UNIVERSITI PUTRA MALAYSIA

A CLASS OF DIAGONALLY PRECONDITIONED LIMITED MEMORY QUASI-NEWTON METHODS FOR LARGE-SCALE UNCONSTRAINED OPTIMIZATION

CHEN CHUEI YEE

FS 2009 29

A CLASS OF DIAGONALLY PRECONDITIONED LIMITED MEMORY QUASI-NEWTON METHODS FOR LARGE-SCALE UNCONSTRAINED OPTIMIZATION

CHEN CHUEI YEE

MASTER OF SCIENCE
UNIVERSITI PUTRA MALAYSIA

2009

A CLASS OF DIAGONALLY PRECONDITIONED LIMITED MEMORY QUASI-NEWTON METHODS FOR LARGE-SCALE UNCONSTRAINED OPTIMIZATION

By

CHEN CHUEI YEE

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Master of Science

September 2009

Abstract of thesis presented to the Senate of Universiti Putra Malaysia
in fulfilment of the requirement for the degree of Master of Science

A CLASS OF DIAGONALLY PRECONDITIONED LIMITED MEMORY QUASI-NEWTON METHODS FOR LARGE-SCALE UNCONSTRAINED OPTIMIZATION

By

CHEN CHUEI YEE

September 2009

Chairman: Leong Wah June, PhD

Faculty: Science

The focus of this thesis is to diagonally precondition on the limited memory quasi-Newton method for large scale unconstrained optimization problem. Particularly, the centre of discussion is on diagonally preconditioned limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method.

L-BFGS method has been widely used in large scale unconstrained optimization due to its effectiveness. However, a major drawback of the L-BFGS method is that it can be very slow on certain type of problems. Scaling and preconditioning have been used to boost the performance of the L-BFGS method.

In this study, a class of diagonally preconditioned L-BFGS method will be proposed. Contrary to the standard L-BFGS method where its initial inverse Hessian approximation is the identity matrix, a class of diagonal preconditioners has been derived based upon the weak-quasi-Newton relation with an additional parameter. Choosing different parameters leads the research to some well-known diagonal updating formulae which enable the *R*-linear convergent for the L-BFGS method.

Numerical experiments were performed on a set of large scale unconstrained minimization problem to examine the impact of each choice of parameter. The computational results suggest that the proposed diagonally preconditioned L-BFGS methods outperform the standard L-BFGS method without any preconditioning.

Finally, we discuss on the impact of the diagonal preconditioners on the L-BFGS method as compared to the standard L-BFGS method in terms of the number of iterations, the number of function/gradient evaluations and the CPU time in second.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia
sebagai memenuhi keperluan untuk ijazah Sarjana Sains

SUATU KAEDAH KELAS KUASI-NEWTON INGATAN TERHAD DENGAN PRAPENSYARAT PEPENJURU BAGI PENGOPTIMUMAN TAK BERKEKANGAN BERSKALA BESAR

Oleh

CHEN CHUEI YEE

September 2009

Pengerusi: Leong Wah June, PhD

Fakulti: Sains

Tumpuan tesis ini adalah mencari prapensyarat pepenjuru untuk kaedah kuasi-Newton ingatan terhad bagi menyelesaikan masalah pengoptimuman tak berkekangan berskala besar. Khususnya, tumpuan perbincangan adalah kepada kaedah Broyden-Fletcher-Goldfarb-Shanno ingatan terhad (L-BFGS) dengan prapensyarat pepenjuru.

Kaedah L-BFGS telah digunakan secara meluas dalam pengoptimuman tak berkekangan berskala besar disebabkan oleh keberkesanannya. Walau bagaimanapun, satu kelemahan utama kaedah L-BFGS adalah ia boleh menjadi perlahan bagi sesetengah masalah. Penskalaan dan prapensyarat telah digunakan untuk meningkatkan prestasi kaedah L-BFGS.

Dalam kajian ini, beberapa cara telah diperiksa untuk menprasyarat kaedah L-BFGS secara pepenjuru. Bertentangan dengan kaedah L-BFGS yang piawai di mana penghampiran songsangan Hessian permulaan merupakan matriks identiti, suatu kelas prapensyarat pepenjuru telah diperolehi berdasarkan kepada hubungan kuasi-Newton lemah dengan satu parameter tambahan. Pemilihan parameter yang berlainan dalam penyelidikan ini membawa kepada beberapa formula kemaskini secara pepenjuru yang terkemuka dan juga mengekalkan penumpuan *R*-linear.

