International Journal of Pure and Applied Mathematics
Volume 63 No. 2 2010, 223-241

SOME REMARKS ON NEWTON'S ALGORITHM

Melisa Hendrata[1] §, P.K. Subramanian[2]
[1,2]Department of Mathematics
College of Natural and Social Sciences
California State University
5151, Los Angeles, State University Drive, Los Angeles, CA 90032-8103, USA
[1]e-mail: mhendra@calstatela.edu
[2]e-mail: mani@calstatela.edu

Abstract: Newton's algorithm and some of its variations are often used to find global minima of real valued functions. We propose another such variation and prove its local quadratic convergence. We combine this with Armijo type line search method to produce global convergence, which is eventually quadratic, for the important class of strictly convex functions. Computational performance on some standard test problems is presented, which shows that the proposed model may be viable.

AMS Subject Classification: 90xxx, 90-08
Key Words: Newton's method, Armijo line search, global optimization

1. Introduction

Let $f : \mathbb{R}^n \to \mathbb{R}$ and consider the problem of minimizing $f(x)$:

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

Given $x_0 \in \mathbb{R}^n$, most algorithms generate a sequence of points

$$x_{i+1} = x_i + \lambda_i p_i, \quad i \geq 0,$$

where λ_i is the stepsize chosen along the direction p_i (Ortega and Rheinboldt [12]). If $f(x)$ is differentiable, then p_i is a *descent direction* if $\nabla f(x_i)^T p_i < 0$ and in this case $f(x)$ decreases in a neighborhood of x_i along p_i. The sequence $\{f(x_i)\}$ so generated is decreasing, and if $f(x)$ has a minimizer x^*, then $x_i \to x^*$

Received: July 12, 2010
§Correspondence author

under appropriate conditions.

If $f(x)$ is twice continuously differentiable, the well known Newton's algorithm (Ortega and Rheinboldt [12]) is given by

$$x_{i+1} = x_i - \{\nabla^2 f(x_i)\}^{-1} \nabla f(x_i), \quad i \geq 0, \tag{1.2}$$

assuming $\nabla f(x_i) \neq 0$, the algorithm terminating otherwise. In this case $\lambda_i \equiv 1$ and $p_i = -\{\nabla^2 f(x_i)\}^{-1} \nabla f(x_i)$. This requires that the Hessian $\nabla^2 f(x_i)$ must be invertible. Moreover, if the Hessian is positive definite, then p_i is a descent direction. Algorithm (1.2) does not use this fact however, and the sequence $\{f(x_i)\}$ need not be decreasing. If $f(x)$ has a minimizer x^* in an open convex set D, $\nabla f(x^*) = 0$ and if $\nabla^2 f(x^*)$ is positive definite, then algorithm (1.2) guarantees (with additional strong conditions) convergence of the sequence $\{x_i\}$ to x^* quadratically. One of the conditions is that x_0 be sufficiently close to x^* so that the convergence is essentially local. To relax this condition and to insure one has global convergence, algorithm (1.2) is modified as

$$x_{i+1} = x_i - \lambda_i \{\nabla^2 f(x_i)\}^{-1} \nabla f(x_i), \quad i \geq 0, \tag{1.3}$$

so that the stepsize λ_i is not always one, but chosen carefully by using one of several line search methods.

In practice however, one does not know a priori if the sequence of Hessians $\{\nabla^2 f(x_i)\}$ will be positive definite for every i and algorithm (1.2) (as well as (1.3)) may very well fail. One alternative, see Dennis and Schnabel [11], would be to check if $\nabla^2 f(x_i)$ is positive definite for each i using its Cholesky decomposition or other means, and if it is not, to use $\nabla^2 f(x_i) + \mu_i I$ instead, where μ_i is a carefully chosen real number and I is the $n \times n$ identity matrix.

In this paper we suggest dispensing with the checking of the Hessian at each iteration. We also suggest using $\mu_i = \|\nabla f(x_i)\|$ and replacing the Hessian $\nabla^2 f(x_i)$ by the matrix $A(x_i) = \nabla^2 f(x_i) + \|\nabla f(x_i)\| I$ in algorithms (1.2) and (1.3). Admittedly there is still no guarantee that $A(x_i)$ will always be positive definite; however, we are encouraged by our computational experience (Section 4) that the modified algorithm is quite viable and works especially well when algorithm (1.3) fails. One reason why this happens may be that since $\|\nabla f(x_i)\| \to 0$ as $x_i \to x^*$, and quadratically so if x_i converges quadratically, in the vicinity of x^*, $A(x_i)$ closely mimics the behavior of the Hessian. Our experience with other choices for μ_i such as $\mu_i = 2^{-i}$ has not been very satisfactory often resulting in failure.

Instead of using $\|\nabla f(x_i)\| I$, more generally, one could define

$$A(x_i) = \nabla^2 f(x_i) + E(x_i),$$

where $E(x_i)$ is a continuous positive definite symmetric matrix for $x_i \neq x^*$ such

that $E(x^*) = 0$. Then $A(x_i)$ is positive definite for a large class of functions (such as strictly convex functions) and the *Modified Newton's Algorithm*

$$x_{i+1} = x_i - \{A(x_i)\}^{-1}\nabla f(x_i), \tag{1.4}$$

is well defined at all points. We shall prove in Theorem 2.1 that with roughly the same conditions as are required for quadratic convergence of the Newton's algorithm (1.2), the modified algorithm (1.4) guarantees local quadratic convergence as well.

Algorithm (1.4) is also essentially local, and one could use as in (1.3), a line search method to determine an appropriate stepsize λ_i to obtain global convergence. In this paper we suggest the use of a *backtracking* line search method generally known as the Armijo method or Armijo-Goldstein method [1]. Using this technique, in Theorem 3.3 we prove that the algorithm

$$x_{i+1} = x_i - \lambda_i \{A(x_i)\}^{-1}\nabla f(x_i), \tag{1.5}$$

guarantees global convergence under appropriate conditions. More importantly, in the case of strictly convex functions, the algorithm morphs itself into algorithm (1.4) in the vicinity of x^*, thus guaranteeing *eventual* quadratic convergence.

