Chapter 2

Solving nonlinear equations in one variable

Contents

In this section we shall be concerned with finding a root of the equation

$$f(x) = 0, \tag{2.1}$$

where f is a nonlinear function of x — if it is linear ($ax + b = 0$) we can solve trivially. We shall only be concerned with finding real roots of (2.1) though, of course, there is no guarantee that all, or indeed any, of the roots have to be real.

For example, we might wish to solve equations of the form

$$\begin{aligned} x - 2^{-x} &= 0, \\ \mathrm{e}^x - x^2 + 3x - 2 &= 0, \\ x\cos x + 2x^2 + 3x - 2 &= 0. \end{aligned}$$

None of these equations have solutions that can be written in a nice closed form, and so numerical approach is the only way.

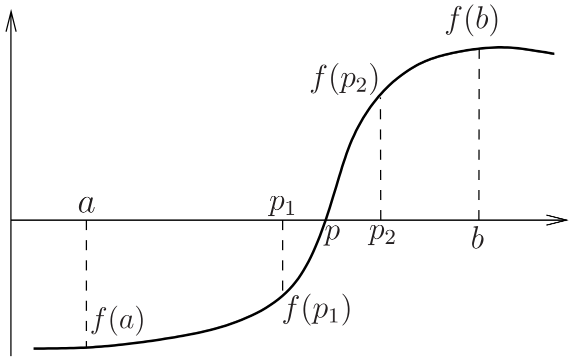
In this section, we shall describe iterative methods, i.e. methods that generate successive approximations to the exact solution, by applying a numerical algorithm to the previous approximation.

2.1 Bisection method

Suppose $f(x)$ is a continuous function on some interval $[a, b]$ with $f(a)$ and $f(b)$ of opposite sign[1]. Then, by the intermediate value theorem (IVT), $\exists p \in (a, b)$ with $f(p) = 0$. (For

[1] The product of two non-zero numbers $f(a)f(b) < 0$ if $f(a)$ and $f(b)$ are of opposite signs; $f(a)f(b) > 0$ if $f(a)$ has the same sign as $f(b)$.

simplicity, let's assume there exists only one root in (a,b).)



Bisection gives a simple and robust method for bracketing the root p. The algorithm works as follows:

Take the midpoint $p_1 = (a+b)/2$ as a first guess and calculate $f(p_1)$. If $f(p_1)$ and $f(b)$ are of opposite sign then the root lies in the interval $[p_1, b]$. (Conversely, if $f(p_1)$ has the opposite sign to $f(a)$ then the root lies in the interval $[a, p_1]$.)

This procedure can be repeated iteratively, e.g. taking a second guess $p_2 = (p_1 + b)/2$ from the example above.

At each iteration we halve the size of the interval $[a_n, b_n]$ that contains the root p and after n iterations of the procedure we have reduced the uncertainty in p down to,

$$E_n = |p - p_n| \leq \frac{b-a}{2^n}. \tag{2.2}$$

Example 2.1

Material covered in class. Please, see your lecture notes.

Advantages of the method :

i. provided the function is *continuous* on an interval $[a,b]$, with $f(a)f(b) < 0$, bisection is guaranteed to work (up to round-off error);

ii. the number of iterations needed to achieve a specific accuracy is known in advance.

Disadvantage of the method :

i. the method is slow to converge. (Reducing the error interval to 10^{-4} requires 16 iterations in the example 2.1);

ii. the errors in p_n and in $f(p_n)$ do not necessarily decrease between iterations. Note that, in the example 2.1, $p_2 = 1.5$ is closer to p (and $f(p_n)$ closer to 0) than the next 3 iterates. Also, no advantage is taken of intermediate good approximations.

2.2 Fixed point iteration

A fixed point of a function $g(x)$ is a value p such that $g(p) = p$.

Fixed point iteration procedure. Let g be a continuous function. If the sequence $p_n = g(p_{n-1})$ converges to p as $n \to \infty$ then $g(p) = p$. So, p is a stable fixed point of g.

Thus, provided p_0 is sufficiently close to p, we can use the simple iteration

$$p_n = g(p_{n-1}) \tag{2.3}$$

to find stable fixed points.

Root finding. A root finding problem, $f(x) = 0$, can be easily converted to a fixed point iteration by choosing, e.g., $g(x) = x - f(x)$, $g(x) = x + 3f(x)$, or $g(x) = \sqrt{x^2 + f(x)}$, etc.

It is important to note that there exists an infinite number of such functions g but the precise form of the g we choose will prove crucial for the convergence of the fixed point iteration.

Example 2.2

Material covered in class. Please, see your lecture notes.

