Chapter 3

Nonlinear equations

3.1 Introduction

Very few nonlinear equations can be solved analytically. For example, it is easy to solve the equation $x^2 + 2x + 1 = 0$: the left hand side is $(x + 1)^2$ and therefore there is a single root $x = -1$ with multiplicity two. There are other equations whose solution is not so easily found, like, for example,

$$e^x + x = 2.$$

From the graph of this equation it is clear that there is one and only one solution. However, there is no formula to obtain its exact value. The purpose of numerical methods for the solution of nonlinear equations is to fill in this gap: to give approximate, but accurate, solutions of nonlinear equations that cannot be solved analytically.

In this unit we give an introduction to this area of numerical analysis by discussing some simple algorithms. It should be made clear that the algorithms currently used in commercial packages and in research laboratories are far more advanced than anything that we will discuss. However, they are based on the same ideas and differ mainly in details of the implementation and in using various tricks to guarantee a fast and global convergence.

We will focus our attention mainly on the solution of a single nonlinear equation in one variable. However, the methods that we will discuss can be extended to systems of nonlinear equations in more than one variable: we will discuss briefly how to do this at the end of this chapter.

In all that follows we assume that we have to solve the nonlinear equation $f(x) = 0$, where x is a real number and $f(x)$ is a real differentiable function.

3.2 A simple example: The bisection method

The bisection method is by far the simplest method and, even though it is not used very much in practice, it is sturdy and reliable. Moreover, we can use it to introduce some general comments on the practical implementation of root finding algorithms.

Very briefly, the method assumes that by suitable inspired guesses two points $x_0^{(L)}$ and $x_0^{(R)}$ have been chosen such that

$$f(x_0^{(L)})f(x_0^{(R)}) \leq 0, \tag{3.1}$$

i.e. the function $f(x)$ changes sign in the interval $[x_0^{(L)}, x_0^{(R)}]$. If the product is zero then one of the two factors is zero and the problem is solved. We therefore assume that the inequality in (3.1) is strict. Since the function $f(x)$ is continuous, by the intermediate value theorem there exists a point $s \in (x_0^{(L)}, x_0^{(R)})$ such that $f(s) = 0$, i.e. s is a root of the equation $f(x) = 0$. The idea behind this method is that the mid point between $x_0^{(L)}$ and $x_0^{(R)}$, $x_0^{(M)}$, is an approximation of the root s. If a more accurate estimate is needed we can refine the approximation by checking in which half interval $(x_0^{(L)}, x_0^{(M)})$ or $(x_0^{(M)}, x_0^{(R)})$ the function $f(x)$ changes sign. We then discard the other half-interval and repeat the procedure.

More formally, the iteration procedure involves first constructing a new point, the centre of the interval and estimate of the root,

$$x_n^{(M)} = \frac{x_n^{(L)} + x_n^{(R)}}{2}, \quad n = 0, 1, 2, \ldots$$

and evaluating $f(x_n^{(M)})$. If this estimate of the root is not sufficiently accurate a new set of left and right points are chosen according to the sign of $f(x_n^{(M)})$:

$$\begin{cases} x_{n+1}^{(L)} & = x_n^{(L)}, \\ x_{n+1}^{(R)} & = x_n^{(M)} \end{cases} \qquad \text{if } f(x_n^{(L)})f(x_n^{(M)}) < 0, \tag{3.2}$$

$$\begin{cases} x_{n+1}^{(L)} & = x_n^{(M)}, \\ x_{n+1}^{(R)} & = x_n^{(R)} \end{cases} \qquad \text{if } f(x_n^{(L)})f(x_n^{(M)}) > 0. \tag{3.3}$$

The procedure is repeated until a stopping condition is reached. There are three conditions that must be checked by any iteration procedure: if any of them is satisfied the iteration must stop.

1. The number of iterations has exceeded a predetermined value: this is used to avoid cases where the convergence is exceedingly slow or, for any reason, the algorithm is going in an infinite loop.

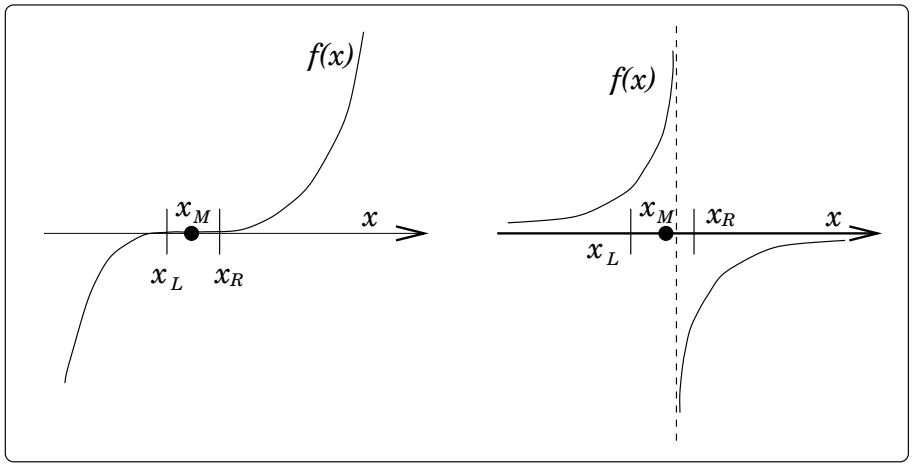

Figure 3.1: *In the left hand case the criterion* $\left|x_n^{(R)} - x_n^{(L)}\right| < \delta$ *fails, in the right hand case the criterion* $\left|f(x_n^{(M)})\right| < \varepsilon$ *fails.*

2. The absolute value of the function at the estimated root, $f(x_n^{(M)})$ is smaller than a predetermined number ε (usually fixed by the number of significant digits of the floating point representation).

3. The difference between two successive values of the estimated root (or the difference between $x_n^{(R)}$ and $x_n^{(L)}$ in the case of the bisection method) is smaller that a predetermined number δ, the requested accuracy of the roots.

Figure 3.1 shows two pathological cases where one of the last two criteria is satisfied, but not the other. In the left hand case there is a multiple root (a bane of root finding algorithms): the value of the function is very small, but the the left and right hand point are not close. For many numerical methods the speed of convergence is proportional to the slope of $f(x)$ at its root. Therefore their convergence is extremely slow at multiple roots where $f(x)$ is flat. This is not the case of the bisection method because its rate of convergence is independent of the slope of the function. However, if the value of the function is very close to zero then numerical errors may introduce spurious zeros and force the algorithm to converge to a spurious root.

In the right hand case the bisection interval is very small so that $\left|x^{(R)} - x^{(L)}\right| < \delta$ but the function is not small. While it is true that in this case the function is not continuous, it is also true that it may not be easy to determine whether the function whose zeros we wish to compute is continuous and so we must make a root finding algorithm capable of handling cases as pathological as these examples.

Remark 1 - The error in the location of the root at the n-th step is smaller than

$(x_0^{(R)} - x_0^{(L)})/2^n$.

Remark 2 - This method is easy to code and it always converges. However, it is rather slow.

