Vol.3.Issue.4.2015 (Oct-Dec.)

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

http://www.bomsr.com

ISSN:2348-0580

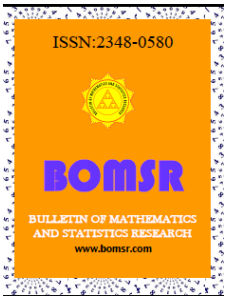
RESEARCH ARTICLE

THE CONVERGENCE OF BFGS METHOD UNDER STRONGLY CONVEX OPTIMIZATION PROBLEM

Dr. EMAN A. HUSSIAN[1], MAZIN H. SUHHIEM[2]

[1]Dept. of Mathematics, College of Sciences, AL-Mustansiriyah University,Baghdad,Iraq.
(dr_ emansultan@yahoo.com)

[2]Dept. of statistics, College of Adm. and Econ., University of sumar,Al-Rifaee, Iraq
(mz80m@yahoo.com)



ABSTRACT

In this paper, we discussed BFGS quasi – Newton method for solving unconstrained optimization problems. we showed the properties of BFGS technique with inexact line search algorithms such as Wolfe line search and backtracking line search. Some theorems that related to convergence of BFGS method have been introduced. Finally, we presented and proved a number of an important theorems that ensure the global convergence of BFGS technique under strongly convex optimization problem.

Keywords: BFGS method, Convex function, Strongly convex function, Optimization problem, Wolfe conditions.

1. INTRODUCTION

The most well-known optimization technique for unconstrained problems is Newton method. It is effective and robust. The second derivative need to be calculated analytically and supplied to the algorithm by the user.But Newton methods has the disadvantage of being computationally expensive, the inverse of the Hessian matrix has to be calculated in every iteration, and that is rather costly. Moreover, in some applications, the second derivatives may be unavailable. One fix to the problem is to use a finite difference approximation to Hessian. The other fix, which is more widely used, is quasi-Newton methods, where approximate Hessian or inverse Hessian updates are updated in each iteration, while the gradients are supplied[6].

Quasi-Newton methods are arguably the most popular class of the non-linear numerical optimization methods. Quasi-Newton methods are based on the Newton's method but don't require calculation of the second derivatives since sometimes, it is very difficult to derive the Hessian H. They update an approximate Hessian matrix at each iteration of the algorithm . i. e., in Quasi-Newton methods, Hessian matrix is estimated by using successive gradient vectors[7,8].

There are many different quasi-Newton update formulae, which are the most popular algorithms, namely : BFGS, DFB, PSB,SR1,etc. From these algorithms, BFGS is the most effective quasi-Newton method [9].

BFGS is the most popular quasi-Newton method, it is the most effective algorithm. The advantage of BFGS method is that this method preserves the structural properties needed for the line search direction methods in optimization which are the symmetry and the positive definiteness [1].

To compute the new step update in BFGS method one can use either the line search or the Trust Region strategies.These two strategies have different properties and are best used with specific Hessian approximation or inverse Hessian approximation updates. The line search will need the Hessian approximation or inverse Hessian approximation to be symmetric and positive definite [5].

Since, in practical computation, theoretically exact optimal step length generally cannot be found, and it is also expensive to find almost exact step length, therefore the inexact line search with less computation load is highly popular (such as : Wolfe line search, Goldstein line search,backtracking line search,etc) [2].

In this work we will discuss the global convergence properties of BFGS technique with inexact line search algorithms.Also we will prove that If the optimization function is strongly convex then the global convergence will always hold.

2. Newton Methods [6]

The most well-known minimization technique for unconstrained problems is Newton method. It is effective and robust. In each iteration, the step update is :

$$x_{k+1} = x_k - (\nabla^2 f_k)^{-1} \nabla f_k.$$

The second derivative need to be calculated analytically and supplied to the algorithm by the user.

2.1. Univariate Function [6]

Starting form initial guess x_0, Newton's method finds a sequence of x_k that converges to an x such that f(x) = 0, i.e., the root of equation. By Taylor's series expansion,

$$f(x + \Delta x) = f(x) + f'(x)\,\Delta x + \frac{1}{2} f''(x)\,\Delta x^2 + \cdots \quad (1)$$

For small enough Δx, and for well-behaved functions, 2nd- and higher-order terms are negligible. When method converges, $f(x + \Delta x) \cong 0$.

So, $$\Delta x = -\frac{f(x)}{f'(x)} \quad (2)$$

That is, update x as follows: $$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \quad (3)$$

Newton's method can also be used to find the max or min of f(x).

in this case, we find x such that $f'(x + \Delta x) \cong 0$.

So, the update equation is : $$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} \quad (4)$$

2.2. Multivariate Function[7,10]

For multivariate function f(x),

• f′(x) is replaced by the gradient of f:

$$g = \nabla f = \left[\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \quad \cdots \quad \frac{\partial f}{\partial x_n}\right]^T. \quad (5)$$

• $(f''(x))^{-1}$ is replaced by inverse of Hessian matrix H of f:

$$H = \nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \quad (6)$$

(Note that H is the Jacobian of g.However, H has more structure than a Jacobian for a general non-linear function. If f is twice continuously differentiable function, then the Hessian matrix is symmetric [1]).

