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

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

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

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

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

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 non-convexity, 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.

©2024 Jiang Hu, Kangkang Deng, Jiayuan Wu and Quanzheng Li.
License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/ Attribution requirements are provided at http://jmlr.org/papers/v25/23-0371.html.
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:

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

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 $C \supset \text{dom}(h)$, a positive constant $\gamma$, and a norm function $\| \cdot \|$, if the proximal operator $\text{prox}_{th}(\cdot) := \arg \min_u th(u) + \frac{1}{2} \| \cdot - u \|_2^2$ is single-valued and Lipschitz continuous over the closed $\gamma$-neighborhood of $C$, denoted as $\{x + tv : x \in C \subset \mathbb{R}^n, v \in \mathbb{R}^n \text{ with } \|v\| = 1, 0 \leq t \leq \gamma\}$.

We call the above definition strong prox-regularity due to the uniform $\gamma$ for all $x \in C$, 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 $C \subset \mathbb{R}^n$ 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 $h(x) + \frac{\rho}{2} \| x \|_2^2$ is convex. By using the same idea for the convex functions, one can verify that $\text{prox}_{th}$ is single-valued and Lipschitz continuous when $t < \frac{1}{\rho}$. Thus, $h$ is strongly prox-regular with $C = \mathbb{R}^n$, $\gamma = t$, and the $\ell_2$-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

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

We say that $\mathcal{X}$ is $r$-proximally smooth if the nearest-point projection $\text{proj}_{\mathcal{X}}$ 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 $C = \mathcal{X}$, $\gamma = r$, 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 $\ell_1$ 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:

$$
\min_{X \in \text{Ob}(n,p)} \|X^T A^T AX - D^2\|_F^2 + \lambda\|X\|_1,
$$

(3)

where Ob$(n,p) = \{X \in \mathbb{R}^{n \times p} : \text{diag}(X^T X) = 1_p\}$ with $\text{diag}(B)$ 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

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

(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 $\text{prox}_{\lambda h}(X)$, denoted by $(\text{prox}_{\lambda h}(X))_i$, is

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

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

$$
\min_{X \in \text{Ob}^{+}(n,p)} \|X^T A^T AX - D^2\|_F^2,
$$

(5)
where $\text{Ob}^+(n, p) := \text{Ob}(n, p) \cap \{ X \in \mathbb{R}^{n \times p} : X_{ij} \geq 0 \}$ and $D$ is defined as in (3). Note that a more general formulation with smooth objective function over $\text{Ob}^+(n, p)$ has been considered in (Jiang et al., 2022). Problem (5) falls into (1) by letting
\[ h(X) = \delta_{\text{Ob}^+(n, p)}(X) \] is the indicator function of $\text{Ob}^+(n, p)$. Due to the separable structure, the $i$-th column of $\text{prox}_h(X)$, denoted by $(\text{prox}_h(X))_i$, is
\[ (\text{prox}_h(X))_i = \begin{cases} (0, \ldots, 0, 1, 0, \ldots, 0), & \text{if } \max(X_i) \leq 0, \\ X_i^+ / \|X_i^+\|_2, & \text{otherwise}, \end{cases} \]
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:
\[
\min_{y \in \mathbb{R}^n} \frac{1}{2} \|Ay - b\|^2_2, \quad \text{s.t. } y \in \Delta_n,
\] where $\Delta_n = \{ y \in \mathbb{R}^n : y \geq 0, 1^T_n y = 1 \}$, $A \in \mathbb{R}^{m \times n}$, 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
\[ y \in \Delta_n \iff x \in \text{Ob}(n, 1). \]
Adding a sparsity constraint on $x$ leads to the following optimization problem
\[
\min_{x \in \mathbb{R}^n} \frac{1}{2} \|A(x \odot x) - b\|^2_2 + \lambda \|x\|_1, \quad \text{s.t. } x \in \text{Ob}(n, 1).
\] By taking $h(x) = \lambda \|x\|_1 + \delta_{\text{Ob}(n, 1)}$, 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 2014).
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 $\varphi$. 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 \( \ell_1 \) 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 \( \ell_1 \) norm, respectively. The symbol \( B \) will denote the closed unit ball in \( \mathbb{R}^n \), while \( B(x, \epsilon) \) 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 \( \varphi \) is defined as \( \text{dom}(\varphi) = \{ x \in \mathbb{R}^n : \varphi(x) < +\infty \} \). A vector \( v \in \mathbb{R}^n \) is said to be a Fréchet subgradient of \( \varphi \) at \( x \in \text{dom}(\varphi) \) if

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

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

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

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

\[
\hat{\partial} \delta_S(x) = \left\{ v \in \mathbb{R}^n : \limsup_{y \to x, y \in S} \frac{\langle v, y - x \rangle}{\|y - x\|_2} \leq 0 \right\} \quad \text{and} \quad \partial \delta_S(x) = N_S(x)
\]

for any \( x \in S \), where \( N_S(x) \) is the normal cone to \( S \) at \( x \).