3.3 Contraction mappings

3.3.1 Introduction

Many of the methods that we will discuss for solving nonlinear equations like

$$f(x) = 0 \tag{3.4}$$

are *iterative* and can be written in the form

$$x_{n+1} = g(x_n), \tag{3.5}$$

for some suitable function $g(x)$ and initial approximation x_0. The aim of the method is to find a suitable function $g(x)$ such that the sequence has a limit and that the limit is a root of $f(x)$:

$$\lim_{n\to\infty} x_n = s \quad \text{and} \quad f(s) = 0.$$

Note that if the limit exists then it is also a fixed point of the map $g(x)$:

$$s = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} g(x_n) = g\left(\lim_{n\to\infty} x_n\right) = g(s).$$

For example, in the case of the nonlinear problem $f(x) = 0$ we can define the function $g(x)$ to be

$$g(x) = x - f(x)$$

and use the mapping (3.5) to attempt finding the roots of $f(x)$. Methods of this kind are called *functional iterations methods* or *fixed point methods*.

Graphically, the solutions of (3.4) or the fixed points of (3.5) are the intersections between the graph of $g(x)$ and the line $y = x$ (see Figure 3.2). The iteration of the map (3.5) can be represented on the same graph (see Figure 3.2): in the case of the solid line path the method is converging to the fixed point, while for the dashed line path the method is diverging.

There are some general theorems that state under what condition the mapping (3.5) converges. Before studying them, however, it is worthwhile to make some general remarks.

- Usually iterative methods are valid for real and complex roots. However, in the latter case complex arithmetic must be incorporated into the appropriate computer codes and the initial estimate of the root must usually be complex.

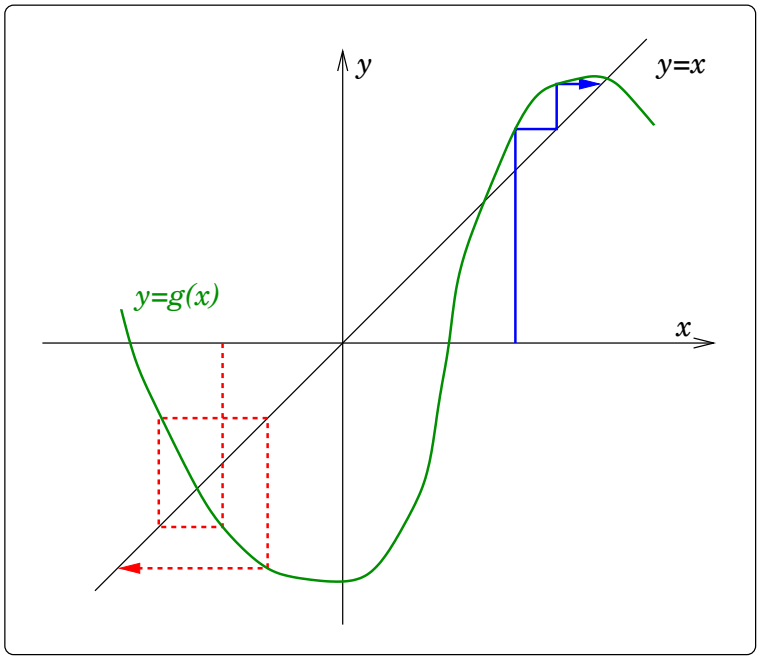

Figure 3.2: *Graphical representation of the functional iteration. The straight line paths are the graphical representation of the mapping (3.5).*

- The iterative methods require at least one initial estimate or guess at the location of the root being sought. If this initial estimate is "sufficiently close" to a root, then, in general, the procedure will converge. The problem of how to obtain such a "good" estimate is unsolved in general.
- As a general empirical rule, the schemes which converge more rapidly (i.e. higher order methods) require closer estimates. In practice, these higher order schemes may require the use of more significant digits in order that they converge as theoretically predicted. Thus, it is frequently a good idea to use a simple method to start with and then, when fairly close to the root, to use some higher order method for just a few iterations.

3.3.2 Geometrical interpretation of fixed point schemes

Before discussing formally what properties a map must have in order for the iteration scheme (3.5) to converge, we can obtain an approximate idea by considering the four maps in Figure 3.3. From the top two maps it is clear that in order to have a fixed point we must require $g(x)$ to be continuous (top left) and, moreover, that the range is contained in the domain (top right). These requirements are not enough to have a unique fixed point as it is shown in the bottom left corner: if the slope of the map is too high there may be two or more fixed points. It is only if the slope is smaller than unity that there can be only one fixed point (bottom right corner). Note than we do not require that map to be differentiable: it can have as many corners as it wish. The case of the two bottom maps is illustrated pictorially in Figure 3.4: at each iteration the image of the starting set I gets smaller and smaller until it reduces to a point (see Figure 3.4).

3.3.3 Definitions

We must now phrase these intuitive results in more formal terms. We start by defining a contracting map.
Definition - A continuous map $g(x)$ from an interval $I = [a, b] \subseteq \mathbb{R}$ into $\mathbb{R}$ is *contracting* if

1. the image of I is contained in I:

$$g(I) \subseteq I \quad \Leftrightarrow \quad g(x) \in I \, \forall x \in I.$$

2. the function $g(x)$ is Lipschitz continuous in I with Lipschitz constant $L < 1$:

$$|g(x) - g(y)| \leq L|x - y| \quad \forall x, y \in I.$$

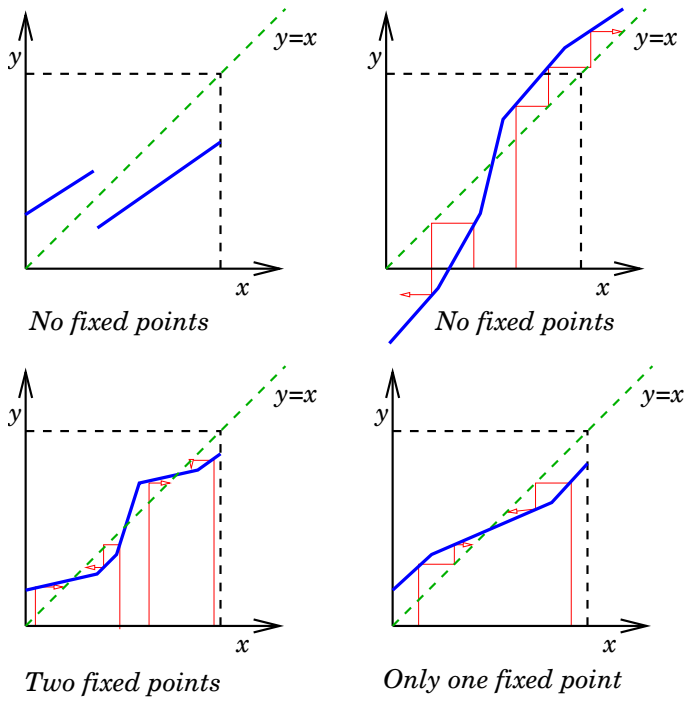

Figure 3.3: *Examples of maps and their convergence properties.*

In other words, the distance between the images is smaller than the distance between the two starting points.

Remark 1 - A function that is Lipschitz continuous is "more" than continuous, but "less" than differentiable. For example, the function $f_1(x) = \sqrt{x}$ is continuous in the interval $[0, 1]$, but it is not Lipschitz. On the other hand, the function $f_2(x) = |x|$ is Lipschitz in the interval $[-1, 1]$, but it is not differentiable at $x = 0$.
Remark 2 - If the function $g(x)$ is differentiable and Lipschitz continuous with constant L then

$$\left|\frac{\mathrm{d}g}{\mathrm{d}x}\right| \leq L.$$

3.3.4 Convergence theorems

A map that is contracting is also called a *contraction mapping*. From the definition and Figures 3.3 and 3.4 we can intuitively understand that if the map $g(x)$ in (3.5) is contracting, then we are guaranteed convergence. All this is expressed more formally (and more clearly) in the following sets of theorems.

