Damped Inexact Quasi-Newton Methods*

by

Trond Steihaug+

Department of Mathematical Sciences
Rice University
Houston, Texas

MASC TR 81-3
Revised May 1984

*Presented at the NATO Advanced Research Symposium, Cambridge, U.K., July 1981

+This research was supported by the Norwegian Research Council for Science and the Humanities and DoE Contract #EY-76-S-05-5046. Current address: Statoil, Stavanger, Norway.
Abstract

The inexact quasi-Newton methods are very attractive methods for large scale optimization since they require only an approximate solution of the linear system of equations for each iteration. In this paper we discuss the use of the conjugate gradient method to find an approximate solution of the linear system. To achieve global convergence results, we adjust the step using a backtracking strategy. We discuss the backtracking strategy in detail and show that this strategy has similar convergence properties as one obtains by using line searches with the Goldstein-Armijo conditions. The combination of backtracking and inexact quasi-Newton method with the conjugate gradient method is particularly attractive since the conditions for convergence are easily met. We give conditions for Q-linear and Q-superlinear convergence.
1. Introduction

An important nonlinear problem is the unconstrained optimization problem. Namely the problem of finding a local minimizer \( x^* \) of a function \( f: \mathbb{R}^n \rightarrow \mathbb{R} \), i.e.,

\[
f(x^*) \leq f(x) \text{ for all } ||x-x^*|| < \varepsilon
\]  \hspace{1cm} (1.1)

for some \( \varepsilon > 0 \). In this work we need the following conditions on the function \( f \): (Standard hypothesis)

A1. \( f \) is bounded below on \( \mathbb{R}^n \);
A2. \( f \) is continuously differentiable on \( \mathbb{R}^n \);
A3. The gradient, \( g \), of \( f \) is Lipschitz continuous\(^1\), i.e. there exists \( L \geq 0 \) such that

\[
||g(x)-g(y)|| \leq L||x-y|| \text{ for all } x \text{ and } y \text{ in } \mathbb{R}^n.
\]

A necessary condition for \( x^* \) to be a local minimizer is that

\[
g(x^*) = 0. \hspace{1cm} (1.2)
\]

An attractive way to solve the nonlinear system of equations (1.2) is the use of a quasi-Newton method or inexact quasi-Newton method. These methods generate a sequence of iterates \( \{x_k\} \) that approximate a solution of (1.2). Let \( \{\nu_k\} \) be a sequence of non negative real numbers and let \( B_k \) be a symmetric matrix that approximates the Hessian matrix of \( f \) at \( x_k \). The inexact quasi-Newton method of [15] is given below.

Given \( x_0 \) and \( B_0 \)

FOR \( k = 0 \) STEP 1 UNTIL Convergence DO

Find some \( p_k \) that satisfies

\[
||B_k p_k + g(x_k)|| \leq \nu_k ||g(x_k)||
\]  \hspace{1cm} (1.3)

Let \( x_{k+1} = x_k + p_k \)

Update to obtain \( B_{k+1} \).

An inexact quasi-Newton method will typically apply an iterative method\(^2\) to the linear system \( B_k p = -g(x_k) \) and terminate when the residual \( r = B_k p + g(x_k) \) satisfies (1.3). How-

\(^1\)For the analysis to follow uniform continuity of \( g \) would suffice.
\(^2\)For convenience, we call the overall iterations generating \( x_k \) the outer iterations and the iterations used to calculate \( p_k \) the inner iterations.
ever, this approach is only locally convergent \([9,15]\), i.e. we can only guarantee convergence when \(x_0\) is close to \(x^*\) and \(B_k, k \geq 0\), are close to the Hessian matrix of \(f\) at \(x^*\). If \(B_k\) is the Hessian matrix of \(f\) at \(x_k\) (i.e. the Jacobian matrix of \(g\) at \(x_k\)), then we have the inexact Newton method [5]. For this case Dembo and Steihaug [6] used conjugate gradient inner iterations and needed to monitor a Raleigh quotient to guarantee that the method was well defined in the case when the Hessian matrix was indefinite. In Section 2 we discuss the Conjugate gradient method that was used in [6].

In order to enlarge the region of convergence, we will take advantage of the additional property that \(g\) is the gradient of the function \(f\) and force \(\{f(x_k)\}\) to be decreasing. Recall that a descent method generates for each iterate \(x_k\) a direction \(p_k\) of local descent in the sense that there exists \(\lambda_k^*\) so that

\[
 f(x_k + \lambda_k p_k) < f(x_k), \quad \text{for } 0 < \lambda_k \leq \lambda_k^*.
\]

The next iterate \(x_{k+1}\) is of the form

\[
 x_{k+1} = x_k + \lambda_k p_k. \tag{1.4}
\]

The framework of the algorithms we will consider is then

Given \(x_0\):

FOR \(k = 0\) STEP 1 UNTIL Convergence DO

Compute descent direction \(p_k\);

Compute step correction \(\lambda_k\);

Let \(x_{k+1} = x_k + \lambda_k p_k\).

In this paper, we will discuss relations between the direction \(p_k\) and the step correction \(\lambda_k\). The importance of the results is that we show how to balance the conditions on the line search procedure and the inexact quasi-Newton method with Hessian approximations so that we may achieve the desired convergence properties and rate without losing efficiency in the implementation. In Section 5 we discuss the above framework and describe a global inexact quasi-Newton method.
If \( g(x_k) \neq 0 \) it can be shown that \( p_k \) is a direction of local descent or descent direction if and only if \( g(x_k)^T p_k < 0 \). A natural decrease condition for the descent direction \( p_k \) is

\[
f(x_k + \lambda_k p_k) \leq f(x_k) + \alpha \lambda_k g(x_k)^T p_k .
\]

(1.5)

For some fixed if \( 0 < \alpha < 1 \) we know that such \( \lambda_k \) exists. To preclude excessively small steplengths, we may also require a "bottom line" condition on \( \lambda_k \) like

\[
f(x_k + \lambda_k p_k) \geq f(x_k) + \beta \lambda_k g(x_k)^T p_k ,
\]

(1.6)

\[
g(x_k + \lambda_k p_k)^T p_k \geq \beta g(x_k)^T p_k ,
\]

(1.7)

or

\[
|g(x_k + \lambda_k p_k)^T p_k| \leq - \beta g(x_k)^T p_k .
\]

(1.8)

We must require that \( \alpha \) and \( \beta \) satisfy \( 0 < \alpha < \beta < 1 \) to ensure that such a \( \lambda_k \) can be found [8]. We will choose \( \alpha < 1/2 \). This will guarantee that if \( f \) is a quadratic function, then the exact minimizer \( \lambda_k \) along \( p_k \) will satisfy (1.5). It is a well known fact [8] that a descent method combining (1.5) and either (1.6), (1.7), or (1.8) will satisfy

\[
\lim_{k \to \infty} \frac{g(x_k)^T p_k}{||p_k||} = 0 .
\]

(1.9)

The conclusion that

\[
\lim_{k \to \infty} g(x_k) = 0
\]

(1.10)

usually requires additional properties of the sequence \( \{p_k\} \). The sequence \( \{p_k\} \) is often said to be gradient related to \( \{x_k\} \) if (1.9) implies (1.10) [13]. We say\(^3\) that the method is convergent if (1.10) holds and the method is said to be globally convergent if (1.10) holds for any \( x_0 \in \mathbb{R}^n \) [11].

