Journal of Computational Mathematics
Vol.xx, No.x, 2024, 1–30.

http://www.global-sci.org/jcm
doi:10.4208/jcm.2404-m2023-0128

AN SQP-TYPE PROXIMAL GRADIENT METHOD FOR COMPOSITE OPTIMIZATION PROBLEMS WITH EQUALITY CONSTRAINTS*

Pinzheng Wei[1)] and Weihong Yang
School of Mathematical Sciences, Fudan University, Shanghai 200433, China
Emails: pzwei21@m.fudan.edu.cn, whyang@fudan.edu.cn

Abstract

In this paper, we present an SQP-type proximal gradient method (SQP-PG) for composite optimization problems with equality constraints. At each iteration, SQP-PG solves a subproblem to get the search direction, and takes an exact penalty function as the merit function to determine if the trial step is accepted. The global convergence of the SQP-PG method is proved and the iteration complexity for obtaining an ϵ-stationary point is analyzed. We also establish the local linear convergence result of the SQP-PG method under the second-order sufficient condition. Numerical results demonstrate that, compared to the state-of-the-art algorithms, SQP-PG is an effective method for equality constrained composite optimization problems.

Mathematics subject classification: 90C30, 65K05.
Key words: Composite optimization, Proximal gradient method, SQP method, Semismooth Newton method.

1. Introduction

In this paper, we consider the following problem:

$$\begin{aligned} &\min_X \psi(X) := f(X) + h(X) \\ &\text{s.t.} \quad c(X) = 0, \end{aligned} \tag{1.1}$$

where $f : \mathbf{E} \to \mathbb{R}$ is a smooth function, $h : \mathbf{E} \to \mathbb{R}$ is a convex nonsmooth function, and c is a continuously differentiable mapping from $\mathbf{E}$ to $\mathbf{F}$. Here $\mathbf{E}$ and $\mathbf{F}$ are Euclidean spaces. The feasible set for (1.1) is denoted by $\Omega := \{X \in \mathbf{E} : c(X) = 0\}$.

If $\mathbf{E} = \mathbb{R}^{n\times p}$ and $c : \mathbf{E} \to S^p$ is defined by $c(X) = X^\top X - I_p$, where S^p denotes the space of $p \times p$ symmetric matrices, then Ω is just the Stiefel manifold $\mathrm{St}(n, p)$, and (1.1) becomes the composite optimization problem over $\mathrm{St}(n, p)$. Such problems have wide applications in many fields such as machine learning, signal processing and numerical linear algebra. For more details about these applications, we refer the readers to [1, 2, 11] and the references therein.

Recently, many numerical algorithms have been proposed for solving composite optimization problems over the Stiefel manifold. These methods can be classified into four categories: subgradient methods, operator splitting methods, augmented Lagrangian (AL) methods and

* Received June 15, 2023 / Revised version received September 26, 2023 / Accepted April 16, 2024 / Published online June 11, 2024 /
1) Corresponding author

proximal-type methods. For detailed discussions of these methods, we refer the reader to [11,46]. Here we only review them briefly. Grohs and Hosseini [21] propose an ϵ-subgradient method and establish the global converge result. Zhang *et al.* [44] extend the smoothing steepest descent method for nonconvex and non-Lipschitz optimization from Euclidean space to Riemannian manifolds. This method can be applied to solve composite optimization problems. Lai *et al.* [26] propose a splitting method for orthogonality constrained problems. There are several excellent works denoted to AL methods for composite optimization. Gao *et al.* [18] propose a parallelized proximal linearized AL algorithm. Zhou *et al.* [46] present a manifold-based AL method to solve composite optimization problems. The AL methods in [13, 30] can also be used to solve (1.1). For proximal gradient methods, Huang and Wei [25] propose a Riemannian proximal gradient (RPG) method and they analyze the iteration complexity of their method under some assumptions. Chen *et al.* [11] present a proximal gradient method, named ManPG, which can be viewed as an inexact RPG method. To accelerate the ManPG method, Wang and Yang [40] propose a proximal quasi-Newton method.

For manifold-based methods, an important operation is the so-called retraction (see [2, Chapter 4]), which maps a tangent vector to a point on the manifold. For the general equality constraint $c(X) = 0$, if we use manifold-based methods to solve (1.1), we can only use the nearest-point projection as the retraction (see [3, Theorem 15]). Usually, computing the nearest-point is expensive for large-scale problems, which results in that the total computational cost of manifold-based methods is very large.

In this paper, we use a different approach and propose an SQP-type method, named SQP-PG, to solve the problem (1.1). Sequential quadratic programming (SQP) methods were first proposed in 1963 by Wilson [41] and were developed in the 1970s by Garcia-Palomares and Mangasarian [19], Han [22, 23], and Powell [34, 35]. For recent developments in SQP methods, the reader is referred to [7, 8, 10, 15, 20, 28, 29]. We also refer to the monograph [32] and the references therein for detailed discussions on SQP methods.

At the k-th iteration, SQP-PG solves the following subproblem:

$$\begin{aligned} &\min \frac{1}{2}\langle V, \mathcal{B}_k[V]\rangle + \langle \nabla f(X_k), V\rangle + h(X_k + V) \\ &\text{s.t.} \quad c(X_k) + Dc(X_k)[V] = 0, \end{aligned} \tag{1.2}$$

where $\mathcal{B}_k$ is an approximated Hessian operator on $\mathbf{E}$, and $Dc(X_k)$ is the derivative of the mapping c at X_k. Similar to the traditional SQP methods, SQP-PG takes an exact penalty function as the merit function to determine if the trial step $X_{k+1} = X_k + \alpha_k V_k$ is accepted or not, where V_k is a non-zero solution of (1.2). An appealing feature of our method is that, compared to the Riemannian manifold optimization method, it does not involve the computation of retraction. Numerical experiments demonstrate that the SQP-PG method is quite efficient especially when the retraction to $\Omega = \{X \in \mathbf{E} : c(X) = 0\}$ is expensive.

The organization of the paper is shown as follows. In Section 2, we introduce some notations and definitions that will be used throughout the paper. In Section 3, we propose the SQP-PG algorithm in detail. The global convergence of SQP-PG is proved and the iteration complexity for obtaining an ϵ-stationary point is analyzed in Section 4. Under the second-order sufficient condition, we also establish the local linear convergence result of SQP-PG in this section. In Section 5, we compare the SQP-PG method with some state-of-the-art methods in the numerical experiments. The paper ends with some conclusions and a short discussion on possible future works.

2. Notions and Preliminaries

In this section, we introduce the notations, definitions, and preliminary results which will be used throughout the paper. We use $\mathbf{E}$ and $\mathbf{F}$ to denote Euclidean spaces. If there is no confusion, the inner products in $\mathbf{E}$ and $\mathbf{F}$ are uniformly denoted by $\langle\cdot\,,\cdot\rangle$. The norms in $\mathbf{E}$ and $\mathbf{F}$ induced by the inner products are denoted by $\|\cdot\|$.

We also assume that $\mathbf{F}$ is a linear subspace of R^t, where $t > 0$ is an integer. The norms $\|\cdot\|$, $\|\cdot\|_1$ and $\|\cdot\|_\infty$ defined on $\mathbf{F}$ are inherited from R^t. Thus, we have

$$\|\Lambda\|_\infty \le \|\Lambda\| \le \|\Lambda\|_1$$

for all $\Lambda \in \mathbf{F}$. For example, the space of $p\times p$ symmetric matrices S^p is a linear subspace of $\mathbb{R}^{p^2}$. For $\Lambda = (\lambda_{ij}) \in S^p$,

$$\|\Lambda\| = \sqrt{\sum_{i,j}\lambda_{ij}^2}, \quad \|\Lambda\|_1 = \sum_{i,j}|\lambda_{ij}|, \quad \|\Lambda\|_\infty = \max_{i,j}|\lambda_{ij}|.$$

The constraint function $c(X)$ of the problem (1.1) is a continuously differentiable mapping from $\mathbf{E}$ to $\mathbf{F}$. For $X \in \mathbf{E}$, the derivative of c at X is denoted by $Dc(X)$, which is a linear operator from $\mathbf{E}$ to $\mathbf{F}$. The adjoint of $Dc(X)$ is denoted by $(Dc(X))^*$, which satisfies

$$\big\langle \Lambda, Dc(X)[V]\big\rangle = \big\langle \big(Dc(X)\big)^*[\Lambda], V\big\rangle, \quad \forall\, V \in \mathbf{E}, \quad \Lambda \in \mathbf{F}.$$

Proposition 2.1. *If $Dc(X)$ is surjective at X, then for any linear positive operator $\mathcal{B}$ defined on $\mathbf{E}$,*

$$Dc(X)\mathcal{B}\big(Dc(X)\big)^* \quad \text{is a surjective linear positive operator on} \quad \mathbf{F}. \tag{2.1}$$

Proof. It is obvious that $\langle z, Dc(X)\mathcal{B}(Dc(X))^*z\rangle \ge 0$ for all $z \in \mathbf{F}$. If

$$\big\langle z, Dc(X)\mathcal{B}\big(Dc(X)\big)^* z\big\rangle = 0,$$

taking account of the positivity of $\mathcal{B}$, we have $(Dc(X))^*z = 0$, which together with the surjectivity of $Dc(X)$ implies $z = 0$. If the range of $Dc(X)\mathcal{B}(Dc(X))^*$ is not $\mathbf{F}$, then there exists a nonzero $z \in \mathbf{F}$ such that $\langle Dc(X)\mathcal{B}(Dc(X))^*w, z\rangle = 0$ for all $w \in \mathbf{E}$, which implies $Dc(X)\mathcal{B}(Dc(X))^*z = 0$. From this, we can derive $z = 0$, yielding a contradiction. □

Definition 2.1. *For a convex function h, the proximal mapping of h is defined by* (see [4, *Definition* 6.1])

$$\mathrm{prox}_h(X) := \arg\min_Y \left\{ h(Y) + \frac{1}{2}\|Y - X\|^2 \right\}.$$

Let $\mathcal{B}$ be a linear positive operator on $\mathbf{E}$, that is $\langle X, \mathcal{B}[X]\rangle > 0$ for all nonzero $X \in \mathbf{E}$. The scaled proximal mapping of h is define by (see [27])

$$\mathrm{prox}_h^{\mathcal{B}}(X) := \arg\min_Y \left\{ h(Y) + \frac{1}{2}\langle Y - X, \mathcal{B}[Y - X]\rangle \right\}. \tag{2.2}$$

Denote $\|X\|_{\mathcal{B}} := \langle X, \mathcal{B}[X]\rangle^{1/2}$. For all $X, Y \in \mathbf{E}$, it holds that (see [27, p. 1424])

$$\big\|\mathrm{prox}_h^{\mathcal{B}}(X) - \mathrm{prox}_h^{\mathcal{B}}(Y)\big\|_{\mathcal{B}} \le \|X - Y\|_{\mathcal{B}}, \tag{2.3}$$

$$\big\langle \mathrm{prox}_h^{\mathcal{B}}(X) - \mathrm{prox}_h^{\mathcal{B}}(Y), \mathcal{B}(X - Y)\big\rangle \ge \big\|\mathrm{prox}_h^{\mathcal{B}}(X) - \mathrm{prox}_h^{\mathcal{B}}(Y)\big\|_{\mathcal{B}}^2. \tag{2.4}$$

Assume that $X^\star$ is a local optimal solution of (1.1) and $Dc(X^\star)$ is surjective. By [14, Corollary 2.4.3], there exists a Lagrange multiplier Λ^* such that the following KKT conditions hold:

$$\begin{aligned} &\nabla f(X^\star) + \xi^\star - \big(Dc(X^\star)\big)^*[\Lambda^\star] = 0, \\ &c(X^\star) = 0, \end{aligned} \tag{2.5}$$

where $\xi^\star \in \partial h(X^\star)$. We say that $X^\star$ is a KKT solution of the problem (1.1).

Definition 2.2 ([36]). *Let $E : \mathbf{F} \to \mathbf{F}$ be a locally Lipschitz continuous mapping at $\Lambda \in \mathbf{F}$. The B-subdifferential of E at Λ is defined by*

$$\partial_B E(\Lambda) := \Big\{ \lim_{k\to\infty} DE(\Lambda_k) \,:\, \Lambda_k \in \mathcal{D}_E, \Lambda_k \to \Lambda \Big\},$$

where $\mathcal{D}_E$ is the set of differentiable points of E in the space $\mathbf{F}$. The set $\partial E(\Lambda) = \mathrm{conv}(\partial_B E(\Lambda))$ is called Clarke's generalized Jacobian of E at Λ, where conv *denotes the convex hull.*

Definition 2.3 ([36]). *Let $E : \mathbf{F} \to \mathbf{F}$ be locally Lipschitz continuous at $\Lambda \in \mathbf{F}$. We say that E is semismooth at $\Lambda \in \mathbf{F}$ if E is directionally differentiable at Λ and for any $J \in \partial E(\Lambda + \Delta\Lambda)$ with $\Delta\Lambda \to 0$,*

$$E(\Lambda + \Delta\Lambda) - E(\Lambda) - J\Delta\Lambda = o(\|\Delta\Lambda\|).$$

We say that E is strongly semismooth at Λ if E is semismooth at Λ and

$$E(\Lambda + \Delta\Lambda) - E(\Lambda) - J\Delta\Lambda = \mathcal{O}(\|\Delta\Lambda\|^2).$$

From [17, Proposition 2.26], if $E : \mathbf{F} \to \mathbf{F}$ is a piecewise $\mathcal{C}^1$ (piecewise smooth) function, then E is semismooth. If E is a piecewise $\mathcal{C}^2$ function, then E is strongly semismooth. It is known that proximal mappings of many interesting functions are piecewise linear or piecewise smooth (see [33, Section 5] and [37, Example 12.31]). Moreover, the proximal mapping of the l_p ($p \geq 1$) norm is strongly semismooth [17, 38].

3. Algorithm Descriptions

In this section, we present an SQP-type proximal gradient method, named SQP-PG, for the composition optimization problem (1.1). First, we give some assumptions that are required throughout this paper.

Assumption 3.1. Assume that

(i) f is smooth, and ∇f is Lipschitz continuous with Lipschitz constant L.

(ii) h is a convex, possibly nonsmooth, and Lipschitz continuous with constant L_h. Moreover, $\mathrm{prox}_h^{\mathcal{B}}$ is a semismooth mapping on $\mathbf{E}$ for any linear positive operator $\mathcal{B}$.

(iii) $c : \mathbf{E} \to \mathbf{F}$ is a differentiable mapping, Dc is a Lipschitz mapping and $Dc(X)$ is a surjective mapping for all $X \in \mathcal{S}$, where $\mathcal{S}$ is the closure of the set $\{X_k : k = 1, 2, \ldots\}$ and X_k is generated by Algorithm 3.1.

In [6, Definition 4.70], if $Dc(X) : \mathbf{E} \to \mathbf{F}$ is a surjective mapping, it is called that the constraint non-degeneracy condition is fulfilled at X.

Assumption 3.2. Let $\mathcal{B}_k$ be the operator from (1.2). There exist two constants $\kappa_2 \geq \kappa_1 > 0$ such that

$$\kappa_1\|V\|^2 \leq \langle V, \mathcal{B}_k[V]\rangle \leq \kappa_2\|V\|^2$$

for all $V \in \mathbf{E}$.

