Contents 31

Numerical Methods of Approximation

Learning outcomes

In this Workbook you will learn about some numerical methods widely used in engineering applications.

You will learn how certain data may be modelled, how integrals and derivatives may be approximated and how estimates for the solutions of non-linear equations may be found.

Polynomial Approximations

31.1

Introduction

Polynomials are functions with useful properties. Their relatively simple form makes them an ideal candidate to use as approximations for more complex functions. In this second Workbook on Numerical Methods, we begin by showing some ways in which certain functions of interest may be approximated by polynomials.

Prerequisites

Before starting this Section you should ...

- revise material on maxima and minima of functions of two variables
- be familiar with polynomials and Taylor series

Learning Outcomes

On completion you should be able to ...

- interpolate data with polynomials
- find the least squares best fit straight line to experimental data

1. Polynomials

A polynomial in x is a function of the form

$$p(x) = a_0 + a_1x + a_2x^2 + \ldots a_nx^n \qquad (a_n \neq 0, \quad n \text{ a non-negative integer})$$

where $a_0, a_1, a_2, \ldots, a_n$ are constants. We say that this polynomial p has **degree** equal to n. (The degree of a polynomial is the highest power to which the argument, here it is x, is raised.) Such functions are relatively simple to deal with, for example they are easy to differentiate and integrate. In this Section we will show ways in which a function of interest can be approximated by a polynomial. First we briefly ensure that we are certain what a polynomial is.

Example 1

Which of these functions are polynomials in x? In the case(s) where f *is* a polynomial, give its degree.

(a) $f(x) = x^2 - 2 - \frac{1}{x}$, (b) $f(x) = x^4 + x - 6$, (c) $f(x) = 1$,

(d) $f(x) = mx + c$, m and c are constants. (e) $f(x) = 1 - x^6 + 3x^3 - 5x^3$

Solution

(a) This is not a polynomial because of the $\frac{1}{x}$ term (no negative powers of the argument are allowed in polynomials).
(b) This is a polynomial in x of degree 4.
(c) This is a polynomial of degree 0.
(d) This straight line function is a polynomial in x of degree 1 if $m \neq 0$ and of degree 0 if $m = 0$.
(e) This is a polynomial in x of degree 6.

Which of these functions are polynomials in x? In the case(s) where f *is* a polynomial, give its degree.

(a) $f(x) = (x-1)(x+3)$ (b) $f(x) = 1-x^7$ (c) $f(x) = 2+3e^x-4e^{2x}$

(d) $f(x) = \cos(x) + \sin^2(x)$

Your solution

Answer

(a) This function, like all quadratics, is a polynomial of degree 2.
(b) This is a polynomial of degree 7.
(c) and (d) These are not polynomials in x. Their Maclaurin expansions have infinitely many terms.

We have in fact already seen, in HELM 16, one way in which some functions may be approximated by polynomials. We review this next.

2. Taylor series

In HELM 16 we encountered Maclaurin series and their generalisation, Taylor series. Taylor series are a useful way of approximating functions by polynomials. The Taylor series expansion of a function $f(x)$ about $x = a$ may be stated

$$f(x) = f(a) + (x-a)f'(a) + \tfrac{1}{2}(x-a)^2 f''(a) + \tfrac{1}{3!}(x-a)^3 f'''(a) + \ldots.$$

(The special case called Maclaurin series arises when $a = 0$.)
The general idea when using this formula in practice is to consider only points x which are near to a. Given this it follows that $(x-a)$ will be small, $(x-a)^2$ will be even smaller, $(x-a)^3$ will be smaller still, and so on. This gives us confidence to simply neglect the terms beyond a certain power, or, to put it another way, to **truncate** the series.

Example 2

Find the Taylor polynomial of degree 2 about the point $x = 1$, for the function $f(x) = \ln(x)$.

Solution

In this case $a = 1$ and we need to evaluate the following terms

$$f(a) = \ln(a) = \ln(1) = 0, \qquad f'(a) = 1/a = 1, \qquad f''(a) = -1/a^2 = -1.$$

Hence

$$\ln(x) \approx 0 + (x-1) - \frac{1}{2}(x-1)^2 = -\frac{3}{2} + 2x - \frac{x^2}{2}$$

which will be reasonably accurate for x close to 1, as you can readily check on a calculator or computer. For example, for all x between 0.9 and 1.1, the polynomial and logarithm agree to at least 3 decimal places.

One drawback with this approach is that we need to find (possibly many) derivatives of f. Also, there can be some doubt over what is the best choice of a. The statement of Taylor series is an extremely useful piece of theory, but it can sometimes have limited appeal as a means of approximating functions by polynomials.

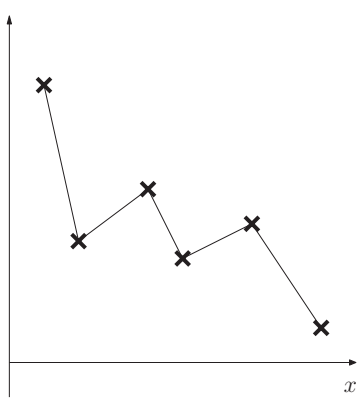
Next we will consider two alternative approaches.

3. Polynomial approximations - exact data

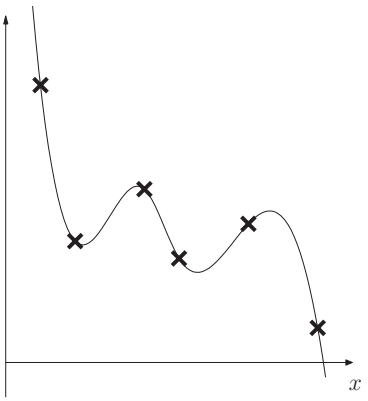
Here and in subsections 4 and 5 we consider cases where, rather than knowing an expression for the function, we have a list of point values. Sometimes it is good enough to find a polynomial that passes *near* these points (like putting a straight line through experimental data). Such a polynomial is an approximating polynomial and this case follows in subsection 4. Here and in subsection 5 we deal with the case where we want a polynomial to pass *exactly* through the given data, that is, an **interpolating polynomial**.

Lagrange interpolation

Suppose that we know (or choose to sample) a function f *exactly* at a few points and that we want to approximate how the function behaves between those points. In its simplest form this is equivalent to a dot-to-dot puzzle (see Figure 1(a)), but it is often more desirable to seek a curve that does not have"corners" in it (see Figure 1(b)).



(a) Linear, or "dot-to-dot", interpolation, with corners at all of the data points.



(b) A smoother interpolation of the data points.

Figure 1

Let us suppose that the data are in the form (x_1, f_1), (x_2, f_2), (x_3, f_3), ..., these are the points plotted as crosses on the diagrams above. (For technical reasons, and those of common sense, we suppose that the x-values in the data are all distinct.)
Our aim is to find a polynomial which passes exactly through the given data points. We want to find $p(x)$ such that

$$p(x_1) = f_1, \quad p(x_2) = f_2, \quad p(x_3) = f_3, \quad \ldots$$

There is a mathematical trick we can use to achieve this. We define **Lagrange polynomials** L_1, L_2, L_3, ... which have the following properties:

$$\begin{array}{rcll} L_1(x) & = & 1, & \text{at } x = x_1, \\ L_2(x) & = & 1, & \text{at } x = x_2, \\ L_3(x) & = & 1, & \text{at } x = x_3, \\ & \vdots & & \end{array} \qquad \begin{array}{rcll} L_1(x) & = & 0, & \text{at } x = x_2, x_3, x_4 \ldots \\ L_2(x) & = & 0, & \text{at } x = x_1, x_3, x_4 \ldots \\ L_3(x) & = & 0, & \text{at } x = x_1, x_2, x_4 \ldots \\ & \vdots & & \end{array}$$

Each of these functions acts like a filter which "turns off" if you evaluate it at a data point other than its own. For example if you evaluate L_2 at any data point other than x_2, you will get zero. Furthermore, if you evaluate any of these Lagrange polynomials at its own data point, the value you get is 1. These two properties are enough to be able to write down what $p(x)$ must be:

$$p(x) = f_1 L_1(x) + f_2 L_2(x) + f_3 L_3(x) + \dots$$

and this does work, because if we evaluate p at one of the data points, let us take x_2 for example, then

$$p(x_2) = f_1 \underbrace{L_1(x_2)}_{=0} + f_2 \underbrace{L_2(x_2)}_{=1} + f_3 \underbrace{L_3(x_2)}_{=0} + \dots = f_2$$

as required. The filtering property of the Lagrange polynomials picks out exactly the right f-value for the current x-value. Between the data points, the expression for p above will give a smooth polynomial curve.

This is all very well as long as we can work out what the Lagrange polynomials are. It is not hard to check that the following definitions have the right properties.

Key Point 1

Lagrange Polynomials

$$L_1(x) = \frac{(x-x_2)(x-x_3)(x-x_4)\dots}{(x_1-x_2)(x_1-x_3)(x_1-x_4)\dots}$$

$$L_2(x) = \frac{(x-x_1)(x-x_3)(x-x_4)\dots}{(x_2-x_1)(x_2-x_3)(x_2-x_4)\dots}$$

$$L_3(x) = \frac{(x-x_1)(x-x_2)(x-x_4)\dots}{(x_3-x_1)(x_3-x_2)(x_3-x_4)\dots}$$

and so on.

The numerator of $L_i(x)$ does not contain $(x - x_i)$.

The denominator of $L_i(x)$ does not contain $(x_i - x_i)$.

In each case the numerator ensures that the filtering property is in place, that is that the functions switch off at data points other than their own. The denominators make sure that the value taken at the remaining data point is equal to 1.

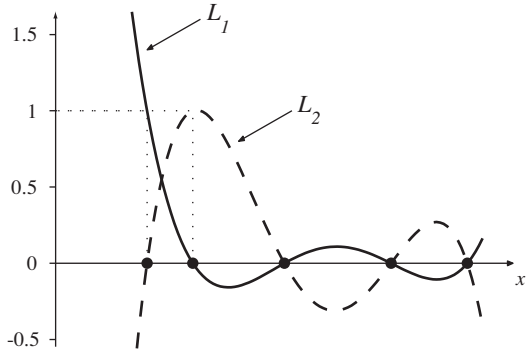


Figure 2

Figure 2 shows L_1 and L_2 in the case where there are five data points (the x positions of these data points are shown as large dots). Notice how both L_1 and L_2 are equal to zero at four of the data points and that $L_1(x_1) = 1$ and $L_2(x_2) = 1$.

In an implementation of this idea, things are simplified by the fact that we do not generally require an expression for $p(x)$. (This is good news, for imagine trying to multiply out all the algebra in the expressions for L_1, L_2, ...) What we *do* generally require is p evaluated at some specific value. The following Example should help show how this can be done.

Example 3

Let $p(x)$ be the polynomial of degree 3 which interpolates the data

x	0.8	1	1.4	1.6
$f(x)$	-1.82	-1.73	-1.40	-1.11

Evaluate $p(1.1)$.

Solution

We are interested in the Lagrange polynomials at the point $x = 1.1$ so we consider

$$L_1(1.1) = \frac{(1.1 - x_2)(1.1 - x_3)(1.1 - x_4)}{(x_1 - x_2)(x_1 - x_3)(x_1 - x_4)} = \frac{(1.1 - 1)(1.1 - 1.4)(1.1 - 1.6)}{(0.8 - 1)(0.8 - 1.4)(0.8 - 1.6)} = -0.15625.$$

Similar calculations for the other Lagrange polynomials give

$$L_2(1.1) = 0.93750, \quad L_3(1.1) = 0.31250, \quad L_4(1.1) = -0.09375,$$

and we find that our interpolated polynomial, evaluated at $x = 1.1$ is

$$\begin{aligned} p(1.1) &= f_1L_1(1.1) + f_2L_2(1.1) + f_3L_3(1.1) + f_4L_4(1.1) \\ &= -1.82 \times -0.15625 + -1.73 \times 0.9375 + -1.4 \times 0.3125 + -1.11 \times -0.09375 \\ &= -1.670938 \\ &= -1.67 \qquad \text{to the number of decimal places to which the data were given.} \end{aligned}$$

Key Point 2

Quote the answer only to the same number of decimal places as the given data (or to less places).

Let $p(x)$ be the polynomial of degree 3 which interpolates the data

x	0.1	0.2	0.3	0.4
$f(x)$	0.91	0.70	0.43	0.52

Evaluate $p(0.15)$.

Your solution

Answer

We are interested in the Lagrange polynomials at the point $x = 0.15$ so we consider

$$L_1(0.15) = \frac{(0.15 - x_2)(0.15 - x_3)(0.15 - x_4)}{(x_1 - x_2)(x_1 - x_3)(x_1 - x_4)} = \frac{(0.15 - 0.2)(0.15 - 0.3)(0.15 - 0.4)}{(0.1 - 0.2)(0.1 - 0.3)(0.1 - 0.4)} = 0.3125.$$

Similar calculations for the other Lagrange polynomials give

$$L_2(0.15) = 0.9375, \quad L_3(0.15) = -0.3125, \quad L_4(0.15) = 0.0625,$$

and we find that our interpolated polynomial, evaluated at $x = 0.15$ is

$$\begin{aligned} p(0.15) &= f_1L_1(0.15) + f_2L_2(0.15) + f_3L_3(0.15) + f_4L_4(0.15) \\ &= 0.91 \times 0.3125 + 0.7 \times 0.9375 + 0.43 \times -0.3125 + 0.52 \times 0.0625 \\ &= 0.838750 \\ &= 0.84, \quad \text{to 2 decimal places.} \end{aligned}$$

The next Example is very much the same as Example 3 and the Task above. Try not to let the specific application, and the slight change of notation, confuse you.

Example 4

A designer wants a curve on a diagram he is preparing to pass through the points

x	0.25	0.5	0.75	1
y	0.32	0.65	0.43	0.10

He decides to do this by using an interpolating polynomial $p(x)$. What is the y-value corresponding to $x = 0.8$?

Solution

We are interested in the Lagrange polynomials at the point $x = 0.8$ so we consider

$$L_1(0.8) = \frac{(0.8 - x_2)(0.8 - x_3)(0.8 - x_4)}{(x_1 - x_2)(x_1 - x_3)(x_1 - x_4)} = \frac{(0.8 - 0.5)(0.8 - 0.75)(0.8 - 1)}{(0.25 - 0.5)(0.25 - 0.75)(0.25 - 1)} = 0.032.$$

Similar calculations for the other Lagrange polynomials give

$$L_2(0.8) = -0.176, \quad L_3(0.8) = 1.056, \quad L_4(0.8) = 0.088,$$

and we find that our interpolated polynomial, evaluated at $x = 0.8$ is

$$\begin{aligned} p(0.8) &= y_1 L_1(0.8) + y_2 L_2(0.8) + y_3 L_3(0.8) + y_4 L_4(0.8) \\ &= 0.32 \times 0.032 + 0.65 \times -0.176 + 0.43 \times 1.056 + 0.1 \times 0.088 \\ &= 0.358720 \\ &= 0.36 \quad \text{to 2 decimal places.} \end{aligned}$$

In this next Task there are five points to interpolate. It therefore takes a polynomial of degree 4 to interpolate the data and this means we must use five Lagrange polynomials.

The hull drag f of a racing yacht as a function of the hull speed, v, is known to be

v	0.0	0.5	1.0	1.5	2.0
f	0.00	19.32	90.62	175.71	407.11

(Here, the units for f and v are N and m s^{-1}, respectively.)

Use Lagrange interpolation to fit this data and hence approximate the drag corresponding to a hull speed of 2.5 m s^{-1}.

Your solution

Answer

We are interested in the Lagrange polynomials at the point $v = 2.5$ so we consider

$$\begin{aligned} L_1(2.5) &= \frac{(2.5-v_2)(2.5-v_3)(2.5-v_4)(2.5-v_5)}{(v_1-v_2)(v_1-v_3)(v_1-v_4)(v_1-v_5)} \\ &= \frac{(2.5-0.5)(2.5-1.0)(2.5-1.5)(2.5-2.0)}{(0.0-0.5)(0.0-1.0)(0.0-1.5)(0.0-2.0)} = 1.0 \end{aligned}$$

Similar calculations for the other Lagrange polynomials give

$$L_2(2.5) = -5.0, \quad L_3(2.5) = 10.0, \quad L_4(2.5) = -10.0, \quad L_5(2.5) = 5.0$$

and we find that our interpolated polynomial, evaluated at $x = 2.5$ is

$$\begin{aligned} p(2.5) &= f_1L_1(2.5) + f_2L_2(2.5) + f_3L_3(2.5) + f_4L_4(2.5) + f_5L_5(2.5) \\ &= 0.00 \times 1.0 + 19.32 \times -5.0 + 90.62 \times 10.0 + 175.71 \times -10.0 + 407.11 \times 5.0 \\ &= 1088.05 \end{aligned}$$

This gives us the approximation that the hull drag on the yacht at 2.5 m s^{-1} is about 1100 N.

The following Example has time t as the independent variable, and two quantities, x and y, as dependent variables to be interpolated. We will see however that exactly the same approach as before works.

Example 5

An animator working on a computer generated cartoon has decided that her main character's right index finger should pass through the following (x, y) positions on the screen at the following times t

t	0	0.2	0.4	0.6
x	1.00	1.20	1.30	1.25
y	2.00	2.10	2.30	2.60

Use Lagrange polynomials to interpolate these data and hence find the (x, y) position at time $t = 0.5$. Give x and y to 2 decimal places.

