RESEARCH ARTICLE
doi: 10.2306/scienceasia1513-1874.2015.41.345

ScienceAsia 41 (2015): **345–349**

A new spectral Polak-Ribière-Polyak conjugate gradient method

Xuesha Wu

College of General Education, Chongqing College of Electronic Engineering, Chongqing 401331 China

e-mail: wuxuesha2013@126.com

Received 29 Jan 2015
Accepted 30 Sep 2015

ABSTRACT: By combining the spectral gradient and Polak-Ribière-Polyak (PRP) methods, a new spectral PRP conjugate gradient method is proposed to solve large-scaled unconstrained optimization problems. The method satisfies the famous conjugacy condition: $d_k^{\mathrm{T}} y_{k-1} = 0$, independent of any line search. The direction at each iteration generated by the proposed method is downward for the general objective function without any line search. Under the standard Wolfe line search, we prove that the proposed method is globally convergent. Finally, the proposed method is compared with the PRP method and the scaled PRP method using a classical set of problems.

KEYWORDS: unconstrained optimization, spectral gradient method, standard Wolfe line search, global convergence

MSC2010: 90C30

INTRODUCTION

Consider the following unconstrained optimization problem:

$$\text{find } \arg\min_{x\in\mathbb{R}^n} f(x), \tag{1}$$

where $f:\mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. The conjugate gradient method is one of the most efficient iterative methods to solve problem (1) especially when n is large. When solving this problem, the iterate is given by

$$x_{k+1} = x_k + \alpha_k d_k, \tag{2}$$

where the step-size $\alpha_k > 0$ is obtained by some line search, and d_k is the search direction computed from

$$d_k = \begin{cases} -g_k, & k = 1, \\ -g_k + \beta_k d_{k-1}, & k \geqslant 2, \end{cases} \tag{3}$$

where $g_k = \nabla f(x_k)$ and β_k is known as the gradient parameter. Plenty of conjugate gradient methods are known and an excellent survey of these methods, with special attention on their global convergence properties, is given by Hager and Zhang[1]. Different conjugate gradient methods correspond to different choices of β_k. In this paper, we are interested in the Polak-Ribière-Polyak (PRP) method[2,3], in which the parameter β_k is computed from

$$\beta_k^{\mathrm{PRP}} = \frac{g_k^{\mathrm{T}} y_{k-1}}{\|g_{k-1}\|^2}, \tag{4}$$

where $y_{k-1} = g_k - g_{k-1}$, and $\|\cdot\|$ stands for the Euclidean norm. When the step-size α_k is very small, y_{k-1} in the numerator of β_k^{PRP} tends to zero. Then β_k^{PRP} becomes small and the direction d_k is very close to the steepest descent direction $-g_k$. Thus the PRP method has a built-in restart feature that directly addresses the jamming problem. This feature means that the PRP method has been one of the most efficient conjugate gradient methods in practical computation for many years. However, Dai[4] constructed an example to indicate that the PRP method may generate an upward direction resulting in the iterative scheme failing even if the objective function is uniformly convex under the strong Wolfe line search. So far, the convergence of the PRP method has not been completely proved under the Wolfe-type line search.

Another popular method to solve problem (1) is the spectral gradient method proposed originally by Barzilai and Borwein[5]. The direction d_k is generated by

$$d_k = -\theta_k g_k + \beta_k s_{k-1}, \tag{5}$$

where $s_{k-1} = \alpha_{k-1} d_{k-1}$ and θ_k is the spectral gradient parameter. In Ref. 6, Raydan introduced the spectral gradient method for large-scale unconstrained optimization problems. An attractive property of this method is that it only needs gradient directions at each line search whereas a non-monotone strategy guarantees the global convergence. Surprisingly, the spectral gradient method outperforms the sophisticated conjugate gradient method in many known problems. Birgin and Martínez[7] proposed a spectral gradient method in which d_k is computed from (5). One parameter θ_k is generated by

$$\beta_k = \frac{\theta_k g_k^{\mathrm{T}} y_{k-1}}{\alpha_{k-1}\theta_{k-1}\|g_{k-1}\|^2}. \tag{6}$$

If $\theta_k = \theta_{k-1} = 1$, this is the classical parameter (4). Motivated by the success of spectral gradient method, they also compute θ_k using

$$\theta_k = s_{k-1}^{\mathrm{T}} s_{k-1} / s_{k-1}^{\mathrm{T}} y_{k-1}. \tag{7}$$

Under the standard Wolfe line search, they show that the scaled PRP method (5)–(7) is very effective. However, the scaled PRP method cannot guarantee the descent direction at each iteration, which may lead to failure of the iterative scheme.