First of all, we prove that if the image of the interval is contained in the interval (without any requirement of Lipschitz continuity) then there is at least one fixed point (but there may be many).

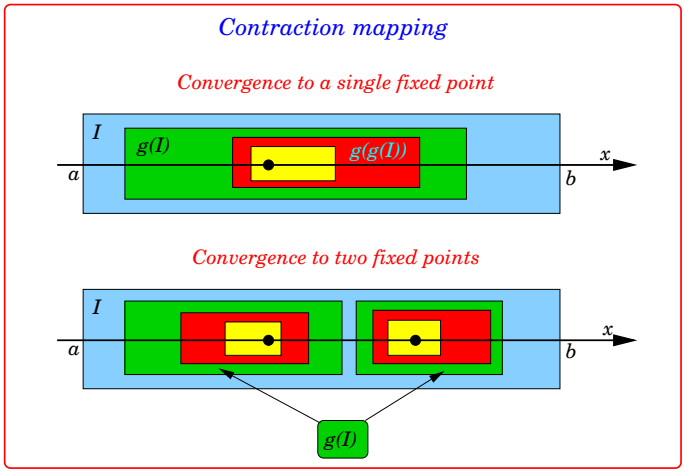

Figure 3.4: *Graphical representation of the contraction mapping principle. At each iteration the image of the interval gets smaller and the iterations of the contraction map converge towards its fixed point(s).*

Theorem 3.3.1 *If the function $g(x)$ is continuous in $I = [a, b]$ and $g(I) \subseteq I$, then $g(x)$ has at least one fixed point in I.*

Proof - Since $g(I) \subseteq I$ we must have

$$a \leq g(a) \leq b \quad \text{and} \quad a \leq g(b) \leq b.$$

If either $g(a) = a$ or $g(b) = b$ then there is one fixed point and the theorem is proved. Otherwise, the following inequalities must hold:

$$g(a) - a \geq 0 \quad \text{and} \quad g(b) - b \leq 0.$$

Define the function $F(x) = g(x) - x$. $F(x)$ is continuous and $F(a) \geq 0$, while $F(b) \leq 0$. Therefore, by the intermediate value theorem there exists a point $c \in [a, b]$ such that

$$F(c) = 0 \implies c = g(c).$$

■

Exercise - Give a graphical representation of this theorem. In particular show that the contraction map that would produce a graph similar to the bottom part of Figure 3.4 must have a jump discontinuity in $[a, b]$.

Theorem 3.3.1 provides us with an important result, but not with enough. Ideally we would like the zero to be unique. In order for this to be true, we must have more stringent requirements on $g(x)$: it must not vary too rapidly. This is assured if $g(x)$ is Lipschitz continuous with a sufficiently small Lipschitz constant.

Theorem 3.3.2 (Contraction mapping theorem) *If $g(x)$ is a contraction mapping in an interval $I = [a, b]$ then there exists one and only one fixed point of the map in $[a, b]$.*

We will prove a slightly less strong version of this theorem: we require that the map $g(x)$ is differentiable in I and that $|g'(x)| \leq L < 1$. Such a function is Lipschitz continuous of Lipschitz constant L: therefore it is a contraction mapping.

Theorem 3.3.3 *If $g(x)$ is a differentiable contraction mapping in an interval $I = [a, b]$, i.e.*

$$|g'(x)| \leq L < 1, \qquad \forall x \in [a, b],$$

then there exists one and only one fixed point of the map in $[a, b]$.

Proof - The existence of the derivative implies that the function $g(x)$ is continuous. Since, by hypothesis $g(I) \subseteq I$, then by Theorem 3.3.1 there is at least one fixed point, s_1. Suppose that there is another one $s_2 : s_2 = g(s_2)$ and $s_1 \neq s_2$. We can use the mean value theorem[1] to prove that this is impossible:

$$|s_2 - s_1| = |g(s_2) - g(s_1)| = |g'(\xi)(s_2 - s_1)| \leq L|s_2 - s_1| < ||s_2 - s_1|.$$

This inequality cannot be true and therefore there can be no other fixed point. (The proof assuming Lipschitz continuity only is essentially identical.) ■

The consequence of this theorem is that the algorithm represented by the mapping (3.5) is guaranteed to have a root in an interval $[a, b]$ if the map is a contraction mapping in this interval. The following theorem tells us how to find it.

Theorem 3.3.4 *Let $I = [a, b]$ and suppose that $g(x)$ is a contraction mapping in I. Then for arbitrary $x_0 \in I$, the sequence $x_n = g(x_{n-1})$, $n = 1, 2, \ldots$ converges to the unique fixed point, s, of the map. Moreover, if the error e_n at the n-th stage is defined by $e_n = x_n - s$ then*

$$|e_n| \leq \frac{L^n}{1 - L}|x_1 - x_0|.$$

[1] **Mean value theorem** - For any differentiable $F(x)$ in $I \subseteq \mathbb{R}$ and any $c, d \in I$ there exists a point $\xi \in [c, d]$ such that

$$F'(\xi) = \frac{F(d) - F(c)}{d - c}.$$

Proof - To prove the convergence to the fixed point, whatever the arbitrary guess $x_0 \in I$, we use the Lipschitz property of the map and the fact that $s = g(s)$ to bound $|e_n|$ from above with a bound that tends to zero as n tends to infinity. As a first step we have:

$$\begin{aligned} |e_n| &= |x_n - s| && (3.6)\\ &= |g(x_{n-1}) - g(s)| && (3.7)\\ &\leq L|x_{n-1} - s| && (3.8)\\ &= L|e_{n-1}|. && (3.9) \end{aligned}$$

By applying this inequality over and over again we obtain

$$\begin{aligned} |e_n| &\leq L|e_{n-1}| && (3.10)\\ &\leq L^2|e_{n-2}| && (3.11)\\ &\leq \dots && (3.12)\\ &\leq L^n|e_0| && (3.13) \end{aligned}$$

By the definition of contraction mapping $L < 1$ and therefore

$$\lim_{n\to\infty} L^n = 0 \implies \lim_{n\to\infty} x_n = s.$$

Equation (3.13) provides a bound on the error of the n-th estimate in terms of the initial error. Unfortunately this quantity is not know, because we do not know the root s of the equation. We therefore must replace $|e_0|$ in (3.13) with an expression that we can compute, namely $|x_0 - x_1|$:

$$\begin{aligned} |x_0 - s| &= |x_0 - x_1 + x_1 - s| && (3.14)\\ &\leq |x_0 - x_1| + |x_1 - s| && (3.15)\\ &\leq |x_0 - x_1| + L|x_0 - s| && (3.16)\\ \implies \quad |e_0| &= |x_0 - s| && (3.17)\\ &\leq \frac{|x_0 - x_1|}{1 - L}. && (3.18) \end{aligned}$$

Using (3.13) we obtain

$$|e_n| \leq L^n|e_0| \leq \frac{L^n}{1 - L}|x_0 - x_1|. \qquad (3.19)$$

■

These four theorems are represented graphically in Figure 3.5. Since the slope of the function $g(x)$ is smaller than unity successive iterations of the map get closer and closer to its fixed point.

