

Newton's method

Mark S. Gockenbach

1 Introduction

I have already briefly introduced Newton's method, which takes two forms, depending on whether the problem is to solve a system of nonlinear equations or to minimize a function of several variables. I will now review the derivation of Newton's method and analyze its convergence.

2 Newton's method for nonlinear systems

A system of nonlinear equations is expressed in the form $F(x) = 0$, where F is a vector-valued function of the vector variable x : $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$. Given an estimate $x^{(k)}$ of a solution x^* , Newton's method computes the (hopefully improved) estimate $x^{(k+1)}$ by setting the local linear approximation to F at $x^{(k)}$ to zero and solving for x :

$$\begin{aligned} F(x^{(k)}) + J(x - x^{(k)}) = 0 &\Rightarrow J(x - x^{(k)}) = -F(x^{(k)}) \\ &\Rightarrow x - x^{(k)} = -J^{-1}F(x^{(k)}) \\ &\Rightarrow x = x^{(k)} - J^{-1}F(x^{(k)}). \end{aligned}$$

In this calculation, $J = J(x^{(k)})$ is the Jacobian matrix of F at $x^{(k)}$. Therefore $x^{(k+1)}$ is defined by the formula

$$x^{(k+1)} = x^{(k)} - J^{-1}F(x^{(k)}), \quad k = 0, 1, 2, \dots \quad (1)$$

If J happens to be singular, then the Newton step is undefined, and a robust algorithm must be prepared to deal with such a situation. I will deal with this issue later. For now, I will simply assume that $J(x^*)$ is nonsingular, in which case the continuity of J will ensure that $J(x^{(k)})$ is nonsingular for any $x^{(k)}$ sufficiently near x^* .

2.1 An example of the convergence of Newton's method

As a concrete example, I define $F : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ by

$$F(x) = \begin{bmatrix} x_1^2 + x_2^2 - 1 \\ x_2 - x_1^2 \end{bmatrix}.$$

The solutions are points of intersection of the circle $x_1^2 + x_2^2 = 1$ and the parabola $x_2 = x_1^2$. A simple graph shows that there are two solutions, and some simple algebra shows that the solution lying in the first quadrant is

$$x^* = \left(\sqrt{\frac{\sqrt{5}-1}{2}}, \frac{\sqrt{5}-1}{2} \right)$$

(the other solution is the reflection of x^* across the x_2 -axis).

Applying Newton's method with $x^{(0)} = (0.5, 0.5)$, I obtain the results shown in Table 1. Several comments about these results are in order. First of all, the computations were carried out in IEEE double precision arithmetic, which translates to about 16 decimal digits of precision. Therefore, for example, where the results show that $\|x^* - x^{(5)}\| = 0$ and $\|F(x^{(5)})\| = 1.1102 \cdot 10^{-16}$, the apparent discrepancy is due to round-off error. Five iterations of Newton's method were sufficient

k	$\ x^* - x^{(k)}\ $	$\ F(x^{(k)})\ $	$x_1^{(k)}$	$x_2^{(k)}$
0	$3.0954 \cdot 10^{-1}$	$5.5902 \cdot 10^{-1}$	0.5000000000000000	5.0000000000000000
1	$8.9121 \cdot 10^{-2}$	$2.1021 \cdot 10^{-1}$	0.8750000000000000	0.6250000000000000
2	$4.5233 \cdot 10^{-3}$	$1.0090 \cdot 10^{-2}$	0.79067460317460314	0.6180555555555558
3	$1.2938 \cdot 10^{-5}$	$2.8769 \cdot 10^{-5}$	0.78616431593458214	0.61803398895790196
4	$1.0646 \cdot 10^{-10}$	$2.3673 \cdot 10^{-10}$	0.78615137786388734	0.61803398874989490
5	0.0000	$1.1102 \cdot 10^{-16}$	0.78615137775742328	0.61803398874989490

Table 1: Results of applying Newton's method to a 2×2 nonlinear system.

to compute the solution exactly to the given precision, but when $F(x^{(5)})$ is computed in floating point arithmetic, round-off error caused the result to differ from zero by a very small amount.

