Chapter 3

Nonlinear equations

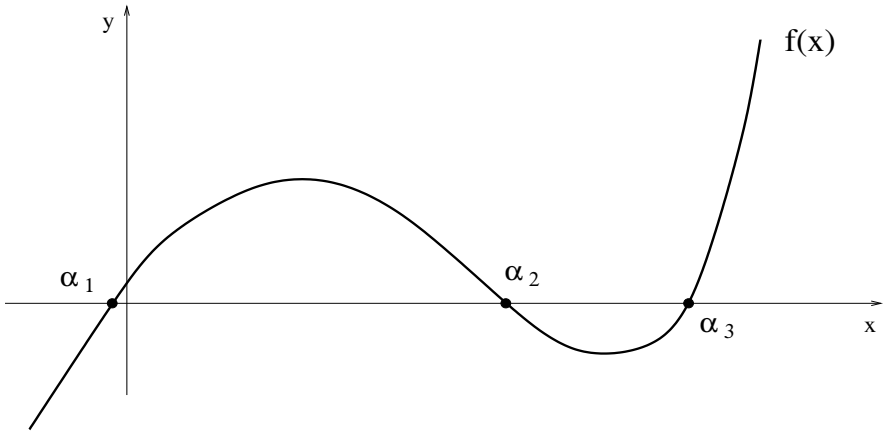
A common problem in scientific computing is to find the zeros of a function $f : \mathbb{R} \to \mathbb{R}$, i.e. the roots of the equation

$$f(x) = 0. \tag{3.1}$$

Note that any equation of the form $h(x) = g(x)$ can be written in the form (3.1) by setting $f(x) = h(x) - g(x)$. So (3.1) covers all equations involving functions of a single real variable.

For some special cases it is possible to write down explicit formulas for the roots of (3.1). But if f is a polynomial of order larger than or equal to five, then the roots cannot in general be expressed in this way. This implies that there is certainly no general explicit solution to the equation $f(x) = 0$ valid for an arbitrary continuous real valued function f.

In this chapter we will discuss iterative methods for the solution of (3.1).



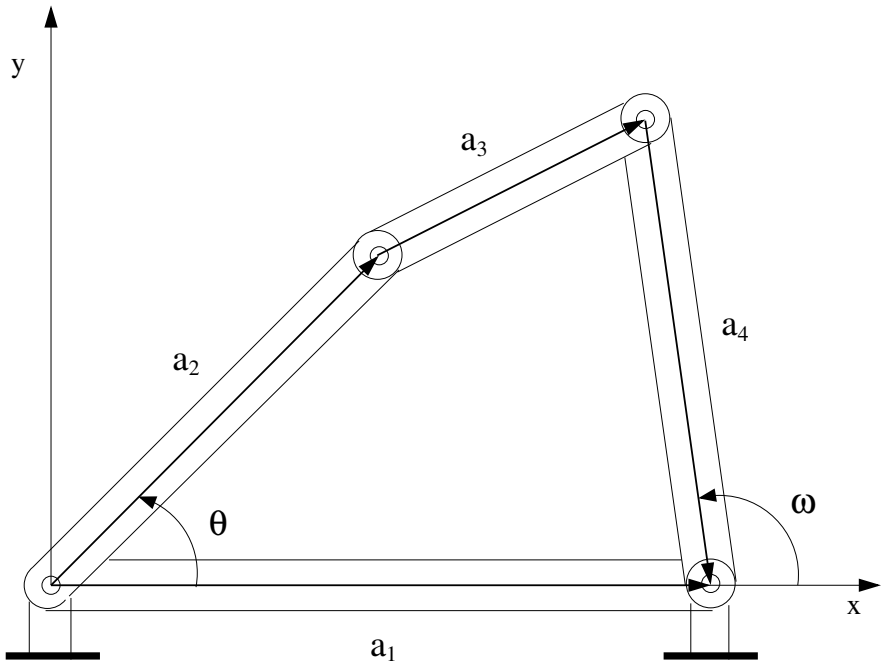


Figure 3.1: *The four joined beams*

3.1 Examples and motivation

Example 3.1.1 (Average interest rate)**.** *Suppose we wish to compute the average interest rate i on an investment over n years. $v = £1000$ has been invested every year and after 5 years the resulting sum is $p = £6000$.*

We know that p, v, i and the number of years n are related by the formula

$$p = v\sum_{k=1}^{n}(1+i)^k = v\frac{1+i}{i}\left[(1+i)^n - 1\right].$$

We may therefore write the problem as: find i such that

$$f(i) = p - v\frac{1+i}{i}\left[(1+i)^n - 1\right] = 0. \tag{3.2}$$

It follows that we must solve a nonlinear equation for which there is no analytical solution.

Example 3.1.2 (Configuration of beams)**.** *Consider the mechanical system of four interconnected beams shown in Figure 3.1. Given ω and the lengths a_i, $i = 1, \ldots, 4$ of the beams suppose that we wish to determine the angle θ between $\mathbf{a}_1$ and $\mathbf{a}_2$. Using the identity*

$$\mathbf{a}_1 - \mathbf{a}_2 - \mathbf{a}_3 - \mathbf{a}_4 = 0,$$

and observing that $\mathbf{a}_1$ is parallel to the x-axis, we find the following relation between ω and θ:

$$\frac{a_1}{a_2}cos(\omega) - \frac{a_1}{a_4}cos(\theta) - cos(\omega - \theta) = -\frac{a_1^2 + a_2^2 - a_3^2 + a_4^2}{2a_2a_4}. \tag{3.3}$$

The equation (3.3) can be solved analytically only for certain values of ω. In the general case the solution must be approximated numerically.

3.2 Bisection method

For continuous [1] functions f the following results of real analysis will be useful.

Theorem 3.2.1 (Intermediate value theorem)**.** *Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$. Then $f(x)$ is bounded on $[a, b]$ and achieves its bounds. If y is any number such that*

$$\min_{x\in[a,b]} f(x) \leq y \leq \max_{x\in[a,b]} f(x),$$

then there exists $\alpha \in [a, b]$ such that $f(\alpha) = y$.

An immediate consequence of the intermediate value theorem is

Theorem 3.2.2. *Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$, and assume that $f(a)f(b) \leq 0$. Then there exists $\alpha \in [a, b]$ such that $f(\alpha) = 0$.*

Proof. If $f(a)f(b) = 0$ then either $f(a) = 0$ or $f(b) = 0$, which yields $\alpha = a$ or $\alpha = b$ respectively. Otherwise we must have $f(a)f(b) < 0$, in which case

$$\min_{x\in[a,b]} f(x) < 0 < \max_{x\in[a,b]} f(x),$$

and the claim follows by the intermediate value theorem. □

Note that the root α may not be unique; there may be multiple roots in the interval $[a, b]$. To guarantee uniqueness one needs to impose additional constraints on f, e.g. requiring that it is differentiable on $[a, b]$ and that $f'(x) \neq 0$ for any $x \in [a, b]$.

These theoretical results naturally suggest a method for root finding for continuous functions f, called the *bisection method*. The idea is to use Theorem 3.2.2 repeatedly, looking for smaller and smaller intervals on which $f(x)$ changes sign, creating a sequence $x_0, x_1, x_2, \ldots$ converging to a root α. Suppose we have a function f satisfying the conditions of Theorem 3.2.2. Then the bisection algorithm takes the following form:

1. Set $k = 0$.

[1] Recall that a function $f : [a, b] \to \mathbb{R}$ is continuous at $x_* \in [a, b]$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that $|x - x_*| < \delta$ implies $|f(x) - f(x_*)| < \epsilon$. Equivalently, if for every sequence $(x_n) \subset [a, b]$ converging to x_*, the sequence $(f(x_n))$ converges to $f(x_*)$. And continuity on $[a, b]$ means continuity at each point $x_* \in [a, b]$.

2. If $f(a)f(b) = 0$ then one of a or b must be a root, and we are done.

3. Set $x_k = \frac{a+b}{2}$.

4. If $f(x_k) = 0$ then x_k is a root and we are done.

5. If $f(x_k) \neq 0$ then either:

   - $f(x_k)f(a) > 0$, in which case $\alpha \in (x_k, b)$, so redefine $a = x_k$, increment k to $k+1$ and return to step 3.
   - $f(x_k)f(a) < 0$, in which case $\alpha \in (a, x_k)$, so redefine $b = x_k$, increment k to $k+1$ and return to step 3.

Using this type of division of the interval, or bisection, we create a sequence $x_0, x_1, x_2, \ldots$ which is guaranteed to converge to a root α (when one exists). In fact it is easy to quantify the rate of convergence, as we can write down an *a priori* error bound.