Conditions (1.7) and (1.8) play an important role in specific quasi-Newton methods called secant methods. In particular if (1.7) or (1.8) holds then the update in many secant methods like the BFGS and DFP methods can be shown to be positive definite (see Dennis and Moré [7] and the references therein). However, the satisfaction of these conditions is not sufficient to guarantee positive definiteness of many other updates like the fixed norm.

\(^3\)Daniel [4] says that the sequence \( \{p_k\} \) is admissible if (1.9) holds, and if (1.10) holds, then \( \{x_k\} \) is a criticizing sequence.
least change secant updates and the updates of Steihaug [16,17]. Dembo and Steihaug [6] used exact Hessian matrices and the conditions (1.5) and (1.7) to achieve global convergence and they showed under reasonable assumptions that the conjugate gradient method generated \( p_k \) for which \( \lambda_k = 1 \) satisfied the conditions (1.5) and (1.6), (1.7), or (1.8) close to a solution of (1.1). The linesearch routine we will used was introduced by Dennis and Schnabel [8] and is discussed in Section 3.

In this paper, we discuss conditions on the line search so that (1.9) holds for descent methods without requiring (1.6), (1.7), or (1.8). Our line search routine is a modification of the backtracking strategy or Armijo [1] algorithm. Dennis and Schnabel [8] have generalized the backtracking strategy, and the method we give in Section 3 is a minor modification of the one in [8]. However, to our knowledge this is the first time its convergence properties have been considered. We show that this routine returns with \( \lambda_k = 1 \) close to a solution for quasi-Newton methods with appropriate Hessian approximations.

In Section 2 we discuss the use of conjugate gradient inner iterations to find an approximate solution (1.3) and in Section 3 we discuss descent methods using the generalized backtracking strategy. We give conditions that imply that the descent method satisfies (1.9) and is globally convergent. We show that for the inexact quasi-Newton methods using conjugate gradient inner iterations and the backtracking strategy, the conditions for global convergence are easily met under very weak conditions on the update sequence of Hessian approximations \( \{B_k\} \). The convergence results in this section are similar to those given in [14,18,20] using a trust-region strategy and the Marwil [12] and Toint [19] update.

In Section 4 we discuss convergence rates. We show that under very weak conditions on the descent directions \( \{p_k\} \), the descent method is Q-linearly convergent. Further, we give conditions which guarantee that for the inexact quasi-Newton method and the backtracking strategy from Section 3 we have that \( \lambda_k = 1 \) and a Q-superlinear convergence rate.

In Section 5, we discuss different methods that satisfy the conditions for convergence developed in the previous sections.
2. Inner Iterations

The underlying assumption in the inexact quasi-Newton methods for the nonlinear system of equations (1.3) is that the residual, \( r_k \), approximates the gradient, \( g \), at the new point \( x_k + p_k \), i.e.

\[ g(x_k + p_k) \approx r_k = B_k p_k + g(x_k). \]

However, since we "globalize" the inexact quasi-Newton method by monitoring the function \( f \) (with gradient \( g \)) it is natural to consider an approximate function \( \varphi_k \) of \( f \) in a neighborhood of \( x_k \), i.e., \( \varphi_k(p) \approx f(x_k + p) \) where \( r_k \) is the gradient of \( \varphi_k \) at \( p_k \), \( g(x_k) \) the gradient at 0, \( B_k \) is the Hessian of \( \varphi_k \), and \( \varphi_k(0) = f(x_k) \). So we choose a quadratic function

\[ \varphi_k(p) = \frac{1}{2} p^T B_k p + p^T g(x_k) + f(x_k). \]

Since we are minimizing \( f \), we will try to minimize the quadratic model \( \varphi_k(p) \). The conjugate gradient method will minimize the quadratic function provided \( B_k \) is positive definite. In the algorithm below, we terminate the inner iteration if a Raleigh quotient is smaller than a constant \( \varepsilon \). This will make the method well defined. The unsubscripted vectors \( p \) and \( g \), and the matrix \( B \) take on the outer iteration subscript \( k \) and should read \( p_k \), \( g(x_k) \), and \( B_k \). The constant \( \varepsilon \) is fixed and does not depend on the outer iteration index.

Conjugate Gradient Inner Iterations:

Given a positive constant \( \varepsilon \);

\( p^0 = 0, r^0 = -g, d^0 = r^0 \).

FOR \( i = 0 \) STEP 1 UNTIL Termination DO

If \( (d^i)^T B d^i \leq \varepsilon \| d^i \|^2 \) then Terminate;

Let \( p^{i+1} = p^i + \alpha_i d^i \) where \( \alpha_i = \frac{\|r^i\|^2}{(d^i)^T B d^i} \);

Let \( r^{i+1} = r^i - \alpha_i B d^i \);

Check Termination;

Let \( d^{i+1} = r^{i+1} + \beta_i d^i \) where \( \beta_i = \frac{\|r^{i+1}\|^2}{\|r^i\|^2} \);

Set \( p := p^i \) if \( i > 0 \) otherwise set \( p := -g \).
We will use the following result found in Dembo and Steihaug [6, Appendix]

**Lemma 2.1:** Let \( \varepsilon > 0 \) and let \( p \) be selected according to the Conjugate Gradient Inner Iterations. Let

\[
\sigma = \min \left\{ \frac{1}{\|B\|} , 1 \right\} \text{ and } \gamma = \max \left\{ \frac{n}{\varepsilon} , 1 \right\}.
\]

Then

\[
g^T p \leq - \sigma \|g\|^2 , \quad \text{and} \quad \|p\| \leq \gamma \|g\|.
\]

If \( i > 0 \) then

\[
(Bp + g)^T p = 0. \tag{2.2}
\]

For simplicity, assume that \( B \) is positive definite and consider the quadratic function

\[
\varphi(p) = \frac{1}{2} p^T B p + g^T p
\]

and define the Cauchy step

\[
p^c = - \frac{\|g\|^2}{g^T B g} g.
\]

Then \( p^c \) satisfies

\[
\varphi(p^c) = \min_{\alpha} \varphi(\alpha p^c) = - \frac{1}{2} \frac{\|g\|^2}{g^T B g} \|g\|^2 \leq - \frac{1}{2 \|B\|} \|g\|^2.
\]

Let \( p \) be any direction so that \( \varphi(p) \leq \varphi(p^c) \) then we have

\[
g^T p \leq - \frac{1}{2 \|B\|} \|g\|^2.
\]

If in addition \( p \) is scaled so that

\[
\varphi(p) = \min_{\alpha} \varphi(\alpha p)
\]

then \( p^T (Bp + g) = 0 \). Finally, if \( \|B^{-1}\| \leq 1/\varepsilon \) and \( p \) satisfies (1.3), i.e., \( \|Bp + g\| \leq \nu \|g\| \) then

\[
\|p\| \leq \frac{1 + \nu}{\varepsilon} \|g\|.
\]

So (2.1) and (2.2) are satisfied if we decrease the quadratic at least as much as what the Cauchy step will, if we scale the direction so that the quadratic is minimized at \( p \) along the direction and we terminate with a relative residual that satisfies (1.3).
3. The Backtracking Strategy

In this section we discuss descent methods using a backtracking strategy. The results are not limited to the case when the descent method is an inexact quasi-Newton method or the inner iterations are Conjugate gradient iterations.

