A NEW DIAGONAL GRADIENT-TYPE METHOD FOR LARGE SCALE UNCONSTRAINED OPTIMIZATION

Mahboubeh Farid[1], Wah June Leong[2] and Lihong Zheng[2]

The main focus of this paper is to derive new diagonal updating scheme via the direct weak secant equation. This new scheme allows us to improve the accuracy of the Hessian's approximation and is also capable to utilize information gathered about the function in previous iterations. It follows by an scaling approach that employs scaling parameter based upon the proposed weak secant equation to guarantee the positive definiteness of the Hessian's approximation. Moreover, we also prove the convergence of the proposed method under a simple monotone strategy. Numerical results show that the method is promising and frequently outperforms its competitors.

Keywords: diagonal updating, weak secant equation, global convergence, large scale problem, unconstrained optimization

MSC2010: 65K10, 90C06, 90C52, 90C30

1. Introduction

To minimize a continuously differentiable function f without constraints,

$$\min f(x), \quad x \in R^n, \tag{1}$$

Barzilai and Borwein method [2] generates the sequence x_k according to the iterative scheme:

$$x_{k+1} = x_k - B_k^{-1} g_k, \tag{2}$$

where $g_k = \nabla f(x_k)$ and $B_k^{-1} = \alpha_k I$. Here, α_k is a stepsize decided by the information obtained at points x_k and x_{k-1}. The two choice of the scalar α_k are given as follows:

$$\alpha_k^{(1)} = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}} \tag{3}$$

and

$$\alpha_k^{(2)} = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}} \tag{4}$$

[1] Department of Mathematics, University Putra Malaysia, 43400 Serdang, Selangor, Malaysia, E-mail: mfarid7@gmail.com

[2] Department of Mathematics, University Putra Malaysia, 43400 Serdang, Selangor, Malaysia, E-mail: wjleong@science.upm.edu.my

[3] School of Computing and Maths, Charles Sturt University, Australia, E-mail: lzheng@csu.edu.au

where $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = g_k - g_{k-1}$. These alternative choices for are related to the quasi-Newton equation (also called the secant equation)

$$B_k s_{k-1} = y_{k-1}, \tag{5}$$

where B_k is an $n \times n$ symmetric positive definite matrix approximating the Hessian matrix $\nabla^2 f(x_k)$. The BB method has received a great deal of studies because it provides an effective and very useful stepsize adaption procedure for unconstrained optimization [4],[10],[15]. However, there are several disadvantages of the BB method. The method does not guarantee a descent in the objective function at each iteration, and the extent of the non-monotonicity depends in some way on the condition of the Hessian matrix. Motivated by these shortcomings, a variant of spectral gradient methods, called diagonal gradient-type method are developed by [8],[9],[11],[12]. This approach replaces $\alpha_k I$ with the diagonal matrix B_k^{-1}, and the spectral information on the B_k is corrected via a weaker form of the quasi-Newton equation (5). In general, the approach consists in finding an updated Hessian approximation B_k, which is restricted to be a diagonal matrix, obeys the weak secant equation of Dennis and Wolkowicz [6], namely

$$s_{k-1}^T B_k s_{k-1} = s_{k-1}^T y_{k-1}, \tag{6}$$

and simultaneously preserves as much information as possible from the current approximation B_{k-1}, which is assumed to be diagonal. Here, the spectral information which is characteristically used in determining B_{k+1} is contained only in vectors s_k and y_k, and does not use information of function values of the objective function. Thus, it is reasonable to construct a new weak secant equation that incorporates information on function values for approximating the curvature. Moreover, one can view the weak secant equation as a projection of the quasi-Newton equation (5) in a direction υ such that $\upsilon^T B_k s_{k-1} = \upsilon^T y_k \neq 0$. It seems that the choice of υ may influent the quality of the curvature information.

