SUPERLINEAR CONVERGENCE AND IMPLICIT FILTERING [*]

T. D. CHOI[†] AND C. T. KELLEY[†]

Abstract. In this note we show how the implicit filtering algorithm can be coupled with the BFGS quasi-Newton update to obtain a superlinearly convergent iteration if the noise in the objective function decays sufficiently rapidly as the optimal point is approached. We show how known theory for the noise-free case can be extended and thereby provide a partial explanation for the good performance of quasi-Newton methods when coupled with implicit filtering.

Key words. noisy optimization, implicit filtering, BFGS algorithm, superlinear convergence

AMS subject classifications. 65K05, 65K10, 90C30

1. Introduction.

In this paper we examine the local and global convergence behavior of the combination of the BFGS [4], [20], [17], [23] quasi-Newton method with the implicit filtering algorithm. The resulting method is intended to minimize smooth functions that are perturbed with low-amplitude noise. Our results, which extend those of [5], [15], and [6], show that if the amplitude of the noise decays sufficiently rapidly near optimality, then the local q-superlinear convergence results of [5] and, under more restrictive assumptions, the global convergence results of [6] hold. The results in this paper are theoretical and require strong assumptions. However, we believe that they represent a important step toward explaining the observations of improvements in performance when quasi-Newton model Hessians are used with implicit filtering.

The quasi-Newton implicit filtering algorithms differ from other methods in the literature that use either inaccurate gradient information, only samples of the function, or difference or interpolatory approximations to gradients and/or Hessians. While we make assumptions on the decay of the noise near optimality, we do not assume that we can control the errors in the function evaluation directly, and therefore our results differ from those of [7] and [8], where it was assumed that control of the errors in function and gradient evaluations was possible and global convergence of a trust region algorithm that managed these errors separately was proved. The superlinearly convergent algorithm in [22], which combines coordinate search with a difference Hessian, is intended for noise-free function evaluations and is not applicable here. Our quasi-Newton algorithms do not attempt to model Hessians with interpolation, as does the trust region/interpolation method of [12], [10], and [11]. We believe that the quasi-Newton approach has an advantage for noisy problems, where the errors in a Hessian formed by differences or interpolation can be large.

In § 2, we review implicit filtering and, using an idealized implementation for a model problem, motivate the assumptions on function/gradient accuracy that we use in § 3. In § 3.1 we apply the results in [15] to show how the local theory for BFGS convergence from [5] can be extended to prove superlinear convergence of the idealized method. In § 3.2 we show how a combination of the assumptions from [19] and the ones in § 3.1 imply extensions of the global superlinear convergence results from [6].

[*] Version of March 18, 1999.

[†] North Carolina State University, Center for Research in Scientific Computation and Department of Mathematics, Box 8205, Raleigh, N. C. 27695-8205, USA (tdchoi@unity.ncsu.edu, Tim_Kelley@ncsu.edu,). This research was supported in part by National Science Foundation grant #DMS-9700569 and a Department of Education GAANN fellowship.

2. Implicit Filtering. Implicit filtering was designed for problems in which the objective function is a high-frequency, low-amplitude, perturbation of a simple smooth problem. The algorithm is a finite difference steepest descent (or quasi-Newton) method in which the difference increment is adjusted as the optimization progresses. In this way the algorithm implicitly filters out the high frequency perturbation. For problems in which the amplitude of the perturbation decreases near optimality, a not uncommon event [24], [13], [9], decreasing the difference increment improves the accuracy of the solution near the optimal point.

Quantitatively we consider an objective function f defined on R^N that is a perturbation of a smooth function f_s by a small function ϕ

$$f(x) = f_s(x) + \phi(x). \tag{2.1}$$

The small oscillations could cause f to have several local minima that would trap any conventional gradient-based algorithms. The perturbation ϕ could be random, [24], and therefore need not even be a function. In this paper, we assume that ϕ is everywhere defined and bounded to make the statement of the results simpler.

Throughout this paper $\|\cdot\|$ will denote the ℓ^2 norm on R^N.

