On convergence and restart conditions for a nonlinear conjugate gradient method

Arnold Neumaier

Institut für Mathematik, Universität Wien
Strudlhofgasse 4, A-1090 Wien, Austria
email: neum@cma.univie.ac.at
www: http://solon.cma.univie.ac.at/~neum

August 1997

Abstract. A global convergence theorem for unconstrained minimization algorithms with an efficient line search is derived. The theorem applies to a new version of the conjugate gradient method derived here in terms of minimizing the effect of zigzagging. The global convergence condition makes much weaker demands on the line search than previous methods.

Local Q-linear convergence in a neighborhood of a strong local minimizer follows under additional conditions that define natural restart conditions based on line search accuracy and loss of conjugacy of successive gradients.

Keywords: local optimization, global convergence, conjugate gradients, conjugate descent, restart conditions, loss of conjugacy, efficient line search, zigzagging

1991 MSC Classification: 90C30

1 Introduction

We consider the problem of finding a local minimizer of a real valued function $f : \mathbb{R}^n \to \mathbb{R}$ in n variables. We shall assume that f is continuously differentiable with computable gradients $g(x) = f'(x)^T$. For the discussion of convergence rates, we shall also assume that the gradients are locally Lipschitz continuous, so that in some neighborhood of a minimizer, there is a constant γ such that

$$\|g(x) - g(x')\| \leq \gamma \|x - x'\|. \tag{1}$$

If n is large and no structural properties of f are available, one often uses conjugate gradient methods (Fletcher & Reeves [2]; cf. the surveys by Powell [11] and Gilbert & Nocedal [4]) or limited memory quasi-Newton methods (e.g., the popular L-BFGS method of Liu & Nocedal [5]) for the minimization. (Alternatively, one may use truncated Newton methods and solve the linear systems by conjugate gradients; see, e.g., Nash [7]. However, truncated Newton methods are unrelated to the present paper.)

These methods use a Hessian approximation B (often diagonal) as a preconditioner, and the search direction may be viewed as a linear combination of the simplified Newton direction

$$p_N = -B^{-1}g \tag{2}$$

and a direction from a low-dimensional subspace that is adapted during the iteration. In both nonlinear conjugate gradient methods and limited-memory quasi-Newton methods, the search direction is motivated by conjugacy properties of the gradients for quadratic objective functions, at least in the initial phase of the algorithm. Also in both cases, restart techniques are needed to recover from a deterioration of this low-dimensional subspace by reverting to the simplified Newton direction (2).

In this paper, we shall look at general methods that calculate their search direction as a linear combination of the simplified Newton direction p_N and another direction p_C, and we discuss conditions on p_C that guarantee global convergence, and approximate conjugacy conditions that imply local Q-linear convergence in a neighborhood of a strong local minimizer. Recall that in a neighborhood of a strong local minimizer, f is uniformly convex, i.e., there is a positive constant γ' such that

$$\gamma'\|x-x'\|^2 \le (g(x)-g(x'))^T(x-x'). \tag{3}$$

The new convergence conditions are derived from a general convergence result discussed in Section 2, and are applied in Section 4 to provide global convergence of a new version of the conjugate gradient method, assuming only that the line search is efficient in the sense of WARTH & WERNER [12]. Virtually all line search procedures that have been used in provably convergent unconstrained minimization algorithms have this property. Local Q-linear convergence in a neighborhood of a strong local minimizer follows under simple restart conditions similar to those discussed by POWELL [10].

Notation. In the following, we consider a sequence x^l ($l = 0, 1, 2, \ldots$) of points in $\mathbb{R}^n$ produced by some algorithm designed for local optimization. To simplify the typography and to improve the readability of the formulae we shall write

$$f_l := f(x^l), \qquad g^l := g(x^l).$$

When l is arbitrary or clear from the context, the letters f, x, g denote f_l, x^l, g^l; the symbols $\bar{f}$, $\bar{x}$, $\bar{g}$ denote f_{l+1}, x^{l+1}, g^{l+1}, and the symbols f_{old}, x_{old}, g_{old} denote f_{l-1}, x^{l-1}, g^{l-1}, respectively. Similar conventions will apply for other quantities indexed by l.