Because of the advantages of the PRP method and the scaled PRP method, we consider a new spectral PRP (SPRP) conjugate gradient method. The proposed SPRP method not only processes the sufficient descent property and global convergence property, but also satisfies the famous conjugacy condition.

The rest of this paper is organized as follows. First, we introduce the SPRP method and prove its descent property without any line search. Second, the global convergence of the SPRP method is established under the standard Wolfe line search. Preliminary numerical results are then presented.

THE SPRP METHOD AND ITS DESCENT PROPERTY

In this paper, we solve problem (1) using a new iterative method, in which the iterative point is generated by (2) and the direction d_k is obtained by

$$d_k = \begin{cases} -g_k, & k = 1, \\ -\theta_k g_k + \beta_k d_{k-1}, & k \geqslant 2, \end{cases} \tag{8}$$

where θ_k is the spectral gradient parameter, and $\beta_k = \beta_k^{\mathrm{PRP}}$. Obviously, if $\theta_k = 1$, it reduces to the PRP method. In our method, the parameter θ_k is selected in such a way that at each iteration the conjugacy condition is satisfied independent of the line search. Multiplying (8) by y_{k-1}^{T}, we have

$$d_k^{\mathrm{T}} y_{k-1} = -\theta_k g_k^{\mathrm{T}} y_{k-1} + \beta_k d_{k-1}^{\mathrm{T}} y_{k-1}.$$

Hence, from the conjugacy condition: $d_k^{\mathrm{T}} y_{k-1} = 0$, we obtain

$$\theta_k = \frac{d_{k-1}^{\mathrm{T}} y_{k-1}}{\|g_{k-1}\|^2}. \tag{9}$$

So the method constructed by (8) and (9) always satisfies the conjugacy condition, and has the structure feature of the spectral gradient method.

In the following, we give the specific iterative algorithm, and refer to it as the SPRP method.

Algorithm 1

Step 1: Data: $x_1 \in \mathbb{R}^n$, $\varepsilon \geqslant 0$. Set $d_1 = -g_1$, if $\|g_1\| \leqslant \varepsilon$, then stop.

Step 2: Compute $\alpha_k > 0$ using the standard Wolfe line search:

$$f(x_k + \alpha_k d_k) \leqslant f(x_k) + \delta \alpha_k g_k^{\mathrm{T}} d_k, \tag{10}$$

$$g(x_k + \alpha_k d_k)^{\mathrm{T}} d_k \geqslant \sigma g_k^{\mathrm{T}} d_k, \tag{11}$$

where $0 < \delta < \sigma < 1$.

Step 3: Let $x_{k+1} = x_k + \alpha_k d_k$, $g_{k+1} = g(x_{k+1})$. If $\|g_{k+1}\| \leqslant \varepsilon$, then stop.

Step 4: Compute β_{k+1} using (4); generate θ_{k+1} using (9).

Step 5: If $d_{k+1}^{\mathrm{T}} g_{k+1} > -10^{-3} \|d_{k+1}\| \cdot \|g_{k+1}\|$ is satisfied, we set $d_{k+1} = -\theta_{k+1} g_{k+1}$; otherwise, we compute d_{k+1} by

$$d_{k+1} = -\theta_{k+1} g_{k+1} + \beta_{k+1} d_k.$$

Step 6: Set $k = k + 1$, go to step 2.

Lemma 1 *Let the sequences $\{g_k\}$ and $\{d_k\}$ be obtained by the SPRP method in which α_k satisfies any line search. Then we have*

$$g_k^{\mathrm{T}} d_k < -\left(\omega \frac{\|d_{k-1}\|}{\|g_{k-1}\|}\right) \|g_k\|^2 \tag{12}$$

where $\omega > 0$.