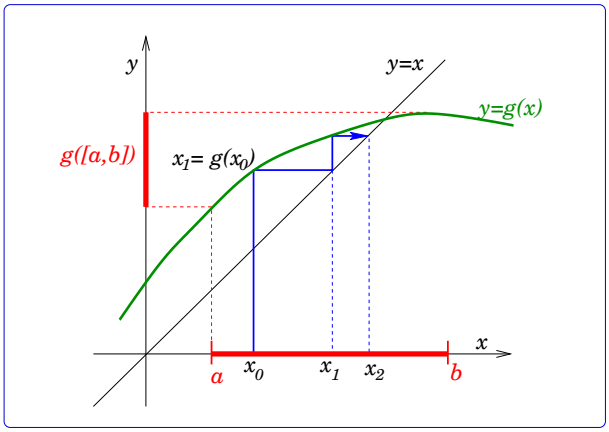

Figure 3.5: *Graphical representation of the contracting mapping theorems. The map $g(x)$ is a contraction mapping in $[a, b]$: the image of the interval is smaller than the original interval. Each iteration of a starting guess x_0 gets closer and closer to the fixed point.*

3.3.5 Speed of convergence

The bound (3.19) on the error provided by Theorem 3.3.4 depends, through the Lipschitz constant L, on the interval chosen to estimate the map. Moreover, it is reasonable to assume that the rate of convergence of the map, i.e. the rate of decrease of the error e_n depends only on the properties of the map in a neighbourhood of the root. We can show that this is indeed the case if the map is differentiable, so that it is possible to expand it in a Taylor polynomial: call $g(x)$ a suitably differentiable contraction mapping in $I = [a, b]$, s its fixed point, $x_0 \in I$ the starting point of the iteration and $e_n = x_n - s$ the error at the n-th iteration. Using the definition of the map and Taylor's expansion we can write:

$$\begin{aligned} e_{n+1} &= x_{n+1} - s \\ &= g(x_n) - g(s) \\ &= g'(s)(x_n - s) + \frac{g''(s)}{2!}(x_n - s)^2 + \ldots + \frac{g^{(k)}(s)}{k!}(x_n - s)^k + R_{n,k} \\ &= g'(s)e_n + \frac{g''(s)}{2!}e_n^2 + \ldots + \frac{g^{(k)}(s)}{k!}e_n^k + R_{n,k}, \end{aligned}$$

where $R_{n,k}$ is the remainder of the expansion:

$$R_{n,k} = \frac{g^{(k)}(\xi)}{k!}(x_n - s)^{k+1}, \quad \xi \in [x_n, s].$$

Assuming that $g'(s) \neq 0$ then

$$e_{n+1} \sim g'(s)e_n,$$

i.e. the error decreases at a constant rate at each iteration: such a method is called *linear* or *first order*. If, instead, $g'(s) = 0$, but $g''(s) \neq 0$ then

$$e_{n+1} \sim g''(s)e_n^2,$$

i.e. the error at each iteration is proportional to the square of the previous error: such a method is called a *quadratic* or *second order* method. The more derivatives of $g(s)$ vanish the higher the order of the method and the faster the convergence. However, it may well be that the method will converge only if the starting point is very close to the root.

3.3.6 Error propagation

The final question that we must answer before discussing practical implementations of the theory we have just studied is "Are these theorems numerically stable?" In other words, what is the effect of the numerical error on the convergence properties of a contraction map? In actual computations it may not be possible, or practical, to evaluate the function $g(x)$ exactly (i.e. only a finite number of decimals may be retained after rounding or $g(x)$ may be given as the numerical solution of a differential equation, etc.). For any value of x we may then represent our approximation to $g(x)$ by $G(x) = g(x)+\delta(x)$ where $\delta(x)$ is the error committed in evaluating $g(x)$. Frequently we may know a bound for $\delta(x)$, i.e. $\delta(x) < \delta$. Thus the actual iteration scheme which is used may be represented as

$$X_{n+1} \equiv G(X_n) = g(X_n) + \delta_n, \quad n = 0, 1, 2, \ldots, \tag{3.20}$$

where the X_n are the numbers obtained from the calculations and the $\delta_n \equiv \delta(X_n)$ satisfy

$$|\delta_n| \leq \delta, \quad n = 0, 1, 2, \ldots \tag{3.21}$$

We cannot expect the computed iterates X_n of (3.20) to converge. However, under proper conditions, it should be possible to approximate a root to an accuracy determined essentially by the accuracy of the computations, δ.

For example, from Figure 3.6 we can see that for the special case of $g(x) = \alpha + L(x - \alpha)$, the uncertainty in the root α is bounded by $\pm\delta/(1 - L)$. We note that if the slope L is close to unity the problem is not "properly posed". The following theorem states quite generally that when the functional iteration scheme is convergent, the presence of errors in computing $g(x)$, of magnitudes bounded by δ, causes the scheme to estimate the root α with an uncertainty bounded by

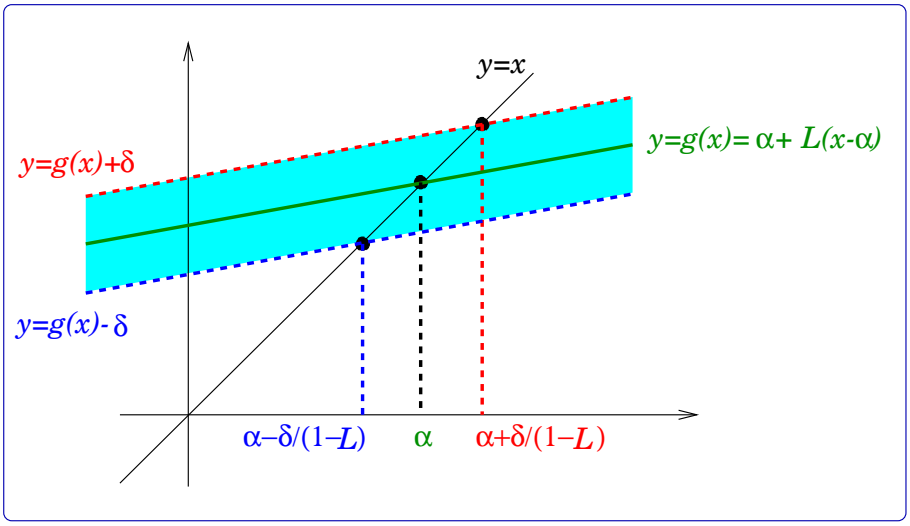

Figure 3.6: *The width of the numerical error band around the fixed point of the iteration map.*

$\pm\delta/(1-L)$, where L is the Lipschitz constant of the contraction mapping. The phrasing of this theorem is slightly different from the previous ones because the numerical error δ forces us to define quite strictly the interval we want to work in: we know that $g(I) \subseteq I$, but it is not generally true that $G(I) \subseteq I$.

Theorem 3.3.5 *Let $g(x)$ be a contraction mapping with fixed point s and let L be its Lipschitz constant in the interval $I(s, r_0) \equiv [s - r_0, s + r_0]$. Let δ be the bound on the numerical errors of the iterates of the numerical map (3.20) as defined in (3.21). Finally, assume that the starting point of the iteration of (3.20) is a point X_0 in the smaller interval*

$$X_0 \in I(s, R_0) \equiv [s - R_0, s + R_0], \tag{3.22}$$

where

$$0 < R_0 \leq r_0 - \frac{\delta}{1-L}.$$

Then the iterates X_n of (3.20) with the errors bounded by (3.21), lie in the interval $I(s, r_0)$, and

$$|s - X_n| \leq \frac{\delta}{1-L} + L^n \left(R_0 - \frac{\delta}{1-L} \right), \tag{3.23}$$

where $L^n \to 0$ as $n \to \infty$.

