New subspace minimization conjugate gradient methods based on regularization model for unconstrained optimization

Ting Zhao¹ · Hongwei Liu¹ · Zexian Liu²,³

Received: 4 March 2019 / Accepted: 20 September 2020 / Published online: 28 October 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract
In this paper, two new subspace minimization conjugate gradient methods based on $p$-regularization models are proposed, where a special scaled norm in $p$-regularization model is analyzed. Different choices of special scaled norm lead to different solutions to the $p$-regularized subproblem. Based on the analyses of the solutions in a two-dimensional subspace, we derive new directions satisfying the sufficient descent condition. With a modified nonmonotone line search, we establish the global convergence of the proposed methods under mild assumptions. $R$-linear convergence of the proposed methods is also analyzed. Numerical results show that, for the CUTEr library, the proposed methods are superior to four conjugate gradient methods, which were proposed by Hager and Zhang (SIAM J. Optim. 16(1):170–192, 2005), Dai and Kou (SIAM J. Optim. 23(1):296–320, 2013), Liu and Liu (J. Optim. Theory. Appl. 180(3):879–906, 2019) and Li et al. (Comput. Appl. Math. 38(1):2019), respectively.

Keywords Conjugate gradient method · $p$-regularization model · Subspace technique · Nonmonotone line search · Unconstrained optimization

Mathematics Subject Classification (2010) 90C30 · 90C06 · 65K05
1 Introduction

Conjugate gradient (CG) methods are of great importance for solving the large-scale unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x),$$

where \( f : \mathbb{R}^n \to \mathbb{R} \) is a continuously differentiable function. The key features of CG methods are that they do not require matrix storage. The iterations \( \{x_n\} \) satisfy the iterative form

$$x_{k+1} = x_k + \alpha_k d_k,$$

where \( \alpha_k \) is the stepsize and \( d_k \) is the search direction defined by

$$d_k = \begin{cases} -g_k, & \text{if } k = 0, \\ -g_k + \beta_k d_{k-1}, & \text{if } k > 0, \end{cases}$$

where \( g_k = \nabla f(x_k) \), \( f(x_k) = f_k \) and \( \beta_k \in \mathbb{R} \) is called the CG parameter.

For general nonlinear functions, various choices of \( \beta_k \) cause different CG methods. Some well-known options for \( \beta_k \) are called FR [19], HS [27], PRP [39], DY [13] and HZ [24], and are given by

\[
\beta_{\text{FR}}^k = \frac{\|g_{k+1}\|^2}{\|g_k\|^2}, \quad \beta_{\text{HS}}^k = \frac{g_k^T y_k}{\|g_k\|^2}, \quad \beta_{\text{PRP}}^k = \frac{g_{k+1}^T y_k}{\|g_k\|^2}, \quad \beta_{\text{DY}}^k = \frac{\|g_{k+1}\|^2}{d_k^T y_k},
\]

and

\[
\beta_{\text{HZ}}^k = \frac{1}{d_k^T y_k} \left( y_k - 2d_k \frac{\|y_k\|^2}{d_k^T y_k} \right)^T g_{k+1},
\]

where \( y_k = g_{k+1} - g_k \) and \( \|\cdot\| \) denotes the Euclidean norm. Recently, other efficient CG methods have been proposed by different ideas, which can be seen in [12, 18, 24, 25, 28, 40, 41, 48].

With the increasing scale of optimization problems, subspace methods have become a class of very efficient numerical methods, because it is not necessary to solve large-scale subproblems at each iteration [49]. Yuan and Stoer [47] first put forward the subspace minimization conjugate gradient (SMCG) method, the search direction of which is computed by solving the following problem:

$$\min_{d \in \Omega_{k+1}} m_{k+1}(d) = g_{k+1}^T d + \frac{1}{2} d^T B_{k+1} d,$$

where \( \Omega_{k+1} = \{g_{k+1}, s_k\} \) and the direction \( d \) is given by

$$d = \mu g_{k+1} + v s_k,$$

where \( B_{k+1} \) is an approximation to Hessian matrix, \( \mu \) and \( v \) are parameters and \( s_k = x_{k+1} - x_k \). The detailed information about subspace technique can be referred to [1, 26, 29, 44, 50]. The SMCG method can be considered as a generalization of CG method and it reduces to the linear CG method when it uses the exact line search condition and objective function is convex quadratic function. Based on the analysis of the SMCG method, Dai and Kou [15] made a theoretical analysis of the BBCG, obtained by combining the Barzilai-Borwein (BB) idea [3] with the SMCG.
Liu and Liu [31] presented an efficient Barzilai-Borwein conjugate gradient method (SMCG_BB) with the generalized Wolfe line search for unconstrained optimization. Li, Liu and Liu [30] deliver a subspace minimization conjugate gradient method based on conic model for unconstrained optimization (SMCG_Conic).

Generally, the iterative methods are often based on a quadratic model because the quadratic model can approximate the objective function well at a small neighborhood of the minimizer. However, when iterative point is far from the minimizer, the quadratic model might not work well if the objective function possess high nonlinearity [42, 46]. In theory, the successive gradients generated by the conjugate gradient method applied to a quadratic function and using exact line search should be orthogonal. However, for some ill-conditioned problems, orthogonality is quickly lost due to the rounding errors, and the convergence is much slower than expected [26]. There are many methods to deal with ill-conditioned problems, among which regularization method is one of the effective methods. Recently, $p$-regularized subproblem plays an important role in more regularization approaches [10, 23, 36] and some $p$-regularization algorithms for unconstrained optimization enjoy a growing interest [4, 7, 10, 11]. Its idea is to incorporate a local quadratic approximation of the objective function with a weighted regularization term $(\sigma_k/p)\|x\|^p$, $p > 2$, and then globally minimize it at each iteration. Interestingly, Cartis et al. [10, 11] proved that, under suitable assumptions, $p$-regularization algorithmic scheme is able to achieve superlinear convergence. The most common choice to regularize the quadratic approximation is $p$-regularization with $p = 3$, which is known as the cubic regularization, since functions of this form are used as local models models (to be minimized) in many algorithmic frameworks for unconstrained optimization [5–11, 17, 21, 23, 36, 38, 43]. The cubic regularization was first introduced by Griewank [23] and was later considered by many authors of global convergence and complexity analysis (see [11, 36, 43]).

Recently, how to obtain the approximate solution of the $p$-regularized subproblem has become a hot research topic. Practical approaches to get an approximate solution are proposed in [7, 22], where the solution of the secular equation is typically approximated over specific evolving subspaces using Krylov methods. The main drawback of such approaches is the large amount of calculation, because they may need to solve multiple linear systems in turn.

In this paper, motivated by [2] and [45], the $p$-regularization with a special scaled norm is analyzed and the solutions of the new $p$-regularization that arise in unconstrained optimization are considered. Based on [2] we propose a method to solve it by using a special scaled norm in the $p$-regularized subproblem. According to the advantages of the new $p$-regularization method with SMCG method, we propose two new subspace minimization conjugate gradient methods. In our algorithms, if the objective function is close to a quadratic, we use a quadratic approximation model in a two-dimensional subspace to generate the direction; otherwise, $p$-regularization model is considered. We prove that the search direction possesses the sufficient descent property and the proposed methods satisfy the global convergence under mild conditions. We present some numerical results, which show that the proposed methods are very promising.
The remainder of this paper is organized as follows. In Section 2, we will state the form of $p$-regularized subproblem and provide how to solve the $p$-regularization problem based on the special $p$-regularization model. Four choices of search direction by minimizing the approximate models, including $p$-regularization and quadratic model on certain subspace, are presented in Section 3. In Section 4, we describe two algorithms and discuss some important properties of the search direction in detail. In Section 5, we establish the convergence of the proposed methods under mild conditions. Some performances of the proposed methods are reported on Section 6. Conclusions and discussions are presented in the last section.

2 The $p$-regularized subproblem

In this section, we will briefly introduce several forms of the $p$-regularized subproblem by using a special scaled norm and provide the solutions to the resulting problems in the whole space and the two-dimensional subspace, respectively. The chosen scaled norm is of the form $\|x\|_A = \sqrt{x^T Ax}$, where $A$ is a symmetric positive definite matrix. After analysis, we will mainly consider two special cases: (I) $A$ is the Hessian matrix. In this case, the $p$-regularized subproblem has a unique solution; (II) $A$ is the identity matrix. In this case, the $p$-regularized subproblem is the same as the general form.

2.1 The form in the whole space

The general form of the $p$-regularized subproblem is:

$$\min_{x \in \mathbb{R}^n} h(x) = c^T x + \frac{1}{2} x^T H x + \frac{\sigma}{p} \|x\|^p,$$

where $p > 2$, $c \in \mathbb{R}^n$, $\sigma > 0$ and $H \in \mathbb{R}^{n \times n}$ is a symmetric matrix.

As for how to solve the above problem, the following theorem is given.

Theorem 2.1 ([45], Thm.1.1) The point $x^*$ is a global minimizer of (6) if and only if

$$(H + \sigma \|x^*\|^{p-2} I)x^* = -c, \quad H + \sigma \|x^*\|^{p-2} I \succeq 0. \quad (7)$$

Moreover, the $l_2$ norms of all the global minimizers are equal.

Now, we give a modified $p$-regularized subproblem with a special scaled norm:

$$\min_{x \in \mathbb{R}^n} h(x) = c^T x + \frac{1}{2} x^T H x + \frac{\sigma}{p} \|x\|_A^p,$$

where $A \in \mathbb{R}^{n \times n}$ is a symmetric positive definite matrix.

By setting $y = A^{\frac{1}{2}} x$, (8) can be arranged as follows:

$$\min_{y \in \mathbb{R}^n} h(y) = (A^{\frac{1}{2}} c)^T y + \frac{1}{2} y^T A^{\frac{1}{2}} H A^{\frac{1}{2}} y + \frac{\sigma}{p} \|y\|^p. \quad (9)$$
According to Theorem 2.1, we know that the point \( y^* \) is a global minimizer of (9) if and only if

\[
(A^{-\frac{1}{2}}HA^{-\frac{1}{2}} + \sigma \|y^*\|^{p-2}I)y^* = -A^{-\frac{1}{2}}c, \tag{10}
\]

\[
A^{-\frac{1}{2}}HA^{-\frac{1}{2}} + \sigma \|y^*\|^{p-2}I \succeq 0. \tag{11}
\]

Let \( V \in \mathbb{R}^{n \times n} \) be an orthogonal matrix such that

\[
V^T (A^{-\frac{1}{2}}HA^{-\frac{1}{2}}) V = Q,
\]

where \( Q = \text{diag} \{ \mu_i \} \) and \( \mu_1 \leq \mu_2 \leq \cdots \leq \mu_n \) are the eigenvalues of \( A^{-\frac{1}{2}}HA^{-\frac{1}{2}} \). Now we introduce the vector \( a \in \mathbb{R}^n \) such that

\[
y = Va. \tag{12}
\]

Denote \( z = \|y\| \) and pre-multiplying (10) by \( V^T \), we get

\[
(Q + \sigma z^{p-2} I)a = -\beta, \tag{13}
\]

where \( \beta = V^T (A^{-\frac{1}{2}} c) \).

