Chapter 2 Numerical Integration

2.1 Introduction

In some simple cases, the calculation of the definite integral

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

(2.1.1)

is directly possible when the primitive (or antiderivative) function $F(x)$ is known

$$\int f(x)dx = F(x)$$

(2.1.2)

hence

$$\int_a^b f(x)dx = F(b) - F(a)$$

(2.1.3)

Most often, this is impossible and the only possible solution is numerical. Frequently, moreover, the function $f(x)$ is only known at a given number of points x_i, i = 0, 1,..., n. In this case, it is possible to search an approximation g(x) of the function $f(x)$ and to proceed to a formal integration.

The interpolation polynomials $P_n(x)$ possess the required approximation properties and are easily integrable. Thus, they will be largely used in numerical integration (also called quadrature).

2.2 Newton and Cotes Closed Integration Formulas

The following integration formulas are called “closed” as they use the two basis points a and b to determine the approximation polynomial.

2.2.1 Global Integration on Interval [a, b]

Consider basis points uniformly distributed on interval [a, b]

$$x_i = a + ih, i = 0, 1, ..., n with h = \frac{b-a}{n}$$

(2.2.1)

Note that n is the degree of the interpolation polynomial $P_n(x)$ such that

$$P_n(x_i) = f(x_i) = f_i, i = 0, 1, ..., n$$

(2.2.2)

For example, a Lagrange polynomial can be chosen as an interpolation polynomial. In this case

$$P_n(x) = \sum_{i=0}^{n} L_i(x) f_i$$

(2.2.3)
with

$$L_i(x) = \Phi_{k=0k\neq i} \frac{x - x_k}{x_i - x_k}$$

(2.2.4)
The variable $t \in [0, n]$ is introduced such that $x = a + ht$. The polynomial $L_i(x)$ becomes

$$L_i(x) = \phi_i(t) = \prod_{k=0k\neq i}^{n} \frac{t-k}{i-k}$$

(2.2.5)
By integrating, we get

$$\begin{array}{c|c|l} \int_a^b P_n(x)dx & = & \sum_{i=0}^n f_i \int_a^b L_i(x)dx \\ & = & h \sum_{i=0}^n f_i \int_0^n \phi_i(t)dt \\ & = & h \sum_{i=0}^n f_i w_i \end{array}$$

(2.2.6)
The coefficients w_i are called weights; they depend only on n, thus they neither depend on the function f nor on the integration limits a and b. Recall that $h = (b - a)/n$.

Example: n = 1

$$beginarraycw_0 = \int_0^1 \frac{t-1}{0-1} dt = \int_0^1 (1-t)dt = \frac{1}{2} w_1 = \int_0^1 \frac{t-0}{1-0} dt = \int_0^1 t dt = \frac{1}{2}$$

(2.2.8)
which gives the following result:

$$\int_a^b P_1(x)dx = \frac{h}{2}(f_0 + f_1) = \frac{h}{2}[f(a) + f(b)]$$

(2.2.9)

corresponding to the trapezoidal rule with h = (b - a) (Figure 2.2).

Example: n = 2

$$w_0 = \int_0^2 \frac{t-1}{0-1}\frac{t-2}{0-2}dt = \frac{1}{2}\int_0^2 (t^2 - 3t + 2)dt = \frac{1}{3}$$

(2.2.10)

$$w_1 = \int_0^2 \frac{t-0}{1-0}\frac{t-2}{1-2}dt = -\int_0^2 (t^2 - 2t)dt = \frac{4}{3}$$

(2.2.11)

$$w_2 = \int_0^2 \frac{t-0}{2-0}\frac{t-1}{2-1}dt = \frac{1}{2}\int_0^2 (t^2 - t)dt = \frac{1}{3}$$

(2.2.12)

which gives the following result:

$$\int_a^b P_2(x)dx = \frac{h}{3}(f_0 + 4f_1 + f_2) = \frac{h}{3}[f(a) + 4f(\frac{a+b}{2}) + f(b)]$$

(2.2.13)

which is Simpson rule with $h = (b-a)/2$ (Figure 2.1).

By continuing, Table 2.1 results for different values of the degree n of the interpo- lation polynomial. From the degree n, the value of s results, then the weights w_i. The values σ_i are introduced only to display a table of integer values instead of fractional weights.

Newton–Cotes formulas thus give the approximation of the integral

$$\int_a^b P_n(x)dx = h\sum_{i=0}^{n} w_i f_i$$

(2.2.14)

with

h = b a n (2.2.15) The weights wi are such that their sum is equal to the degree n of interpolation polynomial

n i=0 wi = n (2.2.16) Let s be the lowest common denominator of the weights wi. The integer numerators i are such that

i = swi (2.2.17)

+2 2 ( ) ( )

+1

Fig. 2.2 Trapezoidal rule Example: For Simpson's rule, according to Table 2.1, we have n = 2, ns = 6, thus s = 3. The weights result w0 = 1/3, w1 = 4/3, w2 = 1/3. Newton–Cotes are now expressed as

b a Pn(x)dx = h n i=0 wi fi = b a ns n i=0 i fi (2.2.18)

The error made by doing the numerical integration is equal to b a Pn(x)dx b a f (x)dx = hp+1K f (p)

() where [a, b] (2.2.19) The values of the degree p and of the constant K only depend on the degree n of the interpolation polynomial. The error being of order p, any polynomial function of degree lower than p will be exactly integrated as the derivative of order p will be zero.

Numerical Methods and Optimization 49 2.2.2 Integration on Subintervals In general, Newton–Cotes formulas are not applied on all the interval [a, b], but on the sequence of subintervals composing [a, b]. The type of subinterval depends on the order of the chosen method. The points xi composing the interval [a, b] are defined by

xi = a + ih , i = 0, 1, ..., N with h = b a

N (2.2.20) It can be noticed that, in the previous formula, the definition of h is different from Equation (2.2.1). N must be chosen in agreement with the order n of the integration formula. • For the trapezoidal rule, a subinterval is defined by [xi, xi+1]. • For Simpson's rule, N is chosen even (the number of calculation points xi is odd), a subinterval is defined by [x2i, x2i+1, x2i+2], i = 0, 1, ..., N/2 1. • For the 3/8 rule, N is a multiple of 3 and a subinterval will be defined by [x3i, x3i+1, x3i+2, x3i+3], i = 0, 1, ..., N/3 1. • Application of the trapezoidal rule: On a subinterval, the trapezoidal rule gives Ii = h 2 [f (xi) + f (xi+1)] (2.2.21)

Applying it to all the interval [a, b], we get I(h) = N 1 i=0 Ii = h

f (a) 2 + f (a + h) + ··· + f (b h) + f (b) 2

= b a 2N

f (a) + f (b) + 2 N 1 i=1 f

a + i b a N

(2.2.22) The function f is assumed to be continuously differentiable. On each subinterval, the error is equal to Ii xi+1 xi f (x)dx = h3 12 f (2) (i) (2.2.23)