Ujikaji berangka telah dijalankan ke atas satu set masalah peminimuman tak berkekangan berskala besar untuk mengkaji impak setiap pilihan parameter. Keputusan pengiraan mencadangkan bahawa kaedah L-BFGS berprapensyarat pepenjuru adalah lebih baik daripada kaedah L-BFGS tanpa sebarang prapensyarat.

Akhirnya, kami membincangkan tentang impak setiap prapensyarat pepenjuru kepada kaedah L-BFGS dengan perbandingan kepada kaedah L-BFGS piawai. Lanjutan bagi penyelidikan masa depan juga diberi.

ACKNOWLEDGEMENT

First and foremost, thank God for His blessing and wisdom.

I would like to thank and express my gratitude to my supervisor, Dr. Leong Wah June, for his guidance and patience throughout the period of this degree. He has indeed given me a lot of insights on this research and did a thorough checking on my thesis.

Apart from that, I would like to express my thanks towards my supervisory committee members, Professor Dr. Malik B. Hj. Abu Hassan and Associate Professor Dr. Fudziah Ismail, for willing to be part of the committee and for their advice throughout the period of my study.

My appreciation also goes to my family members who have always been there for me. Their endurance and love keeps me moving forward all the time. Not forgetting are my beloved friends and colleagues who have offered advices whenever I need one. No word can express my deepest gratitude at the moment.

I certify that a Thesis Examination Committee has met on 29 September 2009 to conduct the final examination of Chen Chuei Yee on her thesis entitled “A Class of Diagonally Preconditioned Limited Memory Quasi-Newton Methods for Large-Scale Unconstrained Optimization” in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Master of Science.

Members of the Examination Committee were as follows:

Isa Daud, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Lee Lai Soon, PhD
Lecturer
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Mansor Monsi, PhD
Lecturer
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Mustafa Mamat, PhD
Lecturer
Faculty of Science and Techonology
Universiti Malaysia Terengganu
(External Examiner)

BUJANG BIN KIM KUAT, Phd
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 15 October 2009

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Master of Science. The members of the Supervisory Committee were as follows:

Leong Wah June, PhD
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Malik Hj. Abu Hassan, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

Fudziah Ismail, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

____

HASANAH MOHD GHAZALI, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 16 November 2009

DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at Universiti Putra Malaysia or any other institution.

CHEN CHUEI YEE

Date: 08 February 2010

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Description
x	real-valued vectors where $x = (x_1, x_2, \dots, x_n)^T$
$f(x)$ or f	function of x
R	set of real numbers
R^n	set of n dimensional real-valued vectors
$\nabla f(x)$ or $g(x)$	first derivative of $f(x)$ where $\nabla f(x) = g(x) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n}\right)^T.$
$\nabla^2 f(x)$ or $G(x)$	second derivative of $f(x)$, called the Hessian, where $G(x) = \nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{pmatrix}.$
I	identity matrix
y_k	the difference between g_{k+1} and g_k where $y_k = g_{k+1} - g_k$
s_k	the difference between x_{k+1} and x_k where $s_k = x_{k+1} - x_k$
B_k	approximation of the Hessian
H_k	approximation of the inverse Hessian
tr	trace operator
FONC	First-Order Necessary Condition
SONC	Second-Order Necessary Condition
SOSC	Second-Order Sufficient Condition
BFGS	Broyden-Fletcher-Goldfarb-Shanno
L-BFGS	limited memory Broyden-Fletcher-Goldfarb-Shanno
DFP	Davidon-Fletcher-Powell

L-DFP	limited memory Davidon-Fletcher-Powell
SR1	Symmetric Rank One
L-SR1	limited memory Symmetric Rank One

CHAPTER 1

INTRODUCTION

1.1 Preliminaries

Optimization, a subject in applied mathematics, is a fascinating blend of heuristics and rigour, of theory and experiment. It is normally studied as a branch of mathematics and yet it has vast applications in almost every branch of science and technology such as engineering, economics, science and military. Optimization is recognized as the science of determining the optimal or best solutions to certain mathematically defined problems. As optimization as a whole, it is the central to any problem involving decision making by which the optimal solutions to the problems can be found from various schemes.