Newton's method as well as its modifications require computation of derivatives, viz., Hessians and gradients throughout the course of an algorithm. Admittedly, these are expensive both in computation as well as computer storage requirements. Where a code for the Hessian or gradient is not readily available, one can use numerical approximations or symbolic derivatives available through several software packages such as *Mathematica, Maple, Macsyma, Matlab.* A more promising new field is that of *automatic differentiation* where one uses computational representation of a function to produce exact values of the derivatives. The interested reader is referred to [2]-[9].

We use the Euclidean norm on $\mathbb{R}^n$. Real valued functions are denoted by lower case letters. We use upper case for operators $F : \mathbb{R}^n \to \mathbb{R}^m$, and in such cases use the operator norm. We write f_i for $f(x_i)$, f_* for $f(x^*)$, p_i for $p(x_i)$, A_i for $A(x_i)$, etc. We say F is *Lipschitz* continuous on $D \subset \mathbb{R}^n$ (and write $F \in \mathrm{Lip}_B(D)$) if

$$\forall\, x, y \in D,\ \|F(y) - F(x)\| \le B\|y - x\|, \tag{1.6}$$

for some constant $B > 0$. In particular, if $\nabla^2 f(x) \in \mathrm{Lip}_B(D)$, an important consequence of Taylor's theorem and (1.6) is the inequality

$$\forall\, x,\ y \in D,\ \|\nabla f(x) - \nabla f(y) - \nabla^2 f(y)(x - y)\| \le \frac{1}{2}B\|x - y\|^2. \tag{1.7}$$

2. Convergence of Modified Newton's Algorithm

In this section we prove that the algorithm (1.4) also converges quadratically under conditions that are very similar to those of algorithm (1.2). This will be useful in Theorem 3.3 where we prove that algorithm (1.5) converges globally with eventual quadratic convergence. See also Ortega and Rheinboldt [12].

Theorem 2.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable on a convex set $D \subset \mathbb{R}^n$. Suppose that the following conditions hold:*

1. *There exists $x^* \in D$ such that $\nabla f_* = 0$ and $\nabla^2 f_*$ is positive definite.*
2. *$\nabla^2 f(x) \in \mathrm{Lip}_K(D)$.*
3. *There exists an $n \times n$ symmetric matrix $E(x)$, continuous on D, positive definite for $x \neq x^*$ such that $E(x^*) = 0$, and $E(x) \in \mathrm{Lip}_C(D)$.*
4. *There exists $\lambda > 0$ such that $\forall\, x,\ y \in D, \lambda x^T x \leq x^T \nabla^2 f(y) x$.*
5. *Let $B = K + C$ and $\delta = 2\lambda/(B+C)$. Let $N = N(x^*, \delta) = \{x : \|x - x^*\| < \delta\}$, $N \subset D$ and let $x_0 \in D$. If $\nabla f(x_i) = 0$, stop, otherwise define $\{x_i\}$ inductively by (1.4) where*

$$A(x) = \nabla^2 f(x) + E(x).$$

Then:

1. *x^* is a local minimizer of f, $x_0 \in N(x^*, \delta) \Rightarrow \{x_i\} \subset N(x^*, \delta)$ and $x_i \to x^*$, and the rate of convergence is quadratic.*
2. *$\|\nabla f_i\| \to 0$ quadratically.*

Proof. We assume the algorithm does not terminate. Since

$$\lambda x^T x \leq x^T (\nabla^2 f_i) x + x^T (E_i) x = x^T A_i x,$$

for all $x \in D$, it follows that $\|A_i\| \geq \lambda$ and that $\|{A_i}^{-1}\| \leq 1/\lambda$.

Since $B = K + C$, it is clear that $A \in \mathrm{Lip}_B(D)$ and we have,

$$\begin{aligned} x_{i+1} - x^* &= x_i - x^* - A_i^{-1}\Big(\nabla f_i - \nabla f_*\Big) \\ &= A_i^{-1}\Big\{A_i(x_i - x^*) - \Big(\nabla f_i - \nabla f_*\Big)\Big\}, \\ \|x_{i+1} - x^*\| &= \|A_i^{-1}\| \|A_i(x_i - x^*) - \nabla f_i + \nabla f_*\| \\ &\leq \lambda^{-1} \|A_i(x_i - x^*) - \nabla f_i + \nabla f_*\|. \end{aligned} \tag{2.1}$$

But from (1.7) we have

$$\|\nabla f_* - \nabla f_i - A_i(x^* - x_i)\| \leq [(B+C)/2]\,\|x^* - x_i\|^2, \tag{2.2}$$

and substituting this in (2.1),

$$\|x_{i+1} - x^*\| \leq [(B+C)/2\lambda]\,\|x^* - x_i\|^2.$$

Since $\delta = 2\lambda/(B+C)$, we get

$$\|x_{i+1} - x^*\| \leq \delta^{-1}\,\|x_i - x^*\|^2, \tag{2.3}$$

which shows that $x_i \in N \Rightarrow x_{i+1} \in N$. In particular, $x_0 \in N \Rightarrow \{x_i\} \subset N$.

Finally, let $e_i = \delta^{-1}\|x_i - x^*\|$ so that from (2.3),

$$e_{i+1} = \delta^{-1}\|x_{i+1} - x^*\| \leq e_i^2 \Rightarrow e_i \leq (e_0)^{2^i}.$$

But $x_0 \in N \Rightarrow e_0 < 1$ and this shows $e_i \to 0$ and that the rate of convergence is quadratic.