Second, the convergence of $\|x^* - x^{(k)}\|$ to zero follows a definite pattern:

$$\frac{\|x^* - x^{(3)}\|}{\|x^* - x^{(2)}\|^2} \doteq 0.6323, \quad \frac{\|x^* - x^{(4)}\|}{\|x^* - x^{(3)}\|^2} \doteq 0.6360,$$

which suggests that the ratio

$$\frac{\|x^* - x^{(k+1)}\|}{\|x^* - x^{(k)}\|^2}$$

is asymptotically constant as $k \rightarrow \infty$. It is difficult to verify this conjecture numerically, since the error so quickly falls below round-off level. For example, I would predict that

$$\|x^* - x^{(5)}\| \doteq 0.63\|x^* - x^{(4)}\|^2 \doteq 7.1 \cdot 10^{-21},$$

but in fact this error is below the precision of the machine, and all I can verify is that $\|x^* - x^{(5)}\|$ is less than about 10^{-16} . However, I will prove below that the conjecture is correct. In this regard, the following definitions are relevant.

Definition 2.1 Suppose $\{x^{(k)}\}$ is a sequence in \mathbb{R}^n that converges to x^* .

1. The sequence is said to converge linearly (or q-linearly) if there exists $c \in (0, 1)$ such that

$$\|x^* - x^{(k+1)}\| \leq c\|x^* - x^{(k)}\| \text{ for all } k \text{ sufficiently large.}$$

2. The sequence is said to converge superlinearly (or q-superlinearly) if

$$\frac{\|x^* - x^{(k+1)}\|}{\|x^* - x^{(k)}\|} \rightarrow 0 \text{ as } k \rightarrow \infty.$$

3. The sequence is said to converge quadratically (or q-quadratically) if there exists $c \in (0, \infty)$ such that

$$\|x^* - x^{(k+1)}\| \leq c\|x^* - x^{(k)}\|^2 \text{ for all } k \text{ sufficiently large.}$$

Cubic convergence, quartic convergence, and so on, are defined analogously (but are rarely used in analysis of optimization algorithms).

It is not difficult to show that quadratic convergence implies superlinear convergence, which in turn implies linear convergence.

Below I will prove the following theorem:

Theorem 2.2 Suppose $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuously differentiable and $F(x^*) = 0$. If

1. the Jacobian $J(x^*)$ of F at x^* is nonsingular, and
2. J is Lipschitz continuous on a neighborhood of x^* ,

then, for all $x^{(0)}$ sufficiently close to x^* , Newton's method produces a sequence $x^{(1)}, x^{(2)}, \dots$ that converges quadratically to x^* .

Lipschitz continuity is a technical condition that is stronger than the mere continuity of J but weaker than the condition that F be twice continuously differentiable. Lipschitz continuity will be defined carefully below.

Quadratic convergence has two important consequences. First of all, it guarantees that if a point close to the solution can be found, then Newton's method will rapidly home in on the exact solution. Second, it provides a *stopping test* for Newton's method. Since Newton's method is an iterative algorithm, it is necessary to have some criteria for deciding whether the current approximation $x^{(k)}$ is sufficiently close to the solution x^* that the algorithm can be halted. There is a simple stopping test for any superlinearly convergent sequence, which I will now derive.

I assume that it is desired to find $x^{(k)}$ such that $\|x^* - x^{(k)}\| < \epsilon$, where ϵ is a given error tolerance, and I also assume that $x^{(k)} \rightarrow x^*$ superlinearly. I will now show that

$$\frac{\|x^{(k)} - x^{(k-1)}\|}{\|x^* - x^{(k-1)}\|} \rightarrow 1 \text{ as } k \rightarrow \infty. \quad (2)$$

This implies that $\|x^{(k)} - x^{(k-1)}\|$ (which is a computable quantity) is a good estimate of $\|x^* - x^{(k-1)}\|$ when $x^{(k-1)}$ is close to x^* , and so it is reasonable to stop the iteration when $\|x^{(k)} - x^{(k-1)}\| < \epsilon$ is satisfied. This is not *guaranteed* to produce an estimate with an error less than ϵ , since it cannot be known for sure that $x^{(k-1)}$ is close enough to x^* that

$$\frac{\|x^{(k)} - x^{(k-1)}\|}{\|x^* - x^{(k-1)}\|} \doteq 1.$$

However, it works well because Newton's method usually does not take small steps unless it is close to the solution. Moreover, having verified that $\|x^{(k)} - x^{(k-1)}\| < \epsilon$ holds, so that $\|x^* - x^{(k-1)}\| < \epsilon$ is expected to hold, the algorithm then returns $x^{(k)}$, which should be much closer to x^* than $x^{(k-1)}$ is. For these reasons, the stopping test

$$\|x^{(k)} - x^{(k-1)}\| < \epsilon \quad (3)$$

is quite reliable.

