Chap 19. Numerics in General

19.1 Introduction

Floating-Point Form of Numbers

The number of significant digits is fixed, whereas the decimal point is floating.

Ex: 0.6247×10^3, 0.1735×10^{-13}, -0.2000×10^{-1}

o. Significant digit

Ex: 1360, 1.360, 0.001360 ⇒ each has 4 significant digits

$$\left(a = \pm m \cdot 10^e, \quad 0.1 \leq m < 1, \quad e: \text{integer} \right)$$

On the computer,

$$\bar{a} = \pm \bar{m} \cdot 10^e, \quad \left(\begin{array}{l} \bar{m} = 0.d_1 d_2 \cdots d_k, \quad d_1 > 0 \\ |e| < M \end{array} \right)$$

m or $\bar{m}$: mantissa, e: exponent

Underflow ⇒ the result is usually set to zero

Overflow ⇒ the computer halts

Roundoff Error: caused by chopping or rounding

Let $\bar{a} = fl(a)$, then $\left| \dfrac{a - \bar{a}}{a} \right| = \left| \dfrac{m - \bar{m}}{m} \right| \leq \underbrace{\dfrac{1}{2} \cdot 10^{1-k}}_{=u}$

$\bar{a} = a(1+\delta)$, $|\delta| \leq u$ $\qquad$ u: rounding unit

$$\frac{\bar{a} - a}{a} = \delta$$

Algorithm, Stability

Numerical instability can be avoided by a better algorithm.

Mathematical instability of a problem is called "ill-conditioning".

Errors of Numeric Results

$a = \hat{a} + \varepsilon$ → error

↳ an approximate value

Relative error: $\varepsilon_r = \dfrac{\varepsilon}{a} = \dfrac{a-\tilde{a}}{a} \approx \dfrac{a-\tilde{a}}{\tilde{a}}$

Error bound: $|\varepsilon| \leqq \beta$, $|a-\tilde{a}| \leqq \beta$

↳ In practice, only an error bound is known

Similarly, for the relative error,

$$|\varepsilon_r| \leqq \beta_r , \quad \left|\frac{a-\tilde{a}}{a}\right| \leqq \beta_r$$

Error Propagation

Th1 $x = \hat{x} + \varepsilon_1$, $y = \hat{y} + \varepsilon_2$, $|\varepsilon_1| \leqq \beta_1$, $|\varepsilon_2| \leqq \beta_2$.

ⓐ Addition and Subtraction: $|\varepsilon| = |\varepsilon_1 \pm \varepsilon_2| \leqq |\varepsilon_1| + |\varepsilon_2| \leqq \beta_1 + \beta_2$

ⓑ Multiplication and division: $|\varepsilon_r| \leqq |\varepsilon_{r_1}| + |\varepsilon_{r_2}| \leqq \beta_{r_1} + \beta_{r_2}$.

<proof>

ⓐ $|\varepsilon| = |x \pm y - (\hat{x} \pm \hat{y})| = |\varepsilon_1 \pm \varepsilon_2| \leqq |\varepsilon_1| + |\varepsilon_2| \leqq \beta_1 + \beta_2$

ⓑ $$|\varepsilon_r| = \left|\frac{xy - \hat{x}\hat{y}}{xy}\right| = \left|\frac{xy - (x-\varepsilon_1)(y-\varepsilon_2)}{xy}\right|$$

$$= \left|\frac{\varepsilon_1 y + \varepsilon_2 x - \varepsilon_1\varepsilon_2}{xy}\right| \approx \left|\frac{\varepsilon_1 y + \varepsilon_2 x}{xy}\right| \leqq \left|\frac{\varepsilon_1}{x}\right| + \left|\frac{\varepsilon_2}{y}\right|$$

$$= |\varepsilon_{r_1}| + |\varepsilon_{r_2}| \leqq \beta_{r_1} + \beta_{r_2}$$

Loss of Significant Digits

Ex1 $x^2 - 40x + 2 = 0$

$\Rightarrow x_1 = 20 + \sqrt{398} = 20.00 + 19.95 = 39.95$

$x_2 = 20 - 19.95 = 0.05$ → Poor!

