Menemui Matematik (Discovering Mathematics)
Vol. 39, No. 1: 22-29 (2017)

Multi-Spectral Gradient Method via Variational Technique under Log-Determinant Norm for Large-Scale Optimization

Hong Seng Sim[1,2], Wah June Leong[2], Chuei Yee Chen[2], and Siti Nur Iqmal Ibrahim[2]
[1]*School of Foundation Studies, Xiamen University Malaysia, 43900 Sepang, Selangor*
[2]*Institute for Mathematical Research, University Putra Malaysia, 43400 UPM Serdang, Selangor*

[1]hssim@xmu.edu.my [2]leongwj@upm.edu.my, cychen@upm.edu.my, iqmal@upm.edu.my

ABSTRACT

The spectral gradient method is popular due to the fact that only the gradient of the objective function is required at each iterate. Besides that, it is more efficient than the quasi-Newton method as the storage of second derivatives (Hessian) approximation are not required especially when the dimension of the problem is large. In this paper, we propose a spectral gradient method via variational technique under log-determinant measure such that it satisfies the weaker secant equation. The corresponding variational problem is solved and the Lagrange multiplier is approximated using the Newton-Raphson method and solved following interior point method that is associated with weaker secant relation. An executable code is developed to test the efficiency of the proposed method with some standard conjugate-gradient methods. Numerical results are presented which suggest a better performance has been achieved.

Keywords: Spectral gradient method, Variational technique, Log-determinant norm, Weak secant relation, Large-scale optimization.

INTRODUCTION

Steepest descent method is the most straightforward optimization tool used to solve large scale unconstrained optimization. However, the steepest descent method is relatively slow as it is closes to minimum. For ill-conditioned problems, the steepest descent directions are exhibiting 'zigzags' as the gradients point nearly orthogonally to the shortest direction to a minimum point. On the other hand, Quasi-Newton method was then introduced to overcome this deficiency. Its popularity is due to no actual Hessian is required for the algorithm but it still needs matrices storage for the Hessian approximation. La Cruz and Raydan (2003) and La Cruz et al (2006) extended the spectral approach to steepest descent direction for unconstrained nonlinear optimization. The main advantage of the spectral methods is that no second order information is needed for the search direction. Therefore, the Hessian is not required explicitly and a low computational cost is expected. On the other hand, a scaling parameter that incorporates certain second order information is used to scale the steepest descent direction. Spectral gradient methods are low-cost nonmonotone schemes for finding the local minimizers. They were first introduced by Barzilai and Borwein in 1988 and have been applied successfully to find local minimizers of large scale unconstrained problems.

Barzilai and Borwein (1988) developed the updating formula for σ_k based on the least square problem as follow:

$$\sigma_k := \arg\min_{\sigma} \frac{1}{2}\left\|\sigma s_k - y_k\right\|^2 \tag{1}$$

where $\sigma_k = \dfrac{s_k^T s_k}{s_k^T y_k}$ is the unique solution of (1).

They also considered

$$\sigma_k^{-1} := \arg\min_{\sigma^{-1}} \frac{1}{2}\left\| s_k - \sigma^{-1} y_k \right\|^2 \tag{2}$$

where $\sigma_k = \dfrac{s_k^T y_k}{y_k^T y_k}$.

Equations σ_k are called BB step sizes. From the solution of problem (1), it can be expressed in the form of

$$\sigma_k = \frac{g_k^T A g_k}{g_k^T g_k}. \tag{3}$$

and can be regarded as a Rayleigh quotient that calculated from the current gradient vector. In this paper, we intend to propose a strategy to obtain a set of spectral parameters, by incorporating the BB and quasi-Newton ideas.

SPECTRAL GRADIENT METHOD FOR CONVEX QUADRATIC MINIMIZATION

Consider the quadratic minimization problem:

$$\min_x f(x) = \frac{1}{2} x^T A x - b^T x,$$

where A is the Hessian matrix that is assumed to be positive definite and symmetry, and the gradient $g_k = Ax_k - b$.