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

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

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 \( \bar{v} \in \partial \varphi(\bar{x}) \), \( \varphi \) is said to be prox-regular at \( \bar{x} \). Note that the inequality (10) holds for all \( x' \in \text{dom}(\varphi) \) and \( v \in \partial \varphi(x) \) with a uniform \( \rho \) if \( \varphi \) 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 \( \varphi : \mathbb{R}^n \to \mathbb{R} \) be proper, lower semicontinuous, and prox-bounded with threshold \( \lambda_{\varphi} \). Suppose \( \varphi \) 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 \( \text{prox}_{\gamma \varphi}(x) \) is single-valued and Lipschitz continuous around \( x + \gamma \bar{v} \) and satisfies the condition \( \text{prox}_{\lambda_{\varphi}}(\bar{x} + \gamma \bar{v}) = \bar{x} \).
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 $C$ with a uniform $\gamma$. As shown later, the uniformity of $\gamma$ 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
  \[ 0 \in \partial \varphi(x) = \nabla f(x) + \partial h(x). \]  (11)

- Fixed point of the proximal mapping:
  \[ x \in \text{prox}_{th}(x - t\nabla f(x)), \]  (12)

where $t > 0$.

It follows from the definition of $\text{prox}_{th}$ 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

\[ \partial_C F(x) = \text{conv} \left( \partial_B F(x) \right), \]

where 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 $\text{prox}_{t h}$ 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 $F(x) := \text{prox}_{t h}(x - t \nabla f(x)) - x$ 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 $\rho$-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

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

where $\lambda$ and $\theta$ 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

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

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

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

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

(i) Note that for any vector $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$, $\|x\|_\infty \geq 1/\sqrt{n}$. Following from the definition of the proximal mapping (4), we have for $t \leq \gamma_1$, the proximal mapping $\text{prox}_{\theta h}$ 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 $\text{prox}_{\theta h}$ are in the set $\mathcal{A} := \{x \in \mathbb{R}^n : \exists i, |x_i| = \lambda t\}$. At a non-differentiable point $x \in \mathcal{A}$, let $d \in \mathbb{R}^n$ be a direction. Without loss of generality, assume $x_i = t\lambda$ and $|x_j| \neq t\lambda$ for all $j \neq i$. If $d_i > 0$, we have $\partial_B \text{prox}_{\theta h}(x + d) = \frac{\text{diag}(1, \ldots, 1, 0)}{\|\tilde{w}^-\|_2} - \frac{\tilde{w}^- (\tilde{w}^-)^\top}{\|\tilde{w}^-\|_2^2} =: J(x + d)$, with $\tilde{w}^- = \min(\lambda t - |x + d|, 0) \odot \text{sign}(x)$. Define $\tilde{d}_j = d_j$ if $j \neq i$ and 0 otherwise. Note that $\text{prox}_{\theta h}(x + d) = \text{prox}_{\theta h}(x + \tilde{d})$, $J(x + d) = J(x + \tilde{d})$ and $J(x + d)d = J(x + \tilde{d})\tilde{d}$. Thus,

$$\|\text{prox}_{\theta h}(x + d) - \text{prox}_{\theta h}(x) - J(x + d)d\|_2$$

$$= \|\text{prox}_{\theta h}(x + \tilde{d}) - \text{prox}_{\theta h}(x) - J(x + \tilde{d})\tilde{d}\|_2$$

$$= o(\|\tilde{d}\|_2) = o(\|d\|_2).$$

One can draw a similar conclusion for the case $d_i < 0$. Combining them together, we conclude that $\text{prox}_{\theta h}$ is semismooth.
(ii) It follows from the definition of the proximal mapping (6) that \( \text{prox}_{th}(X) \) is single-valued and Lipschitz continuous over \( D_2 \). Analogous to the case above, one can prove the semismooth property of \( \text{prox}_{th} \).