Proof - The proof of this theorem involves showing that the numerical error does not push the iteration outside the interval $I(s, r_0)$ where it is defined. In the process of doing so we also derive the bound (3.23) on the error of the numerical

estimate. The proof that the iterations are always in the interval $I(s, r_0)$ is by induction.

The point X_0 is, by hypothesis, inside the interval $I(s, R_0) \subseteq I(s, r_0)$. We now suppose that the iterations $X_0, X_1, \ldots, X_{n-1}$ are in $I(s, r_0)$ and proceed to show that also $X_n \in I(s, r_0)$. By (3.20) and (3.21) we have

$$|s - X_n| \leq |[g(s) - g(X_{n-1})] - \delta_{n-1}| \leq |[g(s) - g(X_{n-1})]| + \delta.$$

Since $g(x)$ is a contraction mapping of Lipschitz constant L we can write

$$\begin{aligned} |s - X_n| &\leq L|s - X_{n-1}| + \delta \\ &\leq L^2|s - X_{n-2}| + L\delta + \delta \\ &\leq L^n|s - X_0| + (L^{n-1} + \ldots 1)\delta \\ &= L^n|s - X_0| + \frac{1 - L^n}{1 - L}\delta. \end{aligned}$$

Hypothesis (3.22) implies that $|s - X_0| \leq R_0$ so that

$$\begin{aligned} |s - X_n| &\leq L^n R_0 + \frac{1 - L^n}{1 - L}\delta \\ &= L^n R_0 + \frac{\delta}{1 - L} - L^n \frac{\delta}{1 - L} \\ &\leq R_0 + \frac{\delta}{1 - L} \\ &\leq r_0. \end{aligned} \tag{3.24}$$

Thus all the iterates are in the interval $I(s, r_0)$ and the iteration process is defined. Moreover (3.24) can be rewritten as (3.23), thus completing the proof. ■

Remark - Theorem 3.3.5 shows that the method is "as convergent as possible", that is, the computational errors which arise from the evaluation of $g(x)$ may cumulatively produce an error of magnitude at most $\delta/(1 - L)$. Moreover, such errors limit the size of the error bound *independently of the number of iterations*. Therefore, it is pointless to iterate the map until $L^n r_0 \ll \delta/(1 - L)$.

3.4 Examples of iteration procedures

3.4.1 Introduction

We have completed the theoretical introduction to the numerical solution of nonlinear equations. We must now discuss some of the algorithms that apply the theory that we have developed.

3.4.2 The chord method (first order)

The chord method and Newton's methods are examples of the application of the contraction mapping theorems. Both these methods can be introduced using a general and elegant framework. As usual we suppose that we have to solve the nonlinear equation

$$f(x) = 0,$$

in some interval $a \leq x \leq b$. To define the different methods to solve this problem we introduce a function $\varphi(x)$ such that

$$0 < \varphi(x) < \infty, \quad x \in [a, b], \tag{3.25}$$

and we use it to construct a contraction mapping $x_{n+1} = g(x_n)$, where

$$g(x) = x - \varphi(x) f(x). \tag{3.26}$$

The fixed points of the map $g(x)$ are the roots of the function $f(x)$.

The simplest choice for $\varphi(x)$ in (3.26) is to take

$$\varphi(x) \equiv m \neq 0, \tag{3.27}$$

$$\implies \quad g(x) = x - mf(x) \tag{3.28}$$

$$\implies \quad x_{n+1} = x_n - mf(x_n) \tag{3.29}$$

where m is a number that we must choose appropriately in order for $g(x)$ to be a contraction mapping in $[a, b]$. The range of m depends on the slope of $f(x)$ in the interval $[a, b]$. From Theorem 3.3.3 we know that $g(x)$ is a contraction mapping if

$$|g'(x)| < 1 \ \forall x \in [a, b]$$

$$\implies \quad |1 - mf'(x)| < 1 \ \forall x \in [a, b]$$

$$\implies \quad 0 < mf'(x) < 2 \ \forall x \in [a, b]. \tag{3.30}$$

Thus m must have the same sign as $f'(x)$, while if $f'(x) = 0$ the inequality cannot be satisfied.

The iterates of (3.29) have a geometrical realisation in which the value x_{n+1} is the x intercept of the line with slope $1/m$ through $(x_n, f(x_n))$ (see Figure 3.7). The inequality (3.30) implies that this slope should be between ∞ (i.e. vertical) and $f'(x)/2$ (i.e. half the slope of the tangent to the curve $y = f(x)$). It is from this geometric description that the name chord method is derived - the next iterate is determined by a chord of constant slope joining a point on the curve to the x-axis.

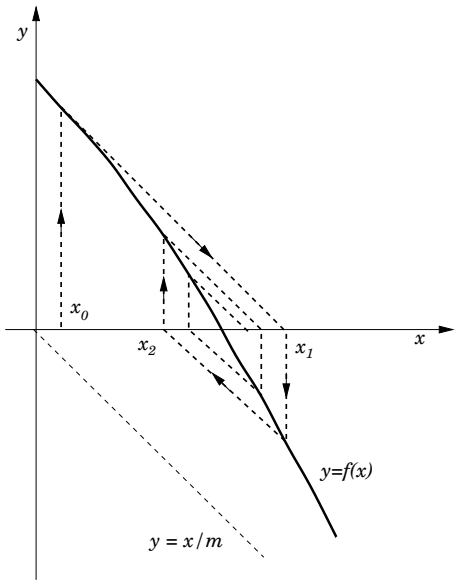

Figure 3.7: *Geometrical representation of the chord method: each new iterate is determined by a chord of constant slope joining a point on the curve to the x-axis.*

3.4.3 Newton's method (second order)

The idea behind Newton's method is to chose the function $\varphi(x)$ in order that the derivative of the iteration mapping $g(x)$ is zero at the root $x = s$. This is ensured by the choice

$$\varphi(x) = \frac{1}{f'(x)} \implies g(x) = x - \frac{f(x)}{f'(x)}, \tag{3.31}$$

so that the iteration procedure is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{3.32}$$

This root finding algorithm is called Newton's method. Theorem 3.3.3 guarantees that the method converges in an interval $[a, b]$ containing the root provided that $|g'(x)| < 1$ in the interval. It is at least second order at the root s of the equation $f(x) = 0$, if $f'(s) \neq 0$ and $f''(x)$ exists, since

$$g'(s) = \frac{f(s)f''(s)}{[f'(s)]^2} = 0. \tag{3.33}$$

The geometrical interpretation of this scheme simply replaces the chord in Figure 3.7 by the tangent to the line to $y = f(x)$ at $x_n, f(x_{n+1})$ (see Figure 3.8).

