A MONOTONE PRECONDITIONED GRADIENT METHOD BASED ON A BANDED TRIDIAGONAL INVERSE HESSIAN APPROXIMATION

Saman Babaie–Kafaki[1]

Based on a tridiagonal memoryless inverse Hessian approximation in a least–squares approach, a preconditioned gradient method is proposed. Conducting an eigenvalue analysis, it is shown that the method possesses the sufficient descent property independent of the line search. Without the convexity assumption on the objective function, the method is established to be globally convergent under the Wolfe line search conditions. Using a set of unconstrained optimization test problems from the CUTEr library, the method is numerically compared with the two–point stepsize gradient method proposed by Barzilai and Borwein. The results of comparisons show that the method is computationally promising in the sense of the Dolan–Moré performance profile.

Keywords: Unconstrained optimization; Preconditioned gradient method; Tridiagonal matrix; Sufficient descent property; Line search

MSC2010: 90C 53, 49M 37, 65K 05

1. **Introduction**

Unconstrained optimization deals with the problem of minimizing an objective function $f : \mathbb{R}^n \to \mathbb{R}$ with no restriction on its variables, that is

$$\min_{x\in\mathbb{R}^n} f(x). \tag{1.1}$$

Here, we assume that f is continuously differentiable and its gradient is available. The problem (1.1) not only directly arises in some applications but also indirectly arises in reformulations of constrained optimization problems; often it is practical to replace the constraints of an optimization problem with penalized terms in the objective function and to solve an unconstrained problem.

As practical tools for solving (1.1), iterative methods define a sequence of approximations that are expected to be closer and closer to the exact solution in a given norm, stopping the iterations using some predefined criterion, and obtaining a vector which is only an approximation of the solution. In a class of such methods, the iterative formula is given by

$$x_0 \in \mathbb{R}^n,\ x_{k+1} = x_k + s_k,\ s_k = \alpha_k d_k,\ k = 0, 1, ..., \tag{1.2}$$

where $\alpha_k \in \mathbb{R}$ is a step length to be computed by a line search along the search direction $d_k \in \mathbb{R}^n$ which is often assumed to satisfy the sufficient descent condition, that is

$$d_k^T g_k \leq -c||g_k||^2,\ k = 0, 1, ..., \tag{1.3}$$

where $g_k = \nabla f(x_k)$, c is a positive constant and $||.||$ stands for the Euclidean norm. Inequality (1.3) plays an important role in the convergence analysis of the iterative method (1.2) [24,25].

Being a first–order optimization algorithm with the attractive features of satisfying (1.3) as well as converging globally, the most fundamental iterative method for solving (1.1)

[1] Department of Mathematics, Faculty of Mathematics, Statistics and Computer Science, Semnan University, P.O. Box: 35195–363, Semnan, Iran, Corresponding author: sbk@semnan.ac.ir, babaiekafaki@gmail.com

is the gradient (or steepest descent) method [6] with the simplest choice $d_k = -g_k$ in (1.2). In spite of such strong theoretical features, the method performs poorly, converges slowly and is badly affected by the ill–conditioning of the Hessian [1, 13, 25].

In an attempt to make a modification on the gradient method based on the quasi–Newton aspects, Barzilai and Borwein [4] (BB) dealt with an effective scaled gradient method with the following search directions:

$$d_0 = -g_0,\ d_k = -\theta_k g_k,\ k \geq 1, \tag{1.4}$$

in which the positive scalar θ_k is computed based on a two–point approximation of the (standard) secant equation [25], that is

$$H_k y_k = s_k, \tag{1.5}$$

where $y_k = g_{k+1} - g_k$, and $H_k \in \mathbb{R}^{n\times n}$ is an approximation of the inverse Hessian $\nabla^2 f(x_k)^{-1}$. More precisely, θ_k is obtained by solving the following least–squares problem:

$$\min_{\theta>0} ||s_k - \theta y_k||. \tag{1.6}$$

This yields

$$\theta_k = \frac{s_k^T y_k}{y_k^T y_k}, \tag{1.7}$$

being positive when the line search fulfills the popular Wolfe conditions [25], i.e.

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \delta \alpha_k d_k^T g_k, & (1.8)\\ d_k^T g_{k+1} &\geq \sigma d_k^T g_k, & (1.9) \end{aligned}$$

with the constants δ and σ satisfying $0 < \delta < \sigma < 1$. By symmetry, another choice for θ_k in (1.4) has been proposed in [4] by solving the following least–squares problem:

$$\min_{\theta>0} ||\frac{1}{\theta} s_k - y_k||,$$

which yields

$$\theta_k = \frac{s_k^T s_k}{s_k^T y_k},$$

being positive under the Wolfe conditions.

