Australian Journal of Basic and Applied Sciences, 5(5): 29-34, 2011
ISSN 1991-8178

An Alternative Scaling Factor in BFGS Method For Unconstrained Optimization

[1]Muhammad Fauzi bin Embong, [2]Mohd Rivaie [3]Mustafa bin Mamat, [4]Ismail bin Mohd

[1]Department of Computer Science and Mathematics UiTM Kuala Terengganu Campus, 21080 Kuala Terengganu, Malaysia
[2]Department of Mathematics Universiti Malaysia Terengganu (UMT) 21030 Kuala Terengganu, Malaysia

Abstract: In order to calculate step size, a suitable line search method can be employed. As the step size usually not exact, the error is unavoidable, thus radically affect quasi-Newton method by as little as 0.1 percent of the step size error. A suitable scaling factor has to be introduced to overcome this inferiority. Self-scaling variable metric algorithms (SSVM's) are commonly used method, where a parameter is introduced, altering BFGS single parameter class of approximations to the inverse Hesssian to a double parameter class. This paper proposes an alternative scaling factor for the algorithms. The alternative scaling factor had been tried on several commonly test functions, and the numerical results shows that the new scaled algorithm shows significant improvement over the standard Broyden's class methods.

Key words: Component; Broyden's class, BFGS, DFP, scaling factor, SSVM, Hessian, Eigenvalues.

INTRODUCTION

The quasi-Newton methods are very popular and efficient methods for solving unconstrained optimization problem

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

where $f : R^n \to R$ is a twice continuously differentiable function. There are a large number of quasi-Newton methods but the BFGS update of Broyden, Fletcher, Goldfarb, and Shanno is more popular. As other quasi-Newton method, the Broyden's method are iterative, whereby at (k+1)-thiteration,

$$x_{k+1} = x_k + \alpha_k d_k \tag{1.2}$$

where d_k denotes the search direction and α_k is its step size. The search direction, d_k , is calculated by using

$$d_k = -H_k^{-1} g_k \tag{1.3}$$

where g_k is the gradient of f evaluated at the current iterate x_k, and H_k^{-1} is the inverse Hessian approximation.

The step size α_k is a positive step length chosen by a line search so that at each iteration either

$$f\left(x_k + \alpha_k d_k\right) \le f\left(x_k\right) - \eta_1 \alpha_k \frac{\left(g_k^T d_k\right)^2}{s_k^T y_k} \tag{1.4}$$

or

Corresponding Author: Muhammad Fauzi bin Embong, Department of Computer Science and Mathematics UiTM Kuala Terengganu Campus, 21080 Kuala Terengganu, Malaysia
E-mail: 1fauziembong@yahoo.com

$$f\left(x_k+\alpha_k d_k\right)\le f\left(x_k\right)-\eta_2 g_k^T d_k \tag{1.5}$$

where η_1 and η_2 are positive constants.

We note that the conditions (1.4) and (1.5) are assumes in Byrd and Nocedal (1989). They cover a large class of line search strategies under suitable conditions. If the gradient of *f* is Lipschitz continuous, then for several well known line search satisfy the Wolfe conditions:

$$f\left(x_k+\alpha_k d_k\right)\le f(x_k)+\delta_1\alpha_k d^T g_k \tag{1.6}$$

and

$$g(x_k+\alpha_k d_k)^T d_k\ge\delta_2 d_k^T g_k \tag{1.7}$$

where $0<\delta_1<\frac{1}{2}$ and $\delta_1<\delta_2<1$, imply (1.4).

Byrd and Nocedal (1989) prove that if the ratio between successive trial values of α is bounded away from zero, the new iteration produced by a backtracking line search satisfies (1.4) and (1.5).
The Hessian approximation is then updated by

$$H_{k+1}^{\phi}=H_k+\left(1+\frac{y_k^T H_k y_k}{s_k^T y_k}\right)\frac{s_k s_k^T}{s_k^T y_k}-\frac{s_k y_k^T+H_k y_k s_k^T}{s_k^T y_k} \tag{1.8}$$

where

$$s_k=x_{k+1}-x_k, \tag{1.9}$$

$$y_k=g_{k+1}-g_k \tag{1.10}$$

Scaling The BFGS Method:
Self-scaling Variable Metric (SSVM) Method:

Many modifications have been applied on quasi-Newton methods in attempt to improve its efficiency. In this section, the discussion will be on the self-scaling variable metric algorithms developed by Oren (1973) and Oren and Luenberger (1974). Multiplying H_k by γ_k, and then replacing $\gamma_k H_k$ in (1.8), the BFGS formula can be written as:

$$H_{k+1}^{\phi}=\left(H_k-\frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k}\right)\gamma_k+\frac{y_k y_k^T}{y_k^T s_k} \tag{2.1}$$

where γ_k is a self-scaling parameter. The formula (2.1) is known as self-scaling variable metric (SSVM) formula. Clearly, when $\gamma_k=1$, the formula (2.1) is reduced to Broyden's class update (1.8).

Choices of the Scaling Factor:

The choice of a suitable scaling factor can be determined by the following theorem.

Theorem (Oren and Luenberger, (1974)):

Let $\phi\in[0,1]$ and $\gamma_k>0$. Let H_k be the Hessian approximation, and H_{k+1}^{ϕ} be defined by (2.1). Let $\lambda_1\ge\lambda_2\ge\ldots\ge\lambda_n$ and $\mu_1^{\phi}\ge\mu_2^{\phi}\ge\ldots\ge\mu_n^{\phi}$ be eigenvalues of H_k and H_{k+1} respectively. Then the

following statements hold.

If $\gamma_k \lambda_n \geq 1$, then $\mu_n^{\phi} = 1$ and $1 \leq \gamma_k \lambda_{i+1} \leq \mu_i^{\phi} \leq \gamma_k \lambda_i$, $i = 1, 2, ..., n-1$.

If $\gamma_k \lambda_1 \leq 1$, then $\mu_n^{\phi} = 1$ and $\gamma_k \lambda_i \leq \mu_i^{\phi} \leq \gamma_k \lambda_{i-1} \leq 1$, $i = 2, 3, ..., n$.

If $\gamma_k \lambda_1 \leq 1 \leq \gamma_k \lambda_1$ and i_0 is an index with $\gamma_k \lambda_{i0} \leq 1 \leq \gamma_k \lambda_{i0}$, then

$$\begin{aligned} \gamma_k \lambda_1 \geq \mu_1^{\phi} \geq \gamma_k \lambda_2 \geq \mu_2^{\phi} \geq ... \geq \gamma_k \lambda_{i0} \\ \geq \mu_{i0} \geq 1 \geq \mu_{i0+1} \\ \geq \gamma_k \lambda_{i0+1} \geq ... \geq \gamma_k \lambda_n \end{aligned}$$

and there is at least one eigenvalue in μ_{i0}^{ϕ} and μ_{i0+1}^{ϕ} which equals 1.

Readers who are interested in the proof for the above theorem may refer Oren and Luenberger (1974), or Sun and Yuan (2006). From the above theorem, it can be shown that

$$\gamma_k = \frac{s_k^T y_k}{s_k^T H_k s_k} \tag{2.2}$$

is a suitable scaling factor (Sun and Yuan, 2006).

Spedicato (1975) has considered the problem of initializing H_0 , suggesting that H_0 be set to a diagonal matrix whose elements are the reciprocals of the true diagonal elements of the Hessian evaluated at the initial approximation x_0. However Shanno and Phua (1978) have found that there is a major drawback to this, namely that it is often expensive to obtain these estimates. Furthermore, computational efficiency is not always improved. They suggested a much simpler initial scaling which require no additional information about the object function than that routinely required by variable metric algorithm. Initially $H_o = I$ may be used to determine x_1 , where α_o is chosen according to some step length or linear search criterion to assure sufficient reduction in the function *f*. Once x_1 has been chosen, but before H_1 is calculated, H_0 now being scaled by

$$\hat{H}_0 = \alpha_0 H_0 \tag{2.3}$$

and

$$H_1^{\phi} = \hat{H}_0 + \frac{s_k s_k^T}{s_k^T y_k} - \frac{\hat{H}_0 y_k y_k^T \hat{H}_0}{y_k^T \hat{H}_0 y_k} + \phi v_k v_k^T \tag{2.4}$$

$$v_k = \left(y_k^T \hat{H}_0 y_k \right)^{\frac{1}{2}} \left(\frac{s_k}{s_k^T y_k} - \frac{\hat{H}_0 y_k}{y_k^T \hat{H}_0 y_k} \right) \tag{2.5}$$