3.1. The SQP-PG algorithm

In this subsection, we give a detailed description of the SQP-PG method. At the current iterate X_k, SQP-PG solves (1.2) to get a search direction. To simplify the notation, we denote $A_k := Dc(X_k)$ and $\|V\|^2_{\mathcal{B}_k} := \langle V, \mathcal{B}_k[V]\rangle$. Then (1.2) can be rewritten as

$$\begin{aligned} &\min_V \langle \nabla f(X_k), V\rangle + \frac{1}{2}\|V\|^2_{\mathcal{B}_k} + h(X_k + V) \\ &\text{s.t.} \quad c(X_k) + A_k[V] = 0. \end{aligned} \tag{3.1}$$

Motivated by the work [11], we use the semismooth Newton method to solve the above problem. We put this part in the next subsection. Following the notation convention of SQP methods, we use V_k to denote the solution and Λ_{k+1} to denote the corresponding Lagrange multiplier.

SQP methods often use a merit function to decide whether a trial step should be accepted. The merit function for problem (1.1) is defined by

$$\Phi(X;\mu) = f(X) + h(X) + \mu\|c(X)\|_1, \tag{3.2}$$

where $\mu > 0$ is a penalty parameter. The following result shows that $\Phi(X;\mu)$ is an exact penalty function for the problem (1.1). We omit the proof since it is the same as that of [24, Theorem 4.4].

Lemma 3.1. *Suppose that $X^\star$ is a strict local solution of* (1.1). *Further suppose Assumptions* 3.1 *and* 3.2 *hold. Then there exists $\mu_0 > 0$ such that $X^\star$ is a local minimizer of $\Phi(X,\mu)$ for all $\mu > \mu_0$.*

The penalty parameter is updated at every iteration and is denoted by μ_k at the k-th iteration. In the theory of the classical SQP method, the condition $\mu_{k+1} > \|\Lambda_{k+1}\|_\infty$ is required (see [32, Eq. (18.32)]) for that it ensures that the computed direction will be a sufficiently good descent direction for the merit function defined with the updated value of μ_{k+1}. Thus, we use the following strategy for updating μ_k:

$$\begin{aligned} &\text{if} \quad \mu_k \geq \|\Lambda_{k+1}\|_\infty + 1, \quad \text{then} \quad \mu_{k+1} = \mu_k, \\ &\text{if} \quad \mu_k < \|\Lambda_{k+1}\|_\infty + 1, \quad \text{then} \quad \mu_{k+1} = \max\{2\mu_k, \|\Lambda_{k+1}\|_\infty + 1\}. \end{aligned} \tag{3.3}$$

Remark 3.1. In the literature, there is an efficient way of updating μ_k based on linear/quadratic models of the merit function, see [5, Eq. (2.7)], [9, Eq. (3.6)] and [32, Eq. (18.36)]. For the classical SQP method, these updating schemes have achieved great success in numerical computation. Since the nonsmooth term h is involved in the objective function of (1.1), the situation becomes more complicated here. It is interesting to consider such updating schemes which utilize linear/quadratic models of the merit function. It is believed that these updating schemes can have better performance in practice for composite optimization. Due to limited space, we do not intend to investigate it here and leave it as a future research topic.

After V_k and μ_{k+1} are obtained, we apply the nonmonotone line search procedure to determine the stepsize α_k. Let $m \geq 0$ be an integer. Denote

$$\Phi_k := \max\Big\{ \max_{\max\{0,k-m\}\leq j\leq k} \Phi(X_j,\mu_j), \Phi(X_k,\mu_{k+1})\Big\}. \tag{3.4}$$

In our method, α_k is set to be γ^{N_k}, where $\gamma \in (0,1)$ and N_k is the smallest nonnegative integer such that

$$\Phi(X_k + \alpha_k V_k, \mu_{k+1}) \leq \Phi_k - \frac{\sigma}{2}\alpha_k \|V_k\|^2_{\mathcal{B}_k}, \tag{3.5}$$

where $\sigma \in (0,1)$.

Now we summarize the SQP-PG method as follows:

If $V_k \neq 0$, it can be served as the search direction of Algorithm 3.1. Now we consider the case of $V_k = 0$. We denote the objective function of (3.1) by ϕ_k, that is

$$\phi_k(V) := \langle \nabla f(X_k), V\rangle + \frac{1}{2}\|V\|^2_{\mathcal{B}_k} + h(X_k + V). \tag{3.6}$$

For $\Lambda \in \mathbf{F}$, the Lagrangian function for the problem (3.1) is

$$\begin{aligned} L_k(V,\Lambda) &= \phi_k(V) - \langle c(X_k) + A_k[V], \Lambda\rangle \\ &= \langle \nabla f(X_k) - A_k^*[\Lambda], V\rangle + \frac{1}{2}\|V\|^2_{\mathcal{B}_k} + h(X_k+V) - \langle c(X_k), \Lambda\rangle. \end{aligned} \tag{3.7}$$

Since V_k is the optimal solution and Λ_{k+1} is the corresponding Lagrange multiplier, there exists $\xi_k \in \partial h(X_k + V_k)$ such that

$$\nabla f(X_k) + \xi_k + \mathcal{B}_k[V_k] - A_k^*[\Lambda_{k+1}] = 0, \tag{3.8}$$

$$c(X_k) + A_k[V_k] = 0. \tag{3.9}$$

We should point out that the pair (ξ_k, Λ_{k+1}), which satisfies (3.8), may not be unique.

Substituting $V_k = 0$ into (3.8) and (3.9), we can derive the following result.

Lemma 3.2. *If $V_k = 0$, then $(X_k, \xi_k, \Lambda_{k+1})$ satisfies (2.5), the KKT condition of the problem (1.1).*

To establish the global convergence of Algorithm 3.1, we need the following boundedness assumption, which is required in the rest of the paper.

Assumption 3.3. The functions $f(X), h(X), c(X)$ and $Dc(X)$ are bounded on the set $\mathcal{S}$, where $\mathcal{S}$ is the closure of the set $\{X_k : k = 1, 2, \ldots\}$ and X_k is generated by Algorithm 3.1.

Algorithm 3.1: The SQP-PG Method.

Input: Initial point $X_0 \in \Omega, \mu_0 > 0, \gamma, \sigma \in (0,1), m > 0$ is an integer.

Choose a sequence $\{\mathcal{B}_k\}$ satisfying Assumption 3.2.
for $k = 0, 1, 2, \ldots$ **do**
  Solve the subproblem (3.1) to get the search direction V_k and the Lagrange multiplier Λ_{k+1}.
  Compute μ_{k+1} by (3.3).
  Set initial stepsize $\alpha_k = 1$.
  while (3.5) is not satisfied **do**
    $\alpha_k = \gamma\alpha_k$.
  end
  Set $X_{k+1} = X_k + \alpha_k V_k$.
end

The boundedness assumption is a standard assumption in the convergence theory of SQP methods.

By Assumption 3.1(iii), we know that $A_k = Dc(X_k)$ is surjective. From (2.1), it follows that $A_k\mathcal{B}_k^{-1}A_k^*$ is invertible. For ease of notation, denote $\Xi_k := A_k\mathcal{B}_k^{-1}A_k^*$. Then, by (3.8) and (3.9), we have

$$\Lambda_{k+1} = \Xi_k^{-1}\big[A_k\mathcal{B}_k^{-1}(\nabla f(X_k) + \xi_k) - c(X_k)\big], \tag{3.10}$$

$$V_k = \mathcal{B}_k^{-1}A_k^*[\Lambda_{k+1}] - \mathcal{B}_k^{-1}\big(\nabla f(X_k) + \xi_k\big). \tag{3.11}$$

From Assumptions 3.1-3.3, we know that $\{\xi_k\}$ is bounded, which together with (3.10) and (3.11) implies that

$$\text{both} \quad \{V_k\} \quad \text{and} \quad \{\Lambda_k\} \quad \text{are bounded.} \tag{3.12}$$

Then there exists a constant $\varrho > 0$ such that $\|V_k\| \le \varrho$ for all $k \ge 0$.

By Assumption 3.1 (iii), Dc is a Lipschitz mapping. Thus, there is a positive constant M_0 such that for all $X \in \mathcal{S}$ and for all $Y \in \mathbf{E}$,

$$\big\|c(X+Y) - c(X) - Dc(X)[Y]\big\|_1 \le M_0\|Y\|^2. \tag{3.13}$$

3.2. The adaptive regularized semismooth Newton method

In this subsection, we give a brief description on the adaptive regularized semismooth Newton (ASSN) method for the subproblem (3.1).

To begin with, we introduce some notations. At the current iterate X_k, for $\Lambda \in \mathbf{F}$, denote

$$B(\Lambda) := X_k - \mathcal{B}_k^{-1}\big(\nabla f(X_k) - \big(Dc(X_k)\big)^*[\Lambda]\big). \tag{3.14}$$

Let $V(\Lambda)$ be the unique minimizer of $L_k(V;\Lambda)$, where L_k is defined by (3.7). By (2.2) and (3.7), we have

$$\begin{aligned} V(\Lambda) &= \mathrm{prox}_h^{\mathcal{B}_k}\big(X_k - \mathcal{B}_k^{-1}\big(\nabla f(X_k) - \big(Dc(X_k)\big)^*[\Lambda]\big)\big) - X_k \\ &= \mathrm{prox}_h^{\mathcal{B}_k}\big(B(\Lambda)\big) - X_k. \end{aligned} \tag{3.15}$$

Substituting (3.15) into the constraint of (3.1) yields

$$\Theta(\Lambda) := c(X_k) + Dc(X_k)[V(\Lambda)] = 0. \tag{3.16}$$

In our method, we use the ASSN method to solve (3.16). To do so, firstly we need to show that the operator Θ is monotone and Lipschitz continuous. For any $\Lambda_1, \Lambda_2 \in \mathbf{F}$, by (3.15) and (3.16), we have

$$\begin{aligned} &\|\Theta(\Lambda_1) - \Theta(\Lambda_2)\| \\ &= \|Dc(X_k)[V(\Lambda_1)] - Dc(X_k)[V(\Lambda_2)]\| \\ &\le \|Dc(X_k)\|\big\|\mathrm{prox}_h^{\mathcal{B}_k}\big(B(\Lambda_1)\big) - \mathrm{prox}_h^{\mathcal{B}_k}\big(B(\Lambda_2)\big)\big\|. \end{aligned} \tag{3.17}$$

From Assumption 3.2, we know that

$$\sqrt{\kappa_1}\|Y\| \le \|Y\|_{\mathcal{B}_k} \le \sqrt{\kappa_2}\|Y\|$$

for all $Y \in \mathbf{E}$. Using this, we can derive that

$$\big\|\mathrm{prox}_h^{\mathcal{B}_k}\big(B(\Lambda_1)\big) - \mathrm{prox}_h^{\mathcal{B}_k}\big(B(\Lambda_2)\big)\big\|$$

$$
\begin{aligned}
&\le \frac{1}{\sqrt{\kappa_1}}\big\|\mathrm{prox}_h^{\mathcal{B}_k}\big(B(\Lambda_1)\big)-\mathrm{prox}_h^{\mathcal{B}_k}\big(B(\Lambda_2)\big)\big\|_{\mathcal{B}_k}\\
&\overset{(2.3)}{\le} \frac{1}{\sqrt{\kappa_1}}\|B(\Lambda_1)-B(\Lambda_2)\|_{\mathcal{B}_k}\\
&\le \frac{\sqrt{\kappa_2}}{\sqrt{\kappa_1}}\|B(\Lambda_1)-B(\Lambda_2)\| \overset{(3.14)}{=} \sqrt{\frac{\kappa_2}{\kappa_1}}\big\|\mathcal{B}_k^{-1}\big(\big(Dc(X_k)\big)^*[\Lambda_1]-\big(Dc(X_k)\big)^*[\Lambda_2]\big)\big\|\\
&\le \sqrt{\frac{\kappa_2}{\kappa_1^3}}\|Dc(X_k)\|\cdot\|\Lambda_1-\Lambda_2\|. \qquad (3.18)
\end{aligned}
$$

Combining (3.17) and (3.18) shows that Θ is Lipschitz continuous.

For any $\Lambda_1,\Lambda_2\in\mathbf{F}$, by (3.14)-(3.16), we have

$$
\begin{aligned}
&\langle\Theta(\Lambda_1)-\Theta(\Lambda_2),\Lambda_1-\Lambda_2\rangle\\
&=\langle Dc(X_k)[V(\Lambda_1)]-Dc(X_k)[V(\Lambda_2)],\Lambda_1-\Lambda_2\rangle\\
&=\big\langle V(\Lambda_1)-V(\Lambda_2),\big(Dc(X_k)\big)^*[\Lambda_1-\Lambda_2]\big\rangle\\
&=\big\langle\mathrm{prox}_h^{\mathcal{B}_k}\big(B(\Lambda_1)\big)-\mathrm{prox}_h^{\mathcal{B}_k}\big(B(\Lambda_2)\big),\mathcal{B}_k\big(B(\Lambda_1)-B(\Lambda_2)\big)\big\rangle\\
&\ge\big\|\mathrm{prox}_h^{\mathcal{B}_k}\big(B(\Lambda_1)\big)-\mathrm{prox}_h^{\mathcal{B}_k}\big(B(\Lambda_2)\big)\big\|_{\mathcal{B}_k}^2\ge 0,
\end{aligned}
$$

where the first inequality is due to (2.4). Then Θ is a monotone operator.

Now we present the iterative scheme of the ASSN method. The iteration number of ASSN is denote by l. At the current iterate Λ_l, ASSN chooses an element $\mathcal{G}_l\in\partial(\Theta(\Lambda_l))$, and computes the Newton direction d_l by solving

$$
(\mathcal{G}_l+\eta_l I)d=-\Theta(\Lambda_l), \qquad (3.19)
$$

where $\eta_l=\lambda_l\|\Theta(\Lambda_l)\|$ and $\lambda_l>0$ is a regularization parameter which is chosen such that $\mathcal{G}_l+\eta_l I$ is invertible. If the size of the problem (3.19) is small, we can solve (3.19) directly; otherwise, we solve it inexactly by the conjugate gradient method. Let d_l be the inexact solution. There is a strategy to decide whether to accept this d_l or not. Compute a trial point $t_l=\Lambda_l+d_l$. If $\|\Theta(t_l)\|$ is sufficient decreased, we take a Newton step. Specifically, let $\xi_0=\|\Theta(\Lambda_0)\|$. Given ξ_l, if $\|\Theta(t_l)\|\le\nu\xi_l$ with $0<\nu<1$, we call the Newton step is successful and set

$$
\Lambda_{l+1}=t_l,\quad \xi_{l+1}=\|\Theta(t_l)\|,\quad \lambda_{l+1}=\lambda_l.
$$

Otherwise, we take a safeguard. We define a ratio $\delta_l:=-\langle\Theta(\Lambda_l),d_l\rangle/\|d_l\|^2$. Select some parameters $0<\theta_1\le\theta_2<1$ and $1<\gamma_1<\gamma_2$. If $\delta_l\ge\theta_1$, the iteration is said to be successful. Otherwise, the iteration is unsuccessful. In summary, we set

$$
\Lambda_{l+1}=\begin{cases}
v_l, & \text{if}\quad \delta_l\ge\theta_1\quad\text{and}\quad \|\Theta(v_l)\|\le\|\Theta(\Lambda_l)\|,\\
w_l, & \text{if}\quad \delta_l\ge\theta_1\quad\text{and}\quad \|\Theta(v_l)\|>\|\Theta(\Lambda_l)\|,\\
\Lambda_l, & \text{if}\quad \delta_l<\theta_1,
\end{cases}
$$

where

$$
v_l=\Lambda_l-\frac{\langle\Theta(t_l),\Lambda_l-t_l\rangle}{\|\Theta(t_l)\|}\Theta(t_l),\quad w_l=\Lambda_l-\alpha\Theta(\Lambda_l),\quad \alpha\in(0,1/\sqrt{\kappa_1}).
$$

The parameters ξ_{l+1} and λ_{l+1} are updated as

$$
\xi_{l+1}=\xi_l,\quad \lambda_{l+1}\in\begin{cases}
(\overline{\lambda},\lambda_l), & \text{if}\quad \delta_l>\theta_2,\\
[\lambda_l,\gamma_1\lambda_l], & \text{if}\quad \theta_1\le\delta_l\le\theta_2,\\
(\gamma_1\lambda_l,\gamma_2\lambda_l], & \text{otherwise},
\end{cases}
$$

where $\overline{\lambda} > 0$ is a small positive constant. For more details, we refer the reader to [43]. A study of the convergence of the ASSM method can be found in [43]. We summarize the ASSN method as follows.