Proof: Multiplying (8) by g_k^{T}, we have

$$g_k^{\mathrm{T}} d_k = -\theta_k \|g_k\|^2 + \beta_k g_k^{\mathrm{T}} d_{k-1}. \tag{13}$$

From (4), (9), and (13), we obtain

$$g_k^{\mathrm{T}} d_k = -\frac{d_{k-1}^{\mathrm{T}} y_{k-1} \|g_k\|^2}{\|g_{k-1}\|^2} + \frac{g_k^{\mathrm{T}} y_{k-1} g_k^{\mathrm{T}} d_{k-1}}{\|g_{k-1}\|^2}$$

$$= \frac{\|g_k\|^2 g_{k-1}^{\mathrm{T}} d_{k-1} - g_k^{\mathrm{T}} g_{k-1} \cdot g_k^{\mathrm{T}} d_{k-1}}{\|g_{k-1}\|^2}. \tag{14}$$

Denote: $\varphi_{k-1} = \angle(g_{k-1}, d_{k-1})$, $\psi_{k-1} = \angle(g_k, g_{k-1})$, $\phi_{k-1} = \angle(g_k, d_{k-1})$. Obviously, $\varphi_{k-1}, \psi_{k-1}, \phi_{k-1} \in$

$(0,\pi)$ and $\varphi_{k-1}=\psi_{k-1}+\phi_{k-1}$. By (14), we obtain

$$\begin{aligned} g_k^{\mathrm T} d_k &= \frac{\|g_k\|^2\cdot\|g_{k-1}\|\cdot\|d_{k-1}\|\cos\varphi_{k-1}}{\|g_{k-1}\|^2} \\ &\quad - \frac{\|g_k\|^2\cdot\|g_{k-1}\|\cdot\|d_{k-1}\|\cos\psi_{k-1}\cos\phi_{k-1}}{\|g_{k-1}\|^2} \\ &= \frac{\|g_k\|^2\cdot\|d_{k-1}\|\cos(\psi_{k-1}+\phi_{k-1})}{\|g_{k-1}\|} \\ &\quad - \frac{\|g_k\|^2\cdot\|d_{k-1}\|\cos\psi_{k-1}\cos\phi_{k-1}}{\|g_{k-1}\|} \\ &= -\frac{\|g_k\|^2\cdot\|d_{k-1}\|\sin\psi_{k-1}\sin\phi_{k-1}}{\|g_{k-1}\|}. \end{aligned}$$

Since $\psi_{k-1},\phi_{k-1}\in(0,\pi)$, there exists a positive constant $\omega>0$ such that

$$\sin\psi_{k-1}\sin\phi_{k-1}>\omega,\quad \forall k\in\mathbb{Z}^+,$$

which implies that (12) holds. □

GLOBAL CONVERGENCE ANALYSIS

In order to establish the global convergence of the SPRP method, we need the following assumptions for the objective function. Assumption H:

(i) The level set $\Phi=\{x\mid f(x)\leqslant f(x_1)\}$ is bounded, where x_1 is the starting point.

(ii) In a neighbourhood Ω of Φ, the objective function is continuously differentiable and its gradient is Lipschitz continuous, i.e., there exists a constant $L>0$ such that

$$\|g(x)-g(y)\|\leqslant L\,\|x-y\|,\quad \forall x,y\in\Omega. \tag{15}$$

These assumptions imply that there is a positive constant γ such that

$$\|g(x)\|\leqslant\gamma,\quad \forall x,y\in\Omega. \tag{16}$$

The Zoutendijk condition[8] is very important for proving the global convergence of the conjugate gradient method. Now we prove that the SPRP method also satisfies the Zoutendijk condition.