To avoid these obstacles, the approach proposed in this paper is based on defining the weak secant equation by interpolation rather than deriving from the secant equation. Moreover, it uses information of two successive function values for approximating the curvature information in higher accuracy. Along this line, a new diagonal updating formula is proposed. The structure of the paper is as follows: in Section 2 we describe the new diagonal updating and details of the proposed algorithm. Section 3 deals with the global convergence of the algorithm and Section 4 presents the result of computational experiments. Finally, Section 5 concludes the paper.

2. Derivation of New Diagonal Updating

Many of the quasi-Newton methods accumulate Hessian information based on the secant equation (5). However, since it is usually difficult to satisfy the secant equation with a nonsingular matrix of the diagonal form, we need some alternatives that can ensure the accumulated curvature information along the step is correct. The alternative we look for is the non-secant updating strategy proposed by Dennis and Wolkowicz [6]. Dennis and Wolkowicz introduced a weaker form of secant equation by projecting the secant equation (5) in a direction υ such that $y_k^T \upsilon \neq 0$ to give

$$\upsilon^T B_k s_{k-1} = \upsilon^T y_k. \tag{7}$$

Particularly, Dennis and Wolkowicz considered $\upsilon = s_{k-1}$, which leads to (6). Under this weak-secant equation, Zhu et al. [16], Hassan et al., [11], and Leong et al. [12] employ independently, a variational technique that is analogue to the one used to derive the Powell Symmetric Broyden (PSB) quasi-Newton update (see, for example Dennis and Schnabel [5]) for approximating the Hessian matrix diagonally. Despite some promising numerical results, it remains unknown on the appropriateness of the choice of υ and one can expect that nontrivial computational experience is required to determine such direction. Moreover, relation (6) does not use information of function values of the objective function, which may be essential to interpolate the curvature information correctly.

Motivated by these drawbacks, we propose an approach that defines a new weak secant equation as view of interpolation rather than deriving from secant equation. The general idea of our approach is given as follows:
Quasi-Newton methods use a local quadratic model of the form

$$f(x_k + s) \approx \phi_k(s) = f(x_k) + g_k^T s + \frac{1}{2} s^T G_k s \tag{8}$$

where G_k is the true Hessian at x_k. Thus, the curvature information carried in $s^T G_k s$ of (11) can be approximated by

$$s^T G_k s \approx 2(f(x_k + s) - f(x_k) - g_k^T s). \tag{9}$$

Since the updated B_{k+1} is supposed to approximate G_k, it is reasonable to having

$$s_k^T B_{k+1} s_k = 2(f_{k+1} - f_k + g_k^T s_k) \tag{10}$$

In fact, this new weak secant equation (10) is superior to the one defined by (6) in the sense that it gives lower error in approximating the curvature information. We shall give some details on this claim. Let us consider the Taylor expansion of $f(x_k + s_k)$ about x_k, and its derivative, respectively:

$$\begin{aligned} f(x_k + s_k) &= f(x_k) + g_k^T s_k + \frac{1}{2} s_k^T G_k s_k + \frac{1}{6} T_k \otimes s_k^3 + O(\|s_k\|^4); & (11) \\ g(x_k + s_k) &= g(x_k) + G(x_k) s_k + \frac{1}{2} T_k \otimes s_k^2 + O(\|s_k\|^3), & (12) \end{aligned}$$

where $T_k \in R^{n \times n \times n}$ is the third order derivative tensor of f at x_k and $\otimes$ is some appropriate tensor product. After multiplying (12) by s_k and using the fact that, $y_k = g(x_k + s_k) - g(x_k)$ we have

$$s_k^T y_k = s_k^T G(x_k) s_k + \frac{1}{2} T_k \otimes s_k^3 + O(\|s_k\|^4). \tag{13}$$

If B_{k+1} and $\hat{B}_{k+1}$ is the Hessian approximation that based upon (10) and (6) (with index k being replaced by $k+1$), respectively. Then we obtain

$$|s_k^T B_{k+1} s_k - s_k^T G_k s_k| = \frac{1}{3}|T_k \otimes s_k^3| + O(\|s_k\|^4); \tag{14}$$

$$|s_k^T \hat{B}_{k+1} s_k - s_k^T G_k s_k| = \frac{1}{2}|T_k \otimes s_k^3| + O(\|s_k\|^4), \tag{15}$$

and these imply that using (10) will eventually give a lower error in the approximation.