Then the sum of the individual errors is I(h) b a f (x)dx = h3 12 N 1 i=0 f (2) (i) = h2 12 b a N N 1 i=0 f (2) (i) (2.2.24)

The summation term can be bounded

min i f (2) (i) 1 N N 1 i=0 f (2) (i) max i f (2) (i) (2.2.25)

As f (2) is continuous, [mini i, maxi i][a, b] such that f (2) () = 1 N N 1 i=0 f (2) (i) (2.2.26)

hence

I(h) b a f (x)dx = b a 12 h2 f (2) () , [a, b] (2.2.27) This result means that the error done when a trapezoidal rule is used decreases like the square of h and thus the method is of order 2. • Application of Simpson's rule: With N even, on each subinterval [x2i, x2i+1, x2i+2], i = 0, 1,..., N/2 1. It gives Ii = h 3 [f (x2i) + 4 f (x2i+1) + f (x2i+2)] with h = b a

N (2.2.28)

By summing these N/2 values, the approximation on [a, b] results I(h) = N /21 i=0 Ii = h 3 [f (a) + 4 f (a + h) + 2 f (a + 2h) + 4 f (a + 3h) + ··· + 2 f (b 2h) + 4 f (b h) + f (b)] = h 3

f (a) + f (b) + 2 N /21 i=1 f (a + 2ih) + 4 N /21 i=0 f (a + (2i + 1)h)

(2.2.29)

The error done is S(h) b a f (x)dx = h5 90 N /21 i=0 f (4) (i) = h4 90 b a 2 2 N N /21 i=0 f (4) (i) (2.2.30) In the same way as for the trapezoidal rule, provided that f is 4 times continuously differentiable, it results that S(h) b a f (x)dx = b a 180 h4 f (4) () (2.2.31)

Thus Simpson's rule is a method of order 4.

2.3 Open Newton and Cotes Integration Formulas The following integration formulas are called "open" as they do not demand one or the other one of the bounds of the integration interval. The interpolation polynomial is of order n 2. Consider n 1 base points regularly spaced x1,..., xn1. It is supposed that the lower integration limit a coincides with x0 = x1 h where h is the spacing between adjacent points. The upper limit b is not fixed. The integration formula is

b a f (x)dx b a Pn2(x)dx (2.3.1)

Thus, by defining

= b x0 h (2.3.2)

we get for = 2

x2 x0 f (x)dx = 2h f (x1) + h3 3 f (2) () (2.3.3)

= 3

x3 x0 f (x)dx = 3h 2 [f (x1) + f (x2)] + 3h3 4 f (2) () (2.3.4)

= 4 x4 x0 f (x)dx = 4h 3 [2 f (x1) f (x2) + 2 f (x3)] + 14h5 45 f (4) () (2.3.5)

= 5 x5 x0 f (x)dx = 5h 24 [11 f (x1) + f (x2) + f (x3) + 11 f (x4)] + 95h5 144 f (4) () (2.3.6) The closed Newton–Cotes formulas are more accurate than the open formulas as soon as a number of points larger than 2 or 3 is used. Thus, in general, it is better to use the closed formulas.

2.4 Conclusions on Newton and Cotes Integration Formulas The formulas with m points for m odd have the same order of accuracy as the formulas with m + 1 points. Their degree of precision is equal to m. A formula of degree of precision m exactly integrates all the polynomials of degree lower than or equal to m. The polynomials of larger degree are not exactly integrated. Thus, Simpson's rule exactly integrates polynomials of degree lower than or equal to 3. Except for the trapezoidal rule used because of its simplicity, it is preferable to use formulas with an odd number of base points than with an even number. The formulas with a number of base points larger than 8 are rarely used. Indeed, the rounding errors become large because of large weight factors with alternate signs. A way to reduce the error is the use of composite integration formulas. Rather than using a formula of a high order, it is often better to choose a formula having a low order, divide the integration interval [a, b] into subintervals, and use the formula of low order separately on each subinterval.

52 Chapter 2. Numerical Integration 2.5 Repeated Integration by Dichotomy and Romberg's Integration Let IN,1 be the estimation of the integral b a f (x)dx (2.5.1) obtained by using the composite trapezoidal rule with a number n of subintervals such that n = 2N . I0,1 is the estimation of the integral obtained by using the simple trapezoidal rule (step = h)

I0,1 = (b a) 1

1 2 [f (a) + f (b)]

(2.5.2) I1,1 is the estimation of the integral obtained by using the simple trapezoidal rule applied two times (step = h/2) I1,1 = (b a) 2

1 2 [f (a) + f (b)] + f

a + (b a) 2

= 1 2

I0,1 + (b a) f

a + (b a) 2 (2.5.3) I2,1 is the estimation of the integral obtained by using the simple trapezoidal rule applied four times (step = h/22) I2,1 = (b a) 4

1 2 [f (a) + f (b)] + 3 i=1 f

a + i (b a) 4

= 1 2 I1,1 + (b a) 2 3 i=1 i=2 f

a + i (b a) 4

(2.5.4)

The recurrence relation relating In,1 (step = h/2n) to In1,1 (step = h/2n1) is thus expressed as

In,1 = 1 2 In1,1 + (b a) 2n1 2n 1 i=1 i=2 f

a + i (b a) 2n

(2.5.5)

The error term corresponding to In,1 is equal to (b a) 3 12(2) 2n f (2) () , [a, b] (2.5.6) Provided that the function f (2) be continuous and bounded, In,1 converges to the exact value of the integral. Richardson's Extrapolation Now, introduce the general technique of Richardson's extrapolation. Given a quantity

Numerical Methods and Optimization 53 gapprox obtained by means of a discretization step h, g can be an integral, a derivative, ..., approximating the exact value gexact. Suppose that the approximation is of order n, hence

gexact = gapprox(h) + 0(hn) (2.5.7)

which could be written as

gexact = gapprox(h) + an hn + an+1 hn+1 + 0(hn+2) (2.5.8) where the coefficients ai depend on the approximation method used. If instead of using a step h, we use h/2, Equation (2.5.8) becomes

gexact = gapprox(h 2) + an hn 2n + an+1 hn+1 2n+1 + 0(hn+2) (2.5.9)

where gapprox(h

2) is in general a better approximation of gexact than gapprox(h). If Equations (2.5.8) and (2.5.9) are combined in order to eliminate the term about hn, the approximation is then improved as the error will be at least about hn+1. We thus get (2n 1) gexact = 2n gapprox(h 2) gapprox(h) an+1 2 hn+1 + 0(hn+2) (2.5.10)

Richardson's extrapolation formula results

gexact = 2n gapprox(h 2) gapprox(h) 2n 1

+ 0(hn+1) (2.5.11) demonstrating that this new formula gives a better result with an approximation of order (n + 1). Application of Richardson's Extrapolation Let us apply Richardson's extrapolation technique to a pair of adjacent elements of the sequence Ii,1 to obtain a better approximation of the integral. Each integral Ik,1 resulting from the trapezoidal rule is obtained with an error of order 2, thus n = 2. The application of Richardson's extrapolation then gives the relation