Theorem 3.2.3. *Let f satisfy the assumptions of Theorem 3.2.2. Then the iterates $x_0, x_1, x_2, \ldots$ of the bisection method converge to a root $\alpha \in [a, b]$ and for all k it holds that*

$$|x_k - \alpha| \leq \frac{b-a}{2^{k+1}}.$$

Proof. The error estimate (and hence convergence) is easily proven by observing that for each iteration the length of the interval is divided by two and that by construction a root α is always in the current interval. □

This is a major advantage of the bisection method: once an interval such that $f(a)f(b) \leq 0$ has been found, the algorithm is guaranteed to converge to a root. Another advantage is that the algorithm only uses values of the function f, and does not require (potentially costly) evaluations of derivatives of f.

The main disadvantages are that convergence is slow, and that not all roots of f can be found. (A root α can only be found using bisection if f changes sign on the interval $(\alpha - \delta, \alpha + \delta)$ for all δ sufficiently small). Also, the method does not generalise in a practical way to higher space dimensions.

Example 3.2.4. *Suppose we wish to find the root of the function $f(x) = sin(2x) - 1 + x$. Figure 3.2 shows a plot of the function f in the interval $[-3, 3]$ obtained using the following Matlab commands:*

```
>> f = @(x)sin(2*x)-1+x;
>> x=[-3:0.1:3];
>> plot(x,f(x))
>> grid on
```

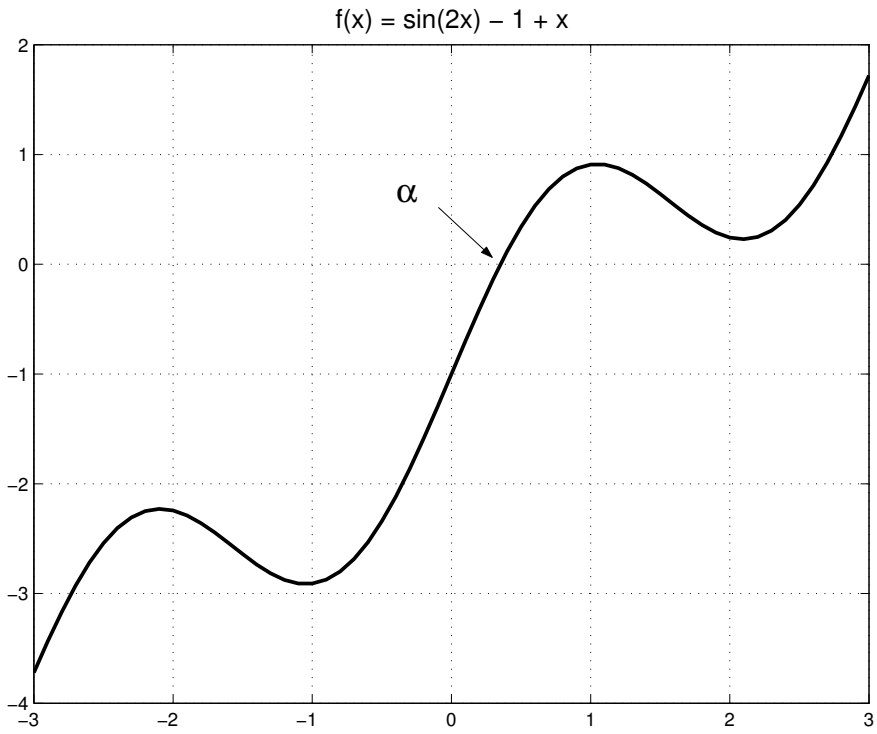


Figure 3.2: *Graph of f from Example 3.2.4.*

Applying the method of bisection in the interval $[-1, 1]$ *with a tolerance of* 10^{-8}

```
>> [zero,res,niter]=bisection(f,-1,1,1e-8,1000)
```

gives the value $\alpha = 0.35228846$ *after 27 iterations.*

3.3 Fixed point methods

A general method to find the roots of a nonlinear equation $f(x) = 0$ is obtained by rewriting it in the form $\phi(x) = x$, for some function $\phi : [a, b] \to \mathbb{R}$ satisfying the "consistency" condition

$$\phi(\alpha) = \alpha \qquad \text{if and only if} \qquad f(\alpha) = 0. \tag{3.4}$$

The point α is called a *fixed point* of the function ϕ. It follows that to find the zeros of f we can determine the *fixed points* of ϕ. Note that one can always find a ϕ satisfying (3.4), because in particular one can take $\phi(x) = x - \beta f(x)$ for any $0 \neq \beta \in \mathbb{R}$. But other choices of ϕ may prove advantageous.

Example 3.3.1 (3.2.4, continued). *Once again consider* $f(x) = \sin(2x) - 1 + x = 0$. *We*

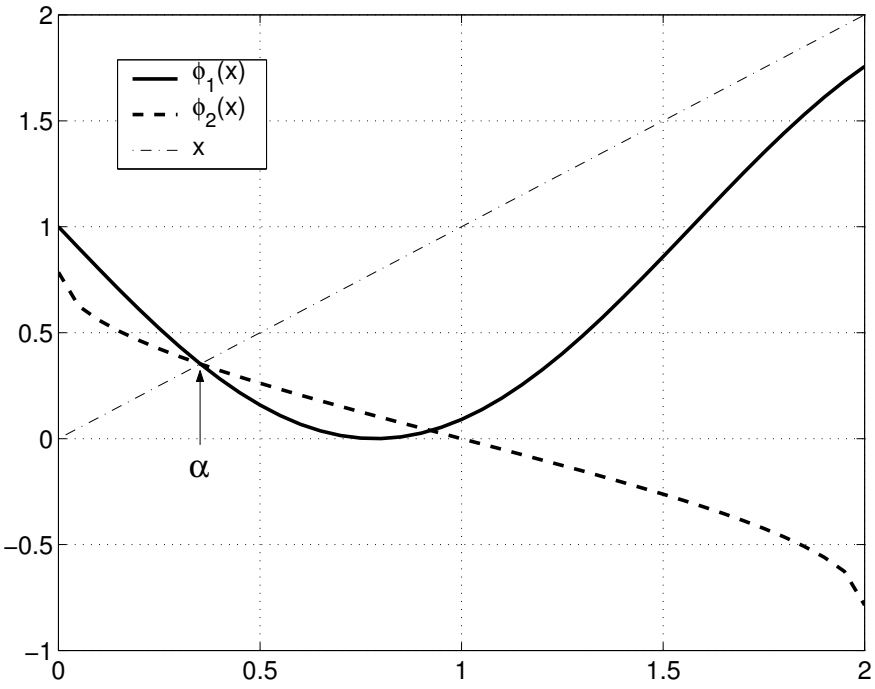


Figure 3.3: *Functions ϕ_1 and ϕ_2 from Example 3.3.1.*

can recast the problem in fixed point form in (at least) two different ways:

$$x = \phi_1(x) = 1 - \sin(2x)$$
$$x = \phi_2(x) = \frac{1}{2}\arcsin(1-x), \qquad 0 \le x \le 2.$$

Figure 3.3 shows the functions ϕ_1, ϕ_2 as well as the line $y = x$. Note that ϕ_1 is of the simple form $\phi(x) = x - \beta f(x)$ mentioned above, with $\beta = 1$. But ϕ_2 is not of this form.

The following theoretical result about fixed points is a simple consequence of Theorem 3.2.2. For an illustration of the setting of the theorem consider the interval $[0, 1]$ in Figure 3.3.

Theorem 3.3.2 (Brouwer's fixed point theorem)**.** *Suppose that $\phi : [a, b] \to \mathbb{R}$ is continuous on $[a, b]$ and satisfies $\phi(x) \in [a, b]$ for all $x \in [a, b]$. Then there exists $\alpha \in [a, b]$ such that $\alpha = \phi(\alpha)$, i.e. α is a fixed point of ϕ.*

Proof. Let $f(x) = x - \phi(x)$. Then, since $\phi(x) \in [a, b]$ for $x \in [a, b]$, we have $f(a) = a - \phi(a) \le 0$ and $f(b) = b - \phi(b) \ge 0$. Hence $f(a)f(b) \le 0$, and then the existence of $\alpha \in [a, b]$ such that $f(\alpha) = 0$ and consequently $\alpha = \phi(\alpha)$ follows from Theorem 3.2.2. □

The importance of the Brouwer fixed point theorem is that it extends to higher dimensions (see Theorem 3.7.3 below). But the proof in the higher-dimensional case uses tools from topology that are beyond the scope of this course.

The fixed point formulation immediately suggests an iterative method for finding a fixed point, defined by the sequence

$$x_{k+1} = \phi(x_k), \quad k \geq 0. \tag{3.5}$$

If the resulting sequence converges to some α and ϕ is continuous at α then α is a fixed point. We formalise this in a lemma.

Lemma 3.3.3. *Let $\phi : \mathbb{R} \to \mathbb{R}$, let $x_0 \in \mathbb{R}$, and define a sequence $(x_k)_{k=0}^{\infty}$ by*

$$x_{k+1} = \phi(x_k), \quad k \geq 0.$$

If $x_k \to \alpha$ for $k \to \infty$ and ϕ is continuous at α then $\alpha = \phi(\alpha)$.

