PART E. NUMERICAL METHODS

CHAPTER 17. Numerical Methods in General

Sec. 17.1 Introduction

Problem Set 17.1. Page 836

7. **Quadratic equation.** Given $x^2 - 30x + 1 = 0$. First use (6), where $a = 1$, $b = -30$, and $c = 1$. Calculating with 4S, you obtain $\sqrt{(-30)^2 - 4} = \sqrt{896} = 29.93$. Hence

$$x_1 = (30 + 29.93)/2 = 59.93/2 = 29.96$$

and

$$x_2 = (30 - 29.93)/2 = 0.07/2 = 0.04.$$

Now use (7). The root x_1 equals 29.96, as before. For x_2 you now obtain

$$x_2 = \frac{c}{ax_1} = 1/29.96 = 0.03338.$$

With 2S the calculations are as follows. You have to calculate the square root of

$$90 \cdot 10^1 - 4 = 90 \cdot 10^1$$

(remember that on the right you may retain only two significant digits) or, differently written,

$$0.90 \cdot 10^3 - 0.40 \cdot 10^1 = 0.90 \cdot 10^3.$$

With 2S, this gives 30. Hence by (6),

$$x_1 = (30 + 30)/2 = 60/2 = 30$$

and

$$x_2 = (30 - 30)/2 = 0.$$

In contrast, from (7) you obtain better results for the second root. You have $x_1 = 30$, as before, and

$$x_2 = 1/x_1 = 1/30 = 0.033.$$

The point of this and similar examples and problems is not to show that calculations with fewer significant digits generally give inferior results (this is fairly plain, although not always the case). The point is to show in terms of simple numbers what will happen in principle, regardless of the number of digits used in a calculation. Here, formula (6) illustrates the loss of significant digits, easily recognizable when we work with pencil (or calculator) and paper, but difficult to spot in a long calculation in which only a few intermediate results are printed out. This explains the necessity of developing programs that are virtually free of possible cancellation effects.

9. **Change of formula.** Given

$$\sqrt{9 + x^2} - 3, \tag{A}$$

where $|x|$ is small. Multiplication and division by

$$\sqrt{9 + x^2} + 3 \tag{B}$$

gives the numerator

$$\left(\sqrt{9 + x^2}\right)^2 - 9 = 9 + x^2 - 9 = x^2$$

and the denominator (B), thus

$$x^2/\left(\sqrt{9 + x^2} + 3\right). \tag{C}$$

For instance, if $x = 0.1$ and you use 4S, you obtain from (A)

$$\sqrt{9.01} - 3 = 3.002 - 3.000 = 0.002.$$

The improved formula (C) gives

$$0.01000/(3.002 + 3.000) = 0.01000/6.002 = 0.001666.$$

The 10S-value is 0.001666 203961.

17. Rounding and adding. For instance, in rounding to, say, 1D, the given numbers $a_1 = 1.03$ and $a_2 = 0.24$ you get $\tilde{a}_1 = 1.0$ and $\tilde{a}_2 = 0.2$, hence the sum 1.2. But if you add first, you obtain 1.27. Rounded to 1D this gives 1.3, which is a more accurate approximation of the true value 1.27 than the approximation 1.2 obtained before. In terms of general formulas you have

$$\tilde{a}_1 = a_1 - \epsilon_1$$

$$\tilde{a}_2 = a_2 - \epsilon_2,$$

where ϵ_1 and ϵ_2 are the errors due to rounding, hence they are less than or equal to 1/2 unit of the last decimal in absolute value. If you round first and add then, you add the rounded numbers $\tilde{a}_1$ and $\tilde{a}_2$, that is,

$$\tilde{a}_1 + \tilde{a}_2 = a_1 + a_2 - (\epsilon_1 + \epsilon_2).$$

You see that in this case the error $\epsilon_1 + \epsilon_2$ is a number between 0 and 1 unit of the last decimal in absolute value. But if you add first, the sum is $a_1 + a_2$, and in rounding it you make an error between 0 and 1/2 unit of the last decimal in absolute value. Similarly for n numbers, where the sum of the rounded numbers is a number with an error between 0 and $n/2$ units of the last decimal in absolute value, whereas in adding and then rounding the error is between 0 and 1/2 unit of the last decimal in absolute value, as before in the case of two numbers.

