-02	0.22	3.55192e-02
30 / 4800	0.11	3.85425e-02	7.67	3.92473e-02	0.59	3.85426e-02	0.81	3.85425e-02
30 / 5000	0.14	3.95414e-02	8.00	4.16688e-02	0.77	3.95439e-02	1.15	3.95438e-02
50 / 3000	0.05	4.14906e-02	4.46	4.16952e-02	0.23	4.14908e-02	0.24	4.14907e-02
50 / 3200	0.03	4.08372e-02	2.60	4.08408e-02	0.17	4.08407e-02	0.18	4.08407e-02
50 / 3400	0.12	4.53565e-02	5.64	4.58502e-02	0.18	4.58502e-02	0.19	4.58501e-02
50 / 3600	0.05	4.52462e-02	8.17	4.57722e-02	0.35	4.52464e-02	0.43	4.52463e-02
50 / 3800	0.06	4.12851e-02	3.47	4.12852e-02	0.19	4.12852e-02	0.23	4.12851e-02
50 / 4000	0.08	4.44167e-02	10.12	4.40984e-02	0.82	4.40979e-02	0.40	4.40983e-02
50 / 4200	0.23	4.16107e-02	10.48	4.16618e-02	0.60	4.16623e-02	1.26	4.16622e-02
50 / 4400	0.13	4.37490e-02	13.72	4.37491e-02	0.62	4.37491e-02	1.13	4.37490e-02
50 / 4600	0.07	4.41428e-02	3.45	4.41463e-02	0.31	4.41463e-02	0.42	4.41462e-02
50 / 4800	0.44	4.40181e-02	13.08	4.49775e-02	0.96	4.40813e-02	1.51	4.40812e-02
50 / 5000	0.18	4.02113e-02	13.16	4.38594e-02	0.91	4.05313e-02	1.35	4.05312e-02

6.3 Nonnegative principal component analysis

In this subsection, we consider the nonnegative PCA model (5) on the oblique manifold. All parameters of our algorithm are the same as those in subsection 6.1. Since AManPG and ARPG cannot achieve our requirement for accuracy in most testing cases, we omit them in this experiment. The possible reason is that the convergence of AManPG and ARPG relies on the Lipschitz continuity of the nonsmooth part, while it is not the case for the indicator function of $\delta_{X\geq 0}$. Hence, we only compare our algorithm with ProxGD. The comparisons are illustrated in Figures 8 and 9 and Table 3 for the computational results. Those results show that ProxSSN achieves better results and converges much faster to highly accurate solutions compared with ProxGD.

Table 3: Computational results of the nonnegative PCA problem (5).

(n,p)	ProxSSN				ProxGD			
	time	obj	err	iter	time	obj	err	iter
500 / 10	0.35	1.166866	1.23e-7	66 (9.2)	2.46	1.166866	2.93e-5	2840
500 / 15	0.22	1.619850	6.55e-7	33 (9.5)	1.96	1.619850	2.60e-5	1838
500 / 20	0.53	1.942255	8.52e-7	64 (9.8)	4.60	1.942255	2.29e-5	3413
500 / 25	0.72	2.300220	7.04e-7	73 (9.8)	10.42	2.300220	1.70e-5	7317
500 / 30	0.89	2.523960	6.56e-7	83 (9.9)	15.41	2.523999	2.69e-4	10000
500 / 5	0.04	0.592243	8.56e-8	13 (8.8)	0.13	0.592244	4.83e-5	189
600 / 10	0.12	1.223420	5.71e-7	20 (9.4)	1.32	1.223420	2.57e-5	1362
600 / 15	0.34	1.894680	3.86e-7	47 (9.7)	5.43	1.894680	2.44e-5	4455
600 / 20	1.23	2.261036	1.19e-6	130 (9.6)	13.43	2.261036	1.63e-5	9617