$x_2 = c/(ax_1) = 2.000/39.95 = 0.05006$ → Better!

(Remark: If $|x_1| >> |x_2|$, try $x_2 = \frac{c}{ax_1}$!)

19.2 Solution of $f(x)=0$ by Iteration

Fixed-Point Iteration

Transform $f(x)=0$ algebraically into $x=g(x)$. → a fixed point of g.

Then choose an x_0 and compute $x_{n+1}=g(x_n)$, $(n=0,1,2,\cdots)$

Th1 (Convergence of Fixed-Point Iteration)

$s=g(s)$, g is C^1-continuous on J, $(s\in J)$

$|g'(x)|\leqq K<1$, for all $x\in J$.

$\Rightarrow \{g(x_n)\}$ converges for any $x_0\in J$.

<proof>

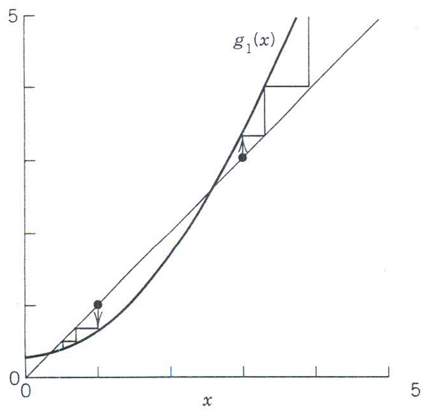
$$|x_n-s|=|g(x_{n-1})-g(s)|=|g'(t)|\,|x_{n-1}-s|\leqq K\cdot|x_{n-1}-s|$$
$$\leqq\cdots\leqq K^n\cdot|x_0-s|\to 0 \text{ as } n\to\infty$$

Ex 1

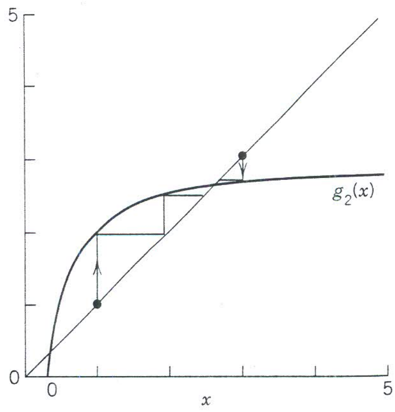
$f(x)=x^2-3x+1=0$

(a) $x=g_1(x)=\frac{1}{3}(x^2+1)$, (b) $x=g_2(x)=3-\frac{1}{x}$

$x_{n+1}=\frac{1}{3}(x_n^2+1)$ $\qquad\qquad$ $x_{n+1}=3-\frac{1}{x_n}$



(a)



(b)

Newton's Method

$$f(x_{n+1}) \approx f(x_n) + (x_{n+1} - x_n) f'(x_n) = 0$$

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

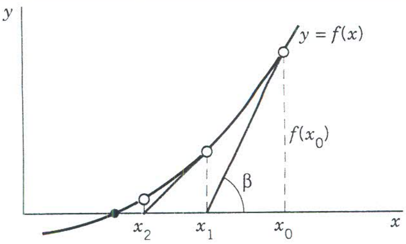


Fig. 425. Newton's method

Ex 3

$$f(x) = x^2 - c, \quad f'(x) = 2x$$

$$\Rightarrow x_{n+1} = x_n - \frac{x_n^2 - c}{2x_n} = \frac{1}{2}\left(x_n + \frac{c}{x_n}\right)$$

Ex 4

$$f(x) = x - 2\sin x, \quad f'(x) = 1 - 2\cos x$$

$$\Rightarrow x_{n+1} = x_n - \frac{x_n - 2\sin x_n}{1 - 2\cos x_n} = \frac{2(\sin x_n - x_n \cos x_n)}{1 - 2\cos x_n}$$