The update equation becomes :

$$x_{k+1} = x_k - [H(x_k)]^{-1} \nabla f(x_k) \quad (7)$$

Usually, the learning rate η is introduced:

$$x_{k+1} = x_k - \eta [H(x_k)]^{-1} \nabla f(x_k) \quad (8)$$

Eq. (8) is called modified Newton method.

The classical modified Newton method:

$$x_{k+1} = x_k - \eta [H(x_0)]^{-1} \nabla (x_k) \quad (9)$$

Note that the Hessian matrix is only evaluated at the initial point x_0.

Newton methods has the disadvantage of being computationally expensive, the inverse of the Hessian matrix has to be calculated in every iteration, and that is rather costly. Moreover, in some applications, the second derivatives may be unavailable. One fix to the problem is to use a finite difference approximation to Hessian (For more details, see [6]).

The other fix, which is more widely used, is quasi-Newton methods, where approximate Hessian or inverse Hessian updates are updated in each iteration, while the gradients are supplied.

3. Quasi-Newton Methods[7,8,10]

Quasi-Newton methods are arguably the most popular class of the non-linear numerical optimization methods. Quasi-Newton methods are based on the Newton's method but don't require calculation of the second derivatives since sometimes, it is very difficult to derive the Hessian H. They update an approximate Hessian matrix at each iteration of the algorithm . i. e., in Quasi-Newton methods, Hessian matrix is estimated using successive gradient vectors.

The approximation B of Hessian matrix is chosen to satisfy:

$$\nabla f(x + \Delta x) = \nabla f(x) + B \Delta x \quad (10)$$

The update equation is :

$$x_{k+1} = x_k - \eta B_k^{-1} \nabla f(x_k) \quad (11)$$

or $$x_{k+1} = x_k - \eta_k B_k^{-1} \nabla f(x_k) \quad (12)$$

The basic idea behind the quasi-Newton formulae is to update B_{k+1} to B_k in some computational cheap ways while ensure the secant condition (eq.13), and the computation of the update should be relative cheap.Therefore, basic requirement for the updating formula is that the secant condition is satisfied in each iteration, i.e.,

$$B_{k+1} p_k = q_k \quad (13)$$

where

$q_k = g_{k+1} - g_k$, $p_k = x_{k+1} - x_k$.

Also,the updated approximation must be symmetric positive definite.

If n = 1,all secant methods reduce to the classical secant method for the single non-linear equation $f'(x) = 0$, i.e.,

$$x_{k+1} = x_k - \frac{f'(x_k)(x_k - x_{k-1})}{f'(x_k) - f'(x_{k-1})} \quad (14)$$

The general structure of quasi-Newton method can be summarized as

Given any starting point $x_0 \in \mathrm{dom}(f)$, B_0 any symmetric positive definite matrix (such as identity matrix I).

For k = 1,2,…, until a stopping criterion is satisfied

1. Compute quasi-Newton direction $d_k = - B_k^{-1} \nabla f(x_k)$
2. Determine step size η_k (e.g.,by backtracking line search)
3. Compute $x_{k+1} = x_k + \eta_k d_k$
4. Compute B_{k+1} .

There are many different quasi-Newton algorithms use different rules for updating B in step 4. The standard quasi-Newton update for non-linear equation is Broyden method [6] :

$$B_{k+1} = B_k + \frac{(q_k - B_k p_k)\, p_k^T}{p_k^T p_k} \tag{15}$$

Broyden method does not preserve the structural properties needed for line search direction methods in optimization, namely, symmetry and positive definiteness.

There are four different quasi-Newton update formulas, which are the most popular algorithms, namely :

BFGS formula (Broyden-Fletcher-Goldfarb-Shanno).

DFB formula (Davidon-Fletcher-Powell).

PSB formula (Powell-Symmetric- Broyden).

SR1 formula (Symmetric-Rank-1).

From the above algorithms, BFGS is the most effective quasi-Newton method. (For more details see [5,6,7,9]).

3.1 BFGS Method[1,7]

BFGS is the most popular quasi-Newton method, it is the most effective algorithm. In this method, if B_k is positive definite,then B_{k+1} is also positive definite. Therefore, if B_0 is chosen positive definite, the rest of the B_k will be positive definite.

The updating formula has the form :

$$B_{k+1} = B_k + \frac{q_k q_k^T}{q_k^T p_k} - \frac{B_k p_k p_k^T B_k}{p_k^T B_k p_k} \tag{16}$$

The updating eq.(16)is satisfy the secant condition (eq.13),since

$$B_{k+1} p_k = B_k p_k + \frac{q_k q_k^T}{q_k^T p_k} p_k - \frac{B_k p_k p_k^T B_k}{p_k^T B_k p_k} p_k = B_k p_k + q_k - B_k p_k = q_k.$$

The advantage of BFGS method is that this method preserves the structural properties needed for the line search direction methods in optimization which are the symmetry and the positive definiteness.

To illustrate this point, the following theorem is presented :

Theorem (1) : Let B_0 (resp. B_0^{-1}) be a positive definite, then $q_k^T p_k > 0$ is a necessary and sufficient condition for BFGS formula to give B_k (resp. B_k^{-1}) positive definite $\forall\, k \in N$.

Proof : see [7] (P. 41).