The expression (13) can be equivalently written as

\[
a_i = \frac{-\beta_i}{\mu_i + \sigma z^{p-2}}, \quad i = 1, 2, \ldots, n,
\]

where \( a_i \) and \( \beta_i \) are the components of vectors \( a \) and \( \beta \), respectively. By the way, if \( \mu_i + \sigma z^{p-2} = 0 \), for any \( i \), we obtain \( \beta = 0 \) from (13).

From (12), we have an equation about \( z \):

\[
z^2 = y^T y = a^T a = \sum_{i=1}^{n} \frac{\beta_i^2}{(\mu_i + \sigma z^{p-2})^2}. \tag{14}
\]

Denote

\[
\phi(z) = \sum_{i=1}^{n} \frac{\beta_i^2}{(\mu_i + \sigma z^{p-2})^2} - z^2.
\]

We can easily obtain

\[
\phi'(z) = \sum_{i=1}^{n} \frac{-2\sigma (p-2) \beta_i^2 z^{p-3} (\mu_i + \sigma z^{p-2})}{(\mu_i + \sigma z^{p-2})^4} - 2z.
\]

It follows from \( p > 2, \ z > 0 \) and \( \sigma > 0 \) that \( \phi'(z) < 0 \), which indicates that \( \phi(z) \) is monotonically decreasing in the interval \( [0, +\infty) \). Moreover, we can observe that \( \phi(0) > 0 \), when \( \beta \neq 0 \), and \( \lim_{z \to \infty} \phi(z) = -\infty \). So, there exists a unique positive solution to (14) when \( \beta \neq 0 \). On the other hand, if \( \beta = 0 \), \( z = 0 \) is the only solution of (14), which means \( x^* = 0 \) is the only global minimizer of (8).

Based on the above derivation and analysis, we can get the following theorem.

**Theorem 2.2** The point \( x^* \) is a global minimizer of (8) if and only if

\[
(H + \sigma (z^*)^{p-2} A)x^* = -c, \tag{15}
\]

\[
H + \sigma (z^*)^{p-2} A \succeq 0, \tag{16}
\]
where $z^*$ is the unique non-negative root of the equation

$$z^2 = \sum_{i=1}^{n} \frac{\beta_i^2}{(\mu_i + \sigma z^{p-2})^2}. \quad (17)$$

Moreover, the $l_A$ norms of all the global minimizers are equal.

Now, let us consider a special case that $H > 0$ and $A = H$. It is clear that $H + \sigma z^{p-2} H$ is always a positive definite matrix since $\sigma > 0$ and $z \geq 0$. So, the global minimizer of (8) is unique.

**Inference 2.3** Let $H > 0$, $A = H$, then the point $x^* = \frac{-1}{1 + \sigma (z^*)^{p-2}} H^{-1} c$ is the only global minimizer of (8) and $z^*$ is a unique non-negative solution to the equation

$$\sigma z^{p-1} + z - \sqrt{c^T H^{-1} c} = 0. \quad (18)$$

**Remark 1**

i) $c = 0$. It is obvious that the equation (18) becomes

$$\sigma z^{p-1} + z = 0,$$

that is

$$z \left( \sigma z^{p-2} + 1 \right) = 0.$$

From $\sigma > 0$, we know $z^* = 0$ is a unique non-negative solution to the equation (18).

ii) $c \neq 0$. Denote

$$\psi(z) = \sigma z^{p-1} + z - \sqrt{c^T H^{-1} c}. \quad (19)$$

We can easily obtain

$$\psi'(z) = \sigma (p-1) z^{p-2} + 1 > 0,$$

which indicates that the $\psi(z)$ is monotonically increasing. From $\psi(0) < 0$ and $\psi(\sqrt{c^T H^{-1} c}) > 0$, we know that $z^*$ is a unique positive solution to the equation (18).

### 2.2 The form in the two-dimensional space

Let $g$ and $s$ be two linearly independent vectors. Denote $\Omega = \{d | d = \mu g + \nu s, \mu, \nu \in \mathbb{R}\}$. In this part, we suppose that $H$ is symmetric and positive definite and $y = Hs$.

We consider the following problem

$$\min_{d \in \Omega} h(d) = c^T d + \frac{1}{2} d^T H d + \frac{\sigma}{p} \|d\|_A^p. \quad (20)$$

Obviously, when $A = H$, problem (20) can be translated into

$$\min_{\mu, \nu \in \mathbb{R}} \left( g^T c \right)^T \left( \begin{array}{c} \mu \\ \nu \end{array} \right) + \frac{1}{2} \left( \begin{array}{c} \mu \\ \nu \end{array} \right)^T B \left( \begin{array}{c} \mu \\ \nu \end{array} \right) + \frac{\sigma}{p} \left\| \begin{array}{c} \mu \\ \nu \end{array} \right\|_B^p. \quad (21)$$
where \( \rho = g^T H g \), and \( B = \left( \begin{array}{c} \rho g^T y \\ g^T y y^T s \end{array} \right) \) is a symmetric and positive definite matrix, since the \( H \) is a symmetric positive definite matrix and the two vectors \( g \) and \( s \) are linearly independent.

By the Inference 2.3, we can obtain the unique solution of (21):

\[
\begin{pmatrix} \mu^* \\ v^* \end{pmatrix} = \frac{-1}{1 + \sigma(z^*)^{p-2}} B^{-1} \begin{pmatrix} g^T c \\ s^T c \end{pmatrix},
\]

(22)

where \( z^* \) is a unique non-negative solution to \( \sigma z^{p-1} + z - \sqrt{\left( \begin{array}{c} g^T c \\ s^T c \end{array} \right)^T B^{-1} \left( \begin{array}{c} g^T c \\ s^T c \end{array} \right)} = 0 \).

When \( A = I \), we obtain from (20) that

\[
\min_{\mu, v \in \mathbb{R}} \left( g^T \frac{c}{s^T c} \right)^T \begin{pmatrix} \mu \\ v \end{pmatrix} + \frac{1}{2} \begin{pmatrix} \mu \\ v \end{pmatrix}^T B \begin{pmatrix} \mu \\ v \end{pmatrix} + \frac{\sigma}{p} \left\| \begin{pmatrix} \mu \\ v \end{pmatrix} \right\|_E^p,
\]

(23)

where \( E = \left( \begin{array}{cc} \|g\|^2 & g^T s \\ g^T s & \|s\|^2 \end{array} \right) \) is positive definite due to the linear independence of vectors \( g \) and \( s \).

By the Theorem 2.2, we can gain the unique solution to (23):

\[
\begin{pmatrix} \mu^* \\ v^* \end{pmatrix} = -\left( B + \sigma(z^*)^{p-2} E \right)^{-1} \begin{pmatrix} g^T c \\ s^T c \end{pmatrix},
\]

(24)

where \( z^* \) is a unique non-negative solution to (17) in which \( 0 < \mu_1 \leq \mu_2 \) are the eigenvalues of \( E^{-\frac{1}{2}} B E^{-\frac{1}{2}} \), \( \beta = V^T \left( E^{-\frac{1}{2}} \left( \begin{array}{c} g^T c \\ s^T c \end{array} \right) \right) \).

3 The search direction and the initial stepsize

In this section, based on the different choices of special scaled norm, we derive two new directions by minimizing the two \( p \)-regularization models of the objective function on the subspace \( \Omega_k = \text{span} \{ g_k, s_k \} \). The selection criteria for how to choose the initial stepsize is also given. For the rest, we assume that \( s_k^T y_k > 0 \) guaranteed by the Wolfe condition (49).

3.1 Derivation of the new search direction

The parameter \( t_k \) by Yuan [48] is used to describe how \( f(x) \) is close to a quadratic function on the line segment between \( x_{k-1} \) and \( x_k \), and is defined by

\[
t_k = \left| \frac{2 \left( f_{k-1} - f_k + g_k^T s_{k-1} \right)}{s_k^T y_{k-1}} - 1 \right|.
\]

(25)
On the other hand, the ratio
\[ \theta_k = \frac{f_{k-1} - f_k}{0.5s_{k-1}^T y_{k-1} - g_k^T s_{k-1}} \] (26)
shows difference between the actual reduction and the predicted reduction for the quadratic model.

If the following condition [33] holds, namely,
\[ t_k \leq c_1 \text{ or } (t_k \leq c_2 \text{ and } t_{k-1} \leq c_2) \] (27)
or
\[ |\theta_k - 1| < \gamma, \] (28)
where \( c_1, c_2 \) and \( \gamma \) are small positive constants, then \( f(x) \) might be very close to a quadratic on the line segment between \( x_{k-1} \) and \( x_k \). We choose the quadratic model.

Moreover, if the conditions [30]
\[ (s_k^T y_k)^2 \leq 10^{-5} ||s_k||^2 ||y_k||^2 \text{ and } (f_{k+1} - f_k - 0.5(g_k^T s_k + g_{k+1}^T s_k))^2 \leq 10^{-6} ||s_k||^2 ||y_k||^2 \] (29)
hold, then the problem might be ill-conditioned, and the current iterative point is far away from the minimizer of problem. At this point, the information might be inaccurate, then we also choose the quadratic model to derive the search direction.

General iterative methods, which are often based on a quadratic model, have been quite successful in solving unconstrained optimization problems, since the quadratic model can approximate the objective function \( f(x) \) well at a small neighborhood of \( x_k \) in many cases. Consequently, when the condition (27), (28) or (29) holds, the quadratic approximation model (4) is preferable. However, when the conditions (27), (28) and (29) do not hold, the iterative point is far away from the minimizer, the quadratic model may not very well approximate the original problem. Thus in this case, we select the \( p \)-regularization model which could include more useful information on the objective function to approximate the original problem.

For general functions, if the condition
\[ \xi_1 \leq \frac{s_{k-1}^T y_{k-1}}{||s_{k-1}||^2} \leq \frac{||y_{k-1}||^2}{s_{k-1}^T y_{k-1}} \leq \xi_2 \] (30)
holds, where \( \xi_1 \) and \( \xi_2 \) are positive constants, then the condition number of the Hessian matrix might be not very large. In this case, we consider the quadratic approximation model or the \( p \)-regularization model.

Now we divide the analysis in the following four cases to derive the search direction.

**Case 1** When the condition (30) holds and any of the conditions (27), (28), (29) do not hold, we consider the following \( p \)-regularized subproblem
\[ \min_{d_k \in \Omega_k} m_k(d_k) = d_k^T g_k + \frac{1}{2} d_k^T H_k d_k + \frac{1}{p} \sigma_k \|d_k\|_A^p, \] (31)
where $H_k$ is a symmetric and positive definite approximation to Hessian matrix satisfying the equation $H_k s_{k-1} = y_{k-1}$, $A_k$ is a symmetric positive definite matrix, $\sigma_k$ is a dynamic non-negative regularization parameter and $\Omega_k = \text{span}\{g_k, s_{k-1}\}$.