I = 22 Ik+1,1 Ik,1 22 1 + 0(h3) (2.5.12)

as Ik+1,1 uses a step two times lower than Ik,1. The approximation results

I 22 Ik+1,1 Ik,1 22 1 (2.5.13) Applied to the previous sequence, Richardson's extrapolation (Figure 2.3) gives for two adjacent elements

Ik,2 = 4Ik+1,1 Ik,1

3 (2.5.14)

54 Chapter 2. Numerical Integration For k = 0 corresponding to the trapezoidal rule applied once for I0,1 on the interval [a, b] and twice for I1,1, Richardson's extrapolation yields

I0,2 = 4I1,1 I0,1 3 = (b a) 6 $f(a) + 4f$

a + (b a) 2

+ f (b)

that is Simpson's rule. The error on Ik,2 is equal to

(b a) 5 2880(2) 4k f (4) () , [a, b] (2.5.16)

1 +1,1 (step = h / 2) (step = h / 2 +1) Richardson 2 Fig. 2.3 Richardson's extrapolation By repeating Richardson's extrapolation, Romberg's extrapolation formula results

Ik,j = 4j1Ik+1,j1 Ik,j1

4j1 1 (2.5.17)

The error corresponding to Ik,j is equal to c(j) 22jk f (2j) () , [a, b] (2.5.18)

where c(j) is a constant depending on a, b, j. It is interesting to note that if Ik,j converges to the exact value of the integral when k increases, I0,j also converges when j increases. Thus, in the particular numerical Example 2.1, I0,j is close to the solution to about 106 for j = 10 whereas it is only the case for I13,1. Example 2.1 : Integral by Richardson's extrapolation The following integral has been calculated: I = 3 0 x exp(x2)dx (2.5.19)

by three different methods:

Numerical Methods and Optimization 55 (a) Gauss–Legendre quadrature with 5 points gives I 3963.45 and with 10 points, it gives I 4051.04. (b) Simpson's rule with 5 base points gives I 6441.21, with 11 base points I 4258.18, with 101 base points I 4051.07. (c) Richardson's extrapolation gives the following Table 2.2. The values of the calculated integrals follow the notations of Equation (2.5.17) with tabulated values of Ik, j obtained by iterative use of Richardson's extrapolation formula. It converges to 4051.042.

Table 2.2 Richardson's extrapolation. Values of Ik, j are given k j 1 2345678 1 36,463.878 12,183.089 6058.419 4295.426 4061.334 4051.171 4051.042 4051.042 2 18,253.286 6441.211 4322.973 4062.248 4051.181 4051.042 4051.042 3 9394.230 4455.363 4066.322 4051.224 4051.043 4051.042 4 5690.080 4090.637

4051.460 4051.043 4051.042 5 4490.498 4053.908 4051.050 4051.042 6 4163.056 4051.228 4051.042 7 4079.185 4051.054 8 4058.087

2.6 Numerical Integration with Irregularly Spaced Points Previously, all developed integration formulas were of the form

b a f (x)dx h n i=0 wi f (xi) (2.6.1) where the n + 1 weights wi are known from the n + 1 values xi. When the points xi are not fixed, there are 2n + 2 unknowns (wi and xi), which allows us to determine a polynomial of degree 2n + 1.

2.6.1 Reminder on Orthogonal Polynomials Two functions gm(x) and gn(x) belonging to a family of functions gi(x) are orthogonal with respect to a weight function w(x) on the interval [a, b] when, for all n

¡ gm—gn ¿= b a w(x)gm(x)gn(x)dx = 0 if n m (2.6.2) ¡ gn—gn ¿= b a w(x)[gn(x)]2dx = c(n) 0 (2.6.3)

56 Chapter 2. Numerical Integration where the notation ¡ f —g ¿ is called scalar product of the functions f and g relative to the weight function w. The scalar product is a number. Two functions are orthogonal when their scalar product is zero. A function is normalized when the scalar product of the function by itself is equal to 1. If all orthogonal functions two by two of the ensemble are normalized, the ensemble is orthonormal. In general, the value of c depends on n. A way to generate an ensemble of orthogonal polynomials for a given weight function w(x) is to use the recurrence relation

P1(x) 0 P0(x) 1 Pn+1(x) = (x an)Pn(x) bnPn1(x) , n = 0, 1, 2,... (2.6.4)

with the coefficients defined by an = ¡ xPn—Pn ¿ ¡ Pn—Pn ¿ , n = 0, 1, 2,... bn = ¡ xPn—Pn1 ¿ ¡ Pn1—Pn1 ¿ , n = 1, 2,... , any b0

(2.6.5) To demonstrate Equation (2.6.5), it suffices to consider Equation (2.6.4) and to multiply by w(x)Pn or w(x)Pn1 respectively, then to take the integral of the new equation and to use the properties of orthogonal polynomials. The polynomials defined by (2.6.4) are monic, i.e. the coefficient of the monomial xn of largest degree of Pn(x) is equal to 1. If each polynomial is divided by ¡ Pn—Pn ¿1/2, the ensemble of polynomials becomes orthonormal. Other orthogonal polynomials can be met with different normalizations. Each polynomial Pn(x) has exactly n distinct roots in the interval [a, b]. Among the known families of orthogonal functions, let us cite the family (sin k x) and the family (cos k x). The monomial functions are not orthogonal. On the opposite, there exist several families of orthogonal polynomials. Legendre polynomials: Legendre polynomials Pn(x) are orthogonal on the interval [1, 1] with the unit weight function w(x) = 1 1 1

Pm(x)Pn(x)dx = 0 if n m (2.6.6)

and moreover

1 1 [Pn(x)]2dx = 2

2n + 1 (2.6.7)

The first Legendre polynomials are

Numerical Methods and Optimization 57

P0(x) = 1 P1(x) = x P2(x) = 1 2(3x2 1) P3(x) = 1 2(5x3 3x) P4(x) = 1 8(35x4 30x2 + 3) ... Pn(x) = 2n 1 n xPn1(x) n 1 n Pn2(x)

(2.6.8)

Example 2.2 : Orthogonal Legendre polynomials To determine the orthogonal Legendre polynomials, consider the recurrence (2.6.4) Pn+1(x) = (x an)Pn(x) bnPn1(x) , n = 0, 1, 2,... (2.6.9)

with P0(x) = 1, P1(x) = x and the equations of the coefficients (2.6.5), that is

an = 1 1 xP2 n(x)dx 1 1 P2 n(x)dx , n = 1, 2,...

bn = 1 1 xPn(x)Pn1(x)dx 1 1 P2 n1(x)dx

, n = 1, 2,...