(**Note :** the condition $q_k^T p_k > 0$ is called the curvature condition[8])

Lemma (2) [1] : Let B_k (resp. B_k^{-1}) be a symmetric positive definite, $q_k^T p_k > 0$, and B_{k+1} given by eq.(16).Then B_{k+1}(resp. B_{k+1}^{-1}) is symmetric positive definite

Proof : From theorem(1) and secant condition.

It is very useful for both theory and practice to express eq.(16)in terms of the inverse matrices. Let $M = B^{-1}$ (inverse Hessian approximation),the following lemma gives the formula for the inverse updating :

Lemma (3) : Let B_k be symmetric positive definite,$q_k^T p_k \neq 0$,and B_{k+1} given by eq.(16).Then B_{k+1}^{-1} (M_{k+1}) is non-singular and

$$M_{k+1} = \left(I - \frac{p_k q_k^T}{q_k^T p_k}\right) M_k \left(I - \frac{q_k p_k^T}{q_k^T p_k}\right) + \frac{p_k p_k^T}{q_k^T p_k} \tag{17}$$

Proof : see [1] (p.72).

Eq.(17) can be rewritten as : $M_{k+1} = M_k + M_k^U$, where M_k^U is the update matrix. Thus, we have :

$$M_{k+1} = M_k + \left(1+\frac{q_k^T M_k q_k}{q_k^T p_k}\right)\frac{p_k p_k^T}{q_k^T p_k} - \frac{p_k q_k^T M_k + M_k q_k p_k^T}{q_k^T p_k} \quad (18)$$

BFGS Quasi-Newton Algorithm :

1) Input x_0, M_0, termination criteria. ($x_0 \in$ dom(f), M_0 is any symmetric positive definite such as identity matrix).

2) For any k, set $S_k = - M_k g_k$.

3) Compute a step size η_k (e.g., by Wolfe line search).

4) Set $x_{k+1} = x_k + \eta_k S_k$

5) Compute the update matrix M_k^U by using the values:

$q_k = g_{k+1} - g_k$, $p_k = x_{k+1} - x_k$, and M_k.

6) Set $M_{k+1} = M_k + M_k^U$.

7) Continue with next k unitil termination criteria are satisfied.

4. Convex Optimization

In this section we will discuss the convex optimization problem on BFGS quasi-Newton method.First, we present the following important basic definitions about the convex functions.

Definition(1),[4] : A subset C of R^n is convex if for any $x, y \in C$ and any $\theta \in [0,1]$, we have :

$$\theta x + (1-\theta)y \in C. \quad (19)$$

(It is clear that the empty set, any singleton set, and the whole space R^n are convex subsets of R^n).

Definition(2),[4] : A function $f: R^n \to R$ is convex if $dom(f)$ is a convex set and if for all $x, y \in dom(f)$ and $\theta \in [0,1]$ we have :

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) \quad (20)$$

a function f is strictly convex if stric inequality holds in (20) whenever $x \neq y$ and $\theta \in (0,1)$.

Remark(1),[4,7] : For differentiable functions, we note :

1) If f is differentiable (i.e., its gradient ∇f exists at each point in $dom(f)$,which is open).Then f is convex if and only if $dom(f)$ is convex and

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) \quad (21)$$

holds for all $x, y \in dom(f)$.

2) f is strictly convex if and only if $dom(f)$ is convex and for all $x, y \in dom(f), x \neq y$ we have

$$f(y) > f(x) + \nabla f(x)^T (y-x) \quad (22)$$

3) If f is twice differentiable (i.e., its Hessian $\nabla^2 f$ exists at each point in $dom(f)$,which is open). Then f is convex if and only if $dom(f)$ is convex and its Hessian is positive semidefinite for all $x \in dom(f)$.

4) If $\nabla^2 f$ is positive definite for all $x \in dom(f)$, then f is strictly convex,but the converse is not necessary true (Indeed, the converse will be true if f is quadratic function).

5) If f is twice differentiable, then f is convex if and only if $dom(f)$ is convex and $(y-$ xT∇2fxy−x≥0,∀x,y∈domf .

6) If f is twice differentiable and $(y-x)^T \nabla^2 f(x)(y-x) > 0, \forall x, y \in dom(f), x \neq y$ then f is strictly convex.

Definition(3),[4] : A function $f: R^n \to R$ is strongly convex if $dom(f)$ is a convex set and if $\exists\, \mu > 0$ *such that* :

$$\left(\nabla f(y) - \nabla f(x)\right)^T (y-x) \geq \mu \|y-x\|^2, \forall x, y \in dom(f).$$

Remark(2),[4,7] : For differentiable functions, we note :

1) If f is differentiable, then if f is strongly convex then f is strictly convex and

$$f(y) - f(x) \geq \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2, \forall x, y \in \mathrm{dom}(f).$$

2) If f is twice differentiable, then :

f is strongly convex iff $y^T \nabla^2 f(x) y \geq \mu \|y\|^2, \forall y \in \mathrm{dom}(f)$.

3) If f is quadratic**,** then :

f is strongly convex iff its Hessian is positive definite.

4) If the function f is convex quadratic then strongly convexity means strictly convexity and vice versa).

Convex functions are of interest in the context of optimization theory for several reasons. They arise frequently and have many significant properties, among which is the fact that a local minimum of a convex function(on a convex domain) is a global minimum. This makes it possible to use local conditions to test for optimality.

