Problems on convergence of unconstrained optimization algorithms

by

Ya-xiang Yuan

Report No. **ICM-98-028** April 1998

Problems on convergence of unconstrained optimization algorithms*

Ya-xiang Yuan

State Key Laboratory of Scientific and Engineering Computing

Institute of Computational Mathematics and Scientific/Engineering Computing

Chinese Academy of Sciences, POB 2719, Beijing 100080, China

April 1998

Abstract

In this paper we give an review on convergence problems of unconstrained optimization algorithms, including line search algorithms and trust region algorithms. Recent results on convergence of conjugate gradient methods are discussed. Some well-known convergence problems of variable metric methods and recent efforts made on these problems are also presented.

Keywords: unconstrained optimization, convergence, line search, trust region.

1 Introduction

Unconstrained optimization is to minimize a nonlinear function $f(x)$, which can be written as

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

Generally, it is assumed that $f(x)$ is continuous. Numerical methods for problem (1.1) are iterative. An initial point x_1 should be given, and at the $k-$th iteration a new iterate point x_{k+1} is to be computed by using the information at the current iterate point x_k and those at the previous points. It is hoped that the sequence $\{x_k\}$ generated will converge to the solution of (1.1).

Most numerical methods for unconstrained optimization can be classified into two groups, namely line search algorithms and trust region algorithms.

*Research partially supported by Chinese NSF grants 19525101, 19731010 and State key project 96-221-04-02-02.

The object of convergence analysis on unconstrained optimization algorithms is to study the properties of the sequence $\{x_k\}$ generated by the algorithms, for example to prove that $\{x_k\}$ converges to a solution or a stationary point point of (1.1), to study the convergence rate of sequence if it is convergent, and to compare the differences between the convergence performances of different algorithms.

The sequence $\{x_k\}$ generated by an algorithm is said to converge to a point x^* if

$$\lim_{k\to\infty} \|x_k - x^*\| = 0. \tag{1.2}$$

In practical computations, the solution x^* is not available, hence it is not possible to use (1.2) to test convergence. One possible replacement of (1.2) is

$$\lim_{k\to\infty} \|x_k - x_{k-1}\| = 0. \tag{1.3}$$

But, unfortunately, the above limit can not guarantee the convergence of $\{x_k\}$. Therefore, global convergence studies on unconstrained optimization algorithms try to prove the following limit

$$\lim_{k\to\infty} \|g_k\| = 0, \tag{1.4}$$

which ensures that x_k is close to the set of the stationary points where $\nabla f(x) = 0$, or

$$\liminf_{k\to\infty} \|g_k\| = 0, \tag{1.5}$$

which ensures that a least a subsequence of $\{x_k\}$ is close to the set of the stationary points. Throughout the paper, we use the notation $g_k = g(x_k) = \nabla f(x_k)$.

Local convergence analyses study the speeds of convergence of the sequence generated by optimization algorithms. When studying local convergences, we normally assume that the sequence $\{x_k\}$ converges to a local minimum x^* at which the second order sufficient condition is satisfied, namely the matrix $\nabla^2 f(x^*)$ is positive definite. Because without the second order sufficient condition, even Newton's method can converge very slowly. Under the second order sufficient condition and the assumption that $x_k \to x^*$, we have that

$$\nabla^2 f(x^*)(x_k - x^*) = g_k. \tag{1.6}$$

Thus, it is easy to see that the sequence converges Q-superlinearly, namely

$$\|x_{k+1} - x^*\| = o(\|x_k - x^*\|) \tag{1.7}$$

if and only if

$$\|\nabla^2 f(x^*)(x_{k+1} - x_k) - g_k\| = o(\|g_k\|). \tag{1.8}$$

The above relation is equivalent to

$$\|(x_{k+1} - x_k) - (\nabla^2 f(x_k))^{-1} g_k\| = o(\|x_{k+1} - x_k\|). \tag{1.9}$$

Therefore, the main technique for proving Q-superlinear convergence of an optimization algorithm is to prove the step is asymptotically very close to the Newton's step.

In this paper, we focus our attentions to global convergence results without discussing those on local convergence.