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

\[
x^{k+1} \in \arg \min_x \left\{ \nabla f(x^k), x - x^k \right\} + \frac{1}{2t_k} \|x - x^k\|^2_2 + h(x) = \text{prox}_{t_k h}(x^k - t_k \nabla f(x^k)),
\]

where \( t_k > 0 \) is the step size depending on the Lipschitz constant of \( \nabla f \). Since \( h \) is nonconvex, \( \text{prox}_{t_k h} \) 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 \( C \supset \text{dom}(h), \gamma, \) and \( \| \cdot \| \), then \( \text{prox}_{th}(x^k - t\nabla f(x^k)) \) is single-valued and Lipschitz continuous (SL) whenever \( \|t\nabla f(x^k)\| \leq \gamma \) and \( x^k \) belongs to the closed set \( C \). To ensure the compactness of the sequence \( \{x^k\} \), one usually investigates the coercive property and the descent property of \( \varphi \). Specifically, any level set \( \{x : \varphi(x) \leq \alpha\} \) with \( \alpha \in \mathbb{R} \) is compact for a coercive \( \varphi \). If the sequence \( \{\varphi(x^k)\} \) is decreasing, \( \{x^k\} \subset \{x : \varphi(x) \leq \varphi(x^0)\} \) is a compact set. Moreover, the norm \( \|\nabla f(x)\| \) is upper bounded by a finite constant \( L > 0 \) over \( \{x : \varphi(x) \leq \varphi(x^0)\} \) due to the smoothness. The proximal operator \( \text{prox}_{th} \) is SL if \( t \leq \frac{\gamma}{2\gamma} \). For this choice of \( t \), we are able to design a second-order method to solve the fixed point equation:

\[
0 = F(x) := x - \text{prox}_{th}(x - t\nabla f(x)),
\]

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

In what follows, we assume that \( \text{prox}_{th} \) 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

\[
(M_k + \mu_k I)d^k = -F(x^k),
\]

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 \( \kappa \). 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

\[
z^k = P_{\text{dom}(h)}(x^k + d^k),
\]

where the projection onto \( \text{dom}(h) \) 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 $\text{prox}_{th}$ 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:

$$\|F(z^k)\|_2 \leq \nu \rho_k, \quad (18)$$
$$\varphi(z^k) \leq \varphi(x^k) + \eta \rho_k^{1-q}\|F(z^k)\|_2^q, \quad (19)$$

where $\rho_k$ 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.,

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

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 $\kappa$ 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 $z^k = P_{\text{dom}(h)}(x^k + d^k)$. 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 $\nabla f$ is Lipschitz continuous with modulus $L > 0$.
- the function $h$ is strongly prox-regular with respect to $C$ and $\gamma$.
- the function $\varphi$ is bounded from below and coercive.

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

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

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

**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

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

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

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 $\varphi$. 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

$$\lim_{k \to \infty} \|F(x^k)\|_2 = 0.$$
Plugging the above inequality into (22) leads to

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

(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

\[ \varphi(x^k) - \varphi(x^{k+1}) \geq c_1 \| F(x^{k+1}) \|_2^2, \]

(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

\[ \varphi(x^k) - \varphi(x^{k+1}) \geq -\eta \rho^{1-q} \| F(x^{k+1}) \|_2^2 \]

\[ = c_1 \| F(x^{k+1}) \|_2^2 - (c_1 \| F(x^{k+1}) \|_2^2 + \eta \rho^{1-q}) \| F(x^{k+1}) \|_2^2 \]

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

\[ \varphi(x^k) - \varphi(x^{k+1}) \geq c_1 \| F(x^{k+1}) \|_2^2 - c_2 \rho^{q}_{k+1}. \]

(24)