Definition(4),[1] : The α −sublevel set of a function $f: R^n \to R$ is defined as:

$$C_\alpha = \{ x \in \mathrm{dom}(f) : f(x) \leq \alpha \} \quad (23)$$

Sublevel sets of a convex function are convex, for any value of α.

Definition(5) : An n x n matrix A is positive semi definite if $x^T Ax \geq 0$ for all $x \in R^n$. A is positive definite if $x^T Ax > 0$ *for all* $x \in R^n, x \neq 0$.

Note : Every positive definite matrix is invertible and its inverse is also positive definite.

4.1. Mathematical Optimization

A mathematical optimization problem, or just optimization problem has the form :

$$\text{minimize } f(x) \quad (24)$$

Here the vector $x = (x_1, x_2, \ldots, x_n)$is the optimization variable of the problem, the function $f: R^n \to R$ is the objective function.We say that the problem(24) is unconstrained because we impose no conditions on the independent variables x and assume that f is defined for all x.

Definition(6), [7]: The vector $x^* \in \mathrm{dom}(f)$ is called optimal, or a solution of the problem (24) if it has the smallest objective value,this mean that :

$$f(x^*) \leq f(z), \forall z \in \mathrm{dom}(f) \quad (25)$$

(In fact, x^* in this case is the global minimizer point for problem(24).

Definition(7), [1,7]: The optimization problem(24) is called a linear problem if the objective function f is linear,i.e., satisfy :

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y), \forall x, y \in R^n, \forall \alpha, \beta \in R \quad (26)$$

If the optimization problem is not linear, it is called a nonlinear problem. In the work, we will focus on a class of the optimization problems called the convex optimization problems.

Definition(8),[1,7] : A convex optimization problems is one in which the objective function f is convex,which means it satisfy

$$f(\alpha x + \beta y) \leq \alpha f(x) + \beta f(y), \forall x, y \in R^n, \forall \alpha, \beta \in R \quad (27)$$

with $\alpha + \beta = 1, \alpha \geq 0, \beta \geq 0$.

We see that the convexity is more general than the linearity.

Remark(3),[4,7] If $\nabla f(x) = 0$,then from eq.(21) we have $f(x) \leq f(y), \forall y \in \mathrm{dom}(f)$ Therefore, the necessary and sufficient condition for x to be a global minimizer for optimization problem(24) is $\nabla f(x) = 0$.

(Therefore, the stopping criterion is usually of the form $\|\nabla f(x)\|_2 \leq \gamma$, where γ is small and positive).

4.2. Quadratic Convex Optimization Problem[4]

The convex unconstrained optimization problem (24) is called a quadratic problem if the objective function f is quadratic function.

Therefore, the optimization problem(24) can be written as follows :

$$\text{minimize } f(x) = \frac{1}{2}\, x^T P x + q^T x + r \tag{28}$$

where P is symmetric positive semidefinite,$q \in R^n$ and $r \in R$.

Remark(4),[6,7] For convex quadratic functions $f: R^n \to R$, we note:

1) f is convex if and only if P is positive semidefinite.

2) f is strictly convex if and only if P is positive definite.

3) The necessary and sufficient condition for x to be a global minimizer for optimization problem(28) is $\nabla f(x) = Px + q = 0$ (29)

4.3. Least-Squares Problem[1]

A Least-Squares Problem is an optimization problem with no constraints and an objective which is a sum of squares of terms of the form $a_i^T x - b_i$:

$$\text{minimize } f(x) = \|Ax - b\|_2^2 = \sum_{i=1}^{k} (a_i^T x - b_i)^2 \tag{30}$$

where $A \in R^{k \times n}$ (with $k \geq n$), a_i^T are the rows of A, and the vector $x \in R^n$ is the optimization variable.

Remark(5),[3,4] One special case of the convex quadratic minimization problem that arises very frequently is the Least-Squares Problem. Therefore, the authors usually used the Least-Squares Problem as error function (minimized error function).

4.4. BFGS Method and Minimization Problem

In this section, we discuss BFGS quasi-Newton method for solving the unconstrained convex quadratic (Least-Squares) minimization problem

$$\text{minimize } f(x) = \frac{1}{2} x^T \nabla^2 f(x) x + \nabla f(x)^T x + r \tag{31}$$

where $\nabla^2 f$ is the Hessian of f (which is symmetric positive semidefinite) and ∇f is the gradient of.We will assume that the problem(31) is solvable,i.e.,there exists an optimal point x^*.

Remark(6),[4] If f is a quadratic(i.e.,differentiable) and convex, we note :

1) The necessary and sufficient condition for the point x^* to be optimal is

$$\nabla f(x^*) = \nabla^2 f(x^*)\, x^* + \nabla f(x^*) = 0 \tag{32}$$

Therefore, the problem (31) can be solved via the optimality condition, $\nabla^2 f(x^*)\, x^* + \nabla f(x^*) = 0$ (It is clear that the secant condition (13) is a special case of the optimality condition).

2) If $\nabla^2 f$ is positive definite,then there is a unique solution

$x^* = -(\nabla^2 f)^{-1} \nabla f$.

Remark(7),[1]:Technique of BFGS Algorithm

Usually the problem(31) must be solved by an iterative algorithms such as Newton methods or quasi-Newton methods. Solving the problem (31) by using BFGS quasi-Newton algorithm means that the algorithm will computes a sequence of points : $x_0, x_1, x_2, \ldots \in \mathrm{dom}(f)$ with $f(x_k) \to f(x^*)$ as $k \to \infty$. Such sequence of points is called a minimizing sequence for the problem (31).The algorithm is terminated when $f(x_k) - f(x^*) \leq \epsilon$, where $\epsilon > 0$ is some specified tolerance.