Our backtracking strategy is a restricted step method where we first take a trial step \( \lambda p_k \) and if the iterate \( x_{k+1} = x_{k+1} + \lambda p_k \) does not satisfy (1.5), then the steplength \( \lambda \) is reduced. At this stage we have the function values \( f(x_k) \) and \( f(x_{k+1}) \) and the directional derivative at \( x_k \) along \( p_k \), \( g(x_k)^T p_k \), and it is natural to use quadratic interpolation of \( h(\lambda) = f(x_k + \lambda p_k) \) at \( \lambda = 0 \) and \( \lambda = \lambda^+ \) with safeguards to reduce the steplength if the step \( x_{k+1} \) does not satisfy (1.5). To find the new steplength, Dennis and Schnabel [8] discuss a strategy where the new steplength \( \lambda^+ \) is chosen so that

\[
\lambda^+ \in [\lambda \mu, \lambda \rho]
\]

where \( 0 < \mu \equiv \rho < 1 \). For suitable values of \( \mu \) and \( \rho \) this allows the use of quadratic or cubic interpolation of \( h(\lambda) \). This process is now repeated until \( x_k + \lambda p_k \) satisfies (1.5) and \( \lambda_k \) is chosen to be the final value of \( \lambda \). The backtracking routine introduced by Armijo [1] and discussed in Ortega and Rheinboldt [13, 14.2.15] uses a fixed reduction and is thus not able to use interpolation to find the new steplength.

In most cases, the previous steplength \( \lambda_{k-1} \) is a reasonable guess for the new steplength. Further, most descent methods are based on some local model which predicts the new function value. So close to a solution, we expect that the step need not to be damped. In this case we will use \( \lambda = 1 \) as initial guess provided the old steplength is not too small. The backtracking strategy that we will use is given below.
Backtracking Routine:

Given positive constants $\tau, \omega, \alpha \leq \frac{1}{2}, \mu \leq \rho < 1$.

Given initial $\lambda$;

Let $\lambda^+ = \min\{1, \max\{\tau, \frac{1}{\omega}, \lambda\}\}$;

Set $\lambda = \lambda^+$;

WHILE $f(x + \lambda p) > f(x) + \lambda \alpha g(x)^T p$ DO

Let $\lambda^+ \in [\mu \lambda, \rho \lambda]$;

Set $\lambda = \lambda^+$.

We have eliminated the subscript referring to the outer iteration number$^4$.

For a descent direction $p$, we know that there exists $\lambda^+$ so that if $\lambda \leq \lambda^+$ then

$f(x + \lambda p) \leq f(x) + \lambda \alpha g(x)^T p$. Hence the backtracking strategy is well defined for any descent direction $p$ and the backtracking strategy terminates in a finite number of iterations.

The new outer iterate is now $x_+ = x + \lambda p$.

The following lemma is a modification of Ortega and Rheinboldt [13, 14.2.15]. In [13] it is assumed$^5$ that $\mu = \rho$. The proof can easily be modified to include $0 < \mu \leq \rho < 1$. A detailed proof is given in the Appendix.

Lemma 3.1: Let the function $f$ satisfy the standard hypothesis. Consider the iteration (1.4) where $\lambda_k$ is selected according to the Backtracking Routine and let $p_k$ be a descent direction. If there exists $\sigma > 0$ so that

$$\|p_k\| \geq \sigma \|g(x_k)\|$$  \hspace{1cm} (3.1)

for all $k$, then

$$\lim_{k \to \infty} \frac{g(x_k)^T p_k}{\|p_k\|} = 0.$$  \hspace{1cm} (3.2)

To show that the descent method is convergent, we need that $\{p_k\}$ is gradient related.

In Section 2 we show that the iterate $p_k$ from Conjugate gradient routine satisfies

$^4$ The constants $\alpha, \mu, \rho, \omega$, and $\tau$ are do not dependent on the outer iteration index.

$^5$ In [13] it is assumed that $g$ is continuous and that for all $x_0$ the level set $L(x_0) = \{x \in \mathbb{R}^n: f(x) \leq f(x_0)\}$ is bounded. These conditions guarantee that $g$ is uniformly continuous on each levelset.
Theorem 3.2: Let the function $f$ satisfy the standard hypothesis. Consider the iteration (1.4) where $\lambda_k$ is selected according to the Backtracking Routine. Let $p_k$ satisfy (3.3). If

$$\sigma_k \geq \sigma > 0 , \quad 0 < \gamma_k \leq \gamma$$

then

$$\lim_{k \to \infty} g(x_k) = 0 .$$

If instead we only have

$$\sigma_k \geq \frac{\sigma}{1 + \sum_{i=0}^{k} \|x_{i+1} - x_i\|} , \quad 0 < \sigma , \quad 0 < \gamma_k \leq \gamma$$

and $\{\|x_{k+1} - x_k\|\}$ is bounded then

$$\lim \inf_{k \to \infty} \|g(x_k)\| = 0 .$$

Proof: From (3.3) we have

$$- \|g(x_k)\| \|p_k\| \leq g(x_k)^T p_k \leq -\sigma_k \|g(x_k)\|^2 \leq -\frac{\sigma_k}{\gamma_k} \|p_k\| \|g(x_k)\|$$

and therefore $\|p_k\| \geq \sigma \|g(x_k)\|$ using (3.4). Hence the conditions in Lemma 3.1 are satisfied and we have that $g(x_k)^T p_k \|p_k\| \to 0$ as $k \to \infty$. From (3.6) and (3.4) we have that

$$\frac{g(x_k)^T p_k}{\|p_k\|} \leq -\frac{\sigma}{\gamma} \|g(x_k)\|$$

so $\|g(x_k)\| \to 0$ as $k \to \infty$ and this proves the first part of the theorem.

We will now prove the second part of the theorem. From (1.5) and (3.6) we have

$$f(x_{k+1}) - f(x_k) \leq \alpha \lambda_k g(x_k)^T p_k \leq -\alpha \frac{\sigma_k}{\gamma_k} \lambda_k \|p_k\| \|g(x_k)\| .$$

Since $f$ is bounded below, we have by summing over all $k$ that

$$\sum_{k=0}^{\infty} \sigma_k \|s_k\| \|g(x_k)\| < +\infty ,$$

where $s_k = x_{k+1} - x_k = \lambda_k p_k$ and using that $\gamma_k \leq \gamma$. We prove the result by obtaining a contradiction. Assume that for some $k_0$
Then the sum
\[ \sum_{k \geq 0} \|s_k\| \leq \sigma < +\infty. \]

Hence using (3.5)
\[ \sum_{k \geq 0} \frac{\|s_k\|}{1 + \sum_{i=0}^{k-1} \|s_i\|} < +\infty. \]

From Fletcher [10, Lemma 8], we have using that \( \|s_k\| \) is bounded that\(^6\)
\[ \sum_{k \geq 0} \|s_k\| < +\infty. \] (3.8)

Since the sum in (3.8) is bounded we can conclude from (3.5) that there exists \( \sigma \) so that \( \sigma_k \geq \sigma > 0 \). From the first part of the theorem, we have
\[ \lim_{k \to \infty} \|g(x_k)\| = 0. \]

However, by assumption, \( \|g(x_k)\| , k \geq k_0 \) is bounded away from zero and we have contradicted this assumption.