A brief review of the literature reveals an abundance of works related to the modified BB methods. As examples, Raydan [22] employed the nonmonotone line search procedure suggested by Grippo et al. [15] and developed a globally convergent modified BB method which is competitive and sometimes preferable to some efficient nonlinear conjugate gradient methods. Instead of using the standard secant equation (1.5) which only employs the gradient information, Dai et al. [8] used the modified secant equations proposed by Yuan [28] and Zhang et al. [29], and Babaie–Kafaki and Fatemi [2] adaptively used the modified secant equations proposed by Li and Fukushima [17], and Zhang et al. [29]. As an interesting feature, the modified BB methods of [2, 8] employ the objective function values in addition to the gradient information. Convergence properties of the BB method have been studied by Raydan [21], Dai and Liao [7], and Fletcher [12]. The interested reader can also study the references [5, 9, 10, 16, 18, 19, 27].

Here, we deal with a tridiagonal memoryless inverse Hessian approximation as extension of the diagonal approximation proposed by Barzilai and Borwein [4]. The method is discussed in Section 2, together with a brief global convergence analysis. In Section 3, the method is numerically compared with the BB method, using the Dolan–Moré performance profile. Conclusions are drawn in Section 4.

2. A tridiagonally preconditioned gradient method

From a matrix point of view, it can be seen that in the BB method with the scaling parameter θ_k given by (1.7), the inverse Hessian is approximated by the diagonal matrix $D_k = \theta_k I$ in (1.6). Extending the approach of [4], here we consider a banded tridiagonal approximation $T_k \in \mathbb{R}^{n\times n}$ for the inverse Hessian as follows:

$$[T_k]_{ij} = \begin{cases} \xi a_k, & i = j, \\ a_k, & |i-j| = 1, \\ 0, & \text{otherwise,} \end{cases}$$

where ξ is a positive constant and $a_k > 0$ is a parameter to be computed based on the secant equation (1.5).

As seen, T_k can be easily saved with a low memory requirement. Also, for an arbitrary vector $b \in \mathbb{R}^n$ with the ith element $b^{(i)}$, $i = 1, 2, ..., n$, $T_k b$ can be effectively computed as follows:

$$T_k b = T_k \begin{bmatrix} b^{(1)} \\ b^{(2)} \\ \vdots \\ b^{(n-1)} \\ b^{(n)} \end{bmatrix} = \begin{bmatrix} \xi a_k b^{(1)} + a_k b^{(2)} \\ a_k b^{(1)} + \xi a_k b^{(2)} + a_k b^{(3)} \\ \vdots \\ a_k b^{(n-2)} + \xi a_k b^{(n-1)} + a_k b^{(n)} \\ a_k b^{(n-1)} + \xi a_k b^{(n)} \end{bmatrix}.$$

Especially, if we let $b^{(0)} = b^{(n+1)} = 0$, then the ith element of the vector $T_k b$ can be generally written as $a_k b^{(i-1)} + \xi a_k b^{(i)} + a_k b^{(i+1)}$, $i = 1, 2, ..., n$.

Now, considering the secant equation (1.5), a_k is computed as a solution of the following least–squares problem:

$$\min_{a_k} ||s_k - T_k y_k||.$$

That is,

$$a_k = \frac{s_k^T p_k}{p_k^T p_k}, \tag{2.1}$$

in which $[p_k]_i = y_k^{(i-1)} + \xi y_k^{(i)} + y_k^{(i+1)}$, $i = 1, 2, ..., n$, with $y_k^{(0)} = y_k^{(n+1)} = 0$. However, a_k given by (2.1) may be nonpositive. So, to guarantee positiveness of the parameter a_k here we let

$$a_k = \frac{|s_k^T p_k|}{p_k^T p_k}. \tag{2.2}$$

Note that since

$$s_k^T p_k = \sum_{i=1}^{n} s_k^{(i)} (y_k^{(i-1)} + y_k^{(i+1)}) + \xi s_k^T y_k,$$

and also, since the Wolfe conditions ensure that $s_k^T y_k > 0$, for enough large values of ξ we have $s_k^T p_k > 0$ and so, (2.1) and (2.2) are equivalent. As will be shown, large values of ξ are more reasonable in the perspective of the conditioning of T_k. In order to avoid computational errors related to the small or large (positive) numbers, in a further modification we consider the following truncation of the parameter a_k given by (2.2) [8, 22]:

$$a_k = \max\left\{\epsilon, \min\left\{\frac{1}{\epsilon}, \frac{|s_k^T p_k|}{p_k^T p_k}\right\}\right\}, \tag{2.3}$$

where ϵ is a small positive constant. Here, the iterative method (1.2) with the search directions

$$d_0 = -g_0, \; d_k = -T_k g_k, \; k \geq 1, \tag{2.4}$$

in which a_k is computed by (2.3) is called a tridiagonal modification of the BB (TMBB) method. Next, we discuss the descent property of the TMBB method.