Hence, by using (10) we shall construct the new updating formula for diagonal approximation of Hessian accordingly. The updating formula that we are looking for is derived based upon the least change updating strategy analogue to that of Leong et al. [12], i.e. the solution of the following problem:

$$\begin{gathered} \min \frac{1}{2} \|B_{k+1} - B_k\|_F^2 , \\ s.t \quad s_k^T B_{k+1} s_k = 2(f_{k+1} - f_k + g_k^T s_k), \\ \text{and } B_{k+1} \text{ is a diagonal matrix} \end{gathered} \tag{16}$$

where $\|.\|_F$ denotes the Frobonius norm. Using the procedure similar to that of Leong et al. [12], the updating formula for B_{k+1} will be generated by the following:

$$B_{k+1} = B_k + \left(\frac{2(f_{k+1} - f_k + g_k^T s_k) - s_k^T B_k s_k}{tr(E_k^2)} \right) E_k \tag{17}$$

where $E_k = diag((s_k^{(1)})^2, (s_k^{(2)})^2, ..., (s_k^{(n)})^2)$ and $s_k^{(i)}$ is the ith component of vector s_k. To safeguard on the possibility of having non-positive-definite updating matrix, we define a scaling for (17) such that B_{k+1} is forced to be positive definite. Note that we will have $B_{k+1} \succ 0$ if the following condition holds:

$$2(f_{k+1} - f_k + g_k^T s_k) - s_k^T B_k s_k > 0. \tag{18}$$

Then, one can employ a scaling β_k such that

$$\beta_k = \min(\rho_k, 1) \tag{19}$$

where

$$\rho_k = \frac{2(f_{k+1} - f_k + g_k^T s_k)}{s_k^T B_k s_k}. \tag{20}$$

We can immediately see that by incorporating such scaling to B_k, before using it to update B_{k+1}, we can guarantee the positive definiteness of B_{k+1} [13]. Accordingly, our updating formula will be as follow:

$$B_{k+1} = \beta_k B_k + \left(\frac{2(f_{k+1} - f_k + g_k^T s_k) - \beta_k s_k^T B_k s_k}{tr(E_k^2)} \right) E_k. \tag{21}$$

We can now state the detailed algorithm corresponding to the updating formula (21) under the monotone strategy of [11].

TSDG Algorithm

Step 0. Choose an initial point $x_0 \in R^n$, and a positive definite symetric matrix $B_0 = I$. Set $k = 0$.

Step 1. Compute g_k. If $\|g_k\| \leq \epsilon$, stop.

Step 2. Set $x_{k+1} = x_k - B_k^{-1} g_k$. Calculate ρ_k, β_k, B_{k+1} by (20),(19),(21), respectively.

Step 3. If $b_{k+1}^M b_k^M > 2(b_k^m)^2$ Set $B_{k+1} = \tau I$ where $\tau = \min\left(\frac{b_k^M}{2(b_k^m)^2}, \frac{s_k^T y_k}{s_k^T s_k}\right)$ in Step 2 with b_k^m, b_k^M, b_{k+1}^M be the smallest and largest diagonal component of B_k and B_{k+1}, respectively.

Step 4. Set $k := k + 1$ and go to Step 2.

3. **Convergence Analysis**

This section is devoted to study the convergence behavior of TSDG method. We shall establish the convergence of the TSDG algorithm when applied to the minimization of a strictly convex quadratic function with constant Hessian.