Remark(8),[11] From above we summarized the following :

1) If x^* is optimal then $\nabla f(x^*) = 0$.

2) If x^* is optimal then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semi definite.

3) If $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite then x^* is optimal .

4) If f is convex and x^* is local optimum then x^* is global optimum .

5) If f is convex and $\nabla f(x^*) = 0$ then x^* is global optimum .

4.5. Initial Point and Sublevel Set [4]

BFGS quasi-Newton method require a suitable starting point x_0.The starting point must lie in $\mathrm{dom}(f)$, and in addition the sublevel set $S = \{x \in \mathrm{dom}(f) : f(x) \leq f(x_0)\}$
must be closed. This condition is satisfied for all $x_0 \in \mathrm{dom}(f)$ if the function f is closed,i.e.,all its sublevel sets are closed.Continuous function with $\mathrm{dom}(f) = R^n$ are closed, so if $\mathrm{dom}(f) = R^n$, the initial sublevel set condition is satisfied by any $x_0 \in R^n$.

4.6.Line Search Strategies and Descents Methods[5]

A line search algorithm searches for decrease in f in a descent direction. To compute the new step update in quasi-Newton methods one can use either the line search or the Trust Region strategies.These two strategies have different properties and are best used with specific Hessian approximation or inverse Hessian approximation updates. The line search will need the Hessian approximation or inverse Hessian approximation to be symmetric and positive definite.

Trust Region methods overcome the problems that line search methods encounter with non-symmetric positive definite approximate Hessians. In particular, a Newton Trust Region strategy allows the use of complete Hessian information even in region where the Hessian has negative curvature (For more details, see[1,6]).

The algorithms described in this work produce a minimizing sequence $\{x_k\}, k = 1,2, \ldots$,where

$$x_{k+1} = x_k + \eta_k d_k \tag{33}$$

where d_k is a vector in R^n called the step or search direction, η_k is the learning rate (step length), with $\eta_k > 0$ (except when x_k is optimal).

Remark(9),[4,5]

1) All the methods we study are descent methods, which means that

$$f(x_{k+1}) < f(x_k) \tag{34}$$

(except when x_k is optimal).This implies that for all k,we have $x_k \in S$ (the initial sublevel set),and in particular we have $x_k \in \mathrm{dom}(f)$.

2) From the convexity we know that$\nabla f(x_k)^T(y - x_k) \geq 0$ implies $f(y) \geq f(x_k)$ so the search direction in a descent method must satisfy :

$$\nabla f(x_k)^T d_k < 0 \tag{35}$$

We call such a direction a descent direction (For f,at x_k).

3) The search directions in the descent methods are called the line search since selection of the step length η determines where along the line $x + \eta d$ the next iterate will be.

Therefore, the descent direction can be defined as

Definition(9),[1] : A vector $d \in R^n$ is a descent direction for f at x if :

$$\frac{df(x+\eta d)}{d\eta}\Big|_{\eta=0} = \nabla f(x)^T d < 0. \tag{36}$$

Remark(10),[4]Convex Quadratic Functions with Descent Directions

Now, we will consider the descent directions based on the convex quadratic models of f of the form

$$m(x) = f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2}(x - x_k)^T H_k (x - x_k) \tag{37}$$

where H_k is the model Hessian (model Hessian approximation with respect to BFGS quasi-Newton method) is symmetric positive definite.

We let $d = x - x_k$ be such that $m(x)$ is minimized. Hence

$$\nabla m(x) = \nabla f(x_k) + H_k(x - x_k) = 0 \tag{38}$$

and therefore, we have

$$d = -H_k^{-1}\nabla f(x_k) \tag{39}$$

The search direction d in (39) is called the Newton step(for f at x_k),also called the quasi-Newton direction(with respect to quasi-Newton method).

Remark(11) The Newton step in eq.(8) (the quasi-Newton direction in eq. (12)) is a descent direction, since

$$\nabla f(x)^T d = -\nabla f(x)^T H_k^{-1} \nabla f(x_k) < 0$$

(from positive definiteness of H_k).

4.6.1. Exact And Inexact Line Search Algorithms[2]

Line search, also called one dimensional search,refers to an optimization procedure for univariate functions. It is the base of multivariable optimization. As stated before, in multivariable optimization algorithms, for given x_k , the iterative scheme is $x_{k+1} = x_k + \eta_k d_k$.

The key is to find the direction vector d_k and a suitable step length η_k.

Let $\emptyset(\eta) = f(x_k + \eta d_k)$. (40)

So, the problem that departs from x_k and finds a step length in the direction d_k such that

$$\emptyset(\eta_k) < \emptyset(0) \quad (41)$$

is just line search about η.

Remark(12),[2]:

1) If we find η_k such that the objective function in the direction d_k is minimized, i.e.,

$$f(x_k + \eta_k d_k) = \min_{\eta>0} f(x_k + \eta d_k) \quad (42)$$

or

$$\emptyset(\eta_k) = \min_{\eta>0} \emptyset(\eta) \quad (43)$$

such a line search is called exact line search or optimal line search, and η_k is called optimal step length.