Substitution of (2.3) into (2.4) yields.

$$H_1^{\phi} = \left(H_0 - \frac{H_0 y_k y_k^T H_0}{y_k^T H_0 y_k} + \phi v_k v_k^T \right) \alpha + \frac{y_k y_k^T}{y_k^T s_k} \tag{2.6}$$

After the initial scaling of H_0 by an appropriate step size α, the approximate Hessian is never rescaled.

Numerical experiments show that the initial scaling is simple and effective for a lot of problems in which the curvature changes smoothly (Sun, 2006).

An Alternative Scale Factor:

In this article, the smallest eigenvalue of Hessian approximation was proposed as an alternative scaling factor of initial scaling on *H* as in (2.3). Replacing step size, α with the smallest eigen value of *H*, λ into (2.6) yields:

$$H_1^{\phi} = \left(H_0 - \frac{H_0 y_k y_k^T H_0}{y_k^T H_0 y_k} + \phi v_k v_k^T \right) \lambda + \frac{y_k y_k^T}{y_k^T s_k}. \tag{2.7}$$

As proposed by Shanno and Phua (1978), this update is also an initial scaling on Hessian approximation. After the initial iteration, the Hessian approximation is never rescaled. The following algorithm is proposed with the smallest eigenvalue of *H*, λ as the scaling factor.

Eigenvalue Scaled Algorithm:

A modification of the BFGS method is obtained by applying this algorithm. For other members of quasi-Newton methods, this algorithm is also applicable.

Step 1: Initialization

Given x_0 , set $k = 0$, and $H_0 = I$.

Step 2: Computing search direction

$d_k = -H_k g_k$. If $g_k = 0$, then stop.

Step 3: Computing step size, α

Step 4: Updating new point, $x_{k+1} = x_k + \alpha_k d_k$

Step 5: Updating approximation of Hessian Matrix, H_{k+1}

For *k*=1, use (2.7), else, use (1.8).

Step 6: Convergent test and stopping criteria

If $f(x_{k+1}) < f(x_k)$ and $\|g_k\| \le \varepsilon$, then stop.

Otherwise go to Step 1 with *k* = *k*+1.

Numerical Results:

A MAPLE subroutine was programmed to test three algorithms, BFGS algorithm without scaling (BFGS), with eigenvalue scaling (2.7) (denoted as ES-BFGS), and with the initial scaling (2.6) (denoted as S-BFGS). The three algorithms was applied on eight commonly tested functions, consist of two variables (n=2) functions and four variables (n=4) functions.

Rosenbrock function with n=2.

$$f(x) = 100(x_2 - x_1^2)^2 + (x_1 - 1)^2$$

Cube function with n=2

$$f(x) = 100(x_2 - x_1^3)^2 + (x_1 - 1)^2$$

Shalow function with n=2

$$f(x) = (x_2 - x_1^2)^2 + (1 - x_1)^2$$

Strait function with n=2

$$f(x) = (x_2 - x_1^2)^2 + 100(x_1 - 1)^2$$

Rosenbrock function with n=4

$$f(x) = 100(x_2 - x_1^2)^2 + (x_1 - 1)^2 + 100(x_4 - x_3)^2 + (x - 1_3)^2$$

Cube function with n=4

$$f(x) = 100(x_2 - x_1^3)^2 + (x_1 - 1)^2 + 100(x_4 - x_3^3)^2 + (x_3 - 1)^2$$

Shalow function with n=4

$$f(x) = (x_2 - x_1^2)^2 + (1 - x_1)^2 + (x_4 - x_3^2)^2 + (1 - x_3)^2$$

Wood function with n =4

$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 + 90(x_4 - x_3^2)^2 + (1 - x_3)^2 + 10\left(x_2 + x_4 - 2\right)^2 + 0.1\left(x_2 - x_4\right)^2$$

The numerical results produced by implementing the three algorithms to the test functions are presented in the Table 1. The efficiency of the algorithm are based on the number of iterations needed to reach the minimum value of the functions. Algorithm with less iteration is considered more efficient. A tolerance of $\varepsilon=10^{-6}$ is set as the stopping criteria.