We assume that a fixed positive definite matrix B is available that can be considered as a *Hessian approximation*. Related to B is the norm

$$\|p\| := \sqrt{p^TBp} \tag{4}$$

and the associated dual norm

$$\|g\|^* = \|B^{-1}g\| = \sqrt{g^TB^{-1}g}, \tag{5}$$

satisfying the *Cauchy-Schwarz inequality*

$$|g^Tp| \le \|g\|^*\|p\| \quad \text{for all } g, p \in \mathbb{R}^n. \tag{6}$$

These norms will be used throughout. In the special case where $B = I$, both norms reduce to the Euclidean norm, and if desired, the general case could be reduced to this case by a coordinate transform.

2 A global convergence theorem

We consider optimization methods based on linear line searches that find for each x and each descent direction p with $g^Tp < 0$ a step $s = \alpha p$ such that $f(x+s) < f(x)$.

We shall call a line search procedure *efficient* if there is a number $\delta > 0$ such that the generated step always satisfies

$$f(x+s) \leq f - \delta(g^Ts)^2/\|s\|^2. \tag{7}$$

Given an efficient linear line search procedure, an optimization method is characterized by the choice of the search direction p at each iteration. The iterate x is modified by adding to x a multiple $s = \alpha p$ ($\alpha > 0$) of the search direction,

$$\bar{x} = x + s = x + \alpha p.$$

Warth & Werner [12] proved that for functions with Lipschitz continuous gradients, efficient linear line searches together with a uniform bound

$$\frac{g^Tp}{\|g\|^*\|p\|} < -\delta' < 0 \tag{8}$$

on the cosines of the angle between gradient and search direction imply local Q-linear convergence in a neighborhood of a strong local minimizer. Condition (8) is very important in theory, but is hardly ever used in practice since for good performance (acceptance of $p = \hat{x} - x$ close to a strong local minimizer $\hat{x}$) the constant δ' cannot be chosen a priori but must depend on the condition of the problem.

The following theorem provides new sufficient conditions for global convergence and local Q-linear convergence that are readily implementable.

2.1. Theorem. *(i) For any optimization method that chooses all its search directions such that, for some positive constant κ,*

$$\frac{p^TBp}{(g^Tp)^2} \leq \frac{p_{\text{old}}^TBp_{\text{old}}}{(g_{\text{old}}^Tp_{\text{old}})^2} + \frac{\kappa}{g^TB^{-1}g} \tag{9}$$

and its steps by an efficient linear line search, we have

$$\inf_{l\geq 0}\|g^l\|=0 \quad \mathit{or} \quad \inf_{l\geq 0} f_l=-\infty. \tag{10}$$

(ii) Suppose that, in addition, there are fixed positive constants η, η' such that

$$\frac{g^T p_{\rm old}}{g_{\rm old}^T p_{\rm old}}\leq \eta, \quad \frac{g^T B^{-1} g_{\rm old}}{g_{\rm old}^T B^{-1} g_{\rm old}}\leq \eta' \leq \frac{1}{2} \tag{11}$$

at each step where the simplified Newton direction is not chosen. If f has Lipschitz continuous gradients and the x^l converge to a strong local minimizer then convergence is locally Q-linear.

Proof. Without loss of generality we may assume that all search directions are scaled such that

$$g^T p=-1, \tag{12}$$

(i) Suppose $\inf\|g^l\|=\epsilon>0$. Then

$$\epsilon\leq\|g\|=\|BB^{-1}g\|\leq\|B\|\,\|B^{-1}g\|,$$

and by (9) and (12),

$$\|p\|^2-\|p_{\rm old}\|^2\leq\kappa/\|B^{-1}g\|^2\leq\kappa\|B\|^2/\epsilon^2. \tag{13}$$

Summation over all steps gives

$$\|p^l\|^2\leq(l+1)\kappa', \qquad \kappa'=\max(\|p^0\|^2,\kappa\|B\|^2/\epsilon^2).$$