As shown in [23], the eigenvalues of T_k are given by

$$\lambda_{k_i} = \xi a_k + 2a_k \cos\frac{i\pi}{n+1},\ i = 1, 2, ..., n. \tag{2.5}$$

Hence, from (2.3) we get

$$\min\{\lambda_{k_i}\}_{i=1}^n = \lambda_{k_n} = \xi a_k + 2a_k \cos\frac{n\pi}{n+1} \geq (\xi - 2)a_k \geq (\xi - 2)\epsilon. \tag{2.6}$$

So, if $\xi > 2$, then from (2.4) for the TMBB method we have

$$d_k^T g_k = -g_k^T T_k g_k \leq -(\xi - 2)\epsilon ||g_k||^2.$$

The following theorem is now immediate.

Theorem 2.1. *For the TMBB method with $\xi > 2$ the sufficient descent condition (1.3) holds.*

Here, we discuss the global convergence of the TMBB method for which the following preliminaries are needed.

Assumption 2.1. The level set $\mathcal{L} = \{x|\ f(x) \leq f(x_0)\}$, with x_0 to be the starting point of the iterative method (1.2), is bounded. Also, in a neighborhood $\mathcal{N}$ of $\mathcal{L}$, f is continuously differentiable and its gradient is Lipschitz continuous; that is, there exists a positive constant L such that

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||,\ \forall x, y \in \mathcal{N}.$$

Lemma 2.1. [24] Suppose that Assumption 2.1 holds. Consider any iterative method in the form of (1.2) for which the sufficient descent condition (1.3) holds and the step length α_k satisfies the Wolfe conditions (1.8) and (1.9). If

$$\sum_{k\geq 0} \frac{1}{||d_k||^2} = \infty,$$

then the method converges in the sense that

$$\liminf_{k\to\infty} ||g_k|| = 0. \tag{2.7}$$

Now, we can prove the following global convergence theorem for the TMBB method, using Lemma 2.1.

Theorem 2.2. *Suppose that Assumption 2.1 holds. For the TMBB method with $\xi > 2$, if the step length α_k is determined such that the Wolfe conditions (1.8) and (1.9) are satisfied, then the method converges in the sense that (2.7) holds.*

Proof. At first, note that from Theorem 2.1 and the Wolfe condition (1.8), we have $\{x_k\}_{k\geq 0} \subseteq \mathcal{L}$. Also, $d_k \neq 0$, $\forall k \geq 0$, and consequently, using Lemma 2.1, to complete the proof it is enough to show that $||d_k||$ is bounded above.

Assumption 2.1 implies that there exists a positive constant γ such that

$$||\nabla f(x)|| \leq \gamma,\ \forall x \in \mathcal{L}, \tag{2.8}$$

(See Proposition 3.1 of [3].) and from (2.5), we have

$$||T_k|| = \max\{\lambda_{k_i}\}_{i=1}^n = \lambda_{k_1} = \xi a_k + 2a_k \cos\frac{\pi}{n+1} \leq (\xi + 2)a_k \leq \frac{\xi + 2}{\epsilon}. \tag{2.9}$$

Thus,

$$||d_k|| = || - T_k g_k|| \leq ||T_k||||g_k|| \leq \frac{\xi + 2}{\epsilon}\gamma,$$

which completes the proof.

□

TABLE 1. Test problems data

Function	n	Function	n
BROYDN7D	500	ENGVAL1	5000
BROYDN7D	1000	SCHMVETT	100
BROYDN7D	5000	SCHMVETT	500
BROYDN7D	10000	SCHMVETT	1000
COSINE	1000	SCHMVETT	5000
COSINE	10000	SPARSQUR	1000
DIXMAANF	1500	SPARSQUR	5000
DIXMAANF	3000	SPARSQUR	10000
DIXMAANF	9000	SPMSRTLS	1000
DIXMAANG	1500	SPMSRTLS	4999
DIXMAANG	3000	SPMSRTLS	10000
DIXMAANG	9000	SROSENBR	1000
DIXMAANJ	1500	SROSENBR	5000
DIXMAANJ	3000	SROSENBR	10000
DIXMAANJ	9000	TOINTGSS	1000
DIXMAANL	1500	TOINTGSS	5000
DIXMAANL	3000	TOINTGSS	10000
DIXMAANL	9000	VAREIGVL	100
ENGVAL1	100	VAREIGVL	500
ENGVAL1	1000	VAREIGVL	1000