2.1. The Basic Algorithm. For $x \in R^N$ and $h \neq 0$ the forward difference gradient of f with *scale* h at x is given by

$$(\nabla_h^f f(x))_i = \frac{f(x+he_i) - f(x)}{h}$$

where e_i is the unit vector in the ith coordinate direction and $(\nabla_h^f f(x))_i$ denotes the ith component of the difference gradient. Similarly, the centered difference gradient with scale h is given by

$$\nabla_h^c f(x) = \frac{\nabla_h^f f(x) + \nabla_{-h}^f f(x)}{2}.$$

We will often refer to the difference gradient $\nabla_h f$ when we are discussing a point that applies to either.

The basic finite difference steepest descent iteration takes a current iteration x_c to the next x_+ by

$$x_+ = x_c - \lambda \nabla_h f(x_c).$$

If the line search is successful the step length λ satisfies the sufficient decrease condition We use a simple Armijo [1] line search and demand that the sufficient decrease condition

$$f(x - \lambda \nabla_h f(x)) - f(x) < -\alpha\lambda \|\nabla_h f(x)\|^2. \tag{2.2}$$

Implicit filtering is based on a finite difference steepest descent algorithm `fdsteep`, which terminates when

$$\|\nabla_h f(x)\| \leq \tau h \tag{2.3}$$

for some $\tau > 0$, when more than $kmax$ iterations have been taken, or when the line search fails by taking more than $amax$ backtracks. Even the failures of `fdsteep` can be used to advantage by triggering a reduction in h. The line search parameters α, β and the parameter τ in the termination criterion (2.3) do not affect the convergence analysis that we present here, but can affect performance.

```
ALGORITHM 2.1. fdsteep($x, f, kmax, \tau, h, amax$)
1. For $k = 1, \dots, kmax$
   (a) Compute $f$ and $\nabla_h f$; terminate if (2.3) holds.
   (b) Find the least integer $0 \le m \le amax$ such that (2.2) holds for $\lambda = \beta^m$. If no such $m$
       exists, terminate.
   (c) $x = x - \lambda \nabla_h f(x)$.
```

Algorithm fdsteep will terminate after finitely many iterations because of the limits on the number of iterations and the number of backtracks. If the set $\{x \mid f(x) \le f(x_0)\}$ is bounded then the iterations will remain in that set. Implicit filtering calls fdsteep repeatedly, reducing h after each termination of fdsteep. Aside from the data needed by fdsteep, a sequence of difference increments, the **scales**, $\{h_k\}_{k=0}^{\infty}$ is needed for the form of the algorithm given here.

```
ALGORITHM 2.2. imfilter1($x, f, kmax, \tau, \{h_k\}, amax$)
1. For $k = 0, \dots$
   Call fdsteep($x, f, kmax, \tau, h_k, amax$)
```

The first order estimate,

$$\|\nabla f_s(x_1) - \nabla_h f(x_1)\| + O(h + h^{-1}\|\phi\|_{S^k}), \tag{2.4}$$

where S^k is the set of points on the difference stencil used to compute $\nabla_{h_k} f$ and

$$\|\phi\|_{S^k} = \max_{z \in S^k} |\phi(z)|,$$

leads to a convergence result [2], [21].

THEOREM 2.1. *Let $h_k \to 0$ and let f satisfy* (2.1). *Let $\{x_k\}$ be the implicit filtering sequence. Assume that* (2.2) *holds (*i. e. *there is no line search failure) for all but finitely many k. Then if*

$$\lim_{k \to \infty} (h_k + h_k^{-1}\|\phi\|_{S^k}) = 0 \tag{2.5}$$

then any limit point of the sequence $\{x_k\}$ is a critical point of f_s.

If, for example, f_s has a unique minimizer x^* and

$$|\phi(x)| \le \epsilon \|x - x^*\|^2 \tag{2.6}$$

for x near x^* and ϵ sufficiently small, then one can prove global and q-linear convergence [19]. The requirement that (2.6) hold is very modest, since (2.6) demands only that the noise be smaller than $f_s(x) - f_s(x^*)$, but allows for the rate of decay (quadratic) to be the same. Superlinear convergence, as one might expect, will need stronger assumptions.