Newton's method has an iterative formula in the form of

$$x_{k+1} = x_k + d_k, \tag{4}$$

where $d_k = -A^{-1} g_k$. Newton's method requires second order information that makes it converges extremely fast near the optimal solution and only one step is needed for quadratic function. However, forming and computing A^{-1} are costly and some modifications are needed when A^{-1} is not positive definite. Motivated from the above, we tend to choose σ_k so that $-\sigma_k^{-1} g_k = -(\sigma_k I)^{-1} g_k$ is used to approximate $-A^{-1} g_k$ in some sense. First, define $s_k = x_{k+1} - x_k$ and $y_k = g_{k+1} - g_k$. Then the matrix A satisfies the relation

$$A s_k = y_k. \tag{5}$$

Since it is inappropriate for a multiple of identity to satisfy (5), we will choose σ_k such that it satisfies some weaker form of (5).

For quadratic function mentioned above, we can assume without loss of generality that an orthogonal transformation is made that transform A to a diagonal matrix that contains only its eigenvalues λ_i. Besides that, if there are any eigenvalues of multiplicity $m > 1$, then we can choose the corresponding eigenvectors so that $g_1^{(i)} = 0$ for at least $m - 1$ corresponding indices

of g_1. Using $A = diag(\lambda_i)$, (4) and the properties of a quadratic function that $g_{k+1} = g_k - \frac{Ag_k}{\sigma_k}$, we have

$$g_{k+1}^{(i)} = \left(1 - \frac{\lambda_i}{\sigma_k}\right) g_k^{(i)}, \quad i = 1,2,3, \quad ,n. \tag{6}$$

It is clear that if $g_k^{(i)} = 0$ for any i and $k := \hat{k}$, then this property will persist for all $k > \hat{k}$. Thus, without any loss of generality, we can assume that A has distinct eigenvalues

$$0 < \lambda_1 < \lambda_2 < \quad < \lambda_n, \tag{7}$$

and that $g_1^{(i)} \neq 0$ for all $i = 1,2,3, \quad ,n$.

From these conditions and (6), we can deduce that first if σ_k is equal to any eigenvalue λ_i, then $g_{k+1}^{(i)} = 0$ and this property persists subsequently. If both

$$g_{k-1}^{(i)} \neq 0 \text{ and } g_{k-1}^{(n)} \neq 0, \tag{8}$$

then from (3) and the external properties of the Rayleigh quotient that

$$\lambda_1 < \sigma_k < \lambda_n. \tag{9}$$

Hence, assuming that σ_1 is not equal to λ_1 or λ_n, then for BB method, a simple inductive argument shows that (8) and (9) hold for all $k > 1$. It also follows that the BB method does not have the property of finite termination. Since the eigenvalues are distinct, it is reasonable to use a set of different σ, so that they can better overlap the spectrum of A. Hence, this motivates us to choose a diagonal matrix, $diag\left(\sigma_k^{(i)}\right)$ to approximate $diag\left(\lambda_k^{(i)}\right)$.

Now, we extend the quadratic optimization problem to nonquadratic unconstrained large scale optimization problem as below:

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

where f is twice continuously differentiable, while n is assumed to be large, says greater than 10000. In order to preserve the descent property, we incorporate the line search to the iterative method in (4) to yield

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

where $\alpha_k > 0$ is a steplength, calculated to satisfy certain line search conditions, such as the Armijo condition. A steplength α_k is said to satisfy the Armijio condition if the following inequality hold:

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c\alpha_k p_k^T \nabla f(x_k), \tag{10}$$

where $0 < c < 1$.

There are various choices on the search direction d_k. In this paper, however, we will only focus on the spectral gradient method where the search direction d_k is given by $d_k = -B_k^{-1} g_k$, where matrix B_k is updated at every iteration and the sequence of matrices $\{B_k\}$ is required to satisfy some weaker form of secant equation.