(2.6.10)

If Pn+1(x) is calculated by Equation (2.6.9), we get a monic polynomial (whose coefficient of monomial xn+1 of highest degree is equal to 1) which does not satisfy Equation (2.6.7), thus we must set the Legendre polynomial equal to

PLegn+1(x) = cn+1Pn+1(x) (2.6.11)

where cn+1 is a coefficient such that

1 1 [PLegn+1(x)]2dx = 2 2(n + 1) + 1 = 1 1 [cn+1Pn+1(x)]2dx (2.6.12)

hence

cn+1 = ”’(2 2(n + 1) + 1 1 1 1 [Pn+1(x)]2dx

(2.6.13)

Thus, Table 2.3 results. Table 2.3 Orthogonal monic polynomials P and Legendre polynomials PLeg n an bn Pn+1(x) PLegn+1(x) 1 0 0.3333 x2 0.3333 1.5x2 0.5 2 0 0.2666 x3 0.6x 2.5x3 1.5x 3 0 0.2570 x4 0.8570x2 + 0.0857 4.375x4 3.750x2 + 0.3750

58 Chapter 2. Numerical Integration Laguerre polynomials: Laguerre polynomials Ln(x) are orthogonal on the interval [0, +[with the weight function w(x) = exp(x) + 0 exp(x)Lm(x)Ln(x)dx = 0 if n m (2.6.14) + 0 exp(x)[Ln(x)]2dx = (n + 1)

n! (2.6.15)

The first Laguerre polynomials are L0(x) = 1 L1(x) = x + 1 L2(x) = x2 4x + 2 L3(x) = x3 + 9x2 18x + 6 ... Ln(x) = (2n 1 x)Ln1(x)(n 1) 2Ln2(x)

(2.6.16)

Chebyshev polynomials of the first kind: Chebyshev polynomials of the first kind Tn(x) are orthogonal on the interval [1, 1] with the weight function w(x) = 1/ 1 x2

+1 1 1 1 x2 Tm(x)Tn(x)dx = 0 if n m

+1 1 1 1 x2 [Tn(x)]2dx = 2 if n 0 if n = 0

(2.6.17)

The first Chebyshev polynomials of the first kind are

T0(x) = 1 T1(x) = x T2(x) = 2x2 1 T3(x) = 4x3 3x ... Tn(x) = 2xTn1(x) Tn2(x)

(2.6.18)

Hermite polynomials: Hermite polynomials Hn(x) are orthogonal on the interval], +[with the weight function w(x) = exp(x2) + exp(x2)Hm(x)Hn(x)dx = 0 if n m (2.6.19) + exp(x2)[Hn(x)]2dx = 2nn! n (2.6.20)

The first Hermite polynomials are

H0(x) = 1 H1(x) = 2x H2(x) = 4x2 2 H3(x) = 8x3 12x ... Hn(x) = 2xHn1(x) 2(n 1)Hn2(x)

(2.6.21)

Each of these orthogonal polynomials of degree n with real coefficients has n distinct roots in its definition interval. Any polynomial of degree n can be represented as a linear combination of functions of any of the previous families.

2.6.2 Gauss–Legendre Quadrature

We estimate the integral in the same way as previously by integration of an approxi- mation polynomial of degree n

b a f (x)dx = b a Pn(x)dx + b a Rn(x)dx (2.6.22)

Rn(x) being the error term. Let us use Lagrange interpolation polynomial f (x) = n i=0 Li(x) f (xi) +

n i=0 (x xi)

f (n+1) () (n + 1)! , a ¡¡ b (2.6.23)

with

Li(x) = n j=0 ji x xj xi xj

(2.6.24) The integration interval [a, b] is transformed into [1, 1] by a change of variable

z = 2x (a + b) b a (2.6.25)

and we define F(z) = f (x), hence F(z) = n i=0 Li(z)F(zi) +
n i=0 (z zi)
F(n+1) () (n + 1)! , 1 ¡¡ 1 (2.6.26)
with
Li(z) = n j=0 ji z zj zi zj
(2.6.27)
Supposing that f (x) is a polynomial of degree 2n + 1, then
60 Chapter 2. Numerical Integration
F(n+1) () (n + 1)! = qn() (2.6.28) is a polynomial of degree n. belonging to the interval [1, 1], but having no known value in this interval, to be able to pursue the calculations, qn() is transformed into a polynomial qn(z) on which it will be possible to work. An estimation of the integral is then given by
1 1 F(z)dz n i=0 F(zi) 1 1 Li(z)dz
= n i=0 wiF(zi)
(2.6.29)
with the weights wi = 1 1 Li(z)dz = 1 1 n j=0 ji z zj zi zj
dz (2.6.30) The integral to be calculated on [a, b] is related to the integral calculated after change of variable by b a f (x)dx = b a 2 1 1 F(z)dz b a 2 n i=0 wiF(zi) (2.6.31) Taking into account the previous remark about and qn(), the error term of the quadrature formula takes the form 1 1
n i=0 (z zi)
qn(z)dz (2.6.32)
The abscissas zi must be chosen in order to minimize the error term. Consider the particular case of Gauss–Legendre quadrature. The polynomials n i=0(z zi) and qn(z) are expressed by means of Legendre polynomials Pi
n i=0 (z zi) = n+1 i=0 biPi(z) (2.6.33)
qn(z) = n i=0 ciPi(z) (2.6.34)
The integral to minimize becomes
Numerical Methods and Optimization 61 1 1
n i=0 (z zi)
qn(z)dz =
1 1 n i=0 n j=0 bicjPi(z)Pj(z) + bn+1 n i=0 ciPi(z)Pn+1(z) dz =
1 1 n i=0 bici [Pi(z)] 2 dz =
n i=0 bici 1 1 [Pi(z)] 2 dz
(2.6.35)

as a result of orthogonality properties. The error term can be rendered equal to zero by imposing that the first n + 1 coefficients bi be zero. There remains only one coefficient bn+1 different from zero so that, from Equation (2.6.33) n i=0 (z zi) = bn+1Pn+1(z) (2.6.36) As the coefficient of the term of degree n of the left-hand polynomial is equal to 1, it results that bn+1 is equal to the inverse of the coefficient of the term of highest degree of Pn+1. From the previous equality, it is obvious that the n + 1 base points zi used in the integration formula are the n+1 roots of Legendre polynomial of degree n+1. Now that the base points are determined, the weights wi can be calculated by Equation (2.6.30). Several different methods have been developed to efficiently calculate the weights. In the particular case of Legendre polynomials (Abramowitz and Stegun 1972), it is possible to use

wi = 2 (1 z2 i)[P n+1(zi)]2 (2.6.37) This constitutes Gauss–Legendre quadrature. Gauss–Legendre quadrature gives an exact integration result when the integrated function f is a polynomial of maximum degree (2n + 1). The values of the roots zi and the corresponding weights wi for a family of given orthogonal polynomials are tabulated (Table 2.4). Remark: The calculation of the integral I = b a f (x)dx (2.6.38) is brought back to the calculation of the integral approximated by the sum

I = b a 2 1 1 F(z)dz b a 2 n i=0 wiF(zi) (2.6.39)