Since line searches are efficient, (7) implies

$$f-\bar{f}\geq\frac{\delta(g^T s)^2}{\|s\|^2}=\frac{\delta(g^T p)^2}{\|p\|^2}=\frac{\delta}{\|p\|^2}\geq\frac{\delta}{\kappa'}\cdot\frac{1}{l+1}.$$

Summation over all steps gives

$$f_l\leq f_0-\frac{\delta}{\kappa'}\sum_{i=1}^{l}\frac{1}{i}\to-\infty.$$

(ii) To prove the second part, we only need to verify a bound (8) on the cosines. For simplified Newton steps, this bound holds by (6) for any $\delta' < 1$. For the other steps we may use (11), which gives

$$\begin{array}{rcl} \|B^{-1}(g - g_{\text{old}})\|^2 & = & (g - g_{\text{old}})^T B^{-1}(g - g_{\text{old}}) \\ & = & g^T B^{-1} g - 2g^T B^{-1} g_{\text{old}} + g_{\text{old}}^T B^{-1} g_{\text{old}} \\ & \geq & g^T B^{-1} g. \end{array}$$

On the other hand, the Lipschitz condition (1) gives

$$\|g - g_{\text{old}}\| = \|g(x^l) - g(x^{l-1})\| \leq \gamma \|x^l - x^{l-1}\| = \gamma \|s_{\text{old}}\|,$$

so that

$$\begin{array}{rcl} \|B^{-1}g\| & = & \sqrt{g^T B^{-1} g} \leq \|B^{-1}(g - g_{\text{old}})\| \\ & \leq & \|B^{-1}\| \|g - g_{\text{old}}\| \leq \gamma \|B^{-1}\| \|s_{\text{old}}\|. \end{array}$$

Since f is uniformly convex in a neighborhood of a strong local minimizer, we also have

$$\begin{array}{rcl} \gamma' \|s_{\text{old}}\|^2 & = & \gamma' \|x^l - x^{l-1}\|^2 \\ & \leq & (g(x^l) - g(x^{l-1}))^T (x^l - x^{l-1}) \\ & = & (g - g_{\text{old}})^T s_{\text{old}}. \end{array}$$

If we multiply the last two inequalities, we find

$$\gamma' \|s_{\text{old}}\|^2 \|B^{-1}g\| \leq \gamma \|B^{-1}\| \|s_{\text{old}}\| (g - g_{\text{old}})^T s_{\text{old}}.$$

With the constant $\sigma = \gamma \|B^{-1}\| / \gamma'$, this simplifies to

$$\|s_{\text{old}}\| \|B^{-1}g\| \leq \sigma (g - g_{\text{old}})^T s_{\text{old}}.$$

By scaling, the same holds with p_{old} in place of s_{old}. Therefore, using (11) and (12), we find

$$\|p_{\text{old}}\| \|B^{-1}g\| \leq \sigma (g - g_{\text{old}})^T p_{\text{old}} \leq \sigma(\eta + 1) = \sigma_1.$$

If we solve this for $\|p_{\text{old}}\|$ and use (13) we get

$$\|p\|^2 \leq (\sigma_1^2 + \kappa) / \|B^{-1}g\|^2,$$

and with (12) we find the cosine bound (8) with $\delta = \sqrt{\sigma_1^2 + \kappa}$. Therefore we have local Q-linear convergence. □

A major virtue of the theorem is that the violation of condition (11) defines a practical restart criterion which ensures that we will not generate search directions that are much poorer than those obtained by the simplified Newton direction.

The first inequality is a Wolfe-like condition that involves the directional derivative at x in the old search direction and measures the accuracy of the old line search. In particular, the quotient vanishes when the last line search was exact.

