UNIVERSITI PUTRA MALAYSIA

MODIFIED QUASI-NEWTON METHODS FOR LARGE-SCALE UNCONSTRAINED OPTIMIZATION

LEONG WAH JUNE

FSAS 2003 60

MODIFIED QUASI-NEWTON METHODS FOR LARGE-SCALE UNCONSTRAINED OPTIMIZATION

By

LEONG WAH JUNE

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Partial Requirement for the Degree of Doctor of Philosophy

January 2003

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy.

MODIFIED QUASI-NEWTON METHODS FOR LARGE-SCALE UNCONSTRAINED OPTIMIZATION

By

LEONG WAH JUNE

January 2003

Chairman: **Associate Professor Malik Hj. Abu Hassan, Ph.D.**

Faculty: **Science and Environmental Studies**

The focus of this thesis is on finding the unconstrained minimizer of a function, when the dimension n is large. Specifically, we will focus on the well-known class of optimization methods called the *quasi-Newton* methods. First we briefly give some mathematical background. Then we discuss the quasi-Newton's methods, the fundamental method in underlying most approaches to the problems of large-scale unconstrained optimization, as well as the related so-called line search methods. A review of the optimization methods currently available that can be used to solve large-scale problems is also given.

The main practical deficiency of quasi-Newton methods is the high computational cost for search directions, which is the key issue in large-scale unconstrained optimization. Due to the presence of this deficiency, we introduce a variety of techniques for improving the quasi-Newton methods for large-scale problems, including scaling the SR1 update, matrix-storage free methods and the extension of modified BFGS updates to limited-memory scheme. Comprehensive theoretical and experimental results are also given.

Finally we comment on some achievements in our researches. Possible extensions are also given to conclude this thesis.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan ijazah Doktor Falsafah

KAEDAH KUASI-NEWTON TERUBAHSUAI UNTUK PENGOPTIMUMAN TAK BERKEKANGAN BERSKALA BESAR

Oleh

LEONG WAH JUNE

Januari 2003

Pengerusi: Profesor Madya Malik Hj. Abu Hassan, Ph.D.

Fakulti: Sains dan Pengajian Alam Sekitar

Penumpuan tesis ini adalah untuk mencari peminimum tak berkekangan bagi suatu fungsi, apabila dimensi n besar. Khususnya, kami akan menumpu kepada suatu k elas kaedah pengoptimuman terkenal yang dipanggil kaedah kuasi-Newton. Pertama, kami memberi secara ringkas sedikit latarbelakang matematik. Kemudian kami membincang kaedah kuasi-Newton, iaitu kaedah asas yang memperihalkan kebanyakan pendekatan kepada masalah pengoptimuman tak berkekangan berskala besar, bersama-sama dengan sesuatu yang berkait dengan kaedah gelintaran garis. Satu sorotan bagi kaedah pengoptimuman sedia ada yang boleh digunakan untuk menyelesaikan masalah berskala besar juga diberi.

Kekurangan utama secara praktik kaedah kuasi-Newton ialah kos pengiraan yang tinggi bagi arah gelintaran, yang menjadi isu utama dalam pengoptimuman tak berkekangan berskala besar. Oleh kerana wujudnya kekurangan tersebut, kami memperkenalkan pelbagai teknik bagi memperbaiki kaedah kuasi-Newton untuk masalah berskala besar, temasuk menskala rumus kemaskini SR1, kaedah bebas-

storan matrik dan lanjutan rumus kemaskini BFGS terubahsuai kepada skema ingatan terhad. Keputusan teori dan berangka yang menyeluruh juga diberikan.

Akhirnya kami memberi komen tentang beberapa pencapaian dalam penyelidikan kami. Kemungkinan lanjutan juga diberi untuk mengakhiri tesis ini.

ACKNOWLEDGEMENTS

I would like to express my most gratitude and sincere appreciation to my chairman, Associate Professor Dr. Malik Hj. Abu Hassan and Dr. Mansor Monsi for their untiring guidance, valuable advice, support and comments. Their patience and persistent encouragement during the course of my research is instrumental to the completion of this thesis. I am also grateful to Dr. Mohd. Rizam Abu Bakar for serving in the supervisory committee.

Special thanks is given to the Head of Department, academic and general staffs of the Department of Mathematics, Universiti Putra Malaysia, for their assistance in various capacities. I am also acknowledged the financial support given to me by Universiti Putra Malaysia under the Graduate Studies Scholarship Scheme.