Obviously $\nabla f_i \to 0$. From equation (1.4), $A_i(x_{i+1} - x_i) + \nabla f_i = 0$. It follows from (2.2) that

$$\begin{aligned}\|\nabla f_{i+1}\| = \|\nabla f_{i+1} - \big(\nabla f_i + A_i(x_{i+1} - x_i)\big)\| &\leq ((B+C)/2)\|x_{i+1} - x_i\|^2 \\ &\leq ((B+C)/2)\|A_i^{-1}\|^2\|\nabla f_i\|^2 \leq (1/\lambda\delta)\|\nabla f_i\|^2,\end{aligned}$$

so that $\nabla f_i \to 0$ quadratically completing the proof. □

3. Line Search Methods and the Modified Newton's Algorithm

In what follows given $x \in \mathbb{R}^n$ we call $p \in \mathbb{R}^n$ *gradient related* if

$$\left|\frac{\nabla f(x)^T p}{\|p\|}\right| \geq \sigma(\|\nabla f(x)\|), \tag{3.1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a forcing function, that is, $\sigma(0) = 0, \sigma(x) > 0\ \forall x > 0$, and $\sigma(x_i) \to 0 \Rightarrow x_i \to 0$ (Ortega and Rheinboldt [12]). If p_i is normalized so that $\|p_i\| = 1$, equation (3.1) can be rewritten as

$$\left|\nabla f(x)^T p\right| \geq \sigma(\|\nabla f(x)\|). \tag{3.2}$$

The following theorem is due to Wolfe [13], [14] and usually proved using Wolfe's or equivalent conditions (Dennis and Schnabel [11]). We give a brief proof using the Armijo backtracking line search method.

Theorem 3.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $x_0 \in \mathbb{R}^n$, $S = \{x | f(x) \leq f(x_0)\}$, a level set of f. Let $\nabla f \in \mathrm{Lip}_K(S)$. Assume that $\delta \in (0,1), \bar{\lambda}, \{\bar{\lambda}_i\}_{i\geq 1}$ be positive real numbers such that $\bar{\lambda}_i \geq \bar{\lambda} > 0$. Let $\{p(x)\}$*

be a sequence of gradient related descent directions and define the sequence $\{x_i\}$ *as follows:*

If $\nabla f(x_i) = 0$, *stop. Else define*

$$x_{i+1} = x_i + \lambda_i p_i,$$

where the stepsize $\lambda_i = \max_j \left\{ 2^{-j} \bar{\lambda}_i \right\}$ *is chosen so that:*

$$f_i - f_{i+1} \geq \delta \lambda_i (-\nabla f_i^T p_i). \tag{3.3}$$

Then either

1. $f(x)$ *is unbounded below on* S *or,*
2. $\dfrac{\nabla f_i^T p_i}{\|p_i\|} \to 0$ *and if* $\{x_i\}$ *has a limit point* x^*, $\nabla f_* = 0$.

Proof. If $\nabla f_i = 0$ there is nothing to prove, so assume the contrary. Without loss of generality we can assume p_i is normalized so that $\|p_i\| = 1$. Since p_i is a descent direction, $\nabla f_i^T p_i < 0$. An application of (1.7) shows

$$\left| f(x_i + \lambda_i p_i) - f(x_i) - \nabla f_i \lambda_i p_i \right| \leq (K/2) \|\lambda_i p_i\|^2 = (K/2) \lambda_i^2,$$

whence

$$\begin{aligned} f_i - f(x_i + \lambda_i p_i) &\geq \lambda_i \Big[- \nabla f_i^T p_i - (K/2) \lambda_i \Big] \\ &= \lambda_i \Big[- \delta \nabla f_i^T p_i - (1 - \delta) \nabla f_i^T p_i - (K/2) \lambda_i \Big] \\ &\geq \lambda_i \Big[- \delta \nabla f_i^T p_i \Big]. \end{aligned} \tag{3.4}$$

Hence $f_i \geq f_{i+1}$, provided $-(1 - \delta) \nabla f_i^T p_i - (K/2) \lambda_i \geq 0$, that is,

$$\lambda_i \leq \frac{2(1 - \delta)}{K} (-\nabla f_i^T p_i), \tag{3.5}$$

which is satisfied for sufficiently small λ_i. Hence, the sequence $\{f_i\}$ is decreasing.

By the choice of λ_i, either $\lambda_i = \bar{\lambda}_i \geq \bar{\lambda} > 0$, or $2\lambda_i$ violates inequality (3.5), that is, $2\lambda_i > (2(1 - \delta)/K)\,(-\nabla f_i^T p_i)$. It follows that

$$\lambda_i \geq \min \big(\bar{\lambda}, \frac{(1 - \delta)}{K} (-\nabla f_i^T p_i) \big). \tag{3.6}$$

From (3.4) we also have that

$$f(x_0) - f_i = \sum_{j=0}^{i-1} f(x_j) - f(x_{j+1}) \geq \delta \sum_{j=0}^{i-1} \lambda_j \Big(- \nabla f(x_j)^T p_j \Big).$$

Hence, either f is unbounded below on S, or $\sum_i \lambda_i \big(- \nabla f_i^T p_i \big)$ converges, that

is,

$$\lambda_i\big(-\nabla f_i^T p_i\big) \to 0.$$

From (3.6),

$$\begin{aligned} 0 &= \lim_{i\to\infty} \lambda_i\big(-\nabla f_i^T p_i\big) \\ &\geq \min\big(\bar{\lambda}\big(-\nabla f_i^T p_i\big), \frac{(1-\delta)}{K}(-\nabla f_i^T p_i)^2\big) \geq 0, \end{aligned}$$

that is,

$$\nabla f_i^T p_i \to 0,$$

and this proves the first part of the second conclusion since p_i is assumed to be normalized. Since p_i is gradient related, $\sigma(\|\nabla f_i\|) \to 0$ for some forcing function σ. It is obvious that if $\{x_i\}$ has a limit point x^*, $\nabla f_* = 0$, completing the proof. □

The following theorem is an adaptation of an important theorem due to Dennis and Moré [10], and is crucial for much of what follows. Although their original proof is in the context of Wolfe conditions, we prove it for backtracking line search algorithm (1.5).