2 Line Search Algorithms

A line search algorithm chooses or computes a search direction d_k at the $k-$th iteration, and it sets the next iterate point by

$$x_{k+1} = x_k + \alpha_k d_k, \tag{2.1}$$

where α_k is computed by carrying certain line search techniques. Normally the search direction d_k is so chosen that it is a descent direction unless $\nabla f(x_k) = 0$. Namely,

$$d_k^T \nabla f(x_k) < 0 \tag{2.2}$$

if $\nabla f(x_k) \neq 0$. There are two categories of line searches: the exact line search and inexact line searches. In the exact line search, α_k is computed to satisfy

$$f(x_k + \alpha_k d_k) = \min_{\alpha \geq 0} f(x_k + \alpha d_k). \tag{2.3}$$

One commonly used inexact line search is the Wolfe line search which finds a $\alpha_k > 0$ satisfying

$$f(x_k + \alpha_k d_k) \leq f(x_k) + c_1 \alpha_k d_k^T \nabla f(x_k) \tag{2.4}$$

and

$$d_k^T \nabla f(x_k + \alpha_k d_k) \geq c_2 d_k^T \nabla f(x_k), \tag{2.5}$$

where $c_1 \leq c_2$ are two constants in $(0, 1)$. Usually $c_1 \leq 0.5$. The strong Wolfe line search requires α_k satisfying (2.4) and

$$|d_k^T \nabla f(x_k + \alpha_k d_k)| \leq -c_2 d_k^T \nabla f(x_k). \tag{2.6}$$

Another famous inexact line search is the Armijo line search [1] which sets $\alpha_k = \delta^t \bar{\alpha}$ where $\bar{\alpha} > 0$ is a positive constant and t is the smallest non-negative integer for which

$$f(x_k + \delta^t \bar{\alpha} d_k) \leq f(x_k) + c_1 \delta^t \bar{\alpha} d_k^T \nabla f(x_k). \tag{2.7}$$

Assume that the gradient $g(x) = \nabla f(x)$ is Lipschitz continuous, that is

$$\|g(x) - g(y)\| \leq L\|x - y\|, \quad \forall x, y \in R^n, \tag{2.8}$$

all the inexact line searches imply that

$$\alpha_k \geq c_3 \frac{-d_k^T g_k}{\|d_k\|^2} \tag{2.9}$$

where c_3 is some positive constant. Thus it follows from (2.9), (2.7) and (2.4) that

$$f(x_k) - f(x_{k+1}) \geq c_1 c_3 \frac{(-d_k^T g_k)^2}{\|d_k\|^2}. \tag{2.10}$$

For the exact line search, if (2.8) holds, there also exists positive constant c_4 such that

$$f(x_k) - f(x_{k+1}) \geq c_4 \frac{(-d_k^T g_k)^2}{\|d_k\|^2}. \tag{2.11}$$

Thus, if $\{f(x_k)\}$ is bounded below (which is always true when we assume that (1.1) is bounded below), (2.10) and (2.11) show that

$$\sum_{k=1}^{\infty} \frac{(-d_k^T g_k)^2}{\|d_k\|^2} < \infty. \tag{2.12}$$

Let θ_k be the angle between the steepest descent direction and the search direction d_k. By definition, we have that

$$\cos^2 \theta_k = \cos^2 \langle -g_k, d_k \rangle = \frac{(-d_k^T g_k)^2}{\|d_k\|^2 \|g_k\|^2}. \tag{2.13}$$

This relation and inequality (2.12) show that

$$\sum_{k=1}^{\infty} \|g_k\|^2 \cos^2 \theta_k < \infty \tag{2.14}$$

if $f(x_k)$ is bounded below.

From the above inequality it is easy to establish the following convergence results.

Theorem 2.1 *([12, 13, 15]) Let $\{x_k\}$ be the sequence generated by a line search algorithm under the exact line search, or any inexact line search such that (2.10) holds. If*

$$\sum_{k=1}^{\infty} \cos^2 \theta_k = \infty, \tag{2.15}$$

then the sequence is convergent in the sense that

$$\liminf_{k \to \infty} \|g_k\| = 0. \tag{2.16}$$

Furthermore, if there exists a positive constant η such that

$$\cos^2 \theta_k \geq \eta \tag{2.17}$$

for all k, then the sequence is strongly convergent in the sense that

$$\lim \|g_k\| = 0. \tag{2.18}$$

The above theorem is the most fundamental and also important result in the convergence analysis of line search algorithms that use gradients. For line search algorithms that use

$$d_k = -B_k^{-1} g_k, \tag{2.19}$$

where B_k is some positive definite matrix. It is easy to see that

$$\cos^2\theta_k = \frac{(g_k^T B_k^{-1} g_k)^2}{\|g_k\|^2 \|B_k^{-1} g_k\|^2} \geq \frac{1}{Tr(B_k)Tr(B_k^{-1})}. \tag{2.20}$$