Sec. 17.2 Solution of Equations by Iteration

Problem Set 17.2. Page 847

1. Nonmonotonicity (as in Example 2) occurs if $g(x)$ is monotone decreasing, that is,

$$g(x_1) \leqq g(x_2) \quad \text{if} \quad x_1 > x_2. \tag{A}$$

(Make a sketch to better understand the reasoning.) Then

$$g(x) \geqq g(s) \quad \text{if and only if} \quad x \leqq s \tag{B}$$

and

$$g(x) \leqq g(s) \quad \text{if and only if} \quad x \geqq s. \tag{C}$$

Start from an $x_1 > s$. Then $g(x_1) \leqq g(s)$ by (C). If $g(x_1) = g(s)$ (which could happen if $g(x)$ is constant between s and x_1), then x_1 is a solution of $f(x) = 0$, and you are done. If $g(x_1) < g(s)$, then by the definition of x_2 (formula (3) in the text) and since s is a fixed point ($s = g(s)$), you obtain

$$x_2 = g(x_1) < g(s) = s \quad \text{so that} \quad x_2 < s.$$

Hence by (B),

$$g(x_2) \geqq g(s).$$

The equality sign would give a solution, as before. Strict inequality and the use of (3) in the text give

$$x_3 = g(x_2) > g(s) = s, \quad \text{so that} \quad x_3 > s,$$

and so on. This gives a sequence of values that are alternatingly larger and smaller than s, as illustrated in Fig. 395 of the text.

11. Newton's method. The derivation of this and similar formulas is schematical. Denote the quantity to be computed by x, that is,

$$x = \sqrt[3]{7}.$$

Then try to find an equation for x, in many cases an equation by which x is (explicitly or implicitly)

defined. In the present problem, using the definition of a cube root, you have

$$x^3 = 7.$$

The equation obtained is written as $f(x) = 0$, simply by collecting all the terms of the equation on the left side. In our case, $f(x) = x^3 - 7 = 0$. You also need $f'(x) = 3x^2$. With this, you can now set up the basic relation of Newton's method. This is equation (5) in the algorithm in Sec. 17.2,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^3 - 7}{3x_n^2} = \frac{2}{3}x_n + \frac{7}{3x_n^2}.$$

Some computational operations are avoided by pulling out the factor 1/3,

$$x_{n+1} = \frac{1}{3}\left(2x_n + \frac{7}{x_n^2}\right).$$

21. Secant method. You have $f(x) = \cos x \cosh x - 1$. Hence formula (10) gives

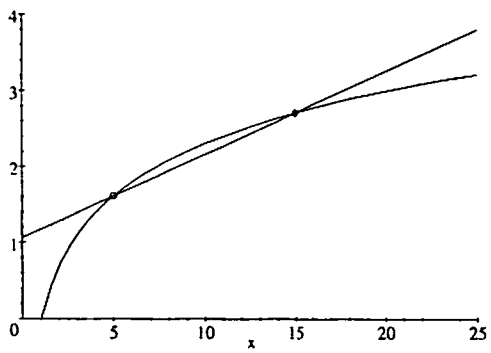
$$x_{n+1} = x_n - (\cos x_n \cosh x_n - 1)\frac{x_n - x_{n-1}}{\cos x_n \cosh x_n - \cos x_{n-1} \cosh x_{n-1}}.$$

In the answer on p. A38 in Appendix 2 the first value listed is the suggested $x_1 = 5$. From $x_0 = 4$ and $x_1 = 5$ you obtain $x_2 = 4.48457$, and so on. The convergence is slower than in Prob. 17 for Newton's method. The sequence of approximate values is not monotone, in contrast to that in Prob. 17 (but these properties are not typical, they depend on the kind of curve you are dealing with).

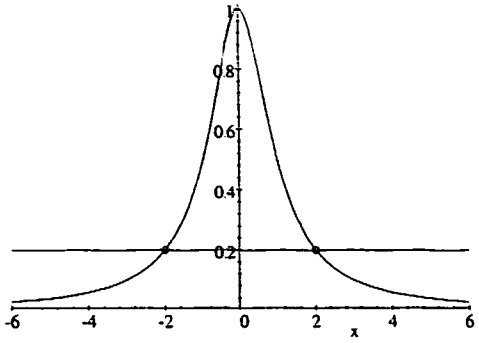
Sec. 17.3 Interpolation

Problem Set 17.3. Page 860

7. Extrapolation. In the case of extrapolation the various factors tend to be larger than in the case of interpolation because in the latter case the point of interpolation lies more "in the middle" between the nodes. In general, interpolation will give better results than extrapolation far enough away from the nodes. However, our simple figures illustrate that we cannot make statements that are always true. In Fig. A, interpolation gives better results than extrapolation at points much smaller than 5 or much larger than 15. In Fig. B, extrapolation is more accurate than interpolation near $x = 0$. Of course, these naive examples should merely make you aware of similar possibilities in more complicated cases in which you cannot see immediately what is going on.