Proof. Using the definition of the sequence we may write

$$\alpha = \lim_{k\to\infty} x_{k+1} = \lim_{k\to\infty} \phi(x_k) = \phi\left(\lim_{k\to\infty} x_k\right) = \phi(\alpha).$$

Here the assumption that ϕ is continuous at α is what allows us to move the limit inside the argument of the function in the third equality. □

But under what conditions does the sequence obtained using (3.5) converge? This is not true for every function ϕ, but requires some additional assumptions.

Definition 3.3.4. *We say that the function $\phi : [a, b] \to \mathbb{R}$ is a strict contraction on the interval $[a, b]$ if there exists a constant $0 < \Lambda < 1$ such that*

$$|\phi(x) - \phi(x')| \leq \Lambda |x - x'|, \quad \forall x, x' \in [a, b]. \tag{3.6}$$

Note that any strict contraction is necessarily a Lipschitz continuous function on $[a, b]$. A useful sufficient condition for a function to be a strict contraction is provided by the following lemma.

Lemma 3.3.5. *Let $\phi : [a, b] \to \mathbb{R}$ be differentiable on $[a, b]$, and suppose there exists a constant $0 < \Lambda < 1$ such that $|\phi'(x)| \leq \Lambda$ for all $x \in [a, b]$. Then ϕ is a strict contraction on $[a, b]$.*

Proof. The mean value theorem implies that for any $x, x' \in [a, b]$ there exists ξ between x and x' such that

$$\phi(x) - \phi(x') = \phi'(\xi)(x - x').$$

The bound (3.6) then follows by taking absolute values and noting that $|\phi'(\xi)| \leq \Lambda < 1$. □

Theorem 3.3.6 (Contraction mapping theorem). *Let $\phi : [a,b] \to \mathbb{R}$ be a strict contraction on $[a,b]$ with contraction constant $0 < \Lambda < 1$. (In particular, by Lemma 3.3.5 this holds if ϕ is differentiable on $[a,b]$ and there exists a constant $0 < \Lambda < 1$ such that $|\phi'(x)| \leq \Lambda$ for all $x \in [a,b]$.) Suppose also that $\phi(x) \in [a,b]$ for all $x \in [a,b]$. Then*

(i) ϕ admits a unique fixed point α in $[a,b]$;

(ii) $\forall x_0 \in [a,b]$ the sequence (x_k) defined by $x_{k+1} = \phi(x_k)$, $k \geq 0$, converges to α as $k \to \infty$;

(iii) The iterates satisfy the error estimate

$$|x_{k+1} - \alpha| \leq \Lambda |x_k - \alpha|, \quad \forall k \geq 0. \tag{3.7}$$

Furthermore, if ϕ is continuously differentiable [a] on $[a,b]$ and $x_k \neq \alpha$ for any k then

$$\lim_{k\to\infty} \frac{x_{k+1} - \alpha}{x_k - \alpha} = \phi'(\alpha).$$

[a] A function is continuously differentiable on $[a,b]$ if f is differentiable on $[a,b]$ and the derivative f' is a continuous function.

Proof. As noted above, the assumption that ϕ is a strict contraction implies that ϕ is continuous. Hence the existence of a fixed point α follows from the Brouwer fixed point theorem (Theorem 3.3.2). (Note that one can also prove the existence of a fixed point directly from the strict contraction property, by showing that the sequence (x_k) is Cauchy. We leave this as an exercise.) To show uniqueness, suppose that $\alpha_1 \neq \alpha_2$ are two distinct fixed points. Then since ϕ is a strict contraction there exists $0 < \Lambda < 1$ such that

$$|\alpha_1 - \alpha_2| = |\phi(\alpha_1) - \phi(\alpha_2)| \leq \Lambda |\alpha_1 - \alpha_2| < |\alpha_1 - \alpha_2|.$$

But this is a contradiction, and hence the fixed point must be unique.

The convergence result and error estimate follow by similar arguments. Indeed, for each $k \geq 0$,

$$|x_{k+1} - \alpha| = |\phi(x_k) - \phi(\alpha)| \leq \Lambda |x_k - \alpha|,$$

and by induction it follows that $|x_k - \alpha| \leq \Lambda^k |x_0 - \alpha|$, so that $x_k \to \alpha$ as $k \to \infty$.

Finally, in the case when ϕ is continuously differentiable, the mean value theorem implies that, for each $k \geq 0$,

$$x_{k+1} - \alpha = \phi(x_k) - \phi(\alpha) = \phi'(\xi_k)(x_k - \alpha),$$

for some ξ_k between α and x_k. By the continuity of ϕ' and the convergence of x_k to α it follows that

$$\frac{x_{k+1} - \alpha}{x_k - \alpha} \to \phi'(\alpha) \qquad \text{as } k \to \infty.$$

□

Remark 3.3.7. *The condition "$\phi(x) \in [a,b]$ for all $x \in [a,b]$" in the hypotheses of the CMT is crucial and cannot in general be omitted. So don't forget to check it!*

The CMT provides sufficient conditions under which the fixed point iteration converges linearly, in the sense of Definition 2.2.1, for *any* choice of initial guess $x_0 \in [a,b]$. But it requires "global" assumptions on the behaviour of ϕ on the whole interval $[a,b]$. The following theorem provides a "local" convergence result which only requires a "local" assumption about ϕ.

Theorem 3.3.8 (linear convergence in a neighbourhood of α)**.** *Let $\phi : [a,b] \to \mathbb{R}$ be continuously differentiable and let $\alpha \in (a,b)$ be a fixed point of ϕ such that $|\phi'(\alpha)| < 1$. Then there exists $\delta > 0$ such that for all initial guesses $x_0 \in [\alpha - \delta, \alpha + \delta]$ the sequence (x_k) defined by $x_{k+1} = \phi(x_k)$ converges linearly to α as $k \to \infty$, with*

$$\lim_{k\to\infty} \frac{x_{k+1} - \alpha}{x_k - \alpha} = \phi'(\alpha).$$

Proof. Since $|\phi'(\alpha)| < 1$ and ϕ' is continuous, there exists $\delta > 0$ such that $|\phi'(x)| \leq \Lambda$ for all $x \in [\alpha - \delta, \alpha + \delta]$, where $0 < \Lambda = (1 + |\phi'(\alpha)|)/2 < 1$. Furthermore, ϕ maps $[\alpha - \delta, \alpha + \delta]$ to $[\alpha - \delta, \alpha + \delta]$ because α is a fixed point of ϕ (exercise: use the MVT to check this!). Applying Theorem 3.3.6 on the interval $[\alpha - \delta, \alpha + \delta]$ we conclude that the sequence x_k converges to α as $k \to \infty$ for any initial guess $x_0 \in [\alpha - \delta, \alpha + \delta]$.

□

In Figures 3.4 and 3.5 we show in some examples how the value of $|\phi'(\alpha)|$ influences the convergence of the fixed point iteration.

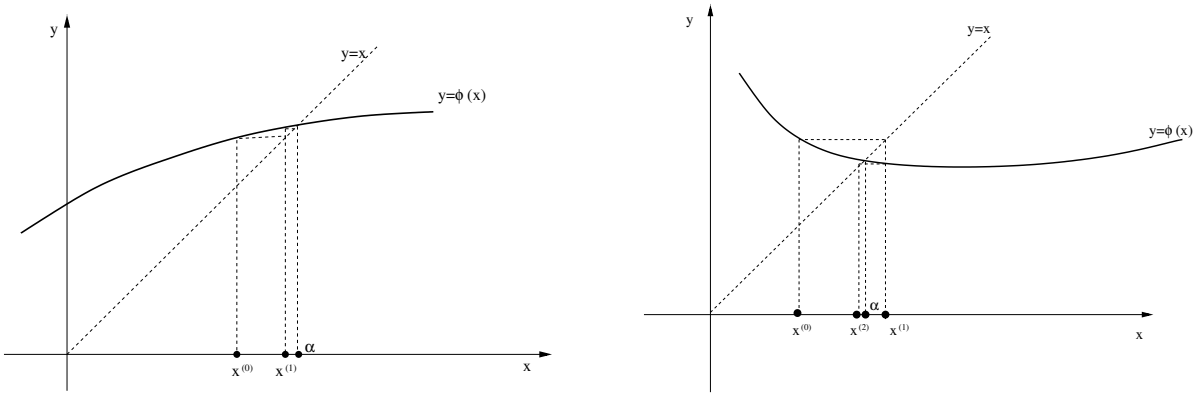


Figure 3.4: *Fixed point iterations; convergent case. Left $0 < \phi'(\alpha) < 1$, right $-1 < \phi'(\alpha) < 0$.*

Example 3.3.9 (Examples 3.2.4 and 3.3.1 continued). *We apply the fixed point method to ϕ_1 and ϕ_2 starting from the initial value $x_0 = 0.7$.*