In modern line search codes it is customary to use a two-sided Wolfe condition

$$\left|\frac{g^T p_{\rm old}}{g_{\rm old}^T p_{\rm old}}\right| \le \eta$$

to terminate the line search. However, satisfying the two-sided Wolfe condition may be very expensive in the following two cases: (i) at starting points where the directional derivative is very small, and (ii) in directions where the function is nearly linear in an extended region and then suddenly becomes very steep. The line search by MORÉ & THUENTE [6], for example, used in many modern optimization codes, returns for such directions the warning "WARNING: INTERVAL TOO SMALL FOR FURTHER PROGRESS", but only after having spent a huge number of function values to bracket the minimizer along the line very accurately. (Interesting univariate functions to study this poor behavior are $x^2/(x^2+1)$, $\sqrt{x^2+10^{-4}}$, $x^2+x^3+10^{-6}x^4$ and $|x|^{1.1}$ for large starting values.)

In a subsequent paper (NEUMAIER [8]) we shall give a robust efficient line search that is free of this defect by not enforcing the Wolfe condition. The above convergence theorem is interesting in this context since, instead of enforcing this condition in the line search, we may use the more liberal line search and regard a failure of this condition (after an initial quadratic interpolation step that is mandatory for conjugate gradient methods) as a sign that the function is so strongly nonquadratic along the search direction that a restart is appropriate.

The second inequality is a measure of deviation from conjugacy; the quotient vanishes for conjugate gradient methods with exact line search on quadratic objective functions. At the same time, the second inequality serves as a measure of lack of progress: it is violated whenever very little progress is made so that $x \approx x_{\text{old}}$, $g \approx g_{\text{old}}$ and the quotient is nearly 1.

Similar restart procedures were suggested by Powell [10] and Nocedal [9].

3 A minimum deviation principle

In this section we consider what happens when we choose each search direction p as the vector with

$$g^T p = g_{\text{old}}^T p_{\text{old}} \tag{14}$$

closest with respect to the norm (4) to some (unnormalized) designed vector p_C. This keeps $g^T p$ constant throughout the iteration; hence if the first search direction is chosen such that $g^T p = \gamma < 0$, all search directions will be descent directions.

3.1. Theorem. *Among all $p \in \mathbb{R}^n$ with* (14)*, the distance $\|p - p_C\|$ becomes minimal for $p = \hat{p}$, where*

$$\hat{p} = p_C - \lambda B^{-1} g, \tag{15}$$

$$\lambda = \frac{g^T p_C - g_{\text{old}}^T p_{\text{old}}}{g^T B^{-1} g}. \tag{16}$$

Proof. λ is chosen such that the vector (15) satisfies $g^T \hat{p} = g_{\text{old}}^T p_{\text{old}}$. For an arbitrary vector p with (14), write $q = p - p_C$. Then

$$g^T q = g_{\text{old}}^T p_{\text{old}} - g^T p_C = -\lambda g^T B^{-1} g,$$

so that

$$\begin{aligned} \|p - \hat{p}\|^2 &= \|q + p_C - \hat{p}\|^2 = \|q + \lambda B^{-1} g\|^2 \\ &= (q + \lambda B^{-1} g)^T B (q + \lambda B^{-1} g) \\ &= q^T B q + 2\lambda g^T q + \lambda^2 g^T B^{-1} g = \|q\|^2 - \lambda^2 g^T B^{-1} g. \end{aligned}$$

Hence

$$\|p - p_C\|^2 = \|p - \hat{p}\|^2 + \lambda^2 g^T B^{-1} g$$

for all p with (14), and this expression is minimal for $p = \hat{p}$. □

Let us call an unconstrained minimization method that uses efficient line searches along search directions defined by (15) a *corrected simplified Newton method* if, after a line search in direction p, the next search direction is based on a vector p_C with

$$\|p_C\| \leq \|p\|.$$

The two most natural choices are $p_C = 0$ (used at the start and at restarts) and $p_C = p_{\text{old}}$ (discussed in the next section).