Besides the weighted Frobenius norm, Byrd and Nocedal (1989) simplified proofs the convergence for BFGS update by working simultaneously with the trace and determinant of B_k. For this purpose, they defined, for any positive definite matrix B, the function, $\varphi(B) = tr(B) - \ln(\det(B))$, defines a measure of matrices, where $\ln$ denotes the natural logarithm.

Motivated by this measure, we propose to derive some spectral gradient type updating such that it satisfied the weaker secant relation defined by Dennis and Wolkowicz in 1993 as follow:

$$s_k^T B_{k+1} s_k = s_k^T y_k.$$

FORMULATION FOR MULTI-SPECTRAL PARAMETERS

Suppose the updating formula B_{k+1} is diagonal and positive definite. Our purpose is to construct B_{k+1} such that it satisfies

$$s_k^T B_{k+1} s_k = s_k^T y_k,$$

where $y_k = g_{k+1} - g_k$ and $s_k = x_{k+1} - x_k$.
Hence, we consider the following:

$$\begin{aligned} &\min tr\left(B_{k+1}\right) - \ln\left(\det\left(B_{k+1}\right)\right) \\ &s.t.\ s_k^T B_{k+1} s_k = s_k^T y_k \end{aligned}. \tag{11}$$

Let $B_{k+1} = diag\left(r_k^{(1)}, r_k^{(2)}, \quad , r_k^{(n)}\right)$. Then, the minimization becomes

$$\begin{aligned} &\min\left(\sum_{i=1}^{n} r_k^{(i)}\right) - \ln\left(\prod_{i=1}^{n} r_k^{(i)}\right) \\ &s.t.\ \sum_{i=1}^{n}\left(s_k^{(i)}\right)^2 r_k^{(i)} - \omega = 0 \end{aligned}. \tag{12}$$

where $\omega = \sum_{i=1}^{n} s_k^{(i)} y_k^{(i)}$.

The Lagrangian formed from equation (12) will become

$$L = \left(r_k^{(1)} + r_k^{(2)} + \quad + r_k^{(n)}\right) - \ln\left(r_k^{(1)} r_k^{(2)} \quad r_k^{(n)}\right) - \lambda\left(\sum_{i=1}^{n}\left(s_k^{(i)}\right)^2 r_k^{(i)} - \omega\right) \tag{13}$$

In order to get the minimizer, we differentiate equation (13) with respect to $r_k^{(1)}, r_k^{(2)}, \quad , r_k^{(n)}$ and setting the resulting derivation to 0:

$$\frac{\delta L}{\delta r_k^{(i)}} = 1 - \frac{1}{r_k^{(i)}} + \lambda\left(s_k^{(i)}\right)^2 = 0, \tag{14}$$

which yields

$$r_k^{(i)} = \frac{1}{1 + \lambda\left(s_k^{(i)}\right)^2}. \tag{15}$$

Now, by substituting equation (15) into the constraint (11), we have

$$F(\lambda) = \sum_{i=1}^{n} \frac{\left(s_k^{(i)}\right)^2}{1 + \lambda\left(s_k^{(i)}\right)^2} - \sum_{i=1}^{n} s_k^{(i)} y_k^{(i)}.$$

Thus, λ can be obtained by solving the nonlinear equation $F(\lambda) = 0$. It is not practical to solve the equation accurately, hence, we would approximate the solution by using only one Newton-Raphson iteration from $\overline{\lambda} = 0$ and interior point method. Hence, the Lagrange multiplier, λ_k is approximated by

$$\lambda_k = \frac{\sum_{i=1}^{n}\left(s_k^{(i)}\right)^2 - \sum_{i=1}^{n} s_k^{(i)} y_k^{(i)}}{\sum_{i=1}^{n}\left(s_k^{(i)}\right)^4}.$$

Finally, we obtain the updating formula for B_{k+1} as follows:

$$B_{k+1} = diag\left(r_k^{(i)}\right), \tag{16}$$

where

$$r_k^{(i)} = \frac{1}{1 + \frac{\sum_{i=1}^{n}\left(s_k^{(i)}\right)^2 - \sum_{i=1}^{n} s_k^{(i)} y_k^{(i)}}{\sum_{i=1}^{n}\left(s_k^{(i)}\right)^4}\left(s_k^{(i)}\right)^2}.$$

The algorithm for solving the optimization problems is the same as the spectral gradient algorithm, the only different is that instead of using θI as the updating formula, we use B_{k+1}. Now, we provide our algorithm for solving large scale unconstrained nonlinear optimization problem.