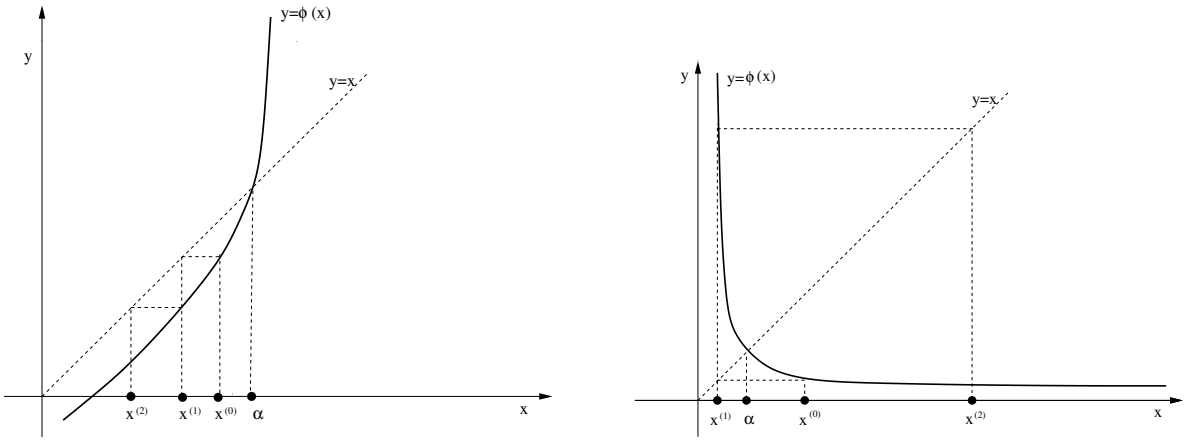


Figure 3.5: *Fixed point iterations; divergent case. Left* $\phi'(\alpha) > 1$*, right* $\phi'(\alpha) < -1$.

```
>>  phi1=@(x)1-sin(2*x);
>>  phi2=@(x).5*asin(1-x);
>>  [fixp,res,niter]=fixpoint(phi1,0.7,1e-8,1000)
>>  [fixp,res,niter]=fixpoint(phi2,0.7,1e-8,1000)
```

We find that the first method does not converge whereas the second converges to $\alpha = 0.35228846$ *in* 44 *iterations. Computing the derivative of the iteration functions at the root* α *we see that* $\phi_1'(\alpha) = -1.5237713$ *and* $\phi_2'(\alpha) = -0.65626645$*. So our numerical results are consistent with the result of Theorem 3.3.8.*

In light of Theorem 3.3.8, one might imagine that if $\phi'(\alpha) = 0$ the convergence should be super-linear (i.e. faster than linear). Under suitable assumptions on ϕ this is indeed the case.

Theorem 3.3.10 (quadratic convergence in a neighbourhood of α)**.** *Let* $\phi : [a, b] \to \mathbb{R}$ *be twice continuously differentiable and let* $\alpha \in (a, b)$ *be a fixed point of* ϕ *satisfying* $\phi'(\alpha) = 0$*. Then there exists* $\delta > 0$ *such that for all initial guesses* $x_0 \in [\alpha - \delta, \alpha + \delta]$ *the sequence* (x_k) *defined by* $x_{k+1} = \phi(x_k)$ *converges quadratically to* α *as* $k \to \infty$*, with*

$$\lim_{k\to\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^2} = \frac{\phi''(\alpha)}{2}.$$

Proof. A Taylor expansion of $\phi(x_k)$ around $x = \alpha$ gives

$$x_{k+1} - \alpha = \phi(x_k) - \phi(\alpha) = \phi'(\alpha)(x_k - \alpha) + \frac{\phi''(\xi_k)}{2}(x_k - \alpha)^2,$$

for some point ξ_k between x_k and α. Since $\phi'(\alpha) = 0$ and ϕ'' is continuous it follows that

$$\lim_{k\to\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^2} = \lim_{k\to\infty} \frac{\phi''(\xi_k)}{2} = \frac{\phi''(\alpha)}{2}.$$

□

We study an example of a quadratically convergent fixed point method in the next section.

3.4 Newton's method

Let us return to our original problem of finding a zero of a given function $f : \mathbb{R} \to \mathbb{R}$, i.e. a solution α of the equation $f(\alpha) = 0$. Suppose that f is continuously differentiable, and that we have an initial guess x_k.

Then by Taylor's theorem we know that

$$0 = f(\alpha) = f(x_k) + f'(x_k)(\alpha - x_k) + R_1(\alpha; x_k)$$

with a remainder term $R_1(\alpha; x_k) = o(|\alpha - x_k|)$ as $|\alpha - x_k| \to 0$. If $f'(x_k) \neq 0$ then we can re-arrange for α as

$$\alpha = x_k - \frac{f(x_k)}{f'(x_k)} - \frac{R_1(\alpha; x_k)}{f'(x_k)} \approx x_k - \frac{f(x_k)}{f'(x_k)}, \tag{3.8}$$

where the approximation on the right-hand side is obtained by dropping the remainder term. This suggests the iterative method known as Newton's method:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \qquad k = 0, 1, 2, \ldots. \tag{3.9}$$

Another way of interpreting Newton's method is to notice that x_{k+1} is the unique root of the local linear approximation to f at the point x_k (i.e. the intersection of the tangent to the graph of f at the point x_k with the x axis) - see Figure 3.6. This tangent is the straight line defined by the function

$$r(x) = f'(x_k)(x - x_k) + f(x_k),$$

and solving the equation $r(x_{k+1}) = 0$ gives the Newton step (3.9).

Newton's method can be written as a fixed point method $x_{k+1} = \phi(x_k)$ using the function

$$\phi(x) = x - \frac{f(x)}{f'(x)}.$$

Therefore we can analyse the convergence of Newton's method using the theory of fixed point iterations. In particular, provided that f is a smooth enough function, then

$$\phi'(x) = 1 - \frac{[f'(x)]^2 - f(x)f(x)}{[f'(x)]^2} = \frac{f(x)f''(x)}{[f'(x)]^2}.$$

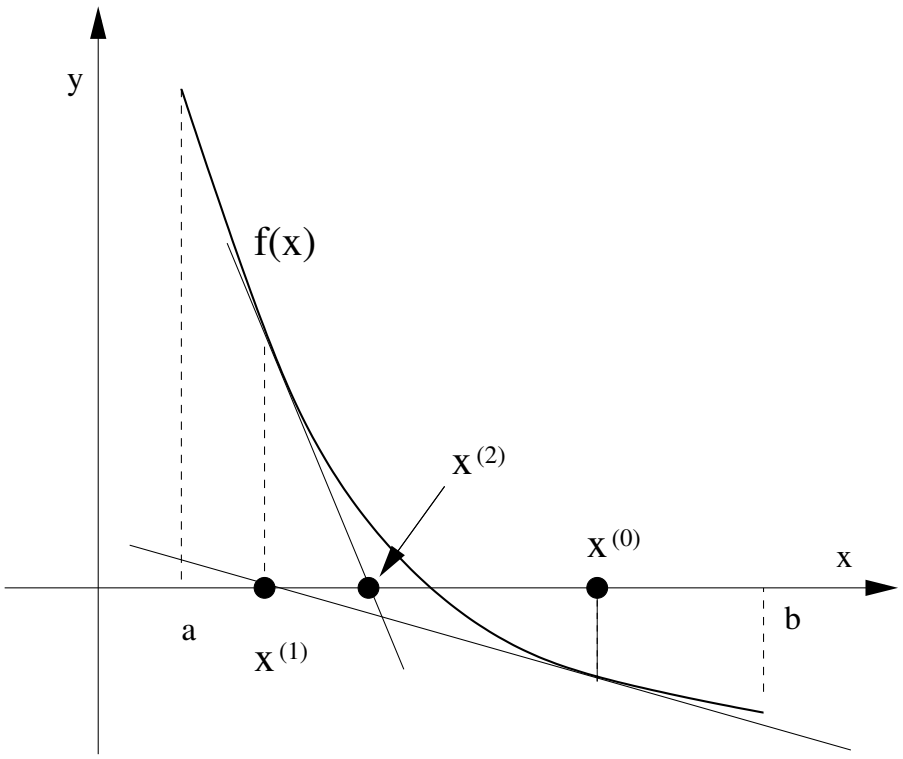


Figure 3.6: *Newton's method. Starting from x_0, the sequence (x_k) converges to the root of f.*

In particular, if α is a root of f with $f(\alpha) = 0$ and $f'(\alpha) \neq 0$ then

$$\phi'(\alpha) = 0.$$

Hence, if f is smooth enough, then Theorem 3.3.10 shows that the iterations converge for a sufficiently close initial guess x_0 and the convergence is quadratic, and if $x_k \neq \alpha$ for any k then

$$\lim_{k\to\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^2} = \frac{f''(\alpha)}{2f'(\alpha)}. \tag{3.10}$$

In particular, the application of Theorem 3.3.10 would require the hypothesis that f is three times continuously differentiable, so that ϕ be twice continuously differentiable. However, the smoothness assumption on f can be relaxed as shown in the following result which is based on a more direct analysis.