Denote

$$d_k = \mu_k g_k + \nu_k s_{k-1}, \quad (32)$$

where $\mu_k$ and $\nu_k$ are parameters to be determined.

In the following, we will discuss that $A_k = H_k$ and $A_k = I$ in two parts.

(I) $A_k = H_k$

It is easy to see the problem (31) is similar to the problem (21), so we obtain

$$\min_{\mu_k, \nu_k \in \mathbb{R}} \left( \frac{\|g_k\|^2}{g_k^T s_{k-1}} \right)^T \left( \begin{array}{c} \mu_k \\ \nu_k \end{array} \right) + \frac{1}{2} \left( \begin{array}{c} \mu_k \\ \nu_k \end{array} \right)^T B_k \left( \begin{array}{c} \mu_k \\ \nu_k \end{array} \right) + \frac{\sigma_k}{p} \left\| \left( \begin{array}{c} \mu_k \\ \nu_k \end{array} \right) \right\|_B^p, \quad (33)$$

where $\rho_k \approx g_k^T H_k g_k$ and $B_k = \left( \begin{array}{cc} \rho_k & g_k^T y_{k-1} \\ g_k^T y_{k-1} & s_{k-1}^T y_{k-1} \end{array} \right)$.

The choice of the two parameters $\rho_k$ and $\sigma_k$ in (33) is very important.

Motivated by the Barzilai-Borwein method, Dai and Kou [15] proposed the BBCG3 method with the very efficient parameter $\rho_{k,BBCG3}^k = \frac{3}{2} \frac{\|y_{k-1}\|^2}{s_{k-1}^T y_{k-1}} \|g_k\|^2$ and considered it a good estimation of the $g_k^T H_k g_k$. So in this paper, we choose $\rho_k = \rho_{k,BBCG3}^k$ in the above function that will make $B_k$ positive definite, which guarantees the uniqueness of the solution to (33).

There are many ways [10, 21] to get the value of $\sigma_k$, and the interpolation condition is one of them. Here, we use the interpolation condition to get it. By imposing the following interpolation condition:

$$f_{k-1} = f_k - g_k^T s_{k-1} + \frac{1}{2} s_{k-1}^T y_{k-1} + \frac{\sigma_k}{p} (s_{k-1}^T y_{k-1})^p,$$

we obtain

$$\sigma_k = \frac{p(f_{k-1} - f_k + g_k^T s_{k-1} - \frac{1}{2} s_{k-1}^T y_{k-1})}{(s_{k-1}^T y_{k-1})^p}.$$

In order to ensure that $\sigma_k \geq 0$, we set

$$\sigma_k = \frac{p \left| f_{k-1} - f_k + g_k^T s_{k-1} - \frac{1}{2} s_{k-1}^T y_{k-1} \right|}{(s_{k-1}^T y_{k-1})^p}.$$

From (22), we can get the unique solution to (33):

$$\mu_k = \frac{1}{(1 + \sigma_k(z^*)^p - 2)} \Delta_k \left( g_k^T y_{k-1} g_k^T s_{k-1} - s_{k-1}^T y_{k-1} \|g_k\|^2 \right), \quad (34)$$

$$\nu_k = \frac{1}{(1 + \sigma_k(z^*)^p - 2)} \Delta_k \left( g_k^T y_{k-1} \|g_k\|^2 - \rho_k g_k^T s_{k-1} \right), \quad (35)$$
where $\Delta_k = \begin{vmatrix} \rho_k & g_k^T y_{k-1} \\ g_k^T y_{k-1} & s_{k-1}^T y_{k-1} \end{vmatrix} = \rho_k s_{k-1}^T y_{k-1} - (g_k^T y_{k-1})^2 > 0$ and $z^*$ is the unique positive solution to

$$\sigma_k z^{p-1} + z - \sqrt{\left( \frac{\|g_k\|^2}{g_k^T s_{k-1}} \right)^T B_{k-1}^{-1} \left( \frac{\|g_k\|^2}{g_k^T s_{k-1}} \right)} = 0. \quad (36)$$

We denote $\tilde{q} = \sqrt{\left( \frac{\|g_k\|^2}{g_k^T s_{k-1}} \right)^T B_{k-1}^{-1} \left( \frac{\|g_k\|^2}{g_k^T s_{k-1}} \right)}$. Substituting $\tilde{q}$ into (36), we get

$$\sigma_k z^{p-1} + z - \tilde{q} = 0. \quad (37)$$

Since it is difficult to obtain the exact root of (37) when $p$ is large, we only consider $p = 3$ and $p = 4$ for simplicity.

(i) $p = 3$. It is not difficult to know the unique positive solution to (37)

$$z^* = \frac{2\tilde{q}}{1 + \sqrt{1 + 4\sigma_k \tilde{q}}}. \quad (38)$$

(ii) $p = 4$. According to Cardano’s formula for the roots on cubic equation and $z > 0$, the unique positive solution to (37) can be obtained

$$z^* = \sqrt[3]{\frac{\tilde{q}}{2\sigma_k}} + \sqrt[3]{\frac{\tilde{q}^2}{4\sigma_k^2} + \left( \frac{1}{3\sigma_k} \right)^3} + \sqrt[3]{\frac{\tilde{q}}{2\sigma_k} - \sqrt[3]{\frac{\tilde{q}^2}{4\sigma_k^2} + \left( \frac{1}{3\sigma_k} \right)^3}}. \quad (39)$$

For ensuring the sufficient descent condition of the direction produced by (34) and (35), if $\sigma_k (z^*)^{p-2} > 1$, we set $\sigma_k (z^*)^{p-2} = 1$, where $z^*$ is determined by (38) or (39).

(II) $A_k = I$

Based on the analysis of (I), we can get the following problem similarly:

$$\min_{\mu_k, \nu_k \in \mathbb{R}} \left( \frac{\|g_k\|^2}{g_k^T s_{k-1}} \right)^T \begin{pmatrix} \mu_k \\ \nu_k \end{pmatrix} + \frac{1}{2} \begin{pmatrix} \mu_k \\ \nu_k \end{pmatrix}^T B_k \begin{pmatrix} \mu_k \\ \nu_k \end{pmatrix} + \sigma_k \frac{\|\mu_k\|}{p} \left\| \begin{pmatrix} \mu_k \\ \nu_k \end{pmatrix} \right\|_{E_k}^p, \quad (40)$$

where $E_k = \left( \frac{\|g_k\|^2}{g_k^T s_{k-1}} \right) s_{k-1}^T s_{k-1}$ and $\rho_k, B_k$ are the same as those in problem (33).

Similarly, we still use the interpolation condition to determine $\sigma_k$. By

$$f_{k-1} = f_k - g_k^T s_{k-1} y_{k-1} + \frac{1}{2} s_{k-1}^T y_{k-1} + \frac{\sigma_k}{p} \left\| s_{k-1} \right\|_{E_k}^p,$$

we get

$$\sigma_k = \frac{p \left\| f_{k-1} - f_k + g_k^T s_{k-1} - \frac{1}{2} s_{k-1}^T y_{k-1} \right\|_{E_k}^p}{\left\| s_{k-1} \right\|_{E_k}^p}. \quad (41)$$

According to (24), the unique solution to (40) can be obtained:

$$\hat{\mu}_k = \frac{1}{\Delta_k} \left( g_k^T y_{k-1} g_k^T s_{k-1} - s_{k-1}^T y_{k-1} g_k \right)^2 + \lambda \left( g_k^T s_{k-1} \right)^2 - \lambda \left\| s_{k-1} \right\|_{E_k}^2 \left\| g_k \right\|^2, \quad (41)$$

$$\hat{\nu}_k = \frac{1}{\Delta_k} \left( g_k^T y_{k-1} \left\| g_k \right\|^2 - \rho_k g_k^T s_{k-1} \right), \quad (42)$$
where
\[
\Delta_k = (\rho_k + \lambda \|g_k\|^2)(s_{k-1}^Ty_{k-1} + \lambda \|s_{k-1}\|^2) - (s_k^Ty_{k-1} + \lambda s_k^Tg_{k}s_{k-1})^2,
\]
(43)
and $z^*$ satisfies the equation (17), which can be solved by tangent method [37]. For ensuring the sufficient descent of the direction produced by (41) and (42), if
\[
\sigma_k(z^*) > \frac{\|y_{k-1}\|^2}{s_{k-1}^Ty_{k-1}},
\]
we set $\lambda = \frac{\|y_{k-1}\|^2}{s_{k-1}^Ty_{k-1}}$.

Remark 2 It is worth emphasizing that in the process of finding the direction, $(\|g_k\|^2) = (g_k^Tg_{k-1})$ is equivalent to the problem (8) which $c \neq 0$.

Case 2 When the condition (30) holds and at least one of the conditions (27), (28), (29) holds, we choose the quadratic model which corresponds to (33) with $\sigma_k = 0$. So the parameters in (32) are generated by solving (34) and (35) with $\sigma_k = 0$:
\[
\bar{\mu}_k = \frac{1}{\Delta_k}(g_k^Ty_{k-1}g_k^Ts_{k-1} - s_{k-1}^Ty_{k-1} \|g_k\|^2),
\]
(44)
\[
\bar{v}_k = \frac{1}{\Delta_k}(g_k^Ty_{k-1} \|g_k\|^2 - \rho_kg_k^Ts_{k-1}).
\]
(45)

Case 3 If the exact line search is used, the direction in Case 2 is parallel to the HS direction with convex quadratic functions. It is known that the conjugate condition, namely, $d_{k+1}^Ty_k = 0$, still holds whether the line search is exact or not for HS conjugate gradient method.

If the condition (30) does not hold and the conditions
\[
\frac{|g_k^Ty_{k-1}g_k^Ts_{k-1}|}{s_{k-1}^Ty_{k-1} \|g_k\|^2} \leq \xi_3 \quad \text{and} \quad \xi_1 \leq \frac{s_{k-1}^Ty_{k-1}}{\|s_{k-1}\|^2}
\]
(46)
hold, where $0 \leq \xi_3 \leq 1$, then $\bar{\mu}_k$ in Case 2 is close to 0, then we use the HS conjugate gradient direction. Besides, with the finite-termination property of the HS method for exact convex quadratic programming, such choice of the direction might lead to a rapid convergence rate of our algorithm.

Case 4 If the condition (30) does not hold and the condition (46) does not hold, then we choose the negative gradient as the search direction, namely,
\[
d_k = -g_k.
\]
(47)

In conclusion, the new search direction can be stated as
\[
d_k = \begin{cases} 
\begin{aligned}
\bar{\mu}_kg_k + \bar{v}_ks_{k-1}, & \text{if (30) holds and any of (27), (28), (29) do not hold}, \\
\bar{\mu}_kg_k + \bar{v}_ks_{k-1}, & \text{if (30) holds and at least one of (27), (28), (29) holds}, \\
-g_k + \beta^H_kd_{k-1}, & \text{if (30) does not hold and (46) holds}, \\
-g_k, & \text{if (30) does not hold and (46) does not hold}.
\end{aligned}
\end{cases}
\]
where $\mu_k, \nu_k$ are given by (34), (35) or (41), (42) and $\tilde{\mu}_k, \tilde{\nu}_k$ are given by (44), (45), respectively.