To prove (2), I simply use the triangle inequality,¹ the reverse triangle inequality,² and the definition of superlinear convergence. First of all,

$$\begin{aligned} \frac{\|x^{(k)} - x^{(k-1)}\|}{\|x^* - x^{(k-1)}\|} &\leq \frac{\|x^{(k)} - x^*\| + \|x^* - x^{(k-1)}\|}{\|x^* - x^{(k-1)}\|} \\ &= 1 + \frac{\|x^{(k)} - x^*\|}{\|x^* - x^{(k-1)}\|}. \end{aligned}$$

Second,

$$\begin{aligned} \frac{\|x^{(k)} - x^{(k-1)}\|}{\|x^* - x^{(k-1)}\|} &\geq \frac{\left| \|x^{(k)} - x^*\| - \|x^{(k-1)} - x^*\| \right|}{\|x^* - x^{(k-1)}\|} \\ &= \left| \frac{\|x^{(k)} - x^*\|}{\|x^* - x^{(k-1)}\|} - 1 \right|. \end{aligned}$$

¹The triangle inequality, which is one of the defining properties of a norm, states that

$$\|x + y\| \leq \|x\| + \|y\| \text{ for all } x, y \in \mathbb{R}^n.$$

²The reverse triangle inequality, which can be proved from the triangle inequality, states that

$$\|x - y\| \geq \left| \|x\| - \|y\| \right| \text{ for all } x, y \in \mathbb{R}^n.$$

Therefore,

$$\left| \frac{\|x^{(k)} - x^*\|}{\|x^* - x^{(k-1)}\|} - 1 \right| \leq \frac{\|x^{(k)} - x^{(k-1)}\|}{\|x^* - x^{(k-1)}\|} \leq 1 + \frac{\|x^{(k)} - x^*\|}{\|x^* - x^{(k-1)}\|},$$

and so

$$\frac{\|x^{(k)} - x^*\|}{\|x^* - x^{(k-1)}\|} \rightarrow 0 \text{ as } k \rightarrow \infty$$

yields (2).

2.2 Proof of quadratic convergence of Newton's method

To prove Theorem 2.2 requires some background from linear algebra and multivariable calculus, which I will now review.

I need to apply the following result, which can be easily proved from the Fundamental Theorem of Calculus:

Theorem 2.3 *Suppose $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is continuously differentiable and $a, b \in \mathbb{R}^n$. Then*

$$F(b) = F(a) + \int_0^1 J(a + \theta(b - a))(b - a) d\theta, \quad (4)$$

where J is the Jacobian of F .

The integral of a vector-valued function, as in (4), is interpreted as the vector whose components are the integrals of the components of the integrand.³ I also need the *triangle inequality for integrals*:

Theorem 2.4 *If $F : \mathbb{R} \rightarrow \mathbb{R}^n$ is integrable over the interval $[a, b]$, then*

$$\left\| \int_a^b F(t) dt \right\| \leq \int_a^b \|F(t)\| dt. \quad (5)$$

In order to estimate the errors in Newton's method, I will need to use a matrix norm. The reader should recall the following definition:

Definition 2.5 *A norm $\|\cdot\|$ for a vector space X is a real-valued function defined on X satisfying the following properties:*

1. $\|x\| \geq 0$ for all $x \in X$, and $\|x\| = 0$ if and only if $x = 0$;
2. $\|\alpha x\| = |\alpha| \|x\|$ for all $x \in X$ and all scalars α ;
3. $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$ (the triangle inequality).

The space $\mathbb{R}^{m \times n}$ of $m \times n$ matrices is a vector space, since such matrices can be added and multiplied by scalars in a fashion analogous to Euclidean vectors. Many norms could be defined on $\mathbb{R}^{m \times n}$, but, as I will show, the following operator norm has significant advantages for analysis:

Definition 2.6 *Given any $A \in \mathbb{R}^{m \times n}$, the norm of A is defined by*

$$\|A\| = \max \left\{ \frac{\|Ax\|}{\|x\|} : x \in \mathbb{R}^n, x \neq 0 \right\} \quad (6)$$

The vector norms used on the right-hand side of (6) are the Euclidean norms on \mathbb{R}^m and \mathbb{R}^n , and the matrix norm is called the operator norm induced by the Euclidean norm.