Algorithm 3.2: The ASSN method.

Input: Current iteration point $X_k \in \mathbf{E}, \mathcal{B}_k^{-1}, \Lambda_0 \in \mathbf{F}$, maximum iteration number N, tolerance ϵ_0.

1 $\Lambda_l = \Lambda_0$, $l = 0$.
2 **while** $\|\Theta(\Lambda_l)\| > \epsilon_0$ and $l < N$ **do**
3 Compute $\Theta(\Lambda_l)$ by (3.16).
4 Choose an element $\mathcal{G}_l \in \partial(\Theta(\Lambda_l))$.
5 Select the regularization parameter η_l.
6 Solve the subproblem (3.19) to get the Newton direction d_l.
7 Decide whether to accept this d_l or not and compute Λ_{l+1}.
8 $l \leftarrow l + 1$.
9 **end**
10 Set $\Lambda_{k+1} = \Lambda_l$ and $V_k = V(\Lambda_{k+1})$, which is computed by (3.15).

4. Convergence Analysis

In this section, we study the convergence properties of the SQP-PG method. Under the conditions of Assumptions 3.1-3.3, we prove the global convergence of Algorithm 3.1. We also analyze the local convergence rate of the SQP-PG method. It is proved that the iterates of the algorithm converge locally linearly to the local minimum point at which the second-order sufficient condition holds.

4.1. Global convergence

Recall that V_k is the optimal solution of (3.1) and the corresponding Lagrange multiplier Λ_{k+1} satisfies (3.8). Let ϕ_k be defined by (3.6). We can establish the following important inequality for ϕ_k.

Lemma 4.1. *Suppose Assumption* 3.1 *holds. Let ϕ_k be defined by* (3.6). *For any $\alpha \in [0, 1]$, it holds that*

$$\phi_k(\alpha V_k) - \phi_k(0) \leq \alpha \left(\frac{\alpha - 2}{2} \|V_k\|^2_{\mathcal{B}_k} - \langle \Lambda_{k+1}, c(X_k) \rangle \right).$$

Proof. By (3.8), there exists $\xi_k \in \partial h(X_k + V_k)$ such that

$$\nabla f(X_k) + \xi_k + \mathcal{B}_k[V_k] - \big(Dc(X_k)\big)^*[\Lambda_{k+1}] = 0. \tag{4.1}$$

From $\nabla f(X_k) + \xi_k \in \partial(\phi_k - 0.5\|\cdot\|^2_{\mathcal{B}_k})(V_k)$, it follows that

$$\begin{aligned} \phi_k(0) - \phi_k(V_k) &\geq \langle \nabla f(X_k) + \xi_k, -V_k \rangle - \frac{1}{2}\|V_k\|^2_{\mathcal{B}_k} \\ &= \langle \nabla f(X_k) + \xi_k + \mathcal{B}_k[V_k], -V_k \rangle + \frac{1}{2}\|V_k\|^2_{\mathcal{B}_k} \end{aligned}$$

$$\begin{aligned} &\overset{(4.1)}{=} \left\langle \big(Dc(X_k)\big)^*[\Lambda_{k+1}], -V_k \right\rangle + \frac{1}{2}\|V_k\|^2_{\mathcal{B}_k} \\ &\overset{(3.9)}{=} \langle \Lambda_{k+1}, c(X_k)\rangle + \frac{1}{2}\|V_k\|^2_{\mathcal{B}_k}. \end{aligned} \tag{4.2}$$

Since h is a convex function, for all $0 \le \alpha \le 1$, we have

$$h(X_k + \alpha V_k) - h(X_k) \le \alpha\big(h(X_k + V_k) - h(X_k)\big). \tag{4.3}$$

Then, we can derive that

$$\begin{aligned} \phi_k(\alpha V_k) - \phi_k(0) &= \langle \nabla f(X_k), \alpha V_k\rangle + \frac{1}{2}\|\alpha V_k\|^2_{\mathcal{B}_k} + h(X_k + \alpha V_k) - h(X_k) \\ &\overset{(4.3)}{\le} \alpha\left(\langle \nabla f(X_k), V_k\rangle + \frac{\alpha}{2}\|V_k\|^2_{\mathcal{B}_k} + h(X_k + V_k) - h(X_k)\right) \\ &\overset{(3.6)}{=} \alpha\left(\phi_k(V_k) - \phi_k(0) + \frac{\alpha - 1}{2}\|V_k\|^2_{\mathcal{B}_k}\right) \\ &\overset{(4.2)}{\le} \alpha\left(\frac{\alpha - 2}{2}\|V_k\|^2_{\mathcal{B}_k} - \langle \Lambda_{k+1}, c(X_k)\rangle\right). \end{aligned} \tag{4.4}$$

The proof is complete. □

The following result tells us that $\{\mu_k\}$, the sequence of penalty parameters in Algorithm 3.1, is nondecreasing and μ_k will become a positive constant when k is large enough.

Lemma 4.2. *Suppose Assumptions* 3.1-3.3 *hold. Then there exists a positive constant* $\overline{\mu}$ *and an integer* $K > 0$ *such that* $\mu_k = \overline{\mu}$ *for all* $k \ge K$.

Proof. From (3.3) and (3.12), it follows that $\{\mu_k\}$ is bounded. By (3.3) again, we know that $\{\mu_k\}$ is nondecreasing. Then, there exists $\overline{\mu} > 0$ such that $\overline{\mu} = \lim_{k\to\infty} \mu_k$. This and (3.3) implies $\mu_k = \overline{\mu}$ for sufficiently large k. Thus, the assertion holds. □

In the following result, we show that the stepsize α_k has a lower bound. Denote

$$\overline{\alpha} := \min\left\{1, \frac{(2-\sigma)\kappa_1}{L + 2\overline{\mu}M_0}\right\}, \tag{4.5}$$

where L is the Lipschitz constant of $\nabla f, \kappa_1, \overline{\mu}, \sigma$ and M_0 are parameters in Assumption 3.2, Lemma 4.2, (3.5) and (3.13).

Lemma 4.3. *Suppose that Assumptions* 3.1-3.3 *hold. Then* $\alpha_k \ge \gamma\overline{\alpha}$ *for all* $k \ge 0$, *where* $\overline{\alpha}$ *is from* (4.5) *and* γ *is the parameter in step* 7 *of Algorithm* 3.1. *Therefore, the backtracking line search procedure* (*steps* 6-8 *of Algorithm* 3.1) *will terminate in finite steps.*

Proof. Let $k \ge 0$ be an integer. Since ∇f is Lipschitz continuous with constant L, for any $\alpha > 0$, we have

$$f(X_k + \alpha V_k) \le f(X_k) + \langle \nabla f(X_k), \alpha V_k\rangle + \frac{L}{2}\|\alpha V_k\|^2. \tag{4.6}$$

Thus, it holds that

$$\begin{aligned} \psi(X_k + \alpha V_k) &= f(X_k + \alpha V_k) + h(X_k + \alpha V_k) \\ &\overset{(4.6)}{\le} f(X_k) + \langle \nabla f(X_k), \alpha V_k\rangle + \frac{L}{2}\|\alpha V_k\|^2 + h(X_k + \alpha V_k) \end{aligned}$$

$$\begin{aligned}
&\overset{(3.6)}{=} f(X_k) + \phi_k(\alpha V_k) + \frac{L}{2}\|\alpha V_k\|^2 - \frac{1}{2}\|\alpha V_k\|^2_{\mathcal{B}_k} \\
&\leq \psi(X_k) + \phi_k(\alpha V_k) - \phi_k(0) + \left(\frac{L}{2\kappa_1} - \frac{1}{2}\right)\|\alpha V_k\|^2_{\mathcal{B}_k} \\
&\overset{(4.4)}{\leq} \psi(X_k) + \left(\frac{L}{2\kappa_1} - \frac{1}{\alpha}\right)\|\alpha V_k\|^2_{\mathcal{B}_k} - \alpha\langle \Lambda_{k+1}, c(X_k)\rangle. \qquad (4.7)
\end{aligned}$$

For all $\alpha \in [0,1]$, taking into account $\overline{\mu} \geq \mu_k$ for all $k \geq 0$, we can obtain that

$$\begin{aligned}
&\mu_{k+1}\big(\|c(X_k + \alpha V_k)\|_1 - \|c(X_k)\|_1\big) \\
&\overset{(3.13)}{\leq} \mu_{k+1}\big(\|c(X_k) + \alpha Dc(X_k)[V_k]\|_1 - \|c(X_k)\|_1\big) + \overline{\mu} M_0 \|\alpha V_k\|^2 \\
&\overset{(3.9)}{=} -\alpha\mu_{k+1}\|c(X_k)\|_1 + \overline{\mu} M_0 \|\alpha V_k\|^2. \qquad (4.8)
\end{aligned}$$

By the definition of Φ (see (3.2)), (4.7) and (4.8), we have

$$\begin{aligned}
&\Phi(X_k + \alpha V_k, \mu_{k+1}) - \Phi(X_k, \mu_{k+1}) \\
&= \psi(X_k + \alpha V_k) - \psi(X_k) + \mu_{k+1}\big(\|c(X_k + \alpha V_k)\|_1 - \|c(X_k)\|_1\big) \\
&\leq \left(\frac{L}{2\kappa_1} - \frac{1}{\alpha}\right)\|\alpha V_k\|^2_{\mathcal{B}_k} - \alpha\langle \Lambda_{k+1}, c(X_k)\rangle - \alpha\mu_{k+1}\|c(X_k)\|_1 + \overline{\mu} M_0 \|\alpha V_k\|^2 \\
&\leq \left(\frac{L}{2\kappa_1} + \frac{\overline{\mu} M_0}{\kappa_1} - \frac{1}{\alpha}\right)\|\alpha V_k\|^2_{\mathcal{B}_k} - \alpha\langle \Lambda_{k+1}, c(X_k)\rangle - \alpha\mu_{k+1}\|c(X_k)\|_1 \\
&\leq \left(\frac{L}{2\kappa_1} + \frac{\overline{\mu} M_0}{\kappa_1} - \frac{1}{\alpha}\right)\|\alpha V_k\|^2_{\mathcal{B}_k} - \alpha(\mu_{k+1} - \|\Lambda_{k+1}\|_\infty)\|c(X_k)\|_1 \\
&\leq \left(\frac{L + 2\overline{\mu} M_0}{2\kappa_1}\alpha - 1\right)\alpha\|V_k\|^2_{\mathcal{B}_k}, \quad \forall\, \alpha \in (0,1], \qquad (4.9)
\end{aligned}$$

where the last inequality uses (3.3). Thus, if $\alpha \in (0, \overline{\alpha}]$, it holds that

$$\begin{aligned}
\Phi(X_k + \alpha V_k, \mu_{k+1}) - \Phi_k &\overset{(3.4)}{\leq} \Phi(X_k + \alpha V_k, \mu_{k+1}) - \Phi(X_k, \mu_{k+1}) \\
&\overset{(4.9)}{\leq} \left(\frac{L + 2\overline{\mu} M_0}{2\kappa_1}\overline{\alpha} - 1\right)\alpha\|V_k\|^2_{\mathcal{B}_k} \\
&\overset{(4.5)}{\leq} -\frac{\sigma}{2}\alpha\|V_k\|^2_{\mathcal{B}_k},
\end{aligned}$$

which implies that (3.5) holds for any $\alpha \in (0, \overline{\alpha}]$. Thus, steps 6-8 of Algorithm 3.1 must be repeated a finite number of times at each iteration, and the stepsize must satisfy $\alpha_k \geq \gamma\overline{\alpha}$. □

Next, we prove a useful result prepared for our main results.

Theorem 4.1. *Suppose Assumptions* 3.1-3.3 *hold. Let V_k be the optimal solution of* (3.1). *Then* $\lim_{k\to\infty} V_k = 0$. *Let $\overline{\mu}$ be the constant in Lemma* 4.2. *Then there exists $\Phi^\star \in \mathbb{R}$ such that*

$$\Phi^\star = \lim_{k\to\infty} \Phi(X_k, \overline{\mu}) = \lim_{k\to\infty} \Phi(X_k, \mu_k).$$

Proof. By Lemma 4.2, there exists an integer $K > 0$ such that $\mu_k = \overline{\mu}$ for all $k \geq K$. Thus, by the definition of Φ_k (see (3.4)), we have

$$\Phi_k = \max_{k-m\leq j\leq k} \Phi(X_j, \overline{\mu}), \quad \forall\, k \geq K + m.$$

Denote

$$l(k) := \arg\max_{k-m\le j\le k} \Phi(X_j, \overline{\mu}), \quad \forall\, k \ge K+m. \tag{4.10}$$

Then $\Phi_k = \Phi(X_{l(k)}, \overline{\mu})$. Combining it with (3.5) yields

$$\Phi(X_{k+1}, \overline{\mu}) - \Phi\big(X_{l(k)}, \overline{\mu}\big) \le -\frac{\sigma}{2}\alpha_k \|V_k\|^2_{\mathcal{B}_k}, \quad \forall\, k \ge K+m. \tag{4.11}$$

Thus, we have

$$\begin{aligned}
\Phi\big(X_{l(k+1)}, \overline{\mu}\big) &= \max_{k+1-m\le j\le k+1} \Phi(X_j, \overline{\mu}) \\
&= \max\Big\{\Phi(X_{k+1}, \overline{\mu}), \max_{k+1-m\le j\le k} \Phi(X_j, \overline{\mu})\Big\} \\
&\le \max\Big\{\Phi(X_{l(k)}, \overline{\mu}) - \frac{\sigma}{2}\alpha_k \|V_k\|^2_{\mathcal{B}_k}, \Phi(X_{l(k)}, \overline{\mu})\Big\} \\
&\le \Phi(X_{l(k)}, \overline{\mu}), \quad \forall\, k \ge K+m,
\end{aligned} \tag{4.12}$$

which implies that $\{\Phi(X_{l(k)}, \overline{\mu})\}$ is a nonincreasing sequence for all k sufficiently large. By Assumption 3.3, we know that $\Phi(X_{l(k)}, \overline{\mu})$ is bounded also. Thus, $\lim_{k\to\infty} \Phi(X_{l(k)}, \overline{\mu})$ exists and there exists a scalar $\Phi^\star$ such that

$$\Phi^\star = \lim_{k\to\infty} \Phi\big(X_{l(k)}, \overline{\mu}\big). \tag{4.13}$$

Replacing k by $l(k)-1$ in (4.11), we can deduce that

$$\begin{aligned}
\Phi\big(X_{l(k)}, \overline{\mu}\big) &\le \Phi\big(X_{l(l(k)-1)}, \overline{\mu}\big) - \frac{\sigma}{2}\alpha_{l(k)-1}\|V_{l(k)-1}\|^2_{\mathcal{B}_{l(k)-1}} \\
&\le \Phi\big(X_{l(l(k)-1)}, \overline{\mu}\big) - \frac{\sigma}{2}\kappa_1\alpha_{l(k)-1}\|V_{l(k)-1}\|^2, \quad \forall\, k \ge K+2m+1,
\end{aligned}$$

which together with (4.13) and $\alpha_k \ge \gamma\overline{\alpha}$ (see Lemma 4.3) implies

$$\lim_{k\to\infty} V_{l(k)-1} = 0. \tag{4.14}$$

By (4.13) and (4.14), taking account of the assumption $\{X_k\}$ is bounded, we have

$$\begin{aligned}
\Phi^\star &= \lim_{k\to\infty} \Phi\big(X_{l(k)}, \overline{\mu}\big) = \lim_{k\to\infty} \Phi\big(X_{l(k)-1} + \alpha_{l(k)-1}V_{l(k)-1}, \overline{\mu}\big) \\
&= \lim_{k\to\infty} \Phi\big(X_{l(k)-1}, \overline{\mu}\big).
\end{aligned}$$

For all $j \ge 1$, we can prove by induction that

$$\lim_{k\to\infty} V_{l(k)-j} = 0, \quad \lim_{k\to\infty} \Phi\big(X_{l(k)-j}, \overline{\mu}\big) = \Phi^\star. \tag{4.15}$$

The proof is similar to that of [42, Lemma 4] and therefore we omit it. For any $k > 0$, by the definition of $l(k)$, we know that there exists an integer $1 \le j(k) \le m+1$ such that $k = l(k+m+1) - j(k)$, which together with (4.15) implies

$$\lim_{k\to\infty} V_k = \lim_{k\to\infty} V_{l(k+m+1)-j(k)} = 0,$$

and

$$\lim_{k\to\infty} \Phi(X_k, \overline{\mu}) = \lim_{k\to\infty} \Phi\big(X_{l(k+m+1)-j(k)}, \overline{\mu}\big) = \Phi^\star,$$

as desired. The proof is complete. □

By Assumption 3.3, $\{X_k\}$ is a bounded sequence. The following result shows that every accumulation point of $\{X_k\}$ is a KKT solution of the problem (1.1).