Line search directions of Quasi-Newton methods have the form of (2.19). Hence, almost all convergence analyses on quasi-Newton methods are based on the estimations of $Tr(B_k)$ and $Tr(B_k^{-1})$, and the bounds on $Det(B_k)$. An elegant example can be seen in [9] where the BFGS method is proved to be convergent for general convex functions with Wolfe line searches. Powell's result is extented to all methods in Broyden's convex family except the DFP method by Byrd, Nocedal and Yuan [2]. Some progress has been made by Yuan [14] on the global convergence of the DFP method. The DFP method is proved to converge to the solution in some situations when additional conditions are satisfied. For example, the following theorem is one of the results obtained.

Theorem 2.2 *([14]) If the DFP method applied to a uniformly convex function satisfies*

$$\|g_{k+1}\|_2 \leq \|g_k\|_2 \tag{2.21}$$

for all k, then x_k generated by the method converges to the unique minimum of the objective function.

From the proofs in [14], one can easily see that condition (2.21) can be replaced by

$$g_k^T y_k \leq 0. \tag{2.22}$$

By estimating the trace of B_k^2, [14] proves that the DFP method converges if (2.22) is replaced by

$$g_k^T B_k y_k \leq 0. \tag{2.23}$$

A conjecture is that the DFP method converges if the following inequality

$$g_k^T (B_k)^m y_k \leq 0, \tag{2.24}$$

holds for all k, where m is any given positive integer. But the proof for such a result may be very difficult for a general positive integer m, since it would require to study the traces of the matrices B_k^{m+1}.

However, it is still an open question whether the DFP method with Wolfe line search is convergent for all convex functions, without assuming any additional conditions. The answer for this question is still unknown, even if we assume that the objective function is a uniformly convex function.

If we do not assume the convexity of the objective function, the convergence problem of variable metric methods is very difficult. It is still not known what kind of line search conditions can ensure a quasi-Newton method to be convergent. For example, Powell [10] gives a very interesting 2-dimensional example that shows that variable metric methods fail to converge even if the line search condition is that α_k be a local mimum of the function $f(x_k + \alpha d_k)$ that satisfies $f(x_k + \alpha_k d_k) < f(x_k)$.

For conjugate gradient methods, search directions are generated by

$$d_{k+1} = -g_{k+1} + \beta_k d_k. \tag{2.25}$$

Normally it is assumed that the parameter β_k is so chosen that the following sufficient descent condition

$$-d_k^T g_k \geq c_5 \|g_k\|^2 \tag{2.26}$$

is satisfied for some positive constant c_5 (for example, see [7]). The proofs on the convergence of conjugate gradient methods are mainly on the estimation of

$$\sum_{k=1}^{\infty} cos^2\theta_k = \sum_{k=1}^{\infty} \frac{1}{\|d_k\|^2} \left(\frac{(d_k^T g_k)^2}{\|g_k\|^2} \right). \tag{2.27}$$

If $\frac{(d_k^T g_k)^2}{\|g_k\|^2}$ is bounded away from zero, namely there exists a positive constant c_6 such that

$$\frac{(d_k^T g_k)^2}{\|g_k\|^2} \geq c_6, \qquad \forall k, \tag{2.28}$$

it follows from Theorem 2.1 and (2.27) that the condition

$$\sum_{k=1}^{\infty} \frac{1}{\|d_k\|^2} = \infty \tag{2.29}$$

implies the convergence of a conjugate gradient method. Therefore, in the convergence analysis of a conjugate gradient method, a widely used technique is to derive a contradiction by establishing (2.29) if there exists a positive constant c_7 such that

$$\|g_k\| \geq c_7, \qquad \forall k. \tag{2.30}$$

It is easy to see that, under the assumption (2.30) and the boundedness of $\|g_k\|$, (2.28) is equivalent to (2.26). However, it is can be seen that (2.26) is not always necessary. Indeed, we only require (2.26) to be satisfied in the mean value sense. Recent convergence results obtained by Dai and Yuan [4] use the technique that the mean value of $-d_k^T g_k$ over every two consecutive iterations is bounded aways from zero . That is to say, we can replace (2.26) by

$$\frac{(d_k^T g_k)^2}{\|g_k\|^4} + \frac{(d_{k+1}^T g_{k+1})^2}{\|g_{k+1}\|^4} \geq c_5, \qquad \forall k. \tag{2.31}$$

We have the following theorem.