³A digression: Equation (4) is commonly used when one wants to use a Mean Value Theorem (MVT). For a function $f : \mathbb{R} \rightarrow \mathbb{R}$ of one variable, the MVT (a special case of Taylor's theorem) states that if f is sufficiently smooth, then there exists $c \in (a, b)$ such that

$$f(b) = f(a) + f'(c)(b - a).$$

However, the MVT does not hold for vector-valued functions, because the number c is typically different for each component F_i . Equation (4) is usually an adequate substitute.

Theorem 2.7 *The norm defined by (6) has the following properties:*

1. *It is a norm on the space $\mathbb{R}^{m \times n}$;*
2. *$\|Ax\| \leq \|A\|\|x\|$ for all $A \in \mathbb{R}^{m \times n}, x \in \mathbb{R}^n$;*
3. *$\|AB\| \leq \|A\|\|B\|$ for all $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}$.*

The second and third properties of the operator norm are key in analyzing errors, particularly in producing upper bounds.

The next fact I need involves both linear algebra and analysis.

Theorem 2.8 *Suppose $J : \mathbb{R}^m \rightarrow \mathbb{R}^{n \times n}$ is a continuous matrix-valued function. If $J(x^*)$ is nonsingular, then there exists $\delta > 0$ such that, for all $x \in \mathbb{R}^m$ with $\|x - x^*\| < \delta$, $J(x)$ is nonsingular and*

$$\|J(x)^{-1}\| < 2 \|J(x^*)^{-1}\|.$$

This theorem implies that the set of nonsingular matrices is an open set. The second part of the theorem follows from the fact that, if $x \mapsto J(x)$ is continuous, then so is $x \mapsto J(x)^{-1}$ wherever this second map is defined.

Finally, I need to define Lipschitz continuity.

Definition 2.9 *Suppose $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$. Then F is said to be Lipschitz continuous on $S \subset \mathbb{R}^n$ if there exists a positive constant L such that*

$$\|F(x) - F(y)\| \leq L\|x - y\| \text{ for all } x, y \in S.$$

The same definition can be applied to a matrix-valued function $J : \mathbb{R}^n \rightarrow \mathbb{R}^{m \times n}$ (like the Jacobian), using a matrix norm to measure the size of $J(x) - J(y)$. The meaning of Lipschitz continuity is clear: The difference $F(x) - F(y)$ is, roughly speaking, proportional in size to $x - y$.

I can now prove Theorem 2.2. I begin with the definition of the Newton iteration,

$$x^{(k+1)} = x^{(k)} - J(x^{(k)})^{-1}F(x^{(k)}),$$

assuming that $x^{(k)}$ is close enough to x^* that $J(x^{(k)})$ is nonsingular. I then subtract x^* from both sides to obtain

$$x^{(k+1)} - x^* = x^{(k)} - x^* - J(x^{(k)})^{-1}F(x^{(k)}).$$

Since, by assumption, $F(x^*) = 0$, I can write this as

$$x^{(k+1)} - x^* = x^{(k)} - x^* - J(x^{(k)})^{-1}(F(x^{(k)}) - F(x^*)).$$

I now use (4) to estimate $F(x^{(k)}) - F(x^*)$:

$$\begin{aligned} F(x^{(k)}) - F(x^*) &= \int_0^1 J(x^* + \theta(x^{(k)} - x^*))(x^{(k)} - x^*) d\theta \\ &= \int_0^1 J(x^*)(x^{(k)} - x^*) d\theta + \int_0^1 \left(J(x^* + \theta(x^{(k)} - x^*)) - J(x^*) \right) (x^{(k)} - x^*) d\theta \\ &= J(x^*)(x^{(k)} - x^*) + \int_0^1 \left(J(x^* + \theta(x^{(k)} - x^*)) - J(x^*) \right) (x^{(k)} - x^*) d\theta. \end{aligned}$$

Therefore,

$$\begin{aligned} \left\| F(x^{(k)}) - F(x^*) - J(x^*)(x^{(k)} - x^*) \right\| &= \left\| \int_0^1 \left(J(x^* + \theta(x^{(k)} - x^*)) - J(x^*) \right) (x^{(k)} - x^*) d\theta \right\| \\ &\leq \int_0^1 \left\| \left(J(x^* + \theta(x^{(k)} - x^*)) - J(x^*) \right) (x^{(k)} - x^*) \right\| d\theta \end{aligned}$$

$$\begin{aligned}
&\leq \int_0^1 \left\| J(x^* + \theta(x^{(k)} - x^*)) - J(x^*) \right\| \|x^{(k)} - x^*\| d\theta \\
&\leq \int_0^1 L\theta \|x^{(k)} - x^*\|^2 d\theta \\
&= \frac{L}{2} \|x^{(k)} - x^*\|^2.
\end{aligned}$$

(The reader should notice that, without the Lipschitz continuity of J , I can conclude that $F(x^{(k)}) - F(x^*) - J(x^*)(x^{(k)} - x^*) = o(\|x^{(k)} - x^*\|)$, but I need the Lipschitz continuity and the above argument to get the stronger estimate $F(x^{(k)}) - F(x^*) - J(x^*)(x^{(k)} - x^*) = O(\|x^{(k)} - x^*\|^2)$).