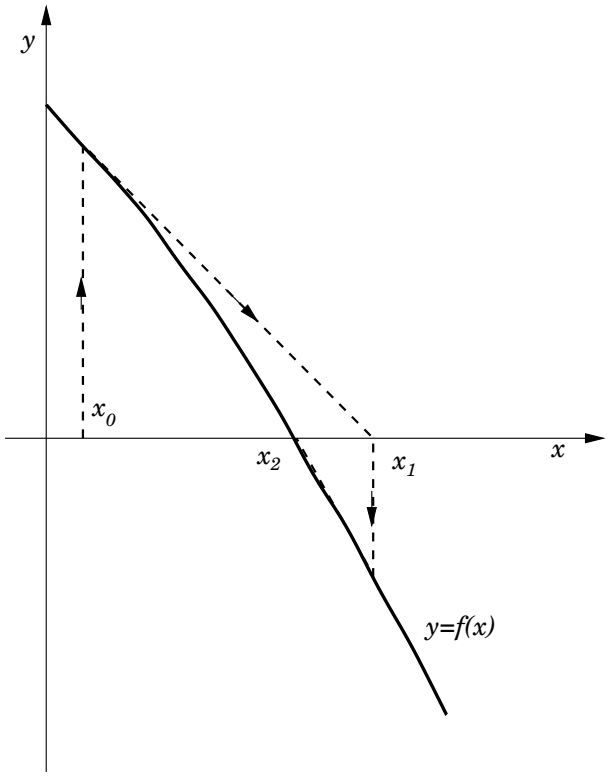

Figure 3.8: *Geometrical representation of Newton's method: each new iterate is intersection of the tangent to the graph of* $f(x)$ *with the* x*-axis.*

Remark 1 - It should be noted that Newton's method may be undefined and the condition (3.25) violated if $f'(x) = 0$ for some $x \in [a, b]$. In particular, if at the root $x = s$, $f'(s) = 0$, the procedure may no longer be of second order since the hypotheses that lead to (3.33) are not satisfied. To examine this case we assume that $f(x)$ has a root of multiplicity p at $x = s$. In other words we can write,

$$f(x) = (x - s)^p h(x), \quad p > 1,$$

where the function $h(x)$ has a second derivative and $h(s) \neq 0$. If we substitute this expression in the definition of $g(x)$ in (3.31) we find that

$$|g'(s)| = 1 - \frac{1}{p}.$$

So only in the case of a linear root, i.e. $p = 1$ is Newton's method second order, but it will converge as a first order method in the general case $p \neq 1$.
Remark 2 - Convergence is quadratic only if we are close to the root s.
Remark 3 - The advantage of Newton's method with respect to the bisection (and chord) method is the faster convergence. The main disadvantage with respect to the bisection method is that we need to start relatively close to the root in order to be sure that the method will converge.

3.4.4 Secant method (fractional order)

Newton's method requires the evaluation of the derivative of the function $f(x)$ whose root we want to find. To do this may be very complicated and time consuming: the secant method obviates this problem by approximating the derivative of $f'(x)$ with the quotient

$$f'(x_n) \simeq \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}. \tag{3.34}$$

This approximation comes directly from the definition of the derivative of $f(x)$ as the limit

$$f'(x) = \lim_{u \to x} \frac{f(u) - f(x)}{u - x}.$$

Substituting (3.34) into the algorithm (3.32) for Newton's method we obtain the secant method, namely:

$$x_{n+1} = x_n - f(x_n) \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}. \tag{3.35}$$

The graphical interpretation of the secant method is similar to that of Newton's method. The tangent line to the curve is replaced by the secant line (see Figure 3.9).

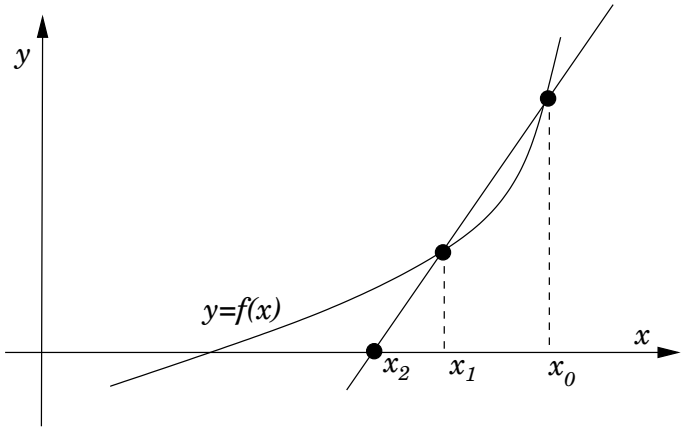

Figure 3.9: *Geometrical representation of the secant method: each new iterate is intersection of the secant to the graph of $f(x)$ with the x-axis.*

Remark 1 - This method requires two initial guesses: x_0 and x_1.
Remark 2 - The convergence of the secant method cannot be analysed using the contraction mapping theorems that we have studied because the method cannot be put in the form $x_{n+1} = g(x_n)$: each new iterate is a function of the previous **two**: $x_{n+1} = g(x_n, x_{n-1})$.
Remark 3 - It is possible to show that the error in the approximation decreases asymptotically as

$$|e_{n+1}| \sim |e_n|^{(1+\sqrt{5})/2}.$$

Since $(1 + \sqrt{5})/2 \simeq 1.62$ the secant method is a super-linear (i.e. faster than linear), but slower than quadratic convergence. This would suggests that the secant method is slower than Newton's method in reaching a given accuracy. However, Newton's method requires two function evaluations per iteration step, while the secant method requires only one. Therefore, when comparing the execution speed of the two methods we should/could compare *two* steps of the secant methods with one step of Newton's method. The rate of decrease of the error in two steps of the secant method is $1.62^2 \simeq 2.6$ which is better than the convergence rate of Newton's method.

3.5 Systems of nonlinear equations

3.5.1 Introduction

Finding the solutions of a systems of nonlinear equations,

$$\begin{cases} f_1(x_1, x_2, \ldots, x_n) = 0, \\ f_2(x_1, x_2, \ldots, x_n) = 0, \\ \vdots \\ f_n(x_1, x_2, \ldots, x_n) = 0, \end{cases} \quad \Leftrightarrow \quad \boldsymbol{f}(\boldsymbol{x}) = 0,$$

is considerably more difficult that solving a single nonlinear equation or solving a system of linear equations:

1. The system may have *no solution*, like for example

$$\begin{cases} x_1^2 + 2x_1x_2 + x_2^2 = 3, \\ x_1^3 + 3x_1^2x_2 + 3x_1x_2^2 + x_2^3 = 4, \end{cases} \Longleftrightarrow \begin{cases} (x_1 + x_2)^2 = 3, \\ (x_1 + x_2)^3 = 4, \end{cases}$$

   or have *no real solutions*, like for example,

$$\begin{cases} x_1^2 + x_2^2 = -5, \\ x_1^2 - 3\,x_2^2 = 11, \end{cases} \Longleftrightarrow \begin{cases} x_1 = \pm\mathrm{i}, \\ x_2 = \pm 2\mathrm{i}, \end{cases}$$

   or have a *unique solution*

$$\begin{cases} x_1^2 + x_2^2 = 0, \\ \cos(x_1x_2) = 1, \end{cases} \Longleftrightarrow \begin{cases} x_1 = 0, \\ x_2 = 0, \end{cases}$$

   or have *many solutions*

$$\begin{cases} x_1^2 + x_2^2 = 1, \\ \cos[\pi(x_1^2 + x_2^2)] + x_1^2 + x_2^2 = 0, \end{cases} \Longleftrightarrow \begin{array}{l} \text{all } x_1 \text{ and } x_2 \text{ that belong to} \\ \text{the circle } x_1^2 + x_2^2 = 1. \end{array}$$

2. If the system is large, even the existence of a solution is quite unclear.

3. For many real world problems the methods used to find the solutions can be rather ad hoc.

3.5.2 Contraction mapping

One can extend to more than one dimension the theorems on contraction mapping that have been discussed so far. The contraction map is now a set of n nonlinear functions in n variables,

$$\boldsymbol{x}_{n+1} = \boldsymbol{g}(\boldsymbol{x}_n).$$

Essentially everything carries over by replacing scalars with vectors and absolute values with norms.