Theorem 2.3 *Let $\{x_k\}$ be the sequence generated by conjugate gradient method (2.25) with the exact line search, or any inexact line search such that (2.10) and (2.6) hold. If $\{f(x_k)\}$ is bounded below, if $\{\beta_k\}$ is bounded, and if (2.29) holds, the the method converges in the sense that (2.16) holds.*

Proof Suppose that (2.16) is not true, there exists a positive constant c_7 such that (2.30) holds. It follows from (2.6) that

$$\frac{d_{k+1}^T g_{k+1}}{\|g_{k+1}\|^2} = -1 + \beta_{k+1} \frac{d_k^T g_{k+1}}{\|g_{k+1}\|^2}, \tag{2.32}$$

which gives that

$$\begin{aligned} 1 &\leq -\frac{d_{k+1}^T g_{k+1}}{\|g_{k+1}\|^2} + |\beta_{k+1}| \frac{|d_k^T g_{k+1}|}{\|g_{k+1}\|^2} \\ &\leq -\frac{d_{k+1}^T g_{k+1}}{\|g_{k+1}\|^2} + c_2 |\beta_{k+1}| \frac{\|g_k\|^2}{\|g_{k+1}\|^2} \frac{|d_k^T g_k|}{\|g_k\|^2} \\ &\leq \sqrt{1 + c_2^2 |\beta_{k+1}|^2 \|g_k\|^2 \|g_{k+1}\|^{-2}} \times \sqrt{\frac{(d_{k+1}^T g_{k+1})^2}{\|g_{k+1}\|^4} + \frac{(d_k^T g_k)^2}{\|g_k\|^4}}. \end{aligned} \tag{2.33}$$

Now the above inequality and our assumptions imply that there exists a positive constant c_5 such that (2.31) holds. It follows from Theorem 2.1 and (2.31) that

$$\sum_{k=1}^{\infty} \min \left[\frac{1}{\|d_{2k-1}\|^2}, \frac{1}{\|d_{2k}\|^2} \right] < \infty, \tag{2.34}$$

which shows that

$$\max[\|d_{2k-1}\|, \|d_{2k}\|] \to \infty. \tag{2.35}$$

This, (2.25) and the boundedness of $\|g_k\|$ show that

$$\|d_{2k}\| \leq \frac{1}{2} \max[\|d_{2k-1}\|, \|d_{2k}\|] + |\beta_{2k-1}| \|d_{2k-1}\|, \tag{2.36}$$

which implies that

$$\|d_{2k}\| \leq (2|\beta_{2k-1}| + 1/2) \|d_{2k-1}\|. \tag{2.37}$$

It follows from the above inequality, (2.34) and the boundedness of β_k that

$$\sum_{k=1}^{\infty} \frac{1}{\|d_{2k}\|^2} < \infty. \tag{2.38}$$

Repeating the above analysis with the indices $2k-1$ and $2k$ replaced by $2k$ and $2k+1$, we can prove that

$$\sum_{k=1}^{\infty} \frac{1}{\|d_{2k+1}\|^2} < \infty. \tag{2.39}$$

Therefore it follows that

$$\sum_{k=1}^{\infty} \frac{1}{\|d_k\|^2} < \infty, \tag{2.40}$$

which contradicts our assumption. Thus the theorem is true. □

From the above theorem, one can easily see that an essential technique for proving convergence of conjugate gradient methods is to obtain some bounds on the increasing rate of $\|d_k\|$ so that (2.29) holds. One normal way to estimate the bounds on $\|d_k\|$ is to use the relation (2.25) recursively. Therefore it is quite often that convergence results are established under certain inequality about β_k. For example, as proved in [3], a conjugate gradient method is convergent if there exists a positive constant c_8 such that

$$\|g_k\|^2 \sum_{j=1}^{k-1} \prod_{i=j}^{k-1} \left(\beta_i \frac{\|g_i\|^2}{\|g_{i+1}\|^2} \right)^2 \leq c_8 k, \tag{2.41}$$

holds.