Theorem 3.4.1. *Let $f : [a,b] \to \mathbb{R}$ be twice continuously differentiable on $[a,b]$ and let $\alpha \in (a,b)$ satisfy $f(\alpha) = 0$ and $f'(\alpha) \neq 0$. Then there exists a $\delta > 0$ such that for all initial guesses $x_0 \in [\alpha - \delta, \alpha + \delta]$ the Newton iteration (3.9) converges to α quadratically, and if $x_k \neq \alpha$ for any k then*

$$\lim_{k\to\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^2} = \frac{f''(\alpha)}{2f'(\alpha)}. \tag{3.11}$$

Proof. Since f' is continuous on $[a,b]$ and since $f'(\alpha) \neq 0$, there is a $\rho > 0$ and an $M > 0$ such that $[\alpha - \rho, \alpha + \rho] \subset [a,b]$ and

$$|f'(x)| \geq M \quad \forall x \in [\alpha - \rho, \alpha + \rho].$$

Therefore $f'(x_k) \neq 0$ and x_{k+1} is well-defined by Newton's method whenever $x_k \in [\alpha - \rho, \alpha + \rho]$. Since f'' is continuous on $[a,b]$, there exists $C \geq 0$ such that $|f''(x)| \leq C$ for all $x \in [a,b]$. We now define the interval I_δ and constant δ by

$$I_\delta = [\alpha - \delta, \alpha + \delta], \quad \delta = \min\left(\rho, \frac{M}{C}\right).$$

Our goal is to show by induction that $x_k \in I_\delta$ for all $k \in \mathbb{N}_0$ whenever $x_0 \in I_\delta$. To this end, suppose that $x_k \in I_\delta$ for some k, and note then that since $\delta \leq \rho$ we have $|f'(x_k)| \geq M > 0$. Hence x_{k+1} is well-defined. Moreover, recalling equation (3.8) we see that

$$\alpha = \underbrace{x_k - \frac{f(x_k)}{f'(x_k)}}_{=x_{k+1}} - \frac{R_1(\alpha; x_k)}{f'(x_k)},$$

where Taylor's Theorem for the twice differentiable function f shows that $R_1(\alpha; x_k) = \frac{1}{2}f''(\xi_k)(\alpha - x_k)^2$ for some ξ_k between x_k and α. Therefore we get

$$x_{k+1} - \alpha = \frac{R_1(\alpha; x_k)}{f'(x_k)} = \frac{f''(\xi_k)}{2f'(x_k)}(x_k - \alpha)^2 \tag{3.12}$$

Hence using the bounds above for f' and f'' and using $|x_k - \alpha| \leq \delta \leq \frac{M}{C}$ we get

$$|x_{k+1} - \alpha| \leq \frac{C}{2M}|x_k - \alpha|^2 \leq \frac{1}{2}|x_k - \alpha| \leq \frac{1}{2}\delta, \tag{3.13}$$

which implies that $x_{k+1} \in I_\delta$. This completes the inductive step and thus shows that $x_k \in I_\delta$ for all $k \in \mathbb{N}$ whenever $x_0 \in I_\delta$. Furthermore the second inequality in (3.13) implies the (linear) convergence of the iterates x_k to α as $k \to \infty$, and consequently also the convergence of ξ_k to α as $k \to \infty$ since ξ_k is between x_k and α for all $k \in \mathbb{N}$. Finally the first inequality in (3.13) implies that the convergence is furthermore quadratric, and if we suppose that $x_k \neq \alpha$ for any k then the continuity of the first and second derivatives implies that

$$\lim_{k\to\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^2} = \lim_{k\to\infty} \frac{f''(\xi_k)}{2f'(x_k)} = \frac{f''(\alpha)}{2f'(\alpha)},$$

which proves (3.11). □

Remark 3.4.2. *If $f'(\alpha) = 0$, then the convergence of Newton's method is only linear. In order to recover quadratic convergence we may consider the modified Newton's method:*

$$x_{k+1} = x_k - p\frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2 \ldots, \tag{3.14}$$

where $p \in \mathbb{N}$ is the smallest integer such that $f^{(p)}(\alpha) \neq 0$.

3.5 The secant method

A major attraction of Newton's method is its super-linear (quadratic) convergence. However, one disadvantage of the method is that it requires knowledge of the derivative of $f(x)$. In some applications this might be difficult or costly to compute. One way of obtaining a derivative-free method is to replace $f'(x)$ by a suitable difference quotient. A natural choice is

$$f'(x_k) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}},$$

which leads to the iteration scheme (note that both x_0 and x_1 must now be supplied)

$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}, \quad k = 1, 2 \ldots. \tag{3.15}$$

This approximate Newton method is known as the *secant method*. For sufficiently smooth f it still exhibits super-linear convergence, albeit not as fast as Newton. One can prove that the order of convergence of the secant method is $p = (1 + \sqrt{5})/2 \approx 1.63$.

3.6 The chord method

An even simpler method is obtained by approximating $f'(x_k)$ by some fixed number $q \in \mathbb{R}$ (the same for each iteration). This leads to the iteration

$$x_{k+1} = x_k - \frac{f(x_k)}{q}, \quad k = 0, 1, 2 \ldots, \tag{3.16}$$

which is known as the *chord method* (or sometimes the *relaxation method*). One may for instance take $q = f'(x_0)$, or, if we are looking for a root in the interval $[a, b]$, $q = (f(b) - f(a))/(b - a)$.

The chord method is a fixed point method for

$$\phi(x) = x - \frac{f(x)}{q}.$$

It follows that $\phi'(x) = 1 - \frac{f'(x)}{q}$. Hence by Theorem 3.3.8, the method converges for initial guesses sufficiently close to the root, provided that

$$|1 - \frac{f'(\alpha)}{q}| < 1.$$

But in general the convergence is only linear.

Example 3.6.1. *(3.2.4 continued) We apply the chord method and Newton's method to find the root of the function f.*
Chord method in the interval $[-1, 1]$, using $q = (f(b) - f(a))/(b - a)$ with $a = -1$, $b = 1$, and starting from $x_0 = 0.7$:

```
>>  [zero,res,niter]=chord(f,-1,1,0.7,1e-8,1000)
```

the script returns the root after 15 iterations.
Newton's method starting from the same x_0 :

```
>>  df = @(x)2*cos(2*x)+1;
>>  [zero,res,niter]=newton(f,df,0.7,1e-8,1000)
```

The zero is found after only 5 iterations. In Figure 3.7 the graph of the error $|x_k - \alpha|$ is shown as a function of the number of iterations k for the bisection method, the fixed point method ϕ_2, the chord method and Newton's method. We observe the linear convergence of the fixed point method and the chord method. We also observe that the error of Newton's method reduces much faster, reflecting the quadratic convergence. Note that the convergence of the bisection method is non-monotone.

3.7 Systems of nonlinear equations

[2] We now consider higher-dimensional generalisations. Let $X \subset \mathbb{R}^n$ and $\boldsymbol{f} : X \to \mathbb{R}^n$, and consider the problem: find $\boldsymbol{x} \in X$ such that

$$\boldsymbol{f}(\boldsymbol{x}) = \mathbf{0}. \tag{3.17}$$

This problem is equivalent to the system of equations

$$f_i(x_1, \ldots, x_n) = 0 \text{ for } i = 1, \ldots, n,$$

where f_i and x_i are the components of $\boldsymbol{f}$ and $\boldsymbol{x}$ respectively.

[2] The material in this section uses results from §4.2 on matrix norms. So you might want to postpone reading it until you have read §4.2.