Section 17.3. Problem 7. Fig. A. Interpolation and extrapolation. (Logarithmic curve)

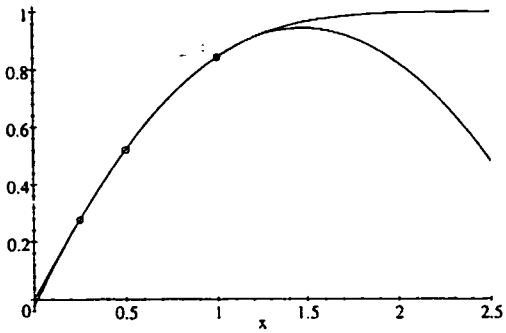


Section 17.3. Problem 7. Fig. B. Interpolation and Extrapolation ($y = 1/(1+x^2)$).

9. **Lagrange polynomial for the error function.** From (3) and the given data you obtain the Lagrange polynomial

$$p_2(x) = \frac{(x-0.5)(x-1.0)}{-0.25(-0.75)}0.27633 + \frac{(x-0.25)(x-1.0)}{0.25(-0.5)}0.52050 + \frac{(x-0.25)(x-0.5)}{0.75 \cdot 0.5}0.84270.$$

Expanding and simplifying, you obtain the answer given on p. A38 of Appendix 2. The approximate value $p_2(0.75) = 0.70929$ is not very accurate. The exact 5D-value is erf (0.75) = 0.71116.



Section 17.3. Problem 9. erf (x) and Lagrange polynomial $p_2(x)$ (lower curve)

11. **Newton's forward difference formula (14)** applies to the given data since these are equally spaced, with $h = 0.02$. Set up a difference table as in Example 5, but containing one column less because you have only three given x-values $x_0 = 1.00$, $x_1 = 1.02$, $x_2 = 1.04$ and corresponding function values of the gamma function rounded to 4D. In (14) you need $\Gamma(1.00) = 1$, $\Delta^1 = -0.0112$, and $\Delta^2 = 0.0008$. With this you can read $p_2(x)$ in the answer on p. A38 of Appendix 2 in terms of r directly from (14). Then calculate $r = (x - x_0)/h = (x - 1.00)/0.02 = 50(x - 1)$. In this r you can substitute $x = 1.01, 1.03, 1.05$ and then calculate p_2 by using the corresponding $r = 0.5, 1.5, 2.5$, respectively. Or you can convert p_2 from r to x (which amounts to expanding p_2 in powers of x, as shown in the answer) and then substitute the x-values into this polynomial in x. The 4D-values in the answer are correct, also the last one (obtained by extrapolation).

15. **Newton's divided difference formula (10)** is less frequently used in practice than Newton's formulas for equally spaced data, which occur more often. Example 4 illustrates that the difference table contains the divided differences needed in calculating those that appear in (10); the latter are circled. In Prob. 15 use erf (0.25) = 0.27633. Then calculate the two first divided differences. The first of them is

$$f[0.25, 0.50] = \text{erf}\,(0.5) - \text{erf}\,(0.25)/(0.50 - 0.25)$$
$$= 4(0.52050 - 0.27633)$$
$$= 0.97668$$

and appears in (10). The second of them is

$$f[0.50, 1.00] = \text{erf}\,(1.0) - \text{erf}\,(0.5)/(1.0 - 0.5)$$
$$= 2(0.84270 - 0.52050)$$
$$= 0.6444$$

and is needed for calculating the second divided difference (you have only one because you have only three nodes, three function values). You obtain

$$f[0.25, 0.50, 1.00] = (f[0.50, 1.00] - f[0.25, 0.50])/(1.00 - 0.25)$$
$$= (0.6444 - 0.97668)/0.75 = -0.44304.$$

This is the last coefficient needed in (10). From this and (10) you obtain the expression for $p_2(x)$ given in the answer. Developing it in powers of x, you obtain the same polynomial in x as in Prob. 9 obtained by Lagrange's method. This illustrates the fact mentioned in the text, that the Lagrange's and Newton's formulas merely give different forms of the same interpolation polynomial, which is uniquely determined by the given data.