Combining (23) and (24), we have

\[ \varphi(x^0) - \varphi(x^{K+1}) = \sum_{i=1}^{K} (\varphi(x^k) - \varphi(x^{k+1})) \geq c_1 \sum_{k=0}^{K} \| F(x^{k+1}) \|_2^2 - c_2 \sum_{k \in K_N} \rho^{q}_{k+1}, \]

where \( K_N \subset \{1, 2, \ldots, K+1\} \) consists of the indices where the projected semismooth Newton updates are accepted. It is easy to see that \( \sum_{k \in K_N} \rho^{q}_{k+1} \leq \rho^{q}_0 \sum_{k=1}^{K+1} \nu q \leq \frac{\rho^{q}_0 (1 - \nu q (K+1))}{1 - \nu q} \leq \frac{\rho^{q}_0}{1 - \nu q}. \) Therefore,

\[ c_1 \sum_{k=0}^{K} \| F(x^{k+1}) \|_2^2 \leq \varphi(x^0) - \varphi(x^{K+1}) + \frac{c_2 \rho^{q}_0}{1 - \nu q}. \]

Since \( \varphi \) 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 $P_{\text{dom}(\kappa)}$ 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 $\|M^{-1}\|_2 \leq C$.

(A4) The function $\varphi$ is Lipschitz continuous over $\text{dom}(h)$ with modulus $L_\varphi$, i.e., for all $x, y \in \text{dom}(h)$,

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

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 $P_{\text{dom}(h)}$ 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 $y \in \text{dom}(h)$, it holds $\|P_{\text{dom}(h)}(x) - y\|_2 \leq 2\|x - y\|_2$.

**Proof** Following the definition of $P_{\text{dom}(h)}$, we have

$$\|P_{\text{dom}(h)}(x) - y\|_2 \leq \|P_{\text{dom}(h)}(x) - x\|_2 + \|x - y\|_2 \leq 2\|x - y\|_2.$$ 

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 $\gamma_F \in \left(0, \min \left\{ \frac{1}{8C}, \frac{\nu}{32C^2L_F}, \frac{\eta}{32C^2(L_\varphi^2C')^{-\frac{1}{2}}} \right\} \right)$, 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 $x \in B(x^*, \varepsilon)$ and $M \in M(x)$,

$$\|F(x) - F(x^*) - (M + \kappa\|F(x)\|_2 I)(x - x^*)\|_2 \leq \gamma_F \|x - x^*\|_2,$$

$$\|(M + \kappa\|F(x)\|^{-1})\|_2 \leq 2C.$$ 

(25)
For the projected semismooth Newton update \(z^k = P_{\text{dom}(h)}(x^k - (M_k + \mu_k I)^{-1} F(x^k))\), it hold that
\[
\|z^k - x^*\|_2 = \|P_{\text{dom}(h)}(x^k - (M_k + \mu_k I)^{-1} F(x^k)) - x^*\|_2 \\
\leq 2\| (M_k + \mu_k I)^{-1} (F(x^k) - F(x^*)) - (M_k + \mu_k I)(x^k - x^*)\|_2 \\
\leq 4\gamma_F C \|x^k - x^*\|_2,
\]
where we assume \(x^k \in \mathbb{B}(x^*, \varepsilon)\). Due to the choice of \(\gamma_F\), we have \(z^k \in \mathbb{B}(x^*, \varepsilon)\). Note that
\[
\|x^k - x^*\|_2 \leq \|z^k - x^*\|_2 + \|z^k - x^k\|_2 \leq 4\gamma_F C \|x^k - x^*\|_2 + 4C\|F(x^k)\|_2.
\]
Then
\[
\|x^k - x^*\|_2 \leq \frac{4C}{1 - 4\gamma_F C}\|F(x^k)\|_2.
\]
Combining (26) and (28) implies
\[
\|z^k - x^*\|_2 \leq \frac{16\gamma_F C^2}{1 - 4\gamma_F C}\|F(x^k)\|_2.
\]
Hence,
\[
\|F(z^k)\|_2 = \|F(z^k) - F(x^*)\|_2 \leq L_F \|z^k - x^*\|_2 \leq \frac{16\gamma_F C^2 L_F}{1 - 4\gamma_F C}\|F(x^k)\|_2 \leq \nu\|F(x^k)\|_2.
\]
In addition, note that
\[
\|z^k - x^*\|_2 = \| (M_k + \mu_k I)^{-1} (F(z^k) - F(x^*) - (M_k + \mu_k I)(z^k - x^*) - F(z^k)) \|_2 \\
\leq 2\gamma_F C \|z^k - x^*\|_2 + 2C\|F(z^k)\|_2.
\]
This gives
\[
\|z^k - x^*\|_2 \leq \frac{2C}{1 - 2\gamma_F C}\|F(z^k)\|_2.
\]
The changes between \(\varphi(z^k)\) and \(\varphi(x^k)\) can be estimated by
\[
\|\varphi(z^k) - \varphi(x^k)\| \leq \|\varphi(z^k) - \varphi(x^*)\|_2 \\
= L_\varphi \|z^k - x^*\|_2^{1-q} \|z^k - x^*\|_2^q \\
\leq L_\varphi \left(\frac{16\gamma_F C^2}{1 - 4\gamma_F C}\right)^{1-q} \left(\frac{2C}{1 - 2\gamma_F C}\right)^q \|F(x^k)\|_2^{1-q} \|F(z^k)\|^q_2 \\
\leq \eta\|F(x^k)\|_2^{1-q} \|F(z^k)\|^q_2.
\]
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
\[
\|x^{k+1} - x^k\|_2 = \|P_{\text{dom}(h)}(x^k + M_k + \mu_k I)^{-1}F(x^k) - x^*\|_2 \\
\leq 4C\|F(x^k) - F(x^*) - (M_k + \mu_k I)(x^k - x^*)\|_2 \\
= o(\|x^k - x^*\|_2),
\]
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. \( \blacksquare \)