Theorem 3.1. *Assume that $f(x)$ is a strictly convex quadratic function. Let $\{x_k\}$ be a sequence generated by the TSDG method and x^* is a unique minimizer of f. Then either $g_k = 0$ holds for some finite $k \geq 1$, or $\lim_{k\to\infty} \|g_k\| = 0$.*

Proof. Denote $G = \nabla^2 f$. Let b_k^m, b_k^M, b_{k+1}^m and b_{k+1}^M be the smallest and largest diagonal elements of B_k and B_{k+1}, respectively where B_{k+1} is obtained is step 4 of TSDG Algorithm. Consider the Taylor expansion of the strictly convex function, f at x_{k+1} :

$$f(x_k - B_k^{-1} g_k) = f(x_k) - g_k^T B_k^{-1} g_k + \frac{1}{2} g_k^T B_k^{-1} G B_k^{-1} g_k. \tag{22}$$

Since $Gs_k = y_k$, it follows that $s_k^T G s_k = g_k^T B_k^{-1} B_{k+1} B_k^{-1} g_k$.Thus

$$f(x_{k+1}) \leq f(x_k) - c\,\|g_k\|^2, \tag{23}$$

where $c = (b_k^M)^{-1} - \frac{(b_k^m)^{-2} b_{k+1}^M}{2} > 0$. If $c > 0$, we have $f(x_{k+1}) \leq f(x_k)$ for all k. Else if $c < 0$, then we let $B_{k+1} = \vartheta I$ where $\vartheta = \min(\frac{b_k^M}{2(b_k^m)^2}, \frac{s_k^T y_k}{s_k^T s_k})$. Hence (23) becomes

$$f(x_{k+1}) \leq f(x_k) - \bar{c}\,\|g_k\|^2,$$

where $\bar{c} = b_k^m - \left((b_k^M)^2 \vartheta\right)/2$. With our choice of ϑ, we have that $\bar{c} \geq 0$. This implies that $f(x_{k+1}) \leq f(x_k)$ for all k and since f is bounded below, it follows that

$$\lim_{k\to\infty} f(x_k) - f(x_{k+1}) = 0.$$

As $f(x_k) - f(x_{k+1}) \to 0$, and $c > 0$ then $\lim_{k\to\infty} \|g_k\| = 0$, i.e. x_k converges to x^*. □

Table 1. Test problem and its dimension

Problem	Dimension
Diagonal 5, Extended Himmelblau, Generalized Rosenbrock, Generalized PSC1, Extended PSC1,Generalized Tridiagonal 1, Extended three Exponential terms, Generalized Tridiagonal 2, Broydan Tridiagonal, Extended Block Diagonal BD1, Extended Freudenstein and Roth, Extended Trigonometric, Extended Beale, Quadratic Diagonal Perturbed, Quadratic QF2, Extended Tridiagonal 2, Penalty 1, Penalty 2, Diagonal 4, Full Hessian FH1, Raydan 2,	10,100,1000,10000
Perturbed Quadratic,Raydan 1, Diagonal 1, Diagonal 2,Diagonal 3, Hager, EG2, Almost perturbed Quadratic, Quadratic QF1	10,100,1000

4. Numerical Results

In this section we analysis the effectiveness of TSDG algorithm and compare it to the BB and MDGRAD [11] method. The algorithms are coded in Matlab 7 and executed by a PC with Core Duo CPU. For all runs in our numerical experience, the iteration counts are limited as 1000. In addition, the algorithms are stopped if the maximum norm of the final gradient is below 10^{-4}, that is

$$\|g(x_k)\| \leq 10^{-4}.$$

We solved 30 problems where the name and dimensions of these tested problems are listed in Table 1.

Figure 1. Performance profile based on Iteration for all problems.