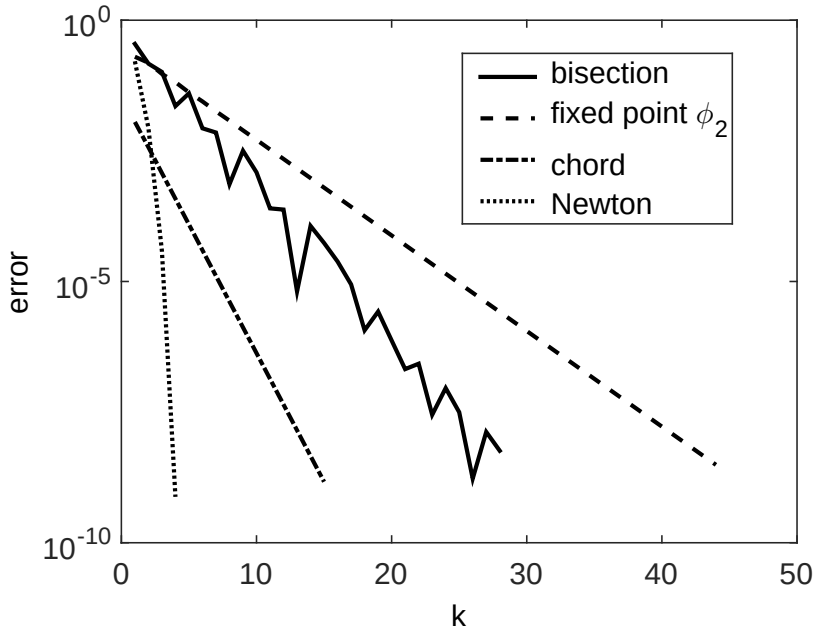


Figure 3.7: *Error as a function of the number of iterations for the bisection method, the fixed point method ϕ_2, the chord method and Newton's method - see Example 3.6.1. Note that the y-axis has logarithmic scale.*

Example 3.7.1.

$$\begin{cases} f_1(x_1, x_2) = x_1^2 + x_2^2 - 1 = 0 \\ f_2(x_1, x_2) = 5x_1^2 + 21x_2^2 - 9 = 0 \end{cases} \tag{3.18}$$

The first equation defines a circle in $\mathbb{R}^2$ and the second equation defines an ellipse in $\mathbb{R}^2$. The solutions of the system are the intersection points of the circle and ellipse (verify by drawing the two curves), namely $(-\frac{\sqrt{3}}{2}, \frac{1}{2})$, $(\frac{\sqrt{3}}{2}, \frac{1}{2})$, $(-\frac{\sqrt{3}}{2}, -\frac{1}{2})$ and $(\frac{\sqrt{3}}{2}, -\frac{1}{2})$.

3.7.1 Simultaneous iteration

The extension of the fixed point iteration method to the case of higher dimensions is known as *simultaneous iteration*. For the problem (3.17), the first step is to find an equivalent problem

$$\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{x}), \quad \boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^n, \tag{3.19}$$

such that

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0} \quad \Leftrightarrow \quad \boldsymbol{x} = \boldsymbol{g}(\boldsymbol{x}).$$

Example 3.7.2. *In the spirit of the chord method one may always take*

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x} - \beta \boldsymbol{f}(\boldsymbol{x}),$$

for some $0 \neq \beta \in \mathbb{R}$.

The existence of a solution for the fixed point formulation can be shown using a Brouwer fixed point theorem.

Theorem 3.7.3 (Brouwer's fixed point theorem in multiple dimensions)**.** *Assume that X is a non-empty, closed, bounded and convex subset of $\mathbb{R}^n$. Suppose that $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^n$ is a continuous mapping such that $\boldsymbol{g}(X) \subset X$. Then there exists $\boldsymbol{\alpha} \in X$ such that $\boldsymbol{\alpha} = \boldsymbol{g}(\boldsymbol{\alpha})$.*

An iterative procedure for the construction of an approximating sequence can be defined similarly as in the one-dimensional case, with the difference that in this case we iterate over all the unknowns simultaneously (hence the name).

Definition 3.7.4 (Simultaneous iteration)**.** *Let $\boldsymbol{g}$ satisfy the hypotheses of Theorem 3.7.3. Given $\boldsymbol{x}^0 \in X \subset \mathbb{R}^n$, define $\boldsymbol{x}^k$, $k = 1, 2, \ldots$, recursively by*

$$\boldsymbol{x}^k = \boldsymbol{g}(\boldsymbol{x}^{k-1}).$$

The convergence of simultaneous iterations can be proved using the contraction mapping theorem, which, when it applies, also gives the existence of a unique solution to the fixed point formulation (3.19).

Theorem 3.7.5 (Contraction mapping theorem in multiple dimensions)**.** *Suppose that X is a closed subset of $\mathbb{R}^n$ and that $\boldsymbol{g} : X \to \mathbb{R}^n$ satisfies*

- *$\boldsymbol{g}(X) \subset X$;*
- *$\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y})\| \leq \Lambda \|\boldsymbol{x} - \boldsymbol{y}\|$, for some $0 < \Lambda < 1$ and some norm $\|\cdot\|$ on $\mathbb{R}^n$. (contraction)*

Then there exists a unique $\boldsymbol{\alpha} \in X$ such that $\boldsymbol{\alpha} = \boldsymbol{g}(\boldsymbol{\alpha})$. Moreover, the sequence $(\boldsymbol{x}^k)$ defined by simultaneous iteration converges to $\boldsymbol{\alpha}$ as $k \to \infty$.

Proof. Follows the one-dimensional case. □

Unfortunately, checking that the function $\boldsymbol{g}(\boldsymbol{x})$ is a contraction may not be that easy in general. As in the one-dimensional case, things are easier when the function $\boldsymbol{g}$ is continuously differentiable in a neighbourhood of the root. In the multi-dimensional case we need to consider the behaviour of the *Jacobian matrix* of $\boldsymbol{g}(\boldsymbol{x})$, evaluated at the root $\boldsymbol{\alpha}$.

Definition 3.7.6 (Jacobian matrix)**.** *Let $\boldsymbol{g} = (g_1, g_2, \ldots, g_n)^T : \mathbb{R}^n \to \mathbb{R}^n$ be a function such that the partial derivatives*

$$\frac{\partial g_i}{\partial x_j}, \quad j = 1, \ldots, n,$$

of g_i exist at the point $\boldsymbol{x} \in \mathbb{R}^n$ for each $i = 1, \ldots, n$. The Jacobian matrix $J_{\boldsymbol{g}}(\boldsymbol{x})$ of $\boldsymbol{g}$ at $\boldsymbol{x}$ is then defined by the $n \times n$ matrix with elements

$$[J_{\boldsymbol{g}}(\boldsymbol{x})]_{ij} = \frac{\partial g_i}{\partial x_j}(\boldsymbol{x}), \quad i, j = 1, \ldots, n,$$

i.e.

$$J_{\boldsymbol{g}}(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial g_1}{\partial x_1}(\boldsymbol{x}) & \ldots & \frac{\partial g_1}{\partial x_n}(\boldsymbol{x}) \\ \vdots & & \\ \frac{\partial g_n}{\partial x_1}(\boldsymbol{x}) & \ldots & \frac{\partial g_n}{\partial x_n}(\boldsymbol{x}) \end{bmatrix}.$$

Other common notations for the Jacobian matrix include $D\boldsymbol{g}(\boldsymbol{x})$ and $\partial \boldsymbol{g}/\partial \boldsymbol{x}$.

We now state and prove a technical lemma. Here the assumption that X is "convex" simply means that if $\boldsymbol{x}, \boldsymbol{y} \in X$ then $t\boldsymbol{x} + (1-t)\boldsymbol{y} \in X$ for all $t \in [0, 1]$, i.e. the straight line connecting $\boldsymbol{x}$ and $\boldsymbol{y}$ also lies in X.

Lemma 3.7.7. *Assume that the partial derivatives of $\boldsymbol{g}$ are well defined and continuous on some convex set $X \subset \mathbb{R}^n$. Then, for $\boldsymbol{x}, \boldsymbol{y} \in X$,*

$$\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y})\|_\infty \leq \max_{t \in [0,1]} \|J_{\boldsymbol{g}}(t\boldsymbol{x} + (1-t)\boldsymbol{y})\|_\infty \|\boldsymbol{x} - \boldsymbol{y}\|_\infty.$$

Proof. Define $\varphi_i(t) = g_i(t\boldsymbol{x} + (1-t)\boldsymbol{y})$, $i = 1, \ldots, n$. By the mean value theorem there holds, for some $\eta_i \in [0, 1]$,

$$g_i(\boldsymbol{x}) - g_i(\boldsymbol{y}) = \varphi_i(1) - \varphi_i(0) = \varphi_i'(\eta_i)(1 - 0) = \varphi_i'(\eta_i).$$

By differentiating φ_i we obtain

$$\varphi_i'(\eta_i) = \sum_{j=1}^{n} \frac{\partial g_i}{\partial x_j}(\eta_i \boldsymbol{x} + (1 - \eta_i)\boldsymbol{y}) \times (x_j - y_j).$$

It follows that

$$|g_i(\boldsymbol{x}) - g_i(\boldsymbol{y})| \leq \|\boldsymbol{x} - \boldsymbol{y}\|_\infty \sum_{j=1}^{n} \left| \frac{\partial g_i}{\partial x_j}(\eta_i \boldsymbol{x} + (1 - \eta_i)\boldsymbol{y}) \right|.$$