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
\[
\beta_k^1 = \frac{\langle s^k, s^k \rangle}{\langle s^k, y^k \rangle}, \quad \text{and} \quad \beta_k^2 = \frac{|\langle s^k, y^k \rangle|}{\langle y^k, y^k \rangle}.
\]

Given \( \varrho, \delta \in (0, 1) \), the nonmonotone Armijo line search is to find the smallest nonnegative integer \( \ell \) satisfying
\[
\varphi(\text{prox}_{t_k(\ell)h}(x^k - t_k(\ell)\nabla f(x^k))) \leq C_k + \frac{\varrho}{2t_k(\ell)}\|\text{prox}_{t_k(\ell)h}(x^k - t_k(\ell)\nabla f(x^k)) - x_k\|_2^2.
\]
Here, \( t_k(\ell) := \beta_k(\ell) \), \( \beta_k(\ell) \) is set to \( \beta_k^1 \) and \( \beta_k^2 \) alternatively, and the reference value \( C_k \) is calculated via \( C_k = (\varrho Q_{k-1} - 1 + \varphi(x^k)) / Q_k \) where \( \varphi \in [0, 1] \), \( C_0 = \varphi(x^0) \), \( Q_k = \varrho Q_{k-1} + 1 \) and \( Q_0 = 1 \). Once \( \ell \) is obtained, we set \( t_k = \beta_k(\ell) \) and the next iterate is then given by \( x^{k+1} = \text{prox}_{t_k h}(x^k - t_k \nabla f(x^k)) \).

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](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 \( f(X) := \|X^TAX - D^2\|_F^2 \). AManPG solves the following subproblem in each iteration:

\[
\eta^{X^k} = \arg \min_{\eta \in T_{X^k} \text{Ob}(n,p)} \left( \langle \text{grad} f(X^k), \eta \rangle + \frac{\tilde{L}}{2} \|\eta\|_F^2 + \lambda \|X^k + \eta\|_1 \right),
\]

where \( \tilde{L} > L \) with \( L \) being the Lipschitz constant of \( f \), \( \text{grad} f(X^k) \) denotes the Riemannian gradient of \( f \) at \( X^k \), and \( T_{X^k} \text{Ob}(n,p) \) is the tangent space to \( \text{Ob}(n,p) \) 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:

\[
\eta^{X^k} = \arg \min_{\eta \in T_{X^k} \text{Ob}(n,p)} \left( \langle \text{grad} f(X^k), \eta \rangle + \frac{\tilde{L}}{2} \|\eta\|_F^2 + \lambda \|\mathcal{R}_{X^k}(\eta)\|_1 \right),
\]

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

\[
\|\tilde{L}\eta^{X^k}\|_F^2 \leq \text{tol},
\]

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

\[
X^{k+1} = \text{prox}_{t_k h}(X^k - t_k \nabla f(X^k)).
\]

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

\[
\text{err} := \frac{\|X^k - \text{prox}_{t_k h}(X^k - t_k \nabla f(X^k))\|_F}{t_k (1 + \|X^k\|_F^2)} \leq \text{tol}.
\]