Theorem 4.2. *Suppose Assumptions* 3.1-3.3 *hold. If* $\lim_{k\in\mathcal{K},k\to\infty} X_k = X^\star$ *for some subsequence* $\mathcal{K}$*, then for every accumulation point* $\Lambda^\star$ *of* $\{\Lambda_{k+1}\}_{k\in\mathcal{K}}$*, there exists* $\xi^\star \in \partial h(X^\star)$ *such that* (2.5) *holds.*

Proof. Assume that $\Lambda^\star$ is an accumulation point of $\{\Lambda_{k+1}\}_{k\in\mathcal{K}}$. Then there exists a subsequence $\{k_l\} \subseteq \mathcal{K}$ such that $\lim_{l\to\infty} \Lambda_{k_l+1} = \Lambda^\star$. At X_{k_l}, by (3.8), there exists $\xi_{k_l} \in \partial h(X_{k_l} + V_{k_l})$ such that

$$\xi_{k_l} = \big(Dc(X_{k_l})\big)^*[\Lambda_{k_l+1}] - \nabla f(X_{k_l}) - \mathcal{B}_{k_l}[V_{k_l}]. \tag{4.16}$$

By Assumption 3.1, h is Lipschitz continuous with constant L_h, which implies $\|\xi_{k_l}\| \le L_h$. Thus $\{\xi_{k_l}\}$ has an accumulation point also, which is denoted by $\xi^\star$. Without loss of generality, assume that $\lim_{l\to\infty} \xi_{k_l} = \xi^\star$. By Theorem 4.1, we have $\lim_{l\to\infty} V_{k_l} = 0$. Using this and taking limits on both sides of (4.16), we can see that

$$\nabla f(X^\star) + \xi^\star - \big(Dc(X^\star)\big)^*[\Lambda^\star] = 0.$$

Since $X_{k_l} + V_{k_l} \to X^\star$ and $\xi_{k_l} \to \xi^\star$, by [14, Proposition 2.1.5], it holds that $\xi^\star \in \partial h(X^\star)$. By (3.9) and $\lim_{l\to\infty} V_{k_l} = 0$, we have $c(X_{k_l}) \to 0$, which together with $\lim_{l\to\infty} X_{k_l} = X^\star$ implies $c(X^\star) = 0$. Thus, (2.5) holds for $(X^\star, \Lambda^\star, \xi^\star)$. □

Next, we analyze the iteration complexity of Algorithm 3.1. First, we introduce the definition of an ϵ-stationary point of the problem.

Definition 4.1. *Given* $\epsilon > 0$ *and a point* X_k *generated by Algorithm* 3.1*, we say that* X_k *is an* ϵ*-stationary point of* (1.1) *if the solution* V_k *to* (3.1) *satisfies* $\|V_k\| \le \epsilon$*.*

Proposition 4.1. *Algorithm* 3.1 *will return an* ϵ*-stationary point of* (1.1) *in at most* $(m+1)\lceil\Psi\rceil$ *iterations, where* $\lceil\cdot\rceil$ *denotes rounding up to the next integer and* Ψ *is denoted as*

$$\Psi := \frac{2\big(\Phi(X_0,\mu_0) - \Phi^\star\big) + 4(\overline{\mu} - \mu_0)\max_{X\in\mathcal{S}} \|c(X)\|_1}{\sigma\kappa_1\epsilon^2\gamma\overline{\alpha}},$$

where X_0 *is the initial iteration point,* $\mu_0, \sigma, \gamma, m, \kappa_1, \overline{\mu}, \overline{\alpha}$ *and* $\Phi^\star$ *are parameters in Algorithm* 3.1*, Assumption* 3.2*, Lemmas* 4.2*,* 4.3*, and Theorem* 4.1*, respectively.*

Proof. For $k \ge 0$, let

$$l(k) := \operatorname*{arg\,max}_{\max\{0,k-m\}\le j\le k} \Phi(X_j, \mu_j).$$

From (3.4), we can deduce that

$$\Phi_k \le \Phi\big(X_{l(k)}, \mu_{l(k)}\big) + (\mu_{k+1} - \mu_k)\|c(X_k)\|.$$

Combining it with (3.5) and $\alpha_k \ge \gamma\overline{\alpha}$ (see Lemma 4.3), we have

$$\begin{aligned}
&\Phi(X_{k+1}, \mu_{k+1}) - \Phi\big(X_{l(k)}, \mu_{l(k)}\big) \\
&\le \Phi(X_{k+1}, \mu_{k+1}) - \Phi_k + (\mu_{k+1} - \mu_k)\|c(X_k)\|_1 \\
&\le -\frac{\sigma}{2}\kappa_1\gamma\overline{\alpha}\|V_k\|^2 + (\mu_{k+1} - \mu_k)\max_{X\in\mathcal{S}}\|c(X)\|_1,
\end{aligned} \tag{4.17}$$

where $\mathcal{S}$ is the set in Assumption 3.3. Recall that m is the integer in (3.4). For $j \geq 0$, denote $m_j := j(m+1)$. By (4.17), similar to the proof of [16, Theorem 3.2], we can deduce that for $j \geq 0$,

$$\begin{aligned}&\Phi\big(X_{l(m_{j+1})}, \mu_{l(m_{j+1})}\big) - \Phi\big(X_{l(m_j)}, \mu_{l(m_j)}\big)\\ &\leq \max_{0\leq i\leq m}\Big\{-\frac{\sigma}{2}\kappa_1\gamma\overline{\alpha}\|V_{m_j+i}\|^2\Big\} + (\mu_{m_{j+1}} - \mu_{m_j})\max_{X\in\mathcal{S}}\|c(X)\|_1. \end{aligned} \tag{4.18}$$

Given $\widetilde{K} > 0$, assume that Algorithm 3.1 does not terminate after $m_{\widetilde{K}} = (m+1)\widetilde{K}$ iterations, which means that $\|V_k\| > \epsilon$ for any $0 \leq k \leq m_{\widetilde{K}} - 1$. From the proof of Theorem 4.1, we can see that

$$\begin{aligned}\Phi^\star - (\overline{\mu} - \mu_0)\max_{X\in\mathcal{S}}\|c(X)\|_1 &\leq \Phi\big(X_{l(m_{\widetilde{K}})}, \overline{\mu}\big) - \big(\overline{\mu} - \mu_{l(m_{\widetilde{K}})}\big)\max_{X\in\mathcal{S}}\|c(X)\|_1\\ &\leq \Phi\big(X_{l(m_{\widetilde{K}})}, \mu_{l(m_{\widetilde{K}})}\big).\end{aligned}$$

Combining it with (4.18), we can obtain

$$\begin{aligned}\Phi(X_0,\mu_0) - \Phi^\star &\geq \Phi\big(X_{l(0)}, \mu_{l(0)}\big) - \Phi\big(X_{l(m_{\widetilde{K}})}, \mu_{l(m_{\widetilde{K}})}\big) - (\overline{\mu} - \mu_0)\max_{X\in\mathcal{S}}\|c(X)\|_1\\ &\geq \sum_{j=0}^{\widetilde{K}-1}\min_{0\leq i\leq m}\Big\{\frac{\sigma}{2}\kappa_1\gamma\overline{\alpha}\|V_{m_j+i}\|^2\Big\} - 2(\overline{\mu} - \mu_0)\max_{X\in\mathcal{S}}\|c(X)\|_1\\ &> \frac{\sigma}{2}\kappa_1\epsilon^2\widetilde{K}\gamma\overline{\alpha} - 2(\overline{\mu} - \mu_0)\max_{X\in\mathcal{S}}\|c(X)\|_1,\end{aligned}$$

which implies $\widetilde{K} \leq \lceil\Psi\rceil - 1$ by the definition of Ψ. Therefore, Algorithm 3.1 will find an ϵ-stationary point in at most $(m+1)\lceil\Psi\rceil$ iterations. □

4.2. Local linear convergence

For composite optimization with orthogonality constraints, a local linear convergence rate is obtained for the proximal quasi-Newton method under some second-order sufficient condition in [40, Theorem 4.3]. For composite optimization with general equality constraints, we can also obtain a local linear convergence rate for the SQP-PG method.

Let $X^\star$ be a KKT solution of the problem (1.1). Denote

$$\Gamma^\star = \big\{\Lambda^\star : \text{there exists } \xi^\star \in \partial h(X^\star) \text{ such that } (X^\star, \xi^\star, \Lambda^\star) \text{ satisfies (2.5)}\big\}. \tag{4.19}$$

Note that $\Gamma^\star$ may not be a singleton. It is easy to see that $\Gamma^\star$ is a compact convex set.

For ease of notation, let $\mathcal{F}(X,\Lambda) := f(X) - \langle c(X), \Lambda\rangle$. Assume that f and c are twice continuously differentiable. Now we introduce the second-order sufficient condition: For all $\Lambda^\star \in \Gamma^\star$,

$$\big\langle \nabla^2_{XX}\mathcal{F}(X^\star,\Lambda^\star)[V], V\big\rangle > 0, \quad \forall\, V \neq 0 \quad \text{with} \quad Dc(X^\star)[V] = 0. \tag{4.20}$$

Since $\Gamma^\star$ is compact and $\mathcal{F}$ is twice continuously differentiable, there exists $\eta > 0$ such that for all $\Lambda^\star \in \Gamma^\star$,

$$\big\langle \nabla^2_{XX}\mathcal{F}(X^\star,\Lambda^\star)[V], V\big\rangle > 2\eta\|V\|^2, \quad \forall\, V \neq 0 \quad \text{with} \quad Dc(X^\star)[V] = 0. \tag{4.21}$$

For $\rho \geq 0$, define

$$\mathcal{L}_\rho(X,\Lambda) := f(X) - \langle c(X), \Lambda\rangle + \frac{\rho}{2}\|c(X)\|^2, \quad \forall\, X \in \mathbf{E}, \quad \forall\, \Lambda \in \mathbf{F}.$$

For a fixed $\Lambda^\star \in \Gamma^\star$, since $c(X^\star) = 0$, we have

$$\nabla_X \mathcal{L}_\rho(X^\star, \Lambda^\star) = \nabla f(X^\star) - \big(Dc(X^\star)\big)^*[\Lambda^\star] = \nabla_X \mathcal{F}(X^\star, \Lambda^\star),$$
$$\nabla^2_{XX} \mathcal{L}_\rho(X^\star, \Lambda^\star) = \nabla^2_{XX} \mathcal{F}(X^\star, \Lambda^\star) + \rho\big(Dc(X^\star)\big)^* Dc(X^\star).$$

Under the condition (4.21), using the fact that $\Gamma^\star$ is compact, it is easy to prove that there exists $\overline{\rho} > 0$ such that for all $\Lambda^\star \in \Gamma^\star$,

$$\big\langle \nabla^2_{XX} \mathcal{L}_{\overline{\rho}}(X^\star, \Lambda^\star)[V], V\big\rangle > 2\eta\|V\|^2, \quad \forall\, V \in \mathbf{E}, \quad V \neq 0. \tag{4.22}$$

Lemma 4.4. *Suppose Assumptions* 3.1-3.3 *hold. Further suppose that the second-order sufficient condition* (4.20) *holds at $X^\star$, which is a KKT solution of* (1.1). *Let $\overline{\rho}$ and η be the real numbers such that* (4.22) *holds. Then there exists $\tilde{\epsilon} > 0$ such that*

(i) $\mathcal{L}_{\overline{\rho}}(X, \Lambda^\star)$ is convex on the set $\{X : \|X - X^\star\| < \tilde{\epsilon}\}$ for all $\Lambda^\star \in \Gamma^\star$,

(ii) for all $\mu \geq \sup_{\Lambda^\star \in \Gamma^\star} \|\Lambda^\star\|_\infty + 1$,

$$\Phi(X, \mu) - \Phi(X^\star, \mu) \geq \eta\|X - X^\star\|^2, \quad \forall\, X \quad \textit{satisfying} \quad \|X - X^\star\| < \tilde{\epsilon}. \tag{4.23}$$

Proof. (i) Since $\nabla^2_{XX} \mathcal{L}_{\overline{\rho}}(X, \Lambda)$ is continuous and $\Gamma^\star$ is compact, by (4.22), there exists $\tilde{\epsilon} > 0$ such that for all $\Lambda^\star \in \Gamma^\star$, if $\|X - X^\star\| < \tilde{\epsilon}$, then

$$\big\langle \nabla^2_{XX} \mathcal{L}_{\overline{\rho}}(X, \Lambda^\star)[V], V\big\rangle \geq 2\eta\|V\|^2, \quad \forall\, V \in \mathbf{E}, \quad V \neq 0, \tag{4.24}$$

which implies $\mathcal{L}_{\overline{\rho}}(X, \Lambda^\star)$ is convex on $\{X : \|X - X^\star\| < \tilde{\epsilon}\}$.