Last but not least, I would like to thank my family and friends for their understanding support and encouragement throughout the course of this study.

This thesis submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degrees of Doctor of Philosophy. The members of the Supervisory Committee are as follows:

MALIK HJ. ABU HASSAN, Ph.D.
Associate Professor
Faculty of Science and Environmental Studies
Universiti Putra Malaysia
(Chairman)

MANSOR MONSI, Ph.D.
Faculty of Science and Environmental Studies
Universiti Putra Malaysia
(Member)

MOHD. RIZAM ABU BAKAR, Ph.D.
Faculty of Science and Environmental Studies
Universiti Putra Malaysia
(Member)

AINI IDERIS, Ph.D.
Professor/Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

TABLE OF CONTENTS

		Page
ABSTRACT		2
ABSTRAK		4
ACKNOWLEDGEMENTS		6
APPROVAL		7
DECLARATION		9
LIST OF TABLES		13
LIST OF FIGURES		14
LIST OF NOTATIONS		15
CHAPTER		
I	INTRODUCTION	16
	Preliminaries	16
	Problems of Optimization	16
	Existence and Uniqueness of Solutions	18
	Necessary and Sufficient Conditions for Unconstrained Optimization	20
	Convexity	23
	Outline of Thesis	25
II	AN OVERVIEW OF QUASI-NEWTON METHODS	27
	Introduction	27
	Rate of Convergence	28
	Line Search Methods	30
	Quasi-Newton Methods: Theory and Algorithm	34
	Newton's Method	34
	Quasi-Newton Methods: Theory and Convergence	37
	Quasi-Newton Methods: Practical Line Search Algorithms	46
	Conclusions	53
III	LARGE-SCALE UNCONSTRAINED OPTIMIZATION: A SURVEY ON THE METHODS	54
	Introduction	54
	A Review on the Existing Methods	55
	Steepest Descent Method	56
	Newton-type Methods	56
	Nonlinear Conjugate Gradient Methods	61
	Limited Memory BFGS Method (L-BFGS)	63
	Partitioned quasi-Newton Method	65
	Conclusions	66

IV POSITIVE-DEFINITE SCALED SYMMETRIC RANK ONE METHOD 67
Introduction 67
Restart Procedure for SR1 Update 68
Scaling the Identity 71
  Preliminaries 71
  Volume as A Measure for Least Change 73
  The σ – optimal Update 75
Convergence of the SR1 Without Uniform Linear Independence 83
Computational Results 91
Conclusions 93

V MATRIX-STORAGE FREE QUASI-NEWTON METHODS FOR LARGE-SCALE OPTIMIZATION 94
Introduction 94
Matrix-Storage Free Quasi-Newton Methods 94
Preliminary Numerical Experiments on MF Methods 98
Scaling the MF-BFGS and MF-SR1 Method 101
  Properties of the Scaling Factor 102
  MF Methods with Scaling 105
Convergence of the MF Method 107
Computational Results 110
  Comparison between MF-BFGS and MF-SR1 Method 110
  Comparison between MF-BFGS and L-BFGS Method 113
  Comparison between MF-BFGS and Conjugate Gradient Methods 119
  Solving Extremely Large Problems 121
Conclusions 122

VI LIMITED MEMORY PARTIALLY MODIFIED BFGS METHODS FOR LARGE-SCALE OPTIMIZATION 124
Introduction 124
Modified BFGS Methods Using Different Function Interpolation 124
Limited Memory Modified BFGS Methods 127
Convergence of Limited Memory BFGS Update with Modifications 131
  Limited Memory Variations of Quasi-Newton Algorithms 132
  Termination of Limited Memory Methods 137
  Examples of Methods that Do Not Reproduce the Conjugate Gradient Iterates 145
Computational Results 146
Conclusions 154

VII CONCLUSIONS AND SUGGESTIONS FOR FUTURE STUDIES 155
Conclusions 155
Future Studies 157