62 Chapter 2. Numerical Integration Rather than making the change of variable x → z to determine the function F(z) from f (x), it is simpler to calculate the roots xi on [a, b] corresponding to the roots zi on [1, 1] and to use the equality f (xi) = F(zi). Thus, we get

I = b a 2 n i=0 wi f (xi) (2.6.40)

Table 2.4 Gauss–Legendre quadrature formulas

Gauss–Legendre quadrature 1 1 F(z)dz n i=0 wiF(zi)

Roots zi Type Weights wi ±1/ 3 Formula with 2 points (n = 1) 1.000000000000000 0.000000000000000 Formula with 3 points (n = 2) 8/9 ± 15/5 5/9 ±) 525 70 30/35 Formula with 4 points (n = 3) (18 + 30)/36

±) 525 + 70 30/35 (18 30)/36 0.000000000000000 Formula with 5 points (n = 4) 0.568888888888888 ±0.538469310105683 0.478628670499366 ±0.906179845938664 0.236926885056189 ±0.238619186083197 Formula with 6 points (n = 5) 0.467913934572691 ±0.661209386466265 0.360761573048139 ±0.932469514203152 0.171324492379170 ±0.148874338981631 Formula with 10 points (n = 9) 0.295524224714753 ±0.433395394129247 0.269266719309996 ±0.679409568299024 0.219086362515982 ±0.865063366688985 0.149451349150581 ±0.973906528517172 0.066671344308688

Example 2.3 : Gauss–Legendre quadrature Calculate numerically the following integral:

Numerical Methods and Optimization 63

I = 3 2 x4 dx If it is integrated analytically, the exact value of this integral is I=55. We perform an integration according to Gauss–Legendre quadrature with 3 points (n = 2). We proceed in the following way: • Calculation of the abscissas xi on interval [2, 3] corresponding to the tabulated zeros zi which are in [1, 1]. • Calculation of the values of the function f (xi). • Evaluation of the integral according to Equation (2.6.40). Thus, Table 2.5 results. Table 2.5 Gauss–Legendre quadrature

zi xi f (xi) wi wi f (xi) 0 0.5 0.0625 8 9 0.05555

15 5 2.4364915 35.2419 5 9 19.5788

15 5 1.4364915 4.2580 5 9 2.36555

Finally

I 5 2 (0.05555 + 19.5788 + 2.36555) 55

Notice that, as the polynomial function to integrate had a degree lower than (2n + 1) with n = 2, the integration is exact. Comparison with the trapezoidal rule and Simpson's rule without subintervals: Trapezoidal rule:

I 5 [1 2 (2) 4 + 1 2 (3) 4) 242.5

Simpson's rule:

I 5 2 [1 3 (2) 4 + 4 3 (0.5) 4 + 1 3 (3) 4)] 81.04

The interest of Gauss–Legendre quadrature is obvious with respect to the trapezoidal rule and Simpson's rule.

2.6.3 Gauss–Laguerre Quadrature

Gauss–Laguerre quadrature is based on the same principle as Gauss–Legendre quadra- ture. However, the integration formula takes into account the weight function under

the form + 0 exp(z)F(z)dz = n i=0 wiF(zi) = + 0 G(z)dz = n i=0 wi exp(zi)G(zi) (2.6.41)

64 Chapter 2. Numerical Integration where the zi are the roots of Laguerre polynomial Ln and the weights are equal to

wi = (n!) 2zi (n + 1) 2)[Ln+1(zi)]2 (2.6.42)

2.6.4 Gauss–Chebyshev Quadrature In the case of Chebyshev polynomials of the first kind, Gauss–Chebyshev quadrature gives +1 1 F(z) * 1 z2 dz = n i=0 wiF(zi) = + G(z)dz = n i=0 wi) 1 z2 i G(zi) (2.6.43) where the zi are the roots of Chebyshev polynomial of the first kind Tn equal to

zi = cos (2i 1) 2n

(2.6.44)

and the weights are equal to

wi = n (2.6.45)

2.6.5 Gauss–Hermite Quadrature Gauss–Hermite quadrature gives + exp(z2)F(z)dz = n i=0 wiF(zi) = + G(z)dz = n i=0 wi exp(z2 i)G(zi) (2.6.46) where the zi are the roots of Hermite polynomial Hn and the weights are equal to (by using the orthonormal ensemble of Hermite polynomials) wi = 2 [H n(zi)]2 (2.6.47)

2.7 Discussion and Conclusion Just as in approximation methods, the use of irregularly spaced points imposed by the quadrature method clearly demonstrates its advantage with respect to the accuracy of the result of numerical integration at the expense of a slightly more important work

Numerical Methods and Optimization 65 of understanding and design. Gauss–Legendre quadrature is the most frequently used. At a comparable level of precision, the number of calculations required to evaluate the integral is considerably lower than for the rules based on regularly spaced points. These latter are still used by many users who do not want to invest time, do a simple program, have the impression to master the method, in particular the trapezoidal rule. This shows the interest to use numerical libraries which are available and would provide the precision to them with a relatively reduced investment.

2.8 Exercise Set Exercise 2.8.1 (Easy) Calculate the following integral: 3 2 x4dx (2.8.1) first by the trapezoidal method, then Simpson's method, and finally Gauss–Legendre quadrature with 3 points (n = 2), by using in the three cases the bounds of the integration interval as calculation interval. Comment. Exercise 2.8.2 (Easy) The error function erf(x) is defined by erf(x) = 2 x 0 exp(t 2)dt = 1 2 + x exp(t 2)dt (2.8.2) 1. Using Gauss–Legendre quadrature with 4 points, give an approximation of both integrals for x = 1. For the calculation of the second integral, it will be useful to think about the choice of the bound to use to replace + and the influence of the value of the term exp(t

2). It may be useful to make a few trials to better understand this influence. Discuss the results thus obtained. Deduce an approximation of erf(x). 2. Do the same estimation with the first integral by Simpson's method with a step h = 0.1. Compare the result thus obtained. Remark: The error function is used in many problems of physics. Consider a solid plate where the one-dimensional heat transfer is ruled by Fourier's law

T t = 2 T x2 (2.8.3) subjected to a constant temperature at the interface T(x = 0, t) = Ts. Let T(x, t = 0) = T0 be the initial temperature. The temperature along time (Incropera and DeWitt 1996) is expressed as

T(x, t) Ts T0 Ts = erf x 2 t

(2.8.4)

66 Chapter 2. Numerical Integration where is the thermal diffusivity. Exercise 2.8.3 (Medium) The fugacity f (atm) of a gas at a pressure P (atm) and at temperature T is given by

ln f P = P 0 Z 1 P dP (2.8.5) with the compressibility factor Z = PV/(RT), R gas constant, and V molar volume (Smith et al. 2018; Vidal 1997). For methane, the experimental data of the compressibility factor Z with respect to pressure are given in Table 2.6. Table 2.6 Compressibility factor Z of methane with respect to pressure P