### 3.2 Choices of the initial stepsize and the Wolfe line search

It is universally acknowledged that the choice of the initial stepsize and the Wolfe line search are of great importance for an optimization method. In this section, we introduce a strategy to choose the initial stepsize and develop a modified nonmonotone Wolfe line search.

#### 3.2.1 Choices of the initial stepsize

Denote $\phi_k(\alpha) = f(x_k + \alpha d_k), \alpha \geq 0$.

(i) The initial stepsize for the search directions in Case 1–Case 3 in Section 3.1.

Similar to [31], we choose the initial stepsize as

$$
\alpha_0^k = \begin{cases} 
\hat{\alpha}_k, & \text{if (27) holds and } \tilde{\alpha}_k > 0, \\
1, & \text{otherwise},
\end{cases}
$$

where

$$
\tilde{\alpha}_k = \min_q(\phi_k(0), \phi_k'(0), \phi_k(1)), \\
\hat{\alpha}_k = \min\{\max\{\tilde{\alpha}_k, \lambda_{\min}\}, \lambda_{\max}\} \text{ and } \lambda_{\max} > \lambda_{\min} > 0.
$$

In the above formula, $q(\phi_k(0), \phi_k'(0), \phi_k(1))$ denotes the interpolation function for the three values $\phi_k(0), \phi_k'(0), \phi_k(1),$ and $\lambda_{\max}$ and $\lambda_{\min}$ represent two positive parameters.

(ii) The initial stepsize for the negative gradient direction (47).

As we all know, the gradient method with the adaptive BB stepsize [52] is very efficient for strictly convex quadratic minimization, especially when the condition number is large. In this paper we choose the strategy in [31]:

$$
\alpha_0^k = \begin{cases} 
\min\{\max\{\tilde{\alpha}_k, \lambda_{\min}\}, \lambda_{\max}\}, & \text{if (27) holds, } d_{k-1} \neq -g_{k-1}, \|g_k\|^2 \leq 1 \text{ and } \tilde{\alpha}_k > 0, \\
\tilde{\alpha}_k, & \text{otherwise},
\end{cases}
$$

where

$$
\tilde{\alpha}_k = \begin{cases} 
\min\{\lambda_k \alpha_{BB}^k, \lambda_{\max}\}, & \text{if } g_k^T s_{k-1} > 0, \text{ and } \tilde{\alpha}_k > 0, \\
\min\{\lambda_k \alpha_{BB}^k, \lambda_{\max}\}, & \text{if } g_k^T s_{k-1} \leq 0, \text{ and } \tilde{\alpha}_k = \min q(\phi_k(0), \phi_k'(0), \phi_k(\hat{\alpha}_k)).
\end{cases}
$$

$\lambda_k$ is a scaling parameter given by $\lambda_k = \begin{cases} 
0.999, & \text{if } n > 10 \text{ and } \text{Numgra} > 12, \\
1, & \text{otherwise},
\end{cases}$

where Numgra denotes the number of the successive use of the negative gradient direction.
3.2.2 Choice of the Wolfe line search

The line search is an important factor for the overall efficiency of most optimization algorithms. In this paper, we pay attention to the nonmonotone line search proposed by Zhang and Hager [51] (ZH line search)

\[ f(x_k + \alpha_k d_k) \leq C_k + \delta \alpha_k \nabla f(x_k)^T d_k, \]  \hfill (48)

\[ \nabla f(x_k + \alpha_k d_k)^T d_k \geq \sigma \nabla f(x_k)^T d_k, \]  \hfill (49)

where \(0 < \delta < \sigma < 1\), \(C_0 = f_0\), \(Q_0 = 1\), and \(C_k\) and \(Q_k\) are updated by

\[ Q_{k+1} = \eta_k Q_k + 1, C_{k+1} = \frac{\eta_k Q_k C_k + f(x_{k+1})}{Q_{k+1}}, \]  \hfill (50)

where \(\eta_k \in [0, 1]\).

It is worth mentioning that some improvements have been made to ZH line search to find a more suitable stepsize and obtain a better convergence result. Specially, \(C_1 = \min\{C_0, f_1 + 1.0\}\), \(Q_1 = 2.0\), \hfill (51)

when \(k \geq 1\), \(C_{k+1}\) and \(Q_{k+1}\) are updated by (50), where \(\eta_k\) is taken as

\[ \eta_k = \begin{cases} \eta, & \text{if } \text{mod} (k, l) = 0, \\ 1, & \text{if } \text{mod} (k, l) \neq 0, \end{cases} \]  \hfill (52)

where \(l = \max(20, n)\), \(\text{mod}(k, l)\) denotes the remainder for \(k\) modulo \(l\) and \(\eta = 0.7\) when \(C_k - f_{k+1} > 0.999 |C_k|\), otherwise \(\eta = 0.999\). Such choice of \(\eta_k\) can be used to control nonmonotonicity dynamically, referred to [32].

4 Algorithms

In this section, according to the different choices of special scaled norm, we will introduce two new subspace minimization conjugate gradient algorithms based on the \(p\)-regularization, and analyze some theoretical properties of the direction \(d_k\).

Denote

\[ r_{k-1} = \left| \frac{f_k}{f_{k-1} + 0.5(g^T_{k-1}s_{k-1} + g^T_k s_k - 1)} - 1 \right|, \]

\[ \tilde{r}_{k-1} = \left| \frac{f_k - f_{k-1} - 0.5(g^T_{k-1}s_{k-1} + g^T_k s_k - 1)}{f_{k-1} + 0.5(g^T_{k-1}s_{k-1} + g^T_k s_k)} \right|. \]

If \(r_{k-1}\) or \(\tilde{r}_{k-1}\) is close to 0, then the function might be close to a quadratic function. If there are continuously many iterations such that \(r_{k-1} \leq \xi_4\) or \(\tilde{r}_{k-1} \leq \xi_5\), where \(\xi_4, \xi_5 > 0\), we restart the method with \(-g_k\). Similarly, if the number of the successive use of CG direction reaches to the threshold MaxRestart, we also restart the method with \(-g_k\).

Firstly, we describe the subspace minimization conjugate gradient method in which the direction of the regularization model is generated by the problem (33), which is called SMCG_PR1.
Remark 3 In Algorithm 1, Numgrad denotes the number of the successive use of the negative gradient direction; Isnotgra denotes the number of the successive use of the CG direction; MaxRestart represents a quantification and when the Isnotgra reaches this value, we restart the method with $-g_k$. MinQuad also represents a quantification and when the IterQuad reaches this value, we restart the method with $-g_k$. These parameters are related to the restart of the algorithm, which has an important impact on the numerical performance of the CG.

Secondly, we describe the subspace minimization conjugate gradient method in which the direction of the regularization model is generated by the problem (40).

If the condition

$$\left( g_k^T s_{k-1} \right)^2 > (1 - 10^{-5})\|g_k\|^2\|s_{k-1}\|^2$$

holds, the value of $\frac{\left( g_k^T s_{k-1} \right)^2}{\|g_k\|^2\|s_{k-1}\|^2}$ is close to 1, which means that vectors $g_k$ and $s_{k-1}$ may be linearly correlated. So the positive definiteness of the matrix $E_k$ in (40) might not be guaranteed. Therefore, we choose the quadratic model to derive a search direction.
We may consider to use “3.2. If the condition (27) or (28) or (29) holds, compute the search direction $d_{k+1}$ by (32) with (44) and (45). Set $\text{Isnotgra}:=\text{Isnotgra}+1$ and go to Step 4; otherwise, if the condition (53) holds, compute the search direction $d_{k+1}$ by (32), (41) and (42) with $\lambda = 0$, otherwise, compute the search direction $d_{k+1}$ by (32) with (41) and (42). Set $\text{Isnotgra}:=\text{Isnotgra}+1$ and go to Step 4.” to replace the Step 3.2 in Algorithm 1. The resulting method is called SMCG\_PR2. We use SMCG\_PR to denote either SMCG\_PR1 or SMCG\_PR2.

The following two Lemmas show some properties of the direction $d_k$, which are essential to the convergence of SMCG\_PR.

**Lemma 4.1** Suppose the direction $d_k$ is calculated by SMCG\_PR. Then, there exists a positive constant $c_1$ such that

$$g_k^T d_k \leq -c_1 \|g_k\|^2. \quad (54)$$

**Proof** We divide the proof into four cases.

**Case 1** The direction $d_k$ is given by (32) with (34) and (35), as in SMCG\_PR1. Denote $T = \frac{1}{1+\sigma_k(z^*)^p-2}$. Obviously, in this case,

$$\mu_k = T \hat{\mu}_k, \quad v_k = T \hat{v}_k. \quad (30)$$

If $\sigma_k(z^*)^p-2 > 1$, we have $T = \frac{1}{2}$ from the first line after (39). Moreover, $\sigma_k(z^*) \geq 0$. So we can establish that $\frac{1}{2} \leq T \leq 1$. From (3.31) and (3.32) of [15], we can get that

$$g_k^T d_k = T g_k^T (\hat{\mu}_k g_k + \hat{v}_k s_{k-1}) \leq -T \frac{\|g_k\|^4}{\rho_k} \leq -\frac{\|g_k\|^4}{2\rho_k}. \quad (55)$$

Substituting $\rho_k = \frac{3}{2} \frac{\|y_{k-1}\|^2}{s_{k-1}^T y_{k-1}} \|g_k\|^2$ into (55), we deduce that $g_k^T d_k \leq -\frac{\|g_k\|^4}{2\rho_k} = \frac{1}{3} \frac{1}{s_{k-1}^T y_{k-1}} \|g_k\|^2$. From (30), we know $-\frac{1}{\xi_1} \leq -\frac{1}{s_{k-1}^T y_{k-1}} \|g_k\|^2 \leq -\frac{1}{\xi_2}$. Therefore, we get

$$g_k^T d_k \leq -\frac{\|g_k\|^4}{2\rho_k} = -\frac{1}{3} \frac{1}{s_{k-1}^T y_{k-1}} \|g_k\|^2 \leq -\frac{1}{3\xi_2} \|g_k\|^2. \quad (56)$$

On the other hand, if the direction $d_k$ is given by (32) with (41) and (42), we have SMCG\_PR2. Then, by direct calculation

$$g_k^T d_k = \hat{\mu}_k \|g_k\|^2 + \hat{v}_k s_{k-1}^T g_k \quad (57)$$

$$= \|g_k\|^4 \left(\frac{s_{k-1}^T y_{k-1}}{s_{k-1}^T g_k}\right) + \rho_k \left(\frac{s_{k-1}^T y_{k-1}}{\|g_k\|^2}\right)^2 - \lambda s_{k-1}^T g_k s_{k-1} \frac{\|g_k\|^2}{\|s_{k-1}\|^2} + \lambda s_{k-1}^T g_k s_{k-1} \frac{\|g_k\|^2}{\|s_{k-1}\|^2} \leq -\frac{\|g_k\|^4}{3\xi_2} \|g_k\|^2 \leq -\frac{1}{3\xi_2} \|g_k\|^2. \quad (58)$$
Due to $0 \leq \lambda \leq \frac{\|y_{k-1}\|^2}{s_{k-1}^T y_{k-1}}$, we have $\frac{3}{2} \frac{\|y_{k-1}\|^2}{s_{k-1}^T y_{k-1}} \leq \frac{3}{2} \|y_{k-1}\|^2 + \lambda \leq \frac{5}{2} \|y_{k-1}\|^2$. So,
\[-\frac{2}{3} \frac{s_{k-1}^T y_{k-1}}{\|y_{k-1}\|^2} \leq \frac{-1}{\frac{3}{2} \|y_{k-1}\|^2 + \lambda} \leq \frac{-2}{\frac{3}{2} \|y_{k-1}\|^2}.
\] From (30), we know $-\frac{1}{\xi_1} \leq -\frac{s_{k-1}^T y_{k-1}}{\|y_{k-1}\|^2}$.
Therefore, the last inequality is established.