Lemma 2 *Suppose Assumption H holds. Let the sequences $\{g_k\}$ and $\{d_k\}$ be obtained by the SPRP method. Then we have*

$$\sum_{k\geqslant1}(g_k^{\mathrm T}d_k)^2/\|d_k\|^2<+\infty. \tag{17}$$

Proof: From (11) and Assumption H(ii), we have

$$\begin{aligned} -(1-\sigma)d_k^{\mathrm T}g_k &\leqslant d_k^{\mathrm T}(g_{k+1}-g_k) \\ &\leqslant \|d_k\|\cdot\|g_{k+1}-g_k\|\leqslant L\alpha_k\,\|d_k\|^2. \end{aligned}$$

Then we have

$$\alpha_k\geqslant\frac{\sigma-1}{L}\frac{d_k^{\mathrm T}g_k}{\|d_k\|^2}.$$

By (10), we see that

$$f(x_k)-f(x_k+\alpha_k d_k)\geqslant\frac{\delta(1-\sigma)}{L}\frac{(d_k^{\mathrm T}g_k)^2}{\|d_k\|^2}.$$

By Assumption H(i), and combining this inequality, we have

$$\sum_{k\geqslant1}(g_k^{\mathrm T}d_k)^2/\|d_k\|^2<+\infty.$$

□

Lemma 3 *Suppose Assumption H holds. Let the sequences $\{g_k\}$ and $\{d_k\}$ be obtained by the SPRP method. If there exists a constant $r>0$ such that*

$$\|g_k\|\geqslant r,\qquad \forall k\geqslant1. \tag{18}$$

then we have

$$\sum_{k\geqslant1}\frac{\|d_{k-1}\|^2}{\|d_k\|^2}<+\infty. \tag{19}$$

Proof: From (17), we know that $d_k\neq0$, $\forall k\in N^+$. From (17), (12), (16), and (18), we have

$$\begin{aligned} \frac{\omega^2r^4}{\gamma^2}\sum_{k\geqslant1}\frac{\|d_{k-1}\|^2}{\|d_k\|^2} &= \omega^2\sum_{k\geqslant1}\frac{r^4\,\|d_{k-1}\|^2}{\gamma^2\,\|d_k\|^2} \\ &\leqslant \omega^2\sum_{k\geqslant1}\frac{\|g_k\|^4\cdot\|d_{k-1}\|^2}{\|d_k\|^2\cdot\|g_{k-1}\|^2}\leqslant\sum_{k\geqslant1}(g_k^{\mathrm T}d_k)^2/\|d_k\|^2 \end{aligned}$$

which is finite. □

Theorem 1 *Suppose Assumption H holds. Let the sequences $\{g_k\}$ and $\{d_k\}$ be obtained by the SPRP method. Then we obtain*

$$\liminf_{k\to+\infty}\|g_k\|=0. \tag{20}$$

Proof: Suppose that (20) does not hold. We have (18). Obviously, (19) also holds. From (9), (15), and (18), we get

$$\|\theta_k\|=\frac{\left|d_{k-1}^{\mathrm T}y_{k-1}\right|}{\|g_{k-1}\|^2}\leqslant\frac{\|d_{k-1}\|\cdot L\,\|s_{k-1}\|}{r^2}. \tag{21}$$

By (4), (15), (16), and (18), we also have

$$\left|\beta_k^{\mathrm{PRP}}\right|=\frac{\left|g_k^{\mathrm T}y_{k-1}\right|}{\|g_{k-1}\|^2}\leqslant\frac{\|g_k\|\cdot\|y_{k-1}\|}{\|g_{k-1}\|^2}$$

$$\leqslant \frac{\gamma \cdot L}{r^2} \cdot \|s_{k-1}\| = A\|s_{k-1}\|, \tag{22}$$

where $A = \gamma L/r^2$. From (8), (15), (21), and (22), we obtain

$$\begin{aligned}\|d_k\|^2 &= \|-\theta_k g_k + \beta_k \cdot d_{k-1}\|^2 \\ &\leqslant 2\theta_k^2 \|g_k\|^2 + 2{\beta_k}^2 \|d_{k-1}\|^2 \\ &\leqslant \frac{2\|d_{k-1}\|^2 \cdot L^2 \|s_{k-1}\|^2}{r^4} \cdot \gamma^2 \\ &\quad + 2A^2 \|s_{k-1}\|^2 \cdot \|d_{k-1}\|^2 \\ &\leqslant \left(\frac{2L^2D^2\gamma^2}{r^4} + 2A^2D^2\right)\|d_{k-1}\|^2 \\ &\leqslant \rho\,\|d_{k-1}\|^2,\end{aligned}$$

where $\rho = (2L^2D^2\gamma^2/r^4) + 2A^2D^2$, and D is the diameter of Ω. Then we have

$$\frac{\|d_{k-1}\|^2}{\|d_k\|^2} \geqslant \frac{1}{\rho}, \tag{23}$$

which means that $\|d_{k-1}\|^2 / \|d_k\|^2$ is not bounded and contradicts (19). Hence the assumption does not hold and the claim (20) is proved. □