One of the well-known methods in optimization is the Newton's method. However, the evaluation of the Hessian matrix or its inverse is considered to be impractical and costly. Moreover, convergence to a solution cannot be guaranteed from an arbitrary initial point for a general nonlinear objective functions. With this, the quasi-Newton methods came to rise by which the Hessian matrix (or its inverse) of the function to be minimized need not be computed. A number of algorithms have been proposed under the quasi-Newton scheme.

Since the origin of limited memory methods in 1977, many attentions have been given to limited memory quasi-Newton methods for solving large scale unconstrained optimization problems. Despite the fact that the limited memory methods are often very effective, they can be slow and have limited accuracy, especially for ill-conditioned problem. It is well-known that the limited-memory methods can be greatly accelerated by preconditioning on the initial inverse Hessian approximation.

The centre of this research revolves around solving large scale unconstrained minimization problems using limited memory quasi-Newton methods. Particularly, we focus on solving the problems by diagonally preconditioning on the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method.

Before we begin with the main idea of this research, it is more appropriate to look at the mathematical background of an optimization problem first. Note that the mathematical background discussed in this chapter can be found in Gill et al. (1981), Fletcher (2004), Nocedal and Wright (2006) and Chong and Żak (2008). Additional information on this chapter can be obtained in those references too.

1.2 Optimization Problem

We consider the optimization problem

$$\begin{array}{c} \text{minimize } f(x) \\ \text{subject to } x \in \Omega \end{array} \tag{1.1}$$

where $f: R^n \to R$ is the objective/cost function, the vector $x = (x_1, x_2, \dots, x_n)^T \in R^n$ is a n-vector of independent variables and the set $\Omega \subset R^n$ is the constraint/feasible set.

If $\Omega = R^n$, the optimization problem is regarded as unconstrained optimization problem by which we

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in R^n. \end{array} \tag{1.2}$$

The problem of maximizing f is the same as minimizing $-f$ since there is no loss of generality.

A constrained optimization problem can be written as

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & c_i \geq 0, & i = 1, 2, \dots, k, \\ & c_i = 0, & i = k+1, k+2, \dots, m. \end{array} \tag{1.3}$$

If the objective function and the constraint functions are in the form of linear functions, the problem is regarded as linear programming. Otherwise, the problem becomes a nonlinear programming problem.

Following, we shall discuss on the mathematical background of large scale unconstrained optimization. Since we aim to solve minimization problems, it is necessary to include the definitions of types of minimizer.

Definition 1.1: Suppose that $f: R^n \rightarrow R$ is a real-valued function defined on some set $\Omega \subset R^n$. A point $x^* \in \Omega$ is a local minimizer of f over Ω if there exists $\varepsilon > 0$ such that $f(x) \geq f(x^*)$ for all $x \in \Omega \backslash \{x^*\}$ and $\|x - x^*\| < \varepsilon$.

Definition 1.2: Suppose that $f: R^n \rightarrow R$ is a real-valued function defined on some set $\Omega \subset R^n$. A point $x^* \in \Omega$ is a strict local minimizer of f over Ω if there exists $\varepsilon > 0$ such that $f(x) > f(x^*)$ for all $x \in \Omega \backslash \{x^*\}$ and $\|x - x^*\| < \varepsilon$.

Definition 1.3: Suppose that $f: R^n \rightarrow R$ is a real-valued function defined on some set $\Omega \subset R^n$. A point $x^* \in \Omega$ is a global minimizer of f over Ω if $f(x) \geq f(x^*)$ for all $x \in \Omega \backslash \{x^*\}$.

1.3 Functions and Derivatives

A function $f: R^n \rightarrow R$ is said to be continuously differentiable at $x \in R^n$, if $\frac{\partial f(x)}{\partial x_i}$ exists and is continuous for $i = 1, 2, \dots, n$. The gradient of f at x is given by

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n}\right)^T. \qquad (1.4)$$

A function $f: \Omega \rightarrow R^n$, $\Omega \in R^n$, is said to be continuously differentiable on Ω if it is differentiable on Ω, and the components of f have continuous partial derivatives. It is denoted by $f \in C^1$. It will be denoted as $f \in C^p$ if the components of f have continuous partial derivatives of order p.