(ii) Select a $\Lambda^\star \in \Gamma^\star$, where $\Gamma^\star$ is defined in (4.19). By Taylor's theorem, we have

$$\begin{aligned}
\mathcal{L}_{\overline{\rho}}(X, \Lambda^\star) &= \mathcal{L}_{\overline{\rho}}(X^\star, \Lambda^\star) + \langle \nabla_X \mathcal{L}_{\overline{\rho}}(X^\star, \Lambda^\star), X - X^\star\rangle \\
&\quad + \frac{1}{2}\big\langle \nabla^2_{XX} \mathcal{L}_{\overline{\rho}}(X^\star + t(X - X^\star), \Lambda^\star)[X - X^\star], X - X^\star\big\rangle \\
&= \mathcal{L}_{\overline{\rho}}(X^\star, \Lambda^\star) + \big\langle \nabla_X \mathcal{F}(X^\star, \Lambda^\star), X - X^\star\big\rangle \\
&\quad + \frac{1}{2}\big\langle \nabla^2_{XX} \mathcal{L}_{\overline{\rho}}(X^\star + t(X - X^\star), \Lambda^\star)[X - X^\star], X - X^\star\big\rangle,
\end{aligned} \tag{4.25}$$

where $t \in (0, 1)$. Since $\Lambda^\star \in \Gamma^\star$, there exists $\xi^\star \in \partial h(X^\star)$ such that (2.5) holds, which implies

$$\xi^\star = \big(Dc(X^\star)\big)^*[\Lambda^\star] - \nabla f(X^\star) = -\nabla_X \mathcal{F}(X^\star, \Lambda^\star).$$

Since h is convex and $\xi^\star \in \partial h(X^\star)$, we have

$$h(X) \geq h(X^\star) + \langle \xi^\star, X - X^\star\rangle, \quad \forall\, X \in \mathbf{E}. \tag{4.26}$$

By (4.25) and (4.26), taking into account $\nabla_X \mathcal{F}(X^\star, \Lambda^\star) + \xi^\star = 0$, we can obtain

$$\begin{aligned}
\mathcal{L}_{\overline{\rho}}(X, \Lambda^\star) + h(X) &\geq \mathcal{L}_{\overline{\rho}}(X^\star, \Lambda^\star) + h(X^\star) \\
&\quad + \frac{1}{2}\big\langle \nabla^2_{XX} \mathcal{L}_{\overline{\rho}}\big(X^\star + t(X - X^\star), \Lambda^\star\big)[X - X^\star], X - X^\star\big\rangle.
\end{aligned}$$

By the above inequality and (4.24), we know that for all $\Lambda^\star \in \Gamma^\star$, if $\|X - X^\star\| < \tilde{\epsilon}$, then

$$\mathcal{L}_{\overline{\rho}}(X, \Lambda^\star) + h(X) \geq \mathcal{L}_{\overline{\rho}}(X^\star, \Lambda^\star) + h(X^\star) + \eta\|X - X^\star\|^2. \tag{4.27}$$

Since c is continuous at $X^\star$, there exists $\bar{\epsilon} > 0$ such that if $\|X - X^\star\| < \bar{\epsilon}$, then $\|c(X)\| \le 2/\overline{\rho}$. Without loss of generality, assume that $\tilde{\epsilon} < \bar{\epsilon}$. Thus, if $\|X - X^\star\| < \tilde{\epsilon}$ and $\mu \ge \sup_{\Lambda^\star \in \Gamma} \|\Lambda^\star\|_\infty + 1$, taking account of $\|c(X)\|_1 \ge \|c(X)\|$, we can derive that

$$\begin{aligned}\Phi(X,\mu) - \big(\mathcal{L}_{\overline{\rho}}(X,\Lambda^\star) + h(X)\big) &= \mu\|c(X)\|_1 + \langle c(X), \Lambda^\star\rangle - \frac{\overline{\rho}}{2}\|c(X)\|^2 \\ &\ge \|c(X)\|_1 - \frac{\overline{\rho}}{2}\|c(X)\|^2 \ge 0,\end{aligned}$$

which together with (4.27) implies

$$\begin{aligned}\Phi(X,\mu) - \Phi(X^\star,\mu) &\ge \mathcal{L}_{\overline{\rho}}(X,\Lambda^\star) + h(X) - \big(\mathcal{L}_{\overline{\rho}}(X^\star,\Lambda^\star) + h(X^\star)\big) \\ &\ge \eta\|X - X^\star\|^2.\end{aligned}$$

Thus, the assertion holds. □

Remark 4.1. If the condition $\overline{\mu} \ge \sup_{\Lambda^\star\in\Gamma} \|\Lambda^\star\|_\infty + 1$ in (ii) is replaced by $\overline{\mu} > \sup_{\Lambda^\star\in\Gamma} \|\Lambda^\star\|_\infty$, the assertion holds also. In fact, we only need to make a slight modification to the proof of Lemma 4.4(ii). We omit the details.

For $X \in \mathbf{E}$ and a positive operator $\mathcal{B} : \mathbf{E} \to \mathbf{E}$, we use $V_{X,\mathcal{B}}$ to denote the unique solution of the problem

$$\begin{aligned}&\min \frac{1}{2}\langle V, \mathcal{B}[V]\rangle + \langle \nabla f(X), V\rangle + h(X+V) \\ &\text{s.t.} \quad c(X) + Dc(X)[V] = 0,\end{aligned}$$

and use $\Gamma_{X,\mathcal{B}}$ to denote the set of multipliers corresponding to the constraints, that is

$$\begin{aligned}\Gamma_{X,\mathcal{B}} = \big\{\Lambda \,:\, &\text{there exists } \xi \in \partial h(X+V) \text{ such that} \\ &\nabla f(X) + \xi + \mathcal{B}[V_{X,\mathcal{B}}] - \big(Dc(X)\big)^*[\Lambda] = 0,\ c(X) + Dc(X)[V_{X,\mathcal{B}}] = 0\big\}.\end{aligned}$$

It is easy to see that $\Gamma_{X,\mathcal{B}}$ is a compact convex set. In the following result, we use the notation

$$\mathbb{B} := \big\{\mathcal{B} : \kappa_1\|V\|^2 \le \langle V, \mathcal{B}[V]\rangle \le \kappa_2\|V\|^2\big\},$$

where κ_1 and κ_2 are constants in Assumption 3.2.

Lemma 4.5. *Suppose Assumptions* 3.1-3.3 *hold. Further suppose* (4.20) *holds at* $X^\star$. *For any* $\epsilon > 0$, *there exists* $\delta > 0$ *such that if* $\|X - X^\star\| < \delta$, *then* $\mathrm{dist}(\Lambda, \Gamma^\star) < \epsilon$ *for all* $\Lambda \in \Gamma_{X,\mathcal{B}}$ *and for all* $\mathcal{B} \in \mathbb{B}$, *where*

$$\mathrm{dist}(\Lambda, \Gamma^\star) := \inf\{\|\Lambda - \Lambda'\| : \Lambda' \in \Gamma^\star\}.$$

Proof. Let $\{X_k\}$ be a sequence which converges to $X^\star$, and let $\{\mathcal{B}_k\} \subset \mathbb{B}$ be an operator sequence. By the definition of $V_{X_k,\mathcal{B}_k}$, there exist $\xi_k \in \partial h(X_k + V_{X_k,\mathcal{B}_k})$ and Λ_{k+1} such that

$$\begin{aligned}&\nabla f(X_k) + \xi_k + \mathcal{B}[V_{X_k,\mathcal{B}_k}] - \big(Dc(X_k)\big)^*[\Lambda_{k+1}] = 0, \\ &c(X_k) + Dc(X_k)[V_{X_k,\mathcal{B}_k}] = 0.\end{aligned} \tag{4.28}$$

First, we prove $V_{X_k,\mathcal{B}_k}$ converges to 0. Similar to the proof (3.12), we know that $\{V_{X_k,\mathcal{B}_k}\}$ and $\{\Lambda_k\}$ are bounded. Let $V^\star$ be an accumulation point of $\{V_{X_k,\mathcal{B}_k}\}$. Then there exists

a subsequence $\mathcal{K}$ such that $\lim_{k\in\mathcal{K},k\to\infty} V_{X_k,\mathcal{B}_k} = V^\star$. Without loss of generality, assume that $\lim_{k\in\mathcal{K}} \mathcal{B}_k = \mathcal{B}^\star, \lim_{k\in\mathcal{K}} \xi_k = \xi^\star$ and $\lim_{k\in\mathcal{K}} \Lambda_{k+1} = \Lambda^\star$. From this and (4.28), we can obtain that $\xi^\star \in \partial h(X^\star + V^\star)$ and

$$\nabla f(X^\star) + \xi^\star + \mathcal{B}^\star[V^\star] - \big(Dc(X^\star)\big)^*[\Lambda^\star] = 0. \tag{4.29}$$

Using the above equality, similar to the proof of (4.9), when $\mu \geq \sup_{\Lambda^\star\in\Gamma} \|\Lambda^\star\|_\infty + 1$, we can deduce that

$$\Phi(X^\star + \alpha V^\star, \mu) - \Phi(X^\star, \mu) \leq \left(\frac{L + 2\overline{\mu}M_0}{2\kappa_1}\alpha - 1\right)\alpha\|V^\star\|^2_{\mathcal{B}^\star}, \quad \forall\, \alpha \in (0, 1],$$

which together with (4.23) implies $V^\star = 0$. That is, any accumulation point of $\{V_{X_k,\mathcal{B}_k}\}$ must be 0, which implies $V_{X_k,\mathcal{B}_k}$ converges to 0.

Next, we prove the assertion by contradiction. If it is not true, there exist $\epsilon_1 > 0$, sequences $\{X_k\}, \{\mathcal{B}_k\} \subset \mathbb{B}$, and $\{\Lambda_k\}$ with $\Lambda_{k+1} \in \Gamma_{X_k,\mathcal{B}_k}$ such that $\|X_k - X^\star\| \to 0$ and

$$\operatorname{dist}(\Lambda_k, \Gamma^\star) \geq \epsilon_1, \quad \forall\, k.$$

Since X_k converges to $X^\star$, by the discussion above, we know that $\lim_{k\to\infty} V_{X_k,\mathcal{B}_k} = 0$. Note that $\{\mathcal{B}_k\}, \{\xi_k\}$ and $\{\Lambda_k\}$ are bounded. Then there exists a subsequence $\mathcal{K}$ such that $\lim_{k\in\mathcal{K}} \mathcal{B}_k = \mathcal{B}^\star$, $\lim_{k\in\mathcal{K}} \xi_k = \xi^\star$ and $\lim_{k\in\mathcal{K}} \Lambda_{k+1} = \Lambda^\star$, which together with (4.28) implies $\Lambda^\star \in \Gamma^\star$, yielding a contradiction. □

Let $X^\star$ be an accumulation point of $\{X_k\}$. The following result tells us that if the second-order sufficient condition holds and $\overline{\mu} \geq \sup_{\Lambda^\star\in\Gamma^\star} \|\Lambda^\star\|_\infty + 1$, then X^* is the limit point of $\{X_k\}$.

Lemma 4.6. *Suppose Assumptions* 3.1-3.3 *hold, and* $\overline{\mu} \geq \sup_{\Lambda^\star\in\Gamma^\star} \|\Lambda^\star\|_\infty + 1$, *where* $\overline{\mu}$ *is as in Lemma* 4.2. *Let* $\{X_k\}$ *be the sequence generated by Algorithm* 3.1. *Further suppose that* $X^\star$ *is an accumulation point of* $\{X_k\}$ *and* (4.20) *holds at* $X^\star$. *Then* X_k *converges to* $X^\star$.

Proof. Let $\hat{X}$ be any accumulation point of X_k. By Theorem 4.1, we have

$$\Phi(\hat{X}, \overline{\mu}) = \lim_{k\to\infty} \Phi(X_k, \mu_k) = \Phi^\star. \tag{4.30}$$

From Lemma 4.4, we know that $X^\star$ is the unique minimizer of $\Phi(X, \overline{\mu})$ in a neighborhood of $X^\star$, which together with (4.30) implies that $X^\star$ is an isolated accumulation point of $\{X_k\}$. By Theorem 4.1, we have $V_k \to 0$, and therefore $X_{k+1} - X_k \to 0$. Using these facts, by [31, Lemma 4.10], we can obtain that X_k converges to $X^\star$. □

Using previous results, we can prove the main result of this section, which establishes the local linear convergence of Algorithm 3.1.

Theorem 4.3. *Under the assumptions of Lemma* 4.6, *there exists an integer* $K_0 > 0, \beta > 0$ *and* $\tau \in (0, 1)$ *such that*

$$\Phi(X_k, \overline{\mu}) - \Phi(X^\star, \overline{\mu}) \leq \beta\tau^{k-K_0}\big(\Phi_{K_0} - \Phi(X^\star, \overline{\mu})\big), \quad \forall\, k \geq K_0.$$

Proof. We separate our proof into three parts.

Part (I). First, we derive some inequalities which will be used in part (II). From (4.6), it follows that

$$\psi(X_{k+1}) \le f(X_k) + \langle \nabla f(X_k), \alpha_k V_k\rangle + \frac{L}{2}\|\alpha_k V_k\|^2 + h(X_k + \alpha_k V_k).$$

Since h is convex and $\alpha_k \in (0,1]$, we have

$$h(X_k + \alpha_k V_k) \le \alpha_k h(X_k + V_k) + (1-\alpha_k)h(X_k).$$

Combining the above two inequalities with (4.8) and using the definition of ϕ_k (see (3.6)), we can obtain

$$\begin{aligned}\Phi(X_{k+1},\overline{\mu}) &= \psi(X_{k+1}) + \overline{\mu}\|c(X_k+\alpha_k V_k)\|_1 \\ &\le f(X_k) + \alpha_k\langle \nabla f(X_k), V_k\rangle + \frac{L}{2}\|\alpha_k V_k\|^2 + \alpha_k h(X_k+V_k) + (1-\alpha_k)h(X_k) \\ &\quad + \overline{\mu}\|c(X_k)\|_1 - \alpha_k\overline{\mu}\|c(X_k)\|_1 + \overline{\mu}M_0\|\alpha_k V_k\|^2 \\ &\le (1-\alpha_k)\Phi(X_k,\overline{\mu}) + \alpha_k\big(f(X_k)+\phi_k(V_k)\big) + \left(\frac{L}{2} + \overline{\mu}M_0 - \frac{\kappa_1}{2\alpha_k}\right)\|\alpha_k V_k\|^2. \quad (4.31)\end{aligned}$$

By Lemma 4.4, there exists $\tilde{\epsilon} > 0$ such that $\mathcal{L}_{\overline{\rho}}(X,\Lambda^\star)$ is a convex function on the set $\{X : \|X - X^\star\| < \tilde{\epsilon}\}$ for all $\Lambda^\star \in \Gamma^\star$. From Lemma 4.6, we know that $X_k \to X^\star$, and therefore there is $K_1 > 0$ such that $\|X_k - X^\star\| < \tilde{\epsilon}$ for all $k > K_1$. Thus, we have

$$\begin{aligned}&f(X^\star) - \left(f(X_k) - \big\langle c(X_k),\Lambda^\star\big\rangle + \frac{\overline{\rho}}{2}\|c(X_k)\|^2\right) \\ &= \mathcal{L}_{\overline{\rho}}(X^\star,\Lambda^\star) - \mathcal{L}_{\overline{\rho}}(X_k,\Lambda^\star) \\ &\ge \big\langle \nabla f(X_k) - \big(Dc(X_k)\big)^*[\Lambda^\star] + \overline{\rho}\big(Dc(X_k)\big)^*[c(X_k)], X^\star - X_k\big\rangle, \quad \forall\, k > K_1, \quad \forall\,\Lambda^\star \in \Gamma^\star.\end{aligned}$$

For any $\theta \in [0,1]$, by rearranging the above inequality, we can deduce that

$$\begin{aligned}&\theta\big\langle \nabla f(X_k), X^\star - X_k\big\rangle \\ &\le \theta\big(f(X^\star) - f(X_k)\big) + \theta\big\langle Dc(X_k)[X^\star - X_k] + c(X_k), \Lambda^\star\big\rangle \\ &\quad - \theta\left(\overline{\rho}\big\langle Dc(X_k)[X^\star - X_k], c(X_k)\big\rangle + \frac{\overline{\rho}}{2}\|c(X_k)\|^2\right), \quad \forall\,\Lambda^\star \in \Gamma^\star. \quad (4.32)\end{aligned}$$