NUMERICAL RESULTS

In this section, we compare the performance of the SPRP method with that of the PRP method and the scaled PRP method on a set of 640 unconstrained optimization problems under the standard Wolfe line search. From the CUTE library[9] and Ref. 10, we selected 64 large-scaled problems in extended or generalized form. Each problem is tested 10 times for a gradually increasing number of variables: $n = 1000, 2000, \dots 10000$. All codes were written in double precision FORTRAN and run on a PC with 2.0 GHz CPU and 512 MB memory under Windows XP.

All methods implement the standard Wolfe line search with $\sigma = 0.5$ and $\delta = 10^{-4}$, and the initial step-size α is computed from

$$\alpha = \begin{cases} 1, & k = 1, \\ \alpha_{k-1}\|d_{k-1}\|/\|d_k\|, & k \geqslant 2. \end{cases}$$

If $\|g_k\| \leqslant 10^{-6}\max\{1, |f(x_k)|\}$ is satisfied, we terminate the iteration; if this condition is not satisfied after 30 000 iterations, we terminate the iteration.

Let f_i^{M1} and f_i^{M2} be the optimal value found by the M1 and M2 methods for the ith problem, respectively. We say that for the particular the ith problem, the performance of M1 is better than the performance of M2 if $f_i^{\mathrm{M1}} < f_i^{\mathrm{M2}} + 10^{-3}$, and the CPU time, or the number of iterations of M1 was less than the number of iterations, or the CPU time of M2, respectively.

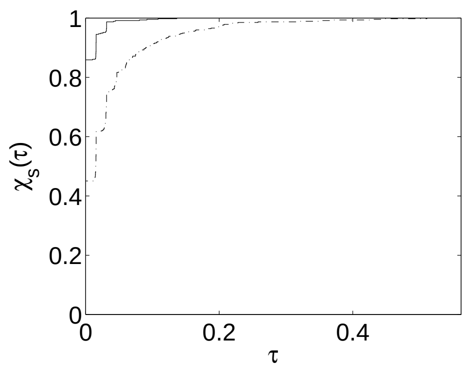

Fig. 1 SPRP method versus PRP method on CPU time. In this and remaining figures, solid line: SPRP method; dash-dot line: PRP method.

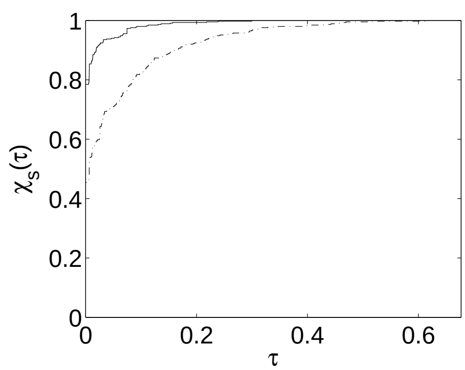

Fig. 2 SPRP method versus PRP method on the number of iterations.

In order to overall evaluate these methods in the CPU time, we also use the profiles of Dolan and Moré[11]. That is, the performance profiles with respect to CPU time mean that for each method we plot the fraction of problems for which the method is within a factor of the best time. The left side of the figure gives the percentage of the test problems out of 640 for which the method is the fastest; the right side gives the percentage of the test problems that are successfully solved by each of the methods. The top curve is the method that solved the most problems in a time that was within a factor of the best time.