Sec. 17.4 Splines

Problem Set 17.4. Page 867

3. **Derivation of (7) and (8) from (6).** The point of the problem is that you minimize a chance of errors by introducing suitable short notations. For instance, for the expressions involving x you may set

$$X_j = x - x_j, \qquad X_{j+1} = x - x_{j+1},$$

and for the occurring constant quantities in (6) you may choose the short notations

$$A = f(x_j)c_j^2, \quad B = 2c_j, \quad C = f(x_{j+1})c_j^2, \quad D = k_j c_j^2, \quad E = k_{j+1}c_j^2.$$

Then formula (6) becomes simply

$$p_j(x) = AX_{j+1}^2(1 + BX_j) + CX_j^2(1 - BX_{j+1}) + DX_jX_{j+1}^2 + EX_j^2X_{j+1}.$$

Differentiate this twice with respect to x, applying the product rule for the second derivative, that is,

$$(uv)'' = u''v + 2u'v' + uv'',$$

and noting that the first derivative of X_j is simply 1, and so is that of X_{j+1}. (Of course, you may do the differentiations in two steps if you want.) You obtain

$$p_j''(x) = A(2(1 + BX_j) + 4X_{j+1}B + 0) + C(2(1 - BX_{j+1}) + 4X_j(-B) + 0) \qquad \text{(I)}$$
$$+ D(0 + 4X_{j+1} + 2X_j) + E(2X_{j+1} + 4X_j + 0),$$

where $4 = 2 \cdot 2$ with one 2 resulting from the product rule and the other from differentiating a square. And the zeros arise from factors whose second derivative is zero. Now calculate p_j'' at $x = x_j$. Since $X_j = x - x_j$, you see that $X_j = 0$ at $x = x_j$. Hence in each line the term containing X_j disappears. This gives

$$p_j''(x_j) = A(2 + 4BX_{j+1}) + C(2 - 2BX_{j+1}) + 4DX_{j+1} + 2EX_{j+1}.$$

Also, when $x = x_j$, then $X_{j+1} = x_j - x_{j+1} = -1/c_j$ (see the formula without number between (4) and (5), which defines c_j). Inserting this as well as the expressions for $A, B, \ldots, E$, you obtain (7). Indeed,

$$p_j''(x_j) = f(x_j)c_j^2\left(2 + 2 \cdot \frac{4c_j}{-c_j}\right) + f(x_{j+1})c_j^2\left(2 - 2 \cdot \frac{2c_j}{-c_j}\right) + \frac{4k_jc_j^2}{-c_j} + \frac{2k_{j+1}c_j^2}{-c_j}$$

and cancellation of some of the factors c_j gives

$$p_j''(x_j) = -6f(x_j)c_j^2 + 6f(x_{j+1})c_j^2 - 4k_j c_j - 2k_{j+1}c_j.$$

The derivation of (8) is similar. For $x = x_{j+1}$ you have $X_{j+1} = x_{j+1} - x_{j+1} = 0$, so that (I) simplifies to

$$p_j''(x_{j+1}) = A(2 + 2BX_j) + C(2 - 4BX_j) + 2DX_j + 4EX_j.$$

Furthermore, for $x = x_{j+1}$ you have $X_j = x_{j+1} - x_j = 1/c_j$, and by substituting $A, \ldots, E$ into the last equation you obtain

$$p_j''(x_{j+1}) = f(X_j)c_j^2\left(2 + \frac{4c_j}{c_j}\right) + f(x_{j+1})c_j^2\left(2 - \frac{8c_j}{c_j}\right) + \frac{2k_j c_j^2}{c_j} + \frac{4k_{j+1}c_j^2}{c_j}.$$

Cancellation of some factors c_j and simplification fianlly gives (8), that is,

$$p_j''(x_{j+1}) = 6c_j^2 f(x_j) - 6c_j^2 f(x_{j+1}) + 2c_j k_j + 4c_j k_{j+1}.$$

11. Determination of a spline. Proceed as in Example 1. Arrange the given data in a table for easier work.

j	x_j	$f(x_j)$	k_j
0	−1	0	0
1	0	4	
2	1	0	0

Since there are three nodes, the spline will consist of two polynomials, $p_0(x)$ and $p_1(x)$. The polynomial $p_0(x)$ gives the spline for x from −1 to 0, and $p_1(x)$ gives the spline for x from 0 to 1.

Step 1. Since $n = 2$, you have just one equation in (12), from which you can determine k_1. The equation is obtained by taking $j = 1$ and noting that $h = 1$; thus

$$k_0 + 4k_1 = \frac{3}{1}(f_2 - f_0) = 0.$$

Hence $k_1 = 0$. Geometrically this means that at $x = 0$ the spline will have a horizontal tangent.