Select a constant ϵ' which satisfies $\epsilon' < \eta/M_0$. Let Λ_{k+1} be the Lagrange multiplier obtained in the step 3 of Algorithm 3.1. By Lemmas 4.5 and 4.6, we know that there exists $K_2 > 0$ such that $\mathrm{dist}(\Lambda_{k+1},\Gamma^\star) < \epsilon'$ for all $k > K_2$. Select a $\Lambda^\star \in \Gamma^\star$ such that

$$\|\Lambda_{k+1} - \Lambda^\star\| < \epsilon'. \quad (4.33)$$

Let $L_k(V,\Lambda)$ be defined by (3.7). Since V_k is the solution of (3.1) and Λ_{k+1} is the Lagrange multiplier, we have $V_k = \arg\min_V L_k(V,\Lambda_{k+1})$. For any $\theta \in [0,1]$, using $L_k(V_k,\Lambda_{k+1}) \le L_k(\theta(X^\star - X_k),\Lambda_{k+1})$ and $\phi_k(V_k) = L_k(V_k,\Lambda_{k+1})$, we can deduce that

$$\begin{aligned}&f(X_k) + \phi_k(V_k) \\ &= f(X_k) + \langle \nabla f(X_k), V_k\rangle + \frac{1}{2}\|V_k\|^2_{\mathcal{B}_k} + h(X_k+V_k) - \langle Dc(X_k)[V_k] + c(X_k), \Lambda_{k+1}\rangle \\ &\le f(X_k) + \langle \nabla f(X_k), \theta(X^\star - X_k)\rangle + \frac{1}{2}\|\theta(X^\star - X_k)\|^2_{\mathcal{B}_k} + h\big(X_k + \theta(X^\star - X_k)\big)\end{aligned}$$

$$
\begin{aligned}
&\quad - \langle Dc(X_k)[\theta(X^\star - X_k)] + c(X_k), \Lambda_{k+1}\rangle \\
&\overset{(4.32)}{\leq} \theta\Phi(X^\star,\overline{\mu}) + (1-\theta)\Phi(X_k,\overline{\mu}) \underbrace{-(1-\theta)\overline{\mu}\|c(X_k)\|_1}_{①} + \frac{1}{2}\theta^2\|X^\star - X_k\|^2_{\mathcal{B}_k} \\
&\quad \underbrace{-\langle Dc(X_k)[\theta(X^\star - X_k)] + c(X_k), \Lambda_{k+1}\rangle}_{②} + \underbrace{\theta\langle Dc(X_k)[X^\star - X_k] + c(X_k), \Lambda^\star\rangle}_{③} \\
&\quad \underbrace{-\theta\left(\overline{\rho}\langle Dc(X_k)[X^\star - X_k], c(X_k)\rangle + \frac{\overline{\rho}}{2}\|c(X_k)\|^2\right)}_{④}. \qquad (4.34)
\end{aligned}
$$

By $c(X^\star) = 0$ and (3.13), we have

$$
\|Dc(X_k)[X^\star - X_k] + c(X_k)\|_1 \leq M_0\|X^\star - X_k\|^2. \qquad (4.35)
$$

It follows from (4.33) and (4.35) that

$$
\begin{aligned}
&①+②+③ \\
&= \theta\langle Dc(X_k)[X^\star - X_k] + c(X_k), \Lambda^\star - \Lambda_{k+1}\rangle - (1-\theta)\big(\overline{\mu}\|c(X_k)\|_1 + \langle c(X_k), \Lambda_{k+1}\rangle\big) \\
&\leq \theta M_0\|\Lambda_{k+1} - \Lambda^\star\|_\infty\|X^\star - X_k\|^2 - (1-\theta)(\overline{\mu} - \|\Lambda_{k+1}\|_\infty\|)\|c(X_k)\|_1 \\
&\leq \theta M_0\epsilon'\|X^\star - X_k\|^2 \underbrace{-(1-\theta)a_0\|c(X_k)\|_1}_{⑤}, \qquad (4.36)
\end{aligned}
$$

where $a_0 := \overline{\mu} - \|\Lambda_{k+1}\|_\infty$. By (3.3), we know that

$$
\overline{\mu} \geq \mu_{k+1} \geq \|\Lambda_{k+1}\|_\infty + 1,
$$

and therefore $a_0 \geq 1$. By (4.35) again, it holds that

$$
\begin{aligned}
④ &= -\theta\left(\overline{\rho}\big\langle Dc(X_k)[X^\star - X_k], c(X_k)\big\rangle + \frac{\overline{\rho}}{2}\|c(X_k)\|^2\right) \\
&\leq \underbrace{\frac{\theta\overline{\rho}}{2}\|c(X_k)\|^2 + \theta\overline{\rho}M_0\|X^\star - X_k\|^2\|c(X_k)\|}_{⑥}. \qquad (4.37)
\end{aligned}
$$

Since $X_k \to X^\star$ and $c(X_k) \to 0$, we know that there exists an integer $K_3 > 0$ such that ⑤ + ⑥ < 0 for all $k > K_3$ and for all $\theta \in (0, 3/4]$. Substituting (4.36) and (4.37) into (4.34) yields

$$
\begin{aligned}
f(X_k) + \phi_k(V_k) \leq{}& \theta\Phi(X^\star,\overline{\mu}) + (1-\theta)\Phi(X_k,\overline{\mu}) + \frac{1}{2}\theta^2\|X^\star - X_k\|^2_{\mathcal{B}_k} \\
&+ \theta M_0\epsilon'\|X_k - X^\star\|^2, \quad \forall\,\theta \in (0, 3/4]. \qquad (4.38)
\end{aligned}
$$

Note that $\|X_k - X^\star\| < \tilde{\epsilon}$ for all $k > K_1$. By (4.23) and Assumption 3.2, we have

$$
\Phi(X_k,\overline{\mu}) - \Phi(X^\star,\overline{\mu}) \geq \eta\|X_k - X^\star\|^2 \geq \frac{\eta}{\kappa_2}\|X_k - X^\star\|^2_{\mathcal{B}_k}, \quad \forall\, k > K_1. \qquad (4.39)
$$

Recall that K is the parameter in Lemma 4.2 and m is the integer used in (3.4). Let

$$
K_0 := \max\{K_1, K_2, K_3, K + m\}.
$$

By (4.31), (4.38) and (4.39), for all $\theta \in (0, 3/4]$, we have

$$\begin{aligned}\Phi(X_{k+1},\overline{\mu}) &\le (1-\alpha_k)\Phi(X_k,\overline{\mu})\\ &\quad+\alpha_k\left[\theta\Phi(X^\star,\overline{\mu})+(1-\theta)\Phi(X_k,\overline{\mu})+\left(\frac{\kappa_2}{2\eta}\theta^2+\frac{\epsilon' M_0}{\eta}\theta\right)\big(\Phi(X_k,\overline{\mu})-\Phi(X^\star,\overline{\mu})\big)\right]\\ &\quad+\left(\frac{L}{2}+\overline{\mu}M_0\right)\|\alpha_k V_k\|^2\\ &= \Phi(X_k,\overline{\mu})+\alpha_k\big(\Phi(X_k,\overline{\mu})-\Phi(X^\star,\overline{\mu})\big)\left[\frac{\kappa_2}{2\eta}\theta^2+\left(\frac{\epsilon' M_0}{\eta}-1\right)\theta\right]\\ &\quad+\overline{L}\alpha_k^2\|V_k\|^2,\quad \forall\, k>K_0, \end{aligned}\tag{4.40}$$

where $\overline{L} := L/2+\overline{\mu}M_0$ and $\alpha_k \in (0,1]$. Define

$$q(\theta) := 1+\alpha_k\left(\frac{\kappa_2}{2\eta}\theta^2+\left(\frac{\epsilon' M_0}{\eta}-1\right)\theta\right).$$

Without loss of generality, we can assume that $\eta < \kappa_2/2$ and $\epsilon' < \eta/M_0$. Thus $q(\theta) > 0$ for all θ. Subtracting $\Phi(X^\star,\overline{\mu})$ from both sides of (4.40) yields

$$\begin{aligned}\Phi(X_{k+1},\overline{\mu})-\Phi(X^\star,\overline{\mu}) &\le \big(\Phi(X_k,\overline{\mu})-\Phi(X^\star,\overline{\mu})\big)q(\theta)+\overline{L}\|V_k\|^2\\ &\le \big(\Phi_k-\Phi(X^\star,\overline{\mu})\big)q(\theta)+\overline{L}\|V_k\|^2,\quad \forall\, k>K_0. \end{aligned}\tag{4.41}$$

Part (II). Next we prove there exists $\nu \in (0,1)$ such that

$$\Phi(X_{k+1},\overline{\mu})-\Phi(X^\star,\overline{\mu}) \le \nu\big(\Phi_k-\Phi(X^\star,\overline{\mu})\big),\quad \forall\, k>K_0.\tag{4.42}$$

Choose a real number ω which satisfies

$$\omega < \min\left\{\frac{2}{\sigma\gamma\overline{\alpha}\kappa_1},\frac{\gamma\overline{\alpha}(\eta-\epsilon' M_0)^2}{2\kappa_2\eta\overline{L}}\right\},$$

where $\overline{\alpha}$ is from (4.5) and γ is the parameter of Algorithm 3.1. We can prove (4.42) in two different situations.

(i) $\|V_k\|^2 \ge \omega(\Phi_k-\Phi(X^\star,\overline{\mu}))$. By $\alpha_k \ge \gamma\overline{\alpha}$ (see Lemma 4.3), we have

$$\begin{aligned}\frac{2}{\sigma\gamma\overline{\alpha}\kappa_1}\big(\Phi_k-\Phi(X_{k+1},\overline{\mu})\big) &\ge \frac{2}{\sigma\alpha_k\kappa_1}\big(\Phi_k-\Phi(X_{k+1},\overline{\mu})\big)\\ &\ge \|V_k\|^2 \ge \omega\big(\Phi_k-\Phi(X^\star,\overline{\mu})\big),\end{aligned}$$

where the second inequality uses (3.5). From the above inequality, we obtain that

$$\Phi(X_{k+1},\overline{\mu})-\Phi(X^\star,\overline{\mu}) \le \left(1-\frac{\sigma\gamma\overline{\alpha}\kappa_1\omega}{2}\right)\big(\Phi_k-\Phi(X^\star,\overline{\mu})\big),$$

which implies (4.42).

(ii) $\|V_k\|^2 < \omega(\Phi_k-\Phi(X^\star,\overline{\mu}))$. Combining it with (4.41) yields

$$\Phi(X_{k+1},\overline{\mu})-\Phi(X^\star,\overline{\mu}) \le \big(\Phi_k-\Phi(X^\star,\overline{\mu})\big)\big(q(\theta)+\overline{L}\omega\big).\tag{4.43}$$

By the definitions of Φ_k and $l(k)$ (see (3.4) and (4.10)), we know that $\Phi_k = \Phi(X_{l(k)}, \overline{\mu})$. Denote $r_k := \Phi(X_k, \overline{\mu}) - \Phi(X^\star, \overline{\mu})$ for all k. It follows from (4.43) that

$$r_{k+1} \le \big(q(\theta) + \overline{L}\omega\big) r_{l(k)}, \tag{4.44}$$

Let $\theta_{\min} = \arg\min q(\theta)$. Then $\theta_{\min} = (\eta - \epsilon' M_0)/\kappa_2 \in (0, 1/2)$. Since $\omega < \gamma\overline{\alpha}(\eta - \epsilon' M_0)^2/(2\kappa_2\eta\overline{L})$, we have

$$\begin{aligned} q(\theta_{\min}) + \overline{L}\omega &= 1 - \frac{\alpha_k(\eta - \epsilon' M_0)^2}{2\kappa_2\eta} + \overline{L}\omega \\ &\le 1 - \frac{\gamma\overline{\alpha}(\eta - \epsilon' M_0)^2}{2\kappa_2\eta} + \overline{L}\omega < 1. \end{aligned}$$

By substituting $\theta_{\min}$ into (4.44), we can see that (4.42) holds.

Part (III). For any $k \ge K_0$, there exists an integer $i \ge 1$ such that

$$(i-1)(m+1) < k - K_0 \le i(m+1),$$

where m is from (3.4). We write (4.42) as $r_{k+1} \le \nu r_{l(k)}$. By the definition of $l(k)$, it holds that $k - m \le l(k)$. Then, by using (4.42) and (4.12) recursively, we can derive that

$$\begin{aligned} r_k &\le r_{l(k)} \le \nu r_{l(l(k)-1)} \le \nu r_{l(k-m-1)} \le \cdots \\ &\le \nu^{i-1} r_{l(k-(i-1)(m+1))} \le \nu^{i-1} r_{l(K_0)} \le \nu^{\frac{k-K_0}{m+1}-1} r_{l(K_0)}. \end{aligned}$$

Taking $\beta := 1/\nu$ and $\tau := \nu^{1/(m+1)}$ in the above inequality, we can obtain $r_k \le \beta\tau^{k-K_0} r_{l(K_0)}$, which completes the proof. □

Remark 4.2. From the proof of Theorem 4.3, we can see that there exists a neighborhood of $X^\star$ such that when X_k enters the neighborhood and $\mu_k = \overline{\mu}, \Phi(X_k, \mu_k) - \Phi(X^\star, \overline{\mu})$ begins to converge at a R-linear rate. If the condition $\overline{\mu} > \sup_{\Lambda^\star \in \Gamma} \|\Lambda^\star\|_\infty + 1$ is replaced by $\overline{\mu} > \sup_{\Lambda^\star \in \Gamma} \|\Lambda^\star\|_\infty$, the assertion holds also.

Remark 4.3. In Theorem 4.3, we only prove the local linear convergence of the SQP-PG method. For classical SQP methods, local superlinear (or quadratic) convergence can be established; see [32, Chapter 18] and the references therein. Due to the nonsmooth term h, it is hard to prove local superlinear convergence of SQP-type methods for composite optimization problems with equality constraints. The difficulty of proving superlinear convergence lies in the fact that the quadratic approximation to h does not exist. Thus, the technique used in SQP methods does not apply here. We leave the challenging task of designing an SQP-type method with local superlinear convergence as our future work.

Corollary 4.1. *Under the assumptions of Theorem* 4.3*, there exists a constant* $C_{K_0} > 0$ *such that*

$$\|X_k - X^\star\| \le C_{K_0} \sqrt{\tau}^{\,k} \tag{4.45}$$

for all $k \ge K_0$*, where* $\tau \in (0, 1)$ *and* K_0 *are as in Theorem* 4.3*.*

Proof. It follows from Theorem 4.3 and (4.39) that

$$\begin{aligned}\|X_k - X^\star\|^2 &\le \frac{1}{\eta}\big(\Phi(X_k,\overline{\mu}) - \Phi(X^\star,\overline{\mu})\big)\\ &\le \frac{1}{\eta}\beta\tau^{k-K_0}\big(\Phi_{K_0} - \Phi(X^\star,\overline{\mu})\big)\\ &= \tau^k C,\end{aligned}$$

where

$$C := \frac{1}{\eta}\beta\tau^{-K_0}\big(\Phi_{K_0} - \Phi(X^\star,\overline{\mu})\big).$$

By letting $C_{K_0} = \sqrt{C}$, we can obtain (4.45) from the above inequality. □

5. Numerical Results

In this section, we report the numerical experiments comparing SQP-PG with ManPG and ManPG-Ada in [11], and the LSq methods in [46] equipped with two different line search methods. Our test problems include the compressed modes (CM) problems, the sparse principle component analysis (Sparse PCA) problems, the quadratic constrained composite optimization (QCCO) problems and the sparse multiview canonical correlation analysis (Sparse MCCA) problems. All of these experiments were conducted in MATLAB R2021b on a PC using Windows 11 (64 bit) system with AMD Ryzen 7 5800H CPU (3.20 GHz) and 16 GB memory.