Q.E.D.

In the first part of Theorem 3.2 we assumed (3.4). From Section 2 we have that if \( \{B_k\} \) is bounded and the inner iterations are Conjugate Gradient Inner Iterations then (3.4) holds. In the second part we relax (3.4), but we require that \( \|s_k\| \) is bounded. Since \( \|f(x_k)\| \) is monotonically decreasing, we have that \( \|s_k\| \) is bounded if the levelset
\[ L(x_0) = \{z \in \mathbb{R}^n : f(z) \leq f(x_0)\} \] is bounded or if \( \{p_k\} \) is bounded. We discuss assumption (3.5) in Section 5. The next theorem shows that if \( \sigma_k \geq \sigma > 0 \) and \( 0 < \gamma_k \leq \gamma \) then the number of iterations in the backtracking routine is uniformly bounded.

**Theorem 3.3:** Let the function \( f \) satisfy the standard hypothesis. Consider the iteration (1.4) where \( \lambda_k \) is selected according to the Backtracking Routine. Let \( p_k \) satisfy (3.3). If \( \sigma_k \geq \sigma > 0 \) and \( 0 < \gamma_k \leq \gamma \), then there exists \( 0 < \lambda_{\min} \) so that the backtracking routine

---

\(^6\) Let \( \{\alpha_i\} \) be a sequence of bounded non negative numbers such that \( \sum_{k \geq 0} \frac{\alpha_k}{\sum_{i=0}^{k-1} \alpha_i} < +\infty \), then
\[ \sum_{k \geq 0} \alpha_k < +\infty. \]
terminates with $\lambda_k \geq \lambda_{\min}$ for all $k$.

**Proof:** Eliminate the outer subscript $k$. From the mean value theorem (c.f. [13, 3.2.2]) we have

$$f(x + \lambda p) - f(x) = \lambda g(x + \theta \lambda p)^T p$$

where $0 < \theta < 1$. Consider

$$f(x + \lambda p) - f(x) - \alpha \lambda g(x)^T p = \lambda g(x + \theta \lambda p)^T p - \alpha \lambda g(x)^T p$$

$$= \lambda (1 - \alpha) g(x)^T p + \lambda [g(x + \theta \lambda p) - g(x)]^T p$$

$$\leq -\lambda [(1 - \alpha) \sigma \|g(x)\|^2 - L \theta \lambda \|p\|^2]$$

where $L$ is the Lipschitz constant in the A.3 and using (3.3) and $\theta < 1$. Hence for

$$\lambda \leq \frac{(1 - \alpha) \sigma}{L \gamma^2}$$

$x + \lambda p$ satisfies (1.5). By choosing

$$\lambda_{\min} = \mu \frac{(1 - \alpha) \sigma}{L \gamma^2}$$

we have the desired result.

Q.E.D.

Since the initial $\lambda_k$ is less than or equal to 1 we have that the number of iterations in the Backtracking Routine is uniformly bounded by $|\log \lambda_{\min}|/\log \rho I$. In Section 5 we state an algorithm and give examples of updates $\{B_k\}$ so that we have global convergence.

### 4. Rates of Convergence

In this section, we first study the rate of convergence of the descent methods that have been discussed in the previous sections. We show that under reasonable smoothness assumptions we have a $Q$-linear rate of convergence. This results strengthen a result of Ortega and Rheinboldt [13, 14.1.6] where an $R$-linear rate of convergence is shown. We give conditions which imply that $\lambda_k = 1$ for $k$ sufficiently large and that the iterates $\{x_k\}$ are $Q$-superlinearly convergent. In this section we will replace the Standard Hypothesis of $f$ with
local properties of \( f \) in a neighborhood of a solution of (1.1). We will assume that: (Local Hypothesis)

\( f \) is twice continuously differentiable in an open neighborhood \( \Omega \) of \( x^* \) where \( g(x^*) = 0; \)

The Hessian matrix at \( x^* \), \( H(x^*) \), is positive definite.

We first show that under these assumptions and the assumptions of Theorem 3.3 we have a Q-linear rate of convergence. From Theorem 3.3 we have that \( \lambda_k \geq \lambda_{\min} > 0 \) hence

\[
\begin{equation}
(4.1) \quad f(x_{k+1}) \leq f(x_k) + \alpha \lambda_k g(x_k)^T p_k \leq f(x_k) - \alpha \lambda_{\min} \sigma \| g(x_k) \|^2
\end{equation}
\]

Consider the weighted norm \( \| \cdot \| \), where \( \| y \|^2 = y^T H(x^*) y \). Ortega and Rheinboldt [13, 14.1.6] have shown that if (4.1) holds then \( \{x_k\} \) converges to \( x^* \) R-linearly. We will show that we have a Q-linear rate of convergence in the weighted norm \( \| \cdot \|. \) We first need a technical lemma.

**Lemma 4.1** Let \( f \) satisfy the Local Hypothesis. Let \( \gamma = \frac{1}{2} \| H(x^*)^{-1} \|^{-1/2} \) and \( 0 < \delta < 1 \). Then there exists \( \epsilon > 0 \) so that

\[
\begin{equation}
(4.2) \quad \frac{1}{2} (1-\delta) \| x - x^* \|^2 \leq f(x) - f(x^*) \leq \frac{1}{2} (1+\delta) \| x - x^* \|^2,
\end{equation}
\]

and

\[
\| g(x) \| \geq \gamma \| x - x^* \|.
\]

for all \( \| x - x^* \| \leq \epsilon \).

**Proof** Since \( H \) is continuous in \( \Omega \) and \( H(x^*) \) is nonsingular we can find \( \epsilon > 0 \) so that for all \( \| x - x^* \| \leq \epsilon \) we have

\[
\| H(x) - H(x^*) \| \leq \delta \| H(x^*)^{-1} \|
\]

and

\[
\| g(x) - g(x^*) - H(x^*)(x - x^*) \| \leq \gamma \| x - x^* \|.
\]

Then for \( \| x - x^* \| \leq \epsilon \) we have for some \( 0 < \theta < 1 \)

\[
\begin{equation}
(4.3) \quad f(z) - f(x^*) = \frac{1}{2} (z - x^*)^T H(x^*)(z - x^*) + \frac{1}{2} (z - x^*)^T [H(x^* + \theta (z - x^*)) - H(x^*)](z - x^*)
\end{equation}
\]

Hence
\[ f(x) - f(x^*) \leq \frac{1}{2} \|x - x^*\|^2 + \frac{1}{2} \|x - x^*\|^2 \delta/\|H(x^*)^{-1}\| \]
\[ \leq \frac{1}{2} \|x - x^*\|^2 + \frac{1}{2} \|x - x^*\|^2 \|H(x^*)^{-1}\| \delta/\|H(x^*)^{-1}\| = \frac{1}{2} (1+\delta) \|x - x^*\|^2. \]

Similarly,
\[ f(x) - f(x^*) \geq \frac{1}{2} (1-\delta) \|x - x^*\|^2 \]
and we have shown (4.2).