(n,p)	ProxSSN				ProxGD			
	time	obj	err	iter	time	obj	err	iter
600 / 25	0.90	2.238704	9.12e-7	91 (9.8)	14.71	2.241462	1.54e-3	10000
600 / 30	0.92	2.510240	1.19e-6	94 (9.8)	16.34	2.510453	1.81e-5	10000
600 / 5	0.04	0.782584	8.94e-8	12 (9.2)	0.52	0.782584	2.43e-5	757
700 / 10	0.48	1.332547	3.06e-7	66 (9.5)	4.74	1.332547	2.90e-5	5031
700 / 15	0.23	1.891921	4.99e-7	32 (9.7)	3.09	1.891922	1.61e-5	2448
700 / 20	0.35	2.232710	7.31e-7	38 (9.7)	4.07	2.232710	1.86e-5	2787
700 / 25	1.00	2.578730	1.39e-6	90 (9.9)	18.99	2.578745	1.61e-4	10000
700 / 30	2.01	2.997021	1.90e-6	124 (9.8)	14.99	3.025005	1.36e-5	6748
700 / 5	0.09	0.751121	1.23e-7	19 (9.1)	1.06	0.751121	4.40e-5	1475
800 / 10	0.62	1.361048	6.58e-7	57 (9.5)	5.33	1.361048	2.53e-5	3805
800 / 15	1.07	1.837726	5.13e-7	99 (9.7)	17.68	1.839436	7.48e-4	10000
800 / 20	1.50	2.262145	1.20e-6	115 (9.5)	18.80	2.262147	5.92e-5	10000
800 / 25	2.30	2.621645	1.51e-6	158 (9.8)	19.63	2.623857	1.42e-4	10000
800 / 30	1.58	2.943294	2.20e-6	122 (9.7)	19.78	2.944495	1.71e-3	10000
800 / 5	0.07	0.754357	1.28e-7	10 (9.0)	0.31	0.754357	4.35e-5	257
900 / 10	0.10	1.374185	3.69e-7	14 (9.3)	1.51	1.374185	2.40e-5	1200
900 / 15	0.86	1.933525	1.06e-6	93 (9.7)	7.89	1.933525	1.07e-5	5513
900 / 20	1.18	2.360027	1.36e-6	107 (9.7)	17.38	2.360027	1.74e-5	10000
900 / 25	1.88	2.773641	1.69e-6	153 (9.8)	19.07	2.777065	3.41e-4	10000
900 / 30	1.60	3.157731	2.02e-6	121 (9.8)	21.05	3.159992	3.55e-3	10000
900 / 5	0.27	0.770418	2.27e-7	62 (7.9)	1.54	0.770418	2.59e-5	1672
1000 / 10	0.64	1.376750	2.84e-7	81 (9.0)	3.50	1.376750	3.47e-5	2719
1000 / 15	0.19	2.049750	1.08e-6	21 (9.5)	2.65	2.049750	2.16e-5	1673
1000 / 20	1.05	2.581317	1.41e-6	85 (9.7)	18.11	2.581318	2.08e-5	10000
1000 / 25	1.30	3.043254	1.18e-6	101 (9.8)	20.56	3.045420	4.82e-4	10000
1000 / 30	1.79	3.516861	2.84e-6	129 (9.8)	23.79	3.517976	1.25e-3	10000
1000 / 5	0.05	0.804861	2.75e-7	10 (9.0)	0.33	0.804861	3.53e-5	390

6.4 Bose-Einstein condensates

In this subsection, we consider the Bose-Einstein condensates (BEC) problem (Aftalion and Du, 2001; Bao and Cai, 2013; Wu et al., 2017). The total energy in the BEC problem is defined as

$$E(\psi) = \int_{\mathbb{R}^n} \left[\frac{1}{2}|\nabla\psi(x)|^2 + V(x)|\psi(x)|^2 + \frac{\beta}{2}|\psi(x)|^4 - \Omega\bar{\psi}(x)L_z(x) \right] \mathrm{d}x, \tag{37}$$

where $x \in \mathbb{R}^d$ is the spatial coordinate vector, $\bar{\psi}$ denotes the complex conjugate of ψ, $L_z = -i(x\partial_y - y\partial_x)$, $V(x)$ is an external trapping potential, $\Omega \in \mathbb{R}$ is an angular velocity, and β is a given constant. Then, the ground state of a BEC is usually defined as the minimizer of the following nonconvex minimization problem

$$\min_{\psi \in S} E(\psi), \tag{38}$$

where S is the spherical constraint and is defined as

$$S := \left\{ \psi \mid E(\psi) < \infty, \int_{\mathbb{R}^d} |\psi(x)|^2 dx = 1 \right\}.$$

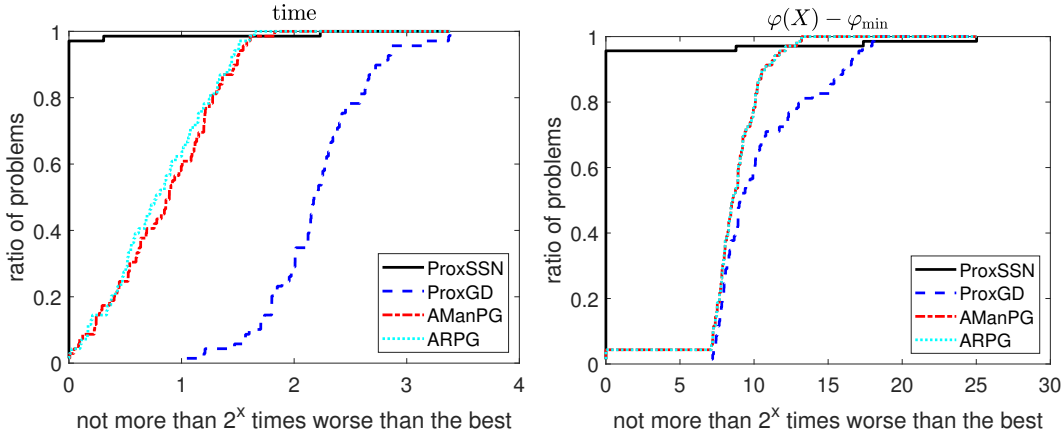

Figure 4: The performance profiles on the sparse PCA problem (3).