Solution

In this case t is the independent variable, and there are two dependent variables: x and y. We are interested in the Lagrange polynomials at the time $t = 0.5$ so we consider

$$L_1(0.5) = \frac{(0.5-t_2)(0.5-t_3)(0.5-t_4)}{(t_1-t_2)(t_1-t_3)(t_1-t_4)} = \frac{(0.5-0.2)(0.5-0.4)(0.5-0.6)}{(0-0.2)(0-0.4)(0-0.6)} = 0.0625$$

Similar calculations for the other Lagrange polynomials give

$$L_2(0.5) = -0.3125, \quad L_3(0.5) = 0.9375, \quad L_4(0.5) = 0.3125$$

Solution (contd.)

These values for the Lagrange polynomials can be used for both of the interpolations we need to do. For the x-value we obtain

$$\begin{aligned} x(0.5) &= x_1L_1(0.5) + x_2L_2(0.5) + x_3L_3(0.5) + x_4L_4(0.5) \\ &= 1.00 \times 0.0625 + 1.20 \times -0.3125 + 1.30 \times 0.9375 + 1.25 \times 0.3125 \\ &= 1.30 \quad \text{to 2 decimal places} \end{aligned}$$

and for the y value we get

$$\begin{aligned} y(0.5) &= y_1L_1(0.5) + y_2L_2(0.5) + y_3L_3(0.5) + y_4L_4(0.5) \\ &= 2.00 \times 0.0625 + 2.10 \times -0.3125 + 2.30 \times 0.9375 + 2.60 \times 0.3125 \\ &= 2.44 \quad \text{to 2 decimal places} \end{aligned}$$

Error in Lagrange interpolation

When using Lagrange interpolation through n points $(x_1, f_1), (x_2, f_2), \ldots, (x_n, f_n)$ the error, in the estimate of $f(x)$ is given by

$$E(x) = \frac{(x - x_1)(x - x_2)\ldots(x - x_n)}{n!} f^{(n)}(\eta) \qquad \text{where } \eta \in [x, x_1, x_n]$$

N.B. The value of η is not known precisely, only the interval in which it lies. Normally x will lie in the interval $[x_1, x_n]$ (that's **interpolation**). If x lies outside the interval $[x_1, x_n]$ then that's called **extrapolation** and a larger error is likely.

Of course we will not normally know what f is (indeed no f may exist for experimental data). However, sometimes f can at least be estimated. In the following (somewhat artificial) example we will be told f and use it to check that the above error formula is reasonable.

Example 6

In an experiment to determine the relationship between power gain (G) and power output (P) in an amplifier, the following data were recorded.

P	5	7	8	11
G	0.00	1.46	2.04	3.42

(a) Use Lagrange interpolation to fit an appropriate quadratic, $q(x)$, to estimate the gain when the output is 6.5. Give your answer to an appropriate accuracy.

(b) Given that $G \equiv 10\log_{10}(P/5)$ show that the actual error which occurred in the Lagrange interpolation in (a) lies withing the theoretical error limits.

Solution

For a quadratic, $q(x)$, we need to fit **three** points and those most appropriate (nearest 6.5) are for P at $5, 7, 8$:

$$\begin{aligned} q(6.5) &= \frac{(6.5-7)(6.5-8)}{(5-7)(5-8)} \times 0.00 \\ &+ \frac{(6.5-5)(6.5-8)}{(7-5)(7-8)} \times 1.46 \\ &+ \frac{(6.5-5)(6.5-7)}{(8-5)(8-7)} \times 2.04 \\ &= 0 + 1.6425 - 0.5100 \\ &= 1.1325 \qquad \text{working to } 4 \text{ d.p.} \\ &\approx 1.1 \qquad \text{(rounding to sensible accuracy)} \end{aligned}$$

(b) We use the error formula

$$E(x) = \frac{(x-x_1)\ldots(x-x_n)}{n!} f^{(n)}(\eta), \qquad \eta \in [x, x_1, \ldots, x_n]$$

Here $f(x) \equiv G(x) = \log_{10}(P/5)$ and $n = 3$:

$$\begin{aligned} \frac{d(\log_{10}(P/5))}{dP} &= \frac{d(\log_{10}(P) - \log_{10}(5))}{dP} = \frac{d(\log_{10}(P))}{dP} \\ &= \frac{d}{dP}\left(\frac{\ln P}{\ln 10}\right) = \frac{1}{\ln 10}\frac{1}{P} \end{aligned}$$

So $\dfrac{d^3}{dP^3}\left(\log_{10}(P/5)\right) = \dfrac{1}{\ln 10}\dfrac{2}{P^3}$.

Substituting for $f^{(3)}(\eta)$:

$$\begin{aligned} E(6.5) &= \frac{(6.5-6)(6.5-7)(6.5-8)}{6} \times \frac{10}{\ln 10} \times \frac{2}{\eta^3}, \qquad \eta \in [5, 8] \\ &= \frac{1.6286}{\eta^3} \qquad \eta \in [5, 8] \end{aligned}$$

Taking $\eta = 5$: $\qquad E_{\text{max}} = 0.0131$

Taking $\eta = 8$: $\qquad E_{\text{min}} = 0.0031$

Taking $x = 6.5$: $\qquad E_{\text{actual}} = G(6.5) - q(6.5) = 10\log_{10}(6.5/5) - 1.1325$

$$= 1.1394 - 1.1325$$

$$= 0.0069$$

The theory is satisfied because $E_{\text{min}} < E_{\text{actual}} < E_{\text{max}}$.

(a) Use Lagrange interpolation to estimate $f(8)$ to appropriate accuracy given the table of values below, by means of the appropriate cubic interpolating polynomial

x	2	5	7	9	10
$f(x)$	0.980067	0.8775836	0.764842	0.621610	0.540302

Your solution

Answer

The most appropriate cubic passes through x at $5, 7, 9, 10$

$x = 8 \quad x_1 = 5, \quad x_2 = 7, \quad x_3 = 9, \quad x_4 = 10$

$$\begin{aligned} p(8) &= \frac{(8-7)(8-9)(8-10)}{(5-7)(5-9)(5-10)} \times 0.877583 \\ &+ \frac{(8-5)(8-9)(8-10)}{(7-5)(7-9)(7-10)} \times 0.764842 \\ &+ \frac{(8-5)(8-7)(8-10)}{(9-5)(9-7)(9-10)} \times 0.621610 \\ &+ \frac{(8-5)(8-7)(8-9)}{(10-5)(10-7)(10-9)} \times 0.540302 \\ &= -\frac{1}{20} \times 0.877583 + \frac{1}{2} \times 0.764842 + \frac{3}{4} \times 0.621610 - \frac{1}{5} \times 0.540302 \\ &= 0.6966689 \end{aligned}$$

Suitable accuracy is 0.6967 (rounded to 4 d.p.).

(b) Given that the table in (a) represents $f(x) \equiv \cos(x/10)$, calculate theoretical bounds for the estimate obtained:

Your solution

Answer

$$E(8) = \frac{(8-5)(8-7)(8-9)(8-10)}{4!} f^{(4)}(\eta), \qquad 5 \leq \eta \leq 10$$

$$f(\eta) = \cos\left(\frac{\eta}{10}\right) \qquad \text{so} \qquad f^{(4)}(\eta) = \frac{1}{10^4}\cos\left(\frac{\eta}{10}\right)$$

$$E(8) = \frac{1}{4 \times 10^4}\cos\left(\frac{\eta}{10}\right), \qquad \eta \in [5, 10]$$

$$E_{\text{min}} = \frac{1}{4 \times 10^4}\cos(1) \qquad E_{\text{max}} = \frac{1}{4 \times 10^4}\cos(0.5)$$

This leads to

$$0.696689 + 0.000014 \leq \text{True Value} \leq 0.696689 + 0.000022$$

$$\Rightarrow \qquad 0.696703 \leq \text{True Value} \leq 0.696711$$

We can conclude that the True Value is 0.69670 or 0.69671 to 5 d.p. or 0.6967 to 4 d.p. (actual value is 0.696707).

4. Polynomial approximations - experimental data

You may well have experience in carrying out an experiment and then trying to get a straight line to pass as near as possible to the data plotted on graph paper. This process of adjusting a clear ruler over the page until it looks "about right" is fine for a rough approximation, but it is not especially scientific. Any software you use which provides a "best fit" straight line must obviously employ a less haphazard approach.

Here we show one way in which best fit straight lines may be found.

Best fit straight lines

Let us consider the situation mentioned above of trying to get a straight line $y = mx + c$ to be as near as possible to experimental data in the form (x_1, f_1), (x_2, f_2), (x_3, f_3), ...

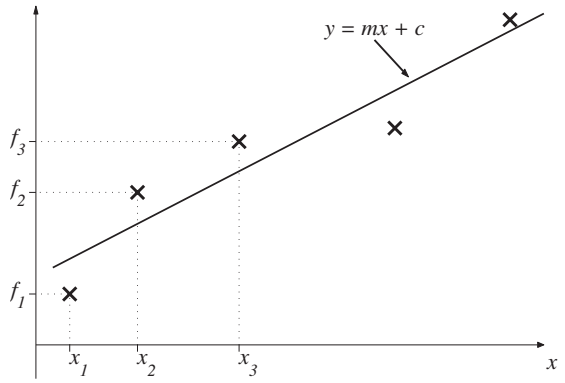


Figure 3

We want to minimise the overall distance between the crosses (the data points) and the straight line. There are a few different approaches, but the one we adopt here involves minimising the quantity

$$R = \underbrace{(mx_1 + c - f_1)^2}_{\substack{\text{vertical distance between line and the point } (x_1, f_1)}} + \underbrace{(mx_2 + c - f_2)^2}_{\text{second data point distance}} + \underbrace{(mx_3 + c - f_3)^2}_{\text{third data point distance}} + \ldots$$

$$= \sum \Big(mx_n + c - f_n\Big)^2.$$

Each term in the sum measures the vertical distance between a data point and the straight line. (Squaring the distances ensures that distances above and below the line do not cancel each other out. It is because we are minimising the distances squared that the straight line we will find is called the **least squares** best fit straight line.)

In order to minimise R we can imagine sliding the clear ruler around on the page until the line looks right; that is we can imagine varying the slope m and y-intercept c of the line. We therefore think of R as a function of the two variables m and c and, as we know from our earlier work on maxima and minima of functions, the minimisation is achieved when

$$\frac{\partial R}{\partial c} = 0 \qquad \text{and} \qquad \frac{\partial R}{\partial m} = 0.$$

(We know that this will correspond to a minimum because R has no maximum, for whatever value R takes we can always make it bigger by moving the line further away from the data points.) Differentiating R with respect to m and c gives

$$\begin{aligned}
\frac{\partial R}{\partial c} &= 2\left(mx_1 + c - f_1\right) + 2\left(mx_2 + c - f_2\right) + 2\left(mx_3 + c - f_3\right) + \ldots \\
&= 2\sum\Big(mx_n + c - f_n\Big) \qquad \text{and} \\
\frac{\partial R}{\partial m} &= 2\left(mx_1 + c - f_1\right)x_1 + 2\left(mx_2 + c - f_2\right)x_2 + 2\left(mx_3 + c - f_3\right)x_3 + \ldots \\
&= 2\sum\Big(mx_n + c - f_n\Big)\,x_n,
\end{aligned}$$

respectively. Setting both of these quantities equal to zero (and cancelling the factor of 2) gives a pair of simultaneous equations for m and c. This pair of equations is given in the Key Point below.

Key Point 3

The least squares best fit straight line to the experimental data

$$(x_1, f_1), \quad (x_2, f_2), \quad (x_3, f_3), \quad \ldots (x_n, f_n)$$

is

$$y = mx + c$$

where m and c are found by solving the pair of equations

$$\begin{aligned}
c\left(\sum_1^n 1\right) + m\left(\sum_1^n x_n\right) &= \sum_1^n f_n, \\
c\left(\sum_1^n x_n\right) + m\left(\sum_1^n x_n^2\right) &= \sum_1^n x_n f_n.
\end{aligned}$$

(The term $\displaystyle\sum_1^n 1$ is simply equal to the number of data points, n.)

Example 7

An experiment is carried out and the following data obtained:

x_n	0.24	0.26	0.28	0.30
f_n	1.25	0.80	0.66	0.20

Obtain the least squares best fit straight line, $y = mx + c$, to these data. Give c and m to 2 decimal places.

Solution

For a hand calculation, tabulating the data makes sense:

x_n	f_n	x_n^2	$x_n f_n$
0.24	1.25	0.0576	0.3000
0.26	0.80	0.0676	0.2080
0.28	0.66	0.0784	0.1848
0.30	0.20	0.0900	0.0600
1.08	2.91	0.2936	0.7528

The quantity $\sum 1$ counts the number of data points and in this case is equal to 4.
It follows that the pair of equations for m and c are:

$$\begin{aligned} 4c + 1.08m &= 2.91 \\ 1.08c + 0.2936m &= 0.7528 \end{aligned}$$

Solving these gives $c = 5.17$ and $m = -16.45$ and we see that the least squares best fit straight line to the given data is

$$y = 5.17 - 16.45x$$

Figure 4 shows how well the straight line fits the experimental data.

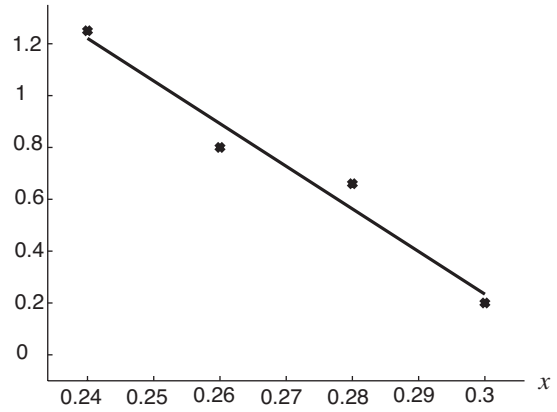


Figure 4

Example 8

Find the best fit straight line to the following experimental data:

x_n	0.00	1.00	2.00	3.00	4.00
f_n	1.00	3.85	6.50	9.35	12.05

Solution

In order to work out all of the quantities appearing in the pair of equations we tabulate our calculations as follows

	x_n	f_n	x_n^2	$x_n f_n$
	0.00	1.00	0.00	0.00
	1.00	3.85	1.00	3.85
	2.00	6.50	4.00	13.00
	3.00	9.35	9.00	28.05
	4.00	12.05	16.00	48.20
$\sum$	10.00	32.75	30.00	93.10

The quantity $\sum 1$ counts the number of data points and is in this case equal to 5.

Hence our pair of equations is

$$5c + 10m = 32.95$$

$$10c + 30m = 93.10$$

Solving these equations gives $c = 1.03$ and $m = 2.76$ and this means that our best fit straight line to the given data is

$$y = 1.03 + 2.76x$$

An experiment is carried out and the data obtained are as follows:

x_n	0.2	0.3	0.5	0.9
f_n	5.54	4.02	3.11	2.16

Obtain the least squares best fit straight line, $y = mx + c$, to these data. Give c and m to 2 decimal places.

Your solution

Answer

Tabulating the data gives

	x_n	f_n	x_n^2	$x_n f_n$
	0.2	5.54	0.04	1.108
	0.3	4.02	0.09	1.206
	0.5	3.11	0.25	1.555
	0.9	2.16	0.81	1.944
$\sum$	1.9	14.83	1.19	5.813

The quantity $\sum 1$ counts the number of data points and in this case is equal to 4.
It follows that the pair of equations for m and c are:

$$\begin{aligned} 4c + 1.9m &= 14.83 \\ 1.9c + 1.19m &= 5.813 \end{aligned}$$

Solving these gives $c = 5.74$ and $m = -4.28$ and we see that the least squares best fit straight line to the given data is

$$y = 5.74 - 4.28x$$

Power output P of a semiconductor laser diode, operating at $35°$C, as a function of the drive current I is measured to be

I	70	72	74	76
P	1.33	2.08	2.88	3.31

(Here I and P are measured in mA and mW respectively.)

It is known that, above a certain threshold current, the laser power increases linearly with drive current. Use the least squares approach to fit a straight line, $P = mI + c$, to these data. Give c and m to 2 decimal places.

Your solution

Answer

Tabulating the data gives

	I	P	I^2	$I \times P$
	70	1.33	4900	93.1
	72	2.08	5184	149.76
	74	2.88	5476	213.12
	76	3.31	5776	251.56
	292	9.6	21336	707.54

The quantity $\sum 1$ counts the number of data points and in this case is equal to 4.
It follows that the pair of equations for m and c are:

$$\begin{aligned} 4c + 292m &= 9.6 \\ 292c + 21336m &= 707.54 \end{aligned}$$

Solving these gives $c = -22.20$ and $m = 0.34$ and we see that the least squares best fit straight line to the given data is

$$P = -22.20 + 0.34I.$$

5. Polynomial approximations - splines

We complete this Section by briefly describing another approach that can be used in the case where the data are exact.

Why are splines needed?

Fitting a polynomial to the data (using Lagrange polynomials, for example) works very well when there are a small number of data points. But if there were 100 data points it would be silly to try to fit a polynomial of degree 99 through all of them. It would be a great deal of work and anyway polynomials of high degree can be very oscillatory giving poor approximations between the data points to the underlying function.

What are splines?

Instead of using a polynomial valid for all x, we use one polynomial for $x_1 < x < x_2$, then a different polynomial for $x_2 < x < x_3$ then a different one again for $x_3 < x < x_4$, and so on.
We have already seen one instance of this approach in this Section. The "dot to dot" interpolation that we abandoned earlier (Figure 1(a)) is an example of a **linear spline**. There is a different straight line between each pair of data points.