REFERENCES 160
APPENDICES 167
BIODATA OF AUTHORS 169

LIST OF TABLES

Table		Page
1	Results for NS-SR1 algorithm	70
2	Comparison of the SSR1 and the BFGS methods	92
3	Comparison of MF-SR1 and MF-BFGS methods	100
4	Comparison of MF-SR1-S and MF-BFGS-S	111
5	Storage locations	114
6	MF-BFGS vs L-BFGS	115
7	MF-BFGS vs L-BFGS in term of **ICL**	116
8	MF-BFGS vs CG	120
9	MF-BFGS vs CG in term of **ICL**	121
10	Results of MF-BFGS ($n = 10^6$)	122
11	Comparative results of limited memory Biggs' BFGS, Yuan's BFGS and L-BFGS methods with $m = 5$	148
12	Comparative results of limited memory Biggs' BFGS, Yuan's BFGS and L-BFGS methods with $m = 10$	149
13	Comparative results of limited memory Biggs' BFGS, Yuan's BFGS and L-BFGS methods with $m = 30$	150
14	Comparison of Biggs' BFGS and L-BFGS with $m = 5$	151
15	Comparison of Biggs' BFGS and L-BFGS with m=10	152
16	Comparison o f Biggs' BFGS and LBFGS with m=30	153

LIST OF FIGURES

Figure		Page
1	'Storage-up' vs 'speed-up' with MF-BFGS and L-BFGS (*m*=3)	117
2	'Storage-up' vs 'speed-up' with MF-BFGS and L-BFGS (*m*=7)	118

LIST OF NOTATIONS

1. $\Re^n$ denotes the linear $n-$dimensional Real space.

2. g is the $n \times 1$ gradient vector of a function f, with components

$$g^{(i)} = \frac{\partial f}{\partial x^{(i)}},\ i = 1,2,\ldots,n .$$

3. G is the $n \times n$ Hessian matrix of f, that is the (i,j)th element of G is given by

$$G^{(i,j)} = \frac{\partial^2 f}{\partial x^{(i)} \partial x^{(j)}},\ i = 1,2,\ldots,n \text{ and } j = 1,2,\ldots,n .$$

4. x_k is the k th approximation to $x*$, a minimum of f.

5. g_k is the gradient vector of f at x_k.

6. B_k is an $n \times n$ k th matrix approximation to G.

7. H_k is an $n \times n$ matrix that is a k th approximation to G^{-1}.

8. A^T denotes the transpose of matrix A.

9. $\|y\|$ denotes an arbitrary norm of y.

10. min denotes the minimum.

CHAPTER I

INTRODUCTION

Preliminaries

Many application problems in engineering, decision sciences, and operations research can be formulated as optimization problems. Such applications include digital processing, structural optimization, engineering design, database design and processing, mechanical engineering and chemical process control. Optimal solutions in these applications have significant economical and social impact. Better designs often result in lower implementation and more robust operation under a variety of operating conditions.

Problems of Optimization

Optimization problems are made up of three basic components: a set of unknowns or variables, an objective function to be minimized or maximized, and a set of constraints that specify feasible values of the variables. The optimization problem entails finding values of the variables that optimize (minimize or maximize) the objective function while satisfying the constraints.

In the following, we formally define optimization problems and identify the classes of problems addressed in this thesis.

Problem 1.1. Optimization Problems

A general minimization problem is defined as follows:

Given a set D and a function $f : D \rightarrow P$, find at least one point $x^* \in D$ that satisfies $f(x^*) \leq f(x)$ for all $x \in D$, or show the non-existence of such a point.

A mathematical formulation of a minimization problem is as follows:

$$\begin{aligned} &\text{Minimize } f(x), \\ &\text{subject to } x \in D. \end{aligned} \tag{1.1}$$

In this formulation, $x = (x_1, x_2, \ldots, x_n)^T$ is an $n-$dimensional vector of unknowns. The function f is the objective function of the problem, and D is the feasible domain of x specified by constraints.

Definition 1.1. A vector $x^* \in D$, satisfying $f(x^*) \leq f(x)$ for all $x \in D$ is called a *global minimizer* of f over D. The corresponding value of f is called a *global minimum.*

Definition 1.2. A vector $x^* \in D$ is called a *local minimizer* of f over D if $f(x^*) \leq f(x)$ for all $x \in D$ closed to x^*. The corresponding value of f is called a *local minimum.*

Note that since max $f(x) = -\min(-f(x))$, maximization problems can be transformed into minimization problems shown in (1.1). We use optimization and minimization interchangeably in this thesis.

Optimization problems are further classified into constrained optimization and unconstrained optimization based on the presence of constraints. Problems without constraints fall into the class of *unconstrained optimization*; $D = \Re^n$

The scope of this thesis is limited to unconstrained optimization problems, in which the variables are continuous, f is differentiable, and a local minimizer provides a satisfactory solution. This probably reflects available software as well as the needs of practical applications.