DS Algorithm:

Step 1: Set $k = 0$; select the initial guess points x_0 and ξ as a stopping condition. Set $B_0 = I$ where I is the $n \times n$ identity matrix.

Step 2: For $k \geq 0$, compute $f(x_k)$ and $g(x_k)$. If $\|g(x_k)\| \leq \xi$, stop, else compute B_{k+1} where B_{k+1} is defined as follows:

$$B_{k+1} = \begin{cases} B_{k+1}, & if\ B_{k+1} > 0 \\ B_k, & if\ B_{k+1} \leq 0 \end{cases},$$

where B_{k+1} is defined by equation (16).

Step 3: Compute $d_k = -B_k^{-1} g_k$.

Step 4: Compute the steplength α_k such that it satisfies the Armijo condition.

Step 5: Compute $x_{k+1} = x_k + \alpha_k d_k$.

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

NUMERICAL RESULTS AND DISCUSSION

In this paper, we compared our proposed methods with a list of conjugate gradient methods as follow:

(1). DS-NR – diagonal spectral gradient method with λ is approximated by Newton-Raphson method.

(2). DS-IP – diagonal spectral gradient method with λ is approximated by interior point method.

(3). CG-PR – conjugate gradient method proposed by Polak & Ribiere (1969) and Polyak (1969).

(4). CG-CD – conjugate gradient conjugate descent method proposed by Fletcher (1987).

(3). CG-LS – conjugate gradient method proposed by Liu and Storey (1991).

(3). CG-HZ – conjugate gradient method proposed by Hager and Zhang (2005).

In order to test the efficiency of our proposed method, a set of 96 tested problems given by CUTE, presented in Andrei (2008) has been tested with dimensions varying from 10 to 100000. we compared the methods in terms of the number of iterations, function calls and CPU times with the list of conjugate gradient method. Default values are used for all the other parameters, and the stopping criterion is set to be

$$\|g_k\| \leq 10^{-4}.$$

We also set our upper bound of the iterations to be 10000. Therefore, whenever the number of iterations exceed the upper bound, we declare that this run as a failure. The codes are written in Matlab software.