2.2. Quasi-Newton Methods.

Typically the performance of implicit filtering is greatly improved by using a quasi-Newton model Hessian [24], [19], [21]. The SR1 [3], [16] or BFGS quasi-Newton methods are defaults in the implicit filtering codes described in [18] and [21].

The finite difference quasi-Newton iteration is

$$x_+ = x_c + \lambda d_c,$$

where H_c is a quasi-Newton approximation to the Hessian of f_s and

$$d_c = -H_c^{-1} \nabla_h f(x_c).$$

The test for sufficient decrease is

$$f(x + \lambda d_c) - f(x) < \alpha\lambda\nabla_h f(x)^T d_c. \tag{2.7}$$

We will update the model Hessian with the BFGS formula

$$H_+ = H_c + \frac{yy^T}{y^T s} - \frac{(H_c s)(H_c s)^T}{s^T H_c s}. \tag{2.8}$$

The algorithmic description is taken from [21].

ALGORITHM 2.3. fdbfgs$(x, f, H, kmax, \tau, h, amax)$

```
1. For $k = 1, \dots, kmax$
    (a) Compute $f$, $\nabla_h f$. $d = -H^{-1}\nabla_h f$; terminate if (2.3) holds.
    (b) Find the least integer $0 \le m \le amax$ such that (2.7) holds for $\lambda = \beta^m$.
    (c) $x = x + \lambda d$.
    (d) Update $H$ with (2.8).
```

If the BFGS update of H fails to be positive definite, one must replace H with a positive definite matrix, for example by skipping the update or setting $H = I$.

As in the noise-free case, if the model Hessians remain positive definite, well-conditioned, and bounded, a simple convergence theorem holds.

THEOREM 2.2. *Let $h_k \to 0$ and let f satisfy* (2.1)*. Let $\{x_k\}$ be the BFGS/implicit filtering sequence and $\{H_k\}$ the model Hessians. Assume that* (2.7) *holds (*i. e. *there is no line search failure) for all but finitely many k. Assume that all H_k are symmetric positive definite, the sequences $\{\|H_k\|\}$ and $\{\|H_k^{-1}\|\}$ are bounded, and* (2.5) *holds. Then any limit point of the sequence $\{x_k\}$ is a critical point of f_s.*

2.3. A Model Problem and Idealized Method.

In this section we show how local and global assumptions,the structure of f_s, and the size of ϕ lead to idealized methods for which h can be computed as a function of x as the iteration progresses. This leads to estimates on the error in the gradients which will be used in 3.

As a model problem, we consider the case where f_s has a unique minimizer x^* and no other critical points. We assume that $\nabla^2 f_s(x*)$ is positive definite and that $\nabla^2 f_s$ is uniformly Lipschitz continuous. Hence there is $C_g > 0$ such that

$$C_g^{-1}\|x - x^*\| \le \|\nabla f_s(x)\| \le C_g\|x - x^*\|. \tag{2.9}$$

Assumption (2.6) implies that $\nabla_h\phi = O(\|x - x^*\|)$ near x^* with a small constant in the O-term. Superlinear convergence can not be proved under these conditions. However if

$$|\phi(x)| = O(\|x - x^*\|^{2+2p}) \tag{2.10}$$

for some $p > 0$ and the sequence of scales $\{h_k\}$ is managed properly, one can prove both local and global superlinear convergence.

2.3.1. Local Estimates.

Therefore, if (2.10) holds then

$$\nabla_h f(x) = \nabla f_s(x) + O(h^d + h^{-1}\|\nabla f_s(x)\|^{2+2p}), \tag{2.11}$$

where $d = 1$ for forward differences and $d = 2$ for centered differences.