3 Trust Region Algorithms

Trust region algorithms do not carry out line searches. A trust region algorithm generates a new point which lies in the trust region, and decides whether it accepts the new point or rejects it. At each iteration, the trial step s_k is normally calculated by solving the "trust region subproblem":

$$\min_{d \in \Re^n} g_k^T d + \frac{1}{2} d^T B_k d = \phi_k(d) \tag{3.1}$$

$$s.\ t. \qquad ||d||_2 \leq \Delta_k \tag{3.2}$$

where B_k is an $n \times n$ symmetric matrix which approximates the Hessian of $f(x)$ and $\Delta_k > 0$ is a trust region radius. A trust region algorithm uses

$$r_k = \frac{Ared_k}{Pred_k} = \frac{f(x_k) - f(x_k + s_k)}{\phi_k(0) - \phi_k(s_k)}. \tag{3.3}$$

to decide whether the trial step s_k is acceptable and how the nex trust region radius is chosen. (3.3) is the ratio between the actual reduction and the predicted reduction in the objective function. A general trust region algorithm for unconstrained optimization can be given as follows.

Algorithm 3.1 .

Step 1 *Given* $x_1 \in \Re^n$, $\Delta_1 > 0$, $\epsilon \geq 0$, $B_1 \in \Re^{n \times n}$ *symmetric;*
$0 < \tau_3 < \tau_4 < 1 < \tau_1$, $0 \leq \tau_0 \leq \tau_2 < 1$, $\tau_2 > 0$, $k := 1$.

Step 2 *If* $||g_k||_2 \leq \epsilon$ *then stop;*
Find an approximate solution of (3.1)-(3.2), s_k.

Step 3 Compute r_k;

$$x_{k+1} = \begin{cases} x_k & \textit{if } r_k \le \tau_0 \ , \\ x_k + s_k & \textit{otherwise} \ ; \end{cases} \tag{3.4}$$

Choose Δ_{k+1} that satisfies

$$\Delta_{k+1} \in \begin{cases} [\tau_3||s_k||_2, \ \tau_4\Delta_k] & \textit{if } r_k < \tau_2, \\ [\Delta_k, \ \tau_1\Delta_k] & \textit{otherwise;} \end{cases} \tag{3.5}$$

Step 4 Update B_{k+1};
$k := k+1$; go to Step 2.

The constants τ_i (i=0,..,4) can be chosen by users. Typical values are $\tau_0 = 0, \tau_1 = 2, \tau_2 = \tau_3 = 0.25, \tau_4 = 0.5$. For other choices of those constants, please see [6], [5], [8], [11], etc.. The values of constants τ_i (i=1,..,4) make no difference in the convergence proofs of trust region algorithms. However, whether $\tau_0 > 0$ or $\tau_0 = 0$ will lead to very different convergence results and, more important, requires different techniques in the proofs.

Theorem 3.2 *Assume that $f(x)$ is differentiable and $\nabla f(x)$ is uniformly Lipschitz continuous. Let x_k be generated by Algorithm 3.1 with s_k satisfies*

$$\phi_k(o) - \phi_k(s_k) \ge \tau \|g_k\| \min\{\Delta_k, ||g_k||_2/||B_k||_2\} \ , \tag{3.6}$$

where τ is some positive constant. If M_k defined by

$$M_k = 1 + \max_{1 \le i \le k} ||B_k||_2 \tag{3.7}$$

satisfy that

$$\sum_{k=1}^{\infty} \frac{1}{M_k} = \infty, \tag{3.8}$$

if $\epsilon = 0$ is chosen in Algorithm 3.1, and if $\{f(x_k)\}$ is bounded below, then it follows that

$$\liminf_{k \to \infty} \quad ||g_k||_2 = 0. \tag{3.9}$$

Moreover, if $\tau_0 > 0$ and if $\{\|B_k\|\}$ is bounded, then

$$\lim_{k \to \infty} \quad ||g_k||_2 = 0. \tag{3.10}$$

Proof For the proof of the theorem when $\tau_0 = 0$ and under the condition (3.8), please see Powell [11]. We only prove the easier part of the theorem, namely we only consider the case when $\tau_0 > 0$ and $\{\|B_k\|\}$ is bounded. Under these assumptions, we can easily see that there exists a positive constant τ_5 such that

$$Ared_k \ge \tau_5 \|g_k\| \min[\Delta_k, \|g_k\|]. \tag{3.11}$$

Because $f(x_k)$ is bounded below, the above inequality implies that

$$\sum_{k=1}^{\infty} \|g_k\| \min[\Delta_k, \|g_k\|] < \infty. \tag{3.12}$$

The uniformly Lipschitz continuity of $\nabla f(x)$ and the above relation give that

$$\sum_{k=1}^{\infty} \|g_k\| \min[|\|g_{k+1}\| - \|g_k\||, \|g_k\|] < \infty. \tag{3.13}$$