Case 2 $d_k = \tilde{\mu}_k g_k + \tilde{v}_k s_{k-1}$, where $\tilde{\mu}_k$ and $\tilde{v}_k$ are calculated by (44) and (45), respectively. From (55) and (56), we can get that
\[g_k^T d_k = g_k^T (\tilde{\mu}_k g_k + \tilde{v}_k s_{k-1}) \leq -\frac{2}{3 \xi_2} \|g_k\|^2. \tag{57}\]

Case 3 If the direction $d_k$ is given by (3) where $\beta_k = \beta_k^{HS}$, (54) is satisfied by setting $c_1 = 1 - \xi_3$. The proof is similar to Lemma 3 in [29].

Case 4 As $d_k = -g_k$, we can easily derive $g_k^T d_k = -\|g_k\|^2$ which satisfies (54) by setting $c_1 = \frac{1}{2}$.

To sum up, the sufficient descent condition (54) holds by setting
\[c_1 = \min \left\{ \frac{1}{2}, 1 - \xi_3, \frac{2}{3 \xi_2}, \frac{1}{\xi_1}, \frac{2}{3 \xi_2}, \frac{1}{5 \xi_2} \right\}, \]
which completes the proof.

Lemma 4.2 Suppose the direction $d_k$ is calculated by SMCG\_PR. Then, there exists a constant $c_2 > 0$ such that
\[\|d_k\| \leq c_2 \|g_k\|. \tag{58}\]

Proof The proof is also divided into four parts. \qed

Case 1 The direction $d_k$ is given by (32) with (34) and (35), as in SMCG\_PR1. From (3.12) in [29] and $T \leq 1$, we obtain
\[\|d_k\| = T \|\tilde{\mu}_k g_k + \tilde{v}_k s_{k-1}\| \leq \frac{20}{\xi_1} \|g_k\|.
\]

On the other hand, if the direction $d_k$ is given by (32) with (41) and (42), as in SMCG\_PR2. At first, we give a lower bound of $\Delta_k$. From (43), we have
\[
\Delta_k = \lambda^2 \left( \|g_k\|^2 \|s_{k-1}\|^2 - \left( s_k^T s_{k-1} \right)^2 \right) + \lambda \left( \rho_k \|s_{k-1}\|^2 + s_{k-1}^T y_{k-1} \|g_k\|^2 - 2g_k^T y_{k-1} s_k^T s_{k-1} \right)
+ \rho_k s_{k-1}^T y_{k-1} - \left( s_k^T y_{k-1} \right)^2.
\]
Moreover, using the Cauchy inequality and average inequality, we have
\[
\rho_k \| s_{k-1} \|^2 + s_{k-1}^T y_{k-1} g_k \|^2 - 2 g_k^T y_{k-1} g_k s_{k-1} \\
\geq \frac{3}{2} \left( \frac{\| y_{k-1} \|}{s_{k-1}^T y_{k-1}} \right)^2 \| s_{k-1} \|^2 + s_{k-1}^T y_{k-1} \| g_k \|^2 - 2 \| s_{k-1} \| \| y_{k-1} \| \| g_k \|^2 \\
= \left( \frac{1}{2} \frac{\| s_{k-1} \|}{s_{k-1}^T y_{k-1}} \| y_{k-1} \| + \frac{\| s_{k-1} \|}{s_{k-1}^T y_{k-1}} \| y_{k-1} \| + \frac{s_{k-1}^T y_{k-1}}{\| s_{k-1} \| \| y_{k-1} \|} - 2 \right) \| s_{k-1} \| \| y_{k-1} \| \| g_k \|^2 \\
\geq \left( \frac{1}{2} \frac{\| s_{k-1} \|}{s_{k-1}^T y_{k-1}} \| y_{k-1} \| + 2 - 2 \right) \| s_{k-1} \| \| y_{k-1} \| \| g_k \|^2 \\
\geq \frac{1}{2} \| s_{k-1} \| \| y_{k-1} \| \| g_k \|^2 \geq 0.
\]
It follows from (30) that \( s_{k-1}^T y_{k-1} \geq \xi_1 \| s_{k-1} \|^2 \). By \( \rho_k = \frac{3}{2} \frac{\| y_{k-1} \|}{s_{k-1}^T y_{k-1}} \| g_k \|^2 \), \( \lambda \geq 0 \) and the Cauchy inequality, we obtain a lower bound of \( \tilde{\Delta}_k \), that is
\[
\tilde{\Delta}_k \geq \rho_k s_{k-1}^T y_{k-1} - (g_k^T y_{k-1})^2 = s_{k-1}^T y_{k-1} \left( \rho_k - \frac{(g_k^T y_{k-1})^2}{s_{k-1}^T y_{k-1}} \right) \\
\geq \xi_1 \| s_{k-1} \|^2 \left( \rho_k - \frac{(g_k^T y_{k-1})^2}{s_{k-1}^T y_{k-1}} \right) \\
\geq \frac{1}{2} \xi_1 \| s_{k-1} \|^2 \frac{\| y_{k-1} \|^2}{s_{k-1}^T y_{k-1}} \| g_k \|^2.
\]
Using the triangle inequality, Cauchy inequality, \( \rho_k = \frac{3}{2} \frac{\| y_{k-1} \|}{s_{k-1}^T y_{k-1}} \| g_k \|^2 \), \( 0 \leq \lambda \leq \frac{\| y_{k-1} \|^2}{s_{k-1}^T y_{k-1}} \) and the last relation, we have
\[
\| d_k \| = \| \tilde{\mu}_k g_k + \tilde{\nu}_k s_{k-1} \| \\
= \| \frac{1}{\tilde{\Delta}_k} \left( g_k^T y_{k-1} - s_{k-1}^T y_{k-1} \| g_k \|^2 + \lambda \left( g_k^T s_{k-1} \| g_k \|^2 \right) \right) \\
\times g_k + \left( g_k^T y_{k-1} - s_{k-1}^T y_{k-1} \| g_k \|^2 - \rho_k s_{k-1}^T y_{k-1} \right) s_{k-1} \| \| \\
\leq \frac{1}{\tilde{\Delta}_k} \left( \| g_k^T y_{k-1} \| g_k \|^2 + \frac{\| y_{k-1} \|^2}{s_{k-1}^T y_{k-1}} \| g_k \|^2 + \lambda \left( g_k^T s_{k-1} \| g_k \|^2 \right) \right) \\
\times g_k + \left( g_k^T y_{k-1} - s_{k-1}^T y_{k-1} \| g_k \|^2 - \rho_k s_{k-1}^T s_{k-1} \| s_{k-1} \| \right) s_{k-1} \| \\
\leq \frac{1}{\tilde{\Delta}_k} \left( \| s_{k-1} \| \| y_{k-1} \| + 2 \| y_{k-1} \|^2 \| s_{k-1} \|^2 \| s_{k-1} \| \right) \| g_k \|^3 \\
+ \left( \| s_{k-1} \| \| y_{k-1} \| + \frac{\rho_k \| g_k \|^2}{\| s_{k-1} \| \| s_{k-1} \|} \right) \| g_k \|^3 \\
= \frac{1}{\tilde{\Delta}_k} \left( \| s_{k-1} \| \| y_{k-1} \| + \frac{7}{2} \| y_{k-1} \|^2 \| s_{k-1} \|^2 \right) \| g_k \|^3 \\
\leq \frac{13}{\xi_1} \| g_k \|.
\]

**Case 2** \( d_k = \tilde{\mu}_k g_k + \tilde{\nu}_k s_{k-1} \), where \( \tilde{\mu}_k \) and \( \tilde{\nu}_k \) are calculated by (44) and (45), respectively. From (3.12) in [29], we can get (58) is satisfied by setting \( c_2 = \frac{20}{\xi_1} \).

Springer
Case 3  If the direction $d_k$ is given by (3) where $\beta_k = \beta_{HS}^k$, (58) is satisfied by setting $c_2 = 1 + \frac{L}{\xi_1}$. The proof is same as Lemma 4 in [29].

Case 4  As $d_k = -g_k$, we can easily establish that $\|d_k\| = \|g_k\|$

In summary, we easily obtain the fact that (58) holds by

$$c_2 = \max \left\{ 1, 1 + \frac{L}{\xi_1}, \frac{20}{\xi_1} \right\},$$

which completes the proof.

5 Convergence analysis

In this section, we establish the global convergence and $R$-linear convergence of SMCG PR. We assume that $\|g_k\| \neq 0$ for each $k$; otherwise, there is a stationary point for some $k$.

At first, we suppose that the objective function $f$ satisfies the following assumptions. Define $\Theta$ as an open neighborhood of the level set $L(x_0) = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$, where $x_0$ is the initial point.

Assumption 1  $f$ is continuously differentiable and bounded from below in $\Theta$.

Assumption 2  The gradient $g$ is Lipchitz continuous in $\Theta$, namely, there exists a constant $L > 0$ such that $\|g(x) - g(y)\| \leq L \|x - y\|$, $\forall x, y \in \Theta$.

Lemma 5.1  Suppose Assumption 1 holds and the iterative sequence $\{x_k\}$ is generated by the SMCG PR. Then, we have $f_k \leq C_k$ for each $k$.

