PART E Numeric Analysis

The subdivision into three chapters has been retained. All three chapters have been updated in the light of computer requirements and developments. A list of suppliers of software (with addresses, etc.) can be found at the beginning of Part E of the book and another list at the beginning of Part G.

CHAPTER 19 Numerics in General

SECTION 19.1. Introduction, page 788

Purpose. To familiarize the student with some facts of numerical work in general, regardless of the kind of problem or the choice of method.

Main Content, Important Concepts

Floating-point representation of numbers, overflow, underflow

Roundoff

Concept of algorithm

Stability

Errors in numerics, their propagation, error estimates

Loss of significant digits

Short Courses. Mention the roundoff rule and the definitions of error and relative error.

SOLUTIONS TO PROBLEM SET 19.1, page 794

1. $-0.7644 \cdot 10^{2}$, $0.6010 \cdot 10^{6}$, $-0.1000 \cdot 10^{-4}$

Numerics needs practical experience. Experience cannot be *taught* but it must be *gained.* However, the present problem set should serve as an eye opener, illustrating various aspects and some (not all!) unexpected facts occurring in numerics.

8. Check backward: $(0.10011)_2 = 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 0 \cdot 2^{-3} + 1 \cdot 2^{-4} + 1 \cdot 2^{-5} = 0.59375$.

10. A binary machine number is a (finite!) sum of terms each of which has a finite decimal representation.

The converse is not true. Take 0.1, for example.

Prove that a number a is a binary machine number if and only if $a = m/2^n$, where m and n are integers. Indeed, if its binary representation is finite, you can convert it to this form by taking the common denominator. And conversely.

11. No, because $\bar{a}_m$ (the computer number for $1/m$) is zero from some m on.

This is typical. Indeed, more generally, since convergence of a series of numbers implies that the terms must approach zero, the corresponding computer numbers must be zero from some m on, which means convergence.

SECTION 19.2. Solution of Equations by Iteration, page 795

Purpose. Discussion of the most important methods for solving equation $f(x) = 0$, a very important task in practice.

Main Content, Important Concepts

Solution of $f(x) = 0$, by iteration (3) $x_{n+1} = g(x_n)$

Condition sufficient for convergence (Theorem 1)

Newton (–Raphson) method (5)

Speed of convergence, order

Secant, bisection, false position methods

Comments on Content

Fixed-point iteration gives the opportunity to discuss the idea of a **fixed point**, which is also of basic significance in modern theoretical work (existence and uniqueness of solutions of differential, integral, and other functional equations).

The less important *method of bisection* and *method of false position* are included in the problem set.

SOLUTIONS TO PROBLEM SET 19.2, page 804

1. $x = \sqrt[4]{x + 0.12}$; 1, 1.02874, $\cdots$ 1.03717 (6S exact, 7 steps)

4. 2, 1.537902, 1.557099, 1.557146, 1.557146

SECTION 19.3. Interpolation, page 805

Purpose. To discuss methods for interpolating (or extrapolating) given data $(x_0, f_0), \cdots, (x_n, f_n)$, all x_j different, arbitrarily or equally spaced, by polynomials of degree not exceeding n.

Main Content, Important Concepts

Lagrange interpolation (4) (arbitrary spacing)

Error estimate (5)

Newton's divided difference formula (10) (arbitrary spacing)

Newton's difference formulas (14), (18) (equal spacing)

Short Courses. Lagrange's formula briefly, Newton's forward difference formula (14).

Comment on Content

For given data, the interpolation polynomial $p_n(x)$ is unique, regardless of the method by which it is derived. Hence the error estimate (5) is generally valid (provided f is $n + 1$ times continuously differentiable).

SOLUTIONS TO PROBLEM SET 19.3, page 816

1. From

$$L_0(x) = x^2 - 20.5x + 104.5$$

$$L_1(x) = \frac{1}{0.75}\,(-x^2 + 20x - 99)$$

$$L_2(x) = \frac{1}{3}\,(x^2 - 18.5x + 85.5)$$

(see Example 2) and the 5S-values of the logarithm in the text we obtain

$$p_2(x) = -0.005233x^2 + 0.205017x + 0.775950.$$

This gives the values and errors

2.2407, error 0
2.3028, error −0.0002
2.3517, error −0.0003
2.4416, error 0.0007
2.4826, error 0.0024.