We define an *interval* in $\mathbb{R}^n$ as a set $I = \{\boldsymbol{x} \in \mathbb{R}^n | \, a_j < x_j < b_j, \, j = 1, 2, \ldots, n\}$ where a_j and b_j are given constants. A *contraction map* in $\mathbb{R}^n$ is defined as:

Definition - A continuous map $\boldsymbol{g}(\boldsymbol{x})$ from an interval $I \subseteq \mathbb{R}^n$ into $\mathbb{R}^n$ is *contracting* if

1. the image of I is contained in I:

$$\boldsymbol{g}(I) \subseteq I \quad \Leftrightarrow \quad \boldsymbol{g}(x) \in I \quad \forall \boldsymbol{x} \in I.$$

2. the function $\boldsymbol{g}(\boldsymbol{x})$ is Lipschitz continuous in I with Lipschitz constant $L < 1$:

$$\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\| \quad \forall \boldsymbol{x}, \boldsymbol{y} \in I.$$

The convergence theorems are modified as follows:

Theorem 3.5.1 *If the function $\boldsymbol{g}(\boldsymbol{x})$ is continuous in $I \subseteq \mathbb{R}^n$ and $\boldsymbol{g}(I) \subseteq I$, then $\boldsymbol{g}(\boldsymbol{x})$ has at least one fixed point in I.*

Theorem 3.5.2 (Contraction mapping theorem in $\mathbb{R}^n$) *If $\boldsymbol{g}(\boldsymbol{x})$ is a contraction mapping in an interval $I \subseteq \mathbb{R}^n$ then there exists one and only one fixed point of the map in I.*

Theorem 3.5.3 *If $\boldsymbol{g}(\boldsymbol{x})$ is a differentiable contraction mapping in an interval $I \subseteq \mathbb{R}^n$, i.e.*

$$\left|\frac{\partial g_i}{\partial x_j}\right| \leq \frac{L}{n}, \quad \forall \boldsymbol{x} \in I, \quad \forall i, j \quad L < 1,$$

then there exists one and only one fixed point of the map in I.

Remark - The condition on the derivative can be relaxed somewhat. For example the theorem holds if $\|J(\boldsymbol{x})\|_\infty \leq L$, where $J(\boldsymbol{x})$ is the Jacobian matrix of the map $\boldsymbol{g}(\boldsymbol{x})$.

Theorem 3.5.4 *Let $I \subseteq \mathbb{R}^n$ be an interval in $\mathbb{R}^n$ and suppose that $\boldsymbol{g}(\boldsymbol{x})$ is a contraction mapping in I with Lipschitz constant $L < 1$.. Then for arbitrary $\boldsymbol{x}_0 \in I$, the sequence $\boldsymbol{x}_n = \boldsymbol{g}(\boldsymbol{x}_{n-1})$, $n = 1, 2, \ldots$ converges to the unique fixed point, $\boldsymbol{s}$, of the map. Moreover, if the error $\boldsymbol{e}_n$ at the n-th stage is defined by $\boldsymbol{e}_n = \boldsymbol{x}_n - \boldsymbol{s}$ then*

$$\|\boldsymbol{e}_n\|_\infty \leq \frac{L^n}{1-L}\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|_\infty.$$

Exercise - Show that

$$\boldsymbol{g}(\boldsymbol{x}) = \begin{cases} g_1(x_1, x_2, x_3) = \dfrac{1}{3}\cos(x_2 x_3) + \dfrac{1}{6}, \\ g_2(x_1, x_2, x_3) = \dfrac{1}{9}\sqrt{x_1^2 + \sin(x_3) + 1.06} - 0.1, \\ g_3(x_1, x_2, x_3) = -\dfrac{1}{20}e^{-x_1 x_2} - \left(\dfrac{10\pi - 3}{60}\right), \end{cases} \tag{3.36}$$

satisfies all the conditions of theorem 3.5.3 in $-1 < x_i < 1$.

Remark 1 - In practice it may be rather tough to prove that $g(I) \subseteq I$.
Remark 2 - There is a "Gauss-Seidel" version of this method, where each iterate is used as soon as it becomes available.
Remark 3 - The analogy with linear systems extends to the S.O.R. method. It is often hard to get the direct iteration to converge. However, the convergence can be helped by using "relaxation" (the opposite of S.O.R.). This method introduces an "under-relaxation" parameter $\omega < 1$ in the iteration. Assuming that we know the iterate $\boldsymbol{x}_n$ we compute a first estimate of the iterate $n+1$, $\hat{\boldsymbol{x}}_{n+1}$ using the iteration map $\boldsymbol{g}(\boldsymbol{x})$:

$$\hat{\boldsymbol{x}}_{n+1} = \boldsymbol{g}(\boldsymbol{x}_n).$$

Use this estimate to obtain $\boldsymbol{x}_{n+1}$ according to

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \omega(\hat{\boldsymbol{x}}_{n+1} - \boldsymbol{x}_n).$$

This procedure effectively multiplies the derivatives of the map $\boldsymbol{g}(\boldsymbol{x})$ by $\omega < 1$. If $\omega = 1$ this procedure reduces to the standard contraction mapping.

The drawback of this method is that sometimes ω has to be made so small that convergence is slow and too many iterations are needed.

3.5.3 Newton's method

In discussing Newton's method we shall use only 2×2 systems, but most of what we say generalises immediately to $n \times n$ systems. The model problem that we