The above inequality shows that either (3.10) is true or

$$\lim_{k\to\infty} \|g_k\| > 0. \tag{3.14}$$

If (3.10) is not true, it follows from (3.12) and (3.14) that

$$\sum_{k=1}^{\infty} \Delta_k < \infty, \tag{3.15}$$

which yields that $\Delta_k \to 0$. This would imply that

$$Ared_k/Pred_k \to 1, \tag{3.16}$$

which shows that $\Delta_{k+1} \geq \Delta_k$ for all sufficiently large k. This contradicts (3.15). The contradiction shows that (3.10) holds. $\square$

It follows from the above theorem that a trust region algorithm converges if there exists a positive constant $\bar{\tau}$ such that

$$\|B_k\| \leq \bar{\tau} k \tag{3.17}$$

holds for all k. The estimation of the predicted reduction (3.6) is crucial in the convergence analyses. If we assume that there is a positive constant $\hat{\tau}$ such that the computed trial step s_k satisfies

$$\phi_k(o) - \phi_k(s_k) \geq \hat{\tau}[\phi_k(0) - \min_{\alpha\in[0,\Delta/\|g_k\|]} \phi_k(-\alpha g_k)] \tag{3.18}$$

for all k. Then, we can see that condition (3.17) can be replaced by the following weaker inequality

$$g_k^T B_k g_k \leq \bar{\tau} k \|g_k\|^2, \tag{3.19}$$

because (3.19) and (3.18) implies that

$$\phi_k(o) - \phi_k(s_k) \geq \frac{\hat{\tau}}{2}\|g_k\| \min\{\Delta_k, \|g_k\|/(\bar{\tau}k)\}. \tag{3.20}$$

References

[1] L. Armijo, "Minimization of functions having Lipschitz continuous partial derivatives", *Pacific J. Math.* 16(1966) 1–3.

[2] R. Byrd, J. Nocedal and Y. Yuan, "Global convergence of a class of variable metric algorithms", *SIAM J. Numer. Anal.* 24(1987) 1171-1190.

[3] Y. Dai, J. Han, G. Liu, D. Sun, H. Yin and Y. Yuan, "Convergence properties of nonlinear conjugate gradient methods", Research Report. ICMSEC-98-24, Institute of Comp. Math. and Scientific/Engineering Computing, Chinese Academy of Sciences, Beijing, China, 1998.

[4] Y. H. Dai and Y. Yuan, "Convergence properties of the Fletcher-Reeves method", *IMA J. Numer. Anal.* 16 (1996) 155–164.

[5] R. Fletcher, "A model algorithm for composite NDO problem", *Math. Prog. Study* 17(1982) 67-76. (1982a)

[6] R. Fletcher, *Practical Methods of Optimization* (second edition) (John Wiley and Sons, Chichester, 1987)

[7] J. C. Gilbert and J. Nocedal, "Global convergence properties of conjugate gradient methods for optimization", *SIAM J. Optimization* 2 (1992) 21–42.

[8] J.J. Moré, "Recent developments in algorithms and software for trust region methods", in: A. Bachem, M. Grötschel and B. Korte, eds., *Mathematical Programming: The State of the Art* (Springer-Verlag, Berlin, 1983) pp. 258-287.

[9] M.J.D. Powell, "Some global convergence properties of a variable metric algorithm for minimization without exact line searches", in: R.W. Cottle and C.E. Lemke, eds., *Nonlinear Programming, SIAM-AMS Proceeding vol. IX* (SIAM publications, Philadelphia, 1976) pp. 53-72.

[10] M.J.D. Powell, "Nonconvex minimization calculations and the conjugate gradient method", in: D.F. Griffiths, ed., *Numerical Analysis* Lecture Notes in Mathematics 1066 (Springer-Verlag, Berlin, 1984) pp. 122-141.

[11] M.J.D. Powell, "On the global convergence of trust region algorithm for unconstrained optimization", *Math. Prog.* 29(1984) 297–303.

[12] P. Wolfe, "Convergence conditions for ascent methods", *SIAM Rev.* 11 (1969) 226–235.

[13] P. Wolfe, "Convergence conditions for ascent methods II: some corrections", *SIAM Rev.* 13 (1971) 185–188.

[14] Y. Yuan, “Convergence of DFP algorithm”, *Science in China(Series A)* 38(1995) 1281–1294.

[15] G. Zoutendijk, “Nonlinear programming, computational methods”, in: J. Abadie, ed., *Integer and Nonlinear Programming* (North-holland, Amsterdam, 1970), pp. 37–86.