Consider
\[ g(x) = H(x^*)(x - x^*) + [g(x) - g(x^*) - H(x^*)(x - x^*)] \]
using that \( g(x^*) = 0 \), hence
\[ \|g(x)\| \geq \frac{1}{\|H(x^*)^{-1}\|^{1/2}} \|x - x^*\|, \quad -\gamma \|x - x\|, \quad \geq \gamma \|x - x^*\|. \]
and we have shown (4.3).

**Theorem 4.2** Let \( f \) satisfy the Local Hypothesis and let \( \{x_k\} \) be a sequence such that
\[ x_k \to x^* \text{ as } k \to \infty. \]
Assume there exists \( \sigma > 0 \) so that
\[ f(x_{k+1}) - f(x_k) \leq -\sigma \|g(x_k)\|^2 \quad (4.4) \]
Then there exist \( 1 > t > 0 \) and \( k_0 \) so that
\[ \|x_{k+1} - x^*\| \leq t \|x_k - x^*\|, \text{ for } k \geq k_0 \quad (4.5) \]
where \( \|y\|^2 = y^T H(x^*) y \).

**Proof:** Let \( \gamma \) be as in Lemma 4.1 and if necessary reduce \( \sigma > 0 \) so that \( 1 - 2\sigma \gamma^2 > 0 \). Choose \( t \) and \( 1 > \delta > 0 \) so that
\[ 0 < \frac{1 - 2\sigma \gamma^2 + \delta}{1 - \delta} \leq t^2 < 1 \quad (4.6) \]
and choose \( \varepsilon > 0 \) so that (4.2) and (4.3) hold. Let \( \|x + s - x^*\| \leq \varepsilon \) and \( \|x - x^*\| \leq \varepsilon \) and assume
\[ f(x + s) \leq f(x) - \sigma \|g(x)\|^2. \quad (4.7) \]
Then
\[ \|x + s - x^*\|^2 \leq 2(f(x + s) - f(x^*)) + \delta \|x + s - x^*\|^2 \]
\[ \leq 2(f(x) - f(x^*)) - 2\sigma \|g(x)\|^2 + \delta \|x + s - x^*\|^2. \]
\[
\leq \|x-x^*\|^2 + \delta \|x-x^*\|^2 - 2\sigma \|g(x)\|^2 + \delta \|x+s-x\|^2
\]
\[
\leq (1-2\sigma\gamma^2 + \delta)\|x-x^*\|^2 + \delta \|x+s-x^*\|^2
\]
using (4.2) and (4.7) twice and (4.3). Hence
\[
\|x+s-x^*\|^2 \leq \frac{1-2\sigma\gamma^2 + \delta}{1-\delta} \|x-x^*\|^2. 
\] (4.8)
Choose \(k_0\) so that \(\|x_k - x^*\| \leq \varepsilon\) for \(k \geq k_0\). Let \(s_k = x_{k+1} - x_k\) then from the choice of \(t\) in (4.6) and (4.8)
\[
\|x_{k+1} - x^*\| \leq t \|x_k - x^*\|. 
\]
Q.E.D.

The next lemma is a minor reformulation of Theorem A.4 of Dembo and Steihaug [6] and is needed to prove our main result.

**Lemma 4.3:** Let \(0<\alpha<\frac{1}{2}\), \(0<\beta<1\) and \(1>\beta_1>\frac{1}{2}\). Let \(f\) satisfy the Local Hypothesis. There exist positive numbers \(\varepsilon\) and \(\delta\) so that for all \(x\) and \(p\) so that
\[
\|x-x^*\| \leq \varepsilon, \quad \text{and} \quad \frac{(H(x)p + g(x))^T p}{p^T p} \leq \delta
\]
then
\[
f(x+p) \leq f(x) + \alpha p^T g(x),\]
\[
g(x+p)^T p \geq \beta g(x)^T p, \quad |g(x+p)^T p| \leq -\beta g(x)^T p, \quad \text{and} \quad f(x+p) \geq f(x) + \beta_1 p^T g(x)
\]
\[\square\]

If \(B_k\) is the Hessian matrix at \(x_k\) and \(p_k\) satisfies the condition \(p_k^T (B_k p_k + g(x_k)) = 0\) then for \(x_k\) sufficiently close to \(x^*\), the point \(x_k+p_k\) satisfies all the linesearch conditions discussed in the Introduction and in particular (1.5). We note that \(p_k^T (B_k p_k + g(x_k)) = 0\) if \(p_k\) is found using Conjugate Gradient Inner Iterations. We now study the globalized inexact quasi-Newton method and we will show that there exists a \(k_0\) such that for all \(k \geq k_0\) then \(\lambda_k = 1\) under reasonable assumptions on \(\{p_k\}\) and \(\{B_k\}\).

**Theorem 4.4:** Let \(f\) satisfy the Local Hypothesis and consider the iteration (1.4) where \(\lambda_k\) is selected according to the Backtracking Routine where \(\tau<1\) and \(\omega<1\). Let \(x_k \to x^*\) as \(k \to \infty\) and let \(\{B_k\}\) be a sequence of real symmetric matrices such that
\[
\lim_{k \to \infty} \frac{\|g(x_k + s_k) - g(x_k) - B_k s_k\|}{\|s_k\|} = 0 
\] (4.9)
where $s_k = x_{k+1} - x_k$. Let the sequence of descent directions $\{p_k\}$ satisfy $\lim_{k \to \infty} p_k = 0$ and
\[
\lim_{k \to \infty} \frac{\tau_k p_k}{p_k d_k} = 0
\] (4.10)
where $\tau_k = B_k p_k + g(x_k)$. Then there exists $k_0$ so that for $k \geq k_0$ we have $\lambda_k = 1$.