It illustrates that in extrapolation one may usually get less accurate values than one does in interpolation. $p_2(x)$ would change if we took more accurate values of the logarithm.

Small changes in initial values can produce large differences in final values.

Examples in which the difference in accuracy between interpolation and extrapolation is larger can easily be constructed.

3. The difference table is

x_j	$f(x_j)$	1st Diff.	2nd Diff.	3rd Diff
1.0	0.94608			
		0.37860		
1.5	1.32468		−0.09787	
		0.28073		−0.00975
2.0	1.60541		−0.10762	
		0.17311		
2.5	1.77852			

The interpolating values and errors are

$$p_1(1.25) = f(1.0) + 0.5 \cdot 0.37860 = 1.13538\ (\epsilon = 0.011017)$$

$$p_2(1.25) = p_1(1.25) + \frac{0.5(-0.5)}{2} \cdot (-0.09787) = 1.14761\ (\epsilon = -0.00116)$$

$$p_3(1.25) = p_2(1.25) + \frac{0.5(-0.5)(-1.5)}{6} \cdot (-0.00975) = 1.14700\ (\epsilon = -0.00055)$$

Note the decrease of the error.

4. $p_2(x) = 1.0000 - 0.0112r + 0.0008r\,(r-1)/2 = x^2 - 2.580x + 2.580$,
$r = (x-1)/0.02$; 0.9943, 0.9835, 0.9735, exact to 4S

6. From (10) and the difference table

j	x	$f(x)$	$f[x_j, x_{j+1}]$	$f[x_j, x_{j+1}, x_{j+2}]$
0	0.25	0.27633		
			0.97667	
1	0.5	0.5250		−0.44304
			0.64640	
2	1.0	0.84270		

we obtain

$$p_2(0.75) = 0.2763 + (0.75 - 0.25) \cdot 0.9767 + (0.75 - 0.25)(0.75 - 0.5) \cdot (-0.4430) = 0.7093,$$

in agreement with Prob. 9, except for a roundoff error.

8. With the change in j the difference table is

j	x_j	$f_j = \cosh x_j$	∇f_j	$\nabla^2 f_j$	$\nabla^3 f_j$
−3	0.5	1.127626			
			0.057839		
−2	0.6	1.185465		0.011865	
			0.069704		0.000697
−1	0.7	1.255169		0.012562	
			0.082266		
0	0.8	1.337435			

From this and (18) we obtain

$$p_3(x) = 1.337435 + 0.082266 \cdot \frac{x - 0.8}{0.1} + 0.0012562 \cdot \frac{(x - 0.8)(x - 0.7)}{0.01 \cdot 2!} + 0.000697 \cdot \frac{(x - 0.8)(x - 0.7)(x - 0.6)}{0.001 \cdot 3!}$$

and with $x = 0.56$ this becomes

$$1.337435 + 0.082266(-2.4) + 0.012562(-2.4)(-1.4)/2 + 0.000697(-2.4)(-1.4)(-0.4)/6 = 1.160945.$$

This agrees with Example 5. The correct last digit is 1 (instead of 5 here or 4 in Example 5).

SECTION 19.4. Spline Interpolation, page 817

Purpose. Interpolation of data $(x_0, f_0), \cdots, (x_n, f_n)$ by a (cubic) spline, that is, a twice continuously differentiable function

$$g(x)$$

which, in each of the intervals between adjacent nodes, is given by a polynomial of third degree at most,

$$\text{on } [x_0, x_1] \text{ by } q_0(x), \qquad \text{on } [x_1, x_2] \text{ by } q_1(x) \qquad \cdots, \qquad \text{on } [x_{n-1}, x_n] \text{ by } q_{n-1}(x).$$

Short Courses. This section may be omitted.