Step 2 for $p_0(x)$. Determine the coefficients of the spline from (14). You see that in general, $j = 0, \ldots, n-1$, so that in the present case you have $j = 0$ (this will give the spline from −1 to 0) and $j = 1$ (which will give the other half of the spline, from 0 to 1). Take $j = 0$. Then (14) gives

$$a_{00} = p_0(p_0) = f_0 = 0$$

$$a_{01} = p_0'(x_0) = k_0 = 0$$

$$a_{02} = \frac{1}{2}p_0''(x_0) = \frac{3}{1^2}(f_1 - f_0) - \frac{1}{1}(k_1 - 2k_0) = 3\cdot 4 - 0 = 12$$

$$a_{03} = \frac{1}{6}p_0'''(x_0) = \frac{2}{1^3}(f_0 - f_1) + \frac{1}{1^2}(k_1 + k_0) = 2\cdot(-4) + 0 = -8.$$

With these Taylor coefficients you obtain from (13) the first half of the spline in the form

$$\begin{aligned} p_0(x) &= a_{00} + a_{01}(x - x_0) + a_{02}(x - x_0)^2 + a_{03}(x - x_0)^3 \\ &= 0 + 0 + 12(x - (-1))^2 - 8(x - (-1))^3 \\ &= 12x^2 + 24x + 12 - 8(x^3 + 3x^2 + 3x + 1) = 4 - 12x^2 - 8x^3. \end{aligned}$$

Step 2 for $p_1(x)$. This is slightly simpler because $x_j = x_1 = 0$, so that (13) will give powers of x directly. From the given data and (14) with $j = 1$ you obtain the Taylor coefficients

$$a_{10} = p_1(x_1) = f_1 = 4$$

$$a_{11} = p_1'(x_1) = k_1 = 0$$

$$a_{12} = \frac{1}{2}p_1''(x_1) = \frac{3}{1^2}(f_2 - f_1) - \frac{1}{1}(k_2 + 2k_1) = 3\cdot(-4) - 0 = -12$$

$$a_{13} = \frac{1}{6}p_1'''(x_1) = \frac{2}{1^3}(f_2 - f_1) + \frac{1}{1^2}(k_2 + k_1) = 2\cdot 4 + 0 = 8.$$

With these coefficients and $x_1 = 0$ you obtain from (13) with $j = 1$ the polynomial

$$p_1(x) = 4 - 12x^2 + 8x^3,$$

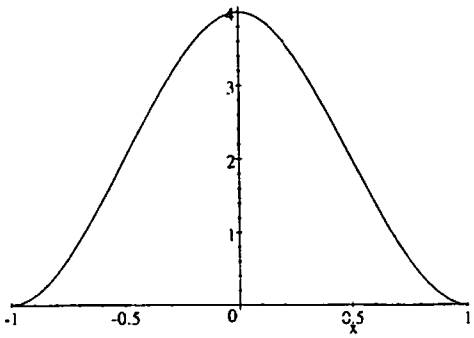
giving the spline on the interval from 0 to 1. As a check of the answer, you should verify that the spline gives the function values $f(x_j)$ and the values k_j of the derivatives in the table at the beginning. Also make sure that the first and second derivatives of the spline at 0 are continuous by verifying that

$$p_0'(0) = p_1'(0) = 0 \quad \text{and} \quad p_0''(0) = p_1''(0) = -24.$$

The third derivative is no longer continuous,

$$p_0'''(0) = -48 \quad \text{but} \quad p_1'''(0) = 48.$$

(Otherwise the spline would consist of a single cubic polynomial for the whole x-interval from -1 to 1.)



Section 17.4. Problem 11. Spline

Sec. 17.5 Numerical Integration and Differentiation

Problem Set 17.5. Page 880

5. **Error estimate (5) for the trapezoidal rule (2).** In (5) you need two approximate values. Since you calculate the integral

$$J = \int_0^1 \sin\left(\frac{\pi x}{2}\right) dx = -\cos\left(\frac{\pi x}{2}\right) \Big/ \frac{\pi}{2} \Big|_0^1 = \frac{2}{\pi} = 0.63662 \qquad \text{(A)}$$

by (2) for three choices of h, namely, for $h = 1$, 1/2, 1/4, you can make two error estimates (5). Sketch the integrand to see what is going on. Now apply the trapezoidal rule (2). By using the exact 5D-value 0.63662 in (A) you can immediately determine the actual error, which we write after each result obtained from (2). The trapezoidal rule (2) with $h = 1$ gives

$$J_{1.0} = 1.0(0 + (1/2)\cdot 1) = 0.50000. \qquad \text{Error} \quad 0.13662. \qquad \text{(B)}$$