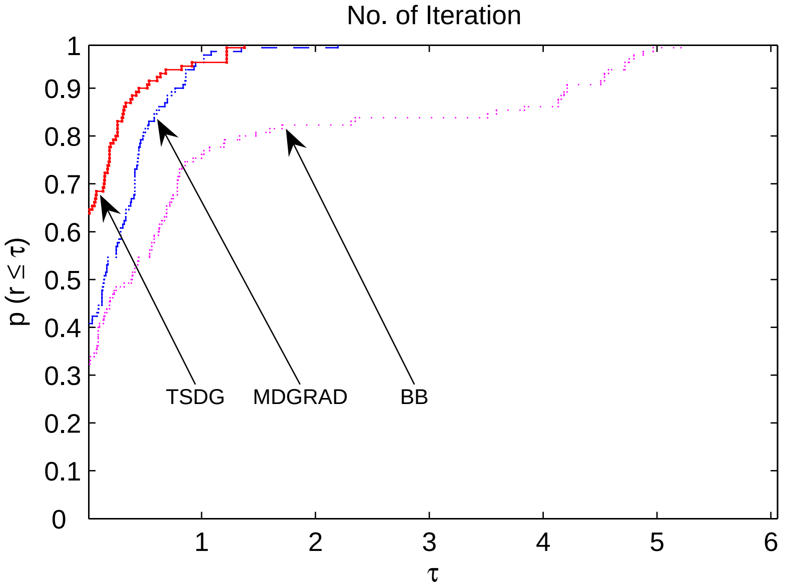

Figure 1 and 2 report the performance profiles of the TSDG, BB and MDGRAD algorithms. These profile graphs compare the number of iteration counts and

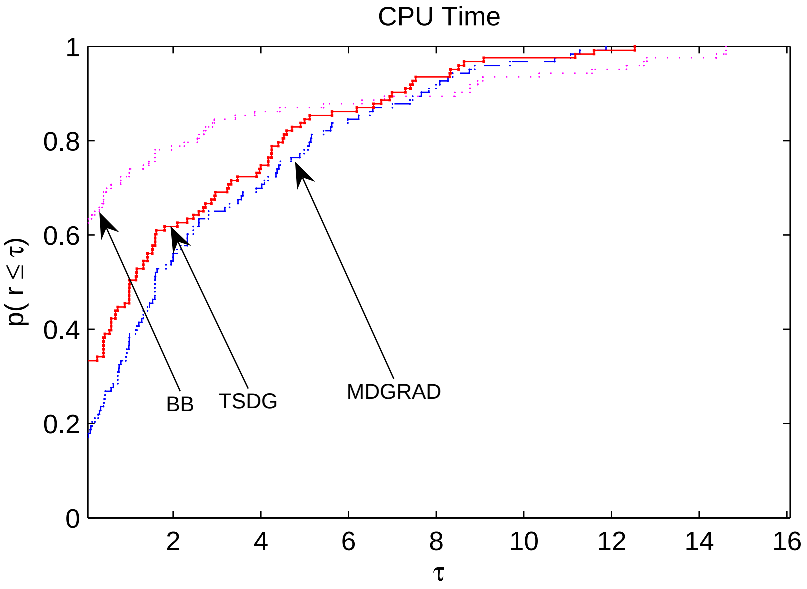

FIGURE 2. Performance profile based on CPU time.