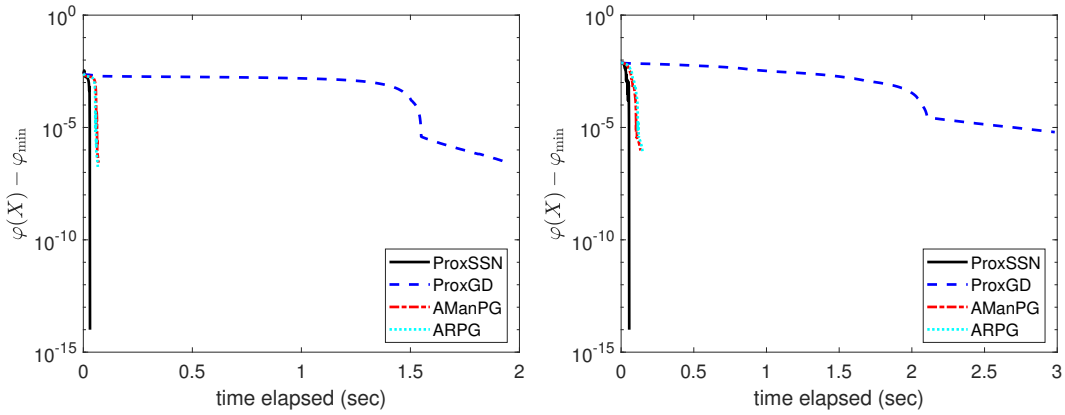

Figure 5: The trajectories of the objective function values with respect to the wall-clock time on the sparse least square regression (8) with $m = 20, \lambda = 0.01$. Left: $n = 2000$; right: $n = 3000$.

By using a suitable discretization, such as finite differences or the sine pseudo-spectral and Fourier pseudo-spectral (FP) method, we can reformulate the BEC problem as follows:

$$\min_{x \in \mathbb{C}^M} \frac{1}{2} x^* A x + \frac{\beta}{2} \sum_{i=1}^{M} |x_i|^4, \quad \text{s.t.} \quad x \in S^M, \tag{39}$$

where $S^M = \{x \in \mathbb{C}^M \mid \|x\|_2 = 1\}$ with a positive integer M and $A \in \mathbb{C}^{M \times M}$ is a Hermitian matrix. We refer to (Wu et al., 2017) for the details.

The ProxGD and ProxSSN are applied to problem (39) by setting

$$f(x) := \frac{1}{2} x^* A x + \frac{\beta}{2} \sum_{i=1}^{M} |x_i|^4, \quad h(x) = \delta_{S^M}(x).$$

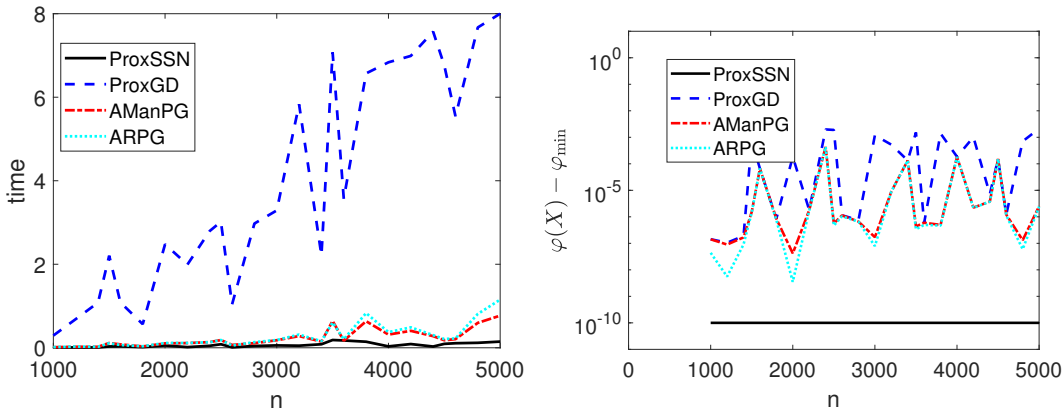

Figure 6: Comparisons of wall-clock time and the objective function values on the sparse least square regression (8) with $m = 30, \lambda = 0.01$ for different n.