Example 2.3

Material covered in class. Please, see your lecture notes.

Consider an iterate close to p, $p_n = p+\varepsilon$, say. Then, $p_{n+1} = g(p_n) = g(p+\varepsilon) = g(p)+\varepsilon g'(p)+O(\varepsilon^2) = p+\varepsilon g'(p)+O(\varepsilon^2)$ (using Taylor series). Thus $|p-p_n| = |\varepsilon|$ and $|p-p_{n+1}| \sim |\varepsilon||g'(p)|$, so if $|g'(p)| > 1$ then p_{n+1} is further away from p than p_n; there is no convergence.

Theorem 2.1 (Fixed point theorem)
If $g \in C[a,b]$ (i.e. is a continuous function on the interval $[a,b]$) and, $\forall x \in [a,b]$, $g(x) \in [a,b]$ then g has a fixed point in $[a,b]$ (existence).

If, in addition, $g'(x)$ exits on (a,b) and there exists a positive constant $K < 1$ such that, $\forall x \in (a,b)$, $|g'(x)| \leq K$ then

i. the fixed point is unique (uniqueness),

ii. for any $p_0 \in [a,b]$ the sequence $p_n = g(p_{n-1})$ converges to this unique fixed point $p \in [a,b]$ (stability).

Proof. The proof is in three parts. First, we prove that a fixed point exists. Second we prove that it is unique and third that the sequence must converge.

Existence. If $g(a) = a$ or $g(b) = b$ then a or b, respectively, is a fixed points. If not, then $g(a) > a$ and $g(b) < b$.

Define $h(x) = g(x) - x$. Then $h \in C[a,b]$ with $h(a) = g(a) - a > 0$ and $h(b) = g(b) - b < 0$. Therefore, the intermediate value theorem implies that there exists $p \in (a,b)$ such that $h(p) = 0$ i.e. $g(p) = p$, so p is a fixed point.

Uniqueness. Suppose that p and q are two distinct fixed points, i.e. $h(p) = h(q) = 0$, with $p < q$. Then by Rolle's theorem, $\exists c \in (p,q)$ such that $h'(c) = 0$.

But $h'(x) = g'(x) - 1 \leq K - 1 < 0$ since $g'(x) \leq K < 1$. This is a contradiction. Since the only assumption we have made is that $p \neq q$, then it follows that this assumption must be wrong. Thus, $p = q$ i.e. the fixed point is unique.

Stability. $|p_n - p| = |g(p_{n-1}) - g(p)|$. But the mean value theorem implies that $\exists \xi$ in the interval between p_{n-1} and p, subset of (a,b), such that $|g(p_{n-1}) - g(p)| = |g'(\xi)||(p_{n-1} - p)| \leq K|(p_{n-1} - p)|$.

Therefore, $|p_n - p| \leq K|(p_{n-1} - p)| \leq K^2|(p_{n-2} - p)| \ldots \leq K^n|(p_0 - p)|$, with $K < 1$. Hence, $|p_n - p| \leq K^n|(p_0 - p)| \to 0$ when $n \to \infty$ and the sequence (p_n) converges to the unique fixed point p.

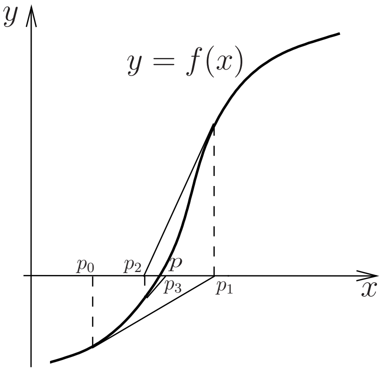
It is important to note that this theorem gives sufficient conditions for convergence, but they are not necessary.

In the example 2.2, when $g(x) = x - x^2/2 + 1$, $g'(x) = 1 - x$. So, $\forall x \in [1, 1.5]$, $|g'(x)| \leq 0.5$ and $g(x) \in [1.375, 1.5] \subset [1, 1.5]$ (subset). Hence, all the conditions for the fixed point theorem are satisfied and the iteration converges.

However, if $g(x) = x + 3x^2/2 - 3$ then $g'(x) = 1 + 3x > 1$ when $x > 0$. So, the fixed point $x = \sqrt{2}$ is unstable.

2.3 Newton-Raphson method

There are various ways of deriving this iterative procedure for solving nonlinear equations $f(x) = 0$ (see appendix C), but the graphical method is particularly instructive.