The ManPG methods and the LSq methods are all manifold-based algorithms. In their framework, the feasible set of the optimization problem is treated as a Riemannian manifold. An important part in manifold-based methods is the numerical computation of retraction. For the QCCO problems and the sparse MCCA problems, we can only use the nearest-point projection to the feasible set as the retraction, which are usually computationally expensive for large-scale problems. The SQP-PG method does not involve retraction operations and is suitable for such problems. Numerical experiments show an advantage of SQP-PG over ManPG and LSq for these problems.

For the stopping criterion, we terminate our algorithm when $\|V_k\|^2 \le \hat{\epsilon}$ or the algorithm reaches the maximum iteration number, where V_k is defined by (3.1), and $\hat{\epsilon} > 0$ is the tolerance. For the CM, sparse PCA and QCCO problems, we set $\hat{\epsilon} = 10^{-5}$, and set $\hat{\epsilon} = 10^{-10}$ for the sparse MCCA problem. The parameters used in ManPG, ManPG-Ada and LSq are set to be the default values in [11] and [46]. For the CM and sparse PCA problems, we use the singular value decomposition (SVD) as the retraction in ManPG, ManPG-Ada, and use the QR decomposition in LSq; for the QCCO and sparse MCCA problems, we use the nearest-point projection as the retraction.

Reported results are averaged over 50 runs with different random initial points. Numerical results are shown in several tables. We compare the running time, loss function (the optimal value of the objective function), iteration number and sparsity of solution of the five methods. In the tables, LSq-I and LSq-II denote the LSq methods with two different line search methods, "T" denotes the CPU time in seconds, "I" represents the number of iterations, "L" denotes the value of the loss function, and "S" denotes the percentage of zeros in the local minimum $X^\star$.

5.1. Numerical results on CM

The CM problem seeks for spatially localized solutions of the independent-particle Schrödinger equation. Let H be a discretized Schrödinger operator. Then the CM problem is formulated as

$$\begin{aligned} \min_{X\in\mathbb{R}^{n\times r}} \ & \mathrm{tr}(X^\top HX) + \Upsilon\|X\|_1 \\ \text{s.t.} \ & X^\top X = I_r. \end{aligned}$$

In our numerical experiments, we use a linear operator $\mathbf{B}_k$ to approximate the linear operator $\mathcal{B}_k$, where $\mathcal{B}_k$ is from (3.1). The operator $\mathbf{B}_k$ is updated by a damped LBFGS method. This strategy is used in [40]. We refer to Section 3.2 of the paper for more details. By [40, Lemma 3.1], we know that $\mathbf{B}_k$ satisfies Assumption 3.2.

We can observe from Tables 5.1 that our SQP-PG method is comparable to the ManPG and LSq methods on the CM problems in terms of accuracy, sparsity and running time. When r grows larger, LSq-I and LSq-II show better performance than other methods. The LSq methods need much less iterations to achieve the same accuracy, and can achieve a solution with slightly better sparsity than other methods for the CM problems.

Table 5.1: Comparison on CM.

(n,r,Υ)	ManPG				ManPG-Ada				LSq-I			
	T	I	L	S	T	I	L	S	T	I	L	S
(128,6,0.4)	0.1082	1622.3	8.544	0.9136	0.0652	671.8	8.544	0.9135	0.4457	49.6	8.544	0.9141
(256,6,0.4)	3.072	14529.2	11.31	0.9170	2.226	9195.6	11.31	0.9169	0.7645	48.1	11.31	0.9215
(512,6,0.4)	5.230	17570.2	14.94	0.9263	7.897	16089.5	14.94	0.9263	3.396	71.6	14.94	0.9295
(1024,6,0.4)	7.552	27838.1	19.72	0.9343	2.175	8838.4	19.72	0.9344	5.016	76.7	19.72	0.9385
(256,4,0.4)	0.5921	5314.6	9.009	0.9274	0.6844	4459.1	9.009	0.9274	0.4960	55.7	9.009	0.9294
(256,8,0.4)	6.015	16245.2	15.08	0.9144	3.067	7425.4	15.08	0.9144	0.8887	46.7	15.08	0.9205
(256,10,0.4)	12.00	21276.2	18.85	0.9107	5.841	8838.4	18.85	0.9107	1.028	45.3	18.85	0.9197
(256,6,0.1)	0.7113	12010.3	3.735	0.8373	0.6432	5184.2	3.735	0.8373	1.865	82.5	3.734	0.8563
(256,6,0.2)	0.5449	8178.6	6.450	0.8837	1.397	6485.2	6.450	0.8836	1.272	62.9	6.499	0.8943
(256,6,0.3)	0.5183	6200.0	8.987	0.9071	1.462	6604.1	8.987	0.9070	0.5630	52.0	8.986	0.9134
(256,6,0.5)	2.320	10590.1	13.51	0.9254	0.9986	4490.3	13.51	0.9254	0.5497	44.5	13.51	0.9283

(n,r,Υ)	LSq-II				SQP-PG			
	T	I	L	S	T	I	L	S
(128,6,0.4)	0.4202	48.0	8.544	0.9131	0.0659	451.6	8.544	0.9119
(256,6,0.4)	0.7629	47.7	11.31	0.9215	0.2347	1439.5	11.31	0.9163
(512,6,0.4)	2.877	70.6	14.94	0.9295	1.385	3432.3	14.94	0.9258
(1024,6,0.4)	4.522	76.3	19.72	0.9385	4.697	9845.2	19.72	0.9346
(256,4,0.4)	0.4907	55.1	9.009	0.9292	0.0990	1065.8	9.009	0.9268
(256,8,0.4)	0.8757	46.6	15.08	0.9205	0.7279	2288.2	15.08	0.9142
(256,10,0.4)	1.011	44.3	18.85	0.9196	2.145	3300.5	18.85	0.9114
(256,6,0.1)	1.267	75.7	3.734	0.8561	0.5601	3122.3	3.735	0.8354
(256,6,0.2)	1.133	61.76	6.499	0.8943	0.3570	2087.9	6.502	0.8838
(256,6,0.3)	0.5561	51.4	8.986	0.9131	0.2362	1340.5	8.991	0.9037
(256,6,0.5)	0.5306	44.4	13.51	0.9273	0.1819	1093.0	13.51	0.9251

5.2. Numerical results on SPCA

Sparse PCA seeks principal components with very few components. For a given data matrix $A \in \mathbb{R}^{50\times n}$, the SPCA problem that seeks the leading r ($r < \min\{50, n\}$) sparse loading vectors can be formulated as

$$\begin{aligned} \min \ & -\mathrm{tr}(X^\top A^\top A X) + \Upsilon\|X\|_1 \\ \text{s.t.} \ & X^\top X = I_r. \end{aligned}$$

When $\Upsilon = 0$, the problem above reduces to calculating the leading r eigenvalues and the corresponding eigenvectors of $A^\top A$. When $\mu > 0$, the l_1 norm $\|X\|_1$ can promote sparsity of the loading vectors.

In our experiments, the random data matrices A were generated in the following way: First, we randomly generate A from the standard Gaussian distribution. Then, we modify the singular values of A to $\{\omega_i^4 + 10^{-5}\}_{i=1}^{50}$ to make A ill-conditioned, where $\{\omega_i\}$ is sampled from the standard Gaussian distribution. Finally, we shift the columns of A to make their mean equal to 0 and normalize the columns of A so that their Euclidean norms are equal to 1. As in Section 5.1, we also use the approximate quasi-Newton strategy proposed in [40] to update the matrix $\mathcal{B}_k$ in (3.1).

Table 5.2: Comparison on SPCA.

(n, r, Υ)	ManPG				ManPG-Ada				LSq-I			
	T	I	L	S	T	I	L	S	T	I	L	S
(500,6,1)	0.8936	4160.0	-337.1	0.2275	0.5009	1807.0	-337.1	0.2273	1.378	43.5	-337.1	0.2286
(1000,6,1)	2.613	7956.7	-732.0	0.1919	1.382	3227.8	-732.0	0.1919	3.623	48.5	-732.0	0.1922
(1500,6,1)	4.429	9813.5	-1140.3	0.1460	2.489	4182.5	-1140.3	0.1460	3.953	41.1	-1140.3	0.1466
(2000,6,1)	5.838	11900.6	-1560.6	0.1280	3.452	5130.6	-1560.6	0.1280	6.153	38.7	-1560.6	0.1283
(2500,6,1)	7.010	13250.1	-1919.4	0.1015	4.066	5597.1	-1919.4	0.1015	11.23	61.4	-1919.4	0.1030
(3000,6,1)	10.34	17368.5	-2387.2	0.1156	6.431	7670.4	-2387.2	0.1156	18.15	75.6	-2387.2	0.1159
(1500,4,1)	1.382	5344.9	-1033.2	0.0795	0.7805	2286.1	-1033.2	0.0795	4.705	52.3	-1033.2	0.0800
(1500,8,1)	8.370	15265.2	-1155.2	0.2165	4.810	6512.7	-1155.2	0.2165	7.964	46.4	-1155.2	0.2185
(1500,10,1)	16.33	22034.8	-1205.4	0.3807	9.821	9639.2	-1205.4	0.3803	7.688	49.2	-1205.4	0.3824
(1500,6,0.6)	5.643	12528.6	-1210.2	0.0996	3.133	5352.6	-1210.2	0.0995	5.650	49.6	-1210.2	0.0999
(1500,6,0.8)	4.586	9715.6	-1166.7	0.1197	2.264	4204.2	-1166.7	0.1197	4.400	40.9	-1166.7	0.1203
(1500,6,1.2)	3.853	8180.4	-1113.6	0.1847	2.110	3392.6	-1113.6	0.1847	5.935	48.6	-1113.6	0.1858

(n, r, Υ)	LSq-II				SQP-PG			
	T	I	L	S	T	I	L	S
(500,6,1)	1.358	43.7	-337.3	0.2286	0.8714	1985.7	-337.0	0.2239
(1000,6,1)	3.283	48.2	-732.1	0.1922	2.615	3559.5	-731.9	0.1906
(1500,6,1)	4.304	41.3	-1140.5	0.1466	4.635	4032.1	-1140.2	0.1435
(2000,6,1)	6.069	39.3	-1560.7	0.1283	6.616	5086.5	-1560.4	0.1276
(2500,6,1)	6.363	47.9	-1919.7	0.1030	8.046	5465.5	-1919.3	0.1008
(3000,6,1)	14.08	68.7	-2387.3	0.1158	17.12	8935.6	-2387.0	0.1149
(1500,4,1)	4.652	45.8	-1033.4	0.0799	1.213	2094.9	-1033.2	0.0798
(1500,8,1)	5.082	40.8	-1155.5	0.2185	11.85	7658.2	-1155.1	0.2149
(1500,10,1)	7.393	44.6	-1205.6	0.3824	18.28	8103.9	-1205.2	0.3801
(1500,6,0.6)	4.866	53.3	-1210.3	0.0999	6.191	5171.7	-1210.1	0.0983
(1500,6,0.8)	4.405	37.8	-1167.0	0.1203	5.517	4572.6	-1166.6	0.1183
(1500,6,1.2)	5.655	42.1	-1113.6	0.1858	4.662	3918.6	-1113.4	0.1844

From Table 5.2, we can observe that our SQP-PG is comparable to the ManPG and LSq methods on the SPCA problems. When r is small, ManPG-Ada is the best method among these methods. When r grows larger, LSq-II shows better performance than other methods. In most cases, the LSq methods can achieve the best optimal value of the loss function, and the best solution sparsity among these methods for the SPCA problems.

5.3. Numerical results on QCCO problems

Quadratic constrained quadratic programming (QCQP) is an important optimization problem in optimization and has wide applications in many fields. The QCCO problem is a variant of QCQP, which is proposed to seek sparse solutions of QCQP. Sparse canonical correlation analysis (Sparse CCA) can be viewed as a special case of the QCCO problem. For more details of CCA and SCCA, the reader is referred to [12,39]. The QCCO problem can be formulated as

$$\begin{aligned} &\min\ x^\top Hx + \Upsilon\|x\|_1 \\ &\text{s.t.}\quad \frac{1}{2}x^\top A_i x + b_i^\top x + c_i = 0, \quad i = 1,\dots,m, \end{aligned} \tag{5.1}$$

where $x, b_i \in \mathbb{R}^n$ and $H, A_i \in \mathbb{R}^{n\times n}$ for $i = 1, 2, \dots, m$.

In our experiments, the matrix H in (5.1) is generated in the following way: We first generate a diagonal matrix D by using Matlab function $D = \text{diag}(\text{rand}(n, 1))$ and a random orthogonal matrix by $P = \text{orth}(\text{rand}(n))$, and then let $H = P^\top DP$. The matrices $A_i, i = 1, \dots, m$, are generated in the same way as H. In the implementation of SQP-PG, we use $\text{diag}(\mathbf{B}_k)$ to approximate $\mathcal{B}_k$ in (3.1), where $\mathbf{B}_k$ is updated by the damped LBFGS method. Similar to the proof of [40, Lemma 3.1], we can prove that $\text{diag}(\mathbf{B}_k)$ satisfies Assumption 3.2.

The feasible set of (5.1), denoted by Ω, is a submanifold of $\mathbb{R}^n$. For this manifold, we use the nearest-point projection to Ω as the retraction. The problem of finding the projection point is solved by Newton's method, whose computational cost grows rapidly when n becomes larger.

Thus, the manifold-based methods will require more time to converge for large-scale problems than the SQP-PG method. This phenomenon can be observed from Table 5.3. We can see that our method is much faster than other methods especially when n is large. ManPG methods can achieve a solution with slightly better sparsity than SQP-PG, and SQP-PG shows better solution sparsity than LSq methods.

5.4. Numerical results on Sparse MCCA

As an extension of the principal component analysis (PCA) and the traditional canonical correlation analysis (CCA), the multiview canonical correlation analysis (MCCA) is a statistical approach that deals with the relation between multiset random variables. For more details about MCCA, we refer to [45] and the references therein. The aim of Sparse MCCA is to find sparse solutions of the MCCA problem. For each $x \in \mathbb{R}^n$, we partition it in the form $x = [x_1^\top, \dots, x_m^\top]^\top$, where $x_i \in \mathbb{R}_i^n$ and $\sum_{i=1}^m n_i = n$. The covariance matrix Σ can be partitioned accordingly, and the (i, j)-th submatrix of Σ is $\Sigma_{ij} = \text{cov}(x_i, x_j) \in \mathbb{R}^{n_i\times n_j}$. Then the Sparse MCCA problem can be formulated as

$$\begin{aligned} &\min_x\ -x^\top \Sigma x + \Upsilon\|x\|_1 \\ &\text{s.t.}\quad x_i^\top \Sigma_{ii} x_i = 1, \quad i = 1,\dots,m. \end{aligned}$$

Table 5.3: Comparison on QCCO.