3.2. Algorithm. *Corrected simplified Newton method*

Purpose: finds local minimizer of $f(x)$ (or a stationary point only)

Input: x (initial guess), B (preconditioner)

start:
  $f = f(x)$, $g = g(x)$,
  $p_N = -B^{-1}g$, $\omega = -g^T p_N$
  if $\omega \leq 0$ stop ! x stationary
  $p_C = 0$, $\alpha = 1$, $\gamma = -\omega$

repeat:
  $\lambda = (g^T p_C - \gamma)/\omega$
  $p = p_C + \lambda p_N$
  modify α by an efficient line search
  $f_{\text{old}} = f$, $\omega_{\text{old}} = \omega$
  $x = x + \alpha p$, $f = f(x)$, $g = g(x)$
  $p_N = -B^{-1}g$, $\omega = -g^T p_N$
  pick p_C such that $\|p_C\| \leq \|p\|$
until $\omega \leq 0$ or ($f = f_{\text{old}}$ and $\omega \geq \omega_{\text{old}}$)

A practical implementation must of course use better stopping tests.

3.3. Theorem. *Every corrected simplified Newton method satisfies the convergence condition* (9) *of Theorem* 2.1 *with* $\kappa = 1$. *In particular, the method is globally convergent, and it converges Q-linearly in a neighborhood of a strong local minimizer if* $p_C = 0$ *whenever the conditions* (11) *are violated.*

Proof. By (15) and (16) we have

$$\begin{aligned} \|p\|^2 &= p^TBp = (p_C - \lambda B^{-1}g)^T B(p_C - \lambda B^{-1}g) \\ &= p_C^T B p_C - 2\lambda g^T p_C + \lambda^2 g^T B^{-1} g \\ &= \|p_C\|^2 + (\gamma^2 - (g^T p_C)^2)/g^T B^{-1} g \\ &\leq \|p_C\|^2 + \gamma^2/\|B^{-1}g\|^2. \end{aligned}$$

Since $g^Tp = g_{\text{old}}^T p_{\text{old}} = -\gamma < 0$ and $\|p_C\| \leq \|p_{\text{old}}\|$ in a corrected simplified Newton method, (9) holds with $\kappa = 1$. □

4 A new conjugate gradient method

The oldest and most basic optimization method, steepest descent, exhibits in long steep valleys the phenomenon of zigzagging, a frequent strong change of search direction, that is a main source of inefficiency. In order to reduce the amount of zigzagging we shall choose the search direction p as the vector with $g^Tp = -1$ closest with respect to the norm (4) to the previous search direction; thus we use the preceding results with $p_C = p_{\text{old}}$ and $\gamma = 1$.

Surprisingly, the resulting algorithm is a new version of the nonlinear conjugate gradient method. Indeed, the traditional conjugate gradient search directions have the form

$$p = -B^{-1}g + \beta p_{\text{old}}, \tag{17}$$

with the original Fletcher-Reeves formula

$$\beta = \frac{g^T B^{-1} g}{g_{\text{old}}^T B^{-1} g_{\text{old}}} \tag{18}$$

(Fletcher & Reeves [2]) or various other formulas for calculating β that are equivalent to (18) for exact line searches and quadratic objective functions. For details see the surveys by Powell [11] and Gilbert & Nocedal [4].

4.1. Theorem. *For exact line searches, the corrected simplified Newton method with $p_C = p_{\text{old}}$ and $g^Tp = -1$ is equivalent with the Fletcher-Reeves method.*

Proof. Since $g^T p_{\text{old}} = 0$ for exact line searches, the expression (16) becomes $\lambda = 1/g^T B^{-1} g > 0$, and the Fletcher-Reeves formula (18) can be written as $\beta = \lambda_{\text{old}}/\lambda > 0$. Thus the conjugate gradient direction (17) can be scaled to give

$$\lambda p = \beta \lambda p_{\text{old}} - \lambda B^{-1} g = \lambda_{\text{old}} p_{\text{old}} - \lambda B^{-1} g.$$

A comparison with the corrected simplified Newton method (15) shows that for $p_C = p_{\text{old}}$, the search directions (17) and (15) are related by $\hat{p} = \lambda p$ and point in the same direction since $\lambda > 0$. □

However, for inexact line searches, the Fletcher-Reeves direction and other traditional conjugate gradient search directions are descent directions only under stronger assumptions on the line search (Wolfe condition with certain parameter restrictions). The first global convergence result for a conjugate gradient method (indeed for the Fletcher-Reeves method) with a practical line search was obtained by Al-Baali [1], and again a Wolfe condition is needed to guarantee convergence of these versions.