Proof  Due to (48) and descent direction $d_{k+1}$, $f_{k+1} < C_k$ always holds. Through (51), we can get $C_1 = C_0$ or $C_1 = f_1 + 1.0$. If $C_1 = C_0$, because of the relations $f_{k+1} < C_k$ and $C_0 = f_0$, we know $f_1 \leq C_1$. If $C_1 = f_1 + 1.0$, we can easily get $f_1 \leq C_1$. When $k \geq 1$, the updated form of $C_{k+1}$ is (50), according to the conclusion of Lemma 1.1 in [51], we have $f_{k+1} \leq C_{k+1}$. Therefore, $f_k \leq C_k$ holds for each $k$. \hfill $\Box$

Lemma 5.2  Suppose the Assumption 2 holds and the iterative sequence $\{x_k\}$ is generated by the SMCG PR. Then,

$$\alpha_k \geq \left(1 - \frac{\sigma}{L}\right) \frac{|g_k^T d_k|}{\|d_k\|^2}. \quad (59)$$

Proof  By (49) and Assumption 2, we have that

$$(\sigma - 1) g_k^T d_k \leq (g_{k+1} - g_k)^T d_k \leq \alpha_k L \|d_k\|^2.$$

Since $d_k$ is a descent direction and $\sigma < 1$, (59) follows immediately. \hfill $\Box$
Theorem 5.3 Suppose Assumptions 1 and 2 hold. If the iterative sequence \( \{x_k\} \) is generated by the SMCG PR, it follows

\[
\lim_{k \to \infty} \|g(x_k)\| = 0. \tag{60}
\]

**Proof** By (48), Lemma 5.2, Lemma 4.1, and Lemma 4.2, we get that

\[
f_{k+1} \leq C_k - \frac{\delta (1 - \sigma)}{L} \left( \frac{g_k^T d_k}{\|d_k\|^2} \right)^2 \leq C_k - \frac{\delta (1 - \sigma) c_1^2}{L c_2^2} \|g_k\|^2.
\]

In short, set \( \beta = \frac{\delta (1 - \sigma) c_1^2}{L c_2^2} \), we give the fact that

\[
f_{k+1} \leq C_k - \beta \|g_k\|^2. \tag{61}
\]

Now, we find an upper bound of \( Q_{k+1} \) in (50) with (52). As for \( k \geq 1 \), \( Q_{k+1} \) can be expressed as [32]

\[
Q_{k+1} = \begin{cases} 
1 + (l + 1) \sum_{i=1}^{k/l} \eta^i, & \text{mod} \ (k, l) = 0, \\
1 + (l + 1) + (l + 1) \sum_{i=1}^{|k/l|} \eta^i, & \text{mod} \ (k, l) \neq 0,
\end{cases}
\]

where \( \lfloor \cdot \rfloor \) is the floor function. Then, we obtain

\[
Q_{k+1} \leq 1 + \text{mod}(k, l) + (l + 1) \sum_{i=1}^{\lfloor k/l \rfloor + 1} \eta^i \\
\quad \leq 1 + (l + 1) + (l + 1) \sum_{i=1}^{\lfloor k/l \rfloor + 1} \eta^i \\
\quad \leq 1 + (l + 1) + (l + 1) \sum_{i=1}^{k+1} \eta^i \\
\quad = 1 + (l + 1) \sum_{i=0}^k \eta^i \\
\quad = 1 + \frac{(l+1)(1-\eta^{k+2})}{1-\eta} \\
\quad \leq 1 + l+1 \frac{1}{1-\eta}. \tag{62}
\]

Denote \( M = 1 + \frac{l+1}{1-\eta} \), which gives the fact \( Q_{k+1} \leq M \).

With the updated form of \( C_{k+1} \) in (50), (61) and (62), we obtain

\[
C_{k+1} = C_k + \frac{f_{k+1} - C_k}{Q_{k+1}} \leq C_k - \frac{\beta}{Q_{k+1}} \|g_k\|^2 \leq C_k - \frac{\beta}{M} \|g_k\|^2. \tag{63}
\]

According to (51), we know \( C_1 \leq C_0 \) which implies that \( C_k \) is monotonically decreasing. Due to Assumption 1 and Lemma 5.1, we can get \( C_k \) is bounded from below. Then

\[
\sum_{k=0}^{\infty} \frac{\beta}{M} \|g_k\|^2 < \infty.
\]
therefore,
\[ \lim_{k \to \infty} \|g(x_k)\| = 0, \]
which completes the proof.

Moreover, $R$-linear convergence of SMCG PR will be established as followed. In order to establish $R$-linear convergence of SMCG PR, we introduce Definition 1 and assume that the set $\chi^*$ is nonempty.

**Definition 1** The continuously differentiable function $f$ has a global error bound on $\mathbb{R}^n$, if there exists a constant $\kappa_f > 0$ such that for any $x \in \mathbb{R}^n$ and $x = [x]_{\chi^*}$, we have
\[ \|g(x)\| \geq \kappa_f \|x - \bar{x}\|, \quad \forall x \in \mathbb{R}^n, \tag{64} \]
where $\bar{x} = [x]_{\chi^*}$ is the projection of $x$ onto the nonempty solution set $\chi^*$. We further denote by $\chi^* = \arg \min_{x \in \mathbb{R}^n} f(x)$ the set of optimal solutions of problem (1).

**Remark 4** By Assumption 2 it is $\|g(x) - g(x^*)\| \leq L \|x - x^*\|$, so that it is also $\|g(x)\| \leq L \|x - x^*\|$, which implies $k_f \leq L$.

**Remark 5** [35] If $f$ is strongly convex on $\mathbb{R}^n$, it must satisfy Definition 1.

**Remark 6** If $f$ is a convex function and the optimal solution set is nonempty, we have $f(x) = f(x')$, $\forall x, x' \in \chi^*$.

**Theorem 5.4** Suppose that Assumption 2 holds, $f$ is convex with a minimizer $x^*$, being the solution set $\chi^*$ nonempty. Suppose there exists $\alpha > 0$ such that $0 < \alpha \leq \alpha_k$ for all $k$. Let $f$ satisfy Definition 1 with constant $\kappa_f > 0$. Then there exists $\theta \in (0, 1)$ such that
\[ f_k - f(x^*) \leq \theta^k (f_0 - f(x^*)). \]

**Proof** In what follows, we only consider the case of $\|g_k\| \neq 0$, $\forall k \geq 0$. From Lemma 5.1, we can get $f_{k+1} \leq C_{k+1}$. Due to Remark 6 and $\|g_k\| \neq 0$, $\forall k \geq 0$, we know $x_{k+1}$ is not the optimal solution. So, we have $f(x^*) < f_{k+1}$. From (63) and $\|g_k\| \neq 0$, $\forall k \geq 0$, we have that $C_{k+1} < C_k$. Therefore, we get $f(x^*) < f_{k+1} \leq C_{k+1} < C_k$, which means $f(x^*) < C_{k+1} < C_k$. It follows
\[ 0 < \frac{C_{k+1} - f(x^*)}{C_k - f(x^*)} < 1, \quad \forall k \geq 0. \tag{65} \]

Set
\[ r = \lim_{k \to \infty} \sup_{\|g_k\| \neq 0} \frac{C_{k+1} - f(x^*)}{C_k - f(x^*)}, \tag{66} \]
then, $0 \leq r \leq 1$.

First of all, we consider the case of $r = 1$. According to (66), there exists a subsequence $\{x_{k_j}\}$ such that
\[ \lim_{j \to \infty} \frac{C_{k_{j+1}} - f(x^*)}{C_{k_j} - f(x^*)} = 1. \tag{67} \]
Because of (62), there exists \( q > 0 \), \( 0 < q \leq \frac{1}{Q_{k+1}} \leq 1 \) holds. Hence, there exists a subsequence of \( \{x_{kj}\} \) such that the corresponding subsequence of \( \left\{ \frac{1}{Q_{k+1}} \right\} \) is convergent. Without loss of generality, we assume that

\[
\lim_{j \to \infty} \frac{1}{Q_{k+1}} = r_1.
\]  

Clearly, \( 0 < r_1 \leq 1 \).

By the updating formula of \( C_{k+1} \) in (50), we obtain

\[
\frac{C_{k+1} - f(x^*)}{C_k - f(x^*)} = \left( 1 - \frac{1}{Q_{k+1}} \right) + \frac{1}{Q_{k+1}} \frac{f_{k+1} - f(x^*)}{C_{k+1} - f(x^*)}.
\]

It follows from (67), (68) and finding the limit of upper formula that

\[
\lim_{j \to \infty} \frac{f_{k+1} - f(x^*)}{C_k - f(x^*)} = 1.
\]  

Using convexity of \( f \), the solution set \( \chi^* \) is nonempty and by Remark 6, we know \( f(x^*) = f(\bar{x}) \), where \( \bar{x} \) is introduced in Definition 1. So, we have that

\[
\frac{f_{k+1} - f(x^*)}{C_{k+1} - f(x^*)} = (\nabla f_{k+1}, x_{k+1} - \bar{x}) \leq \frac{1}{\kappa_f} \| g_{k+1} \|^2.
\]

Therefore, we get

\[
f_{k+1} - f(x^*) = f_{k+1} - f(\bar{x}) \leq (\nabla f_{k+1}, x_{k+1} - \bar{x}) \leq \frac{1}{\kappa_f} \| g_{k+1} \|^2. \tag{70}
\]

According to the Lipschitz continuity of \( g \), \( \alpha_k \leq \bar{\alpha} \) and (58), we have

\[
\| g_{k+1} \| \leq \| g_{k+1} - g_k \| + \| g_k \| \leq L \| x_{k+1} - x_k \| + \| g_k \| \leq (1 + L\bar{\alpha}c_2) \| g_k \|,
\]

together with (70), it implies that

\[
f_{k+1} - f(x^*) \leq \frac{1}{\kappa_f} (1 + L\bar{\alpha}c_2)^2 \| g_k \|^2.
\]

Dividing the above inequality by \( C_k - f(x^*) \), we have

\[
0 < \frac{f_{k+1} - f(x^*)}{C_k - f(x^*)} \leq \frac{(1 + L\bar{\alpha}c_2)^2 \| g_k \|^2}{\kappa_f (C_k - f(x^*))}. \tag{71}
\]

Based on (61)

\[
f_{k+1} - f(x^*) \leq C_k - f(x^*) - \beta \| g_k \|^2.
\]

Dividing both sides of above inequality by \( C_k - f(x^*) \), we get

\[
\frac{f_{k+1} - f(x^*)}{C_k - f(x^*)} \leq 1 - \frac{\beta \| g_k \|^2}{C_k - f(x^*)}.
\]
Combining with (69), then
\[
\lim_{j \to \infty} \frac{\|g_{k_j}\|^2}{C_{k_j} - f(x^*)} = 0,
\]
due to (71), it follows
\[
\lim_{j \to \infty} \frac{f_{k_j + 1} - f(x^*)}{C_{k_j} - f(x^*)} = 0,
\]
which contradicts with (69). Therefore, the case of \( r = 1 \) does not occur, that is,
\[
\lim_{k \to \infty} \sup C_{k+1} - f(x^*) = r < 1.
\]
Then, there exists an integer \( k_0 > 0 \) such that
\[
\frac{C_{k+1} - f(x^*)}{C_k - f(x^*)} < r + \frac{1 - r}{2} = \frac{1 + r}{2} < 1, \quad \forall k > k_0.
\]
From (65), we know that
\[
0 < \max_{0 \leq k \leq k_0} \left\{ \frac{C_{k+1} - f(x^*)}{C_k - f(x^*)} \right\} = \bar{r} < 1.
\]
Let \( \theta = \max \left\{ \frac{1 + r}{2}, \bar{r} \right\} \). Clearly, \( 0 < \theta < 1 \). It follows from (72) that
\[
C_{k+1} - f(x^*) \leq \theta \left( C_k - f(x^*) \right),
\]
which indicates that
\[
C_{k+1} - f(x^*) \leq \theta^k \left( C_0 - f(x^*) \right).
\]
In addition, due to \( f_{k+1} \leq C_{k+1} \) in Lemma 5.1 and \( C_0 = f_0 \), we can deduce that
\[
(f_k - f(x^*)) \leq \theta^k \left( f_0 - f(x^*) \right),
\]
which completes the proof.