A number of books give substantial attention to unconstrained optimization and are recommended to readers who desire additional information on this topic. These include Ortega and Rheinboldt (1970), Fletcher (1980), Gill et al. (1981), and Dennis and Schnabel (1983).

Existence and Uniqueness of Solutions

Let g denotes the n component *gradient* column vector of first partial derivatives of f; in general

$$[g(x)]_j = \frac{\partial f(x)}{\partial x_j}, \quad j = 1,\mathrm{K}\ ,n. \qquad (1.2)$$

Also, G denotes the $n \times n$ Hessian matrix of second partial derivatives of f; in general

$$[G(x)]_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \ i = 1,\mathrm{K}\ ,n, \ j = 1,\mathrm{K}\ ,n. \tag{1.3}$$

Note that G is symmetric if f is twice continuously differentiable.

Definition 1.3. A function f of the $n-$vector x is said to be *Lipschitz continuous* with constant γ in an open neighbourhood $D \subset \Re^n$, written $f \in Lip_\gamma(D)$, if for all $x, y \in D$,

$$|f(x) - f(y)| \leq \gamma \|x - y\| \tag{1.4}$$

where $\|\cdot\|$ is an appropriate norm.

Most methods for optimizing nonlinear differentiable functions of continuous variables rely heavily upon Taylor series expansions of these functions. We will briefly review the Taylor series expansions used in unconstrained optimization and a few mathematical properties of these expansions.

The fundamental Taylor series expansion used in unconstrained optimization is the first three terms of the Taylor series of f around $\hat{x}$,

$$m(\hat{x} + d) = f(\hat{x}) + g(\hat{x})^T d + \frac{1}{2} d^T G(\hat{x}) d. \tag{1.5}$$

The standard Taylor series with remainder results from the calculus of one variable can also be extended to single valued functions of multiple variables. For any direction $d \in \Re^n$ there exist $\tau_1, \tau_2 \in [0, \tau]$ for which

$$f(x+d)=f(x)+g(x+\tau_1 d)^T d \qquad (1.6)$$

and

$$f(x+d)=f(x)+g(x)^T d+\frac{1}{2}d^T G(x+\tau_2 d)d\,. \qquad (1.7)$$

These results are the keys to the necessary and sufficient conditions for unconstrained minimization that we consider next.

Necessary and Sufficient Conditions for Unconstrained Optimization

Algorithms for solving the unconstrained minimization problem are based upon the first and second order conditions for a point x^* to be a local minimizer of f. These conditions are briefly reviewed in this subsection.

Theorem 1.1. First Order Condition

Let $f:\Re^n \to \Re$ be continuously differentiable, and let $y \in \Re^n$. If $g(y) \neq 0$, then y is not a local minimizer of $f(x)$.

Proof. If $g(y)\neq 0$, then there exist directions $d \in \Re^n$ for which $g(y)^T d<0$; an example is $d=-g(y)$. For any such direction d, we have from (1.6) that

$$f(y+\tau d)-f(y)=\tau d^T g(x+\tau_1 d) \qquad (1.8)$$

for some $\tau_1 \in (0,\tau)$. Also by the continuity of $f(x)$, there exists $\delta>0$ such that $g(x+\tau_1 d)^T d<0$ for any $\tau_1 \in (0,\tau)$. Thus for any step-length $\tau<\delta$,

$$f(y+\tau d)<f(y). \qquad (1.9)$$

Therefore y cannot be a local minimizer of f. □

Directions d for which $g(y)^T d < 0$ are called *descent directions* for f at y. Descent directions play an important role in the numerical methods for unconstrained optimization and will be discussed further in other sections of this thesis.

The above argument in Theorem 1.1 is equivalent to say that for x^* to be a local minimizer, it is necessary that $g(x^*) = 0$. Any point $x^* \in \Re^n$ such that $g(x^*) = 0$ is called *stationary point* of f. To distinguish between minimizers and other stationary points it is necessary to consider the second derivative matrix G defined in (1.3). First we need the definition of a *positive definite* matrix.

Definition 1.4. Let $G \in \Re^{n \times n}$ be symmetric. Then G is positive definite if $v^T G v > 0$ for all nonzero $v \in \Re^n$.