2) If we choose η_k such that the objective function has acceptable descent amount, i.e.,such that the descent

$$f(x_k) - f(x_k + \eta_k d_k) > 0$$

is acceptable by users,such a line search is called inexact line search,or approximate line search,or acceptable line search.

3) Since, in practical computation, theoretically exact optimal step length generally cannot be found, and it is also expensive to find almost exact step length, therefore the inexact line search with less computation load is highly popular (such as : Wolfe line search, Goldstein line search,backtracking line search,etc).

4) If f is a convex quadratic, $f(x) = \frac{1}{2} x^T Q x + b^T x + c$, its one-dimensional minimizer along the ray $x_k + \eta d_k$ can be computed analytically and is given by :

$$\eta_k = -\frac{\nabla f(x_k)^T d_k}{d_k^T Q d_k} \quad (44)$$

4.6.1.1. BFGS quasi-Newton Method with Wolfe Line Search[1,8]

There are many ways to do a line search in a given direction to find an acceptable step length η_k.Some require the Wolfe conditions to be satisfied(eq.45),others require the Goldstein conditions(more details in[3,6]). The most commonly used line search method is to find the step length that satisfies the Wolfe conditions. The Wolfe line search conditions ensure that the gradients are sampled at points where the model captures important curvature information.

In this work, we will present some theorems that ensure the convergence of the BFGS quasi Newton algorithm with Wolfe line search procedure under strongly convex optimization function.

We require that the step length η_k satisfies **Wolfe conditions** :

(1) $f(x_k + \eta_k d_k) \le f(x_k) + c_1 \eta_k d_k^T \nabla f(x_k)$,

(**2**) $d_k^T \nabla f(x_k + \eta_k d_k) \geq c_2 d_k^T \nabla f(x_k)$. (45)

where $0 < c_1 < c_2 < 1$.

where the first Wolfe condition is called the general sufficient decrease condition).

The following theorems ensure the convergence of the BFGS quasi -Newton algorithm with Wolfe line search procedure.

Theorem (4) : Suppose that f is C^1 and bounded from below. Then Wolfe's line-search terminates (i.e., the number of the line-search iterations is finite).

Proof : see [7], (P. 29). □

Theorem (5) : Given x_0 , let f be a convex function, such that the set $\{x: f(x) \leq f(x_0)\}$ is bounded and such that f has contiuous second derivatives in this set. Let B_0 be any positive definite matrix. Then the BFGS method with step length chosen to satisfy the Wolfe conditions generates a sequence $\{x_k\}$ for $k = 0,1, \ldots$ such that $f(x_k)$ for $k = 0,1, \ldots$ converges to a minimum of f.

Proof : see [5], (P.99).

Note : If the function f is convex and quadratic then f is bounded from below [1].

To derive the Wolfe conditions,it is necessary to present the following theorem :

Theorem (6) [1] : Let f be twice continuously differentiable function in a neighborhood of a point $x \in R^n$. Then for $e \in R^n$ and $\|e\|_2$ sufficiently small,

$$f(x + e) = f(x) + \nabla f(x)^T e + \frac{1}{2} e^T \nabla^2 f(x) e + o(\|e\|_2^2)$$

Remark(13),[2]: The framework of the Wolfe line search procedure is as follows. First, determine or give an initial search interval which contains the desirable step lengths ; then employ a bisection or interpolation formula to compute a good step length within this interval,i.e.,to reduce this interval iteratively until the length of the interval is less than some given tolerance.

It is not difficult to prove that there exist step lengths that satisfy the Wolfe conditions for every function f that is smooth and bounded below.

Theorem (7) : Existence of Acceptable η

Suppose that $f: R^n \to R$ is continuously differentiable. Let d_k be a descent direction at x_k, and assume that f is bounded below along the ray $\{x_k + \eta d_k : \eta > 0\}$. Then if $0 < c_1 < c_2 < 1$,there exist intervals of step lengths satisfying the Wolfe conditions.

Proof : see [2], (P.40). □

Wolfe line search algorithm :

INITIALIZATION : Choose $0 < c_1 < c_2 < 1$, and set $\alpha = 0, \eta = 1,$ and $\beta = +\infty$.

REPEAT

If $f(x + \eta d) > f(x) + c_1 \eta d^T \nabla f(x)$,

set $\beta = \eta$ and reset $\eta = \frac{1}{2}(\alpha + \beta)$.

Else if $d^T \nabla f(x + \eta d) < c_2 d^T \nabla f(x)$,

set $\alpha = \eta$ and reset $\eta = \begin{cases} 2\alpha, & \text{if } \beta = +\infty \\ \frac{1}{2}(\alpha + \beta), & \text{otherwise.} \end{cases}$

Else, STOP.

End REPEAT .

The parameters c_1 and c_2 are typically chosen to be $c_1 = 0.0001$ and $c_2 = 0.9$. (For more details, see [2].

4.6.1.2 BFGS Method with Backtracking Line Search .

Backtracking line search is very simple line search method and quite effective. It depends on two constants α, β where $\alpha \in (0,0.5)$ and $\beta \in (0,1)$.This line search is called backtracking because it starts with unit step length and then reduces it by a factor β until the stopping condition $f(x_k + \eta_k d_k) \leq f(x_k) + \alpha\eta_k \nabla f(x_k)^T d_k$ holds.