The most commonly used splines are **cubic splines**. We use a different polynomial of degree three between each pair of data points. Let $s = s(x)$ denote a cubic spline, then

$$\begin{aligned}
s(x) &= a_1(x-x_1)^3 + b_1(x-x_1)^2 + c_1(x-x_1) + d_1 \qquad (x_1 < x < x_2)\\
s(x) &= a_2(x-x_2)^3 + b_2(x-x_2)^2 + c_2(x-x_2) + d_2 \qquad (x_2 < x < x_3)\\
s(x) &= a_3(x-x_3)^3 + b_3(x-x_3)^2 + c_3(x-x_3) + d_3 \qquad (x_3 < x < x_4)\\
&\vdots
\end{aligned}$$

And we need to find a_1, b_1, c_1, d_1, a_2, ... to determine the full form for the spline $s(x)$. Given the large number of quantities that have to be assigned (four for every pair of adjacent data points) it is possible to give s some very nice properties:

- $s(x_1) = f_1$, $s(x_2) = f_2$, $s(x_3) = f_3$, ... This is the least we should expect, as it simply states that s interpolates the given data.
- $s'(x)$ is continuous at the data points. This means that there are no "corners" at the data points - the whole curve is smooth.
- $s''(x)$ is continuous. This reduces the occurrence of points of inflection appearing at the data points and leads to a smooth interpolant.

Even with all of these requirements there are still two more properties we can assign to s. A **natural cubic spline** is one for which s'' is zero at the two end points. The natural cubic spline is, in some sense, the smoothest possible spline, for it minimises a measure of the curvature.

How is a spline found?

Now that we have described what a **natural cubic spline** is, we briefly describe how it is found. Suppose that there are N data points. For a natural cubic spline we require $s''(x_1) = s''(x_N) = 0$ and values of s'' taken at the other data points are found from the system of equations in Key Point 4.

Key Point 4

Cubic Spline Equations

$$\begin{bmatrix} k_2 & h_2 & & & \\ h_2 & k_3 & h_3 & & \\ & \ddots & \ddots & \ddots & \\ & & h_{N-3} & k_{N-2} & h_{N-2} \\ & & & h_{N-2} & k_{N-1} \end{bmatrix} \begin{bmatrix} s''(x_2) \\ s''(x_3) \\ \vdots \\ s''(x_{N-2}) \\ s''(x_{N-1}) \end{bmatrix} = \begin{bmatrix} r_2 \\ r_3 \\ \vdots \\ r_{N-2} \\ r_{N-1} \end{bmatrix}$$

in which

$$h_1 = x_2 - x_1, \quad h_2 = x_3 - x_2, \quad h_3 = x_4 - x_3, \quad h_4 = x_5 - x_4, \ldots$$

$$k_2 = 2(h_1 + h_2), \quad k_3 = 2(h_2 + h_3), \quad k_4 = 2(h_3 + h_4), \ldots$$

$$r_2 = 6\left(\frac{f_3 - f_2}{h_2} - \frac{f_2 - f_1}{h_1}\right), \quad r_3 = 6\left(\frac{f_4 - f_3}{h_3} - \frac{f_3 - f_2}{h_2}\right), \ldots$$

Admittedly the system of equations in Key Point 4 looks unappealing, but this is a "nice" system of equations. It was pointed out at the end of HELM 30 that some applications lead to systems of equations involving matrices which are **strictly diagonally dominant**. The matrix above is of that type since the diagonal entry is always twice as big as the sum of off-diagonal entries.

Once the system of equations is solved for the second derivatives s'', the spline s can be found as follows:

$$a_i = \frac{s''(x_{i+1}) - s''(x_i)}{6h_i}, \quad b_i = \frac{s''(x_i)}{2}, \quad c_i = \frac{f_{i+1} - f_i}{h_i} - \left(\frac{s''(x_{i+1}) + 2s''(x_i)}{6}\right) h_i, \quad d_i = f_i$$

We now present an Example illustrating this approach.

Example 9

Find the natural cubic spline which interpolates the data

x_j	1	3	5	8
f_j	0.85	0.72	0.34	0.67

Solution

In the notation now established we have $h_1 = 2$, $h_2 = 2$ and $h_3 = 3$. For a *natural* cubic spline we require s'' to be zero at x_1 and x_4. Values of s'' at the other data points are found from the system of equations given in Key Point 4. In this case the matrix is just 2×2 and the pair of equations are:

$$h_1 \underbrace{s''(x_1)}_{=0} + 2(h_1 + h_2)s''(x_2) + h_2 s''(x_3) = 6\left(\frac{f_3 - f_2}{h_2} - \frac{f_2 - f_1}{h_1}\right)$$

$$h_2 s''(x_2) + 2(h_2 + h_3)s''(x_3) + h_3 \underbrace{s''(x_4)}_{=0} = 6\left(\frac{f_4 - f_3}{h_3} - \frac{f_3 - f_2}{h_2}\right)$$

In this case the equations become

$$\begin{pmatrix} 8 & 2 \\ 2 & 10 \end{pmatrix} \begin{pmatrix} s''(x_2) \\ s''(x_3) \end{pmatrix} = \begin{pmatrix} -0.75 \\ 1.8 \end{pmatrix}$$

Solving the coupled pair of equations leads to

$$s''(x_2) = -0.146053 \qquad s''(x_3) = 0.209211$$

We now find the coefficients a_1, b_1, etc. from the formulae and deduce that the spline is given by

$$\begin{aligned} s(x) &= -0.01217(x-1)^3 - 0.016316(x-1) + 0.85 \qquad (1 < x < 3) \\ s(x) &= 0.029605(x-3)^3 - 0.073026(x-3)^2 - 0.162368(x-3) + 0.72 \qquad (3 < x < 5) \\ s(x) &= -0.01162(x-5)^3 + 0.104605(x-5)^2 - 0.099211(x-5) + 0.34 \qquad (5 < x < 8) \end{aligned}$$

Figure 5 shows how the spline interpolates the data.

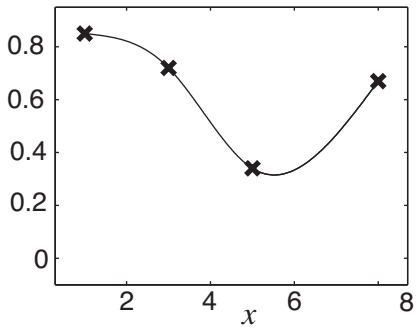


Figure 5

Task

Find the natural cubic spline which interpolates the data

x_j	1	2	3	5
f_j	0.1	0.24	0.67	0.91

Your solution

Answer

In the notation now established we have $h_1 = 1$, $h_2 = 1$ and $h_3 = 2$. For a *natural* cubic spline we require s'' to be zero at x_1 and x_4. Values of s'' at the other data points are found from the system of equations

$$\begin{aligned} h_1 \underbrace{s''(x_1)}_{=0} + 2(h_1+h_2)s''(x_2) + h_2 s''(x_3) &= 6\left(\frac{f_3-f_2}{h_2} - \frac{f_2-f_1}{h_1}\right) \\ h_2 s''(x_2) + 2(h_2+h_3)s''(x_3) + h_3 \underbrace{s''(x_4)}_{=0} &= 6\left(\frac{f_4-f_3}{h_3} - \frac{f_3-f_2}{h_2}\right) \end{aligned}$$

In this case the equations become

$$\begin{pmatrix} 4 & 1 \\ 1 & 6 \end{pmatrix} \begin{pmatrix} s''(x_2) \\ s''(x_3) \end{pmatrix} = \begin{pmatrix} 1.74 \\ -1.86 \end{pmatrix}$$

Solving the coupled pair of equations leads to $s''(x_2) = 0.534783$ $s''(x_3) = -0.399130$

We now find the coefficients a_1, b_1, etc. from the formulae and deduce that the spline is

$$\begin{aligned} s(x) &= 0.08913(x-1)^3 + 0.05087(x-1) + 0.1 \qquad (1 < x < 2) \\ s(x) &= -0.15565(x-2)^3 + 0.267391(x-2)^2 + 0.318261(x-2) + 0.24 \qquad (2 < x < 3) \\ s(x) &= 0.033261(x-3)^3 - 0.199565(x-3)^2 + 0.386087(x-3) + 0.67 \qquad (3 < x < 5) \end{aligned}$$

Exercises

1. A political analyst is preparing a dossier involving the following data

x	10	15	20	25
$f(x)$	9.23	8.41	7.12	4.13

   She interpolates the data with a polynomial $p(x)$ of degree 3 in order to find an approximation $p(22)$ to $f(22)$. What value does she find for $p(22)$?

2. Estimate $f(2)$ to an appropriate accuracy from the table of values below by means of an appropriate quadratic interpolating polynomial.

x	1	3	3.5	6
$f(x)$	99.8	295.5	342.9	564.6

3. An experiment is carried out and the data obtained as follows

x_n	2	3	5	7
f_n	2.2	5.4	6.5	13.2

   Obtain the least squares best fit straight line, $y = mx + c$, to these data. (Give c and m to 2 decimal places.)

4. Find the natural cubic spline which interpolates the data

x_j	2	4	5	7
f_j	1.34	1.84	1.12	0.02

Answers

1. We are interested in the Lagrange polynomials at the point $x = 22$ so we consider

$$L_1(22) = \frac{(22-x_2)(22-x_3)(22-x_4)}{(x_1-x_2)(x_1-x_3)(x_1-x_4)} = \frac{(22-15)(22-20)(22-25)}{(10-15)(10-20)(10-25)} = 0.056.$$

Similar calculations for the other Lagrange polynomials give

$$L_2(22) = -0.288, \quad L_3(22) = 1.008, \quad L_4(22) = 0.224,$$

and we find that our interpolated polynomial, evaluated at $x = 22$ is

$$\begin{aligned} p(22) &= f_1L_1(22) + f_2L_2(22) + f_3L_3(22) + f_4L_4(22) \\ &= 9.23 \times 0.056 + 8.41 \times -0.288 + 7.12 \times 1.008 + 4.13 \times 0.224 \\ &= 6.197 \\ &= 6.20, \quad \text{to 2 decimal places,} \end{aligned}$$

which serves as the approximation to $f(22)$.

2.

$$\begin{aligned} f(2) &= \frac{(2-1)(2-3)}{(3.5-1)(3.5-3)} \times 342.9 + \frac{(2-1)(2-3.5)}{(3-1)(3-3.5)} \times 295.5 + \frac{(2-3)(2-3.5)}{(1-3)(1-3.5)} \times 99.8 \\ &= -274.32 + 443.25 + 29.94 \\ &= 198.87 \end{aligned}$$

Estimate is 199 (to 3 sig. fig.)

3. We tabulate the data for convenience:

	x_n	f_n	x_n^2	x_nf_n
	2	2.2	4	4.4
	3	5.4	9	16.2
	5	6.5	25	32.5
	7	13.2	49	92.4
$\sum$	17	27.3	87	145.5

The quantity $\sum 1$ counts the number of data points and in this case is equal to 4.
It follows that the pair of equations for m and c are as follows:

$$\begin{aligned} 4c + 17m &= 27.3 \\ 17c + 87m &= 145.5 \end{aligned}$$

Solving these gives $c = -1.67$ and $m = 2.00$, to 2 decimal places, and we see that the least squares best fit straight line to the given data is

$$y = -1.67 + 2.00x$$

Answers

4. In the notation now established we have $h_1 = 2$, $h_2 = 1$ and $h_3 = 2$. For a *natural* cubic spline we require s'' to be zero at x_1 and x_4. Values of s'' at the other data points are found from the system of equations

$$\begin{aligned} h_1 \underbrace{s''(x_1)}_{=0} + 2(h_1+h_2)s''(x_2) + h_2 s''(x_3) &= 6\left(\frac{f_3-f_2}{h_2} - \frac{f_2-f_1}{h_1}\right) \\ h_2 s''(x_2) + 2(h_2+h_3)s''(x_3) + h_3 \underbrace{s''(x_4)}_{=0} &= 6\left(\frac{f_4-f_3}{h_3} - \frac{f_3-f_2}{h_2}\right) \end{aligned}$$

In this case the equations become

$$\begin{pmatrix} 6 & 1 \\ 1 & 6 \end{pmatrix} \begin{pmatrix} s''(x_2) \\ s''(x_3) \end{pmatrix} = \begin{pmatrix} -5.82 \\ 1.02 \end{pmatrix}$$

Solving the coupled pair of equations leads to

$$s''(x_2) = -1.026857 \qquad s''(x_3) = 0.341143$$

We now find the coefficients a_1, b_1, etc. from the formulae and deduce that the spline is given by

$$\begin{aligned} s(x) &= -0.08557(x-2)^3 + 0.592286(x-2) + 1.34 \qquad (2 < x < 4) \\ s(x) &= 0.228(x-4)^3 - 0.513429(x-4)^2 - 0.434571(x-4) + 1.84 \qquad (4 < x < 5) \\ s(x) &= -0.02843(x-5)^3 + 0.170571(x-5)^2 - 0.777429(x-5) + 1.12 \qquad (5 < x < 7) \end{aligned}$$

Numerical Integration 31.2

Introduction

In this Section we will present some methods that can be used to approximate integrals. Attention will be paid to how we ensure that such approximations can be guaranteed to be of a certain level of accuracy.

Prerequisites

Before starting this Section you should ...

- review previous material on integrals and integration

Learning Outcomes

On completion you should be able to ...

- approximate certain integrals
- be able to ensure that these approximations are of some desired accuracy

1. Numerical integration

The aim in this Section is to describe numerical methods for approximating integrals of the form

$$\int_a^b f(x)\,dx$$

One motivation for this is in the material on probability that appears in HELM 39. Normal distributions can be analysed by working out

$$\int_a^b \frac{1}{\sqrt{2\pi}}\,e^{-x^2/2}\,dx$$

for certain values of a and b. It turns out that it is not possible, using the kinds of functions most engineers would care to know about, to write down a function with derivative equal to $\frac{1}{\sqrt{2\pi}}\,e^{-x^2/2}$ so values of the integral are approximated instead. Tables of numbers giving the value of this integral for different interval widths appeared at the end of HELM 39, and it is known that these tables are accurate to the number of decimal places given. How can this be known? One aim of this Section is to give a possible answer to that question.

It is clear that, not only do we need a way of approximating integrals, but we also need a way of working out the accuracy of the approximations if we are to be sure that our tables of numbers are to be relied on.

In this Section we address both of these points, begining with a simple approximation method.

2. The simple trapezium rule

The first approximation we shall look at involves finding the area under a straight line, rather than the area under a curve f. Figure 6 shows it best.

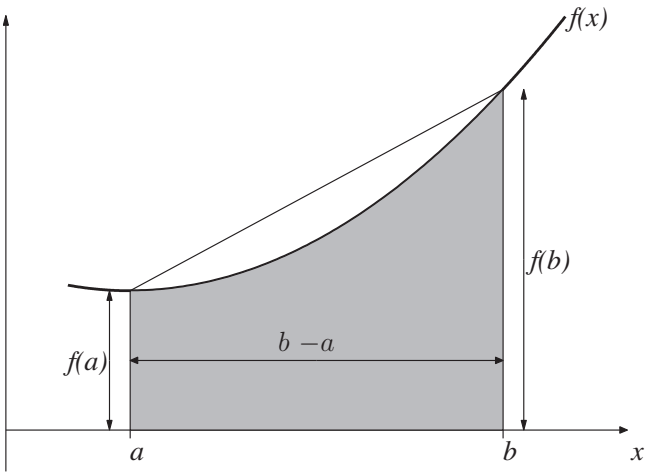


Figure 6

We approximate as follows

$$\begin{aligned}\int_a^b f(x)\,dx &= \text{grey shaded area}\\ &\approx \text{area of the trapezium surrounding the shaded region}\\ &= \text{width of trapezium} \times \text{average height of the two sides}\\ &= \frac{1}{2}(b-a)\Big(f(a)+f(b)\Big)\end{aligned}$$

Key Point 5

Simple Trapezium Rule

The simple trapezium rule for approximating $\int_a^b f(x)\,dx$ is given by approximating the area under the graph of f by the area of a trapezium.

The formula is:

$$\int_a^b f(x)\,dx \approx \frac{1}{2}(b-a)\big(f(a)+f(b)\big)$$

Or, to put it another way that may prove helpful a little later on,

$$\int_a^b f(x)\,dx \approx \frac{1}{2} \times (\text{interval width}) \times \Big(\ f(\text{left-hand end}) \quad + \quad f(\text{right-hand end})\ \Big)$$

Next we show some instances of implementing this method.

Example 10

Approximate each of these integrals using the simple trapezium rule

$$\text{(a)}\ \int_0^{\pi/4} \sin(x)\,dx \qquad \text{(b)}\ \int_1^2 e^{-x^2/2}\,dx \qquad \text{(c)}\ \int_0^2 \cosh(x)\,dx$$

Solution

$$\text{(a)}\ \int_0^{\pi/4} \sin(x)\,dx \approx \frac{1}{2}(b-a)(\sin(a)+\sin(b)) = \frac{1}{2}\left(\frac{\pi}{4}-0\right)\left(0+\frac{1}{\sqrt{2}}\right) = 0.27768,$$

$$\text{(b)}\ \int_1^2 e^{-x^2/2}\,dx \approx \frac{1}{2}(b-a)\left(e^{-a^2/2}+e^{-b^2/2}\right) = \frac{1}{2}(1-0)\left(e^{-1/2}+e^{-2}\right) = 0.37093,$$

$$\text{(c)}\ \int_0^2 \cosh(x)\,dx \approx \frac{1}{2}(b-a)\left(\cosh(a)+\cosh(b)\right) = \frac{1}{2}(2-0)\left(1+\cosh(2)\right) = 4.76220,$$

where all three answers are given to 5 decimal places.

It is important to note that, although we have given these integral approximations to 5 decimal places, this does not mean that they are accurate to that many places. We will deal with the accuracy of our approximations later in this Section. Next are some Tasks for you to try.