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 \( \text{tol} = 10^{-10} \times n \times p \) or the number of iterations exceeds 10000, and denote \( F_{\text{ARPG}} \) as the obtained objective value. The other algorithms are terminated when the objective value satisfies \( F(X^k) \leq F_{\text{ARPG}} + 10^{-6} \) or (35) (or (36)) is satisfied with \( \text{tol} = 10^{-10} \times n \times p \), 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 \( \text{randn}(m,n) \), 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: \( n = 400 \).](image)

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., \( r_{i,s} := \min(t_{i,s}) \leq s \leq n_s \). For \( \tau > 0 \), the value

\[
\pi_{s}(\tau) := \frac{\text{number of problems where } \log_2(r_{i,s}) \leq \tau}{\text{total number of problems}}
\]

indicates that solver \( s \) is within a factor \( 2^\tau \geq 1 \) 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 \( \tau \). 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 $n$.

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

<table>
<thead>
<tr>
<th>$(m, n, p)$</th>
<th>ProxSSN time obj</th>
<th>ProxGD time obj</th>
<th>AManPG time obj</th>
<th>ARPG time obj</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 / 500 / 10</td>
<td>1.38 1.28380</td>
<td>13.59 1.28380</td>
<td>1.99 1.28380</td>
<td>1.75 1.28380</td>
</tr>
<tr>
<td>100 / 500 / 15</td>
<td>1.16 1.85986</td>
<td>7.05 1.85986</td>
<td>1.82 1.85986</td>
<td>1.696 1.85986</td>
</tr>
<tr>
<td>100 / 500 / 20</td>
<td>1.71 2.44963</td>
<td>16.59 2.44963</td>
<td>3.40 2.44963</td>
<td>2.96 2.44963</td>
</tr>
<tr>
<td>100 / 500 / 25</td>
<td>2.36 3.00555</td>
<td>15.66 3.00555</td>
<td>4.05 3.00555</td>
<td>3.97 3.00555</td>
</tr>
<tr>
<td>100 / 500 / 30</td>
<td>0.84 3.58139</td>
<td>12.15 3.58139</td>
<td>3.21 3.58139</td>
<td>3.16 3.58139</td>
</tr>
<tr>
<td>100 / 600 / 10</td>
<td>0.40 1.39524</td>
<td>2.47 1.39524</td>
<td>0.53 1.39524</td>
<td>0.62 1.39524</td>
</tr>
<tr>
<td>100 / 600 / 15</td>
<td>0.19 2.04237</td>
<td>1.45 2.04237</td>
<td>0.51 2.04237</td>
<td>0.48 2.04237</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>$(m, n)$</th>
<th>ProxSSN</th>
<th>ProxGD</th>
<th>AMaPG</th>
<th>ARPG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>time</td>
<td>obj</td>
<td>time</td>
<td>obj</td>
</tr>
<tr>
<td>20 / 3000</td>
<td>0.04</td>
<td>3.43796e-02</td>
<td>2.92</td>
<td>3.43839e-02</td>
</tr>
<tr>
<td>20 / 3200</td>
<td>0.18</td>
<td>3.39873e-02</td>
<td>1.07</td>
<td>3.39886e-02</td>
</tr>
<tr>
<td>20 / 3400</td>
<td>0.11</td>
<td>3.23110e-02</td>
<td>5.61</td>
<td>3.23240e-02</td>
</tr>
<tr>
<td>20 / 3600</td>
<td>0.17</td>
<td>3.17896e-02</td>
<td>5.18</td>
<td>3.20365e-02</td>
</tr>
<tr>
<td>20 / 3800</td>
<td>0.05</td>
<td>3.45052e-02</td>
<td>5.94</td>
<td>3.45061e-02</td>
</tr>
<tr>
<td>20 / 4000</td>
<td>0.09</td>
<td>3.44652e-02</td>
<td>6.07</td>
<td>3.54900e-02</td>
</tr>
<tr>
<td>20 / 4200</td>
<td>0.21</td>
<td>3.67684e-02</td>
<td>6.34</td>
<td>3.67852e-02</td>
</tr>
<tr>
<td>20 / 4400</td>
<td>0.11</td>
<td>3.46042e-02</td>
<td>6.49</td>
<td>3.46059e-02</td>
</tr>
<tr>
<td>20 / 4600</td>
<td>0.13</td>
<td>3.39846e-02</td>
<td>6.59</td>
<td>3.71414e-02</td>
</tr>
<tr>
<td>20 / 4800</td>
<td>0.16</td>
<td>3.40047e-02</td>
<td>6.90</td>
<td>3.40144e-02</td>
</tr>
</tbody>
</table>
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).