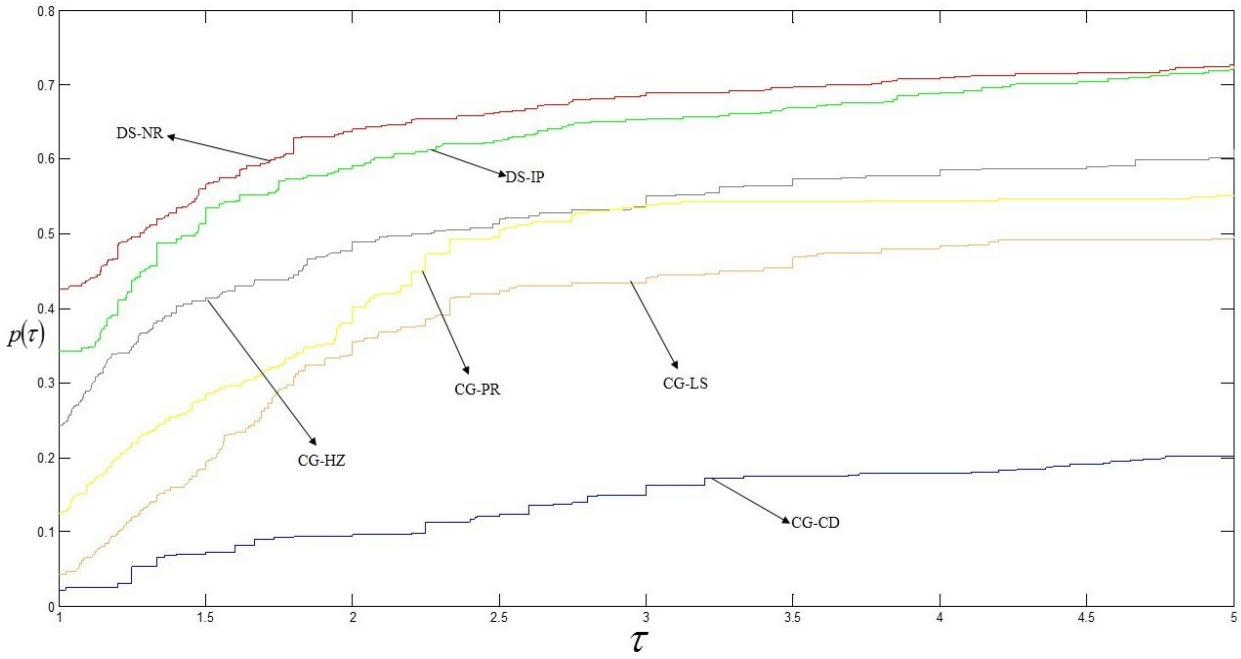

Figure 1: Performance profiling for the proposed methods and the list of CG methods in terms of number of iterations.

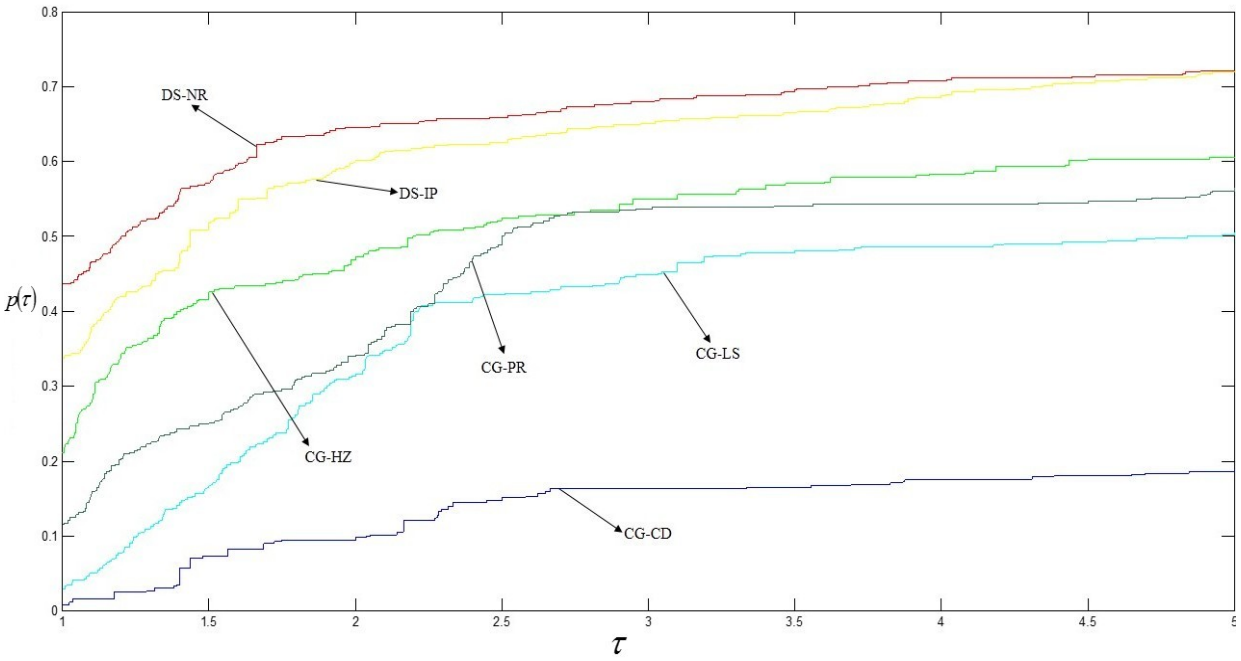

Figure 2: Performance profiling for the proposed methods and the list of CG methods in terms of function calls.