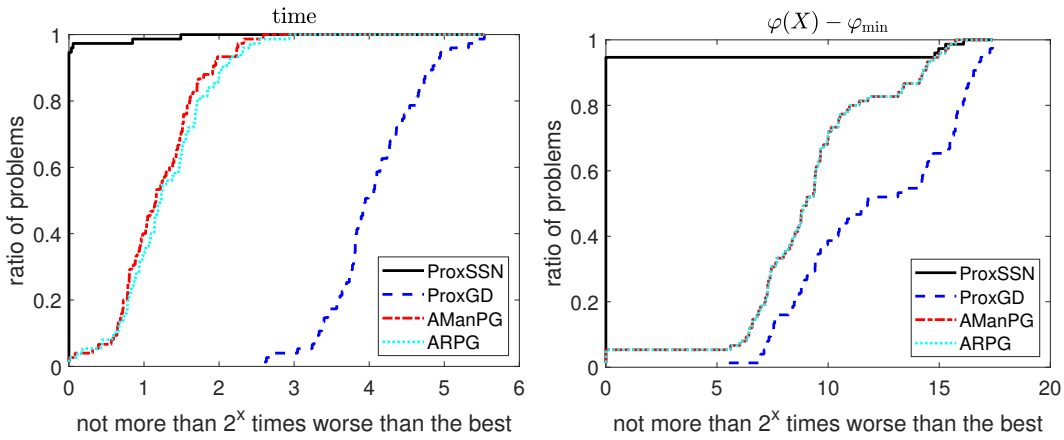

Figure 7: The performance profiles on the sparse least square regression (8).

Since problem (39) can be seen as a smooth problem on the complex sphere, we do comparisons with the adaptive regularized Newton method (ARNT) in (Hu et al., 2018). All parameters of ProxGD and ProxSSN follow the setup discussed in subsection 6.1 except tol $= 10^{-6}$. The parameters of ARNT are the same as in (Hu et al., 2018), we stop ARNT when the Riemannian gradient norm is less than 10^{-6} or the maximum number of iterations 500 is reached. We take $d = 2$ and $V(x, y) = \frac{1}{2}x^2 + \frac{1}{2}y^2$. The BEC problem is discretized by FP on the bounded domain $(-16, 16)^2$ with β ranging from 500 to 1000 and $\Omega = 0, 0.1, 0.25$. Following the settings in (Wu et al., 2017), we use the mesh refinement procedure with the coarse meshes $(2^k + 1) \times (2^k + 1)(k = 2, \cdots, 5)$ to gradually obtain an initial solution point on the finest mesh $(2^6 + 1) \times (2^6 + 1)$. all algorithms are tested with mesh refinement and start from the same initial point on the coarsest mesh with

$$\phi_a(x, y) = \frac{(1 - \Omega)\phi_1(x, y) + \Omega\phi_2(x, y)}{\|(1 - \Omega)\phi_1(x, y) + \Omega\phi_2(x, y)\|}, \quad \phi_b(x, y) = \frac{\phi_1(x, y) + \phi_2(x, y)}{\|\phi_1(x, y) + \phi_2(x, y)\|},$$

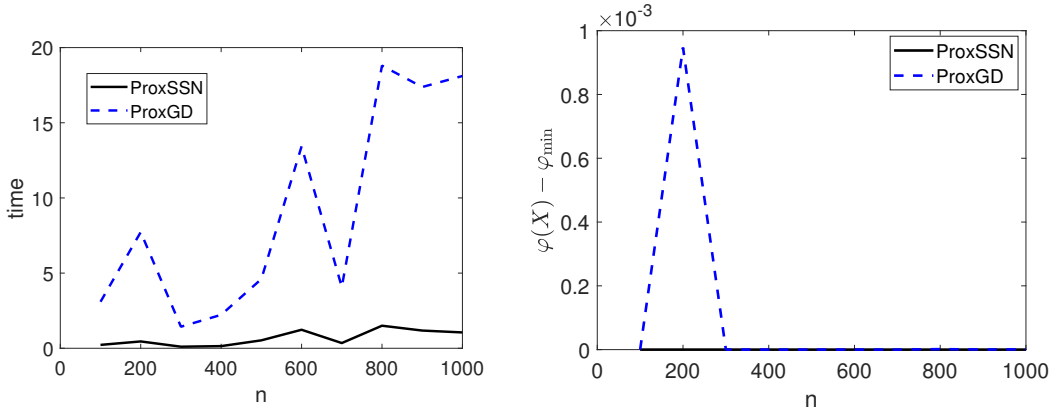

Figure 8: Comparisons of wall-clock time and the objective function values on the nonnegative PCA problem (5) with $p = 20$ for different n.