Approximate the following integrals using the simple trapezium method

(a) $\int_1^5 \sqrt{x}\, dx$ (b) $\int_1^2 \ln(x)\, dx$

Your solution

Answer

(a) $\int_1^5 \sqrt{x}\, dx \approx \frac{1}{2}(b-a)\left(\sqrt{a}+\sqrt{b}\right) = \frac{1}{2}(5-1)\left(1+\sqrt{5}\right) = 6.47214$

(b) $\int_1^2 \ln(x)\, dx \approx \frac{1}{2}(b-a)(\ln(a)+\ln(b)) = \frac{1}{2}(1-0)(0+\ln(2)) = 0.34657$

The answer you obtain for this next Task can be checked against the table of results in HELM 39 concerning the Normal distribution or in a standard statistics textbook.

Use the simple trapezium method to approximate $\int_0^1 \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx$

Your solution

Answer

We find that

$$\int_0^1 \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx \approx \frac{1}{2}(1-0)\frac{1}{\sqrt{2\pi}}(1+e^{-1/2}) = 0.32046$$

to 5 decimal places.

So we have a means of approximating $\int_a^b f(x)\, dx$. The question remains whether or not it is a *good* approximation.

How good is the simple trapezium rule?

We define e_T, the error in the simple trapezium rule to be the difference between the actual value of the integral and our approximation to it, that is

$$e_T = \int_a^b f(x)\,dx \quad - \quad \frac{1}{2}(b-a)\Big(f(a)+f(b)\Big)$$

It is enough for our purposes here to omit some theory and skip straight to the result of interest. In many different textbooks on the subject it is shown that

$$e_T = -\frac{1}{12}(b-a)^3 f''(c)$$

where c is some number between a and b. (The principal drawback with this expression for e_T is that we do not know what c is, but we will find a way to work around that difficulty later.)

It is worth pausing to ask what meaning we can attach to this expression for e_T. There are two factors which can influence e_T:

1. If $b-a$ is small then, clearly, e_T will most probably also be small. This seems sensible enough - if the integration interval is a small one then there is "less room" to accumulate a large error. (This observation forms part of the motivation for the composite trapezium rule discussed later in this Section.)

2. If f'' is small everywhere in $a < x < b$ then e_T will be small. This reflects the fact that we worked out the integral of a straight line function, instead of the integral of f. If f is a long way from being a straight line then f'' will be large and so we must expect the error e_T to be large too.

We noted above that the expression for e_T is less useful than it might be because it involves the unknown quantity c. We perform a trade-off to get around this problem. The expression above gives an exact value for e_T, but we do not know enough to evaluate it. So we replace the expression with one we *can* evaluate, but it will not be exact. We replace $f''(c)$ with a worst case value to obtain an **upper bound** on e_T. This worst case value is the largest (positive or negative) value that $f''(x)$ achieves for $a \leq x \leq b$. This leads to

$$|e_T| \leq \max_{a\leq x\leq b}\Big|f''(x)\Big|\,\frac{(b-a)^3}{12}.$$

We summarise this in Key Point 6.

Key Point 6

Error in the Simple Trapezium Rule

The error, $|e_T|$, in the simple trapezium approximation to $\displaystyle\int_a^b f(x)\,dx$ is bounded above by

$$\max_{a\leq x\leq b}\Big|f''(x)\Big|\,\frac{(b-a)^3}{12}$$

Example 11

Work out the error bound (to 6 decimal places) for the simple trapezium method approximations to

(a) $\displaystyle\int_0^{\pi/4} \sin(x)\,dx$ (b) $\displaystyle\int_0^{2} \cosh(x)\,dx$

Solution

In each case the trickiest part is working out the maximum value of $f''(x)$.

(a) Here $f(x) = \sin(x)$, therefore $f'(x) = -\cos(x)$ and $f''(x) = -\sin(x)$. The function $\sin(x)$ takes values between 0 and $\frac{1}{\sqrt{2}}$ when x varies between 0 and $\pi/4$. Hence

$$e_T < \frac{1}{\sqrt{2}} \times \frac{(\pi/4)^3}{12} = 0.028548 \quad \text{to 6 decimal places.}$$

(b) If $f(x) = \cosh(x)$ then $f''(x) = \cosh(x)$ too. The maximum value of $\cosh(x)$ for x between 0 and 2 will be $\cosh(2) = 3.762196$, to 6 decimal places. Hence, in this case,

$$e_T < (3.762196) \times \frac{(2-0)^3}{12} = 2.508130 \quad \text{to 6 decimal places.}$$

(In Example 11 we used a rounded value of $\cosh(2)$. To be on the safe side, it is best to round this number *up* to make sure that we still have an upper bound on e_T. In this case, of course, rounding up is what we would naturally do, because the seventh decimal place was a 6.)

Work out the error bound (to 5 significant figures) for the simple trapezium method approximations to

(a) $\displaystyle\int_1^{5} \sqrt{x}\,dx$ (b) $\displaystyle\int_1^{2} \ln(x)\,dx$

Your solution

(a)

Answer

If $f(x) = \sqrt{x} = x^{1/2}$ then $f'(x) = \frac{1}{2}x^{-1/2}$ and $f''(x) = -\frac{1}{4}x^{-3/2}$.

The negative power here means that f'' takes its biggest value in magnitude at the left-hand end of the interval $[1, 5]$ and we see that $\max_{1\leq x\leq 5}|f''(x)| = f''(1) = \frac{1}{4}$. Therefore

$$e_T < \frac{1}{4} \times \frac{4^3}{12} = 1.3333 \quad \text{to 5 s.f.}$$

Your solution

(b)

Answer

Here $f(x) = \ln(x)$ hence $f'(x) = 1/x$ and $f''(x) = -1/x^2$.

It follows then that $\max_{1\leq x\leq 2}|f''(x)| = 1$ and we conclude that

$$e_T < 1 \times \frac{1^3}{12} = 0.083333 \quad \text{to 5 s.f.}$$

One deficiency in the simple trapezium rule is that there is nothing we can do to improve it. Having computed an error bound to measure the quality of the approximation we have no way to go back and work out a better approximation to the integral. It would be preferable if there were a parameter we could alter to tune the accuracy of the method. The following approach uses the simple trapezium method in a way that allows us to improve the accuracy of the answer we obtain.

3. The composite trapezium rule

The general idea here is to split the interval $[a, b]$ into a sequence of N smaller subintervals of equal width $h = (b-a)/N$. Then we apply the simple trapezium rule to each of the subintervals.
Figure 7 below shows the case where $N = 2$ (and $\therefore h = \frac{1}{2}(b-a)$). To simplify notation later on we let $f_0 = f(a)$, $f_1 = f(a+h)$ and $f_2 = f(a+2h) = f(b)$.

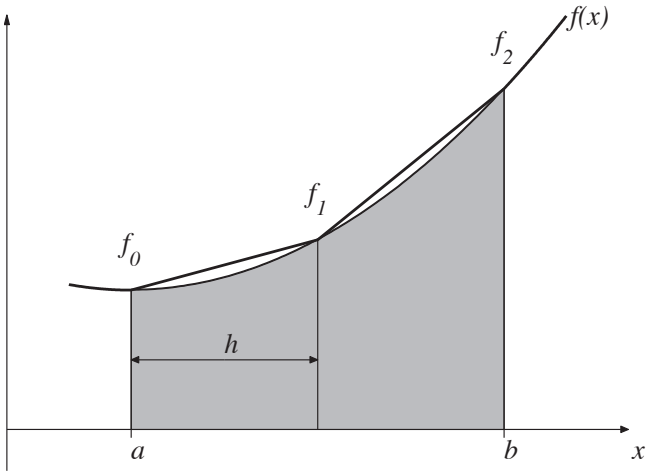


Figure 7

Applying the simple trapezium rule to each subinterval we get

$$\begin{aligned}\int_a^b f(x)\,dx &\approx \text{(area of first trapezium)} + \text{(area of second trapezium)} \\ &= \frac{1}{2}h(f_0+f_1) + \frac{1}{2}h(f_1+f_2) = \frac{1}{2}h\left(f_0 + 2f_1 + f_2\right)\end{aligned}$$

where we remember that the width of each of the subintervals is h, rather than the $b-a$ we had in the simple trapezium rule.

The next improvement will come from taking $N = 3$ subintervals (Figure 8). Here $h = \frac{1}{3}(b-a)$ is smaller than in Figure 7 above and we denote $f_0 = f(a)$, $f_1 = f(a+h)$, $f_2 = f(a+2h)$ and $f_3 = f(a+3h) = f(b)$. (Notice that f_1 and f_2 mean something different from what they did in the $N = 2$ case.)

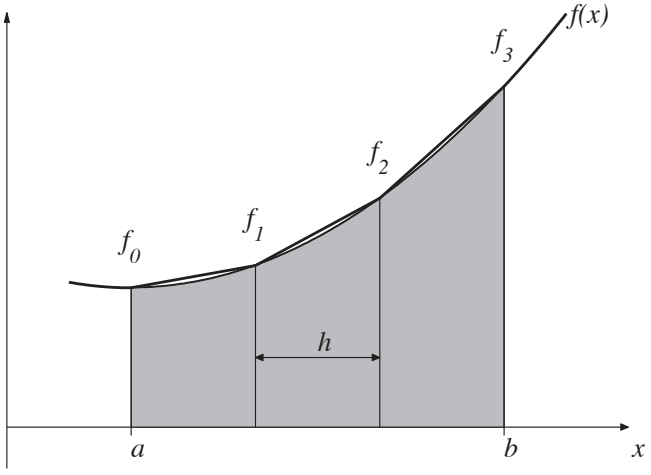


Figure 8

As Figure 8 shows, the approximation is getting closer to the grey shaded area and in this case we have

$$\begin{aligned}\int_a^b f(x)\,dx &\approx \frac{1}{2}h(f_0+f_1)+\frac{1}{2}h(f_1+f_2)+\frac{1}{2}h(f_2+f_3)\\ &= \frac{1}{2}h\left(f_0\ +2\left\{f_1+f_2\right\}+\ f_3\right).\end{aligned}$$

The pattern is probably becoming clear by now, but here is one more improvement. In Figure 9 $N=4$, $h=\frac{1}{4}(b-a)$ and we denote $f_0=f(a)$, $f_1=f(a+h)$, $f_2=f(a+2h)$, $f_3=f(a+3h)$ and $f_4=f(a+4h)=f(b)$.

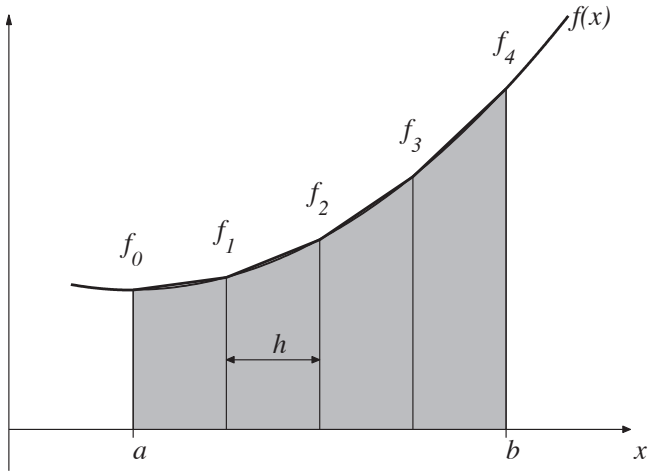


Figure 9

This leads to

$$\begin{aligned}\int_a^b f(x)\,dx &\approx \frac{1}{2}h(f_0+f_1)+\frac{1}{2}h(f_1+f_2)+\frac{1}{2}h(f_2+f_3)++\frac{1}{2}h(f_3+f_4)\\ &= \frac{1}{2}h\left(f_0\ +2\left\{f_1+f_2+f_3\right\}+\ f_4\right).\end{aligned}$$

We generalise this idea into the following Key Point.

Key Point 7

Composite Trapezium Rule

The composite trapezium rule for approximating $\int_a^b f(x)\,dx$ is carried out as follows:

1. Choose N, the number of subintervals,
2. $\int_a^b f(x)\,dx \approx \frac{1}{2}h\left(f_0 \ + 2\{f_1 + f_2 + \cdots + f_{N-1}\} + \ f_N\right)$,

   where

$$h = \frac{b-a}{N}, \qquad f_0 = f(a), \quad f_1 = f(a+h), \ldots, f_n = f(a+nh), \ldots,$$
$$\text{and} \quad f_N = f(a+Nh) = f(b).$$

Example 12

Using 4 subintervals in the composite trapezium rule, and working to 6 decimal places, approximate

$$\int_0^2 \cosh(x)\,dx$$

Solution

In this case $h = (2-0)/4 = 0.5$.
We require $\cosh(x)$ evaluated at five x-values and the results are tabulated below to 6 d.p.

x_n	$f_n = \cosh(x_n)$
0	1.000000
0.5	1.127626
1	1.543081
1.5	2.352410
2	3.762196

It follows that

$$\begin{aligned}\int_0^2 \cosh(x)\,dx &\approx \frac{1}{2}h\left(f_0 + f_4 + 2\{f_1 + f_2 + f_3\}\right)\\ &= \frac{1}{2}(0.5)\left(1 + 3.762196 + 2\{1.127626 + 1.543081 + 2.35241\}\right)\\ &= 3.452107\end{aligned}$$

Using 4 subintervals in the composite trapezium rule approximate

$$\int_1^2 \ln(x)\, dx$$

Your solution

Answer

In this case $h = (2-1)/4 = 0.25$.

We require $\ln(x)$ evaluated at five x-values and the results are tabulated below t0 6 d.p.

x_n	$f_n = \ln(x_n)$
1	0.000000
1.25	0.223144
1.5	0.405465
1.75	0.559616
2	0.693147

It follows that

$$\begin{aligned}\int_1^2 \ln(x)\, dx &\approx \frac{1}{2}h\left(f_0 + f_4 + 2\{f_1 + f_2 + f_3\}\right)\\ &= \frac{1}{2}(0.25)\left(0 + 0.693147 + 2\{0.223144 + 0.405465 + 0.559616\}\right)\\ &= 0.383700\end{aligned}$$

How good is the composite trapezium rule?

We can work out an upper bound on the error incurred by the composite trapezium method. Fortunately, all we have to do here is apply the method for the error in the simple rule over and over again. Let e_T^N denote the error in the composite trapezium rule with N subintervals. Then

$$\left|e_T^N\right| \leq \max_{\text{1st subinterval}} \left|f''(x)\right| \frac{h^3}{12} + \max_{\text{2nd subinterval}} \left|f''(x)\right| \frac{h^3}{12} + \ldots + \max_{\text{last subinterval}} \left|f''(x)\right| \frac{h^3}{12}$$

$$= \frac{h^3}{12} \underbrace{\left(\max_{\text{1st subinterval}} \left|f''(x)\right| + \max_{\text{2nd subinterval}} \left|f''(x)\right| + \ldots + \max_{\text{last subinterval}} \left|f''(x)\right| \right)}_{N \text{ terms}}.$$

This is all very well as a piece of theory, but it is awkward to use in practice. The process of working out the maximum value of $|f''|$ separately in each subinterval is very time-consuming. We can obtain a more user-friendly, if less accurate, error bound by replacing each term in the last bracket above with the biggest one. Hence we obtain

$$\left|e_T^N\right| \leq \frac{h^3}{12} \left(N \max_{a \leq x \leq b} \left|f''(x)\right| \right)$$

This upper bound can be rewritten by recalling that $Nh = b - a$, and we now summarise the result in a Key Point.

Key Point 8

Error in the Composite Trapezium Rule

The error, $\left|e_T^N\right|$, in the N-subinterval composite trapezium approximation to $\displaystyle\int_a^b f(x)\,dx$ is bounded above by

$$\max_{a \leq x \leq b} \left|f''(x)\right| \frac{(b-a)h^2}{12}$$

Note: the special case when $N = 1$ is the simple trapezium rule, in which case $b - a = h$ (refer to Key Point 6 to compare).

The formula in Key Point 8 can be used to decide how many subintervals to use to guarantee a specific accuracy.

Example 13

The function f is known to have a second derivative with the property that

$$|f''(x)| < 12$$

for x between 0 and 4.

Using the error bound given in Key Point 8 determine how many subintervals are required so that the composite trapezium rule used to approximate

$$\int_0^4 f(x)\,dx$$

can be guaranteed to be in error by less than $\frac{1}{2} \times 10^{-3}$.

Solution

We require that

$$12 \times \frac{(b-a)h^2}{12} < 0.0005$$

that is

$$4h^2 < 0.0005.$$

This implies that $h^2 < 0.000125$ and therefore $h < 0.0111803$.
Now $N = (b-a)/h = 4/h$ and it follows that

$$N > 357.7708$$

Clearly, N must be a whole number and we conclude that the smallest number of subintervals which *guarantees* an error smaller than 0.0005 is $N = 358$.

It is worth remembering that the error bound we are using here is a pessimistic one. We effectively use the same (worst case) value for $f''(x)$ all the way through the integration interval. Odds are that fewer subintervals will give the required accuracy, but the value for N we found here will guarantee a good enough approximation.

Next are two Tasks for you to try.

The function f is known to have a second derivative with the property that

$$|f''(x)| < 14$$

for x between -1 and 4.

Using the error bound given in Key Point 8 determine how many subintervals are required so that the composite trapezium rule used to approximate

$$\int_{-1}^{4} f(x)\,dx$$

can be guaranteed to have an error less than 0.0001.

Your solution

Answer

We require that

$$14 \times \frac{(b-a)h^2}{12} < 0.0001$$

that is

$$\frac{70h^2}{12} < 0.0001$$

This implies that $h^2 < 0.00001714$ and therefore $h < 0.0041404$.
Now $N = (b-a)/h = 5/h$ and it follows that

$$N > 1207.6147$$

Clearly, N must be a whole number and we conclude that the smallest number of subintervals which *guarantees* an error smaller than 0.00001 is $N = 1208$.