There are several equivalent characterizations of positive definite matrices; another common one is that a symmetric matrix $G \in \Re^{n \times n}$ is positive definite if and only if all its eigenvalues are positive. If $v^T G v \geq 0$ for all v, G is said to be *positive semi-definite.*

Theorem 1.2. Let $f : \Re^n \to \Re$ be twice continuously differentiable, and let $x^* \in \Re^n$. If $g(x^*) = 0$ and $G(x^*)$ is positive definite, then x^* is a local minimizer of f.

Proof. By (1.7) for any $d \in \Re^n$,

$$f(x^*+d) = f(x^*) + g(x^*)^T d + \frac{1}{2} d^T G(x^*+\tau d) d \qquad (1.10)$$

for some $\tau \in (0,1)$. By the continuity of f and the positive definiteness of $G(x^*)$, there exists $\delta > 0$ such that for any direction d with $\|d\| < \delta$ and any scalar τ with $|\tau| \le 1$, $G(x^*+\tau d)$ is positive definite. Thus for any d with $\|d\| < \delta$, we have from (1.10) and $g(x^*) = 0$ that

$$f(x^*+d) > f(x^*). \qquad (1.11)$$

Therefore x^* is a local minimizer of f. □

By a similar argument it is easy to show that a necessary condition for x^* to be a local minimizer of a twice continuously differentiable $f(x)$ is that $g(x^*) = 0$ and $H(x^*)$ is positive semi-definite; in this case, it is necessary to examine higher order derivatives to determine whether x^* is a local minimizer of $f(x)$.

We can make the following summary for second order conditions:

Property 1.1. Second Order Conditions

The *second order necessary* (and *sufficient*) *condition* for $x^* \in \Re^n$ to be a local minimizer of a twice continuously differentiable function f is that the Hessian matrix $G(x^*)$ is positive semi-definite (and positive definite).

If $g(x^*) = 0$ and $G(x^*)$ has both positive and negative eigenvalues, then x^* is said to be a *saddle point* of f. A saddle point is a local minimizer of some cross-section of f and a local maximizer of some other cross-section.

Convexity

A very important concept in minimization theory is that of convexity.

Definition 1.5. A set $\Omega \subseteq \Re^n$ is said to be *convex* if $z = \lambda x + (1-\lambda)y \in \Omega$, for each $x, y \in \Omega$ and $0 \le \lambda \le 1$.

Definition 1.6. A function $f : D \rightarrow \Re$ is said to be *convex* if D is a convex set and in addition,

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y) \qquad (1.12)$$

for each $x, y \in D$ and $0 \le \lambda \le 1$.

It is called *strictly convex* if the inequality (1.12) becomes a strict inequality.

If a function is twice continuously differentiable, one can provide the following convenient necessary and sufficient conditions for convexity.

Theorem 1.3. If $f:\Re^n \to \Re$ is continuously differentiable on a convex set $D \subseteq \Re^n$, and f is strictly convex then

$$f(y) > f(x) + g(x)^T (y-x) \qquad (1.13)$$

for each $x, y \in D$.

Proof. From (1.12),

$$f(\lambda y + (1-\lambda)x) < \lambda f(y) + (1-\lambda) f(x),\ \lambda \in [0,1],$$

$$\frac{f(x+\lambda(y-x)) - f(x)}{\lambda} < f(y) - f(x). \qquad (1.14)$$

By (1.6), (1.14) becomes

$$\frac{\lambda (y-x)^T g(x + \tau_1 \lambda (y-x))}{\lambda} < f(y) - f(x)$$

for some $\tau_1 \in [0,1]$. When $\lambda \to 0$, we have the result. □

Theorem 1.4. If $f:\Re^n \to \Re$ is twice continuously differentiable on a convex set $D \subseteq \Re^n$, and f is strictly convex then its Hessian matrix has positive eigenvalues for each $x \in D$.

Proof. From (1.13),

$$f(y) - f(x) - g(x)^T (y-x) > 0, \qquad (1.15)$$

for each $x, y \in D$.

Taylor series (1.7) for $x, y \in D$ and $\lambda \in [0,1]$ give us the following:

$$f(y) = f(x) + g(x)^T (y-x) + \frac{1}{2}(y-x)^T G(x + \tau_2 \lambda (y-x))(y-x),$$

with $\tau_2 \in [0,1]$.

Then, for $z = x + \tau_2 \lambda (y-x) \in D$ and $v = y - x \in \Re^n$,