A function $f: R^n \to R$ is said to be twice continuously differentiable at $x \in R^n$, if $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ exists and is continuous for $i = 1, 2, \dots, n$. The second derivative of $f(x)$ is written as

$$G(x) = \nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{pmatrix}. \tag{1.5}$$

The matrix $G(x)$ is called the Hessian matrix of the function $f: R^n \to R$ at x and it is an $n \times n$ symmetric matrix if f is continuous.

Definition 1.4: A vector $d \in R^n$, $d \neq 0$, is a feasible direction at $x \in \Omega$ if there exists $\alpha_0 > 0$ such that $x + \alpha d \in \Omega$ for all $\alpha \in [0, \alpha_0]$.

Let $f: R^n \to R$ be a real-valued function and let d be a feasible direction at $x \in \Omega$, the directional derivative of x in the direction of d is denoted by

$$\frac{\partial f(x)}{\partial d} = \lim_{\alpha \to 0} \frac{f(x+\alpha d) - f(x)}{\alpha} = \nabla f(x)^T d. \tag{1.6}$$

Suppose that x and d are given, then $f(x + \alpha d)$ is a function of α and

$$\frac{\partial f(x)}{\partial d} = \frac{\partial}{\partial \alpha} f(x + \alpha d)\Big|_{\alpha=0}. \tag{1.7}$$

By the chain rule,

$$\frac{\partial f(x)}{\partial d} = \frac{\partial}{\partial \alpha} f(x + \alpha d)\Big|_{\alpha=0} = \nabla f(x)^T d = \langle \nabla f(x), d \rangle = d^T \nabla f(x). \tag{1.8}$$

In short, if $\|d\| = 1$ which means d is a unit vector, then $\frac{\partial f(x)}{\partial d}$ or $\langle \nabla f(x), d \rangle$ is the rate of increase of f at the point x in the direction d.

Following, we will discuss on the first order conditions of functions. Let $f: R^n \to R$ be a continuously differentiable function defined on R^n.

Theorem 1.1 First-Order Necessary Condition (FONC):

Let Ω be a subset of R^n and $f \in C^1$ a real-valued function on Ω. If x^* is a local minimizer of f over Ω, then for any feasible direction d at x^*, we have

$$d^T \nabla f(x^*) \geq 0. \tag{1.9}$$

Proof.

Define

$$x(\alpha) = x^* + \alpha d, \qquad x(\alpha) \in \Omega. \tag{1.10}$$

If $\alpha = 0$, it is clear that $x(0) = x^*$. The composite function is defined as:

$$\phi(\alpha) = f\big(x(\alpha)\big). \tag{1.11}$$

By Taylor's theorem,

$$\begin{aligned} f(x^* + \alpha d) - f(x^*) &= \phi(\alpha) - \phi(0) \\ &= \phi'(0)\alpha + o(\alpha) \\ &= \alpha d^T \nabla f\big(x(0)\big) + o(\alpha), \end{aligned} \tag{1.12}$$

where $\alpha \geq 0$.

Thus if $\phi(\alpha) \geq \phi(0)$, that is

$$\begin{aligned} f(x^* + \alpha d) - f(x^*) &\geq 0, \\ f(x^* + \alpha d) &\geq f(x^*), \end{aligned} \tag{1.13}$$

for sufficiently small values of $\alpha > 0$, then we must have $d^T \nabla f(x^*) \geq 0$. ∎

Corollary 1.1 Interior Case:

Let Ω be a subset of R^n and $f \in C^1$ a real-valued function on Ω. *If* x^* is a local minimizer of f over Ω and if x^* is an interior point of Ω, then

$$\nabla f(x^*) = 0. \tag{1.14}$$

Proof.

Suppose that f has a local minimizer x^* that is an interior point of Ω. Because x^* is an interior point of Ω, the set of feasible directions at x^* is the whole of R^n. Thus, for any $d \in R^n$, $d^T \nabla f(x^*) \geq 0$ and $-d^T \nabla f(x^*) \geq 0$. Hence, $d^T \nabla f(x^*) = 0$ for all $d \in R^n$, which implies that $\nabla f(x^*) = 0$. ∎