6 Numerical results

In this section, numerical experiments are conducted to show the efficiency of the SMCG_PR with \( p = 3 \) and \( p = 4 \). We compare the performance of SMCG_PR with that of CG_DESCENT (5.3) [24], CGOPT [14], SMCG_BB [31], SMCG_Conic [30] and CG_DESCENT 6.8 [26] for the 145 test problems in the CUTEr library [20]. The names and dimensions for the 145 test problems are the same as that of the numerical results in [26]. The codes of CG_DESCENT, CGOPT and SMCG_BB can be downloaded from http://users.clas.ufl.edu/hager/papers/Software, http://coa.amss.ac.cn/wordpress/?page_id=21 and http://web.xidian.edu.cn/xdliuhongwei/paper.html, respectively.

The following parameters are used in SMCG_PR:
\[
\varepsilon = 10^{-6}, \quad \delta = 0.0005, \quad \sigma = 0.9999, \quad \lambda_{\text{min}} = 10^{-30}, \quad \lambda_{\text{max}} = 10^{30}, \quad \gamma = 10^{-5},
\]
\[
\xi_1 = 10^{-7}, \quad \xi_2 = 1.25 \times 10^{4}, \quad \xi_3 = 10^{-5}, \quad \xi_4 = 10^{-9}, \quad \xi_5 = 10^{-11}, \quad c_1 = 10^{-4}, \quad c_2 = 0.080.
\]
CG_DESCENT (5.3), CGOPT, SMCG_BB, SMCG_Conic and CG_DESCENT 6.8 use the default parameters in their codes. All test methods are terminated if \( \|g_k\|_{\infty} \leq \)}
$10^{-6}$ is satisfied or the number of iterations exceeds 200,000. For each comparison, however, we excluded those problems for which different solvers converge to different local minimizers.

The performance profiles introduced by Dolan and Moré [16] are used to display the performances of the test methods. We present four groups of the numerical experiments. They all run in Ubuntu 10.04 LTS which is fixed in a VMware Workstation 10.0 installed in Windows 7. In the following Figs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 and 16 and Table 2, “$N_{iter}$”, “$N_f$”, “$N_g$” and “$T_{cpu}$” represent the number of iterations, the number of function evaluations, the number of gradient evaluations and CPU time(s), respectively.

Fig. 1  Performance profile based on $N_{iter}$ (CUTEr)

Fig. 2  Performance profile based on $N_f$ (CUTEr)
In the first group of numerical experiments, we compare SMCG_PR1 and SMCG_PR2 with $p = 3$ and $p = 4$. All these test methods can successfully solve 139 problems. It is observed from Figs. 1, 2, 3 and 4 that the SMCG_PR1 with $p = 3$ is better than others.

In the second group of numerical experiments, we compare SMCG_PR1 ($p = 3$) with CG_DECENT (5.3) and CGOPT. SMCG_PR1 successfully solves 139 problems, while CG_DECENT (5.3) and CGOPT successfully solve 144 and 134 problems, respectively. After eliminating the problems for which the three variants converge to different local minimizers, 133 problems are left.
Fig. 5 Performance profile based on $N_{iter}(\text{CUTEr})$

Regarding the number of iterations in Fig. 5, we observe that SMCG_PR1 is more efficient than CG_DESCENT (5.3) and CGOPT, since it successfully solves about 50.4% of the test problems with the least number of iterations, while the percentages of solved problems of CG_DESCENT (5.3) and CGOPT are 42.8% and 23.3%, respectively. As shown in Fig. 6, we see that SMCG_PR1 outperforms CG_DESCENT (5.3) and CGOPT for the number of function evaluations.

Figure 7 presents the performance profile relative to the number of gradient evaluations. We can observe that the SMCG_PR1 is the top performance and solves about 54.2% of the test problems with the least number of gradient evaluations, and CG_DESCENT (5.3) solves about 31.6% of the test problems and CGOPT solves
about 21.8% of the test problems. From Fig. 8, we can see that SMCG_PR1 is fastest for about 66.2% of the test problems, while CG_DESCENT (5.3) and CGOPT are fastest for about 8.3% of the test problems and 34.6% of the test problems, respectively. From Figs. 5, 6, 7 and 8, it indicates that SMCG_PR1 outperforms CG_DESCENT (5.3) and CGOPT for the 145 test problems in the CUTEr library.

In the third group of the numerical experiments, we compare SMCG_PR1 \((p = 3)\) with SMCG_BB and SMCG_Conic [30]. SMCG.PR1 successfully solves 139 problems, which are 1 problem more than SMCG_Conic, while SMCG_BB successfully solves 140 problems. There are 137 problems left after the elimination process mentioned above. As shown in Figs. 9, 10, 11 and 12, we can easily observe that
SMCG_PR1 is superior to SMCG_BB and SMCG_Conic for the 145 test problems in the CUTEr library.

Due to limited space, we do not list all detailed numerical results. Instead, we present some numerical results about SMCG_PR1 \((p = 3)\), CG_DESCENT (5.3), CGOPT, SMCG_BB and SMCG_Conic for some ill-conditioned problems. Table 1 illustrates the notations, names and dimensions about the ill-conditioned problems. Table 2 presents some numerical results about SMCG_PR1 \((p = 3)\), CG_DESCENT (5.3), CGOPT, SMCG_BB and SMCG_Conic for the problems in Table 1. As shown in Table 2, the most famous CG software packages CGOPT and CG_DESCENT (5.3) both require many iterations, function evaluations and gradient evaluations when
Fig. 11 Performance profile based on $N_g$(CUTEr)

solving these ill-conditioned problems, though the dimensions of some of these ill-conditioned problems are small. From Table 2, we observe that SMCG_PR1 ($p = 3$) has significant improvements over the other test methods, especially for CGOPT and CG_DESCENT (5.3). It indicates that SMCG_PR1 ($p = 3$) is relatively competitive for ill-conditioned problems compared to other test methods.

In the fourth group of the numerical experiments, we compare SMCG_PR1 ($p = 3$) and CG_DESCENT 6.8 with memory = 11 and memory = 0, denoted by CG_DESCENT 6.8 ($m = 11$) and CG_DESCENT 6.8 ($m = 0$) respectively. For the 83 test problems of the 145 test problems in the CUTEr library, and the dimensions of which vary from 12 and $10^4$. The reason that the test problems whose dimensions

Fig. 12 Performance profile based on $T_{cpu}$(CUTEr)
are less than or equal to 11 are omitted is that CG DESCENT \( (m = 11) \) reduces to the LBFGS method \([34]\) when \( m \leq 11 \).

In Fig. 13, we observe that SMCG_PR1 \( (p = 3) \) performs slightly inferior to CG_DESCENT 6.8 \( (m = 11) \), but better than CG_DESCENT 6.8 \( (m = 0) \) relative to the number of iterations. In Fig. 14, we see that SMCG_PR1 \( (p = 3) \) outperforms slightly CG_DESCENT 6.8 \( (m = 0) \) relative to the number of function evaluations, and is at a disadvantage only for the case of \( \tau < 5.4 \) in contrast with CG_DESCENT 6.8 \( (m = 11) \). Figure 15 indicates that SMCG_PR1 \( (p = 3) \) has a significant improvement over CG_DESCENT 6.8 \( (m = 0) \), and is better than CG_DESCENT 6.8 \( (m = 11) \) for the case of \( \tau < 2.0 \). And we can observe that the
Fig. 15 Performance profile based on $N_g$(CUTEr)

Fig. 16 Performance profile based on $T_{cpu}$(CUTEr)

Table 1 Some ill-conditioned problems in CUTEr

| Notation | Name      | Dimension |
|----------|-----------|-----------|
| P1       | EIGENBLS  | 2550      |
| P2       | EXTROSNB  | 1000      |
| P3       | GROWTHLS  | 3         |
| P4       | MARATOSB  | 2         |
| P5       | NONCVXU2  | 5000      |
| P6       | PALMER1C  | 8         |
| P7       | PALMER1D  | 7         |
| P8       | PALMER2C  | 8         |
| P9       | PALMER4C  | 8         |
| P10      | PALMER6C  | 8         |
| P11      | PALMER7C  | 8         |
### Table 2: Numerical results for some ill-conditioned problems in CUTEr

| Problem | SMCG_PR1  | CG_DESCENT (5.3)  | CGOPT   | SMCG_BB   | SMCG_Conic |
|---------|-----------|--------------------|----------|-----------|------------|
|         | $N_{iter}$ / $N_f$ / $N_g$ | $N_{iter}$ / $N_f$ / $N_g$ | $N_{iter}$ / $N_f$ / $N_g$ | $N_{iter}$ / $N_f$ / $N_g$ | $N_{iter}$ / $N_f$ / $N_g$ |
| P1      | 9190/1832/9192 | 16092/32185/16093 | 19683/39369/19686 | 16040/32066/16041 | 12330/24654/12332 |
| P2      | 3568/6956/3574 | 6879/13839/6975  | 9127/18465/9305 | 8416/16195/8426 | 3733/7466/3735 |
| P3      | 1/2/2      | 441/997/596      | 480/1241/644  | 689/1512/711  | 1/2 |
| P4      | 212/614/389 | 946/2911/2191    | 1411/4185/2213 | 1159/9592/2634 | 3640/13621/5632 |
| P5      | 6096/12174/6098 | 7160/13436/8046 | 6195/12402/6207 | 6722/12800/6723 | 6459/12816/6460 |
| P6      | 1453/2093/1546 | 126827/224532/378489 | Failed | 88047/135548/89509 | 13007/23796/13532 |
| P7      | 445/682/470  | 3971/5428/10036  | 16490/36567/19846 | 2701/3703/2727 | 584/943/635 |
| P8      | 307/440/318  | 21362/21455/42837 | 25716/61275/30492 | 4894/7169/5002 | 695/1386/697 |
| P9      | 54/107/59   | 4421/49913/96429 | 88681/197232/105736 | 1064/1622/1074 | 1055/2025/1071 |
| P10     | 202/323/213  | 14174/14228/28411 | 29118/63118/31844 | 35704/58676/36281 | 1458/2429/1505 |
| P11     | 6288/8757/6576 | 65294/78428/149585 | 98699/220388/119626 | 46397/65692/46929/ | 502/575/514 |
SMCG_PR1 \((p = 3)\) is the top performance and solves about 61.0\% of test problems with the least number of gradient evaluations, and CG_DESCENT 6.8 \((m = 0)\) solves about 26.5\% problems and CG_DESCENT 6.8 \((m = 11)\) solves about 38.0\% problems. From Fig. 16, we observe that SMCG_PR1 \((p = 3)\) performs slightly inferior to CG_DESCENT 6.8 \((m = 11)\), but better than CG_DESCENT 6.8 \((m = 0)\), and we can see that SMCG_PR1 \((p = 3)\) is fastest for about 55.5\% of test problems, while CG_DESCENT 6.8 \((m = 0)\) and CG_DESCENT 6.8 \((m = 11)\) are fastest for about 4.0\% and 46.8\%, respectively. It indicates that SMCG_PR1 \((p = 3)\) is better than CG_DESCENT 6.8 \((m = 0)\), and is inferior to CG_DESCENT 6.8 \((m = 11)\), which is what we expected since CG_DESCENT 6.8 \((m = 11)\) generates the search direction with memory = 11 while SMCG_PR1 \((p = 3)\) only with memory = 2.