P Z 1 0.9940 10 0.9370 20 0.8683 30 0.7928 40 0.7034 50 0.5936 60 0.4515 80 0.3429 100 0.3767 120 0.4259 140 0.4753 160 0.5252 180 0.5752 200 0.6246

We desire to calculate the fugacity at P = 200 atm. 1. A first crude method would consist in using the trapezoidal method by using only the experimental data. Calculate in this way the fugacity with detailed calculation. 2. A second method would consist in using Gauss–Legendre quadrature. For that purpose, we propose an approximation function of the form Z = 1 0.00858 P 0.000463 P ln(P + 1) + 0.0000475 P2 (2.8.6) By explaining the steps of the calculation, without fully explaining Gauss–Legendre quadrature, calculate the fugacity. Exercise 2.8.4 (Easy) Calculate the following integral: I = + 0 1 1 + x4 dx (2.8.7)

Numerical Methods and Optimization 67 by both methods, 1. Simpson's method with the step h = 0.25. 2. Gauss–Legendre quadrature with five points. Remark: To calculate this integral, it is recommended to divide the integration domain as [0, 1] and [1, +[, and then to do a change of variable t = 1/x on the second domain. Exercise 2.8.5 (Medium) Calculate the following integral: I = 1 1 1x2 1+x2) x2 + y2 dx dy (2.8.8) by Gauss–Legendre quadrature with three points, clearly explaining the technique used and giving intermediate results. Exercise 2.8.6 (Easy) Calculate the following integral: I = +2 2 exp(x2)dx (2.8.9)

by 5-point Gauss–Legendre quadrature.

References M. Abramowitz and I. A. Stegun. Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables. Dover, New York, 1972. F. P. Incropera and D. P. DeWitt. Fundamentals of Heat and Mass Transfer. John Wiley, New York, 4th edition, 1996. J. M. Smith, H. C. Van Ness, M. M. Abbott, and M. T. Swihart.Introduction to Chemical Engineering Thermodynamics. McGraw-Hill, New York, 8th edition, 2018.

J. Vidal. Thermodynamique. Technip, Paris, 1997.

Chapter 3 Equation Solving by Iterative Methods

3.1 Introduction The problem is to develop adequate methods to find the solutions of the general equation

f (x) = 0 (3.1.1) The roots will be noted i. In a given number of cases, f (x) will be supposed to be a polynomial of degree n

f (x) = xn + a1 xn1 + ··· + an1 x1 + an (3.1.2)

but some methods are applicable to any type of function. There exist four main classes of methods (Gritton et al. 2001) to find the roots of a nonlinear equation of the form

f (x) = 0 (3.1.3)

1. Local methods that require an initial estimation of the root (e.g. successive sub- stitutions, Newton). Frequently, local methods are designed to search for only one

real root of the nonlinear equation, even if multiple roots exist. Nevertheless, they are often very robust and nearly always converge (e.g. Newton, quasi-Newton), but they present the drawback to need to provide an initial estimation sufficiently close to the root. 2. Global methods that find a root from an arbitrary initial value (e.g. homotopy). They are adapted to the search of multiple roots. 3. Interval methods that find all the roots in a specified domain of x (e.g. dichotomy, regula falsi). They are robust but slow. 4. Graphical methods or spreadsheet that uses a graphical view of f (x) in a specified domain of x. Some of the methods presented below (Graeffe, Bernoulli, Bairstow) are more interesting from a mathematical point of view than for real applications, but they present a historical interest and their exposure may promote future ideas. J.-P. Corriou, Numerical Methods and Optimization, Springer Optimization and Its Applications 187, https://doi.org/10.1007/978-3-030-89366-8$_3$

69

70 Chapter 3. Equation Solving by Iterative Methods 3.2 Graeffe's Method Graeffe's method is a global method, as it gives a simultaneous approximation of all roots. Consider a monic polynomial f of type (3.1.2). To f , the following adjoint function is associated

(x) = (1) n f (x) f (x) = (x2 2 1)(x2 2 2) ... (x2 2

n), (3.2.1) where i are the searched roots, ordered by decreasing modulus. As (x) contains only even powers, a new function is defined f2(x) = (x) = (x 2 1)(x 2 2) ... (x 2 n) (3.2.2) which has the property that its roots are the squares of the roots of f . The operation can be repeated, and we obtain a sequence of polynomials f2, f4, f8 ... such that

fm(x) = (x m 1)(x m 2) ... (x m

n) (3.2.3)

where m is an integer positive of 2 and fm has roots m 1 , m 2 ,...,m n . The aim of this sequence is to form an equation whose roots have very different orders of magnitude, that is, if the roots are real, the ratios —m i1/m i — can be made as small as desired when

m becomes large. fm(x) can be developed fm(x) = xn (m

1 + ...)xn1 + (m 1 m 2 + ...)xn2

(m 1 m 2 m 3 + ...)xn3 + ··· + (1) n(m 1 m 2 ...m n)

= xn A1 xn1 + ··· + (1)

i Ai xni + ··· + (1) nAn

(3.2.4)

the approximations result m 1 = A1 , m 2 = A2 A1 ,... m n = An An1

(3.2.5) hence an approximation of the absolute values or of the moduli of the searched roots by taking the mth root. The sign of the roots is not determined by this method. It must be verified by substitution in the original equation. If multiple or complex roots exist, —i — = —i+1—, the equation

Ai1 x2 Ai x + Ai+1 = 0 (3.2.6)

gives approximations of m i and m i+1.

Graeffe's method presents some numerical difficulties and is not commonly used. Example 3.1 : Graeffe's method Consider a polynomial with real roots, equal to 1, 2, 3, and 4. This polynomial is equal to

Numerical Methods and Optimization 71 P(x) = x4 + 2 x3 13 x2 14 x + 24 (3.2.7)

Table 3.1 Graeffe's method: root finding for a polynomial having real roots im A1/m 1 (A2/A1)

1/m (A3/A2)

1/m (A4/A3) 1/m 1 2 5.4772 3.0166 1.7331 0.8381 2 4 4.3376 2.9409 1.9173 0.9812 3 8 4.0497 2.9787 1.9904 0.9994 4 16 4.0024 2.9984 1.9998 0.9999 5 32 4.0000 3.0000 2.0000 1.0000

Using Graeffe's method, the successive polynomials are found f2(x) = x4 30 x3 + 273 x2 820 x + 576 f4(x) = x4 354 x3 + 26481 x2 357904 x + 331776 f8(x) = x4 72354 x3 + 448510881 x2 110523752704 x + 110075314176

(3.2.8) This shows that the polynomial coefficients increase very rapidly. In Table 3.1, the values of A1/m 1 , (A2/A1) 1/m, ... , have been gathered to highlight the limits that are the ordered root moduli. The convergence is fast. However, it must be noted that the coefficients Ai very rapidly take very large values, which poses huge numerical problems. In the case of a polynomial with complex roots, Graeffe's method is difficult to use. At the best, it allows to have an estimation of m i .