Speed of Convergence

Let $x_n = s - \varepsilon_n$, where ε_n is the error of x_n.

$$x_{n+1} = g(x_n) = g(s) + g'(s)(x_n - s) + \frac{1}{2} g''(s)(x_n - s)^2 + \cdots$$

$$= g(s) - g'(s)\cdot\varepsilon_n + \frac{1}{2} g''(s)\cdot\varepsilon_n^2 + \cdots$$

$$\varepsilon_{n+1} = s - x_{n+1} = g(s) - x_{n+1} = g'(s)\cdot\varepsilon_n - \frac{1}{2} g''(s)\cdot\varepsilon_n^2 + \cdots$$

(a) $\varepsilon_{n+1} \approx g'(s)\cdot\varepsilon_n$ (first order)

(b) $\varepsilon_{n+1} \approx -\frac{1}{2} g''(s)\cdot\varepsilon_n^2$ if $g'(s) = 0$ (second order)

[If $\varepsilon_n = 10^{-k}$, for second order, $\varepsilon_{n+1} = \text{const}\cdot 10^{-2k}$
double in significant digits]

Convergence of Newton's Method

$$g(x) = x - \frac{f(x)}{f'(x)}, \quad g'(x) = \frac{f(x)\cdot f''(x)}{f'(x)^2}, \quad g'(s) = 0 \quad (\because f(s) = 0)$$

$\Rightarrow$ Newton's Method is at least of second order

$$g''(s) = \frac{f''(s)}{f'(s)} \text{ is not zero in general}$$

Th2 (Second-Order Convergence of Newton's Method)

$f(x)$: three times differentiable, $f'(s) \neq 0$, $f''(s) \neq 0$

$\Rightarrow$ Newton's Method is of second order for x_0 (sufficiently close to s)

Comments: $\varepsilon_{n+1} \approx -\dfrac{f''(s)}{2f'(s)} \cdot \varepsilon_n^2$

$\Rightarrow$ s needs to be a simple zero of $f(x)$

(i.e., $f(s)=0$, but $f'(s) \neq 0$)

Ex6: $f(x) = x - 2\sin x$, $x_0 = 2.0$, $x_1 = 1.901$

$\Rightarrow$ Estimate how many iteration steps we need to produce the solution to 5D accuraracy.

<sol> $\dfrac{f''(s)}{2f'(s)} \approx \dfrac{f''(x_1)}{2f'(x_1)} = \dfrac{2\sin x_1}{2(1-2\cos x_1)} \approx 0.57$

$|\varepsilon_{n+1}| \approx 0.57\,\varepsilon_n^2 \approx 0.57^3\,\varepsilon_{n-1}^4 \approx 0.57^M \varepsilon_0^{M+1} \leq 5 \cdot 10^{-6}$

where $M = 2^{n+1} - 1$.

$$\left[\begin{array}{l} \varepsilon_1 - \varepsilon_0 = (\varepsilon_1 - s) - (\varepsilon_0 - s) = -x_1 + x_0 \approx 0.10 \\ \varepsilon_1 = \varepsilon_0 + 0.10 \approx -0.57\,\varepsilon_0^2 \text{ or } 0.57\,\varepsilon_0^2 + \varepsilon_0 + 0.10 \approx 0 \\ \therefore \varepsilon_0 \approx -0.11 \end{array}\right]$$

$\Rightarrow$ $0.57^M \cdot 0.11^{M+1} \leq 5 \cdot 10^{-6}$

Hence, $n=2$ is the smallest possible n.

Ex7: Ill-Conditioned Equation

$f(x) = x^5 + 10^{-4}x = 0$ is ill-conditioned at $x=0$

$f'(0) = 10^{-4}$ is small. at $\tilde{s} = 0.1$, $f(0.1) = 2 \cdot 10^{-5}$ is small.