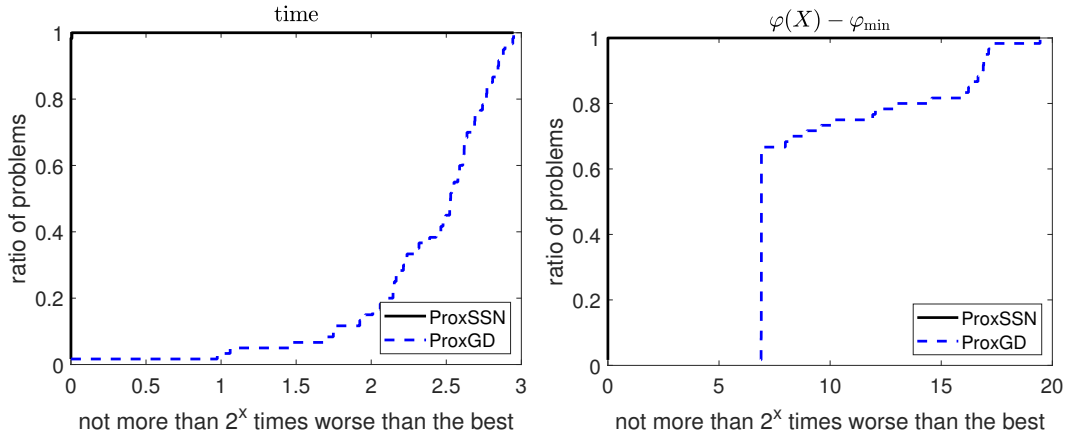

Figure 9: The performance profiles on nonnegative PCA problem (5).

where $\phi_1(x, y) = \frac{1}{\sqrt{\pi}} e^{-(x^2+y^2)/2}$ and $\phi_2(x, y) = \frac{x+iy}{\sqrt{\pi}} e^{-(x^2+y^2)/2}$.

Table 2 gives detailed computational results. For the first column,"Initial" denotes the type of the initial point, "a" and "b" are $\phi_a(x, y)$ and $\phi_b(x, y)$, respectively. For the iteration numbers in our table, "iter" and "siter" denote the outer iterations and the average sub-iterations, respectively. Note that ProxGD reaches the maximum iteration of 1000, which shows that ProxGD does not converge to the required accuracy in all cases. ProxSSN and ARNT find a point with almost the same objective function value, while our algorithm ProxSSN is faster than ARNT in most cases. Figures 10 and 11 demonstrate the superiority of ProxSSN over ARNT and ProxGD.

Table 4: Computational results of BEC

$(\beta, \Omega, Initial)$	ProxSSN			ProxGD			ARNT		
	time	obj	iter (siter)	time	obj	iter	time	obj	iter (siter)
500 / 0.00 / a	0.17	9.38492745	2 (46.0)	10.75	9.38492745	1000	0.33	9.38492745	6 (17.3)

$(\beta, \Omega, Initial)$	ProxSSN			ProxGD			ARNT		
	time	obj	iter (siter)	time	obj	iter	time	obj	iter (siter)
500 / 0.10 / a	0.94	9.38492744	3 (133.3)	15.13	9.38492746	1000	2.45	9.38492744	7 (61.7)
500 / 0.25 / a	1.14	9.38492744	3 (133.3)	12.65	9.38492747	1000	4.74	9.38492744	15 (73.3)
500 / 0.00 / b	0.74	9.38492745	3 (133.3)	14.24	9.38492748	1000	1.21	9.38492745	7 (49.9)
500 / 0.10 / b	0.89	9.38492744	3 (133.3)	14.39	9.38492746	1000	2.23	9.38492744	8 (60.3)
500 / 0.25 / b	0.98	9.38492744	3 (133.3)	12.29	9.38492747	1000	3.50	9.38492744	11 (72.3)
600 / 0.00 / a	0.20	10.60175601	3 (95.0)	11.38	10.60175602	1000	0.36	10.60175601	6 (17.2)
600 / 0.10 / a	1.01	10.60175601	3 (133.3)	13.75	10.60175604	1000	2.99	10.60175601	8 (57.8)
600 / 0.25 / a	1.20	10.60175601	3 (133.3)	12.75	10.60175606	1000	5.09	10.60175601	14 (70.3)
600 / 0.00 / b	0.99	10.60175601	3 (133.3)	16.95	10.60175602	1000	1.75	10.60175601	6 (54.0)
600 / 0.10 / b	1.07	10.60175601	3 (133.3)	14.52	10.60175604	1000	3.64	10.60175601	11 (52.4)
600 / 0.25 / b	1.10	10.60175601	3 (133.3)	12.24	10.60175606	1000	4.61	10.60175601	10 (65.0)
700 / 0.00 / a	0.26	11.75508441	3 (95.0)	11.90	11.75508441	1000	0.42	11.75508441	6 (15.2)
700 / 0.10 / a	1.02	11.75508441	3 (133.3)	12.30	11.75508444	1000	4.05	11.75508441	6 (57.8)
700 / 0.25 / a	1.00	11.75508441	3 (133.3)	11.52	11.75508453	1000	4.81	11.75508441	12 (60.5)
700 / 0.00 / b	0.87	11.75508441	3 (133.3)	16.77	11.75508442	1000	1.78	11.75508441	6 (54.8)
700 / 0.10 / b	0.93	11.75508441	3 (133.3)	11.83	11.75508444	1000	4.95	11.75508441	9 (47.8)
700 / 0.25 / b	1.12	11.75508441	3 (133.3)	11.51	11.75508452	1000	4.93	11.75508441	10 (62.2)
800 / 0.00 / a	0.20	12.85654802	3 (95.3)	11.09	12.85654802	1000	0.47	12.85654802	6 (15.0)
800 / 0.10 / a	1.06	12.85654802	3 (133.3)	14.40	12.85654804	1000	3.51	12.85654802	12 (55.8)
800 / 0.25 / a	1.22	12.85654801	3 (133.3)	12.85	12.85654804	1000	5.24	12.85654801	8 (62.9)
800 / 0.00 / b	0.77	12.85654802	3 (133.3)	16.38	12.85654803	1000	1.71	12.85654802	6 (53.5)
800 / 0.10 / b	1.02	12.85654802	3 (133.3)	14.80	12.85654804	1000	4.52	12.85654802	9 (49.2)
800 / 0.25 / b	1.13	12.85654801	3 (133.3)	12.58	12.85654804	1000	6.63	12.85654801	14 (60.1)
900 / 0.00 / a	0.19	13.91448057	3 (91.3)	10.89	13.91448057	1000	0.57	13.91448057	6 (16.0)
900 / 0.10 / a	1.10	13.91448057	3 (133.3)	14.24	13.91448058	1000	5.89	13.91448057	14 (51.4)
900 / 0.25 / a	1.50	13.91448056	4 (150.0)	14.54	13.91448058	1000	7.22	13.91448056	15 (64.2)
900 / 0.00 / b	0.53	13.91448057	2 (100.0)	17.00	13.91448057	1000	2.21	13.91448057	6 (50.2)
900 / 0.10 / b	1.07	13.91448057	3 (133.3)	15.21	13.91448058	1000	7.55	13.91448057	10 (53.8)
900 / 0.25 / b	1.16	13.91448056	3 (133.3)	12.41	13.91448057	1000	8.21	13.91448056	11 (62.5)
1000 / 0.00 / a	0.23	14.93511997	2 (67.0)	8.41	14.93511997	1000	0.90	14.93511997	6 (22.7)
1000 / 0.10 / a	8.69	14.93511995	4 (150.0)	11.08	14.93511996	1000	10.38	14.93511995	14 (75.7)
1000 / 0.25 / a	4.90	14.93511986	5 (160.0)	11.39	14.93512017	1000	13.75	14.93511986	21 (96.2)
1000 / 0.00 / b	2.60	14.93511997	3 (133.3)	12.94	14.93511997	1000	3.98	14.93511997	10 (76.3)
1000 / 0.10 / b	9.34	14.93511995	4 (150.0)	11.92	14.93511996	1000	12.44	14.93511995	13 (77.9)
1000 / 0.25 / b	2.43	14.93511986	5 (160.0)	11.99	14.93512015	1000	17.62	14.93511986	18 (93.4)