The idealized component of the method is that we assume that h can be controlled so that

$$C_1^{-1}\|x - x^*\|^{1+p} \le h \le C_1\|x - x^*\|^{(1+p)/2}, \tag{2.12}$$

when x is sufficiently near x^*. One way to realize (2.12) in the case of a convergent (to x^*) quasi-Newton iteration $\{x_k\}$ is to set

$$h_{k+1} = \|\nabla_{h_k} f(x_k)\|^{1+p}.$$

Assuming that h_k satisfies (2.12) and expecting the convergence to be at worst q-linear and no better than quadratic, we would have, for some $C_1 > 0$ and n sufficiently large,

$$\begin{aligned} C_1^{-1}\|x_{k+1} - x^*\|^{1+p} &\le C_1^{-1}\|x_k - x^*\|^{1+p} \le h_{k+1} \\ &\le C_1\|x_k - x^*\|^{1+p} \le C_3\|x_{k+1} - x^*\|^{(1+p)/2}. \end{aligned}$$

which is (2.12).

From (2.11) we obtain, using centered differences ($d = 2$),

$$\begin{aligned} \nabla_h^c f(x) &= \nabla f_s(x) + O(\|x - x^*\|^{d(1+p)/2} + \|x - x^*\|^{1+p}) \\ &= \nabla f_s(x) + O(\|x - x^*\|^{1+p}). \end{aligned} \tag{2.13}$$

We base our analysis in § 3.1 on (2.13).

In [19] we used (2.6) and (2.9) (a small relative error in the gradient) to show global convergence. (2.13) is much stronger, but the bound of the relative error in the gradient by a power of the distance from optimality is necessary to prove superlinear convergence.

2.3.2. Global Estimates. We will require estimates for both f_s and the noise on the set

$$D_0 = \{x \,|\, f(x) \le 2f(x_0)\}.$$

In order to extend the global and superlinear convergence theory from [6] we must extend the assumptions from that paper.

ASSUMPTION 2.1.

1. *The set $D_0 = \{x \,|\, f(x) \le 2f(x_0)\}$ is bounded. and contains the convex hull of $D = \{x \,|\, f(x) \le f(x_0)\}$.*
2. *There are $M, m > 0$ such that $m\|u\|^2 \le u^T \nabla^2 f_s(x) u \le M\|u\|^2$ for all $x \in D_0$ and $u \in R^N$.*

For $x \in D_0$, we define, for some $p, \epsilon > 0$,

$$\xi(x : p, \epsilon) = \min\left(\epsilon\|x - x^*\|, \|x - x^*\|^{1+p}\right) \tag{2.14}$$

and assume that

$$|\phi(x)| = O\left(\xi(x : p, \epsilon)^2\right) \tag{2.15}$$

for all $x \in D_0$. For x near x^* (2.15) is equivalent to (2.10). However, when far from x^* (2.15) is much stronger. For sufficiently small ϵ, we have

$$|\phi(x)| \le \epsilon(f_s(x) - f_s(x^*)) \tag{2.16}$$

for all $x \in D_0$.

We replace (2.12) by

$$C_1^{-1}\xi(x : p, \epsilon) \le h \le C_1 \xi(x : p, \epsilon)^{1/2}. \tag{2.17}$$

This could be realized with

$$h_{k+1} = \min(\epsilon \|\nabla^c_{h_k} f(x_k)\|, \|\nabla^c_{h_k} f(x_k)\|^{1+p})$$

assuming a convergent iteration.

The estimates (2.17) and (2.15) imply that

$$\nabla^c_h f(x) - \nabla f_s(x) = O(\xi(x : p, \epsilon)). \tag{2.18}$$

(2.18) combines the small relative error estimate, needed for global convergence, from [19] with (2.13), which is necessary for local q-superlinear convergence.

3. Quadratic Models and Quasi-Newton Methods.

Throughout this section we assume that (2.10) holds and that gradients are approximated by central differences with a difference increment satisfying (2.12), for x near x^*. Following the notation [15] we let g denote an approximation of ∇f_s

$$g(x) = \nabla f_s(x) + N(x). \tag{3.1}$$

For example, in the notation of § 1, $g = \nabla^c_h f$ and $N(x) = O(\|x - x^*\|^{1+p})$.