3.3 Bernoulli's Method Bernoulli's method to find a root k of the polynomial

P(x) = n i=0 ai xni with a0 = 1 (3.3.1) first consists of building a sequence ui by associating to each monomial xnk a term uik (thus i n). To understand the interest of the building of the sequence ui, first consider the fact that the roots i of the polynomial P(x) are supposed to be ordered according to their modulus —1— ¿ ··· ¿ —n—. Express that i is a root a0n 1 + a1n1 1 + ··· + an11 + an = 0 . . . a0n n + a1n1 n + ··· + an1n + an = 0

(3.3.2)

72 Chapter 3. Equation Solving by Iterative Methods Each of the previous equations is then multiplied by an arbitrary coefficient ci, we sum the rows, that is c1(a0n 1 +a1n1 1 +···+an11+an)+···+cn(a0n

n +a1n1 n +···+an1n+an) = 0 (3.3.3)

and then we order again with respect to the coefficients ai, i.e. a0(c1n 1 +···+cnn n)+a1(c1n1 1 +···+cnn1 n)+···+an(c1 +···+cn) = 0 (3.3.4)

We set

ui = c1i 1 + ··· + cni

n , 0 i n (3.3.5)

hence

a0un + a1un1 + ··· + anu0 = 0 (3.3.6)

To simplify the writing, we consider a0 = 1, hence un = a1un1 ··· anu0 = n i=1 aiuni (3.3.7)

By extension, the sequence is defined ui = n j=1 ajuij for i n (3.3.8) Thus, it can be seen that to define this sequence, it suffices to arbitrarily choose the real numbers ui for 0 i ¡ n. From the form of the solutions ui previously given

ui = c1i 1 + ··· + cni

n , 0 i n 1 (3.3.9)

a system of n equations with n unknowns i

j results (Vandermonde determinant). To

find the solution, an initial vector ui as simple as possible can be chosen

ui = 0 , 0 i ¡ n 1 and un1 0 (3.3.10)

Equation (3.3.8) can also be written as

un+k = a1un+k1 ··· anuk (3.3.11)

Now consider the ratio of two successive terms

un+k un+k1 = a1 a2 un+k2 un+k1 ··· an uk un+k1

(3.3.12)

and suppose that this ratio admits a limit l when k →

lim k→ un+k un+k1 = l (3.3.13)

From Equation (3.3.12), we deduce

Numerical Methods and Optimization 73 l = a1 a2 1 l ··· an 1 l n1 l n + a1l n1 + ··· + an = 0 (3.3.14)

hence the limit l is a root of the polynomial. Theorem:

If —1— ¿ —2— , 1 = lim i→ ui ui1 , (3.3.15) thus the limit l corresponds to the root of larger modulus. Bernoulli's method remains valid even when multiple roots exist: 1 = 2 = ··· = j provided we have: —j — ¿ —j+1—. If there are no multiple roots, the unique solution for i 0 is of the form

ui = c1i 1 + ··· + cni

n , 0 i n 1 (3.3.16) the values of the coefficients ci depending on the initial value taken for ui. In the case where a double root exists: 1 = 2 with —2— ¿ —3— , the solution is of the form

ui = c1i 1 + ic2i 1 + c3i 3 + ··· + cni

n (3.3.17) To find the solution, among all possible sequences, it may be useful to associate the two following sequences vi and ti to the sequence ui

vi = u2 i ui+1ui1 (3.3.18)

and

ti = uiui1 ui+1ui2 (3.3.19)

Thus, in the case (in particular, if 1 and 2 are complex) where

—1— = —2— ¿ —3— (3.3.20)

we obtain the following results lim i→ vi vi1 = 12 (3.3.21)

and

lim i→ ti+1 vi = 1 + 2 (3.3.22) Bernoulli's method is rarely employed since the intensive use of computers. Example 3.2 : Bernoulli's method Consider a polynomial with real roots, equal to 1, 2, 3, and 4. This polynomial is P(x) = x4 + 2 x3 13 x2 14 x + 24 (3.3.23) First, the sequence ui is calculated by initializing with u0 = 0, u1 = 0, u2 = 0, u3 = 1. The other terms of this sequence obey Equation (3.3.16). The sequences vi and ti are also calculated according to Equations (3.3.18) and (3.3.19), respectively. The value of 1 results as the root of largest

74 Chapter 3. Equation Solving by Iterative Methods modulus. Equations (3.3.21) and (3.3.22) are also used to calculate the product and the sum of 1 and 2. In Table 3.2, the results of the calculations are gathered. It appears that the ratio ui/ui1 tends toward the root 1 of largest modulus, the ratio vi/vi1 tends toward the product of both roots (12) of larger modulus, and the ratio ti+1/vi tends toward their sum. However, the convergence is slow. Table 3.2 Bernoulli's method: root finding for a polynomial in the case of real roots i ui ui/ui1 vi/vi1 ti+1/vi 0 0 1 0 2 0 3 1 4 2 2 5 17 8.5 15.1538 1.3197 6 46 2.7058 11.7817 0.9358 7 261 5.6739 12.8207 1.1177 8 834 3.1954 11.7885 0.9610 9 4009 4.8069 12.2893 1.0459 10 14102 3.5175 11.8809 0.9796 11 62381 4.4235 12.1143 1.0187 12 231946 3.7182 11.9411 0.9901 13 981201 4.2302 12.0477 1.0079 14 3765918 3.8380 11.9724 0.9953 15 15543061 4.1272 12.0204 1.0034 16 60739538 3.9078 11.9873 0.9978 17 247267193 4.0709 12.0089 1.0014 18 976163494 3.9478 11.9943 0.9990 19 3943413501 4.0397 12.0039 20 15657462810 3.9705

Example 3.3 : Bernoulli's method Consider the following real polynomial

P(x) = x4 + 2 x3 + 3 x2 + 4 x + 5 (3.3.24) This polynomial has complex conjugate roots, equal to 0.2878 ± 1.4160i and 1.2878 ± 0.8578. The respective moduli of these roots are 1.4450 and 1.5474. The sequence ui is initialized exactly like in previous Example 3.2 with u0 = 0, u1 = 0, u2 = 0, u3 = 1. The other terms of this sequence are calculated according to Equation (3.3.16). The sequences vi and ti are also calculated from Equations (3.3.18) and (3.3.19), respectively. In Table 3.3, it can be noticed that, opposite to the previous case where the roots were real, the ratio ui/ui1 is not stabilized, which indicates that the root of largest modulus is complex. Thus, we must deal with the two complex conjugate roots of largest modulus. Equations (3.3.21) and (3.3.22) are used to calculate the product and the sum of 1 and 2 that are conjugate. It appears that the ratio vi/vi1 tends toward the product of both roots (12) (equal to 2.3944) of largest modulus and that the ratio ti+1/vi tends toward their sum (equal to 2.5756). The number of iterations is very large with respect to a very limited convergence.