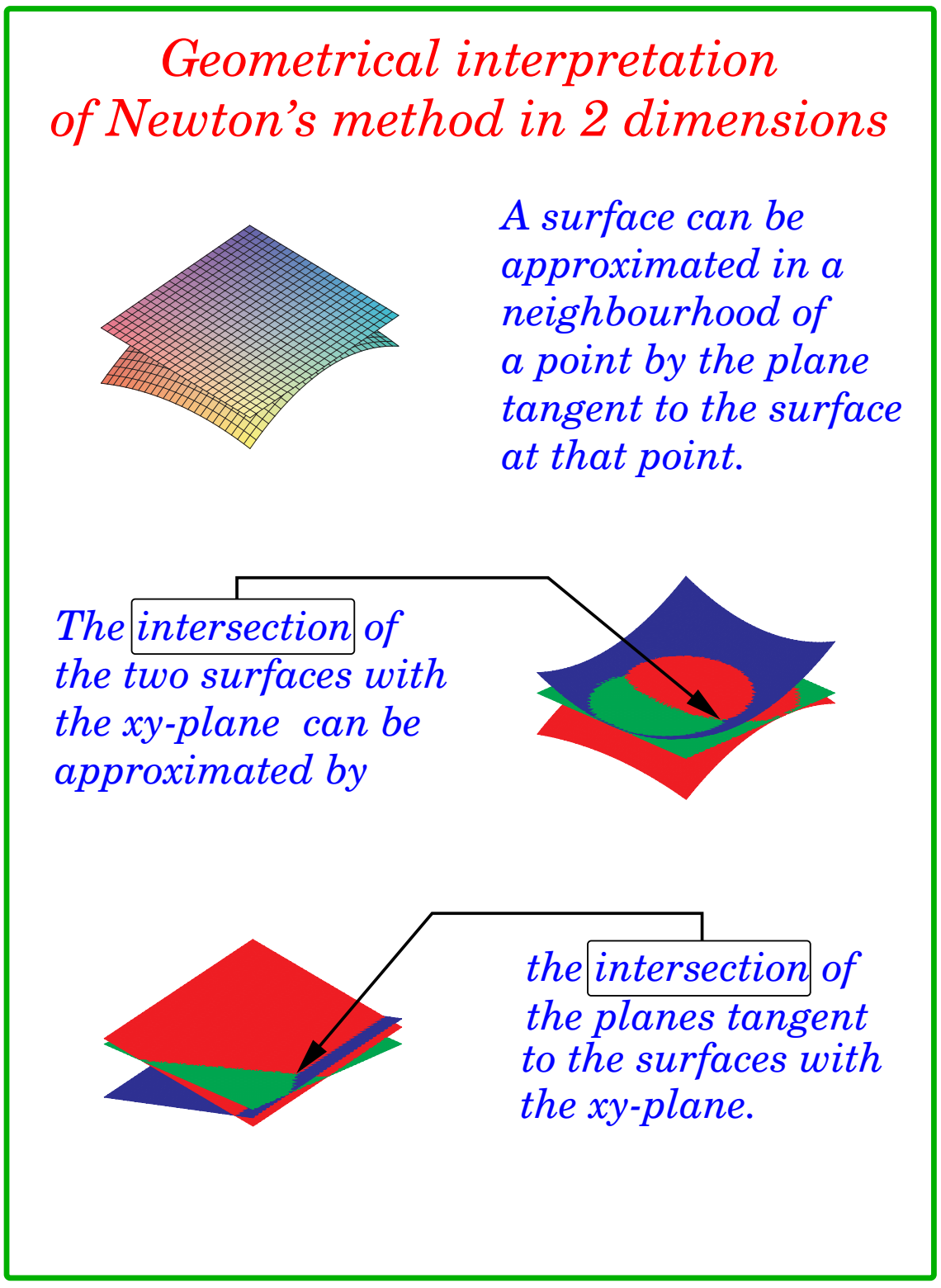

Figure 3.10: *Geometrical meaning of Newton's method to solve a system of two nonlinear equations in two variables. The solutions of the system are the intersections of the two surfaces with the xy-plane.*

wish to solve is

$$\begin{cases} f_1(x,y) = 0, \\ f_2(x,y) = 0, \end{cases} \qquad \Longleftrightarrow \qquad \boldsymbol{f}(\boldsymbol{x}) = 0.$$

The easiest way to obtain an expression for Newton's method in two dimensions is based on the its geometrical interpretation. We start by recapping the one dimension case: suppose that we wish to solve the equation $f(x) = 0$ and that we have evaluated the function at x_n. The next iterate, x_{n+1} is the intersection of the straight line tangent to the graph of $f(x)$ at $(x_n, f(x_n))$ with the x axis. The equation of this line is given by the first two terms of the Taylor expansion of the function $f(x)$ around x_n:

$$y = f(x_n) + f'(x_n)(x - x_n).$$

It intersects the x axis at the point x_{n+1} such that $y = 0$, i.e.

$$0 = f(x_n) + f'(x_n)(x_{n+1} - x_n) \implies x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

The geometrical interpretation of Newton's method in two dimensions is that the iterate (x_{n+1}, y_{n+1}) is the common point where the planes tangent to the graph of $f_1(x,y)$ and $f_2(x,y)$ at (x_n, y_n) intersect the xy-plane, i.e. the plane $z = 0$ (see Figure 3.10). The plane tangent to the two surfaces are

$$z = f_1(x_n, y_n) + (x - x_n)\frac{\partial f_1}{\partial x} + (y - y_n)\frac{\partial f_1}{\partial y},$$
$$z = f_2(x_n, y_n) + (x - x_n)\frac{\partial f_2}{\partial x} + (y - y_n)\frac{\partial f_2}{\partial y}.$$

At the intersection with the xy-plane we have $z = 0$. Therefore the equations for the next iterate are

$$f_1(x_n, y_n) + (x_{n+1} - x_n)\frac{\partial f_1}{\partial x} + (y_{n+1} - y_n)\frac{\partial f_1}{\partial y} = 0, \tag{3.37}$$
$$f_2(x_n, y_n) + (x_{n+1} - x_n)\frac{\partial f_2}{\partial x} + (y_{n+1} - y_n)\frac{\partial f_2}{\partial y} = 0. \tag{3.38}$$

We introduce the Jacobian of the function $\boldsymbol{f}(\boldsymbol{x})$ at (x_n, y_n),

$$J(x_n, y_n) = \begin{pmatrix} \partial_x f_1(x_n, y_n) & \partial_y f_1(x_n, y_n) \\ \partial_x f_2(x_n, y_n) & \partial_y f_2(x_n, y_n) \end{pmatrix}$$

and write (3.37) and (3.38) in matrix notation as

$$J(x_n, y_n) \begin{pmatrix} x_{n+1} - x_n \\ y_{n+1} - y_n \end{pmatrix} = \begin{pmatrix} -f_1(x_n, y_n) \\ -f_2(x_n, y_n) \end{pmatrix},$$

so that the next iterate of Newton's method is

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} x_n \\ y_n \end{pmatrix} - J^{-1} \begin{pmatrix} -f_1(x_n, y_n) \\ -f_2(x_n, y_n) \end{pmatrix}. \tag{3.39}$$

Equation (3.39) generalises to n-dimensions as

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - [J(\boldsymbol{x}_n)]^{-1} \boldsymbol{f}(\boldsymbol{x}_n). \tag{3.40}$$

The scheme (3.40) is numerically not convenient, because it requires the computation of the inverse of the Jacobian. Therefore, one normally defines an additional variable

$$\boldsymbol{z} = \boldsymbol{x}_{n+1} - \boldsymbol{x}_n \implies J\boldsymbol{z} = -\boldsymbol{f}(\boldsymbol{x}_n),$$

and solves the linear problem for $\boldsymbol{z}$ using, for example, Gauss elimination with pivoting. Once $\boldsymbol{z}$ is known we can obtain

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \boldsymbol{z}.$$

Remark 1 - The initial guess is absolutely crucial and the method can be very "touchy". However, if we are close enough to the root and the Jacobian is non singular the convergence is quadratic.
Remark 2 - Newton's method is very computationally intensive: it requires the evaluation of n^2 derivatives and the solution of an $n \times n$ linear problem at each iteration. A multidimensional version of the secant method has been developed to reduce the number of computations per step (Broyden algorithm).

Further reading

Topics covered here are also covered in

- Chapter 7 of Linz & Wang, *Exploring Numerical Methods* (QA297 LIN),
- Chapter 3 of Kincaid & Cheney, *Numerical Analysis* (QA297 KIN),
- Chapters 1 and 4 of Süli & Mayers, *An Introduction to Numerical Analysis* (not in library – includes considerably more analysis of the convergence issues).