Backtracking line search algorithm :

1) Given a descent direction d_k for f at $x \in \mathrm{dom}(f), \alpha \in (0,0.5), \beta \in (0,1)$.

2) $\eta_k := 1$.

3) While $f(x_k + \eta_k d_k) > f(x_k) + \alpha\eta_k \nabla f(x_k)^T d_k$, $\eta_k = \beta\eta_k$.

The parameter α is typically chosen between 0.01 and 0.3 and the parameter β is often chosen to be between 0.1 and 0.8 (For more details, see [4]).

5. Global Convergence of BFGS quasi-Newton Method[10,11]

In our analysis of global convergence for BFGS in the context of an inexact line search, we need the following assumptions :

Assumptions

[1] The objective function f is twice continuously differentiable

[2] The level set $\Omega = \{x \in R^n : f(x) \leq f(x_0)\}$ is convex, and there exist positive constants m_1 and m_2 such that

$$m_1\|z\|^2 \leq z^T \nabla^2 f(x) z \leq m_2\|z\|^2, \text{for all } z \in R^n \text{ and } x \in \Omega. \quad (46)$$

The second assumption implies that the Hessian is positive definite on Ω and that f has a unique minimizer $x^* \in \Omega$.

Also in our analysis of global convergence for BFGS quasi-Newton method, it is necessary to present the following lemma :

Lemma (8) : Let H be symmetric positive definite with smallest and largest eigenvalues $0 < \lambda_s < \lambda_l$. Then for all $z \in R^n$,

$$\lambda_l^{-1}\|z\|^2 \leq z^T H^{-1} z \leq \lambda_s^{-1}\|z\|^2 \quad (47)$$

Proof : see[1],(P.41). □

Theorem (9) : Global Convergence of BFGS

Let B_0 be any symmetric positive definite initial matrix, and let x_0 be a starting point for which the stated assumptions are satisfied.Then the sequence $\{x_k\}$ generated by the BFGS method converges to the minimizer x^* of f.

Proof : see [10] . □

6. Results

In this section we will introduce and prove some important theorems about the convergence of BFGS method with Wolfe line search under strongly convex optimization function .

Theorem(10) : Let f be a convex and differentiable then :

$\nabla f(x_k)^T d_k > 0$ iff $-d_k$ is descent direction for f at x_k

Proof

since f in eq. (31) is quadratic function,then f is differentiable we will prove this theorem by using the descent direction in eq.(39),i.e.,by using the Newton step(resp. the quasi-Newton direction)

$$d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

Firstly, suppose that $-d_k$ is a descent direction for f at x_k

Therefore, we have

$$d_k = \nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

Then this gives

$$\nabla f(x_k)^T d_k = \nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k) > 0$$

(From positive definiteness of $\nabla^2 f(x_k)$).

Therefore, we get $\nabla f(x_k)^T d_k > 0$

Conversely, suppose that $\nabla f(x_k)^T d_k > 0$

Then this gives $-\nabla f(x_k)^T d_k < 0$

This implies that $\nabla f(x_k)^T (-d_k) < 0$

Then by definition(9)

we have $(-d_k)$is descent direction for f at x_k.

Theorem(11) : By successive measurements of the gradient,BFGS quasi-Newton method build a quadratic model of the objective function which is sufficiently good that convergence is achieved.

Proof

The derivation starts with the quadratic model

$$m_k(d) = f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T B_k d$$

where B_k is the Hessian approximation(symmetric positive definite) and d is the quasi-Newton direction (eq. 39), $d_k = -B_k^{-1} \nabla f(x_k)$.

For the equation

$x_{k+1} = x_k + \eta_k d_k = x_k - \eta_k B_k^{-1} \nabla f(x_k)$,

we require that the step length η_k satisfies the Wolfe conditions.

Therefore, for equation, $B_{k+1} = B_K + B_k^U$ we get a new model :

$$m_{k+1}(d) = f(x_{k+1}) + \nabla f(x_{k+1})^T d + \frac{1}{2} d^T B_{k+1} d$$

Clearly, for this to make sense we must impose some conditions on the update[8].We impose two conditions on the new model $m_{k+1}(d)$:

[1,2]: $m_{k+1}(d)$ must match the gradient of the objective function in x_k and x_{k+1}**.**The second condition is satisfied by the construction, since :

$$\nabla m_{k+1}(0) = \nabla f(x_{k+1})$$

(here, 0 is the zero vector).

The first condition given us :

$$\nabla m_{k+1}(-\eta_k d_k) = \nabla f(x_{k+1}) - \eta_k B_{k+1} d_k = \nabla f(x_k)$$

with a little bit of re-arrangement we get :

$$\eta_k B_{k+1} d_k = \nabla f(x_{k+1}) - \nabla f(x_k)$$

Now, from the above equation :

$B_{k+1} \eta_k d_k = \nabla f(x_{k+1}) - \nabla f(x_k) \rightarrow B_{k+1}(x_{k+1} - x_k) = \nabla f(x_{k+1}) - \nabla f(x_k)$,

Then we have the secant condition,

$B_{k+1} p_k = q_k$ where $q_k = \nabla f(x_{k+1}) - \nabla f(x_k)$, $p_k = x_{k+1} - x_k$.