Proof: We will first show that
\[
\lim_{k \to \infty} \frac{\|g(x_k + p_k) - g(x_k) - B_k p_k\|}{\|p_k\|} = 0.
\] (4.11)

Let $x$ and $x + p$ be in $\Omega$ and $0 < \lambda \leq 1$. Consider
\[
g(x + \lambda p) - g(x) - \lambda[g(x + p) - g(x)] =
\]
\[
g(x + \lambda p) - g(x) - H(x)\lambda p - \lambda[g(x + p) - g(x) - H(x)p]
\]
Using the triangle inequality and
\[
\|g(y) - g(x) - H(x)(y - x)\| \leq \sup_{\|\theta\|_1} \|H(x + \theta(y - x)) - H(x)\| \|y - x\|
\]
for $x, y \in \Omega$ ([13, 3.2.12]), we have that
\[
\|g(x + \lambda p) - g(x) - \lambda[g(x + p) - g(x)]\| \leq 2\|p\| \sup_{l \neq 0} \|H(x + \theta p) - H(x)\|.
\]
Hence from the Local Hypothesis and the convergence of $\{s_k\}$ and $\{p_k\}$ to $0$ we have\footnote{Let $\{\xi_k\}$ be any real sequence which converges to $\xi^*$. Given continuous nonnegative functions $g$ and $h$, we write $g(\xi_k) = o(h(\xi_k))$ if $\lim_{k \to \infty} g(\xi_k)/h(\xi_k) = 0$.}
\[
\|g(x_k + \lambda p_k) - g(x_k) - \lambda[e_g(x_k + p_k) - g(x_k)]\| = o(\|s_k\|).
\] (4.13)
Consider
\[
\frac{\|g(x_k + p_k) - g(x_k) - B_k p_k\|}{\|p_k\|} \leq
\]
\[
\frac{\|g(x_k + \lambda p_k) - g(x_k) - \lambda[e_g(x_k + p_k) - g(x_k)]\|}{\|s_k\|} + \frac{\|g(x_k + s_k) - g(x_k) - B_k s_k\|}{\|s_k\|}
\]
and (4.11) follows directly from (4.9) and (4.13).

We now show that $p_k^T (g(x_k) + H(x_k) p_k) = o(\|p_k\|^2)$ holds. From the Local Hypothesis and the convergence of $p_k$ we have
\[
H(x_k)p_k = g(x_k + p_k) - g(x_k) + o(\|p_k\|)
\]
Hence
\[ p_k^T(g(x_k)+H(x_k)p_k) = p_k^Tg(x_k+p_k) + o(\|p_k\|^2) \]
\[ = p_k^Tg(x_k+p_k) - p_k^Tg(x_k) + o(\|p_k\|^2) \]
\[ = p_k^Tg(x_k+p_k) - g(x_k) - B_k p_k + o(\|p_k\|^2) \]
\[ = o(\|p_k\|^2) \quad (4.14) \]

using (4.10) and (4.11).

Let \( \epsilon \) and \( \delta \) be as in Lemma 4.3. Let \( k_1 \) be so that for all \( k \geq k_1 \) then

\[ \|x_k-x^*\| \leq \epsilon, \quad p_k^T(g(x_k)+H(x_k)p_k) \leq \delta p_k^T p_k \]

This can be done in view of the convergence of the sequences \( \{x_k\}, \{p_k\} \) and (4.14). Hence \( x_k + p_k \) will satisfy (1.5). Since \( \lambda_k \leq 1 \), the initial \( \lambda_k \) in the Backtracking Routine will satisfy (1.5) and

\[ \lambda_{k+1} = \min \{1, \max \{\tau, \frac{1}{\omega}\lambda_k\}\} \geq \tau. \]

If for some \( k \geq k_1 \) we have \( \frac{\lambda_k}{\omega} < 1 \) then \( \lambda_{k+1} = \frac{1}{\omega} \lambda_k \). So for some \( k_0 > k_1 \) we will have

\[ \frac{\lambda_{k-1}}{\omega} > 1 \] and \( \lambda_{k_0} = 1 \). It follows from Lemma 4.3 that for \( k \geq k_0 \), we have \( \lambda_k = 1 \).

Q.E.D.

If \( B_k \) is the Hessian matrix at \( x_k \) then (4.9) holds and if the number of conjugate gradient iterations is \( \geq 1 \), then \( p_k^T(B_k p_k + g(x_k)) = 0 \) and (4.10) holds. Further, since \( B_k \) (\( =H(x_k) \)) is nonsingular \( p_k \to 0 \) as \( k \to \infty \) and we have shown that if \( \{x_k\} \) converges to a local minimizer where \( f \) is strictly convex and sufficiently smooth, then \( \lambda_k = 1 \). Hence, from now on, the iterates are those of the inexact Newton method.

Given \( x_{k_0} \) and \( B_{k_0} \)

FOR \( k = k_0 \) STEP 1 UNTIL Convergence DO

Find some \( p_k \) that satisfies

\[ \|B_k p_k + g(x_k)\| \leq \nu_k |g(x_k)| \quad (4.15) \]

Let \( x_{k+1} = x_k + p_k \)

Update to obtain \( B_{k+1} \).

The following result characterizes Q-superlinear convergence in terms of the relative residuals.
Lemma 4.5: Let $f$ satisfy the Local Hypothesis. Let $\{x_k\}$ be determined by an inexact quasi-Newton method for $k \geq k_0$. Let $x_k \to x^*$ as $k \to \infty$ and

$$
\lim_{k \to \infty} \frac{\|g(x_k + p_k) - g(x_k) - B_k p_k\|}{\|p_k\|} = 0
$$

Let $r_k = B_k p_k + g(x_k)$. Then

$$
\lim_{k \to \infty} \frac{\|r_k\|}{\|p_k\|} = 0 \iff \lim_{k \to \infty} \frac{\|r_k\|}{\|g(x_k)\|} = 0 \iff \lim_{k \to \infty} \frac{\|x_k+1 - x^*\|}{\|x_k - x^*\|} = 0.
$$

We note that if $\frac{\|r_k\|}{\|p_k\|} \to 0$ as $k \to \infty$ then (4.10) holds. By choosing the sequence $\{\nu_k\}$ so that $\nu_k \to 0$ as $k \to \infty$ in (4.15) then we have a $Q$-superlinear rate of convergence of the iterates $\{x_k\}$.

5. Applications

In this section, we discuss applications of the results in the previous sections and approximations to the Hessian. In particular, we discuss least change updates in the Frobenius norm and finite difference approximations of the Hessian matrix.

Let $B$ be an affine set of $\mathbb{R}^{n \times n}$. For example $B$ may consist of all matrices $B$ that are symmetric and have a specified sparsity structure. Let $Q(y,s)$ be the set of all matrices $B$ that for given $s$ and $y$ in $\mathbb{R}^n$ satisfy the secant condition

$$
Bs = y.
$$

Let

$$
y = g(x + s) - g(x).
$$

Recall that if $f$ is twice continuously differentiable on $\mathbb{R}^n$ then $M_s = y$ where

$$
M = \int_0^1 H(x + \tau s) d\tau.
$$

Further if $H(x) \in B$ for all $x$ then $A = B \cap Q(y,s)$ is nonempty. A least change update $\tilde{B}$ in Frobenius norm satisfies

$$
\|\tilde{B} - B\|_F = \min \{\|\tilde{B} - B\|_F : \tilde{B} \in A\}.
$$

Since the Frobenius norm is the $l_2$ norm on $\mathbb{R}^{n \times n}$ and if $A$ is nonempty then we have for all
\( M \in A \) that
\[
\|B - M\|_F^2 = \|B - M\|_F^2 - \|B - B\|_F^2
\]  
(5.5)

If the Hessian is Lipschitzcontinuous\footnote{H is Lipschitz continuous if there exist \( L_H \geq 0 \) so that for all \( x,y \in \mathbb{R}^n \) then \( \|H(x) - H(y)\|_F \leq L_H \|x - y\| \)} then we have using (5.5) and \( M \) given by (5.3) that
\[
\|B - H(x + s)\|_F \leq \|B - H(x)\|_F + L_H \|s\|.
\]  
(5.6)

Steinhaug [16, 17] has shown that approximate least change secant updates satisfy (5.6).