Theorem 3.2. *Let:*

1. *$D \subseteq \mathbb{R}^n$ be open and convex and let $f : D \to \mathbb{R}$ be twice continuously differentiable on D with a minimizer x^*. Let $\nabla^2 f_*$ be positive definite and $\nabla^2 f \in Lip_{\gamma_1}(D)$.*

2. *For $x \in D$, let $E(x)$ be an $n \times n$ symmetric, continuous, positive definite matrix for $x \neq x^*$ with $E_* = 0$ and let $E(x) \in Lip_{\gamma_2}(D)$.*

3. *Let $0 < \delta < (1/2), \bar{\lambda}, \{\bar{\lambda_i}\}_{i\geq 1}$ be positive real numbers such that $\bar{\lambda_i} \geq \bar{\lambda} > 0$. Let $\{p_i\}$ be a sequence of gradient related descent directions and define $\{x_i\}, i \geq 0$ by*

$$x_{i+1} = x_i + \lambda_i p_i,$$

where $\lambda_i = \max\limits_{j\geq 1}\{2^{-j+1}\}$ satisfies $f_i - f_{i+1} \geq -\delta\lambda_i \nabla f_i^T p_i$.

4. *Assume that $x_i \to x^*$,*

$$\lim_{i\to\infty} \frac{\|\nabla f_i + \big(\nabla^2 f_i + E_i\big)p_i\|}{\|p_i\|} = 0. \tag{3.7}$$

Then there exists i_0 such that $i \geq i_0 \Rightarrow \lambda_i = 1$, that is, $x_{i+1} = x_i + p_i$.

Proof. We assume that $x_i \to x^*$. Since $\nabla^2 f_*$ is positive definite, there exists a neighborhood $N(x^*)$ such that for $x \in N$, $\nabla^2 f$ is *uniformly* positive definite. Hence, $\exists\, \mu, \nu > 0$ and such that for all $x, y \in N$,

$$\mu \, \|x\|^2 \leq x^T \nabla^2 f(y) x \leq \nu \|x\|^2. \tag{3.8}$$

Let $A(x) = \nabla^2 f(x) + E(x)$. Since $E(x)$ is bounded on N, there exists κ such that $\sup_{x \in N} \|E(x)\| \leq \kappa$. Then from (3.8)

$$\forall\, x, y \in N, \mu \|x\|^2 \leq x^T A(y) x \leq (\nu + \kappa) \|x\|^2. \tag{3.9}$$

Let $\gamma = \gamma_1 + \gamma_2$. By assumptions 1 and 2, $A(x) \in \mathrm{Lip}_\gamma(D)$, that is,

$$\forall\, x, y \in D, \|A(y) - A(x)\| \leq \gamma \|y - x\|. \tag{3.10}$$

Let

$$\sigma_i = \frac{\|\nabla f_i + A_i p_i\|}{\|p_i\|}.$$

We have

$$\begin{aligned} -\nabla f_i^T p_i &= p_i^T A_i \, p_i - \big(\nabla f_i + A_i p_i\big)^T p_i \\ &\geq \mu \|p_i\|^2 - \sigma_i \|p_i\|^2 \\ &= (\mu - \sigma_i) \, \|p_i\|^2 \end{aligned} \tag{3.11}$$

Also,

$$\sigma_i \to 0 \Rightarrow \exists\, i_0 \text{ such that } \forall\, i \geq i_0,\, \sigma_i \leq \frac{1}{2}\, \mu.$$

Hence for such i, from (3.11)

$$\frac{1}{2}\, \mu \, \|p_i\| \leq \left(\frac{-\nabla f_i^T p_i}{\|p_i\|} \right). \tag{3.12}$$

In particular, we have by Theorem 3.1 and (3.11) that

$$p_i \to 0. \tag{3.13}$$

For all $i \geq i_0$, there exists z_i in the line segment $[x_i, x_i + p_i]$ such that

$$f(x_i + p_i) - f_i = \nabla f_i^T p_i + \frac{1}{2}\, p_i^T \nabla^2 f(z_i) p_i.$$

Hence,

$$\begin{aligned} & f(x_i + p_i) - f_i - \frac{1}{2}\, \nabla f_i^T p_i \\ &= \frac{1}{2} \big(\nabla f_i + \nabla^2 f(z_i) p_i\big)^T p_i \\ &= \frac{1}{2} \big(\nabla f_i + A_i p_i\big)^T p_i + \frac{1}{2}\, p_i^T \big(\nabla^2 f(z_i) - A_i\big) p_i \end{aligned}$$

$$= \frac{1}{2}\left(\nabla f_i + A_i p_i\right)^T p_i + \frac{1}{2} p_i^T \big(A(z_i) - A_i - E(z_i)\big) p_i$$
$$\leq \frac{1}{2}\|\nabla f_i + A_i p_i\| \, \|p_i\| + \frac{1}{2}\gamma \, \|z_i - x_i\| \, \|p_i\|^2 + \frac{1}{2}\|E(z_i)\| \, \|p_i\|^2$$

by Cauchy-Schwarz inequality and (3.10).

Now $z_i \in [x_i, x_i + p_i]$ implies $\|z - x_i\| \leq \|p_i\|$. Further, since $x_i \to x^*$, we have $z_i \to x^*$ and $\|E(z_i)\| \to \|E(x^*)\| = 0$. Hence,

$$f(x_i + p_i) - f_i - \frac{1}{2}\nabla f_i^T p_i \leq \frac{1}{2}(\sigma_i + \gamma\|p_i\|) \, \|p_i\|^2 + \frac{1}{2}\|E(z_i)\| \, \|p_i\|^2. \tag{3.14}$$

Since $\sigma_i, p_i, \|E(z_i)\|$ all converge to 0, we can, without loss of generality, assume that for $i \geq i_0$,

$$\sigma_i + \gamma\|p_i\| + \|E(z_i)\| \leq \mu(\frac{1}{2} - \delta). \tag{3.15}$$

Hence, it follows from (3.12), (3.14), and (3.15) that

$$f(x_i + p_i) - f_i \leq \frac{1}{2}\nabla f_i^T p_i + \frac{1}{2}\Big(\mu\Big(\frac{1}{2} - \delta\Big) \cdot \frac{(-2\nabla f_i^T p_i)}{\mu}\Big)$$
$$= \frac{1}{2}\nabla f_i^T p_i - \frac{1}{2}(1 - 2\delta)\nabla f_i^T p_i$$
$$= \frac{1}{2}\nabla f_i^T p_i\big(1 - (1 - 2\delta)\big)$$
$$= \delta \nabla f_i^T p_i.$$

It follows that $i \geq i_0 \Rightarrow \lambda_i = 1$ completing the proof. □

We now prove the main theorem of this paper.