Discussion:

From the calculation using MAPLE software, an interesting relation between step size and eigenvalues is observed. For initial iteration, the value of step size and the smallest eigen value of Hessian approximation is almost identical, but after a number of iterations, the difference increased accordingly. This explains why

the ES-BFGS is effective on initial scaling only, a similarity with the scaled algorithm proposed by Shanno and Phua (1978). When the scaling on Hessian approximation was done on every iteration, the performance of the scaled BFGS deteriorated, or even failed at all.

The numerical results show that the new algorithm (ES-BFGS) performance is comparable to the algorithm with initial scaling (S-BFGS).

Conclusion:

An improvement over unscaled BFGS is achieved, as for most of the cases, the number of iterations are reduced. Further investigation will be carried out using the alternative scaling factor, λ on the other types of quasi-Newton methods. The relationship between the smallest eigenvalue of Hessian approximation and the optimal step size is also of the interest in future research, triggering the possibility of using eigenvalue as a new step size in the quasi-Newton methods.

Table 1: Numerical results produced by the three tested algorithm, BFGS without scaling (BFGS), with eigenvalue scaling (ES-BFGS) and step size scaling (S-BFGS)

Test function	Initial point	*Number of iterations*		
		BFGS	ES-BFGS	S-BFGS
Rosenbrock	(-2,-2)	17	15	16
(*n*=2)	(-100,100)	30	28	28
	(10 000,10 000)	133	11	131
Cube	(-1.2,1.6)	17	8	8
(*n*=2).	(1.5,-150)	62	54	63
	(100,50)	44	44	35
Shalow	(5,5)	9	9	9
(*n* = 2)	(-100,100)	15	13	13
	(1000,-5000)	11	9	9
Strait	(2,2)	5	5	5
(*n* = 2).	(100,100)	9	13	9
	(1000,1000)	12	13	13
Rosenbrock	(-2,-2,-2,-2)	16	18	17
(*n*= 4).	(-100,100,100,100)	30	29	28
	(100,100,100,1.5)	156	102	137
Cube	(1.5,-1.5,1.5,-1.5)	42	40	39
(n=4).	(10,-10,10,-10)	18	17	18
	(15,-15,15,-15)	45	28	24
Shalow	(2,4,2,4)	7	8	7
(n = 4)	(-200,400,200,400)	326	23	23
	(2000,2000,2000,2000)	78	74	99
Wood	(2,-2,2,-2)	29	24	22
(*n*=4)	(200,-5,200,-5)	43	29	30
	(2000,2000,2000,2000)	80	56	58

REFERENCES

Byrd, R.H. and J. Nocedal, 1989. A tool for the analysis of quasi-Newton methods with application to unconstrained minimization. SIAM J. Numerical Analysis, 26: 727-739.

Chong, E.K.P. and S.H. Zak, 2001. An Introduction to Optimization (2 Ed.). San Diego, U.S.A: John Wiley & Sons, Inc.

Dennis, J.E. and J.J. More, 1977. Quasi-Newton Methods, Motivation and Theory,SIAM Review, 19: 46-89.

Oren, S.S., 1973. Self-scaling Variable Metric Algorithms Algorithms Without Line Search for Unconstrained Minimization. Mathematics of Computation American Mathematical Society, 27(124): 863-874.

Oren, S.S. and D.G. Luenberger, 1974. Self-scaling variable metric (SSVM) Algorithms. Management Science, 20(5): 845-862.

Shanno, D.F. and K.H. Phua, 1978. Matrix Conditioning and Nonlinear Optimization. Mathematical Programming, 14: 149-160.

Spedicato, E., 1975. Computational Experience with quasi-Newton Algorithms for Minimization Problems of Moderately Large Size. Report CISE-N-175, CISE Documentation Service, Segrate. (Milano).

Sun, W. and Y.X. Yuan, 2006. Optimization Theory and Method, Springer Optimization and Its Applications. New York USA, 1.