It is given that the function $e^{-x^2/2}$ has a second derivative that is never greater than 1 in absolute value.

(a) Use this fact to determine how many subintervals are required for the composite trapezium method to deliver an approximation to

$$\int_0^1 \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx$$

that is guaranteed to have an error less than $\frac{1}{2} \times 10^{-2}$.

(b) Find an approximation to the integral that is in error by less than $\frac{1}{2} \times 10^{-2}$.

Your solution

(a)

Answer

We require that $\dfrac{1}{\sqrt{2\pi}} \dfrac{(b-a)h^2}{12} < 0.005$. This means that $h^2 < 0.150398$ and therefore, since $N = 1/h$, it is necessary for $N = 3$ for the error bound to be less than $\pm\frac{1}{2} \times 10^{-2}$.

Your solution

(b)

Answer

To carry out the composite trapezium rule, with $h = \frac{1}{3}$ we need to evaluate $f(x) = \dfrac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$ at $x = 0,\ h,\ 2h,\ 1$. This evaluation gives

$$f(0) = f_0 = 0.39894, \qquad f(h) = f_1 = 0.37738, \qquad f(2h) = f_2 = 0.31945$$

and $\qquad f(1) = f_3 = 0.24197,$

all to 5 decimal places. It follows that

$$\int_0^1 \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx \approx \frac{1}{2}h(f_0 + f_3 + 2\{f_1 + f_2\}) = 0.33910$$

We know from part (a) that this approximation is in error by less than $\frac{1}{2} \times 10^{-2}$.

Example 14

Determine the minimum number of steps needed to guarantee an error not exceeding ± 0.001, when evaluating

$$\int_0^1 \cosh(x^2)\, dx$$

using the trapezium rule.

Solution

$$f(x) = \cosh(x^2) \qquad f'(x) = 2x\sinh(x^2) \qquad f''(x) = 2\sinh(x^2) + 4x^2\cosh(x^2)$$

Using the error formula in Key Point 8

$$E = \left| -\frac{1}{12}h^2\{2\sinh(x^2) + 4x^2\cosh(x^2)\} \right| \qquad x \in [0, 1]$$

$|E|_{\max}$ occurs when $x = 1$

$$\begin{aligned} 0.001 &> \frac{h^2}{12}\{2\sinh(1) + 4\cosh(1)\} \\ h^2 &< 0.012/\{(2\sinh(1) + 4\cosh(1)\} \\ \Rightarrow h^2 &< 0.001408 \\ \Rightarrow h &< 0.037523 \\ \Rightarrow n &\geq 26.651 \\ \Rightarrow n &= 27 \quad \text{needed} \end{aligned}$$

Determine the minimum of strips, n, needed to evaluate by the trapezium rule:

$$\int_0^{\pi/4} \{3x^2 - 1.5\sin(2x)\}\, dx$$

such that the error is guaranteed not to exceed ± 0.005.

Your solution

Answer

$f(x) = 3x^2 - 1.5\sin(2x) \qquad f''(x) = 6 + 6\sin(2x)$

|Error| will be maximum at $x = \dfrac{\pi}{4}$ so that $\sin(2x) = 1$

$$E = -\frac{(b-a)}{12}h^2 f^{(2)}(x) \qquad x \in [0, \frac{\pi}{4}]$$

$$E = -\frac{\pi}{48}h^2 6\{1 + \sin(2x)\}, \qquad x \in [0, \frac{\pi}{4}]$$

$$|E|_{\max} = \frac{\pi}{48}h^2(12) = \frac{\pi h^2}{4}$$

We need $\dfrac{\pi h^2}{4} < 0.005 \qquad \Rightarrow h^2 < \dfrac{0.02}{\pi} \qquad \Rightarrow \qquad h < 0.07979$

Now $nh = (b-a) = \dfrac{\pi}{4}$ so $n = \dfrac{\pi}{4h}$

We need $n > \dfrac{\pi}{4 \times 0.07979} = 9.844$ so $n = 10$ required

4. Other methods for approximating integrals

Here we briefly describe other methods that you may have heard, or get to hear, about. In the end they all amount to the same sort of thing, that is we sample the integrand f at a few points in the integration interval and then take a weighted average of all these f values. All that is needed to implement any of these methods is the list of sampling points and the weight that should be attached to each evaluation. Lists of these points and weights can be found in many books on the subject.

Simpson's rule

This is based on passing a quadratic through three equally spaced points, rather than passing a straight line through two points as we did for the simple trapezium rule. The composite Simpson's rule is given in the following Key Point.

Key Point 9

Composite Simpson's Rule

The composite Simpson's rule for approximating $\int_a^b f(x)\,dx$ is carried out as follows:

1. Choose N, which must be an [even] number of subintervals,
2. Calculate $\int_a^b f(x)\,dx$

$$\approx \tfrac{1}{3}h\left(f_0 + 4\{f_1 + f_3 + f_5 + \cdots + f_{N-1}\} + 2\{f_2 + f_4 + f_6 + \cdots + f_{N-2}\} + f_N\right)$$

where

$$h = \frac{b-a}{N}, \qquad f_0 = f(a), \quad f_1 = f(a+h), \ldots, f_n = f(a+nh), \ldots,$$
$$\text{and} \quad f_N = f(a+Nh) = f(b).$$

The formula in Key Point 9 is slightly more complicated than the corresponding one for composite trapezium rule. One way of remembering the rule is the learn the pattern

$$1\ 4\ 2\ 4\ 2\ 4\ 2\ \ldots\ 4\ 2\ 4\ 2\ 4\ 1$$

which show that the end point values are multiplied by 1, the values with odd-numbered subscripts are multiplied by 4 and the *interior* values with even subscripts are multiplied by 2.

Example 15

Using 4 subintervals in the composite Simpson's rule approximate

$$\int_0^2 \cosh(x)\, dx.$$

Solution

In this case $h = (2-0)/4 = 0.5$.
We require $\cosh(x)$ evaluated at five x-values and the results are tabulated below to 6 d.p.

x_n	$f_n = \cosh(x_n)$
0	1.000000
0.5	1.127626
1	1.543081
1.5	2.352410
2	3.762196

It follows that

$$\begin{aligned}\int_0^2 \cosh(x)\, dx &\approx \frac{1}{3}h\left(f_0 + 4f_1 + 2f_2 + 4f_3 + f_4\right)\\ &= \frac{1}{3}(0.5)\left(1 + 4\times 1.127626 + 2\times 1.543081 + 4\times 2.352410 + 3.762196\right)\\ &= 3.628083,\end{aligned}$$

where this approximation is given to 6 decimal places.

This approximation to $\int_0^2 \cosh(x)\ dx$ is closer to the true value of $\sinh(2)$ (which is 3.626860 to 6 d.p.) than we obtained when using the composite trapezium rule with the same number of subintervals.

Using 4 subintervals in the composite Simpson's rule approximate

$$\int_1^2 \ln(x)\, dx.$$

Your solution

Answer

In this case $h = (2-1)/4 = 0.25$. There will be five x-values and the results are tabulated below to 6 d.p.

x_n	$f_n = \ln(x_n)$
1.00	0.000000
1.25	0.223144
1.50	0.405465
1.75	0.559616
2.00	0.693147

It follows that

$$\begin{aligned}\int_1^2 \ln(x)\,dx &\approx \frac{1}{3}h\left(f_0 + 4f_1 + 2f_2 + 4f_3 + f_4\right)\\ &= \frac{1}{3}(0.25)\left(0 + 4\times 0.223144 + 2\times 0.405465 + 4\times 0.559616 + 0.693147\right)\\ &= 0.386260 \quad \text{to 6 d.p.}\end{aligned}$$

How good is the composite Simpson's rule?

On page 39 (Key Point 8) we saw a formula for an upper bound on the error in the composite trapezium method. A corresponding result for the composite Simpson's rule exists and is given in the following Key Point.

Key Point 10

Error in Composite Simpson's Rule

The error in the N-subinterval composite Simpson's rule approximation to $\displaystyle\int_a^b f(x)\,dx$ is bounded above by

$$\max_{a\le x\le b}\left|f^{(iv)}(x)\right|\frac{(b-a)h^4}{180}$$

(Here $f^{(iv)}$ is the fourth derivative of f and h is the subinterval width, so $N\times h = b-a$.)

The formula in Key Point 10 can be used to decide how many subintervals to use to guarantee a specific accuracy.

Example 16

The function f is known to have a fourth derivative with the property that

$$\left|f^{(iv)}(x)\right| < 5$$

for x between 1 and 5. Determine how many subintervals are required so that the composite Simpson's rule used to approximate

$$\int_1^5 f(x)\,dx$$

incurs an error that is guaranteed less than 0.005 .

Solution

We require that

$$5 \times \frac{4h^4}{180} < 0.005$$

This implies that $h^4 < 0.045$ and therefore $h < 0.460578$.
Now $N = 4/h$ and it follows that

$$N > 8.684741$$

For the composite Simpson's rule N must be an *even* whole number and we conclude that the smallest number of subintervals which *guarantees* an error smaller than 0.005 is $N = 10$.

The function f is known to have a fourth derivative with the property that

$$\left|f^{(iv)}(x)\right| < 12$$

for x between 2 and 6. Determine how many subintervals are required so that the composite Simpson's rule used to approximate

$$\int_2^6 f(x)\,dx$$

incurs an error that is guaranteed less than 0.0005 .

Your solution

Answer
We require that

$$12 \times \frac{4h^4}{180} < 0.0005$$

This implies that $h^4 < 0.001875$ and therefore $h < 0.208090$.
Now $N = 4/h$ and it follows that

$$N > 19.222491$$

N must be an *even* whole number and we conclude that the smallest number of subintervals which *guarantees* an error smaller than 0.0005 is $N = 20$.

The following Task is similar to one that we saw earlier in this Section (page 42). Using the composite Simpson's rule we can achieve greater accuracy, for a similar amount of effort, than we managed using the composite trapezium rule.

It is given that the function $e^{-x^2/2}$ has a fourth derivative that is never greater than 3 in absolute value.

(a) Use this fact to determine how many subintervals are required for the composite Simpson's rule to deliver an approximation to

$$\int_0^1 \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx$$

that is guaranteed to have an error less than $\frac{1}{2} \times 10^{-4}$.

Your solution

Answer
We require that $\dfrac{3}{\sqrt{2\pi}} \dfrac{(b-a)h^4}{180} < 0.00005$.

This means that $h^4 < 0.00751988$ and therefore $h < 0.294478$. Since $N = 1/h$ it is necessary for $N = 4$ for the error bound to be guaranteed to be less than $\pm\frac{1}{2} \times 10^{-4}$.

(b) Find an approximation to the integral that is in error by less than $\frac{1}{2} \times 10^{-4}$.

Your solution

Answer

In this case $h = (1-0)/4 = 0.25$. We require $\dfrac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$ evaluated at five x-values and the results are tabulated below to 6 d.p.

x_n	$\dfrac{1}{\sqrt{2\pi}}\, e^{-x_n^2/2}$
0	0.398942
0.25	0.386668
0.5	0.352065
0.75	0.301137
1	0.241971

It follows that

$$\begin{aligned}\int_0^1 \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx &\approx \frac{1}{3}h\left(f_0 + 4f_1 + 2f_2 + 4f_3 + f_4\right)\\ &= \frac{1}{3}(0.25)\left(0.398942 + 4 \times 0.386668 + 2 \times 0.352065\right.\\ &\qquad \left. +4 \times 0.301137 + 0.241971\right)\\ &= 0.341355 \quad \text{to 6 d.p.}\end{aligned}$$

We know from part (a) that this approximation is in error by less than $\frac{1}{2} \times 10^{-4}$

Example 17

Find out how many strips are needed to be sure that

$$\int_0^4 \sinh(2t)\, dt$$

is evaluated by Simpson's rule with error less than ± 0.0001

Solution

$$E = -\frac{(b-a)}{180} h^4 (16) \sinh(2x) \qquad 0 < x < 4$$

$$|E| \leq \frac{64h^2 \sinh(8)}{180} \leq 0.0001$$

$$\Rightarrow h^4 \leq \frac{0.0180}{64 \sinh(8)} \Rightarrow h \leq 0.0208421$$

$$nh = b - a \quad \Rightarrow \quad n \geq \frac{4}{0.0208421} = 191.92$$

So $n = 192$ is needed (minimum even number).

Engineering Example 1

Plastic bottle design

Introduction

Manufacturing containers is a large and varied industry and optimum packaging can save companies millions of pounds. Although determining the capacity of a container and amount of material needed can be done by physical experiment, mathematical modelling provides a cost-effective and efficient means for the designer to experiment.

Problem in words

A manufacturer is designing a new plastic bottle to contain 900 ml of fabric softener. The bottle is circular in cross section, with a varying radius given by

$$r = 4 + 0.5z - 0.07z^2 + 0.002z^3$$

where z is the height above the base in cm.

(a) Find an expression for the volume of the bottle and hence show that the fill level needs to be approximately 18 cm.

(b) If the wall thickness of the plastic is 1 mm, show that this is always small compared to the bottle radius.

(c) Hence, find the volume of plastic required to manufacture a bottle which is 20 cm tall (include the plastic in the base and side walls), using a numerical method.

A graph the radius against z is shown below:

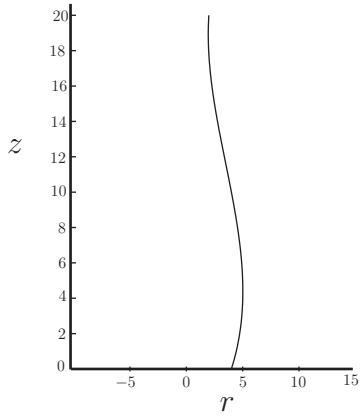


Figure 10

Mathematical statement of problem

Calculate all lengths in centimetres.

(a) The formula for the volume of a solid of revolution, revolved round the z axis between $z = 0$ and $z = d$ is $\displaystyle\int_0^d \pi r^2\, dz$. We have to evaluate this integral.

(b) To show that the thickness is small relative to the radius we need to find the minimum radius.

(c) Given that the thickness is small compared with the radius, the volume can be taken to be the surface area times the thickness. Now the surface area of the base is easy to calculate being $\pi \times 4^2$, but we also need to calculate the surface area for the sides, which is much harder.

For an element of height dz this is $2\pi z \times$(the slant height) of the surface between z and $z+dz$. The slant height is, analytically $\left(\sqrt{1+\left(\frac{dr}{dz}\right)^2}\right) \times dz$, or equivalently the distance between $(r(z), z)$ and $(r(z+dz), z+dz)$, which is easier to use numerically.

Analytically the surface area to height 20 is $\int_0^{20} 2\pi r\sqrt{1+\left(\frac{dr}{dz}\right)^2}\, dz$; we shall approximate this numerically. This will give the area of the side surface.

Mathematical analysis

(a) We could calculate this integral exactly, as the volume is $\int_0^d \pi(4+0.5z-0.07z^2+0.002z^3)^2\, dz$ but here we do this numerically (which can often be a simpler approach and possibly is so here). To do that we need to keep an eye on the likely error, and for this problem we shall ensure the error in the integrals is less than 1 ml. The formula for the error with the trapezium rule, with step h and integrated from 0 to 20 (assuming from the problem that we shall not integrate over a larger range) is $\frac{20}{12}h^2 \max|f''|$. Doing this crudely with $f = \pi g^2$ where $g(z) = 4 + 0.5z - 0.07z^2 + 0.002z^3$ we see that

$$|g(z)| \leq 4 + 10 + 28 + 16 = 58 \quad \text{(using only positive signs and } |z| \leq 20)$$

and $$|g'(z)| \leq 0.5 + 0.14z + 0.006z^2 \leq 0.5 + 2.8 + 2.4 = 5.7 < 6,$$

and $$|g''(z)| \leq 0.14 + 0.012z \leq 0.38.$$

Therefore

$$f'' = 2\pi(gg'' + (g')^2) \leq 2(58 \times 0.38 + 6^2)\pi < 117\pi, \quad \text{so} \quad \frac{20}{12}h^2 \max|f''| \leq 613h^2.$$

We need $h^2 < 1/613$, or $h < 0.0403$. We will use $h = 0.02$, and the error will be at most 0.25.

The approximation to the integral from 0 to 18 is

$$\frac{1}{2}\pi g^2(0)0.02 + \sum_{i=1}^{899} \pi g^2(0.02i)0.02 + \frac{1}{2}\pi g^2(18)0.02$$

(recalling the multiplying factor is a half for the first and last entries in the trapezium rule). This yields a value of 899.72, which is certainly within 1 ml of 900.

(b) From the graph the minimum radius looks to be about 2 at about $z = 18$. Looking more exactly (found by solving the quadratic to find the places where the derivative is zero, or by plotting the values and by inspection), the minimum is at $z = 18.93$, when $r = 1.948$ cm. So the thickness is indeed small (always less than 0.06 of the radius at all places.)

(c) For the area of the side surface we shall calculate $\displaystyle\int_0^{20} 2\pi r\sqrt{1+\left(\frac{dr}{dz}\right)^2}\,dz$ numerically, using the trapezium rule with step 0.02 as before. $\sqrt{1+\left(\frac{dr}{dz}\right)^2}\,dz = \sqrt{(dz)^2+(dr)^2}$, which we shall approximate at point z_n by $\sqrt{(z_{n+1}-z_n)^2+(r_{n+1}-r_n)^2}$, so evaluating $r(z)$ at intervals of 0.02 gives the approximation

$$\pi r(0)\sqrt{(0.02)^2+(r(0.02)-r(0))^2}+\sum_{i=1}^{999}2\pi r(0.02i)\sqrt{(0.02)^2+(r(0.02(i+1))-r(0.02i))^2}$$

$$+\pi r(20)\sqrt{(0.02)^2+(r(20)-r(19.98))^2}.$$

Calculating this gives 473 cm^2. Approximating the analytical expression by a direct numerical calculation gives 474 cm^2. (The answer is between 473.5 and 473.6 cm^2, so this variation is understandable and does not indicate an error.) The bottom surface area is $16\pi = 50.3$ cm^2, so the total surface area we may take to be $474+50 = 524$ cm^2, and hence the volume of plastic is $524\times 0.1 = 52.4$ cm^3.