Comments on Content

Higher order polynomials tend to oscillate between nodes—the polynomial $P_{10}(x)$ in Fig. 434 is typical—and splines were introduced to avoid that phenomenon. This motivates their application.

It is stated in the text that splines also help lay the foundation of **CAD (computer-aided design)**.

If we impose the additional condition (3) with given k_0 and k_n (tangent direction of the spline at the beginning and at the end of the total interval considered), then for given data the cubic spline is unique.

SOLUTIONS TO PROBLEM SET 19.4, page 822

5. This derivation is simple and straightforward.

$p_2(x) = x^2$ $[f(x) - p_2(x)]' = 4x^3 - 2x = 0$ gives the points of maximum deviation $x = \pm 1/\sqrt{2}$ and by inserting this, the maximum deviation itself,

$$|f(1/\sqrt{2}) - p_2(1/\sqrt{2})| = |\tfrac{1}{4} - \tfrac{1}{2}| = \tfrac{1}{4}.$$

For the spline $g(x)$ we get, taking $x \geqq 0$,

$$[f(x) - g(x)]' = 4x^3 + 2x - 6x^2 = 0.$$

A solution is $x = \frac{1}{2}$. The corresponding maximum deviation is

$$f(\tfrac{1}{2}) - g(\tfrac{1}{2}) = \tfrac{1}{16} - (-\tfrac{1}{4} + 2 \cdot \tfrac{1}{8}) = \tfrac{1}{16},$$

which is merely 25% of the previous value.

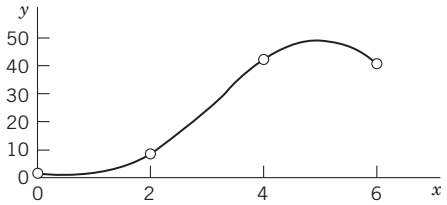
6. $n = 3$, $h = 2$, so that (14) is

$$\begin{aligned} k_0 + 4k_1 + \ \ k_2 \qquad &= \tfrac{3}{2}(f_2 - f_0) = 60 \\ k_1 + 4k_2 + k_3 &= \tfrac{3}{2}(f_3 - f_1) = 48. \end{aligned}$$

Since $k_0 = 0$ and $k_3 = -12$, the solution is $k_1 = 12$, $k_2 = 12$.

In (13) with $j = 0$ we have $a_{00} = f_0 = 1$, $a_{01} = k_0 = 0$,

$$\begin{aligned} a_{02} &= \tfrac{3}{4}(9 - 1) - \tfrac{1}{2}(12 + 0) = 0 \\ a_{03} &= \tfrac{2}{8}(1 - 9) + \tfrac{1}{4}(12 + 0) = 1. \end{aligned}$$



Section 19.4. Spline in Prob. 6

From this and, similarly, from (13) with $j = 1$ and $j = 2$ we get the spline $g(x)$ consisting of the three polynomials (see the figure)

$$q_0(x) = 1 + x^3 \qquad (0 \leqq x \leqq 2)$$

$$q_1(x) = 9 + 12(x - 2) + 6(x - 2)^2 - 2(x - 2)^3 = 25 - 36x + 18x^2 - 2x^3 \qquad (2 \leqq x \leqq 4)$$

$$q_2(x) = 41 + 12(x - 4) - 6(x - 4)^2 = -103 + 60x - 6x^2 \qquad (4 \leqq x \leqq 6).$$

8. $q_0(x) = 2 + x^3$, $q_1(x) = 3 + 3(x - 1) + 3(x - 1)^2 - (x - 1)^3$, $q_2(x) = 8 + 6(x - 2) - 2(x - 2)^3$. Note that this is not a natural spline because $g''(3) = -12 \neq 0$.

10. We obtain

$$q_0 = 2 + x^2 - x^3$$
$$q_1 = -2 - 8(x-2) - 5(x-2)^2 + 5(x-2)^3$$
$$q_2 = 2 + 32(x-4) + 25(x-4)^2 - 11(x-4)^3.$$

The data of Prob. 10 are obtained from those of Prob. 9 by subtracting 2 from the f-values, leaving k_0 and k_3 as they were. Hence, to obtain the answer to Prob. 10, subtract 2 from each of the three polynomials in the answer to Prob. 9.