From the profiles of Dolan and Moré[11], Figs. 1 and 2 show that the SPRP method is more efficient than the PRP method in terms of CPU time and the number of iterations. Fig. 3 shows that the

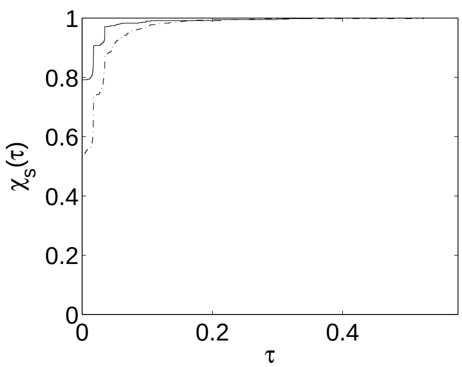

Fig. 3 SPRP method versus the scaled PRP method on CPU time.

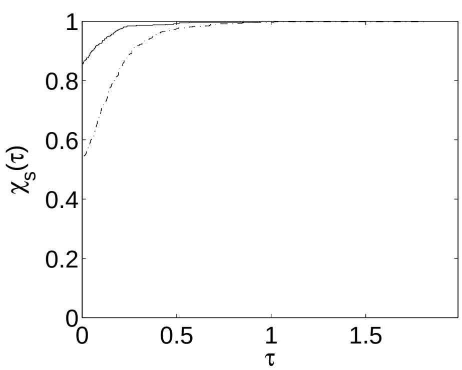

Fig. 4 SPRP method versus the scaled PRP method on the number of iterations.

SPRP method is similar to the scaled PRP method in terms of CPU time, and Fig. 4 indicates that the SPRP method is more efficient than the scaled PRP method in terms of the number of iterations.

CONCLUSIONS

In this paper we propose a new spectral PRP conjugate gradient method in which the direction d_k is computed from $d_k = -\theta_k g_k + \beta_k^{\mathrm{PRP}} d_{k-1}$. Applying the conjugacy condition, we obtain the spectral parameter θ_k as $\theta_k = d_{k-1}^{\mathrm{T}} y_{k-1} / \|g_{k-1}\|^2$. The new method overcomes the drawbacks of PRP and scaled PRP methods, and has stable descent and convergence properties. What is more, numerical results also show that the new method outperforms PRP method and the scaled PRP method. In view of the SPRP method's advantages, by applying the same technique to the LS method[12], we also consider a spectral LS conjugate gradient method in which the spectral parameter θ_k is computed from $\theta_k = d_{k-1}^{\mathrm{T}} y_{k-1} / d_{k-1}^{\mathrm{T}} g_{k-1}$.

Acknowledgements: The author wishes to express their heartfelt thanks to the referees and editors for their detailed and helpful suggestions for revising the manuscript.

REFERENCES

1. Hager WW, Zhang H (2006) A survey of nonlinear conjugate gradient methods. *Pac J Optim* **2**, 35–58.
2. Polak E, Ribière G (1969) Note sur la convergence de directions conjugées. *Rev Fran Informat Recherche Op* **16**, 35–43.
3. Polyak BT (1969) The conjugate gradient method in extremal problems. *USSR Comput Math Math Phys* **9**, 94–112.
4. Dai YH (1997) Analyses of conjugate gradient methods. PhD thesis, Institute of Computational Mathematics and Scientific/ Engineering Computing, Chinese Academy of Sciences [in Chinese].
5. Barzilai J, Borwein JM (1988) Two-point step size gradient methods. *IMA J Numer Anal* **8**, 141–8.
6. Raydan M (1997) The Barzilai and Borwein gradient method for the large unconstrained minimization problem. *SIAM J Optim* **7**, 26–33.
7. Birgin EG, Martínez JM (2001) A spectral conjugate gradient method for unconstrained optimization. *Appl Math Optim* **43**, 117–28.
8. Zoutendijk G (1970) Nonlinear programming, computational methods. In: Abadie J (ed) *Integer and Nonlinear Programming*, North-Holland, Amsterdam, pp 37–86.
9. Bongartz KE, Conn AR, Gould NIM, Toint PL (1995) CUTE: constrained and unconstrained testing environments. *ACM Trans Math Software* **1**, 123–60.
10. Andrei N (2008) An unconstrained optimization test functions collection. *Adv Model Optim* **10**, 147–61.
11. Dolan ED, Moré JJ (2002) Benchmarking optimization software with performance profiles. *Math Program* **91**, 201–13.
12. Liu Y, Story C (1992) Efficient generalized conjugate gradient algorithms. *J Optim Theor Appl* **69**, 129–37.