I now have

$$\begin{aligned}
x^{(k+1)} - x^* &= x^{(k)} - x^* - J(x^{(k)})^{-1} (F(x^{(k)}) - F(x^*)) \\
&= x^{(k)} - x^* - J(x^{(k)})^{-1} \left(J(x^*)(x^{(k)} - x^*) + F(x^{(k)}) - F(x^*) - J(x^*)(x^{(k)} - x^*) \right) \\
&= \left(I - J(x^{(k)})^{-1} J(x^*) \right) (x^{(k)} - x^*) - \\
&\quad J(x^{(k)})^{-1} \left(F(x^{(k)}) - F(x^*) - J(x^*)(x^{(k)} - x^*) \right),
\end{aligned}$$

and so

$$\begin{aligned}
\|x^{(k+1)} - x^*\| &\leq \left\| \left(I - J(x^{(k)})^{-1} J(x^*) \right) (x^{(k)} - x^*) \right\| + \\
&\quad \left\| J(x^{(k)})^{-1} \left(F(x^{(k)}) - F(x^*) - J(x^*)(x^{(k)} - x^*) \right) \right\| \\
&\leq \left\| I - J(x^{(k)})^{-1} J(x^*) \right\| \|x^{(k)} - x^*\| + \\
&\quad \left\| J(x^{(k)})^{-1} \right\| \left\| F(x^{(k)}) - F(x^*) - J(x^*)(x^{(k)} - x^*) \right\| \\
&\leq \left\| I - J(x^{(k)})^{-1} J(x^*) \right\| \|x^{(k)} - x^*\| + \frac{L}{2} \left\| J(x^{(k)})^{-1} \right\| \|x^{(k)} - x^*\|^2.
\end{aligned}$$

I use the Lipschitz continuity of J again, this time to estimate the size of $I - J(x^{(k)})^{-1} J(x^*)$:

$$\begin{aligned}
\left\| I - J(x^{(k)})^{-1} J(x^*) \right\| &= \left\| J(x^{(k)})^{-1} \left(J(x^{(k)}) - J(x^*) \right) \right\| \\
&\leq \left\| J(x^{(k)})^{-1} \right\| \left\| J(x^{(k)}) - J(x^*) \right\| \\
&\leq L \left\| J(x^{(k)})^{-1} \right\| \|x^{(k)} - x^*\|.
\end{aligned}$$

I have now obtained

$$\|x^{(k+1)} - x^*\| \leq \frac{3L}{2} \left\| J(x^{(k)})^{-1} \right\| \|x^{(k)} - x^*\|^2.$$

The final step is to recognize that, for all $x^{(k)}$ sufficiently close to x^* ,

$$\left\| J(x^{(k)})^{-1} \right\| \leq 2M, \tag{7}$$

where $M = \left\| J(x^*)^{-1} \right\|$. Then, for $x^{(k)}$ sufficiently close to x^* ,

$$\|x^{(k+1)} - x^*\| \leq 3LM \|x^{(k)} - x^*\|^2. \tag{8}$$

If

$$\|x^{(k)} - x^*\| < \frac{1}{6LM}, \tag{9}$$

then

$$\|x^{(k+1)} - x^*\| < \frac{1}{2} \|x^{(k)} - x^*\|. \tag{10}$$

I have now proved Theorem 2.2: If $x^{(0)}$ is chosen close enough to x^* that (7) and (9) both hold, then (10) shows that $x^{(k)} \rightarrow x^*$ and (8) shows that the convergence is quadratic.

k	$\ x^* - x^{(k)}\ $	$\ \nabla f(x^{(k)})\ $	$x_1^{(k)}$	$x_2^{(k)}$
16	$1.0765 \cdot 10^{-3}$	$1.9962 \cdot 10^{-8}$	1.0007612194201734	1.9992387805798266
17	$7.1768 \cdot 10^{-4}$	$5.9145 \cdot 10^{-9}$	1.0005074796134474	1.9994925203865526
18	$4.7846 \cdot 10^{-4}$	$1.7525 \cdot 10^{-9}$	1.0003383197422977	1.9996616802577023
19	$3.1897 \cdot 10^{-4}$	$5.1925 \cdot 10^{-10}$	1.0002255464948688	1.9997744535051312
20	$2.1265 \cdot 10^{-4}$	$1.5385 \cdot 10^{-10}$	1.0001503643299121	1.9998496356700879

Table 2: Results of applying Newton's method minimize a function of two variables.