As known, an essential factor which plays an important role in the sensitivity analysis of a numerical problem related to a matrix, is the matrix condition number [26]. A matrix with a large condition number is called an ill–conditioned matrix since instability may occur in the computations related to the matrix. About the computational stability of the TMBB method with $\xi > 2$, note that from (2.6) and (2.9) the spectral condition number of the matrix T_k is given by

$$\kappa(T_k) = \frac{\lambda_{k_1}}{\lambda_{k_n}} = \frac{\xi + 2\cos\dfrac{\pi}{n+1}}{\xi + 2\cos\dfrac{n\pi}{n+1}}.$$

As seen, for large values of ξ, $\kappa(T_k)$ tends to 1. That is, large values of ξ make T_k to be a well–conditioned matrix, enhancing the numerical stability of the TMBB method.

3. Numerical experiments

Here, we present some numerical results obtained by applying MATLAB implementations of the TMBB method with $\xi = 100$ and $\epsilon = 10^{-10}$, and the BB method with the parameter (1.7). The runs were performed on a set of 40 unconstrained optimization test problems of the CUTEr collection [14] with the minimum dimension being equal to 100, as specified in Table 1, using a computer Intel(R) Core(TM)2 Duo CPU 2.00 GHz with 1 GB of RAM. In the line search procedure, the Wolfe conditions (1.8) and (1.9) were used with $\delta = 0.0001$ and $\sigma = 0.9$, and the step length α_k was computed using Algorithm 3.5 of [20]. All attempts for finding an approximation of the solution were terminated by reaching maximum of 10000 iterations or achieving a solution with $||g_k||_\infty < 10^{-6}(1 + |f(x_k)|)$.

Efficiency comparisons were drawn using the Dolan–Moré performance profile [11], on the running time and the total number of function and gradient evaluations being equal to $N_f + 3N_g$, where N_f and N_g respectively denote the number of function and gradient

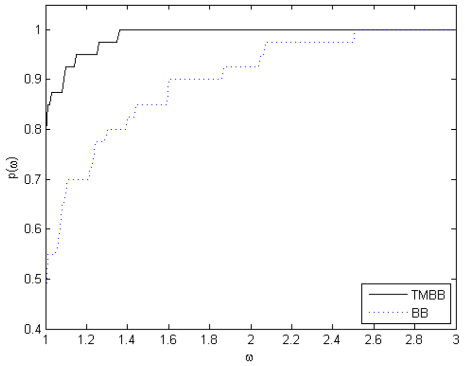

Figure 1. Total number of function and gradient evaluations performance profiles

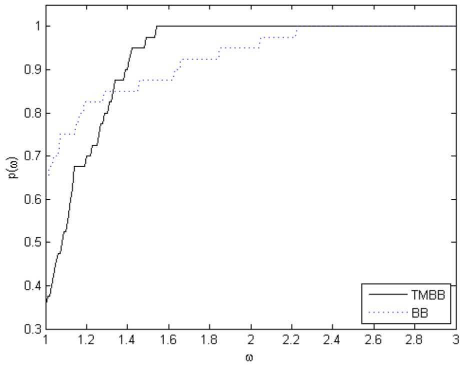

Figure 2. CPU time performance profiles

evaluations. The performance profile gives, for every $\omega \geq 1$, the proportion $p(\omega)$ of the test problems that each considered algorithmic variant has a performance within a factor of ω of the best. Figures 1 and 2 demonstrate the results of comparisons. As the figures show, TMBB outperforms BB with respect to the total number of function and gradient evaluations while BB is at times preferable to TMBB with respect to the running time. This seems reasonable since in contrast to BB, the search direction computation in TMBB needs more time.

4. **Conclusions**

As an extension of the Barzilai–Borwein approach, a preconditioned gradient method is proposed using a banded tridiagonal memoryless approximation of the inverse Hessian. Based on an eigenvalue analysis, a sufficient descent property has been established for the method which leads to the global convergence. Preliminary numerical experiments on a set of CUTEr unconstrained optimization test problems showed that the method turns out to be computationally promising, especially with respect to the number of function evaluations.

Acknowledgements

This research was supported by Research Council of Semnan University. The author is grateful to Professor Michael Navon for providing the line search code. He also thanks the anonymous referee for his/her valuable comments helped to improve the presentation.

REFERENCES

[1] *H. Akaike*, On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method, Ann. Inst. Statist. Math. Tokyo, **11**(1959), 1–16.