Theorem 3.3. *Let $D \subset \mathbb{R}^n$ be an open convex set, $f : D \to \mathbb{R}$, strictly convex and twice continuously differentiable on D and continuous in the closure of D. Given $x_0 \in D$, let*

$$S = \{x \,|\, f(x) \leq f(x_0)\}$$

be the level set of f. Assume that $\nabla^2 f(x) \in \mathrm{Lip}_\gamma(S)$.

1. *Suppose that f has a (unique) minimizer $x^* \in D$ and let $\nabla^2 f_*$ be positive definite.*
2. *Let $E(x)$ be a symmetric continuous $n \times n$ matrix for $x \in S$, positive definite for all $x \in S, x \neq x^*$, and satisfying $E(x^*) = 0$. Assume further that $E(x) \in \mathrm{Lip}_\mu(S)$ and define*
$$A_i = \nabla^2 f_i + E_i.$$
3. *Let $0 < \delta < (1/2), \bar{\lambda}, \{\bar{\lambda}_i\}_{i \geq 1}$ be positive real numbers such that $\bar{\lambda}_i \geq \bar{\lambda} > 0$ and let $x_0 \in D$. For $i \geq 0$, if $\nabla f_i = 0$ stop. Else define $\{x_i\}$ inductively*

by

$$x_{i+1} = x_i + \lambda_i p_i, \quad p_i = -A_i^{-1}\nabla f_i,$$

where $\lambda_i = \max_{j\geq 1}\{2^{-j+1}\}$ *satisfies* $f_i - f_{i+1} \geq \delta\,\lambda_i\big(-\nabla f_i^T p_i\big)$.

Then $x_i \to x^*$, *eventually quadratically, that is,* $x_i \to x^*$ *and there exists a neighborhood* $N(x^*)$ *of* x^* *and* i_0 *such that for all* $i \geq i_0, x_i \in N(x^*), \lim_k x_{i_0+k} = x^*$ *and the convergence is quadratic.*

Proof. Since x^* is the unique minimizer of f, the level set $S^* = \{x : f(x) \leq f_*\} = \{x^*\}$ is bounded. Hence, every level set of f, in particular S is bounded (see Ortega and Rheinboldt [12]), and therefore compact. Hence, $\|\nabla^2 f(x)\|$ is bounded on S, say by κ and $\nabla f(x) \in \mathrm{Lip}_\kappa(S)$. By Theorem 3.1, the sequence $\{f(x_i)\}$ is decreasing and the limit points of $\{x_i\}$ are stationary points and hence global minimizers. Since x^* is the unique global minimizer of f, it is the only limit point of $\{x_i\}$. That is, $x_i \to x^*$.

It remains to show that the convergence is eventually quadratic. By the strict convexity of f, the Hessian $\nabla^2 f(x)$ is positive semi-definite on D, and since $E(x), x \neq x^*$ is positive definite on S, A_i is positive definite on S, and since S is compact, A_i is uniformly positive definite on S. Hence, p_i is gradient related (Ortega and Rheinboldt [12]). Clearly

$$\lim_{i\to\infty} \frac{\|\nabla f_i + \big(\nabla^2 f_i + E_i\big)\, p_i\|}{\|p_i\|} = \lim_{i\to\infty} \frac{\|\nabla f_i + A_i\, p_i\|}{\|p_i\|}$$
$$= \lim_{i\to\infty} \frac{\|\nabla f_i - A_i(A_i)^{-1}\nabla f_i\|}{\|p_i\|} \equiv 0$$

and this shows that the hypotheses of Theorem 3.2 are satisfied. Thus, there exists a neighborhood $N_1(x^*)$ and i_0 such that for $i \geq i_0,\ x_i \in N_1(x^*)$ and

$$x_{i+1} = x_i + p_i = x_i - A_i^{-1}\nabla f_i.$$

Since $A(x) \in \mathrm{Lip}_{\mu+\gamma}(S)$, the conditions of Theorem 2.1 apply in $N_1(x^*)$. Thus, there exists another neighborhood $N(x^*) \subseteq N_1(x^*)$ and $i_1 \geq i_0$ such that $x_{i+i_1} \in N(x^*)$ and $x_{i+i_1} \to x^*$ quadratically completing the proof. □

4. Computational Experience

In this section we describe our computational experience with algorithm (1.5) which we will refer to as *Modified Newton's Algorithm* henceforth. For comparison purposes, we also include the performance of algorithm (1.3) which will

simply be referred to as *Newton's Algorithm*, and occasionally the standard Newton's Algorithm (1.2) ($\lambda_i = 1$). In the case of algorithm (1.5) we have chosen $A(x_i) = \nabla^2 f_i + \|\nabla f_i\| I$, and we have used Armijo backtracking line search to determine λ_i for both (1.3) and (1.5).

We tested the performance of the two algorithms on the following list of well known test problems: *Six Hump Camel Back function, Goldstein-Price function, Extended Rosenbrock function, Beale function*, and *Branin function*.

When the Hessians $\{\nabla^2 f(x_i)\}$ are known to be positive definite, the preferred algorithm would naturally be algorithm (1.2). Not surprisingly it did better than algorithm (1.5) in such cases. For instance, in the case of the ordinary ($n = 2$) *Rosenbrock function*, with the starting point at (-1.5, 2), Newton's Algorithm (1.2) with $\lambda_i = 1$ reached the minimum (1.0, 1.0) in seven iterations; algorithm (1.3) took 23 iterations while the Modified Newton's algorithm (1.5) took 30 iterations. For this reason this function is not considered here.