The numerical results indicate that the SMCG_PR method outperforms CG_DESCENT (5.3), CGOPT, SMCG_BB, SMCG_Conic and CG_DESCENT 6.8 \((m = 0)\).

7 Conclusions

In this paper, we present two new subspace minimization conjugate gradient methods based on the special \(p\)-regularization model for \(p > 2\). In the proposed methods, the search directions satisfy the sufficient descent condition. Under mild conditions, the global convergences of SMCG_PR is established. We also prove that SMCG_PR is \(R\)-linearly convergent. The numerical experiments show that SMCG_PR is very promising.

Acknowledgments We would like to thank the anonymous referees for their useful comments. We also would like to thank Professors Hager, W.W. and Zhang, H.C. for their CG_DESCENT (5.3) code, and thank Professor Dai, Y.H and Dr. Kou, C.X. for their CGOPT code.

Funding This research is supported by the National Science Foundation of China (No. 11901561), Guangxi Natural Science Foundation (No. 2018GXNSFBA281180) and China Postdoctoral Science Foundation (2019M660833).

References

1. Andrei, N.: An accelerated subspace minimization three-term conjugate gradient algorithm for unconstrained optimization. Numer. Algorithm. 65, 859–874 (2014)
2. Andrea, C., Tayebeh, D.N., Stefano, L.: On global minimizers of quadratic functions with cubic regularization. Optim. Lett. 13, 1269–1283 (2019)
3. Barzilai, J., Borwein, J.M.: Two-point step size gradient methods. IMA J. Numer. Anal. 8, 141–148 (1988)
4. Bellavia, S., Morini, B., Cartis, C., Gould, N.I.M., Toint, P.L.: Convergence of a regularized euclidean residual algorithm for nonlinear least-squares. SIAM J. Numer. Anal. 48, 1–29 (2010)
5. Bellavia, S., Morini, B.: Strong local convergence properties of adaptive regularized methods for nonlinear least squares. IMA J. Numer. Anal. 35, 947–968 (2014)
6. Benson, H.Y., Shanno, D.F.: Interior-point methods for nonconvex nonlinear programming:cubic regularization. Comput. Optim. Appl. 58, 323–346 (2014)
7. Bianconcini, T., Liuzzi, G., Morini, B., Sciandrone, M.: On the use of iterative methods in cubic regularization for unconstrained optimization. Comput. Optim. Appl. 60, 35–57 (2015)
8. Bianconcini, T., Sciandrone, M.: A cubic regularization algorithm for unconstrained optimization using line search and nonmonotone techniques. Optim. Methods Softw. 31, 1008–1035 (2016)
9. Birgin, E.G., Gardenghi, J.L., Martinez, J.M., Santos, S.A., Toint, P.L.: Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. Math. Program. 163, 359–368 (2017)
10. Cartis, C., Gould, N.I.M., Toint, P.L.: Adaptive cubic regularisation methods for unconstrained optimization. Part i: motivation, convergence and numerical results. Math. Program. 127, 245–295 (2011)
11. Cartis, C., Gould, N.I.M., Toint, P.L.: Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function-and derivative-evaluation complexity. Math. Program. 130, 295–319 (2011)
12. Dai, Y.H., Liu, L.Z.: New conjugacy conditions and related nonlinear conjugate gradient methods. Appl. Math. Optim. 43, 87–101 (2001)
13. Dai, Y.H., Yuan, Y.: A nonlinear conjugate gradient method with a strong global convergence property. SIAM J. Optim. 10, 177–182 (1999)
14. Dai, Y.H., Kou, C.X.: A nonlinear conjugate gradient algorithm with an optimal property and an improved Wolfe line search. SIAM J. Optim. 23, 296–320 (2013)
15. Dai, Y.H., Kou, C.X.: A Barzilai-Borwein conjugate gradient method. Sci. China Math. 59, 1511–1524 (2016)
16. Dolan, E.D., Moré, J.J.: Benchmarking optimization software with performance profiles. Math. Program. 91, 201–213 (2002)
17. Dussault, J.P.: Simple Unified Convergence Proofs for the Trust-Region and a New ARC variant. Tech. Rep., University of Sherbrooke, Sherbrooke, Canada (2015)
18. Fatemi, M.: A new efficient conjugate gradient method for unconstrained optimization. J. Comput. Appl. Math. 300, 207–216 (2016)
19. Fletcher, R., Reeves, C.M.: Function minimization by conjugate gradients. Comput. J. 7, 149–154 (1964)
20. Gould, N.I.M., Orban, D.: CUTEr and SifDec: a constrained and unconstrained testing environment, revisited. ACM Trans. Math. Softw. 29, 373–394 (2003)
21. Gould, N.I.M., Porcelli, M.: Updating the regularization parameter in the adaptive cubic regularization algorithm. Comput. Optim. Appl. 53, 1–22 (2012)
22. Gould, N.I.M., Robinson, D.P., Thorne, H.S.: On solving trust-region and other regularised subproblems in optimization. Math. Program. Comput. 2, 21–57 (2010)
23. Griewank, A.: The modification of Newton’s method for unconstrained optimization by bounding cubic terms. Technical Report NA/12, Department of Applied Mathematics and Theoretical Physics University of Cambridge (1981)
24. Hager, W.W., Zhang, H.: A new conjugate gradient method with guaranteed descent and an efficient line search. SIAM J. Optim. 16, 170–192 (2005)
25. Hager, W.W., Zhang, H.: A survey of nonlinear conjugate gradient methods. Pac. J. Optim. 2, 35–58 (2006)
26. Hager, W.W., Zhang, H.: The limited memory conjugate gradient method. SIAM J. Optim. 23, 2150–2168 (2013)
27. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear system. J. Res. Natl. Bur. Stand. 49, 409–436 (1952)
28. Li, L.B.: A new algorithm for solving large scale trust region subproblem. Oper. Res. Manag. Sci. 16, 48–52 (2007)
29. Li, M., Liu, H.W., Liu, Z.X.: A new subspace minimization conjugate gradient method with nonmonotone line search for unconstrained optimization. Numer. Algorithms. 79, 195–219 (2018)
30. Li, Y.F., Liu, Z.X., Liu, H.W.: A subspace minimization conjugate gradient method based on conic model for unconstrained optimization. Comput. Appl. Math. 38 (2019)
31. Liu, H.W., Liu, Z.X.: An efficient Barzilai-Borwein conjugate gradient method for unconstrained optimization. J. Optim. Theory Appl. 180, 879–906 (2019)
32. Liu, Z.X., Liu, H.W.: Several efficient gradient methods with approximate optimal stepsizes for large scale unconstrained optimization. J. Comput. Appl. Math. 328, 400–413 (2018)
33. Liu, Z.X., Liu, H.W.: An efficient gradient method with approximate optimal stepsize for large-scale unconstrained optimization. Numer. Algorithms. 78, 21–39 (2018)
34. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Math. Program. 45, 503–528 (1989)
35. Necula, I., Nesterov, Yu., Glineur, F.: Linear convergence of first order methods for non-strongly convex optimization. Math. Program., Ser. A. 175, 69–107 (2018)
36. Nesterov, Y., Polyak, B.T.: Cubic regularization of Newton’s method and its global performance. Math. Program. 108, 177–205 (2006)
37. Nathan, C., Autar, K., Jai, P., Michael, K.: Newton-Raphson Method-Graphical Simulation of the Method. University of South Florida. http://numericalmethods.eng.usf.edu/mws (2003)
38. Nesterov, Y.: Accelerating the cubic regularization of Newton's method on convex problems. Math. Program. 112, 159–181 (2008)
39. Polyak, B.T.: The conjugate gradient method in extreme problems. USSR Comput. Math. Math. Phys. 9, 94–112 (1969)
40. Radosaw, P.: Conjugate Gradient Algorithms in Nonconvex Optimization. Springer, Berlin (2009)
41. Rivaie, M., Mamat, M., Abashar, A.: A new class of nonlinear conjugate gradient coefficients with exact and inexact line searches. Appl. Math. Comput. 268, 1152–1163 (2015)
42. Sun, W.Y.: On nonquadratic model optimization methods. Asia Pac. J. Oper. Res. 13, 43–63 (1996)
43. Weiser, M., Deuflhard, P., Erdmann, B.: Affine conjugate adaptive Newton methods for nonlinear elastomechanics. Optim. Methods Softw. 22, 413–431 (2007)
44. Yang, Y.T., Chen, Y.T., Lu, Y.L.: A subspace conjugate gradient algorithm for large scale constrained optimization. Numer. Algorithm. 76, 813–828 (2017)
45. Hsia, Y., Sheu, R.L., Yuan, Y.X.: Theory and application of p-regularized subproblems for $p > 2$. Optim Methods Softw. 1059–1077 (2017)
46. Yuan, Y.X., Sun, W.Y.: Optimization Theory and Methods. Science Press, Beijing (1997)
47. Yuan, Y.X., Stoer, J.: A subspace study on conjugate gradient algorithms. Z. Angew. Math. Mech. 75, 69–77 (1995)
48. Yuan, Y.X.: A modified BFGS algorithm for unconstrained optimization. IMA J. Numer. Anal. 11, 325–332 (1991)
49. Yuan, Y.X.: A review on subspace methods for nonlinear optimization. In: Proceedings of the International Congress of Mathematics 2014, Seoul, Korea, pp. 807–827 (2014)
50. Yuan, Y.X.: Subspace methods for large scale nonlinear equations and nonlinear least squares. Optim. Eng. 10, 207–218 (2009)
51. Zhang, H., Hager, W.W.: A nonmonotone line search technique and its application to unconstrained optimization. SIAM J. Optim. 14, 1043–1056 (2004)
52. Zhou, B., Gao, L., Dai, Y.H.: Gradient methods with adaptive stepsizes. Comput. Optim. Appl. 35, 69–86 (2006)

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.