In this section we will show that given \( x_j, j = 0, 1, \ldots, k \) the approximate Hessian matrix \( B_k \) satisfies (5.6) then
\[
\|B_k\| \leq \alpha_1 + \alpha_2 \sum_{i=0}^{k-1} \|x_{i+1} - x_i\|
\]  
(5.7)

where \( \alpha_1 \) and \( \alpha_2 \) are non negative numbers that only depend on \( x_0 \) and \( B_0 \). From Lemma 2.1 we have that
\[
g(x_k)^T p_k \leq -\sigma_k \|g(x_k)\|^2.
\]
\[
\sigma_k = \min \{1, \frac{1}{\|B_k\|} \} \geq \frac{\sigma}{1 + \sum_{i=0}^{k-1} \|x_{i+1} - x_i\|}
\]
and
\[
\sigma = \frac{1}{\max \{1 + \alpha_1, \alpha_2 \}}
\]

Hence we have (3.5) provided (5.7) holds.

Let \( \{B_k\} \) be a sequence of updates that satisfies (5.6). Then
\[
\|B_{k+1} - H(x_{k+1})\|_F \leq \|B_k - H(x_k)\|_F + L_H \|s_k\| \leq \|B_0 - H(x_0)\|_F + L_H \sum_{i=0}^{k} \|s_i\|
\]
and
\[
\|B_{k+1} - H(x_0)\|_F \leq \|B_{k+1} - H(x_{k+1})\|_F + \|H(x_{k+1}) - H(x_0)\|_F
\]
\[
\leq \|B_0 - H(x_0)\|_F + 2L_H \sum_{i=0}^{k} \|s_i\|
\]
Hence for given \( x_0 \) and \( B_0 \) the matrix \( B_k \) satisfies (5.7) with
\[
\alpha_1 = \|B_0 - H(x_0)\|_F + \|H(x_0)\|_F \text{ and } \alpha_2 = 2L_H.
\]
Finally consider a finite difference approximation of $H(x_k)$, where we assume that
\[ \|B_k - H(x_k)\| \leq \rho_k. \] (5.8)

Then
\[ \|B_k\| \leq \|B_k - H(x_k)\| + \|H(x_k) - H(x_0)\| + \|H(x_0)\| \leq \rho_k + \|H(x_0)\| + L_h \sum_{i=1}^{k} \|s_i\| \]
so for $\rho_k \leq \rho$ we have (5.7) with
\[ \alpha_1 = \rho + \|H(x_0)\| \quad \text{and} \quad \alpha_2 = L_h. \]

Let $C_i = 1, 2, \ldots, m$ be a partitioning of $\{1, 2, \ldots, n\}$ into $m$ disjoint sets. By choosing the sets $C_i$ appropriately the columns $j \in C_i$ of the Hessian matrix may be approximated by the difference
\[ \frac{1}{\gamma}[g(x_k + \gamma e^{(i)}) - g(x_k)] \]
where $e^{(i)}$ is a sum of the unit vectors $e^j, j \in C_i$. Let $B_k$ be this approximate Hessian matrix at $x_k$. Then
\[ \gamma^2 \|B_k - H(x_k)\|^2 = \sum_{i=1}^{m} \|g(x_k + \gamma e^{(i)}) - g(x_k) - \gamma H(x_k) e^{(i)}\|^2 \]
\[ \leq \sum_{i=1}^{m} (\gamma^2 \|e^{(i)}\|^2 L_h)^2 \leq (\gamma^2 L_h m)^2 \]
and we have (5.8) for $\rho_k \geq L_h \gamma$. If the Hessian matrix is bounded in $\mathbb{R}^n$, then we have that the sequence of finite difference approximations $\{\|B_k\|\}$ is bounded.

Consider the following algorithmic framework:

Given $x_0$ and $B_0$:

FOR $k = 0$ STEP 1 UNTIL Convergence DO

Compute direction $p_k$ using Conjugate Gradient Inner iterations;

Compute step correction $\lambda_k$ using the Backtracking Routine;

Let $x_{k+1} = x_k + \lambda_k p_k$.

Update to obtain $B_{k+1}$.

**Corollary 5.1:** Let $f$ be three times continuously differentiable on $\mathbb{R}^n$ and let for all $x_0 \in \mathbb{R}^n$ the levelset
be bounded. Then the sequence of iterates \( \{x_k\} \) selected according to the above method is well defined. If \( B_k \) is an approximation of the Hessian matrix at \( x_k \) that satisfies (5.8) with \( \rho_k \leq \rho \) then the method is globally convergent, i.e.,

\[
\lim_{k \to \infty} g(x_k) = 0
\]

and every limitpoint of the sequence \( \{x_k\} \) is a stationary point of \( f \). Further, if \( x^* \) is a limit-point of \( \{x_k\} \) and \( H(x^*) \) is positive definite, then \( x_k \to x^* \) as \( k \to \infty \) at least Q-linearly. If \( \rho_k \to 0 \) as \( k \to \infty \) and the Conjugate Gradient Inner Iterations is terminated with (1.3) where \( \nu_k \to 0 \) as \( k \to \infty \) then \( \{x_k\} \) converges Q-superlinearly.

If instead \( B_k \) satisfies (5.7) then

\[
\liminf_{k \to \infty} g(x_k) = 0
\]

and the sequence \( \{x_k\} \) has a limitpoint that is a stationary point of \( f \). Further, if \( x^* \) is a limitpoint of \( \{x_k\} \) and a local minimizer of \( f \) at which \( H(x^*) \) is positive definite, then \( x_k \to x^* \) as \( k \to \infty \). Let \( B_k \) be the least change secant update (5.4) using (5.2) then (4.9) holds and if and the Conjugate Gradient Inner Iterations is terminated with (1.3) where \( \nu_k \to 0 \) as \( k \to \infty \) then \( \{x_k\} \) converges Q-superlinearly.

The next corollary shows convergence of Cauchy's method. Cauchy [2] suggests using (see [3], or [13, NR 8.3-4])

\[
p_k = -\frac{g(x_k)^T g(x_k)}{g(x_k)^T H(x_k) g(x_k)} g(x_k),
\]

i.e. a Newton step along the steepest descent direction at \( x_k \). The new iterate is

\[
x_{k+1} = x_k + p_k
\]

**Corollary 5.2:** Assume \( f \) is twice continuously differentiable in an open neighborhood of a local minimizer \( x^* \) at which \( H(x^*) \) is positive definite. Consider the iteration (5.9) where \( p_k \) is given in (5.8). Then Cauchy's method is well defined and locally convergent, i.e. there exists \( \varepsilon > 0 \) so that for \( \|x_0 - x^*\| \leq \varepsilon \) the iterates \( \{x_k\} \) converge to \( x^* \). Furthermore, there exist \( 0 < t < 1 \) and \( k_0 \) so that
\[ \|x_{k+1} - x^*\| \leq \ell \|x_k - x^*\|, \text{ for } k \geq k_0. \]

where \( \|y\|^2 = y^T H(x^*) y \).