7. Conclusion

This paper introduces a new concept of strong prox-regularity and validates it over many existing interesting applications, including composite optimization problems with weakly convex regularizer, smooth optimization problems on manifolds, and several composite optimization problems on manifolds. Then a projected semismooth Newton method is proposed for solving a class of nonconvex optimization problems equipped with strong prox-regularity. The idea is to utilize the locally single-valued, Lipschitz continuous properties of the residual mapping. The global convergence and local superlinear convergence results of the proposed algorithm are presented under standard conditions. Numerical results have convincingly demonstrated the effectiveness of our proposed method in various nonconvex composite problems, including the sparse PCA problem, the nonnegative PCA problem, the sparse least square regression, and the BEC problem.

Acknowledgments

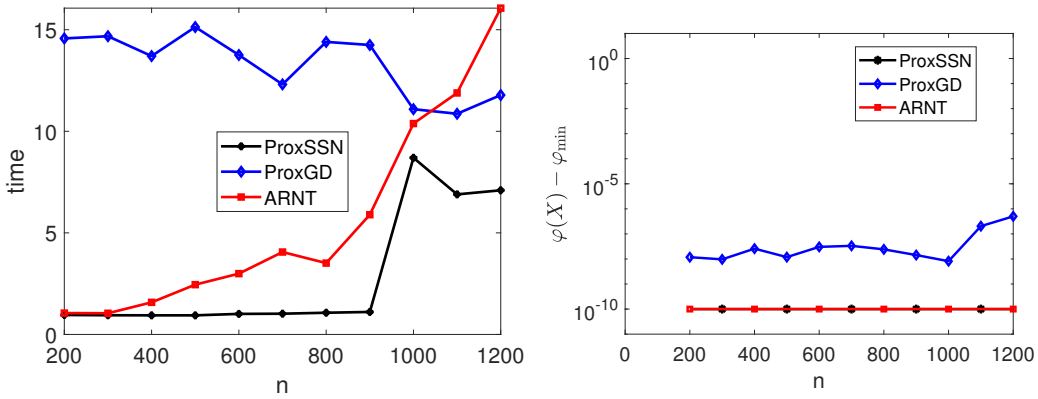

Figure 10: Comparisons of wall-clock time and the objective function values on the BEC problem (39) with $\Omega = 0.2$ and "b".