Because of the equivalence with the Fletcher-Reeves method, a restart in the corrected simplified Newton method with $p_C = p$ implies finite termination for functions that locally become exactly quadratic, when line searches are done with a forced quadratic interpolation step after the first function evaluation (which is mandatory for all conjugate gradient methods). Indeed, for all conjugate gradient methods with exact line searches on a quadratic function we have

$$g^T p_{\text{old}} = g^T B^{-1} g_{\text{old}} = 0,$$

so that the no-restart condition (11) is always satisfied.

Conclusion

A new global and local convergence analysis for unconstrained minimization algorithms was discussed. The sufficient conditions for global convergence are readily implementable, and the sufficient conditions for local Q-linear convergence in a neighborhood of a strong local minimizer translate directly into implementable conditions based on line search accuracy and loss of conjugacy of successive gradients.

The efficiency conditions on the line search needed for convergence are the same as those for algorithms with angles between search direction and gradient bounded away from 90^o, although no such bound is assumed for the global convergence proof. Thus the conditions are considerably weaker than those for traditional convergence analyses of algorithms based on conjugate directions, which need Wolfe conditions that are hard to satisfy on some classes of problems.

The convergence conditions hold for a large class of algorithms called corrected simplified Newton methods and defined by a minimum deviation principle. In particular, one can apply this principle to minimize the amount of zigzagging. Then one obtains an algorithm that – when exact line searches are used – is equivalent to the preconditioned conjugate gradient method with Fletcher-Reeves search directions.

However, for inexact line searches, the Fletcher-Reeves direction and other traditional conjugate gradient search directions are descent directions only under stronger assumptions on the line search (Wolfe condition with certain parameter restrictions). On the other hand, our new conjugate gradient method always produces descent directions, and global convergence holds assuming only the much weaker efficiency condition for the line search.

References

1. M. Al-Baali, Descent property and global convergence of the Fletcher-Reeves method with inexact line search, IMA J. Numer. Anal., 5 (1985), 121-124.

2. R. Fletcher and C.M. Reeves, Function minimization by conjugate gradients, Computer J. 7 (1964), 149–154.

3. R. Fletcher, Practical Methods of Optimization, Wiley, New York 1987.

4. J.C. Gilbert and J. Nocedal, Global convergence properties of conjugate gradient methods for optimization, SIAM J. Optimization 2 (1992), 21-42.

5. D.C. Liu and J. Nocedal, On the limited memory BFGS method for large scale optimization methods, Math. programming 45 (1989), 503-528.

6. J.J. Moré and D.J. Thuente, Line search algorithms with guaranteed sufficient decrease, ACM Trans. Math. Software, 20 (1994), 286–307.

7. S.G. Nash, Solving nonlinear programming problems using truncated-Newton techniques, pp. 119-136 in: Numerical Optimization 1984 (P.T. Boggs et al., eds.), SIAM, Philadelphia 1985.

8. A. Neumaier, A fast and robust efficient line search. In preparation. (See `http://solon.cma.univie.ac.at/~neum/software/ls/` for MATLAB code.)

9. J. Nocedal, Theory of algorithms for unconstrained opimization, pp. 199-242 in: Acta Numerica 1992 (A. Iserles, ed.), Cambridge Univ. Press, Cambridge 1992.

10. M.J.D. Powell, Restart procedures for the conjugate gradient method, Math. Programming 12 (1977), 241-254.

11. M. J. D. Powell, Convergence properties of algorithmus for nonlinear optimization, SIAM Rev. 28 (1986), 487-500.

12. W. Warth and J. Werner, Effiziente Schrittweitenfunktionen bei unrestringierten Optimierungsaufgaben, Computing 19 (1977), 59-72.