Mathematical comment

An alternative to using the trapezium rule is Simpson's rule which will require many fewer steps.

When using a computer program such as Microsoft Excel having an efficient method may not be important for a small problem but could be significant when many calculations are needed or computational power is limited (such as if using a programmable calculator).

The reader is invited to repeat the calculations for (a) and (c) using Simpson's rule.

The analytical answer to (a) is given by

$$\int_0^{18}\pi(16+4z-0.31z^2-0.054z^3+0.0069z^4-0.00028z^5+0.000004z^6)\,dz$$

which gives 899.7223 to 4 d.p.

Exercises

1. Using 4 subintervals in the composite trapezium rule approximate

$$\int_1^5 \sqrt{x}\, dx.$$

2. The function f is known to have a second derivative with the property that

$$|f''(x)| < 12$$

for x between 2 and 3. Using the error bound given earlier in this Section determine how many subintervals are required so that the composite trapezium rule used to approximate

$$\int_2^3 f(x)\, dx$$

can be guaranteed to have an error in it that is less than 0.001.

3. Using 4 subintervals in the composite Simpson rule approximate

$$\int_1^5 \sqrt{x}\, dx.$$

4. The function f is known to have a fourth derivative with the property that

$$\left|f^{(iv)}(x)\right| < 6$$

for x between -1 and 5 . Determine how many subintervals are required so that the composite Simpson's rule used to approximate

$$\int_{-1}^5 f(x)\, dx$$

incurs an error that is less than 0.001.

5. Determine the minimum number of steps needed to guarantee an error not exceeding ± 0.000001 when numerically evaluating

$$\int_2^4 \ln(x)\, dx$$

using Simpson's rule.

Answers

1. In this case $h = (5-1)/4 = 1$. We require $\sqrt{x}$ evaluated at five x-values and the results are tabulated below

x_n	$f_n = \sqrt{x_n}$
1	1
2	1.414214
3	1.732051
4	2.000000
5	2.236068

It follows that

$$\begin{aligned}\int_1^5 \sqrt{x}\,dx &\approx \frac{1}{2}h\left(f_0 + f_4 + 2\{f_1 + f_2 + f_3\}\right)\\ &= \frac{1}{2}(1)\Big(1 + 2.236068 + 2\{1.414214 + 1.732051 + 2\}\Big)\\ &= 6.764298.\end{aligned}$$

2. We require that $12 \times \dfrac{(b-a)h^2}{12} < 0.001$. This implies that $h < 0.0316228$.
Now $N = (b-a)/h = 1/h$ and it follows that

$$N > 31.6228$$

Clearly, N must be a whole number and we conclude that the smallest number of subintervals which *guarantees* an error smaller than 0.001 is $N = 32$.

3. In this case $h = (5-1)/4 = 1$.
We require $\sqrt{x}$ evaluated at five x-values and the results are as tabulated in the solution to Exercise 1. It follows that

$$\begin{aligned}\int_1^5 \sqrt{x}\,dx &\approx \frac{1}{3}h\left(f_0 + 4f_1 + 2f_2 + 4f_3 + f_4\right)\\ &= \frac{1}{3}(1)\left(1 + 4 \times 1.414214 + 2 \times 1.732051 + 4 \times 2.000000 + 2.236068\right)\\ &= 6.785675.\end{aligned}$$

4. We require that $6 \times \dfrac{6h^4}{180} < 0.001$. This implies that $h^4 < 0.005$ and therefore $h < 0.265915$.
Now $N = 6/h$ and it follows that $N > 22.563619$. We know that N must be an *even* whole number and we conclude that the smallest number of subintervals which *guarantees* an error smaller than 0.001 is $N = 24$.

Answers

5. $f(x) = \ln(x) \qquad f^{(4)}(x) = -\dfrac{6}{x^4}$

$$\text{Error} = -\frac{(b-a)h^4 f^{(4)}(x)}{180} \qquad a = 2,\ b = 4$$

$$|E| = \frac{2h^4(6/x^4)}{180} \qquad x \in [2,4]$$

$$|E|_{\max} = \frac{h^4}{15}\frac{1}{2^4} \leq 0.000001$$

$$\Rightarrow h^4 \leq 15 \times 2^4 \times 0.000001 \Rightarrow h \leq 0.124467$$

Now $nh = (b-a)$ so

$$n \geq \frac{2}{0.124467} \Rightarrow n \geq 16.069568 \Rightarrow n = 18 \quad \text{(minimum even number)}$$

Numerical Differentiation

31.3

Introduction

In this Section we will look at ways in which derivatives of a function may be approximated numerically.

Prerequisites

Before starting this Section you should . . .

- review previous material concerning differentiation

Learning Outcomes

On completion you should be able to . . .

- obtain numerical approximations to the first and second derivatives of certain functions

1. Numerical differentiation

This Section deals with ways of numerically approximating derivatives of functions. One reason for dealing with this now is that we will use it briefly in the next Section. But as we shall see in these next few pages, the technique is useful in itself.

2. First derivatives

Our aim is to approximate the slope of a curve f at a particular point $x = a$ in terms of $f(a)$ and the value of f at a nearby point where $x = a + h$. The shorter broken line Figure 11 may be thought of as giving a reasonable approximation to the required slope (shown by the longer broken line), if h is small enough.

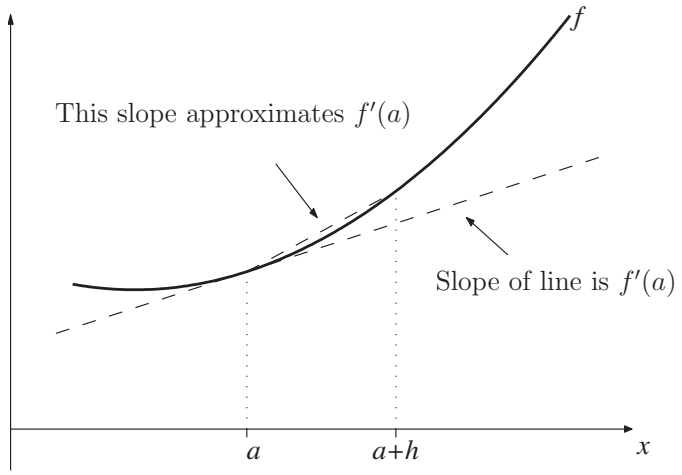


Figure 11

So we might approximate

$$f'(a) \approx \text{slope of short broken line} = \frac{\text{difference in the } y\text{-values}}{\text{difference in the } x\text{-values}} = \frac{f(a+h) - f(a)}{h}.$$

This is called a **one-sided difference** or **forward difference** approximation to the derivative of f.

A second version of this arises on considering a point to the left of a, rather than to the right as we did above. In this case we obtain the approximation

$$f'(a) \approx \frac{f(a) - f(a-h)}{h}$$

This is another **one-sided difference**, called a **backward difference**, approximation to $f'(a)$.

A third method for approximating the first derivative of f can be seen in Figure 12.

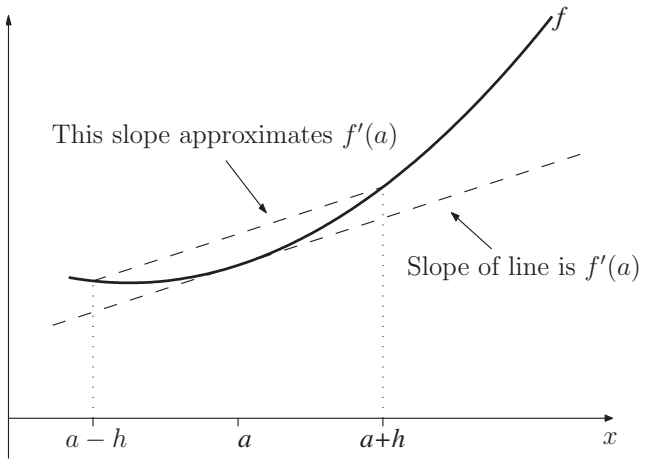


Figure 12

Here we approximate as follows

$$f'(a) \approx \text{slope of short broken line} = \frac{\text{difference in the } y\text{-values}}{\text{difference in the } x\text{-values}} = \frac{f(x+h)-f(x-h)}{2h}$$

This is called a **central difference** approximation to $f'(a)$.

Key Point 11

First Derivative Approximations

Three approximations to the derivative $f'(a)$ are

1. the one-sided (forward) difference $\dfrac{f(a+h)-f(a)}{h}$
2. the one-sided (backward) difference $\dfrac{f(a)-f(a-h)}{h}$
3. the central difference $\dfrac{f(a+h)-f(a-h)}{2h}$

In practice, the central difference formula is the most accurate.

These first, rather artificial, examples will help fix our ideas before we move on to more realistic applications.

Example 18

Use a forward difference, and the values of h shown, to approximate the derivative of $\cos(x)$ at $x = \pi/3$.

(a) $h = 0.1$ (b) $h = 0.01$ (c) $h = 0.001$ (d) $h = 0.0001$

Work to 8 decimal places throughout.

Solution

(a) $f'(a) \approx \dfrac{\cos(a+h) - \cos(a)}{h} = \dfrac{0.41104381 - 0.5}{0.1} = -0.88956192$

(b) $f'(a) \approx \dfrac{\cos(a+h) - \cos(a)}{h} = \dfrac{0.49131489 - 0.5}{0.01} = -0.86851095$

(c) $f'(a) \approx \dfrac{\cos(a+h) - \cos(a)}{h} = \dfrac{0.49913372 - 0.5}{0.001} = -0.86627526$

(d) $f'(a) \approx \dfrac{\cos(a+h) - \cos(a)}{h} = \dfrac{0.49991339 - 0.5}{0.0001} = -0.86605040$

One advantage of doing a simple example first is that we can compare these approximations with the 'exact' value which is

$$f'(a) = -\sin(\pi/3) = -\frac{\sqrt{3}}{2} = -0.86602540 \quad \text{to 8 d.p.}$$

Note that the accuracy levels of the four approximations in Example 15 are:

(a) 1 d.p. (b) 2 d.p. (c) 3 d.p. (d) 3 d.p. (almost 4 d.p.)

The errors to 6 d.p. are:

(a) 0.023537 (b) 0.002486 (c) 0.000250 (d) 0.000025

Notice that the errors reduce by about a factor of 10 each time.

Example 19

Use a central difference, and the value of h shown, to approximate the derivative of $\cos(x)$ at $x = \pi/3$.

(a) $h = 0.1$ (b) $h = 0.01$ (c) $h = 0.001$ (d) $h = 0.0001$

Work to 8 decimal places throughout.

Solution

(a) $f'(a) \approx \dfrac{\cos(a+h)-\cos(a-h)}{2h} = \dfrac{0.41104381 - 0.58396036}{0.2} = -0.86458275$

(b) $f'(a) \approx \dfrac{\cos(a+h)-\cos(a-h)}{2h} = \dfrac{0.49131489 - 0.50863511}{0.02} = -0.86601097$

(c) $f'(a) \approx \dfrac{\cos(a+h)-\cos(a-h)}{2h} = \dfrac{0.49913372 - 0.50086578}{0.002} = -0.86602526$

(d) $f'(a) \approx \dfrac{\cos(a+h)-\cos(a-h)}{2h} = \dfrac{0.49991339 - 0.50008660}{0.0002} = -0.86602540$

This time successive approximations generally have *two* extra accurate decimal places indicating a superior formula. This is illustrated again in the following Task.

Let $f(x) = \ln(x)$ and $a = 3$. Using both a forward difference and a central difference, and working to 8 decimal places, approximate $f'(a)$ using $h = 0.1$ and $h = 0.01$.
(Note that this is another example where we can work out the exact answer, which in this case is $\frac{1}{3}$.)

Your solution

Answer
Using the forward difference we find, for $h = 0.1$

$$f'(a) \approx \frac{\ln(a+h) - \ln(a)}{h} = \frac{1.13140211 - 1.09861229}{0.1} = 0.32789823$$

and for $h = 0.01$ we obtain

$$f'(a) \approx \frac{\ln(a+h) - \ln(a)}{h} = \frac{1.10194008 - 1.09861229}{0.01} = 0.33277901$$

Using central differences the two approximations to $f'(a)$ are

$$f'(a) \approx \frac{\ln(a+h) - \ln(a-h)}{2h} = \frac{1.13140211 - 1.06471074}{0.2} = 0.33345687$$

and

$$f'(a) \approx \frac{\ln(a+h) - \ln(a-h)}{2h} = \frac{1.10194008 - 1.09527339}{0.02} = 0.33333457$$

The accurate answer is, of course, 0.33333333

There is clearly little point in studying this technique if all we ever do is approximate quantities we could find exactly in another way. The following example is one in which this so-called **differencing method** is the best approach.

Example 20

The distance x of a runner from a fixed point is measured (in metres) at intervals of half a second. The data obtained are

t	0.0	0.5	1.0	1.5	2.0
x	0.00	3.65	6.80	9.90	12.15

Use central differences to approximate the runner's velocity at times $t = 0.5$ s and $t = 1.25$ s.

Solution

Our aim here is to approximate $x'(t)$. The choice of h is dictated by the available data given in the table.
Using data with $t = 0.5$ s at its centre we obtain

$$x'(0.5) \approx \frac{x(1.0) - x(0.0)}{2 \times 0.5} = 6.80 \text{ m s}^{-1}.$$

Data centred at $t = 1.25$ s gives us the approximation

$$x'(1.25) \approx \frac{x(1.5) - x(1.0)}{2 \times 0.25} = 6.20 \text{ m s}^{-1}.$$

Note the value of h used.

The velocity v (in m s^{-1}) of a rocket measured at half second intervals is

t	0.0	0.5	1.0	1.5	2.0
v	0.000	11.860	26.335	41.075	59.051

Use central differences to approximate the acceleration of the rocket at times $t = 1.0$ s and $t = 1.75$ s.

Your solution

Answer

Using data with $t = 1.0$ s at its centre we obtain

$$v'(1.0) \approx \frac{v(1.5) - v(0.5)}{1.0} = 29.215 \text{ m s}^{-2}.$$

Data centred at $t = 1.75$ s gives us the approximation

$$v'(1.75) \approx \frac{v(2.0) - v(1.5)}{0.5} = 35.952 \text{ m s}^{-2}.$$

3. Second derivatives

An approach which has been found to work well for second derivatives involves applying the notion of a central difference three times. We begin with

$$f''(a) \approx \frac{f'(a + \frac{1}{2}h) - f'(a - \frac{1}{2}h)}{h}.$$

Next we approximate the two derivatives in the numerator of this expression using central differences as follows:

$$f'(a + \tfrac{1}{2}h) \approx \frac{f(a+h) - f(a)}{h} \qquad \text{and} \qquad f'(a - \tfrac{1}{2}h) \approx \frac{f(a) - f(a-h)}{h}.$$

Combining these three results gives

$$\begin{aligned} f''(a) &\approx \frac{f'(a+\frac{1}{2}h) - f'(a-\frac{1}{2}h)}{h} \\ &\approx \frac{1}{h}\left\{\left(\frac{f(a+h)-f(a)}{h}\right) - \left(\frac{f(a)-f(a-h)}{h}\right)\right\} \\ &= \frac{f(a+h) - 2f(a) + f(a-h)}{h^2} \end{aligned}$$

Key Point 12

Second Derivative Approximation

A central difference approximation to the second derivative $f''(a)$ is

$$f''(a) \approx \frac{f(a+h) - 2f(a) + f(a-h)}{h^2}$$

Example 21

The distance x of a runner from a fixed point is measured (in metres) at intervals of half a second. The data obtained are

t	0.0	0.5	1.0	1.5	2.0
x	0.00	3.65	6.80	9.90	12.15

Use a central difference to approximate the runner's acceleration at $t = 1.5$ s.

Solution

Our aim here is to approximate $x''(t)$.
Using data with $t = 1.5$ s at its centre we obtain

$$\begin{aligned} x''(1.5) &\approx \frac{x(2.0) - 2x(1.5) + x(1.0)}{0.5^2} \\ &= -3.40 \text{ m s}^{-2}, \end{aligned}$$

from which we see that the runner is slowing down.

Exercises

1. Let $f(x) = \cosh(x)$ and $a = 2$. Let $h = 0.01$ and approximate $f'(a)$ using forward, backward and central differences. Work to 8 decimal places and compare your answers with the exact result, which is $\sinh(2)$.

2. The distance x, measured in metres, of a downhill skier from a fixed point is measured at intervals of 0.25 s. The data gathered are

t	0	0.25	0.5	0.75	1	1.25	1.5
x	0	4.3	10.2	17.2	26.2	33.1	39.1

   Use a central difference to approximate the skier's velocity and acceleration at the times $t =$0.25 s, 0.75 s and 1.25 s. Give your answers to 1 decimal place.

Answers

1. Forward: $f'(a) \approx \dfrac{\cosh(a+h) - \cosh(a)}{h} = \dfrac{3.79865301 - 3.76219569}{0.01} = 3.64573199$

   Backward: $f'(a) \approx \dfrac{\cosh(a) - \cosh(a-h)}{h} = \dfrac{3.76219569 - 3.72611459}{0.01} = 3.60810972$

   Central: $f'(a) \approx \dfrac{\cosh(a+h) - \cosh(a-h)}{2h} = \dfrac{3.79865301 - 3.72611459}{0.02} = 3.62692086$

   The accurate result is $\sinh(2) = 3.62686041$.