Suppose we have an initial guess p_o where f and f' are known. Then, p_1, the intersection of the x-axis with the tangent to f in p_0, is an improved estimate of the root.

$$f'(p_0) = \frac{f(p_0)}{p_0 - p_1} \Rightarrow p_1 = p_0 - \frac{f(p_0)}{f'(p_0)}.$$

Repeating this procedure leads to the general iterative scheme:

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \quad n \geq 1. \tag{2.4}$$

Obviously it is vital to have $f'(p_{n-1}) \neq 0$.

Relation with fixed point iteration. The Newton-Raphson method is a particular case of the fixed point iteration algorithm, with the iterative map $p_n = g(p_{n-1})$, where the mapping function $g(x) = x - f(x)/f'(x)$.

Example 2.4

Material covered in class. Please, see your lecture notes.

Advantages of Newton's method.

i. It is fast. (We shall quantify this later.)

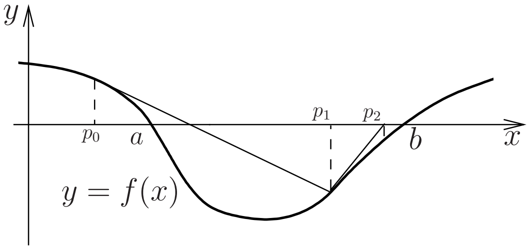
ii. It can be extended to multidimensional problem of finding, e.g. $f_1(x, y) = f_2(x, y) = 0$ for which bisection method cannot be used.

Disadvantages of Newton's method.

i. It requires evaluating the derivative, $f'(x)$, at each iteration.

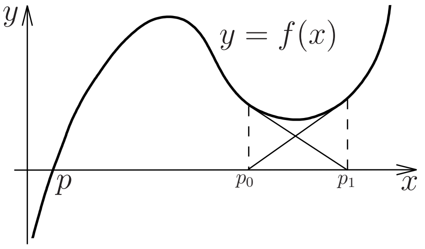
ii. It can fail if the initial guess is not sufficiently close to the solution, particularly if $f'(x) = 0$.

For instance, in the example 2.4 convergence is guaranteed. (Iterations converge to $+\sqrt{2}$ if $p_0 > 0$ and to $-\sqrt{2}$ if $p_0 < 0$.) However if f has the form



even though the initial guess p_0 is closest to the root $x = a$, (p_n) will converge to $x = b$.

Even more troublesome, a function of the form



has a root at $x = p$ but with the initial guess p_0 as shown, the iterations get caught in an endless loop, never converging.

2.4 Secant method

A solution to avoid the calculation of the derivative for Newton-Raphson method is to use the last two iterates to evaluate $f'(x)$.

By definition,

$$f'(p_{n-1}) = \lim_{x \to p_{n-1}} \frac{f(x) - f(p_{n-1})}{x - p_{n-1}},$$

but letting $x = p_{n-2}$ leads to the approximation

$$f'(p_{n-1}) \approx \frac{f(p_{n-2}) - f(p_{n-1})}{p_{n-2} - p_{n-1}}.$$

Rather than (2.4), the iteration scheme for the secant method is

$$p_n = p_{n-1} - f(p_{n-1}) \frac{p_{n-1} - p_{n-2}}{f(p_{n-1}) - f(p_{n-2})}. \tag{2.5}$$

Unlike Newton-Raphson, this method is a two-point method. (Iteration p_n uses both, p_{n-1} and p_{n-2}.) So, initially, two guesses are needed, p_0 and p_1.

Example 2.5

Material covered in class. Please, see your lecture notes.

Secant method is slightly slower than Newton's method, but it does not require the evaluation of a derivative. It does need two initial points but these do not have to straddle the root.

2.5 Rates of convergence

We mentioned, that the bisection method converges slowly, whilst, when the Newton-Raphson algorithm converges, it is fast. In this section we shall quantify the convergence rate of iteration schemes.

Suppose a numerical method producing a sequence of iterations (p_n) that converges to p. The method has order of convergence α if $|p_{n+1}-p| \sim K|p_n-p|^\alpha$ (i.e. $|p_{n+1}-p|$ is asymptotic to $K|p_n-p|^\alpha$), for some $K>0$. Or equivalently if

$$\lim_{n\to\infty} \frac{|p_{n+1}-p|}{|p_n-p|^\alpha} = K.$$

When $\alpha = 1$ the convergence is linear, and when $\alpha = 2$ it is quadratic. If the error of an iterative algorithm is $O(\varepsilon)$ at iteration n then, for a linear methods, it remains $O(\varepsilon)$ at iteration $n+1$, but for a quadratic method the error becomes $O(\varepsilon^2)$. Thus, higher order methods generally converge more rapidly than lower order ones.