(n,m,Υ)	ManPG				ManPG-Ada				LSq-I			
	T	I	L	S	T	I	L	S	T	I	L	S
(100,10,0.4)	0.1244	196.5	23.10	0.4894	0.0787	107.0	23.10	0.4896	0.0813	17.0	23.10	0.4364
(200,10,0.4)	0.4247	291.2	44.15	0.5395	0.2556	138.9	44.15	0.5398	0.2403	15.5	44.15	0.5090
(500,10,0.4)	9.827	679.0	108.9	0.5664	5.259	260.1	108.9	0.5664	3.525	11.2	108.9	0.5131
(1000,10,0.4)	58.59	984.8	217.6	0.5749	30.84	358.3	217.6	0.5751	19.12	9.4	217.6	0.4819
(2000,10,0.4)	323.5	1185.0	428.3	0.5735	173.0	425.2	428.3	0.5735	76.45	8.2	428.3	0.5645
(5000,10,0.4)	4828.0	2667.0	1064.9	0.5808	2659.5	876.1	1064.9	0.5808	817.1	6.5	1065.1	0.5725
(500,1,0.4)	2.589	419.5	95.96	0.5900	1.505	182.2	95.96	0.5900	0.6658	11.0	95.96	0.5731
(500,20,0.4)	7.990	468.1	111.9	0.5461	4.144	195.1	111.9	0.5461	4.471	11.7	111.9	0.5168
(500,50,0.4)	14.86	518.6	117.1	0.5130	8.182	210.8	117.1	0.5130	11.83	12.5	117.1	0.4362
(500,10,0.3)	8.580	714.4	89.38	0.4864	4.911	286.2	89.38	0.4864	2.762	11.6	89.38	0.4278
(500,10,0.5)	7.331	601.0	126.5	0.6131	4.031	231.3	126.5	0.6132	3.163	11.8	126.5	0.5925
(500,10,0.6)	8.696	768.5	143.2	0.6672	4.187	270.2	143.2	0.6672	4.412	11.6	143.2	0.5914
(500,10,0.7)	6.670	602.5	161.0	0.7101	3.486	220.1	161.0	0.7103	3.842	10.3	161.1	0.5535

(n,m,Υ)	LSq-II				SQP-PG			
	T	I	L	S	T	I	L	S
(100,10,0.4)	0.1126	18.3	23.10	0.4398	0.0254	78.1	23.10	0.4870
(200,10,0.4)	0.2976	16.7	44.15	0.5207	0.0687	109.1	44.15	0.5386
(500,10,0.4)	4.521	11.8	108.9	0.5611	1.231	177.5	108.9	0.5649
(1000,10,0.4)	25.54	10.1	217.6	0.5468	6.286	227.1	217.6	0.5688
(2000,10,0.4)	126.1	8.8	428.3	0.5695	26.96	282.4	428.3	0.5712
(5000,10,0.4)	923.9	7.4	1065.1	0.5743	155.7	290.7	1065.2	0.5763
(500,1,0.4)	1.165	11.6	95.96	0.5654	0.1323	111.2	95.96	0.5885
(500,20,0.4)	5.274	12.1	111.9	0.5172	1.451	159.6	111.9	0.5455
(500,50,0.4)	17.62	14.3	117.1	0.4635	4.992	221.3	117.1	0.5111
(500,10,0.3)	2.940	11.7	89.38	0.4787	1.150	202.2	89.38	0.4836
(500,10,0.5)	4.299	12.9	126.5	0.6106	1.092	190.0	126.5	0.6110
(500,10,0.6)	4.848	11.5	143.2	0.6645	1.052	206.9	143.2	0.6659
(500,10,0.7)	4.740	11.1	161.1	0.6492	0.7346	162.1	161.1	0.7045

In our experiments, matrices Σ and Σ_{ii}, $i = 1,\ldots,m$, are generated as follows: We first generate a $n\times n$ symmetric positive definite matrix Σ, and then divide $\{1,\ldots,n\}$ into m groups and partition Σ into a block matrix according to this group, lastly set the diagonal blocks of Σ as $\Sigma_{11},\cdots,\Sigma_{mm}$. In the implementation of SQP-PG, we approximate $\mathcal{B}_k$ (see (3.1)) by the diagonal matrix of $\mathbf{B}_k$, which is updated by the damped LBFGS method. Then $\mathrm{diag}(\mathbf{B}_k)$ satisfies Assumption 3.2.

For the sparse MCCA problem, we use the nearest-point projection to the constrained manifold as the retraction. When the dimension n increases, the total computational cost of manifold-based methods grows rapidly. This phenomenon can be observed from Table 5.4. We can see from these tables that all algorithms need more CPU time to converge as n increases, and SQP-PG is much faster than other methods especial for large-scale problems. These numerical results demonstrate the advantage of SQP-PG over the manifold-based methods for the general equality constrained composite optimization problems.

Table 5.4: Comparison on SMCCA.

(n,m,Υ)	ManPG				ManPG-Ada				LSq-I			
	T	I	L	S	T	I	L	S	T	I	L	S
(100,10,0.1)	0.1229	269.0	-15.89	0.2770	0.1733	194.6	-15.98	0.2762	0.4510	40.0	-16.00	0.2710
(200,10,0.1)	0.4071	408.9	-14.90	0.3940	0.5138	221.1	-14.90	0.3931	0.9277	40.0	-14.86	0.3622
(500,10,0.1)	3.133	550.5	-13.37	0.6069	2.911	238.1	-13.37	0.6058	5.957	40.0	-13.38	0.5386
(1000,10,0.1)	26.00	871.1	-11.96	0.7839	16.53	270.7	-11.96	0.7833	47.83	40.0	-11.89	0.7663
(2000,10,0.1)	135.6	883.4	-10.64	0.9197	60.54	241.2	-10.64	0.9194	236.8	40.0	-10.58	0.9118
(5000,10,0.1)	1487.3	1401.8	-9.612	0.9935	532.1	272.8	-9.612	0.9935	1487.0	40.0	-9.573	0.9925
(500,5,0.1)	3.357	628.8	-5.894	0.7872	1.861	207.6	-5.894	0.7867	5.622	40.0	-5.881	0.7346
(500,15,0.1)	4.068	667.3	-21.13	0.5250	5.291	338.6	-21.13	0.5250	6.710	40.0	-21.14	0.4554
(500,20,0.1)	4.216	675.4	-29.14	0.4587	6.667	396.7	-29.14	0.4577	6.539	40.0	-29.14	0.3977
(500,10,0.05)	3.144	554.8	-16.10	0.3159	5.865	367.3	-16.10	0.3139	5.797	40.0	-16.09	0.2817
(500,10,0.15)	3.353	560.6	-11.48	0.8037	1.829	191.5	-11.48	0.8024	5.931	40.0	-11.49	0.7552
(500,10,0.2)	3.219	546.7	-10.00	0.9127	1.519	178.4	-9.974	0.9130	4.562	40.0	-10.02	0.9038

(n,m,Υ)	LSq-II				SQP-PG			
	T	I	L	S	T	I	L	S
(100,10,0.1)	0.5752	40.0	-15.90	0.2328	0.3365	1364.3	-16.01	0.2778
(200,10,0.1)	1.1712	40.0	-14.75	0.3562	0.4941	1424.7	-14.89	0.3987
(500,10,0.1)	8.598	40.0	-11.37	0.5874	1.220	1384.0	-13.35	0.6089
(1000,10,0.1)	54.89	40.0	-11.88	0.7760	6.316	1810.4	-11.95	0.7865
(2000,10,0.1)	228.5	40.0	-10.62	0.9198	28.53	2000.0	-10.61	0.9217
(5000,10,0.1)	1572.0	40.0	-9.564	0.9925	155.6	1838.6	-9.597	0.9939
(500,5,0.1)	6.673	40.0	-5.896	0.7851	0.8929	1606.5	-5.890	0.7856
(500,15,0.1)	8.630	40.0	-21.07	0.5096	1.684	1471.3	-21.15	0.5250
(500,20,0.1)	9.924	40.0	-29.07	0.4453	1.837	1637.9	-29.12	0.4573
(500,10,0.05)	7.659	40.0	-15.97	0.2707	1.599	1679.6	-16.10	0.3155
(500,10,0.15)	6.952	40.0	-11.40	0.7911	1.111	1632.3	-11.46	0.8009
(500,10,0.2)	5.375	40.0	-9.940	0.9077	1.095	1643.3	9.992	0.9137

6. Conclusion and Future Work

In this paper, we present an SQP-type proximal gradient method, named SQP-PG, for composite optimization problems with general equality constraints. Under some mild conditions, we establish the global convergence of SQP-PG. If the second-order sufficient condition holds at the local minimizer, the local linear convergence of SQP-PG is also proved. An advantage of our method is that, compared to the Riemannian manifold optimization method, it does not involve the computation of retraction. Numerical experiments show that our method is quite efficient especially when the retraction to the feasible set defined by the equality constraints is expensive.

As suggested insightfully by one of the referees, it is important to construct SQP-type algorithms with local superlinear convergence for composite optimization problems with equality constraints. Thus, one of the topics for our future research will be such algorithms. Our future

work also includes considering the modern update method of μ_k which involves linear/quadratic models of the merit function. Since this paper focuses only on composite optimization problems with equality constraints, another topic of our future works will be concentrated on composite optimization with inequality constraints.

Acknowledgements. The authors would like to thank the two anonymous referees and the associate editor for their valuable comments and constructive suggestions, which have greatly improved the quality and the presentation of this paper.

The work of W.H. Yang was supported by the National Natural Science Foundation of China (Grant No. 72394365).

References

[1] P.A. Absil and S. Hosseini, A collection of nonsmooth Riemannian optimization problems, in: *Nonsmooth Optimization and its Applications. International Series of Numerical Mathematics*, Birkhäuser, **170** (2019), 1–15.

[2] P.A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2008.

[3] P.A. Absil and J. Malick, Projection-like retractions on matrix manifolds, *SIAM J. Optim.*, **22** (2012), 135–158.

[4] A. Beck, *First-Order Methods in Optimization*, SIAM, 2017.

[5] A.S. Berahas, F.E. Curtis, D. Robinson, and B. Zhou, Sequential quadratic optimization for nonlinear equality constrained stochastic optimization, *SIAM J. Optim.*, **31** (2021), 1352–1379.

[6] J.F. Bonnas and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, 2000.

[7] J.V. Burke, F.E. Curtis, and H. Wang, A sequential quadratic optimization algorithm with rapid infeasibility detection, *SIAM J. Optim.*, **24** (2014) , 839–872.

[8] J.V. Burke, F.E. Curtis, H. Wang, and J. Wang, Inexact sequential quadratic optimization with penalty parameter updates within the QP solver, *SIAM J. Optim.*, **30** (2020), 1822–1849.

[9] R.H. Byrd, F.E. Curtis, and J. Nocedal, An inexact SQP method for equality constrained optimization, *SIAM J. Optim.*, **19** (2008), 351–369.

[10] R.H. Byrd, F.E. Curtis, and J. Nocedal, Infeasibility detection and SQP methods for nonlinear optimization, *SIAM J. Optim.*, **20** (2010), 2281–2299.

[11] S. Chen, S. Ma, A.M.C. So, and T. Zhang, Proximal gradient method for nonsmooth optimization over the Stiefel manifold. *SIAM J. Optim.*, **30** (2020), 210–239.

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

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

[14] F.H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM, 1990.

[15] F.E. Curtis, T.C. Johnson, D.P. Robinson, and A. Wächter, An inexact sequential quadratic optimization algorithm for nonlinear optimization, *SIAM J. Optim.*, **24** (2014), 1041–1074.

[16] Y.H. Dai, A nonmonotone conjugate gradient algorithm for unconstrained optimization, *J. Syst. Sci. Complex.*, **15** (2002), 139–145.

[17] F. Facchinei and J.S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer, 2007.

[18] B. Gao, X. Liu, and Y.X. Yuan, Parallelizable algorithms for optimization problems with orthogonality constraints, *SIAM J. Sci. Comput.*, **41** (2019), 1949–1983.

[19] U.M. Garcia-Palomares and O.L. Mangasarian, Superlinearly convergent quasi-Newton algorithms for nonlinearly constrained optimization problems, *Math. Program.*, **11** (1976), 1–13.
[20] P.E. Gill, W. Murray, and M.A. Saunders, SNOPT: An SQP algorithm for large-scale constrained optimization, *SIAM J. Optim.*, **12** (2002), 979–1006.
[21] P. Grohs, and S. Hosseini, ϵ-subgradient algorithms for locally Lipschitz functions on Riemannian manifolds, *Adv. Comput. Math.*, **42** (2016), 333–360.
[22] S.P. Han, Superlinearly convergent variable metric algorithms for general nonlinear programming problems, *Math. Program.*, **11** (1976), 263–282.
[23] S.P. Han, A globally convergent method for nonlinear programming, *J. Optim. Theory Appl.*, **22** (1977), 297–309.
[24] S.P. Han and O.L. Mangasarian, Exact penalty functions in nonlinear programming, *Math. Program.*, **17** (1979), 251–269.
[25] W. Huang and K. Wei, Riemannian proximal gradient methods, *Math. Program.*, **194** (2022), 371–413.
[26] R. Lai and S. Osher, A splitting method for orthogonality constrained problems, *J. Sci. Comput.*, **58** (2014), 431–449.
[27] J.D. Lee, Y. Sun, and M.A. Saunders, Proximal Newton-type methods for minimizing composite functions, *SIAM J. Optim.*, **24** (2014), 1420–1443.
[28] X.W. Liu and Y.X. Yuan, A robust algorithm for optimization with general equality and inequality constraints, *SIAM J. Sci. Comput.*, **22** (2000), 517–534.
[29] X.W. Liu and Y.X. Yuan, A sequential quadratic programming method without a penalty function or a filter for nonlinear equality constrained optimization, *SIAM J. Optim.*, **21** (2011), 545–571.
[30] Z. Lu and Y. Zhang, An augmented Lagrangian approach for sparse principal component analysis, *Math. Program.*, **135** (2012), 149–193.
[31] J.J. Moré and D.C. Sorensen, Computing a trust region step, *SIAM J. Sci. Statist. Comput.*, **4** (1983), 553–572.
[32] J. Nocedal and S.J. Wright, *Numerical Optimization*, Springer, 2006.
[33] P. Patrinos, L. Stella, and A. Bemporad, Forward-backward truncated Newton methods for convex composite optimization, *arXiv:1402.6655*, 2014.
[34] M.J.D. Powell, A fast algorithm for nonlinearly constrained optimization calculations, in: *Numerical analysis (Proc. 7th Biennial Conf., Univ. Dundee, Dundee, 1977)*, Springer, 1978, 144–157.
[35] M.J.D. Powell, Algorithms for nonlinear constraints that use Lagrangian functions, *Math. Program.*, **14** (1978), 224–248.
[36] L. Qi, and J. Sun, A nonsmooth version of Newton's method, *Math. Program.*, **58** (1993), 353–367.
[37] R.T. Rockafellar, and R.J.B. Wets, *Variational Analysis*, Springer, 1998.
[38] M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, in: *MOS-SIAM Series on Optimization*, SIAM, 2011.
[39] L. Wang, L.H. Zhang, Z. Bai, and R.C. Li, Orthogonal canonical correlation analysis and applications, *Optim. Methods Softw.*, **35** (2020), 787–807.
[40] Q. Wang, and W.H. Yang, Proximal quasi-Newton method for composite optimization over the Stiefel manifold, *J. Sci. Comput.*, **95** (2023), https://doi.org/10.1007/s10915-023-02165-x.
[41] R.B. Wilson, *A Simplicial Algorithm for Concave Programming*, PhD Thesis, Graduate School of Business Administration, Harvard University, 1963.
[42] S.J. Wright, R.D. Nowak, and M.A. Figueiredo, Sparse reconstruction by separable approximation, *IEEE Trans. Signal Process.*, **57** (2009), 2479–2493.
[43] X. Xiao, Y. Li, Z. Wen, and L.W. Zhang, A regularized semi-smooth Newton method with projection steps for composite convex programs, *J. Sci. Comput.*, **76** (2018), 364–389.
[44] C. Zhang, X. Chen, and S. Ma, A Riemannian smoothing steepest descent method for non-Lipschitz optimization on submanifolds, *arXiv:2104.04199*, 2021.
[45] L.H. Zhang, X. Ma, and C. Shen, A structure-exploiting nested Lanczos-type iteration for the

multi-view canonical correlation analysis, *SIAM J. Sci. Comput.*, **43** (2021), A2685–A2713.
[46] Y. Zhou, C. Bao, C. Ding, and J. Zhu, A semismooth Newton based augmented Lagrangian method for nonsmooth optimization on matrix manifolds, *Math. Program.*, **201** (2023), 1–61.