The quasi-Newton implementation uses g instead of ∇f_s in both the computation of the BFGS step (we take full steps in a local theory)

$$s = -H_c^{-1} g(x_c) \tag{3.2}$$

and in the difference in gradients

$$y = g(x_+) - g(x_c), \tag{3.3}$$

both of which are used in the BFGS update (2.8) of the model Hessian H. In this paper we neglect floating point errors, so in the language of [15] $\epsilon_A = \mu = r_A = 0$.

3.1. Local Theory.

In this section we show how the estimates in [15] and classical analysis in [5] can be extended to problems that satisfy (3.1). We begin with the two main results from [15], specialized to the BFGS update, which satisfies the bounded deterioration inequality.

We make the standard assumptions [14], [21] that x^* is a local minimum of f_s, that $\nabla^2 f(x^*)$ is positive definite, and that $\nabla^2 f_s$ is Lipschitz continuous in a neighborhood of x^* with Lipschitz constant γ.

The basic estimate is a direct consequence of Lemma (2.4) in [15] with $p = 1$ and $\epsilon_A = \mu = r_A = 0$.

THEOREM 3.1. *There are $\delta \in (0, \|\nabla^2 f_s(x^*)^{-1}\|^{-1}/2)$ and $\bar\epsilon_F > 0$ so that if $\|x - x^*\| < \delta$, $\|H_c - \nabla^2 f_s(x^*)\| < \delta$, and $\|N(x)\| \le \bar\epsilon_F \|x - x^*\|$ then*

$$\|x_+ - x^*\| \le r\|x_c - x^*\| \tag{3.4}$$

where

$$r = 2\|\nabla^2 f_s(x^*)^{-1}\|\delta + \gamma\|\nabla^2 f_s(x^*)^{-1}\|\|x_c - x^*\| + \bar\epsilon_F \kappa(\nabla^2 f_s(x^*)). \tag{3.5}$$

In (3.5), κ is the ℓ^2 condition number.

We will need estimates of the difference between the update (2.8) using inaccurate data and a fully accurate BFGS update of H_c,

$$\bar{H}_+ = H_c + \frac{\bar{y}\bar{y}^T}{\bar{y}^T\bar{s}} - \frac{(H_c\bar{s})(H_c\bar{s})^T}{\bar{s}^T H_c \bar{s}}. \tag{3.6}$$

In (3.6) $\bar{s} = -H_c^{-1}\nabla f_s(x_c)$ and $\bar{y} = \nabla f_s(x_c + \bar{s}) - \nabla f_s(x_c)$. We define

$$M_c = H_+ - \bar{H}_+. \tag{3.7}$$

The estimates of M_c will be in terms of $\epsilon_f(x)$, which we define by

$$\|N(x)\| = \|x - x^*\|\epsilon_f(x). \tag{3.8}$$

Our assumptions imply that there are C_ϵ, δ and $p > 0$ such that

$$\epsilon_f(x) \le C_\epsilon \|x - x^*\|^p \tag{3.9}$$

whenever $\|x - x^*\| < \delta$.

LEMMA 3.2. *Assume that* (2.10) *holds and that gradients are approximated by central differences with a difference increment satisfying* (2.12)*. Then there are C_M and $\delta > 0$ such that if $\|x_c - x^*\| < \delta$ and $\|H_c - \nabla^2 f_s(x^*)\| < \delta$ then*

$$\|M_c\| \le C_M \|x_c - x^*\|^p. \tag{3.10}$$

Proof. We write $M_c = M_1 + M_2$ where

$$M_1 = \frac{yy^T}{y^Ts} - \frac{\bar{y}\bar{y}^T}{\bar{y}^T\bar{s}}, \text{ and } M_2 = \frac{(H_c\bar{s})(H_c\bar{s})^T}{\bar{s}^T H_c\bar{s}} - \frac{(H_cs)(H_cs)^T}{s^T H_c s}.$$

We will show that $M_1 = O(\|x_c - x^*\|^p)$. The bound on M_2 can be obtained in a similar fashion.