Often the Hessian may be indefinite at the starting point. This is the case with the *Six Hump Camel Back function* where with starting point (-0.5, 0.2) algorithm (1.3) converges to a saddle point but algorithm (1.5) succeeds in reaching the minimum (-0.09, 0.71). Similar situation occurs in *Goldstein-Price function* with starting point (-0.5, 1) for which algorithm (1.3) fails to converge while the algorithm (1.5) is able to find a local minimizer.

This situation need not improve even if the Hessian at the starting point is positive definite. This is the case with the *Extended Rosenbrock function,* $n = 4$ and the *Beale function*. In both cases, as the tables below show, algorithm (1.3) fails to converge while algorithm (1.5) succeeds in converging to a global minimizer.

For each problem, we give a statement of the problem and a list of local and global minimizers. A table provides the number of iterates computed, and whether the algorithm resulted in a success or failure either due to non-convergence or convergence to a point which is not a minimizer. We also indicate whether $\nabla^2 f(x_0)$ is positive definite or indefinite. In the case of algorithm (1.5) when convergence takes place we indicate the iterate when quadratic convergence has set (as predicted by Theorem 3.3) in italics.

We also provide a contour map showing the path traced by the iterates of both algorithms. All problems were coded in *Matlab* using symbolic derivatives.

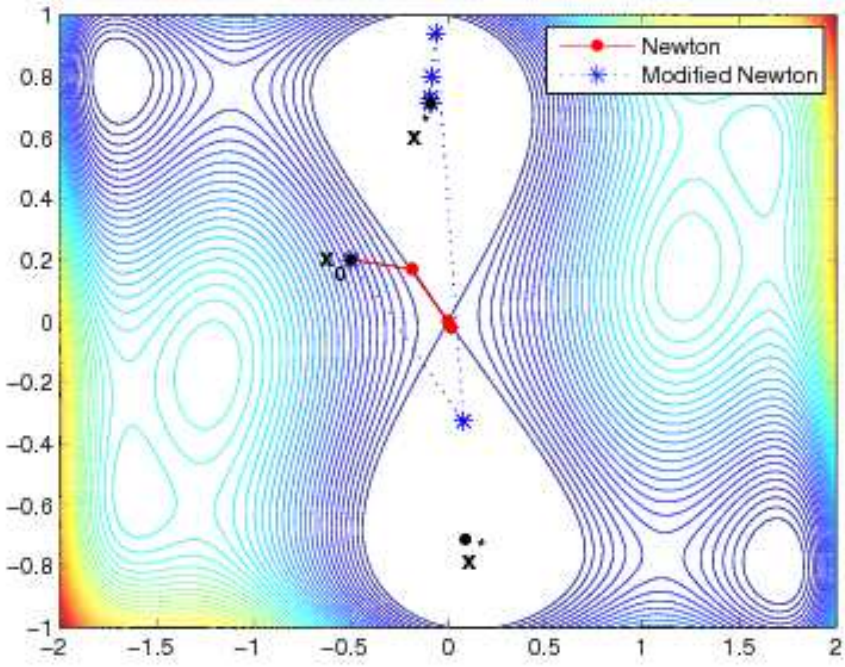

Figure 1: Contour plot of Six Hump Camel Back function and the trajectories taken by Newton's and Modified Newton's Algorithms

4.1. Six Hump Camel Back Function

$$f(x,y) = x^2(4 - 2.1x^2 + (x^4/3)) + xy + y^2(-4 + 4y^2).$$

There are two global minimizers $(-0.0898, 0.7126)$ and $(0.0898, -0.7126)$ and four local minimizers. At the starting point (-0.5,0.2), the Hessian $\nabla^2 f(x_0)$ is indefinite (Table 1).

4.2. Goldstein-Price Function

$$f(x,y) = [1 + (x + y + 1)^2(19 - 14x + 3x^2 - 14y + 6xy + 3y^2)], \cdot$$

Iter	Newton	$\|\nabla f\|$	Modified Newton	$\|\nabla f\|$
0	(-0.5000, 0.2000)	3.43496	(-0.5000, 0.2000)	3.43496
2	(0.0176,-0.0215)	0.22373	(-0.0624, 0.9408)	5.75004
4	(0.0000,-0.0000)	0.00000	*(-0.0907, 0.7300)*	*0.29280*
5			*(-0.0900, 0.7135)*	*0.01418*
6			*(-0.0898, 0.7127)*	*0.00004*
7			*(-0.0898, 0.7127)*	*0.00000*
	Saddle point: (0.0000,-0.0000) Indefinite Hessian		Global minimizer: (-0.0898, 0.7127)	

Table 1

Iter	Newton	$\|\nabla f\|$	Mod. Newton	$\|\nabla f\|$
0	(-0.500, 1.000)	191838.1	(-0.500, 1.000)	191838.1
5	(0.151, 0.233)	2544.711	(-0.868,-0.170)	583.7683
9	(0.151, 0.233)	2544.711	*(-0.603,-0.397)*	*1.9036*
10	(0.151, 0.233)	2544.711	*(-0.600,-0.400)*	*0.0431*
11	(0.151, 0.233)	2544.711	*(-0.600,-0.400)*	*0.0000*
12	(0.151, 0.233)	2544.711	*(-0.600,-0.400)*	*0.0000*
13	(0.151, 0.233)	2544.711	*(-0.600,-0.400)*	*0.0000*
100	(0.151, 0.233)	2544.711		
200	(0.151, 0.233)	2544.711		
	Fails to converge Indefinite Hessian		Local min.: (-0.600,-0.400)	

Table 2

$$[30 + (2x - 3y)^2(18 - 32x + 12x^2 + 48y - 36xy + 27y^2)].$$

There is one global minimizer $x^* = (0, -1)$ and several local minimizers. At the starting point (-0.5,1), the Hessian $\nabla^2 f(x_0)$ is indefinite (Table 2).