With $h = 0.5$ you have the x-values 0, 1/2, 1, for which the integrand has the values 0, $1/\sqrt{2} = 0.70711$, 1, respectively, so that (2) gives

$$J_{0.5} = 0.5(0 + 0.70711 + (1/2)\cdot 1) = 0.60355. \qquad \text{Error} \quad 0.03307. \qquad \text{(C)}$$

With $h = 0.25$ you have the cosine values just used plus 0.38268 at $x = 1/4$ and 0.92388 at $x = 3/4$, so that (2) gives

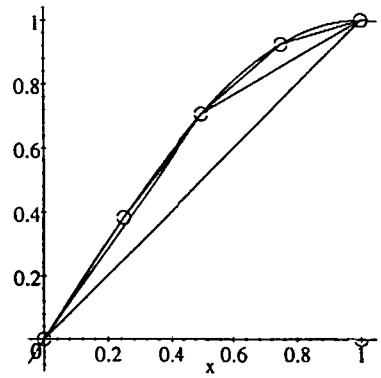
$$J_{0.25} = 0.25(0 + 0.38268 + 0.70711 + 0.92388 + 0.50000) = 0.62842. \qquad \text{Error} \quad 0.00820. \qquad \text{(D)}$$

Note that the error (3) contains the factor h^2. Hence in halving you can expect the error to be multiplied by about $(1/2)^2 = 1/4$. This property is nicely reflected by the numerical values in (B)-(D). Now turn to error estimating by (5). You obtain

$$\epsilon_{0.5} \approx \frac{1}{3}(J_{0.5} - J_{1.0}) = \frac{1}{3}(0.60355 - 0.50000) = 0.03452$$

$$\epsilon_{0.25} \approx \frac{1}{3}(J_{0.25} - J_{0.5}) = \frac{1}{3}(0.62842 - 0.60355) = 0.00829.$$

The agreement of these estimates with the actual value of the errors is very good, Although in other cases the difference between estimate and actual value may be larger, estimation will still serve its purpose. namely, to give an impression of the order of magnitude of the error.



Section 17.5 Problem 5. Given sine curve and approximating polygons in the three trapezoidal rules used

21. Three-eights rule. For the present problem, this rule is very practical because the values of the integrand needed are simple,

$$\cos 30^\circ = \cos \frac{1}{6}\pi = \frac{1}{2}\sqrt{3},$$

$$\cos 60^\circ = \cos \frac{1}{3}\pi = \frac{1}{2}.$$

Also, the fourth derivative in the error term is simply $\cos \hat{t}$. You thus obtain

$$J = \int_0^{\pi/2} \cos x \, dx \approx \frac{3}{8} \cdot \frac{\pi}{6}\left(1 + 3 \cdot \frac{\sqrt{3}}{2} + 3 \cdot \frac{1}{2} + 0\right) - \frac{\pi/2}{80}\left(\frac{\pi}{6}\right)^4 \cos \hat{t} \qquad \text{(E)}$$

$$= \frac{\pi}{16} \cdot 5.098076 - 0.001476 \cos \hat{t} = 1.001005 - 0.001476 \cos \hat{t}.$$

Note that this approximation 1.001005 is much inferior to that in Prob. 23 obtained by Gauss integration with almost as little work as in the present problem. Error bounds are now readily obtained from (E) by noting that in the interval of integration, $\cos \tilde{t}$ varies between 0 and 1. Hence $\cos \pi/2 = 0$ gives the upper bound 0 for the error, and $\cos 0 = 1$ gives the lower bound $-0.001476 \cdot 1 = -0.001476$ for the error. From this and (E) you have $1.001005 - 0.001476 = 0.999529$. Hence bounds for the approximate value $\hat{J} = 1.001005$ of $J = 1$ given by (E) are

$$1.001005 - 0.001476 \cdot 1 = 0.999529 \leqq \hat{J} \leqq 1.001005.$$

23. Gauss integration. The answer on p. A39 shows that the transformation of a given integral to the standard interval $-1 \leqq x \leqq 1$ can often be avoided. This gives an additional reduction of the amount of work involved in this integration. You see that you obtain almost 7D accuracy with very little work. This result is much more accurate than that in Prob. 21 just considered.