<table>
<thead>
<tr>
<th>(n, p)</th>
<th>ProxSSN</th>
<th>ProxGD</th>
<th>AManPG</th>
<th>ARPG</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 / 5000</td>
<td>time obj</td>
<td>time obj</td>
<td>time obj</td>
<td>time obj</td>
</tr>
<tr>
<td>0.06</td>
<td>3.32278e-02</td>
<td>6.90</td>
<td>3.54949e-02</td>
<td>0.86</td>
</tr>
<tr>
<td>0.05</td>
<td>3.73733e-02</td>
<td>3.28</td>
<td>3.85623e-02</td>
<td>0.18</td>
</tr>
<tr>
<td>0.05</td>
<td>3.46184e-02</td>
<td>5.82</td>
<td>3.51463e-02</td>
<td>0.28</td>
</tr>
<tr>
<td>0.08</td>
<td>3.57899e-02</td>
<td>2.21</td>
<td>3.59235e-02</td>
<td>0.14</td>
</tr>
<tr>
<td>0.17</td>
<td>3.73116e-02</td>
<td>3.56</td>
<td>3.76236e-02</td>
<td>0.17</td>
</tr>
<tr>
<td>0.14</td>
<td>3.76258e-02</td>
<td>2.21</td>
<td>3.59235e-02</td>
<td>0.15</td>
</tr>
<tr>
<td>0.08</td>
<td>3.96908e-02</td>
<td>6.83</td>
<td>4.08145e-02</td>
<td>0.31</td>
</tr>
<tr>
<td>0.14</td>
<td>3.95462e-02</td>
<td>7.57</td>
<td>3.95534e-02</td>
<td>0.27</td>
</tr>
<tr>
<td>0.10</td>
<td>3.55181e-02</td>
<td>5.54</td>
<td>3.55193e-02</td>
<td>0.20</td>
</tr>
<tr>
<td>0.11</td>
<td>3.85425e-02</td>
<td>7.67</td>
<td>3.92473e-02</td>
<td>0.59</td>
</tr>
<tr>
<td>0.14</td>
<td>3.95414e-02</td>
<td>8.00</td>
<td>4.16688e-02</td>
<td>0.77</td>
</tr>
<tr>
<td>0.05</td>
<td>4.14906e-02</td>
<td>4.46</td>
<td>4.16952e-02</td>
<td>0.23</td>
</tr>
<tr>
<td>0.03</td>
<td>4.08372e-02</td>
<td>2.60</td>
<td>4.08408e-02</td>
<td>0.17</td>
</tr>
<tr>
<td>0.12</td>
<td>4.55556e-02</td>
<td>5.64</td>
<td>4.55502e-02</td>
<td>0.18</td>
</tr>
<tr>
<td>0.05</td>
<td>4.52462e-02</td>
<td>8.17</td>
<td>4.57722e-02</td>
<td>0.35</td>
</tr>
<tr>
<td>0.06</td>
<td>4.12851e-02</td>
<td>3.47</td>
<td>4.12852e-02</td>
<td>0.19</td>
</tr>
<tr>
<td>0.08</td>
<td>4.44167e-02</td>
<td>10.12</td>
<td>4.40984e-02</td>
<td>0.82</td>
</tr>
<tr>
<td>0.23</td>
<td>4.16107e-02</td>
<td>10.48</td>
<td>4.16618e-02</td>
<td>0.60</td>
</tr>
<tr>
<td>0.13</td>
<td>4.37490e-02</td>
<td>13.72</td>
<td>4.37491e-02</td>
<td>0.62</td>
</tr>
<tr>
<td>0.07</td>
<td>4.41428e-02</td>
<td>3.45</td>
<td>4.41463e-02</td>
<td>0.31</td>
</tr>
<tr>
<td>0.44</td>
<td>4.40181e-02</td>
<td>13.08</td>
<td>4.49775e-02</td>
<td>0.96</td>
</tr>
<tr>
<td>0.18</td>
<td>4.02113e-02</td>
<td>13.16</td>
<td>4.38594e-02</td>
<td>0.91</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>$(n, p)$</th>
<th>( \text{ProxSSN} )</th>
<th>( \text{ProxGD} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>obj</td>
<td>err</td>
</tr>
<tr>
<td>600 / 25</td>
<td>0.90</td>
<td>2.237804</td>
</tr>
<tr>
<td>600 / 30</td>
<td>0.92</td>
<td>2.510240</td>
</tr>
<tr>
<td>600 / 5</td>
<td>0.04</td>
<td>0.782584</td>
</tr>
<tr>
<td>700 / 10</td>
<td>0.48</td>
<td>1.332547</td>
</tr>
<tr>
<td>700 / 15</td>
<td>0.23</td>
<td>1.891921</td>
</tr>
<tr>
<td>700 / 20</td>
<td>0.35</td>
<td>2.232710</td>
</tr>
<tr>
<td>700 / 25</td>
<td>1.00</td>
<td>2.578730</td>
</tr>
<tr>
<td>700 / 30</td>
<td>2.01</td>
<td>2.997021</td>
</tr>
<tr>
<td>800 / 10</td>
<td>0.48</td>
<td>1.332547</td>
</tr>
<tr>
<td>800 / 15</td>
<td>0.23</td>
<td>1.891921</td>
</tr>
<tr>
<td>800 / 20</td>
<td>0.35</td>
<td>2.232710</td>
</tr>
<tr>
<td>800 / 25</td>
<td>1.00</td>
<td>2.578730</td>
</tr>
<tr>
<td>800 / 30</td>
<td>2.01</td>
<td>2.997021</td>
</tr>
<tr>
<td>900 / 10</td>
<td>0.48</td>
<td>1.332547</td>
</tr>
<tr>
<td>900 / 15</td>
<td>0.23</td>
<td>1.891921</td>
</tr>
<tr>
<td>900 / 20</td>
<td>0.35</td>
<td>2.232710</td>
</tr>
<tr>
<td>900 / 25</td>
<td>1.00</td>
<td>2.578730</td>
</tr>
<tr>
<td>900 / 30</td>
<td>2.01</td>
<td>2.997021</td>
</tr>
<tr>
<td>900 / 5</td>
<td>0.07</td>
<td>0.754357</td>
</tr>
<tr>
<td>1000 / 10</td>
<td>0.64</td>
<td>1.376750</td>
</tr>
<tr>
<td>1000 / 15</td>
<td>0.19</td>
<td>2.049750</td>
</tr>
<tr>
<td>1000 / 20</td>
<td>1.05</td>
<td>2.581317</td>
</tr>
<tr>
<td>1000 / 25</td>
<td>1.30</td>
<td>3.043254</td>
</tr>
<tr>
<td>1000 / 30</td>
<td>1.79</td>
<td>3.516861</td>
</tr>
<tr>
<td>1000 / 5</td>
<td>0.05</td>
<td>0.804861</td>
</tr>
</tbody>
</table>