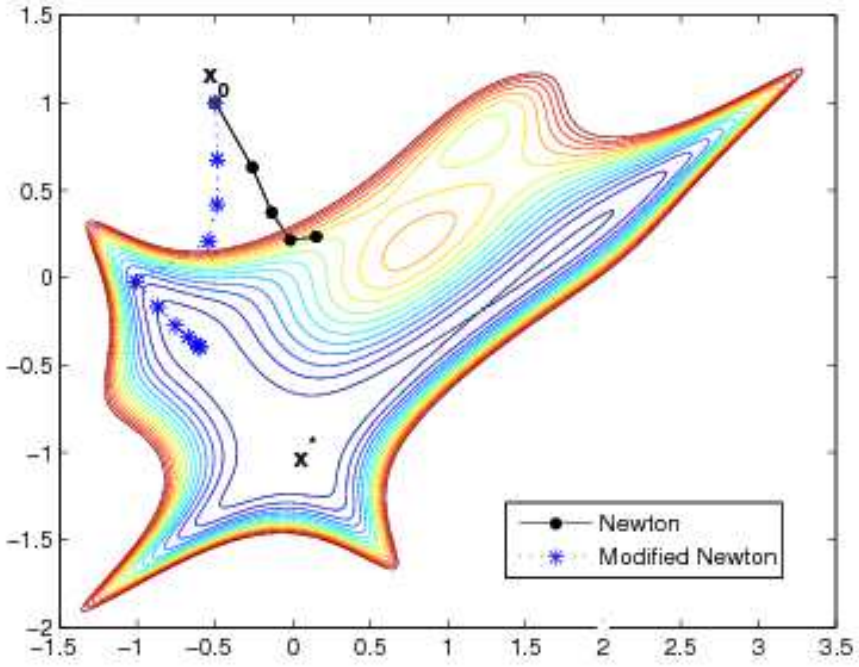

Figure 2: Goldstein-Price function contour plot with the sequence of iterates generated by Newton's and the Modified Newton's Algorithms

4.3. Extended Rosenbrock Function

$$f(x) = \sum_{i=1}^{3} [(1 - x_i)^2 + 100(x_{i+1} - x_i^2)^2].$$

There is a global minimizer $x^* = (1, 1, 1, 1)$ and several local minimizers. At the starting point $x_0 = (0, -2, 5, 2)$, the Hessian $\nabla^2 f(x_0)$ is positive definite (Table 3).

Iter.	Newton	$\|\nabla f\|$	Modified Newton	$\|\nabla f\|$
0	(0.0,-2.0, 5.0, 2.0)	46438	(0.0,-2.0, 5.0, 2.0)	46438
10	(0.0,-0.3, 0.3,-0.2)	116.55	(0.0,-1.0, 1.4, 2.1)	40.97
20	(-0.0, 0.0, 0.0,-0.0)	5.96	(0.8, 0.6, 0.3, 0.1)	2.71
31	(-0.0, 0.0, 0.0,-0.0)	5.96	*(1.0, 1.0, 1.0, 1.0)*	*0.19*
32	(-0.0, 0.0, 0.0,-0.0)	5.96	*(1.0, 1.0, 1.0, 1.0)*	*0.00*
33	(-0.0, 0.0, 0.0,-0.0)	5.96	*(1.0, 1.0, 1.0, 1.0)*	*0.00*
34	(-0.0, 0.0, 0.0,-0.0)	5.96	*(1.0, 1.0, 1.0, 1.0)*	*0.00*
50	(-0.0, 0.0, 0.0,-0.0)	5.96		
100	(-0.0, 0.0, 0.0,-0.0)	5.96		
200	(-0.0, 0.0, 0.0,-0.0)	5.96		
	Fails to converge Indefinite Hessian		Global min.: (1.0, 1.0, 1.0, 1.0)	

Table 3

Iter	Newton	$\|\nabla f\|$	Modified Newton	$\|\nabla f\|$
0	(-0.50000,-0.60000)	18.709	(-0.50000,-0.60000)	18.709
3	(0.49733,-1.46749)	4.4324	(1.29874,-0.41934)	4.0250
6	(0.49733,-1.46749)	4.4324	(2.42156, 0.31968)	0.4860
9	(0.49733,-1.46749)	4.4324	(2.92658, 0.48217)	0.0536
11	(0.49733,-1.46749)	4.4324	*(2.99873, 0.49971)*	*0.0011*
12	(0.49733,-1.46749)	4.4324	*(3.00000, 0.50000)*	*0.0000*
100	(0.49733,-1.46749)	4.4324		
200	(0.49733,-1.46749)	4.4324		
	Fails to converge Indefinite Hessian		Global minimizer: (3.00000, 0.50000)	

Table 4

4.4. Beale Function

$$f(x,y) = [1.5 - x(1-y)]^2 + [2.25 - x(1-y^2)]^2 + [2.625 - x(1-y^3)]^2.$$

Global minimizer $x^* = (3, 0.5)$. The Hessian at the starting point $\nabla^2 f(x_0)$ is positive definite (Table 4).

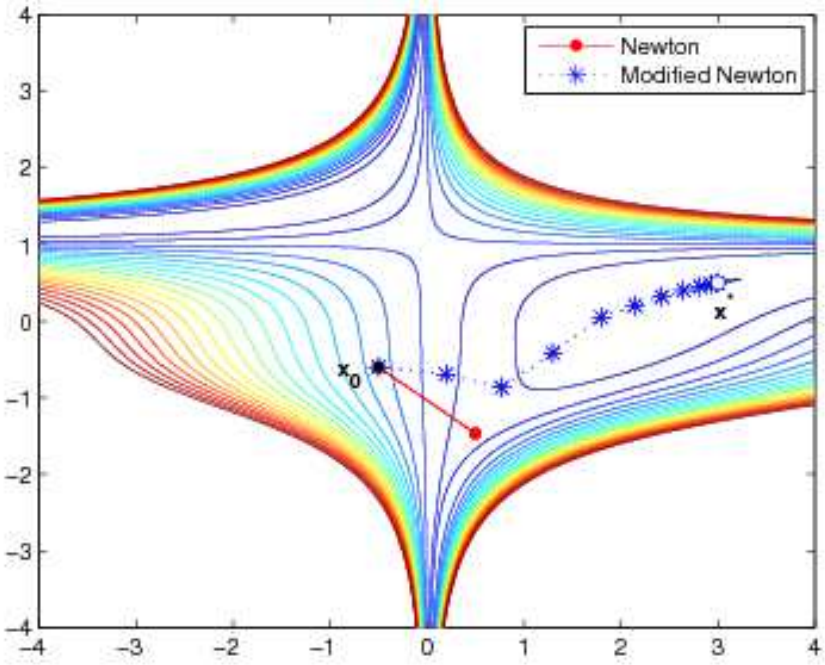

Figure 3: Contour plot of Beale function and the trajectories taken by Newton's and Modified Newton's Algorithms