Let δ be small enough so that the hypotheses of Theorem 3.1 hold with

$$r \le 2\bar{\epsilon}_F \kappa(\nabla^2 f_s(x^*)) \le 2C_\epsilon \delta^p \kappa(\nabla^2 f_s(x^*)) < 1/2.$$

Since

$$y = g(x_+) - g(x_c) = \bar{y} + N(x_+) - N(x_c),$$

and $\|x_+ - x^*\| \le \|x_c - x^*\|/2$, we have

$$\|y - \bar{y}\| \le 2C_\epsilon \|x_c - x^*\|^{p+1}. \tag{3.11}$$

Hence,

$$yy^T = \bar{y}\bar{y}^T + O(\|x_c - x^*\|^{p+2}). \tag{3.12}$$

Similarly

$$y^Ts = \bar{y}^T\bar{s} + O(\|x_c - x^*\|^{p+2}). \tag{3.13}$$

The standard assumptions imply (reducing δ if necessary) that there is c_y

$$\|\bar{y}\| \geq c_y \|x_c - x^*\| \text{ and } |\bar{y}^T \bar{s}| \geq c_y \|x_c - x^*\|^2,$$

hence $M_1 = O(\|x_c - x^*\|^p)$, as asserted. □

One can use Lemma 3.2 to obtain a q-linear convergence from Theorem 3.1 via a bounded deterioration result. To do this we will invoke Theorem (2.5) of [15], which we will state in the context of this paper. In Theorem 3.3 $\|\cdot\|_F$ is the Frobenius norm.

THEOREM 3.3. *Assume that* (2.10) *holds and that gradients are approximated by central differences with a difference increment satisfying* (2.12). *Let* $r \in (0,1)$. *Then there are* $C > 0$ *and* δ *such that if* $\|x_0 - x^*\| < \delta$ *and* $\|H_0 - \nabla^2 f_s(x^*)\| < \delta$ *then for all* $n \geq 0$,

1. *H_n is nonsingular,*
2. $\|H_n - \nabla^2 f_s(x^*)\| \leq C\delta^p$, *and*
3. $\|x_{n+1} - x^*\| \leq r\|x_n - x^*\|$.

Moreover there are $\alpha_1, \alpha_2 > 0$ *such that a bounded deterioration inequality,*

$$\|H_{n+1} - \nabla^2 f_s(x^*)\|_F \leq (1 + \alpha_1 \|x_n - x^*\|)\|H_n - \nabla^2 f_s(x^*)\|_F + \alpha_2 \|x_n - x^*\|^p, \tag{3.14}$$

holds.

In order to obtain superlinear convergence, we need a more refined version of (3.14) and will extend the results of [5].

LEMMA 3.4. *Let the assumptions of Theorem 3.3 hold. Then there are* $\alpha_0, \alpha_1, \alpha_3$ *such that for* δ *sufficiently small and* $\|x_0 - x^*\| < \delta$ *and* $\|H_0 - \nabla^2 f_s(x^*)\| < \delta$

$$\begin{aligned} \|H_{k+1}^{-1} - \nabla^2 f_s(x^*)^{-1}\|_F &\leq \left[\sqrt{1 - \alpha_0 \Theta_k^2} + \alpha_1 \|x_k - x^*\|^p\right] \|H_k^{-1} - \nabla^2 f_s(x^*)^{-1}\|_F \\ &\quad + \alpha_2 \|x_k - x^*\|^p, \end{aligned} \tag{3.15}$$

where

$$\Theta_k = \frac{\|(H_k^{-1} - \nabla^2 f_s(x^*)^{-1}) y_k\|}{\|H_k^{-1} - \nabla^2 f_s(x^*)^{-1}\|_F \|y_k\|}.$$

Proof. The exact BFGS $\bar{H}_+$ update satisfies (3.15) with $p = 1$. The two terms, $\alpha_1 \|x_k - x^*\|^p$ and $\alpha_2 \|x_k - x^*\|^p$, on the right side of (3.15) account for the difference between the exact update (using (3.10)) and H_{k+1}. □

With Lemma 3.4 in hand, the proof of local superlinear convergence in [5] can be used in this case.