2.5.1 Fixed point iteration

Let us consider a fixed point iteration $p_{n+1} = g(p_n)$ producing a sequence of iterations $(p_n) \to p$ as $n \to \infty$, such that $p = g(p)$ (since g is continuous).

Expand $g(p_n)$ as a Taylor series in powers of $(p_n - p)$. For some ξ_n in the interval between p and p_n,

$$g(p_n) = g(p) + (p_n - p)g'(\xi_n) \Rightarrow p_{n+1} = p + (p_n - p)g'(\xi_n).$$

Thus, $|p_{n+1}-p| = |g'(\xi_n)||p_n - p|$, and

$$\lim_{n\to\infty} \frac{|p_{n+1}-p|}{|p_n-p|} = |g'(p)| \Leftrightarrow |p_{n+1}-p| \sim |g'(p)||p_n-p|. \tag{2.6}$$

So, for $0 < |g'(p)| < 1$ fixed point iteration is linearly convergent.

2.5.2 Newton-Raphson

Equation (2.6) shows that the converge of fixed point iterations is best when $g'(p)$ is as small as possible, and preferably zero. Newton-Raphson, which is a fixed point iteration method with the mapping function $g(x) = x - f(x)/f'(x)$, achieves this.

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

Thus, provided $f'(p) \neq 0$ (i.e. p is a simple root, i.e. of multiplicity one), $g'(p) = 0$ since, by definition, $f(p) = 0$.

Expand $g(p_n)$ as a Taylor series in powers of $(p_n - p)$, up to second order. For some ξ_n in the interval between p and p_n,

$$g(p_n) = g(p) + (p_n-p)g'(p) + \frac{(p_n-p)^2}{2}g''(\xi_n) \Rightarrow p_{n+1} = p + \frac{(p_n-p)^2}{2}g''(\xi_n),$$

since $p_{n+1} = g(p_n)$, $g(p) = p$ and $g'(p) = 0$. Thus,

$$|p_{n+1}-p| = \frac{|g''(\xi_n)|}{2}|p_n-p|^2$$

implies

$$\lim_{n\to\infty} \frac{|p_{n+1}-p|}{|p_n-p|^2} = \frac{|g''(p)|}{2} \Leftrightarrow |p_{n+1}-p| \sim \frac{|g''(p)|}{2}|p_n-p|^2, \tag{2.7}$$

so the convergence of Newton's method is quadratic. (Note that $g''(p) = f''(p)/f'(p)$.)

However, in the case when $g'(p) \neq 0$ (i.e. if $f'(p) = 0$),

$$g(p_n) = g(p) + (p_n - p)g'(\xi_n) \Leftrightarrow p_{n+1} - p = (p_n - p)g'(\xi_n).$$

Thus,

$$\lim_{n\to\infty} \frac{|p_{n+1}-p|}{|p_n-p|} = |g'(p)|,$$

So, if $g'(p) \neq 0$ the convergence of Newton's methods is not quadratic but linear (i.e. the same as the general form of fixed point iteration).

2.5.3 Other methods

Bisection. At each step, the size of the interval that contains the root is halved, so $\max(E_n) = \max(E_{n-1})/2$, but the error does not necessarily decrease monotonically. However, if we regard the upper bounds of the errors as an estimate of the error, then bisection is linearly convergent.

Secant. This method requires two steps and is harder to analyse. However, it can be shown in a strict mathematical sens (see appendix D) that $|p_{n+1} - p| = K|p_n - p||p_{n-1} - p|$, giving an order of convergence of $(1 + \sqrt{5})/2 \approx 1.62$ (golden ratio). The convergence is super-linear, i.e. better than linear but poorer than quadratic.

2.6 Evaluation of the error

Generally, the exact value of the root of a function (or the exact value of a fixed point) is unknown. So, it is impossible the evaluate the error of iterative methods, $E_n = |p - p_n|$, but we can find an upper bound on the error using the difference between two iterates.

A sequence (p_n) is called contractive if there exists a positive constant $K < 1$ such that

$$|p_{n+2} - p_{n+1}| \leq K|p_{n+1} - p_n|, \forall n \in \mathbb{N}.$$

Theorem 2.2
If (p_n) is contractive with constant K then,

i. (p_n) is convergent,

$$\lim_{n\to\infty} p_n = p. \tag{2.8}$$

ii. $|p - p_n|$ is bounded above,

$$|p - p_n| \leq \frac{K}{1-K}|p_n - p_{n-1}|. \tag{2.9}$$

Hence, for K given, $|p_n - p_{n-1}|$ provides an the upper bound on $E_n = |p - p_n|$, the error at iteration n.