By pre-multiplying the secant equation by q_k^T we see the curvature condition :

$p_k = B_{k+1}^{-1} q_k \rightarrow q_k^T p_k = q_k^T B_{k+1}^{-1} q_k$,

since $q_k^T B_{k+1}^{-1} q_k > 0$,(from positive definiteness of B_{k+1}),

Then this gives

$q_k^T p_k > 0$ (curvature condition)

Theorem(12) : If f is strongly convex, the curvature condition

$q_k^T p_k > 0$ will be satisfied for any two points x_k and x_{k+1} .

Proof

Since f is quadratic strictly convex, then its Hessian (resp.Hessian approximation) is symmetric positive definite(i.e.,non-singular matrix).

By contradiction, if we assume that $q^T p \leq 0$,

From secant condition(eq.13), we have

$Bp = q \quad \rightarrow \quad p = B^{-1}q$

$\rightarrow q^T p = q^T B^{-1} q$,

since by assumption $q^T p \leq 0$, then we have

$q^T B^{-1} q \leq 0$ C! (since B^{-1} is positive definite).

Therefore, we must have $q^T p > 0$.

From theorem(12) we note that the curvature condition holds if we impose another line search condition on the quasi-Newton direction.

Theorem(13) : Global Result

If $f: R^n \rightarrow R$ is strongly convex,BFGS with backtracking line search (resp. with any inexact line search)converges from any $x_0 \in dom(f)$ and any initial symmetric positive definite B_0.

Proof

we will prove that If $f: R^n \rightarrow R$ is quadratic strongly convex then the stated assumptions(section 5) will always holds, and therefore the global convergence will be satisfied by theorem (9)

(1) To prove that f is twice continuously differentiable **:**

since f is quadratic, then f is differentiable.

In fact, f is a polynomial of second degree

Therefore, $f: R^n \rightarrow R$ is twice continuously differentiable function**.**

(2) To prove that The level set $\Omega = \{x \in R^n : f(x) \leq f(x_0)\}$ is convex, and there exist positive constants m_1 and m_2 such that

$m_1 \|z\|^2 \leq z^T \nabla^2 f(x) z \leq m_2 \|z\|^2, \text{for all } z \in R^n \text{ and } x \in \Omega.$ **:**

By subsection (4.5), the sublevel set Ω is exist and satisfied by any initial point $x_o \in dom(f)$. Moreover, since f is convex function,

then the level set Ω is convex .

(by definition(4),sublevel set of convex function is convex).

Since f is quadratic strongly convex, then its Hessian (resp.Hessian approximation) is symmetric positive definite.

Therefore, by lemma (8), we get :

$\lambda_l^{-1} \|z\|^2 \leq z^T \nabla^2 f(x)\, z \leq \lambda_s^{-1} \|z\|^2, \text{for all } z \in R^n \text{ and } x \in \Omega,$

where λ_s and λ_l are smallest and largest eigenvalues of $\nabla^2 f(x)$ respectively.

Now, For the above equation,

if we replace λ_l^{-1} by m_1 and λ_s^{-1} by m_2

we get

$m_1 \|z\|^2 \leq z^T \nabla^2 f(x) z \leq m_2 \|z\|^2, \text{for all } z \in R^n \text{ and } x \in \Omega.$

7. CONCLUSION

In this work we introduced and proved a number of theorems that ensure the global convergence of BFGS quasi Newton method under strongly convex optimization function.We can conclude that By successive measurements of the gradient,BFGS quasi-Newton method under inexact line search algorithms build a quadratic model of the objective function which is sufficiently good that convergence is achieved.For future work one can study the convergence properties of BFGS method under another type of optimization functions.

REFERENCES

[1]. Kelley C. T.," Iterative Methods For Optimization ", North Carolina State University, Raleigh, North Carolina, USA, 1999.
[2]. Nocedal J.,Wright S. J., "Numerical Optimization", Springer-Verlag New York ,1999.
[3]. Tawfiq L. N. M.,"On Design and Training of Artificial Neural Network For Solving Differential Equations",Ph.D. Thesis, College of Education Ibn AL-Haitham,University of Baghdad, Iraq, 2004.
[4]. Boyd S., Vandenberghe L., "Convex Optimization ",Seventh Edition, University Press, Cambridge,UK, 2009.
[5]. Papakonstantinou J. M.,"Historical Development of the BFGS Secant Method and Its Characterization Properties",Ph.D. Thesis, Rice University, Houston, Texas, USA, 2009.
[6]. Ding Y., Lushi E., et al.,"Investigation of Quasi-Newton Methods For Unconstrained Optimization",International Journal of Computer Application,Vol. 29, 48-58, 2010.
[7]. Nikolova M.," Optimization. Applications to Image Processing ", Montevideo, 2012.
[8]. Blomgren P.,"Numerical Optimization. Quasi-Newton Methods, The BFGS Method", Computational Sciences Research Center, San Diego State University, San Diego, 2013.
[9]. Vandenberghe L.," Quasi-Newton Methods ", EE236C (Spring 2013-14), 1-15, 2013.
[10]. Blomgren P.,"Numerical Optimization. Quasi-Newton Methods, Convergence Analysis ", Computational Sciences Research Center, San Diego State University, San Diego, 2014.
[11]. Blomgren P.,"Numerical Optimization. Convergence ;Line Search Methods ", Computational Sciences Research Center, San Diego State University, San Diego, 2014.