Taking the max over i in this expression, we get

$$\|g(\boldsymbol{x}) - g(\boldsymbol{y})\|_\infty \leq \|\boldsymbol{x} - \boldsymbol{y}\|_\infty \max_{i \in \{1,\ldots,n\}} \sum_{j=1}^{n} \left| \frac{\partial g_i}{\partial x_j}(\eta_i \boldsymbol{x} + (1 - \eta_i)\boldsymbol{y}) \right| \tag{3.20}$$

$$\leq \|\boldsymbol{x} - \boldsymbol{y}\|_\infty \max_{i \in \{1,\ldots,n\}} \max_{t \in [0,1]} \sum_{j=1}^{n} \left| \frac{\partial g_i}{\partial x_j}(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \right| \tag{3.21}$$

$$= \|\boldsymbol{x} - \boldsymbol{y}\|_\infty \max_{t \in [0,1]} \max_{i \in \{1,\ldots,n\}} \sum_{j=1}^{n} \left| \frac{\partial g_i}{\partial x_j}(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \right| \tag{3.22}$$

$$= \|\boldsymbol{x} - \boldsymbol{y}\|_\infty \max_{t \in [0,1]} \|J_{\boldsymbol{g}}(t\boldsymbol{x} + (1-t)\boldsymbol{y})\|_\infty, \tag{3.23}$$

where in the final equality we used formula (4.5) for the infinity norm of a matrix. (The continuity of the partial derivatives ensures that they are all bounded on the segment between $\boldsymbol{x}$ and $\boldsymbol{y}$.) □

We can now state a corollary of the contraction mapping theorem, that says that if our initial guess is close enough to the root, and the Jacobian of the iteration function is small enough in a neighbourhood of the root, then the simultaneous iterations will converge. This result is a higher-dimensional analogue of Theorem 3.3.8.

Theorem 3.7.8. *Let $\boldsymbol{g} = (g_1, g_2, \ldots, g_n)^T : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable on some open set $X \subset \mathbb{R}^n$, i.e. the partial derivatives of g are defined and continuous in X. Let $\boldsymbol{\alpha} \in X$ be such that $\boldsymbol{\alpha} = \boldsymbol{g}(\boldsymbol{\alpha})$. Suppose further that*

$$\|J_{\boldsymbol{g}}(\boldsymbol{\alpha})\|_\infty < 1.$$

Then there exists $\delta > 0$ such that for any initial guess $\boldsymbol{x}^0 \in \bar{B}_\delta(\boldsymbol{\alpha}) = \{\boldsymbol{x} \in X : \|\boldsymbol{x} - \boldsymbol{\alpha}\|_\infty \leq \delta\}$ the sequence $(\boldsymbol{x}^k)$ generated using simultaneous iteration converges to $\boldsymbol{\alpha}$ as $k \to \infty$.

Proof. Our proof naturally mimicks that of Theorem 3.3.8. Define $\Lambda = \frac{1}{2}(1 + \|J_{\boldsymbol{g}}(\boldsymbol{\alpha})\|_\infty)$. Since $\|J_{\boldsymbol{g}}(\boldsymbol{\alpha})\|_\infty < 1$ we have $0 < \|J_{\boldsymbol{g}}(\boldsymbol{\alpha})\|_\infty < \Lambda < 1$. By the continuity of the partial derivatives of $\boldsymbol{g}$ there exists $\delta > 0$ such that $\|J_{\boldsymbol{g}}(\boldsymbol{x})\|_\infty \leq \Lambda$ for all $\boldsymbol{x} \in \bar{B}_\delta(\boldsymbol{\alpha})$, and by Lemma 3.7.7 it follows that

$$\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y})\|_\infty \leq \Lambda \|\boldsymbol{x} - \boldsymbol{y}\|_\infty, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \bar{B}_\delta(\boldsymbol{\alpha}).$$

In particular, since $\boldsymbol{g}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}$, this bound (with $\boldsymbol{y} = \boldsymbol{\alpha}$)) implies that $\boldsymbol{g}(\bar{B}_\delta(\boldsymbol{\alpha})) \subset \bar{B}_\delta(\boldsymbol{\alpha})$, and the desired result then follows from the contraction mapping theorem. □

3.7.2 Newton's method for systems

To generalize the one-dimensional Newton iteration

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \tag{3.24}$$

to the case of a higher-dimensional nonlinear system

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}, \qquad \boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n, \tag{3.25}$$

one replaces the reciprocal of the derivative $f'(x_k)$ by the *inverse* of the Jacobian matrix $J_{\boldsymbol{f}}(\boldsymbol{x}^k)$. Explicitly, the Newton step in multiple dimensions is

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - [J_{\boldsymbol{f}}(\boldsymbol{x}^k)]^{-1} \boldsymbol{f}(\boldsymbol{x}^k), \quad k = 0, 1, 2 \ldots,$$

where $\boldsymbol{x}^k$ denotes the kth iteration. In practice one rarely computes the inverse $[J_{\boldsymbol{f}}(\boldsymbol{x}^k)]^{-1}$ explicitly, and so the iteration is more commonly written implicitly as

$$[J_{\boldsymbol{f}}(\boldsymbol{x}^k)](\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) = -\boldsymbol{f}(\boldsymbol{x}^k), \quad k = 0, 1, 2 \ldots . \tag{3.26}$$

Regarding convergence of the iteration defined by (3.26), similar results hold as in the scalar case. Provided that $\boldsymbol{f}$ is sufficiently smooth, and that its derivatives satisfy appropriate conditions (in particular, the Jacobian matrix should be nonsingular in a neighbourhood of the root), the iterates converge quadratically. But the analysis is more involved and we refer to Süli and Mayers for details.

Theorem 3.7.9. *Let $\boldsymbol{\alpha}$ satisfy $\boldsymbol{f}(\boldsymbol{\alpha}) = \boldsymbol{0}$. Let $\boldsymbol{f}$ be twice continuously differentiable in a neighbourhood of $\boldsymbol{\alpha}$, and suppose that the matrix $J_{\boldsymbol{f}}(\boldsymbol{\alpha})$ is nonsingular. Then the Newton iteration defined by (3.26) converges quadratically to $\boldsymbol{\alpha}$ for a sufficiently good initial guess $\boldsymbol{x}^0$.*

Proof. See Süli and Mayers Theorem 4.4. □

Calculating the one-dimensional iteration (3.24) is usually straightforward (provided we have a good method for evaluating f and f'). But in the multi-dimensional case at each iteration we need to solve a linear system (3.26) governed by the matrices $A_k = J_{\boldsymbol{f}}(\boldsymbol{x}^k)$. This is easier than solving the original nonlinear problem (3.25), but if the dimension n is large (as it often is in practice) the solution of these linear systems is still potentially very costly. In particular, a direct solution method such as Gaussian elimination is often impractical. In the next chapter we will study a number of iterative methods which can provide efficient algorithms for the solution of large linear systems.

Example 3.7.10. *Consider the following system of two nonlinear equations in two unknowns, x_1 and x_2:*

$$\begin{cases} x_1^2 - 2x_1x_2 = 2 \\ x_1 + x_2^2 = -1. \end{cases} \tag{3.27}$$

We can recast the system in the form

$$\boldsymbol{f} = \boldsymbol{0}, \qquad \textit{i.e.} \qquad \begin{cases} f_1(x_1, x_2) = 0 \\ f_2(x_1, x_2) = 0 \end{cases}$$

where $\boldsymbol{f} = (f_1, f_2)$, $f_1(x_1, x_2) = x_1^2 - 2x_1x_2 - 2$ and $f_2(x_1, x_2) = x_1 + x_2^2 + 1$. In Figure 3.8 the two curves $f_1 = 0$ and $f_2 = 0$ are plotted in the square $-6 \le x_1 \le 2$, $-4 \le x_2 \le 4$. This figure has been obtained using the Matlab commands

```
>>  f1 = @(x1,x2)x1.^2-2*x1.*x2-2;
>>  f2 = @(x1,x2)x1+x2.^2+1;
```

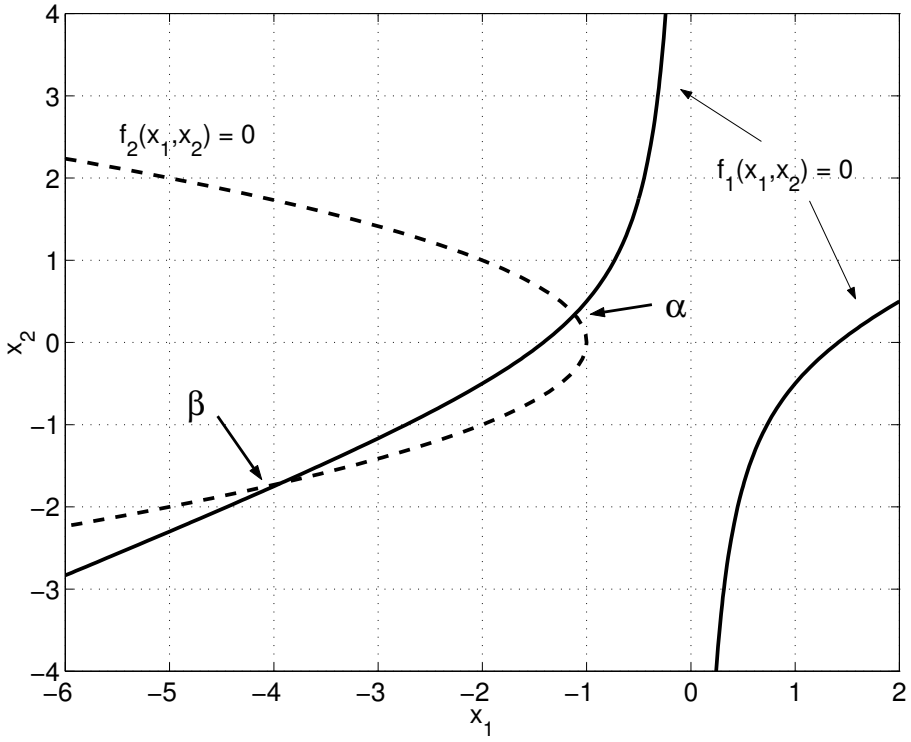


Figure 3.8: *Curves $f_1(x_1, x_2) = 0$ and $f_2(x_1, x_2) = 0$. Their intersection point is a solution to the nonlinear system (3.27).*

```
>>  [x1,x2] = meshgrid(-6:.1:2,-4:.1:4);
>>  contour(x1,x2,f1(x1,x2),[0,0],'b'); hold on;
>>  contour(x1,x2,f2(x1,x2),[0,0],'b--')
```