SECTION 19.5. Numeric Integration and Differentiation, page 824

Purpose. Evaluation of integrals of empirical functions, functions not integrable by elementary methods, etc.

Main Content, Important Concepts

Simpson's rule (7) (most important), error (8), (10)

Trapezoidal rule (2), error (4), (5)

Gaussian integration

Degree of precision of an integration formula

Adaptive integration with Simpson's rule (Example 6)

Numerical differentiation

Short Courses. Discuss and apply Simpson's rule.

Comments on Content

The range of numerical integration includes empirical functions, as measured or recorded in experiments, functions that cannot be integrated by the usual methods, or functions that can be integrated by those methods but lead to expressions whose computational evaluation would be more complicated than direct numerical integration of the integral itself.

Simpson's rule approximates the integrand by quadratic parabolas. Approximations by higher order polynomials are possible, but lead to formulas that are generally less practical.

Numerical differentiation can sometimes be avoided by changing the mathematical model of the problem.

SOLUTIONS TO PROBLEM SET 19.5, page 836

3. 0.693150. Exact to 6S: ln 2 = 0.693147; hence Simpson's rule here gives a 5S-exact value.

4. 0.07392816. Exact to 7S: $0.07392811 = -\frac{1}{2}(\exp(-0.4^2) - 1)$

5. 0.785398514 (9S-exact 0.785398164)

6. $C = -0.5^4/90$ in (9), $-0.000694 \leqq \epsilon \leqq -0.000094$ (actual error -0.000292). In (10),

$$\epsilon_{0.5} \approx \tfrac{1}{15}(0.864956 - 0.868951) = -0.000266.$$

Note that the absolute value of this is less than that of the actual error, and we must carefully distinguish between bounds and approximate values.

8. From (10) and Prob. 1 we obtain

$$0.94608693 + \tfrac{1}{15}(0.94608693 - 0.946145) = 0.946083$$

which is exact to 6S, the error being 7 units of the 8th decimal.

10. We obtain

$$\frac{1}{24}\left(4\sum_{j=1}^{5}\sin\left(\frac{1}{4}j-\frac{1}{8}\right)^2+2\sum_{k=1}^{4}\sin\frac{1}{16}k^2+\sin\frac{25}{16}\right)=0.545941.$$

The exact 6S-value is 0.545962.

13. $x=(t+1)/1.6$ gives $x=0$ when $t=-1$ and $x=1.25$ when $t=1$. Also, $dx=dt/1.6$. The computation gives 0.545963, the 6S-value of the Fresnel integral S(1.25) is 0.545962.

14. Differentiating (14) in Sec. 19.3 with respect to r and using $dr=dx/h$ we get

$$\frac{df(x)}{dr}=hf'(x)\approx\Delta f_0+\frac{2r-1}{2!}\,\Delta^2 f_0+\frac{3r^2-6r+2}{3!}\,\Delta^3 f_0+\cdots.$$

Now $x=x_0$ gives $r=(x-x_0)/h=0$ and the desired formula follows.

SOLUTIONS TO CHAPTER 19 REVIEW QUESTIONS AND PROBLEMS, page 837

3. $-3.145\leqq d\leqq -3.035$

5. $x_1=50+7\sqrt{51}=99.990$, $x_2=0.01$, $x_2=1/99.990=0.010001$

7. 0.5, 0.924207, 0.829106, 0.824146, 0.824132, 0.824132.
Answer: ± 0.824132

9. 0.406 (3S-exact 0.411)

10. $q_0(x)=x^3$, $q_1(x)=1+3(x-1)+3(x-1)^2-(x-1)^3$, $q_2(x)=6+6(x-2)-2(x-2)^3$; $p=-\frac{8}{3}x+\frac{9}{2}x^2-\frac{5}{6}x^3$

12. 0.90450 (5S-exact 0.90452)

13. **(a)** $(-0+0.4^3)/0.4=0.16$, **(b)** $(-0.1^2+0.3^3)/0.2=0.13$, exact 0.12