Acknowledgements.

Parts of this work were performed as a part of the doctoral thesis [15] in the Department of Administrative Sciences at Yale University. I am grateful to Ron Dembo of Department of Administrative Sciences and Stan Eisenstat of Computer Science Department for their guidance. Discussions with John Dennis and Dick Tapia have greatly benefited the results presented. Andy Conn and the referees gave some helpful comments on the original version of this paper.
References


In the appendix we restate the framework and prove Lemma 3.1. The proof is closely related to the proof of Theorem 6.3.3 in Dennis and Schnabel [8].

Given \( x_0 \):

For \( k = 0, 1, 2, \ldots \) do

Compute descent direction \( p_k \):

Compute \( \lambda_k \):

Let \( \lambda_k \geq \tau \):

\[
\text{WHILE } f(x_k + \lambda_k p_k) > f(x_k) + \lambda_k \alpha g(x_k)^T p_k \text{ DO}
\]

Let \( \lambda_k^* \in [\mu \lambda_k, \rho \lambda_k] \):

Set \( \lambda_k := \lambda_k^* \):

Update: \( x_{k+1} = x_k + \lambda_k p_k \).

**Lemma A.1:** Let \( f \) satisfy the standard hypothesis and let \( 0 < \tau, 0 < \mu \leq \rho < 1 \). Let \( \{p_k\} \) be descent directions. If there exists \( \sigma > 0 \) such that for all \( k \geq k_0 \)

\[
\|p_k\| \geq \sigma \|g(x_k)\| \tag{a.1}
\]

then

\[
\lim_{k \to \infty} \frac{g(x_k)^T p_k}{\|p_k\|} = 0
\]

**Proof:** Since \( p_k \) is a descent direction, the backtracking routine is well defined and terminates with a \( \lambda_k \) that satisfies [8]

\[
f(x_k + \lambda_k p_k) \leq f(x_k) + \lambda_k \alpha g(x_k)^T p_k \tag{a.2}
\]

In the proof we will distinguish between the two cases that the initial \( \lambda_k \) satisfies (a.2) and that one or more iterations are done to determine \( \lambda_k \).

Consider the case that the initial \( \lambda_k \) satisfies (a.2). Then

\[
0 > g(x_k)^T p_k = \|p_k\|^2 \frac{g(x_k)^T p_k}{\|p_k\|^2} \geq \|p_k\|^2 \frac{1}{\sigma} \frac{g(x_k)^T p_k}{\|p_k\| \|g(x_k)\|} \geq - \frac{\lambda_k}{\sigma \tau} \|p_k\|^2 \tag{a.3}
\]

using (a.1), \( \lambda_k \geq \tau \), and the Cauchy-Schwarz's inequality.
Let $\lambda_k^+$ be the current $\lambda_k$ such that

$$f(x_k + \lambda_k^+ p_k) > f(x_k) + \alpha \lambda_k^+ g(x_k)^T p_k \tag{a.4}$$

and let $\lambda_k^*$ be the new iterate so that

$$f(x_k + \lambda_k^* p_k) \leq f(x_k) + \alpha \lambda_k^* g(x_k)^T p_k \tag{a.5}$$

Then from the standard mean value theorem

$$f(x_k + \lambda_k^* p_k) - f(x_k + \lambda_k^+ p_k) = \tilde{g}_k^*(\lambda_k^* - \lambda_k^+)(p_k + \lambda_k^+ p_k) - f(x_k)$$

where

$$\tilde{g}_k = g(x_k + [(1 - \theta_k) \lambda_k^* + \theta_k \lambda_k^+]p_k), \text{ for } 0 < \theta_k < 1.$$ 

From (a.5) we have

$$\alpha \lambda_k^* g(x_k)^T p_k \geq f(x_k + \lambda_k^* p_k) - f(x_k)$$

$$= [f(x_k + \lambda_k^* p_k) - f(x_k + \lambda_k^+ p_k)] + [f(x_k + \lambda_k^+ p_k) - f(x_k)]$$

$$= (\lambda_k^* - \lambda_k^+) \tilde{g}_k^* p_k + [f(x_k + \lambda_k^* p_k) - f(x_k)]$$

$$> (\lambda_k^* - \lambda_k^+ \tilde{g}_k^* p_k + \alpha \lambda_k^* g(x_k)^T p_k$$

using (a.6) and (a.4). Hence

$$(\lambda_k^* - \lambda_k^+ \tilde{g}_k^* p_k < \alpha (\lambda_k^* - \lambda_k^+ g(x_k)^T p_k)$$

and we have

$$\tilde{g}_k^* p_k > \alpha g(x_k)^T p_k \tag{a.7}$$

using $\lambda_k^* < \lambda_k^+$. Consider now (a.7) and subtract $g(x_k)^T p_k$ on both sides. Then

$$(\alpha - 1)g(x_k)^T p_k < \tilde{g}_k^* p_k - g(x_k)^T p_k$$

$$\leq \|\tilde{g}_k - g(x_k)\| \|p_k\|$$

$$\leq L[(1 - \theta_k)\lambda_k^* + \theta_k \lambda_k^+]\|p_k\|^2$$

$$\leq L \lambda_k^+ \|p_k\|^2 \leq \frac{L}{\mu} \lambda_k^+ \|p_k\|^2 = \frac{L}{\mu} \lambda_k \|p_k\|^2,$$

using the Lipschitz continuity and $\lambda_k^* \geq \mu \lambda_k$. Combined with (a.3) we have for

$$\beta = \max\left\{\frac{L}{\mu(1-\alpha)}, \frac{1}{\sigma \tau}\right\}$$

that

$$g(x_k)^T p_k \leq -\beta \lambda_k \|p_k\|^2 \tag{a.8}.$$
Consider

\[ f(x_{i+1}) - f(x_i) \leq \alpha \lambda_i g(x_i)^T p_i \]

for \(0 \leq i \leq k\). By summing over all \(i \leq k\)

\[ f(x_{k+1}) - f(x_0) \leq \sum_{i=0}^{k} \alpha \lambda_i g(x_i)^T p_i < 0. \]

Since \(f\) is bounded below, we have

\[ \sum_{i=0}^{\infty} \lambda_i |g(x_i)^T p_i| < +\infty, \]

and

\[ \lambda_i g(x_k)^T p_k \to 0 \text{ as } k \to \infty \]  \hspace{1cm} \text{(a.9)}

But from (a.8)

\[ \lambda_k g(x_k)^T p_k = -\lambda_k ||p_k|| \left( - \frac{g(x_k)^T p_k}{||p_k||} \right) \leq - \frac{1}{\beta} \left( \frac{g(x_k)^T p_k}{||p_k||} \right)^2 < 0. \]

hence from (a.9) we have

\[ \lim_{k \to \infty} \frac{g(x_k)^T p_k}{||p_k||} = 0. \]  \hspace{1cm} \text{Q.E.D.}