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

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

where \( x \in \mathbb{R}^d \) is the spatial coordinate vector, \( \bar{\psi} \) denotes the complex conjugate of \( \psi \), \( L_z = -i(x \partial_y - y \partial_x) \), \( V(x) \) is an external trapping potential, \( \Omega \in \mathbb{R} \) is an angular velocity, and \( \beta \) 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

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

23
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:

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

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

\[
f(x) := \frac{1}{2} x^* Ax + \frac{\beta}{2} \sum_{i=1}^{M} |x_i|^4, \quad h(x) = \delta_{S^M}(x).
\]
A Projected SSN for Nonconvex and Nonsmooth Programs

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 $\text{tol} = 10^{-6}$. 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 $\beta$ 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 $\phi_1(x, y) = \frac{1}{\sqrt{\pi}} e^{-\left(x^2 + y^2\right)/2}$ and $\phi_2(x, y) = \frac{x+iy}{\sqrt{\pi}} e^{-\left(x^2 + y^2\right)/2}$.

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

<table>
<thead>
<tr>
<th>$(\beta, \Omega, \text{Initial})$</th>
<th>ProxSSN</th>
<th>ProxGD</th>
<th>ARNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>time obj iter (siter)</td>
<td>time obj iter</td>
<td>time obj iter (siter)</td>
<td>time obj iter (siter)</td>
</tr>
<tr>
<td>500 / 0.00 / a</td>
<td>0.17 9.38492745 2 (46.0)</td>
<td>10.75 9.38492745 1000</td>
<td>0.33 9.38492745 6 (17.3)</td>
</tr>
</tbody>
</table>
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