Numerical Methods and Optimization 75 Table 3.3 Bernoulli's method: root finding for a real polynomial in the case of complex conjugate roots

i ui ui/ui1 vi/vi1 ti+1/vi 0 0 1 0 2 0 3 1 4 1 2 5 1 0.5 0.3333 0 6 0 id. id. id. 7 0 id. id. id. 8 6 id. id. 2.8333 9 17 2.8333 5.3611 1.2538 10 16 0.9411 0.8860 0.4678 20 1746 1.7962 2.6906 1.3435 30 126,576 2.6744 4.0406 2.1753 40 7,484,506 4.6352 2.9735 2.6754 50 340,135,536 6.5020 2.4605 2.6521 60 0.2271 2.4442 2.6057 70 0.6304 2.4000 2.6199 80 1.0668 2.3533 2.6013 90

1.4228 2.3650 2.5732 100 1.8325 2.3912 2.5670

3.4 Bairstow's Method Bairstow's method allows us to find the complex conjugate roots of a real polynomial by noticing that they correspond to the factorization of a real polynomial of degree 2. To calculate these roots by Newton's method, it would be necessary to use numerical complex calculation. In a general manner, a polynomial P(x) P(x) = n i=0 ai xni (3.4.1)

by setting P0(x) P(x) can be written under the form

P0(x) = P1(x)(x2 r x q) + A0 x + B0 (3.4.2) where the polynomial P1(x) is of degree n 2 and the remainder of the division of P0(x) by P1(x) is equal to A0 x + B0. The coefficients A0 and B0 depend on the values of r and q; thus the issue is to find the values of these coefficients that make A0(r, q) and B0(r, q) equal to zero. Indeed, Bairstow's method makes use of Newton–Raphson method that will be examined in Section 5.12, in the solution of systems of nonlinear equations. The recurrence relation giving r and q is

ri+1 qi+1

=

ri qi

A0 r A0 q B0 r B0 q 1

r=ri q=qi

A0(ri, qi) B0(ri, qi)

(3.4.3)

However, A0(r, q) and B0(r, q) are unknown, thus the elements of the Jacobian matrix also. Consequently, it is necessary to determine those four partial derivatives that are present in the following identities: P0(x) r = (x2 r x q) P1(x) r xP1(x) + x A0 r + B0 r 0 (3.4.4)

P0(x) q = (x2 r x q) P1(x) q P1(x) + x A0 q + B0 q 0 (3.4.5) After this first division by (x2 r x q), the operation can be repeated; hence P1(x) = P2(x)(x2 r x q) + A1 x + B1 (3.4.6)

Supposing that (x2 r x q = 0) has two distinct roots x0, x1, we get

P1(xi) = A1 xi + B1 , i = 0, 1 (3.4.7)

and both identities (3.4.4) and (3.4.5) become

xi(A1 xi + B1) + A0 r xi + B0 r = 0

(A1 xi + B1) + A0 q xi + B0 q = 0 , i = 0, 1 (3.4.8) From the second equation of Equation (3.4.8) and from Equation (3.4.7), we draw

A0 q = A1 , B0

q = B1 (3.4.9)

and consequently the first equation of (3.4.8) becomes

x2 i A0 q + xi(A0 r B0 q) + B0 r = 0 , i = 0, 1 (3.4.10)

As x2 i = r xi + q, it gives xi(A0 r B0 q A0 q r) + B0 r A0 q q = 0 , i = 0, 1 (3.4.11)

hence

A0 r B0 q A0 q r = 0 (3.4.12) B0 r A0 q q = 0 (3.4.13)

Numerical Methods and Optimization 77 and finally

A0 q = A1 , B0 q = B1

A0 r = r A1 + B1 , B0 r = qA1

(3.4.14)

By taking

P0(x) = n i=0 ai xni and P1(x) = n2 i=0 bi xn2i (3.4.15)

it results n i=0 ai xni = (x2 r x q) n2 i=0 bi xn2i + A0 x + B0 (3.4.16) By identifying with respect to the successive powers of x by decreasing order, we get the relations a0 = b0 a1 = b1 r b0 a2 = b2 r b1 q b0 ··· ai = bi r bi1 q bi2 , i = 2,..., n 2 ··· an1 = r bn2 q bn3 + A0 an = q bn2 + B0

(3.4.17)

and the values of A0 and B0 result by means of Horner' scheme

b0 = a0 b1 = b0 r + a1 b2 = b0 q + b1 r + a2 ··· bi = bi2 q + bi1 r + ai , i = 2,..., n 2 ··· A0 = bn3 q + bn2 r + an1 B0 = bn2 q + an

(3.4.18)

The process is repeated with P1 and so on until exhaustion. Example 3.4 : Bairstow's method Consider the following real polynomial

P(x) = x4 + 2 x3 + 3 x2 + 4 x + 5 (3.4.19)

having complex conjugate roots, equal to 0.2878 ± 1.4160i and 1.2878 ± 0.8578i. By applying Bairstow's method, the values of A0(r, q) and B0(r, q) are found by applying Equation (3.4.18)

78 Chapter 3. Equation Solving by Iterative Methods b0 = 1 b1 = r + 2 b2 = q + r (r + 2) + 3 A0 = (r + 2) q + (q + (r + 2) r + 3) r + 4 = 2 q r + 2 q + r3 + 2 r2 + 3 r + 4 B0 = (q + (r + 2)r + 3) q + 5 = q2 + q r2 + 2 q r + 3 q + 5

(3.4.20)

Applying Equation (3.4.3), Newton–Raphson's algorithm follows at iteration i

r q

i+1 =

r q

i

2 q + 3 r2 + 4 r + 3 2 r + 2 2 q r + 2 q 2 q + r2 + 2 r + 3 1 i

A0 B0

i (3.4.21) The initialization is simply done with (r, q) = (0, 0). After a limited number of iterations, Table 3.4 results. Table 3.4 Bairstow's method: roots solving for a real polynomial having complex conjugate roots Iteration ir q A0 B0 1 00 4 5 2 0.2222 1.6667 0.8285 3.4362 3 1.0132 1.3463 4.7119 1.3364 4 0.5601 1.6794 1.2433 0.374 5 0.5514 2.0700 0.0071 0.1630 6 0.5760 2.0880 0.0013 0.0023 7 0.5756 2.0882 0.62 × 106 0.029 × 106

Having found (A0, B0)(0, 0), the searched polynomial results

x2 r x q = x2 0.5756 x + 2.0882 (3.4.22) whose roots are 0.2878±1.4160 i. In a general way, then it would be necessary to continue the method by applying it to the polynomial P1 obtained by exact division of P(x) by (x2 r x q). In the present case, as P1(x) is of degree 2, the solution is immediate.

3.5 Existence of a Root of a Function Before searching for the root of a continuous function f by an iterative method such that f () = 0, it is essential to find an interval [a, b] such that

f (a) f (b) ¡ 0 (3.5.1) This guarantees the existence of a solution in the interval [a, b]. Finding the interval [a, b] is the initial step of bracketing.

To compare different root-finding methods, the following equation will be consid- ered

f (x) = exp(x) x2