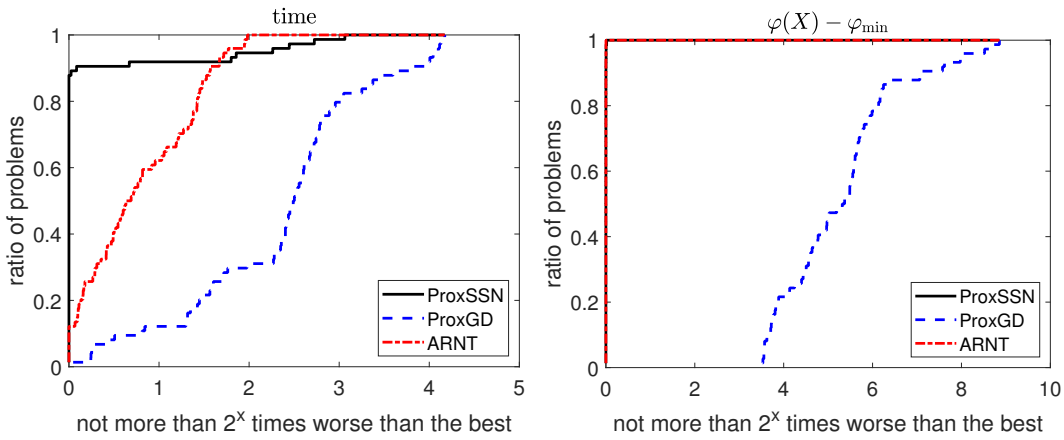

Figure 11: The performance profiles on the BEC problem (39).

The authors are grateful to Prof. Anthony Man-Cho So for his valuable comments and suggestions.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

Amandine Aftalion and Qiang Du. Vortices in a rotating Bose-Einstein condensate: Critical angular velocities and energy diagrams in the thomas-fermi regime. *Physical Review A*, 64(6):063603, 2001.

MV Balashov and AA Tremba. Error bound conditions and convergence of optimization methods on smooth and proximally smooth manifolds. *Optimization*, 71(3):711–735, 2022.

Weizhu Bao and Yongyong Cai. Mathematical theory and numerical methods for Bose-Einstein condensation. *Kinetic & Related Models*, 6(1):1–135, 2013.

Jonathan Barzilai and Jonathan M Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988.

Amir Beck. *First-order Methods in Optimization*. SIAM, 2017.

Axel Böhm and Stephen J Wright. Variable smoothing for weakly convex composite functions. *Journal of Optimization Theory and Applications*, 188(3):628–649, 2021.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Richard H Byrd, Gillian M Chin, Jorge Nocedal, and Figen Oztoprak. A family of second-order methods for convex ℓ_1-regularized optimization. *Mathematical Programming*, 159 (1):435–467, 2016.

Zi Xian Chan and Defeng Sun. Constraint nondegeneracy, strong regularity, and nonsingularity in semidefinite programming. *SIAM Journal on Optimization*, 19(1):370–396, 2008.

Shixiang Chen, Shiqian Ma, Anthony M.-C. So, and Tong Zhang. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. *SIAM Journal on Optimization*, 30(1):210–239, 2020.

Francis H Clarke, RJ Stern, and PR Wolenski. Proximal smoothness and the lower-C2 property. *Journal of Convex Analysis*, 2(1-2):117–144, 1995.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Damek Davis, Dmitriy Drusvyatskiy, Kellie J MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179(3):962–982, 2018.

Kangkang Deng and Zheng Peng. A manifold inexact augmented Lagrangian method for nonsmooth optimization on Riemannian submanifolds in Euclidean space. *IMA Journal of Numerical Analysis*, 2022.

Elizabeth D Dolan and Jorge J Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, 2002.

Dmitriy Drusvyatskiy. The proximal point method revisited. *SIAG/OPT Views and News*, 26:1–7, 2018.

Jianqing Fan. Comments on "wavelets in statistics: A review" by a. antoniadis. *Journal of the Italian Statistical Society*, 6(2):131–138, 1997.

Robert L Foote. Regularity of the distance function. *Proceedings of the American Mathematical Society*, 92(1):153–155, 1984.

Evan S Gawlik and Melvin Leok. Iterative computation of the Fréchet derivative of the polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1354–1379, 2017.

Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Huang, and Jieping Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *International Conference on Machine Learning*, pages 37–45, 2013.

Jiang Hu, Andre Milzarek, Zaiwen Wen, and Yaxiang Yuan. Adaptive quadratically regularized Newton method for Riemannian optimization. *SIAM Journal on Matrix Analysis and Applications*, 39(3):1181–1207, 2018.

Jiang Hu, Xin Liu, Zaiwen Wen, and Yaxiang Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8(2):199–248, 2020.

Wen Huang and Ke Wei. Riemannian proximal gradient methods. *Mathematical Programming*, pages 1–43, 2021.

Bo Jiang, Xiang Meng, Zaiwen Wen, and Xiaojun Chen. An exact penalty approach for optimization with nonnegative orthogonality constraints. *Mathematical Programming*, 2022.