The zeros of the system are given by the intersection points of the two curves. From the figure we see that there are two zeros $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)$.

The Jacobian matrix of the vector valued function $\boldsymbol{f}$ is

$$J_{\boldsymbol{f}}(\boldsymbol{x} = (x_1, x_2)) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 - 2x_2 & -2x_1 \\ 1 & 2x_2 \end{bmatrix},$$

and Newton's method is: given $\boldsymbol{x}^0 = (x_1^0, x_2^0)$, for each $n = 0, 1, 2, \ldots$ find $\boldsymbol{x}^{k+1}$ such that

$$[J_{\boldsymbol{f}}(\boldsymbol{x}^k)](\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) = -\boldsymbol{f}(\boldsymbol{x}^k). \tag{3.28}$$

We detail the first iteration of the algorithm, using the initial vector $\boldsymbol{x}^0 = [1, 1]^T$. Then

$$J_{\boldsymbol{f}}(\boldsymbol{x}^0) = \begin{bmatrix} 0 & -2 \\ 1 & 2 \end{bmatrix},$$

and $\boldsymbol{x}^1$ *is given by the solution of the linear system:*

$$[J_{\boldsymbol{f}}(\boldsymbol{x}^0)](\boldsymbol{x}^1 - \boldsymbol{x}^0) = -\boldsymbol{f}(\boldsymbol{x}^0).$$

The iterations continue using (3.28) until some stopping criteria is met.

3.8 Stopping criteria

An important practical issue concerning iterative methods is determining when to stop the iteration. For simplicity we restrict our attention to the one-dimensional case.

Suppose we are using an iterative method (e.g. the bisection method, or a fixed point iteration $x_{k+1} = \phi(x_k)$) to generate a sequence of approximations x_k converging to a root α of the equation $f(\alpha) = 0$ for some $f : \mathbb{R} \to \mathbb{R}$. We would like to specify a criterion on k that guarantees that the error $e_k := x_k - \alpha$ satisfies

$$|e_k| = |x_k - \alpha| \leq tol,$$

where $tol > 0$ is a user-specified tolerance.

A priori criteria. For some of the numerical methods above, we have *a priori* error bounds, i.e. a bound on the error that can be evaluated *before* any computations are made. For instance, the bisection method produces iterates that satisfy

$$|e_k| = |x_k - \alpha| \leq \frac{b-a}{2^{k+1}}.$$

Therefore the error $|e_k| \leq tol$ provided that

$$k \geq \log_2\left(\frac{b-a}{tol}\right) - 1.$$

Similarly, if we are using a fixed point iteration $x_{k+1} = \phi(x_k)$ where ϕ satisfies the conditions of the Contraction Mapping Theorem, Theorem 3.3.6, then

$$|e_k| = |x_k - \alpha| \leq \Lambda^k |x_0 - \alpha| \leq \Lambda^k (b-a), \qquad \Lambda \in (0,1),$$

so k can also be chosen appropriately to ensure that the error tolerance is satisfied.

A posteriori criteria. *A priori* stopping criteria are worst-case bounds for the behaviour of the algorithm, and can sometimes be pessimistic in some situations. In other cases, an *a priori* bound might not be known. This motivates the use of bounds that are *a posteriori*, i.e. that can be evaluated after an approximation x_k is computed. These typically take the form

$$|x_k - \alpha| \leq R(x_k) \tag{3.29}$$

where $R\colon \mathbb{R} \to \mathbb{R}$ is a function to be determined by analysis of the problem. Once x_k is computed for some $k \in \mathbb{N}$, we can then decide to stop the iteration when

$$R(x_k) \leq tol$$

is satisfied. *A posteriori* bounds typically require some analysis of the problem at hand before they can be used. A simple example of an *a posteriori* bound for root-finding is based on the *residual* $f(x_k)$. The bound is based on the following theorem.

Theorem 3.8.1. *Suppose that the function $f\colon \mathbb{R} \to \mathbb{R}$ is continuously differentiable on $[a, b]$ and that a root α exists in $[a, b]$. Suppose that*

$$L \leq |f'(x)| \leq U \quad \forall x \in [a, b],$$

for some positive constants $L > 0$ and $U > 0$. Then, for any $x \in [a, b]$,

$$\frac{1}{U}|f(x)| \leq |x - \alpha| \leq \frac{1}{L}|f(x)|. \tag{3.30}$$

Proof. Since x and α are both in $[a, b]$, by the MVT there exists $z \in [a, b]$ such that

$$f(x) = f(x) - f(\alpha) = f'(z)(x - \alpha)$$

Since $L > 0$ we have $f'(z) \neq 0$ and thus $x - \alpha = \frac{f(x)}{f'(z)}$. We then obtain (3.30) from the bounds for $L \leq |f'(z)| \leq U$. □

Under the hypotheses of Theorem 3.8.1, provided we know L, we can use the stopping criteria

$$R(x_k) := \frac{1}{L}|f(x_k)| \leq tol.$$

Notice that this *a posteriori* analysis does not depend on the choice of numerical method. However, it does depend on the properties of the function f and may not be always applicable.

Heuristic criteria. If we don't have a rigorous error bound such as an *a priori* or *a posteriori* bound for a particular problem, we might consider heuristic (non-rigorous) approaches. Such approaches don't guarantee that the error is definitely below a certain threshold, but they often indicate that the error is of the desired order of magnitude. Two common heuristics are residual stopping criteria and increment stopping criteria.

If f is continuous then $f(x_k) \to 0$. And if f is differentiable with $f'(\alpha) \neq 0$ then

$$e_k = \frac{f(x_k)}{f'(\alpha)} + o(e_k), \qquad k \to \infty.$$

Hence, for x^k close to α, the quantity $\frac{f(x_k)}{f'(\alpha)}$ will be a good approximation to the error e_k. While we don't generally know $f'(\alpha)$, we might replace $f'(\alpha)$ by a computable approximation, for example $f'(x_k)$. This leads to the heuristic (i.e. non-rigorous) stopping criterion

$$\frac{|f(x_k)|}{|f'(x_k)|} \leq tol.$$

This condition won't in general ensure that $|e_k| \leq tol$. But if f is sufficiently smooth then for large k it will ensure that $|e_k| \leq 2\,tol$, say, so that $|e_k|$ is approximately of the desired order of magnitude. If it is known that $|f'(\alpha)| \approx 1$ then one might even use the simpler condition

$$|f(x_k)| \leq tol.$$

An alternative heuristic stopping criterion is found by considering the iterative increment: the iterations are stopped when

$$|x_k - x_{k+1}| \leq tol. \tag{3.31}$$

To explore the effectiveness of this criterion, consider the following argument, for a fixed point formulation. Recalling that $e_k = x_k - \alpha$, if ϕ is differentiable we can write the error as

$$e^{k+1} = x_{k+1} - \alpha = \phi(x_k) - \phi(\alpha) = \phi'(\xi_k)e_k,$$

for some ξ_k between x_k and α. Therefore

$$x_k - x_{k+1} = (x_k - \alpha) - (x_{k+1} - \alpha) = e^k - e^{k+1} = (1 - \phi'(\xi_k))\, e_k.$$

Assuming that $|\phi'(\alpha)| < 1$, so that $x_k \to \alpha$ as $k \to \infty$, it also follows that $\phi'(\xi_k) \to \phi'(\alpha)$ as $k \to \infty$. Hence for large k we have

$$e_k \approx \frac{x_k - x_{k+1}}{1 - \phi'(\alpha)}.$$

This shows that the criterion (3.31) gives approximate error control to the accuracy tol, provided that $\phi'(\alpha)$ is not close to 1.