3 Newton's method for unconstrained minimization

Since Newton's method for solving

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

is nothing more than Newton's method applied to the nonlinear system

$$\nabla f(x) = 0,$$

the following theorem is a corollary to Theorem 2.2:

Theorem 3.1 *Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is twice continuously differentiable, and $x^* \in \mathbb{R}^n$ satisfies*

1. $\nabla f(x^*) = 0$;
2. $\nabla^2 f(x^*)$ is positive definite (and hence, in particular, nonsingular);
3. $\nabla^2 f$ is Lipschitz continuous on a neighborhood of x^* .

Then x^ is a strict local minimizer of f and, for any $x^{(0)}$ sufficiently close to x^* , Newton's method defines a sequence that converges quadratically to x^* .*

I now give an example that shows that Newton's method can still converge if the hypotheses of the above theorem fail, specifically, if $\nabla^2 f(x^*)$ is positive semidefinite and singular. I define $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ by

$$f(x) = (x_1 + x_2 - 3)^2 + (x_1 - x_2 + 1)^4.$$

Then

$$\begin{aligned} \nabla f(x) &= \begin{bmatrix} 2(x_1 + x_2 - 3) + 4(x_1 - x_2 + 1)^3 \\ 2(x_1 + x_2 - 3) - 4(x_1 - x_2 + 1)^3 \end{bmatrix}, \\ \nabla^2 f(x) &= \begin{bmatrix} 2 + 12(x_1 - x_2 + 1)^2 & 2 - 12(x_1 - x_2 + 1)^2 \\ 2 - 12(x_1 - x_2 + 1)^2 & 2 + 12(x_1 - x_2 + 1)^2 \end{bmatrix}. \end{aligned}$$

An easy calculation show that, with $x^* = (1, 2)$,

$$f(x^*) = 0, \quad \nabla f(x^*) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \nabla^2 f(x^*) = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix},$$

and x^* is the unique global minimizer of f . However, the eigenvalues of $\nabla^2 f(x^*)$ are 0 and 4, and so $\nabla^2 f(x^*)$ is positive semidefinite and singular.

Beginning with $x^{(0)} = (2, 2)$, Newton's method produces the results shown in Table 2. (To save space, I only show iterates $x^{(16)}, \dots, x^{(20)}$.)

The results suggest that Newton's method produces a sequence that converges to x^* . However, the convergence is definitely not quadratic. Indeed, the ratios

$$\frac{\|x^{(17)} - x^*\|}{\|x^{(16)} - x^*\|} \doteq 0.6666,$$

$$\frac{\|x^{(17)} - x^*\|}{\|x^{(16)} - x^*\|} \doteq 0.6667,$$

$$\frac{\|x^{(17)} - x^*\|}{\|x^{(16)} - x^*\|} \doteq 0.6667,$$

$$\frac{\|x^{(17)} - x^*\|}{\|x^{(16)} - x^*\|} \doteq 0.6667$$

strongly suggest that $x^{(k)} \rightarrow x^*$ linearly. A comparison between Tables 1 and 2 shows the desirability of quadratic (or at least superlinear) convergence.

4 Advantages and disadvantages of Newton's method

The results that I have derived above show that Newton's method produces excellent *local* convergence. However, there is no reason to expect that the algorithm will behave well when $x^{(0)}$ is chosen far from x^* . Indeed, the algorithm may not even be defined; when an iterate $x^{(k)}$ is encountered with the property that $J(x^{(k)})$ or $\nabla^2 f(x^{(k)})$ is singular, then Newton's method does not define $x^{(k+1)}$. Moreover, in the case of a minimization problem, the sequence may converge to a stationary point of f that is not a local minimizer, such as a local maximizer or saddle point.

For these reasons, it is necessary to enhance Newton's method to obtain *global* convergence (that is, convergence to a solution from a starting point that may be far away). Whatever techniques are used to ensure global convergence should not ruin the excellent local convergence exhibited by Newton's method.