2. Velocities at the given times approximated by a central difference are:

   20.4 m s^{-1}, 32.0 m s^{-1} and 25.8 m s^{-1}.

   Accelerations at these times approximated by a central difference are:

   25.6 m s^{-2}, 32.0 m s^{-2}and -14.4 m s^{-2}.

Nonlinear Equations 31.4

Introduction

In this Section we briefly discuss nonlinear equations (what they are and what their solutions might be) before noting that many such equations which crop up in applications cannot be solved exactly.

The remainder (and majority) of the Section then goes on to discuss methods for approximating solutions of nonlinear equations.

Prerequisites

Before starting this Section you should ...

- understand derivatives of simple functions
- understand the quadratic formula
- understand exponentials and logarithms

Learning Outcomes

On completion you should be able to ...

- approximate roots of equations by the bisection method and by the Newton-Raphson method
- implement an approximate Newton-Raphson method

1. Nonlinear Equations

A **linear equation** is one related to a straight line, for example $f(x) = mx + c$ describes a straight line with slope m and the linear equation $f(x) = 0$, involving such an f, is easily solved to give $x = -c/m$ (as long as $m \neq 0$). If a function f is not represented by a straight line in this way we say it is **nonlinear**.

The **nonlinear equation** $f(x) = 0$ may have just one solution, like in the linear case, or it may have no solutions at all, or it may have many solutions. For example if $f(x) = x^2 - 9$ then it is easy to see that there are two solutions $x = -3$ and $x = 3$. The nonlinear equation $f(x) = x^2 + 1$ has no solutions at all (unless the application under consideration makes it appropriate to consider complex numbers).

Our aim in this Section is to approximate (real-valued) solutions of nonlinear equations of the form $f(x) = 0$. The definitions of a root of an equation and a zero of a function have been gathered together in Key Point 13.

Key Point 13

If the value x is such that $f(x) = 0$ we say that

1. x is a **root** of the equation $f(x) = 0$
2. x is a **zero** of the function f.

Example 22

Find any (real valued) zeros of the following functions. (Give 3 decimal places if you are unable to give an exact numerical value.)

(a) $f(x) = x^2 + x - 20$ (b) $f(x) = x^2 - 7x + 5$ (c) $f(x) = 2^x - 3$
(d) $f(x) = e^x + 1$ (e) $f(x) = \sin(x)$

Solution

(a) This quadratic factorises easily into $f(x) = (x - 4)(x + 5)$ and so the two zeros of this f are $x = 4$, $x = -5$.

(b) The nonlinear equation $x^2 - 7x + 5 = 0$ requires the quadratic formula and we find that the two zeros of this f are $x = \dfrac{7 \pm \sqrt{7^2 - 4 \times 1 \times 5}}{2} = \dfrac{7 \pm \sqrt{29}}{2}$ which are equal to $x = 0.807$ and $x = 6.193$, to 3 decimal places.

Solution (contd.)

(c) Using the natural logarithm function we see that

$$x\ln(2) = \ln(3)$$

from which it follows that $x = \ln(3)/\ln(2) = 1.585$, to 3 decimal places.

(d) This f has no zeros because $e^x + 1$ is always positive.

(e) $\sin(x)$ has an infinite number of zeros at $x = 0, \pm\pi, \pm 2\pi, \pm 3\pi, \ldots$. To 3 decimal places these are $x = 0.000, \pm 3.142, \pm 6.283, \pm 9.425, \ldots$.

Find any (real valued) zeros of the following functions.

(a) $f(x) = x^2 + 2x - 15$, (b) $f(x) = x^2 - 3x + 3$,

(c) $f(x) = \ln(x) - 2$, (d) $f(x) = \cos(x)$.

For parts (a) to (c) give your answers to 3 decimal places if you cannot give an exact answer; your answers to part (d) may be left in terms of π.

Your solution

Answer

(a) This quadratic factorises easily into $f(x) = (x-3)(x+5)$ and so the two zeros of this f are $x = 3$, $x = -5$.

(b) The equation $x^2 - 3x + 3 = 0$ requires the quadratic formula and the two zeros of this f are

$$x = \frac{3 \pm \sqrt{3^2 - 4 \times 1 \times 3}}{2} = \frac{3 \pm \sqrt{-3}}{2}$$

which are complex values. This f has no real zeros.

(c) Solving $\ln(x) = 2$ gives $x = e^2 = 7.389$, to 3 decimal places.

(d) $\cos(x)$ has an infinite number of zeros at $x = \frac{\pi}{2}, \frac{\pi}{2} \pm \pi, \frac{\pi}{2} \pm 2\pi, \ldots$.

Many functions that crop up in engineering applications do not lend themselves to finding zeros directly as was achieved in the examples above. Instead we approximate zeros of functions, and this Section now goes on to describe some ways of doing this. Some of what follows will involve revision of material you have seen in HELM 12 concerning Applications of Differentiation.

2. The bisection method

Suppose that, by trial and error for example, we know that a single zero of some function f lies between $x = a$ and $x = b$. The root is said to be **bracketed** by a and b. This must mean that $f(a)$ and $f(b)$ are of opposite signs, that is that $f(a)f(b) < 0$.

Example 23

The single positive zero of the function $f(x) = x\tanh(\frac{1}{2}x) - 1$ models the wave number of water waves at a certain frequency in water of depth 0.5 (measured in some units we need not worry about here). Find two points which bracket the zero of f.

Solution

We simply evaluate f at a selection of x-values.

x	$f(x) = x\tanh(\frac{1}{2}x) - 1$	
0	$0 \times \tanh(0) - 1$	$= -1$
0.5	$0.5 \times \tanh(0.25) - 1 = 0.5 \times 0.2449 - 1$	$= -0.8775$
1	$1 \times \tanh(0.5) - 1 = 1 \times 0.4621 - 1$	$= -0.5379$
1.5	$1.5 \times \tanh(0.75) - 1 = 1.5 \times 0.6351 - 1$	$= -0.0473$
2	$2 \times \tanh(1) - 1 = 2 \times 0.7616 - 1$	$= 0.5232$

From this we can see that f changes sign between 1.5 and 2. Thus we can take $a = 1.5$ and $b = 2$ as the bracketing points. That is, the zero of f is in the bracketing interval $1.5 < x < 2$.

The function $f(x) = \cos(x) - x$ has a single positive zero. Find bracketing points a and b for the zero of f. Arrange for the difference between a and b to be equal to $\frac{1}{2}$.
(NB - be careful to use radians on your calculator!)

Your solution

Answer
We evaluate f for a range of values:

x	$f(x)$
0	1
0.5	0.37758
1	-0.459698

Clearly f changes sign between the bracketing values $a = 0.5$ and $b = 1$.
(Other answers are valid of course, it depends which values of f you tried.)

The aim with the bisection method is to repeatedly reduce the width of the bracketing interval $a < x < b$ so that it "pinches" the required zero of f to some desired accuracy. We begin by describing one iteration of the bisection method in detail.

Let $m = \frac{1}{2}(a + b)$, the mid-point of the interval $a < x < b$. All we need to do now is to see in which half (the left or the right) of the interval $a < x < b$ the zero is in. We evaluate $f(m)$. There is a (very slight) chance that $f(m) = 0$, in which case our job is done and we have found the zero of f. Much more likely is that we will be in one of the two situations shown in Figure 13 below. If $f(m)f(b) < 0$ then we are in the situation shown in (a) and we replace $a < x < b$ with the smaller bracketing interval $m < x < b$. If, on the other hand, $f(a)f(m) < 0$ then we are in the situation shown in (b) and we replace $a < x < b$ with the smaller bracketing interval $a < x < m$.

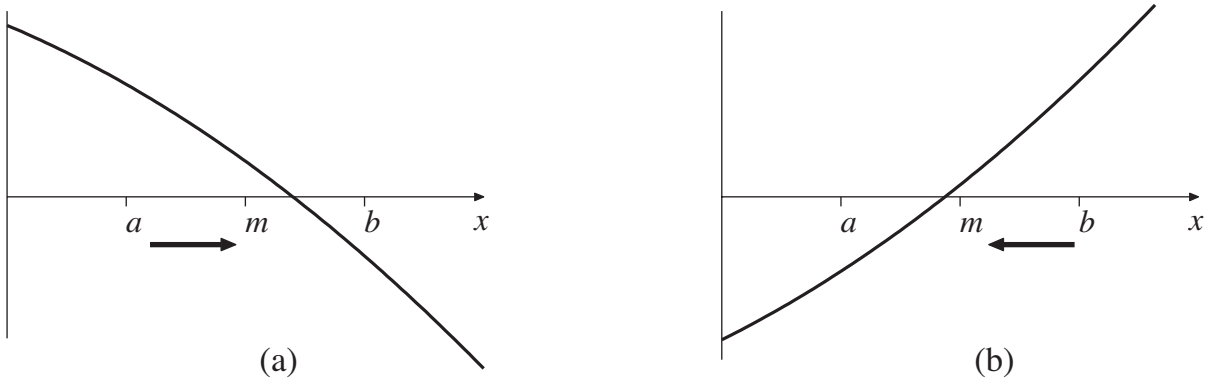


Figure 13

Either way, we now have a bracketing interval that is half the size of the one we started with. We have carried out one iteration of the bisection method. By successively reapplying this approach we can make the bracketing interval as small as we wish.

Example 24

Carry out one iteration of the bisection method so as to halve the width of the bracketing interval $1.5 < x < 2$ for

$$f(x) = x\tanh(\tfrac{1}{2}x) - 1.$$

Solution

The mid-point of the bracketing interval is $m = \frac{1}{2}(a+b) = \frac{1}{2}(1.5+2) = 1.75$. We evaluate

$$f(m) = 1.75 \times \tanh(\tfrac{1}{2} \times 1.75) - 1 = 0.2318,$$

to 4 decimal places. We found earlier (Example 20, page 63) that $f(a) < 0$ and $f(b) > 0$, the fact that $f(m)$ is of the opposite sign to $f(a)$ means that the zero of f lies in the bracketing interval $1.5 < x < 1.75$.

Carry out one iteration of the bisection method so as to halve the width of the bracketing interval $0.5 < x < 1$ for

$$f(x) = \cos(x) - x.$$

Your solution

Answer

Here $a = 0.5, b = 1$. The mid-point of the bracketing interval is $m = \frac{1}{2}(a+b) = \frac{1}{2}(0.5+1) = 0.75$. We evaluate

$$f(m) = \cos(0.75) - 0.75 = -0.0183$$

We found earlier (Task, pages 58-59) that $f(a) > 0$ and $f(b) < 0$, the fact that $f(m)$ is of the opposite sign to $f(a)$ means that the zero of f lies in the bracketing interval $0.5 < x < 0.75$.

So we have a way of halving the size of the bracketing interval. By repeatedly applying this approach we can make the interval smaller and smaller.

The general procedure, involving (possibly) many iterations, is best described as an algorithm:

1. Choose an error tolerance.
2. Let $m = \frac{1}{2}(a+b)$, the mid-point of the bracketing interval.
3. There are three possibilities:
   (a) $f(m) = 0$, this is very unlikely in general, but if it does happen then we have found the zero of f and we can go to step 7,
   (b) the zero is between m and b,
   (c) the zero is between a and m.
4. If the zero is between m and b, that is if $f(m)f(b) < 0$ (as in Figure 13(a)) then let $a = m$.
5. Otherwise the zero must be between a and m (as in Figure 13(b)) so let $b = m$.
6. If $b - a$ is greater than the required tolerance then go to step 2.
7. End.

One feature of this method is that we can predict in advance how much effort is required to achieve a certain level of accuracy.

Example 25

A given problem using the bisection method starts with the bracketing points $a = 1.5$ and $b = 2$. How many iterations will be required so that the error in the approximation is less that $\frac{1}{2} \times 10^{-6}$?

Solution

Before we carry out any iterations we can write that the zero to be approximated is 1.75 ± 0.25 so that the maximum magnitude of the error in 1.75 may be taken to be equal to 0.25.
Each successive iteration will halve the size of the error, so that after n iterations the error is equal to

$$\frac{1}{2^n} \times 0.25$$

We require that this quantity be less than $\frac{1}{2} \times 10^{-6}$. Now,

$$\frac{1}{2^n} \times 0.25 < \frac{1}{2} \times 10^{-6} \quad \text{implies that} \quad 2^n > \frac{1}{2} \times 10^6.$$

The smallest value of n which satisfies this inequality can be found by trial and error, or by using logarithms to see that $n > (\ln(\frac{1}{2}) + 6\ln(10))/\ln(2)$. Either way, the smallest integer which will do the trick is

$$n = 19.$$

It takes 19 iterations of the bisection method to ensure the required accuracy.

A function f is known to have a single zero between the points $a = 3.2$ and $b = 4$. If these values were used as the initial bracketing points in an implementation of the bisection method, how many iterations would be required to ensure an error less than $\frac{1}{2} \times 10^{-3}$?

Your solution

Answer

We require that

$$\frac{1}{2^n} \times \left(\frac{4 - 3.2}{2}\right) < \frac{1}{2} \times 10^{-3}$$

or, after a little rearranging,

$$2^n > \frac{4}{5} \times 10^3.$$

The smallest value of n which satisfies this is $n = 10$. (This can be found by trial-and-error or by using logarithms.)

Pros and cons of the bisection method

Pros

- the method is easy to understand and remember
- the method always works (once you find values a and b which bracket a single zero)
- the method allows us to work out how many iterations it will take to achieve a given error tolerance because we know that the interval will exactly halve at each step

Cons

- the method is **very** slow
- the method cannot find roots where the curve just touches the x-axis but does not cross it (e.g. double roots)

The slowness of the bisection method will not be a surprise now that you have worked through an example or two! Significant effort is involved in evaluating f and then all we do is look at this f-value and see whether it is positive or negative! We are throwing away hard won information.

Let us be realistic here, the slowness of the bisection method hardly matters if all we are saying is that it takes a few more fractions of a second of computing time to finish, when compared with a competing approach. But there are applications in which f may be very expensive (that is, slow) to calculate and there are applications where engineers need to find zeros of a function *many thousands* of times. (Coastal engineers, for example, may employ mathematical wave models that involve finding the wave number we saw in Example 20 at many different water depths.) It is quite possible that you will encounter applications where the bisection method is just not good enough.

3. The Newton-Raphson method

You may recall (e.g. HELM 13.3) that the Newton-Raphson method (often simply called Newton's method) for approximating a zero of the function f is given by

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

where f' denotes the first derivative of f and where x_0 is an initial guess to the zero of f. A graphical way of interpreting how this method works is shown in Figure 14.

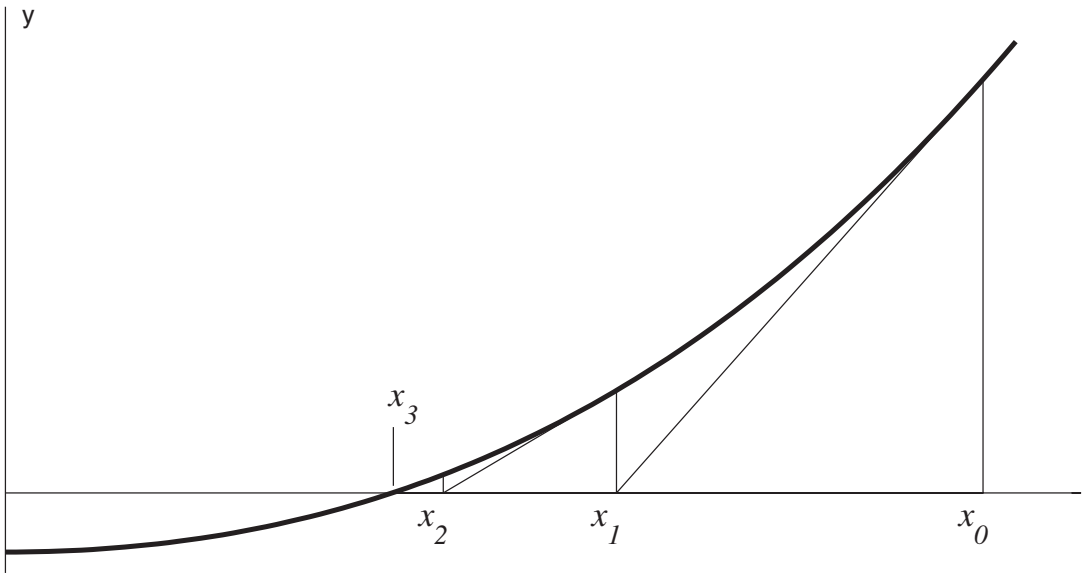


Figure 14

At each approximation to the zero of f we extrapolate so that the tangent to the curve meets the x-axis. This point on the x-axis is the new approximation to the zero of f. As is clear from both the figure and the mathematical statement of the method above, we require that $f'(x_n) \neq 0$ for $n = 0, 1, 2, \ldots$.

Example 26

Let us consider the example we met earlier in Example 24. We know that the single positive zero of

$$f(x) = x\tanh(\tfrac{1}{2}x) - 1$$

lies between 1.5 and 2. Use the Newton-Raphson method to approximate the zero of f.

Solution

We must work out the derivative of f to use Newton-Raphson. Now

$$f'(x) = \tanh(\tfrac{1}{2}x) + x\left(\tfrac{1}{2}\text{sech}^2(\tfrac{1}{2}x)\right)$$

on differentiating a product and recalling that $\dfrac{d}{dx}\tanh(x) = \text{sech}^2(x)$. (To evaluate sech on a calculator recall that $\text{sech}(x) = \dfrac{1}{\cosh(x)}$.)