[2] *S. Babaie–Kafaki and M. Fatemi*, A modified two–point stepsize gradient algorithm for unconstrained minimization, Optim. Methods Softw., **28**(2013), 1040–1050.

[3] *S. Babaie–Kafaki, R. Ghanbari and N. Mahdavi–Amiri*, Two new conjugate gradient methods based on modified secant equations, J. Comput. Appl. Math., **234**(2010), 1374–1386.

[4] *J. Barzilai and J. M. Borwein*, Two–point stepsize gradient methods, IMA J. Numer. Anal., **8**(1988), 141–148.

[5] *E. Birgin, J. M. Martínez and M. Raydan*, Nonmonotone spectral projected gradient methods on convex sets, SIAM J. Optim., **10**(2000), 1196–1211.

[6] *A. Cauchy*, Méthodes générales pour la résolution des systèmes déquations simultanées, C. R. Acad. Sci. Paris, **25**(1847), 536–538.

[7] *Y. H. Dai and L. Z. Liao*, R–linear convergence of the Barzilai and Borwein gradient method, IMA J. Numer. Anal., **22**(2002), 1–10.

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

[9] *Y. H. Dai and H. Zhang*, Adaptive two–point stepsize gradient algorithm, Numer. Algorithms, **27**(2001), 377–385.

[10] *R. De Asmundis, D. Di Serafino, W. W. Hager, G. Toraldo and H. Zhang*, An efficient gradient method using the Yuan steplength, Comput. Optim. Appl., **59**(2014), 541–563.

[11] *E. D. Dolan and J. J. Moré*, Benchmarking optimization software with performance profiles, Math. Programming, **91**(2002), 201–213.

[12] *R. Fletcher*, On the Barzilai–Borwein method, In Optimization and Control with Applications, volume 96 of Appl. Optim., pages 235–256, Springer, New York, 2005.

[13] *G. E. Forsythe*, On the asymptotic directions of the s–dimensional optimum gradient method, Numer. Math., **11**(1968), 57–76.

[14] *N. I. M. Gould, D. Orban and Ph. L. Toint*, CUTEr: a constrained and unconstrained testing environment, revisited, ACM Trans. Math. Softw., **29**(2003), 373–394.

[15] *L. Grippo, F. Lampariello and S. Lucidi*, A nonmonotone line search technique for Newton's method, SIAM J. Numer. Anal., **23**(1986), 707–716.

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

[17] *D. H. Li and M. Fukushima*, A modified BFGS method and its global convergence in nonconvex minimization, J. Comput. Appl. Math., **129**(2001), 15–35.

[18] *F. Luengo, M. Raydan, W. Glunt and T. L. Hayden*, Preconditioned spectral gradient method, Numer. Algorithms, **30**(2002), 241–258.

[19] *Y. Narushima, T. Wakamatsu and H. Yabe*, Extended Barzilai–Borwein method for unconstrained minimization problems, Pacific J. Optim., **6**(2010), 591–613.

[20] *J. Nocedal and S. J. Wright*, Numerical Optimization, Springer, New York, 2006.

[21] *M. Raydan*, On the Barzilai and Borwien choice of steplength for the gradient method, IMA J. Numer. Anal., **13**(1993), 321–326.

[22] *M. Raydan*, The Barzilai and Borwein gradient method for the large–scale unconstrained minimization problem, SIAM J. Optim., **7**(1997), 26–33.

[23] *G. D. Smith*, Numerical Solution of Partial Differential Equations: Finite Difference Methods, Oxford University Press, Oxford, 1985.

[24] *K. Sugiki, Y. Narushima and H. Yabe*, Globally convergent three–term conjugate gradient methods that use secant conditions and generate descent search directions for unconstrained optimization, J. Optim. Theory Appl., **153**(2012), 733–757.

[25] *W. Sun and Y. X. Yuan*, Optimization Theory and Methods: Nonlinear Programming, Springer, New York, 2006.

[26] *D. S. Watkins*, Fundamentals of Matrix Computations, John Wiley and Sons, New York, 2002.

[27] *Z. Yu, J. Zang and J. Liu*, A class of nonmonotone spectral memory gradient method, J. Korean Math. Soc., **47**(2010), 63–70.

[28] *Y. X. Yuan*, A modified BFGS algorithm for unconstrained optimization, IMA J. Numer. Anal., **11**(1991), 325–332.

[29] *J. Z. Zhang, N. Y. Deng and L. H. Chen*, New quasi–Newton equation and related methods for unconstrained optimization, J. Optim. Theory Appl., **102**(1999), 147–167.