THEOREM 3.5. *Let the assumptions of Theorem 3.3 hold. Then there is* $\delta > 0$ *such that if* $\|x_0 - x^*\| < \delta$ *and* $\|H_0 - \nabla^2 f_s(x^*)\| < \delta$ *the quasi-Newton iteration given by* (3.2) *and* (2.8) *converges q-superlinearly to* x^*.

3.2. Global Theory.

In this brief section we describe how some of the results in [6] can be extended to the class of noisy problems considered in this paper. In the noisy case we must treat as assumptions two critical estimates that can be proved in the noise-free case. Having made those assumptions, the proof of global convergence in [6] requires only modest modifications.

The assumptions for the local theory do not imply good performance when far from x^*, even when combined with Assumption 2.1. The reasons for this are that because of the noise in the function the line search may fail and the inequalities

$$y^T s \geq m\|s\|^2 \text{ and } y^T s \geq M^{-1}\|y\|^2 \tag{3.16}$$

may not hold. However, (3.16) follows from Assumption 2.1 and (2.15) if ϵ is sufficiently small. Success of the line search and (3.16) were the critical components of the r-linear convergence result (Theorem 3.1) in [6].

Success of the line search is not guaranteed for noisy problems and convergence theorems for implicit filtering, such as Theorem 2.1 must assume that the line search fails only finitely often and that the scale is reduced when the line search fails. We must make the same assumption in order to prove the r-linear convergence theorem from [6].

With these new assumptions, the proof in [6] of r-linear convergence can be applied in the noisy case.

THEOREM 3.6. *Let Assumption 2.1,* (2.17)*, and* (2.15) *hold with $p > 0$ and ϵ sufficiently small. Let H_0 be symmetric positive definite. Let $\{x_k\}$ be the BFGS/implicit filtering iterations. Assume that the line search fails at most finitely many times. Then $\{x_k\}$ converges r-linearly to x^*.*

Proof. The proof is exactly the same as that in [6] with ϵ chosen small enough so that (2.16) holds and (2.17) and (2.15) used to conclude that convergence of the function values and (2.7) imply convergence of x_k to x^*. □

The success of the line search will follow from (2.17) and (2.15) provided the BFGS model Hessians are uniformly bounded and uniformly well-conditioned and the parameter ϵ in (2.15) is sufficiently small [19]. One could replace the assumption that ϵ be sufficiently small in the statement of Theorem 3.6 with the assumption that (3.16) fails only finitely many times and that the update is skipped when that happens.

The proof of superlinear convergence from [6] can be extended to the noisy case in the same way that the one from [5] was extended in § 3.1 by using (2.15) and (2.17) to derive (3.11), (3.12), and (3.13).

THEOREM 3.7. *Let Assumption 2.1,* (2.17)*, and* (2.15) *hold with $p > 0$ and ϵ sufficiently small. Let H_0 be symmetric positive definite. Let $\{x_k\}$ be the BFGS/implicit filtering iterations. Assume that the line search fails at most finitely many times. Then $\{x_k\}$ converges q-superlinearly to x^*.*

Acknowledgments. Both authors, especially the second, gratefully acknowledge several verbal and electronic communications with Richard Byrd, John Dennis, and Jorge Nocedal that helped us formulate the ideas in this paper.

REFERENCES

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

[2] D. M. BORTZ AND C. T. KELLEY, *The simplex gradient and noisy optimization problems*, in Computational Methods in Optimal Design and Control, J. T. Borggaard, J. Burns, E. Cliff, and S. Schreck, eds., vol. 24 of Progress in Systems and Control Theory, Birkhäuser, Boston, 1998, pp. 77–90.

[3] C. G. BROYDEN, *Quasi-Newton methods and their application to function minimization*, Math. Comp., 21 (1967), pp. 368–381.

[4] ———, *A new double-rank minimization algorithm*, AMS Notices, 16 (1969), p. 670.