We must choose a starting value x_0 for the iteration and, given that we know the zero to be between 1.5 and 2, we take $x_0 = 1.75$. The first iteration of Newton-Raphson gives

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 1.75 - \frac{f(1.75)}{f'(1.75)} = 1.75 - \frac{0.231835}{1.145358} = 1.547587,$$

where 6 decimal places are shown. The second iteration gives

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 1.547587 - \frac{f(1.547587)}{f'(1.547587)} = 1.547587 - \frac{0.004585}{1.09687} = 1.543407.$$

Clearly this method lends itself to implementation on a computer and, for example, using a spreadsheet package, it is not hard to compute a few more iterations. Here is output from Microsoft Excel where we have included the two lines of hand-calculation above:

n	x_n	$f(x_n)$	$f'(x_n)$	x_{n+1}
0	1.75	0.231835	1.145358	1.547587
1	1.547587	0.004585	1.09687	1.543407
2	1.543407	$2.52E-06$	1.095662	1.543405
3	1.543405	$7.69E-13$	1.095661	1.543405
4	1.543405	0	1.095661	1.543405

and all subsequent lines are equal to the last line here. The method has converged (very quickly!) to 1.543405, to six decimal places.

Earlier, in Example 25, we found that the bisection method would require 19 iterations to achieve 6 decimal place accuracy. The Newton-Raphson method gave an answer good to this number of places in just two or three iterations.

Use the starting value $x_0 = 0$ in an implementation of the Newton-Raphson method for approximating the zero of

$$f(x) = \cos(x) - x.$$

(If you are doing these calculations by hand then just perform two or three iterations. Don't forget to use radians.)

Your solution

Answer

The derivative of f is $f'(x) = -\sin(x) - 1$. The first iteration is

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 0 - \frac{1-0}{-0-1} = 1$$

and the second iteration is

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 1 - \frac{\cos(1)-1}{-\sin(1)-1} = 1 - \frac{-0.459698}{-1.841471} = 0.750364,$$

and so on. There is little to be gained in our understanding by doing more iterations by hand, but using a spreadsheet we find that the method converges rapidly:

n	x_n	$f(x_n)$	$f'(x_n)$	x_{n+1}
0	0	1	−1	1
1	1	−0.4597	−1.84147	0.750364
2	0.750364	−0.01892	−1.6819	0.739113
3	0.739113	$-4.6E-05$	−1.67363	0.739085
4	0.739085	$-2.8E-10$	−1.67361	0.739085
5	0.739085	0	−1.67361	0.739085

It is often necessary to find zeros of polynomials when studying transfer functions. Here is a Task involving a polynomial.

The function $f(x) = x^3 + 2x + 4$ has a single zero near $x_0 = -1$. Use this value of x_0 to perform two iterations of the Newton-Raphson method.

Your solution

Answer

Using the starting value $x_0 = -1$ you should find that $f(x_0) = 1$ and $f'(x_0) = 5$. This leads to

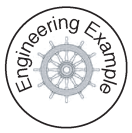
$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = -1 - \frac{1}{5} = -1.2.$$

The second iteration should give you $x_2 = x_1 - \dfrac{f(x_1)}{f'(x_1)} = -1.2 - \dfrac{-0.128}{6.32} = -1.17975.$

Subsequent iterations will home in on the zero of f. Using a computer spreadsheet gives:

n	x_n	$f(x)$	$f'(x)$	x_{n+1}
0	-1	1	5	-1.2
1	-1.2	-0.128	6.32	-1.17975
2	-1.17975	-0.00147	6.175408	-1.17951
3	-1.17951	$-2E-07$	6.173725	-1.17951
4	-1.17951	0	6.173725	-1.17951

where we have recomputed the hand calculations for the first two iterations.
We see that the method converges to the value -1.17951.



Engineering Example 2

Pressure in an ideal multi-component mixture

Introduction

An ideal multi-component mixture consists of

1. n-pentane (5%)
2. n-hextane (15%)
3. n-heptane (50%)
4. n-octane (30%)

In general, the total pressure, P (Pa) of an ideal four-component mixture is related to the boiling point, T (K) through the formula:

$$P = x_1 p_1^* + x_2 p_2^* + x_3 p_3^* + x_4 p_4^*$$

where, for component i, the mole fraction is x_i and the vapour pressure is p_i^*, given by the formula:

$$p_i^* = \exp\left\{A_i - \frac{B_i}{(T + C_i)}\right\} \qquad i = 1, 2, 3, 4$$

Here p_i^* is in mm Hg (1 mm Hg $= 133.32$ Pa), T is the absolute temperature (K) and the constants A_i, B_i and C_i are given in the table below.

i	component	x_i	A_i	B_i	C_i
1	n-pentane	0.05	15.8333	2477.07	-39.94
2	n-hexane	0.15	15.8366	2697.55	-48.78
3	n-heptane	0.50	15.8737	2911.32	-56.51
4	n-octane	0.30	15.9426	3120.29	-63.63

Problem 1

For the liquid compositions, x_i given in the table above, plot a graph of the total pressure, P (Pa) against temperature (K) over the range 250 to 500 K.

Solution

$p_i^* = \exp\left(A_i - \dfrac{B_i}{T + C_i}\right)$, expressed in millimetres of mercury, and so it is 133.32 times that in pascals. Therefore, expressed in pascals, we have

$$P = 133.32 \sum_{i=1}^{4} x_i \exp\left(A_i - \frac{B_i}{T + C_i}\right)$$

Plotting this from $T = 250$ to 500 gives the following graph

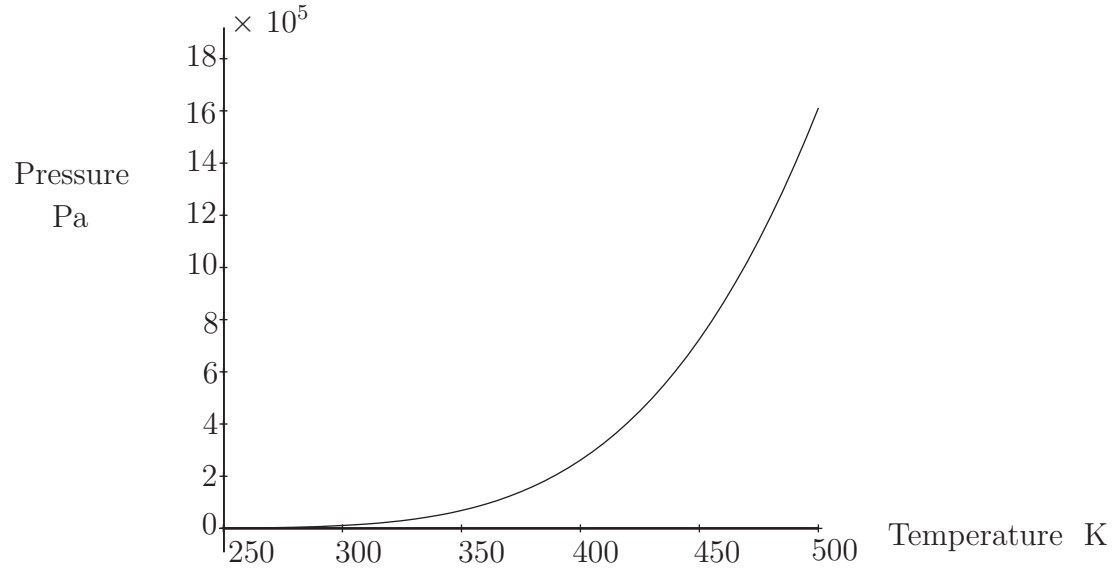


Figure 15

Problem 2

Using the Newton-Raphson method, solve the equations to find the boiling points at total pressures of 1, 2, 5 and 10 bars. Show the sequence of iterations and perform sufficient calculations for convergence to three significant figures. Display these solutions on the graph of the total pressure, P (Pa) against temperature T (K).

Solution

We wish to find T when $P = 1, 2, 5$ and 10 bars, that is, $10^5, 2 \times 10^5, 5 \times 10^5$ and 10×10^5 Pa.

Reading crude approximations to T from the graph gives a starting point for the Newton-Raphson process. We see that for $10^5, 2\times10^5, 5\times10^5$ and 10×10^5 Pa, temperature T is roughly $365, 375, 460$ and 485 degrees K, respectively, so we shall use these values as the start of the iteration.

In this case it is easy to calculate the derivative of P with respect to T exactly, rather than numerically, giving

$$P'(T) = 133.32 \sum_{i=1}^{4} x_i \text{exp}\left(A_i - \frac{B_i}{T + C_i}\right) \times \left(\frac{B_i}{(T + C_i)^2}\right)$$

Therefore to solve the equation $P(T) = y$, we set T_0 to be the starting value above and use the iteration

$$T_{n+1} = T_n - \frac{P(T_n) - y}{P'(T_n)}$$

For $y = 100000$ this gives the iterations

T_0	T_1	T_2	T_3	T_4
365	362.7915	362.7349	362.7349	362.7349

We conclude that, to three significant figures $T = 363°$K when $P = 100000$ Pa.

For $y = 200000$ this gives the iterations

T_0	T_1	T_2	T_3	T_4
375	390.8987	388.8270	388.7854	388.7854

We conclude that, to three significant figures $T = 389°$K when $P = 200000$ Pa.

For $y = 500000$ this gives the iterations

T_0	T_1	T_2	T_3	T_4	T_5
460	430.3698	430.4640	430.2824	430.2821	430.2821

We conclude that, to three significant figures $T = 430°$K when $P = 500000$ Pa.

For $y = 1000000$ this gives the iterations

T_0	T_1	T_2	T_3	T_4	T_5
475	469.0037	468.7875	468.7873	468.7873	468.7873

We conclude that, to three significant figures $T = 469°$K when $P = 1000000$ Pa.

An approximate Newton-Raphson method

The Newton-Raphson method is an excellent way of approximating zeros of a function, but it requires you to know the derivative of f. Sometimes it is undesirable, or simply impossible, to work out the derivative of a function and here we show a way of getting around this.

We approximate the derivative of f. From Section 31.3 we know that

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

is a one-sided (or forward) approximation to f' and another one, using a central difference, is

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}.$$

The advantage of the forward difference is that only one extra f-value has to be computed. If f is especially complicated then this can be a considerable saving when compared with the central difference which requires two extra evaluations of f. The central difference does have the advantage, as we saw when we looked at truncation errors, of being a more accurate approximation to f'.

The spreadsheet program Microsoft Excel has a built in "solver" command which can use Newton's method. (It may be necessary to use the "Add in" feature of Excel to access the solver.) In reality Excel has no way of working out the derivative of the function and must approximate it. Excel gives you the option of using a forward or central difference to estimate f'.

We now reconsider the problem we met in Examples 24 to 26.

Example 27

We know that the single positive zero of $f(x) = x\tanh(\frac{1}{2}x) - 1$ lies between 1.5 and 2. Use the Newton-Raphson method, with an approximation to f', to approximate the zero of f.

Solution

There is no requirement for f' this time, but the nature of this method is such that we will resort to a computer straight away. Let us choose $h = 0.1$ in our approximations to the derivative.
Using the one-sided difference to approximate $f'(x)$ we obtain this sequence of results from the spreadsheet program:

n	x_n	$f(x_n)$	$\frac{f(x+h)-f(x)}{h}$	x_{n+1}
0	1.75	0.231835	1.154355	1.549165
1	1.549165	0.006316	1.110860	1.543479
2	1.543479	$8.16E-05$	1.109359	1.543406
3	1.543406	$1.01E-06$	1.109339	1.543405
4	1.543405	$1.24E-08$	1.109339	1.543405
5	1.543405	$1.53E-10$	1.109339	1.543405
6	1.543405	$1.89E-12$	1.109339	1.543405
7	1.543405	$2.31E-14$	1.109339	1.543405
8	1.543405	0	1.109339	1.543405

And using the (more accurate) central difference gives

n	x_n	$f(x_n)$	$\frac{f(x+h)-f(x-h)}{2h}$	x_{n+1}
0	1.75	0.231835	1.144649	1.547462
1	1.547462	0.004448	1.095994	1.543404
2	1.543404	$-1E-06$	1.094818	1.543405
3	1.543405	$7.95E-10$	1.094819	1.543405
4	1.543405	$-6.1E-13$	1.094819	1.543405
5	1.543405	0	1.094819	1.543405

We see that each of these approaches leads to the same value (1.543405) that we found with the Newton-Raphson method.

Use a spreadsheet to recompute the approximations shown in Example 24, for the following values of h:

$$h = 0.001, \quad 0.00001, \quad 0.000001.$$

Your solution

Answer

You should find that as h decreases, the numbers get closer and closer to those shown earlier for the Newton-Raphson method. For example, when $h = 0.0000001$ we find that for a one-sided difference the results are

n	x_n	$f(x_n)$	$\frac{f(x+h)-f(x)}{h}$	x_{n+1}
0	1.75	0.231835	1.145358	1.547587
1	1.547587	0.004585	1.096870	1.543407
2	1.543407	$2.52E-06$	1.095662	1.543405
3	1.543405	$8.08E-13$	1.095661	1.543405
4	1.543405	0	1.095661	1.543405

and those for a central difference with $h = 0.0000001$ are

n	x_n	$f(x_n)$	$\frac{f(x+h)-f(x-h)}{2h}$	x_{n+1}
0	1.75	0.231835	1.145358	1.547587
1	1.547587	0.004585	1.096870	1.543407
2	1.543407	$2.52E-06$	1.095662	1.543405
3	1.543405	$7.7E-13$	1.095661	1.543405
4	1.543405	0	1.095661	1.543405

It is clear that these two tables very closely resemble the Newton-Raphson results seen earlier.

Exercises

1. It is given that the function

$$f(x) = x^3 + 2x + 8$$

has a single negative zero.

(a) Find two integers a and b which bracket the zero of f.

(b) Perform one iteration of the bisection method so as to halve the size of the bracketing interval.

2. Consider a simple electronic circuit with an input voltage of 2.0 V, a resistor of resistance 1000 Ω and a diode. It can be shown that the voltage across the diode can be found as the single positive zero of

$$f(x) = 1 \times 10^{-14} \exp\left(\frac{x}{0.026}\right) - \frac{2-x}{1000}.$$

Use one iteration of the Newton-Raphson method, and an initial value of $x_0 = 0.75$ to show that

$$x_1 = 0.724983$$

and then work out a second iteration.

3. It is often necessary to find the zeros of polynomials as part of an analysis of transfer functions. The function

$$f(x) = x^3 + 5x - 4$$

has a single zero near $x_0 = 1$. Use this value of x_0 in an implementation of the Newton-Raphson method performing two iterations. (Work to at least 6 decimal place accuracy.)

4. The smallest positive zero of

$$f(x) = x\tan(x) + 1$$

is a measure of how quickly certain evanescent water waves decay, and its value, x_0, is near 3. Use the forward difference

$$\frac{f(3.01) - f(3)}{0.01}$$

to estimate $f'(3)$ and use this value in an approximate version of the Newton-Raphson method to derive one improvement on x_0.

Answers

1. (a) By trial and error we find that $f(-2) = -4$ and $f(-1) = 5$, from which we see that the required bracketing interval is $a < x < b$ where $a = -2$ and $b = -1$.

   (b) For an iteration of the bisection method we find the mid-point $m = -1.5$. Now $f(m) = 1.625$ which is of the opposite sign to $f(a)$ and hence the new smaller bracketing interval is $a < x < m$.

2. The derivative of f is $f'(x) = \dfrac{1 \times 10^{-14}}{0.026} \exp\left(\dfrac{x}{0.026}\right) + \dfrac{1}{1000}$, and therefore the first iteration of Newton-Raphson gives $\quad x_1 = 0.75 - \dfrac{0.032457}{1.297439} = 0.724983.$

   The second iteration gives $\quad x_2 = 0.724983 - \dfrac{0.011603}{0.496319} = 0.701605.$

   Using a spreadsheet we can work out some more iterations. The result of this process is tabulated below

n	x_n	$f(x_n)$	$f'(x_n)$	x_{n+1}
2	0.701605	0.003942	0.202547	0.682144
3	0.682144	0.001161	0.096346	0.670092
4	0.670092	0.000230	0.060978	0.666328
5	0.666328	$1.56E-05$	0.052894	0.666033
6	0.666033	$8.63E-08$	0.052310	0.666031
7	0.666031	$2.68E-12$	0.052306	0.666031
8	0.666031	0	0.052306	0.666031

   and we conclude that the required zero of f is equal to 0.666031, to 6 decimal places.

3. Using the starting value $x_0 = 1$ you should find that $f(x_0) = 2$ and $f'(x_0) = 8$. This leads to

   $$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 1 - \frac{2}{8} = 0.75.$$

   The second iteration should give you $\quad x_2 = x_1 - \dfrac{f(x_1)}{f'(x_1)} = 0.75 - \dfrac{0.171875}{6.6875} = 0.724299.$

   Subsequent iterations can be used to 'home in' on the zero of f and, using a computer spreadsheet program, we find that

n	x_n	$f(x)$	$f'(x)$	x_{n+1}
2	0.724299	0.001469	6.573827	0.724076
3	0.724076	$1.09E-07$	6.572856	0.724076
4	0.724076	0	6.572856	0.724076

   We see that the method converges to the value 0.724076.

Answers

4. We begin with

$$f'(3) \approx \frac{f(3.01) - f(3)}{0.01} = \frac{0.02924345684}{0.01} = 2.924345684,$$

to the displayed number of decimal places, and hence an improvement on $x_0 = 0.75$ is

$$x_1 = 3 - \frac{f(3)}{2.924345684} = 2.804277,$$

to 6 decimal places. (It can be shown that the root of f is 2.798386, to 6 decimal places.)