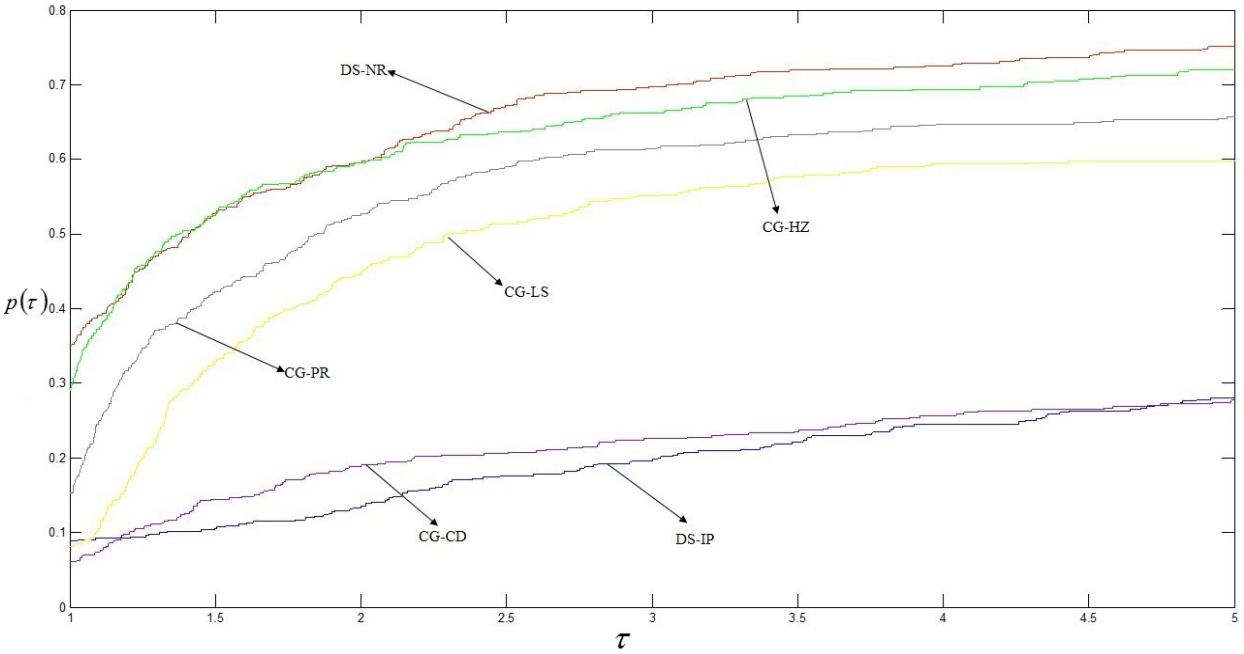

Figure 3: Performance profiling for the proposed methods and the list of CG methods in terms of CPU time.

Figure 1 shows that the profiling graph of our proposed methods and the list of conjugate gradient methods in terms of number of iterations. From figure 1, it is obvious that our proposed methods DS-NR and DS-IP perform better than the conjugate gradient methods. In other words, our proposed methods required less iterations to get the desired minimum points.

Figure 2 shows that the profiling graph of our proposed methods and the list of conjugate gradient methods in terms of number of function calls. Therefore, our proposed methods DS-NR and DS-IP confirm that the number of function calls are lesser.

Figure 3 shows that the profiling graph of our proposed methods and the list of conjugate gradient methods in terms of CPU time. From this figure, the graph of DS-NR appeared to be higher than the list of conjugate gradient methods. Therefore, DS-NR needs a shorter time as compared to other conjugate gradient methods. However, our DS-IP needs a longer time to execute and get the results. This is mainly because of the optimal value of Lagrange multiplier λ need to be computed using Matlab solver.