But, the error $0 - 0.1 = -0.1$ is larger in absolute value than $f(0.1) = 2 \cdot 10^{-5}$ by a factor 5000.

Secant Method

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

$$\Rightarrow x_{n+1} = x_n - f(x_n) \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \qquad (*)$$

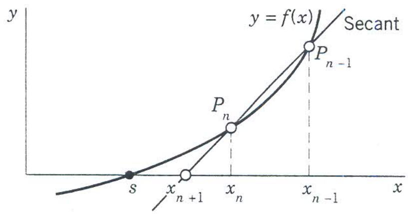


Fig. 426. Secant method

Superlinear Convergence (i.e., $|\varepsilon_{n+1}| \approx \text{const} \cdot |\varepsilon_n|^{1.69}$)

Warning: It is no good to write $(*)$ as follows

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

$\Rightarrow$ This may lead to loss of significant digits if x_n and x_{n-1} are about equal

Ex 8: $f(x) = x - 2\sin x$

$$x_{n+1} = x_n - \frac{(x_n - 2\sin x_n)(x_n - x_{n-1})}{x_n - x_{n-1} + 2(\sin x_{n-1} - \sin x_n)} = x_n - \frac{N_n}{D_n}$$

n	x_{n-1}	x_n	N_n	D_n	$x_{n+1} - x_n$
1	2.000 000	1.900 000	−0.000 740	−0.174 005	−0.004 253
2	1.900 000	1.895 747	−0.000 002	−0.006 986	−0.000 252
3	1.895 747	1.895 494	0		0

$x_3 = 1.895\,494$ is exact to 6D. See Example 4. ■

19.3 Interpolation

Given $f_0 = f(x_0), f_1 = f(x_1), \cdots, f_n = f(x_n)$ at nodes $x_0, \cdots, x_n$, find an interpolation polynomial $P_n(x)$ of degree n (or less) s.t. $P_n(x_0) = f_0, P_n(x_1) = f_1, \cdots, P_n(x_n) = f_n$.

Lagrange Interpolation

$$f(x) \approx P_n(x) = \sum_{k=0}^{n} L_k(x) f_k,$$

where $$L_k(x) = \frac{\prod_{i=0\,(i\neq k)}^{n} (x - x_i)}{\prod_{i=0\,(i\neq k)}^{n} (x_k - x_i)}$$

$$L_k(x_j) = \delta_{k,j} = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{otherwise} \end{cases}$$

Undesirable Oscillations for Large n

If n is large, $P_n(x)$ may tend to oscillate for x between the nodes $x_0, \cdots, x_n$. We must be prepared for numerical instability.

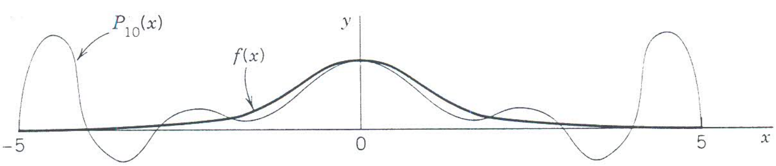


Fig. 431. Runge's example $f(x) = 1/(1 + x^2)$ and interpolating polynomial $P_{10}(x)$

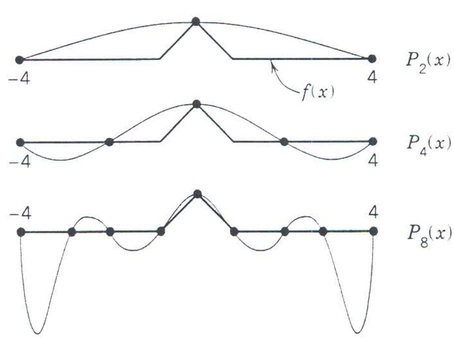


Fig. 432. Piecewise linear function $f(x)$ and interpolation polynomials of increasing degrees

19.4 Spline Interpolation

Th 1 (Existence and Uniqueness of Cubic Splines)

$f(x)$: defined on $[a,b]$ s.t. $a = x_0 < x_1 < \cdots < x_n = b$.