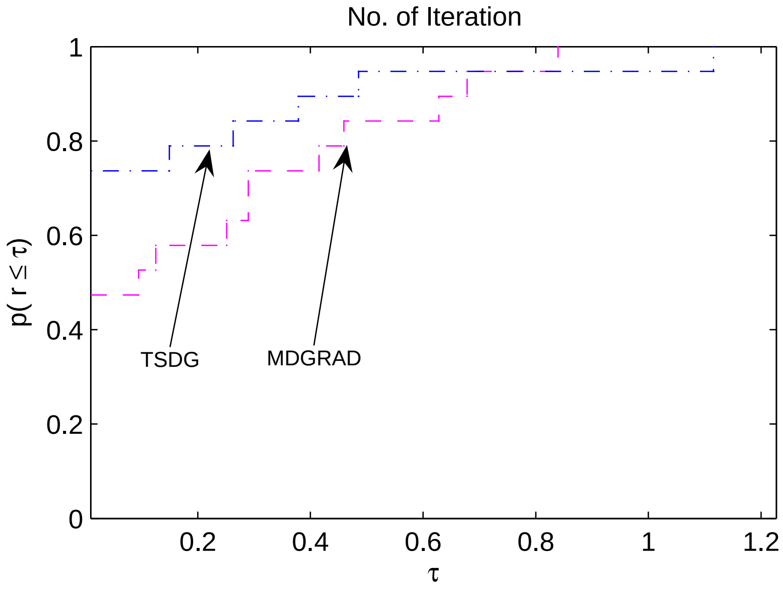

FIGURE 3. Performance profile based on iteration for large scale problems ($n = 10000$).

the computation time of the runs. It can be seen from Fig. 1 that TSDG algorithm is superior to the BB and MDGRAD methods in general. Moreover, all of the algorithms employ a monotone search strategy that only uses one function and gradient evaluation per iteration while TSDG algorithm uses an additional function value in approximating the Hessian diagonally. The additional function value requires only a unit of storage but the overall improvement is worthy.

5. Conclusion

The main contribution of the paper is in proposing a new derivation for weak secant equation. Through this new relation, we have presented a new gradient-type method that estimates Hessian matrix by a diagonal matrix. Our scheme is simple and able to enhance the performance of the gradient-type methods with minimal storage.

REFERENCES

[1] N. Andrei, An unconstrained optimization test functions collection, J. Adv. Model Optim. 10 (2008) 147-161.

[2] J. Barzilai and J.M. Borwein, Two point step size gradient methods, IMA J. Numer. Anal. 8 (1988) 141-148.

[3] Y.H. Dai, J.Y. Yuan and Y. Yuan, Modified two-point stepsize gradient methods for unconstrained optimization, J. Comput. Optim. Appl. 22 (2002) 103-109.

[4] Y.H. Dai and Y. Yuan, Alternative minimization gradient method, IMA J. Numer. Anal. 23 (2003) 373-393.

[5] J.E. Dennis and R.B. Schnabel, Numerical methods for unconstrained optimization and nonlinear equations, Prentice-Hall, Englewood Cliffs, New Jersey, 1983.

[6] J.E. Dennis and H.Wolkowicz, Sizing and least change secant method, SIAM J. Numer. Anal. 30 (1993) 1291-1313.

[7] E.D. Dolan and J.J. More, Benchmarking optimization software with perpormance profiles, Math. Program. 91 (2002) 201-213.

[8] M. Farid, W.J. Leong, M.A. Hassan, A new two-step gradient method for large-scale unconstrained optimization, Comput. Math. Appl. 59 (2010) 3301-3307.

[9] M. Farid, W.J. Leong, An improved multi-step gradient-type method for large-scale unconstrained optimization, Comput. Math. Appl. 61(2011) 3312-3318.

[10] R. Fletcher, On the Barzilai-Borwein method, Research Report NA/207, University of Dundee, UK, 2001.

[11] M.A. Hassan, W.J. Leong, M. Farid, A new gradient method via quasi-Cauchy relation which guarantees descent, J. Comput. Appl. Math. 230 (2009) 300-305.

[12] W.J. Leong, M.A. Hassan, M. Farid, A monotone gradient method via weak secant equation for unconstrained optimization, Taiwanese J. Math. 14(2) (2010) 413-423.

[13] W.J. Leong, M. Farid, M.A. Hassan, Scaling on diagonal quasi-Newton update for large-scale unconstrained optimization, B. Malays Math Sci So. (2)35(2) (2012)247-256.

[14] J.J. More$'$, B.S. Garbow and K.E. Hillstorm, Testing unconstrained optimization software. ACM Trans. Math. Softw. 7 (1981) 17-41.

[15] Y. Yuan, A new stepsize for the steepest descent method, J. Comput. Math. 24 (2006) 149-156.

[16] M. Zhu, J.L. Nazareth and H. Wolkowicz, The quasi-Cauchy relation and diagonal updating, SIAM J. Optim. 9(4) (1999) 1192-1204.