[5] C. G. BROYDEN, J. E. DENNIS, AND J. J. MORÉ, *On the local and superlinear convergence of quasi-Newton methods*, J. Inst. Maths. Applics., 12 (1973), pp. 223–246.

[6] R. H. BYRD AND J. NOCEDAL, *A tool for the analysis of quasi-Newton methods with application to unconstrained minimization*, SIAM J. Numer. Anal., 26 (1989), pp. 727–739.

[7] R. G. CARTER, *On the global convergence of trust region algorithms using inexact gradient information*, SIAM J. Numer. Anal., 28 (1991), pp. 251–265.

[8] ———, *On the global convergence of trust region algorithms using inexact gradient information*, SIAM J. Sci. Comput., 14 (1993), pp. 368–388.

[9] T. D. CHOI, *Bound Constrained Optimization*, PhD thesis, North Carolina State University, Raleigh, North Carolina, 1999 (expected).

[10] A. R. CONN, , K. SCHEINBERG, AND P. L. TOINT, *On the convergence of derivative-free methods for unconstrained optimization*, in Approximation Theory and Optimization: Tributes to M. J. D. Powell, A. Iserles and M. Buhmann, eds., Cambridge, U.K., 1997, Cambridge University Press, pp. 83–108.

[11] ———, *Recent progress in unconstrained optimization without derivatives*, Math. Prog. Ser. B, 79 (1997), pp. 397–414.

[12] A. R. CONN AND P. L. TOINT, *An algorithm using quadratic interpolation for unconstrained derivative-free optimization*, Tech. Rep. 95/6, Facultès Universitaires de Namur, 1995.

[13] J. W. DAVID, C. T. KELLEY, AND C. Y. CHENG, *Use of an implicit filtering algorithm for mechanical system parameter identification*. SAE Paper 960358, 1996 SAE International Congress and Exposition Conference Proceedings, Modeling of CI and SI Engines, pp. 189–194.

[14] J. E. DENNIS AND R. B. SCHNABEL, *Numerical Methods for Nonlinear Equations and Unconstrained Optimization*, no. 16 in Classics in Applied Mathematics, SIAM, Philadelphia, 1996.

[15] J. E. DENNIS AND H. F. WALKER, *Inaccuracy in quasi-Newton methods: Local improvement theorems*, in Mathematical Programming Study 22: Mathematical programming at Oberwolfach II, North–Holland, Amsterdam, 1984, pp. 70–85.

[16] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming*, no. 4 in Classics in Applied Mathematics, SIAM, Philadelphia, 1990.

[17] R. FLETCHER, *A new approach to variable metric methods*, Comput. J., 13 (1970), pp. 317–322.

[18] P. GILMORE, *IFFCO: Implicit Filtering for Constrained Optimization*, Tech. Rep. CRSC-TR93-7, Center for Research in Scientific Computation, North Carolina State University, May 1993. available by anonymous ftp from math.ncsu.edu in pub/kelley/iffco/ug.ps.

[19] P. GILMORE AND C. T. KELLEY, *An implicit filtering algorithm for optimization of functions with many local minima*, SIAM J. Optim., 5 (1995), pp. 269–285.

[20] D. GOLDFARB, *A family of variable metric methods derived by variational means*, Math. Comp., 24 (1970), pp. 23–26.

[21] C. T. KELLEY, *Iterative Methods for Optimization*, no. 18 in Frontiers in Applied Mathematics, SIAM, Philadelphia, 1999.

[22] R. MIFFLIN, *A superlinearly convergent algorithm for minimization without deriviatives*, Math. Prog., 9 (1975), pp. 100–117.

[23] D. F. SHANNO, *Conditioning of quasi-Newton methods for function minimization*, Math. Comp., 24 (1970), pp. 647–657.

[24] D. STONEKING, G. BILBRO, R. TREW, P. GILMORE, AND C. T. KELLEY, *Yield optimization using a GaAs process simulator coupled to a physical device model*, IEEE Transactions on Microwave Theory and Techniques, 40 (1992), pp. 1353–1363.