4.5. Branin Function

$$f(x,y) = \Big[y - \Big(\frac{5.1}{4\pi^2}\Big)x^2 + \frac{5x}{\pi} - 6\Big]^2 + 10\Big(1 - \frac{1}{8\pi}\Big)\cos x + 10.$$

There are several global minimizers including $(-\pi, 12.275)$, $(\pi, 2.275)$, $(9.42478, 2.475)$. The Hessian at the starting point $\nabla^2 f(x_0)$ is positive definite (Table 5).

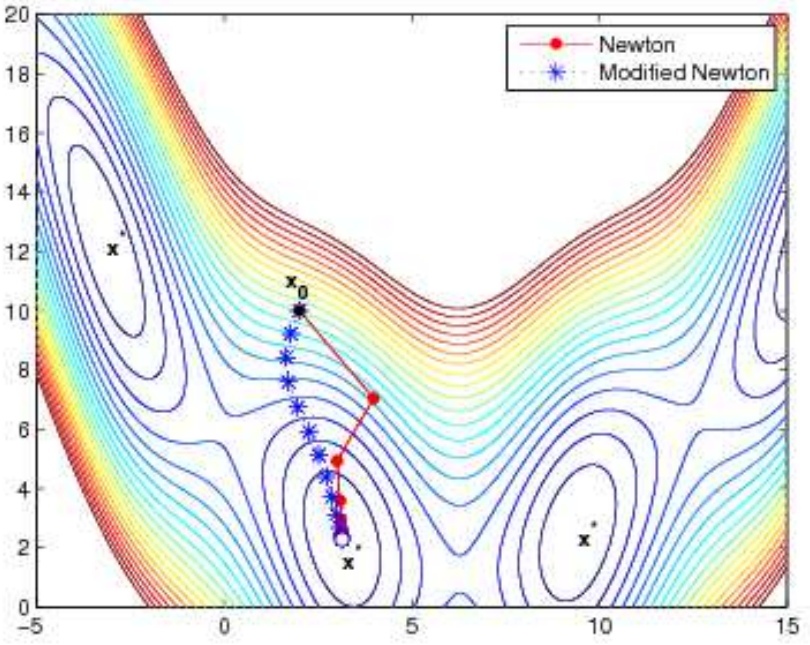

Figure 4: Contour plot of Branin function and the trajectories taken by Newton's and the Modified Newton's Algorithms

Iter	Newton	$\|\nabla f\|$	Modified Newton	$\|\nabla f\|$
0	(2.0000,10.0000)	14.4606	(2.0000,10.0000)	14.4606
4	(3.1155, 2.9301)	1.4733	(1.9560, 6.7382)	6.9034
8	(3.1401, 2.3159)	0.0923	(2.8729, 3.7131)	2.4849
11	(3.1414, 2.2801)	0.0115	*(3.1231, 2.3853)*	*0.1937*
12	(3.1415, 2.2776)	0.0058	*(3.1398, 2.2860)*	*0.0193*
13	(3.1416, 2.2763)	0.0029	*(3.1416, 2.2751)*	*0.0002*
14	(3.1416, 2.2756)	0.0014	*(3.1416, 2.2750)*	*0.0000*
18	(3.1416, 2.2750)	0.0000		
	Global minimizer: (3.1416, 2.2750)		Global minimizer: (3.1416, 2.2750)	

Table 5

References

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

[2] M. Berz, C. Bischof, C.F. Corliss, A. Griewank, *Computational Differentiation: Techniques, Applications, and Tools*, SIAM, Philadephia (1996).

[3] J.R. Birge, F. Louveaux, *Introduction to Stochastic Programming*, Springer-Verlag, New York (1997).

[4] C. Bischof, A. Bouaricha, P. Khademi, J.J. Moré, Computing gradients in large-scale optimization using automatic differentiation, *INFORMS J. Comp.*, **9** (1997), 185-194.

[5] C. Bischof, A. Carle, P. Khademi, A. Mauer, ADIFOR 2.0: Automatic differentiation of FORTRAN 77 programs, *IEEE Comp. Sci Eng.*, **3** (1996), 18-32.

[6] C. Bischof, G. Corliss, A. Griewank, Structured second- and higher-order derivatives through univariate Taylor series, *Optimization Methods and Software*, **2** (1993), 211-232.

[7] C. Bischof, M.R. Haghighat, On hierarchical differentiation, In: *Computational Differentiation: Techniques, Applications, and Tools*, SIAM, Philadelphia (1996), 83-94.

[8] C. Bischof, P. Khademi, A. Bouaricha, A. Carle, Efficient computation of gradients and Jacobians by transparent exploitation of sparsity in automatic differentiation, *Optimization Methods and Software*, **7** (1996), 1-39.

[9] C. Bischof, L. Roh, A. Mauer, ADIC: An extensible automatic differentiation tool for ANSI-C, *Software–Practice and Experience*, **27** (1997), 1427-1456.

[10] J.E. Dennis, J.J. Moré, A characterization of superlinear convergence and its application to quasi-Newton methods, *Math. Comp.*, **28** (1974), 549-560.

[11] J.E. Dennis, R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia (1996).

[12] J.M. Ortega, W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, SIAM, Philadelphia (2000).

[13] P. Wolfe, Convergence conditions for ascent methods, *SIAM Review*, **11** (1969), 226-235.

[14] P. Wolfe, Convergence conditions for ascent methods II: Some corrections, *SIAM Review*, **13** (1971), 185-188.