Christian Kanzow and Theresa Lechner. Globalized inexact proximal Newton-type methods for nonconvex composite functions. *Computational Optimization and Applications*, 78(2): 377–410, 2021.

Pham Duy Khanh, Boris Mordukhovich, and Vo Thanh Phat. A generalized Newton method for subgradient systems. *arXiv:2009.10551*, 2020.

Pham Duy Khanh, Boris Mordukhovich, Vo Thanh Phat, and Ba Dat Tran. Globally convergent coderivative-based generalized Newton methods in nonsmooth optimization. *arXiv:2109.02093*, 2021.

Artiom Kovnatsky, Klaus Glashoff, and Michael M Bronstein. MADMM: a generic algorithm for non-smooth optimization on manifolds. In *European Conference on Computer Vision*, pages 680–696. Springer, 2016.

Rongjie Lai and Stanley Osher. A splitting method for orthogonality constrained problems. *Journal of Scientific Computing*, 58(2):431–449, 2014.

Jason D Lee, Yuekai Sun, and Michael A Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

Qiuwei Li, Daniel McKenzie, and Wotao Yin. From the simplex to the sphere: Faster constrained optimization using the Hadamard parametrization. *arXiv:2112.05273*, 2021.

Xudong Li, Defeng Sun, and Kim-Chuan Toh. A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems. *SIAM Journal on Optimization*, 28(1):433–458, 2018a.

Yongfeng Li, Zaiwen Wen, Chao Yang, and Yaxiang Yuan. A semismooth Newton method for semidefinite programs and its applications in electronic structure calculations. *SIAM Journal on Scientific Computing*, 40(6):A4131–A4157, 2018b.

Huikang Liu, Man-Chung Yue, and Anthony Man-Cho So. On the estimation performance and convergence rate of the generalized power method for phase synchronization. *SIAM Journal on Optimization*, 27(4):2426–2446, 2017a.

Huikang Liu, Man-Chung Yue, Anthony Man-Cho So, and Wing-Kin Ma. A discrete first-order method for large-scale MIMO detection with provable guarantees. In *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5. IEEE, 2017b.

Huikang Liu, Man-Chung Yue, and Anthony Man-Cho So. A unified approach to synchronization problems over subgroups of the orthogonal group. *arXiv:2009.07514*, 2020.

Robert Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM Journal on Control and Optimization*, 15(6):959–972, 1977.

Andre Milzarek and Michael Ulbrich. A semismooth Newton method with multidimensional filter globalization for l_1-optimization. *SIAM Journal on Optimization*, 24(1):298–333, 2014.

Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

Jong-Shi Pang and Liqun Qi. Nonsmooth equations: Motivation and algorithms. *SIAM Journal on Optimization*, 3(3):443–465, 1993.

Liqun Qi. Convergence analysis of some algorithms for solving nonsmooth equations. *Mathematics of Operations Research*, 18(1):227–244, 1993.

Liqun Qi and Defeng Sun. A survey of some nonsmooth equations and smoothing Newton methods. In *Progress in optimization*, pages 121–146. Springer, 1999.

Liqun Qi and Jie Sun. A nonsmooth version of Newton's method. *Mathematical programming*, 58(1):353–367, 1993.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Yueyong Shi, Jian Huang, Yuling Jiao, and Qinglong Yang. A semismooth Newton algorithm for high-dimensional nonconvex sparse learning. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8):2993–3006, 2019.

Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM Journal on Optimization*, 28(3):2274–2303, 2018.

Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1):397–434, 2013.

Xinming Wu, Zaiwen Wen, and Weizhu Bao. A regularized Newton method for computing ground states of Bose–Einstein condensates. *Journal of Scientific Computing*, 73(1):303–329, 2017.

Guiyun Xiao and Zheng-Jian Bai. A geometric proximal gradient method for sparse least squares regression with probabilistic simplex constraint. *arXiv:2107.00809*, 2021.

Xiantao Xiao, Yongfeng Li, Zaiwen Wen, and Liwei Zhang. A regularized semi-smooth Newton method with projection steps for composite convex programs. *Journal of Scientific Computing*, 76(1):364–389, 2018.

Zongben Xu, Xiangyu Chang, Fengmin Xu, and Hai Zhang. $\ell_{1/2}$ regularization: A thresholding representation theory and a fast solver. *IEEE Transactions on Neural Networks and Learning Systems*, 23(7):1013–1027, 2012.

Lei Yang. Proximal gradient method with extrapolation and line search for a class of nonconvex and nonsmooth problems. *arXiv:1711.06831*, 2017.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

Xin-Yuan Zhao, Defeng Sun, and Kim-Chuan Toh. A Newton-CG augmented Lagrangian method for semidefinite programming. *SIAM Journal on Optimization*, 20(4):1737–1765, 2010.

Yuhao Zhou, Chenglong Bao, Chao Ding, and Jun Zhu. A semi-smooth Newton based augmented Lagrangian method for nonsmooth optimization on matrix manifolds. *arXiv:2103.02855*, 2021.