CONCLUSION

From the profiling graphs above, we can see that our proposed methods perform better when compared to the list of conjugate gradient methods. Thus, it is essential that satisfy the weak secant relation when deriving the updating formula for the diagonal matrix B_k. In solving large scale optimization problems, computational times and costs are the main concern among all the other factors. Our proposed methods require only $O(n)$ storage to get the optimal solutions. Hence, it is much more desirable to be the alternative methods than the quasi-Newton methods that required $O(n^2)$ storage. The DS-NR shows an outstanding result in terms of number of iterations, number of function calls and computational times. DS-IP also giving a similar trend as DS-NR except for computational times. DS-IP required more times to obtain the desired results

due to the optimal value of Lagrange multiplier λ need to be executed using the built-in solver in Matlab.

In conclusion, our proposed methods outperform the list of the conjugate gradient methods not only in term of number of iterations but also number of function calls. Thus, our proposed methods can be an economical alternative for solving large-scale problems.

ACKNOWLEDGMENT

The work was supported by the Malaysia FRGS grant (FRGS/2/2013/ST06/UPM/02/1) and the first author would like to acknowledge Yayasan Sultan Iskandar Johor for the financial support.

REFERENCES

B. T. Polyak, (1969), The Conjugate Gradient in Extreme Problems, *USSR Comp. Math. Phys.,* **9**: 94–112.

E. Polak, and G. Ribiere, (1969), Sur La Convergence De Directions Conjugees, *Rev. Francaise Informat Recherche Opertionelle, 3e Annee* 16: 35–43.

J. Barzilai and J. M. Borwein, (1988), Two Point Stepsize Gradient Methods, *IMA J. Numer. Anal.,* **8:** 141–148.

J. Nocedal and D. C. Liu, (1989), On the Limited Memory Method for Large Scale Optimization, *Math. Program. Ser. B,* **45:** 503-528.

J. W. Daniel, (1967), The Conjugate Gradient Method for Linear and Nonlinear Operator Equations, *J. Numer. Anal*., **4:** 10-26.

M. R. Hestenes and E. L. Stiefel, (1952), Methods of Conjugate Gradient for Solving Linear Systems, *J. Research Nat. Bur. Standards.,* **49:** 409-436.

R. Fletcher, (1987), Practical Methods of Optimization Vol 1: Unconstrained Optimization, *John Wiley and Sons, New York.*

R. Fletcher, (2005), On the Barzilai Borwein Method, *Optimization and Control with Applications,* **96:** 235-256.

R. Fletcher and C. Reeves, (1964), Function Minimization by Conjugate Gradients, *Comput. J.,* **7:** 149-154.

R. H. Byrd, J. Nocedal and Y. Yuan, (1987), Global Convergence of a Class of Quasi-Newton Methods on Convex Problems, *SIAM J. Numer. Anal.,* **24:** 1171-1189.

Raydan M., (1997), The Barzilai and Borwein Gradient Method for the Large Unconstrained Minimization Problems, *SIAM Journal on Optimization,* **7:** 26-33.

W. La Cruz, J. M. Martinez, M. Raydan, (2006), Spectral Residual Method Without Gradient Information for Solving Large-Scale Nonlinear Systems of Equations, *Math. Comput.,* **75:** 1449-1466.

W. La Cruz and M. Raydan, (2003), Nonmonotone Spectral Methods for Large-Scale Nonlinear Systems, *Optim. Methods Softw.,* **18:** 583-599.

W. W. Hager and H. Zhang, (2005), A New Conjugate Gradient Method with Guaranteed Descent and an Effifient Line Search, *SIAM J. Optim.,* **16(1):** 170-192.

Y. H. Dai and Y. Yuan, (1999), A Nonlinear Conjugate Gradient Method with a Strong Global Convergence Property, *SIAM J. Optim.,* **10:** 177-182.

Y. Liu and C. Storey, (1991), Efficient Generalized Conjugate Gradient Algorithms, Part 1: Theory, *J. Optim. Theory Appl.,* **69:** 129-137.