Let k_0 and k_n be any given numbers.

$\Rightarrow \exists 1\ g(x)$: cubic spline s.t.

$$g(x_0) = f(x_0) = f_0, \ \cdots, \ g(x_n) = f(x_n) = f_n,$$
$$g'(x_0) = k_0, \quad g'(x_n) = k_n.$$

<proof>

On each interval $I_j = [x_j, x_{j+1}]$, $g(x)$ is defined by a cubic polynomial $p_j(x)$ s.t

$$p_j(x_j) = f_j, \ p_j(x_{j+1}) = f_{j+1}, \ p_j'(x_j) = k_j, \ p_j'(x_{j+1}) = k_{j+1}$$

for some k_j and k_{j+1}

Assume $p_{j-1}''(x_j) = p_j''(x_j) \quad (j = 1, \cdots, n-1)$

$$\Rightarrow \ c_{j-1} k_{j-1} + 2(c_{j-1} + c_j) k_j + c_j k_{j+1} = 3\left[c_{j-1}^2 \nabla f_j + c_j^2 \nabla f_{j+1}\right]$$

where $c_j = \dfrac{1}{x_{j+1} - x_j}$ and $\nabla f_j = f(x_j) - f(x_{j-1})$

$$\Rightarrow \begin{bmatrix} 2(c_0+c_1) & c_1 & & & \\ c_1 & 2(c_1+c_2) & c_2 & & \\ & c_2 & 2(c_2+c_3) & c_3 & \\ & & & \ddots & \\ & & & c_{n-2} & 2(c_{n-2}+c_{n-1}) \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_{n-1} \end{bmatrix}$$

$$= \begin{bmatrix} 3\left[c_0^2 \nabla f_1 + c_1^2 \nabla f_2\right] - c_0 k_0 \\ 3\left[c_1^2 \nabla f_2 + c_2^2 \nabla f_3\right] \\ \vdots \\ 3\left[c_{n-2}^2 \nabla f_{n-1} + c_{n-1}^2 \nabla f_n\right] - c_{n-1} k_n \end{bmatrix}$$

Since $2(c_{j-1} + c_j) > c_{j-1} + c_j$, the matrix is diag. dominant and there is a unique solution.

Clamped Condition: $g'(x_0)=f'(x_0)$, $g'(x_n)=f'(x_n)$
Free or Natural Cond: $g''(x_0)=0$, $g''(x_n)=0$

Equidistant Nodes

$x_0,\ x_1=x_0+h,\ \cdots,\ x_i=x_0+ih,\ \cdots$

$k_{j-1}+4k_j+k_{j+1}=\frac{3}{h}(f_{j+1}-f_{j-1})$ for $j=1,2,\cdots,n-1$.

Ex1: Interpolate $f(x)=x^4$, $-1\leq x\leq 1$,
by a cubic spline $g(x)$, $x_0=-1, x_1=0, x_2=1$.
under clamped condition: $g'(-1)=f'(-1)$, $g'(1)=f'(1)$.

<sol>

$f_0=f(-1)=1$, $f_1=f(0)=0$, $f_2=f(1)=1$

$k_0=f'(-1)=-4$, $k_2=f'(1)=4$

$k_0+4k_1+k_2=\frac{3}{1}(f_2-f_0)=0 \Rightarrow k_1=0$

$$\Rightarrow \begin{cases} p_0(x)=-x^2-2x^3, & -1\leq x\leq 0 \\ p_1(x)=-x^2+2x^3, & 0\leq x\leq 1 \end{cases}$$

Ex2: Interpolate $f_0=f(0)=1$, $f_1=f(2)=9$, $f_2=f(4)=41$, $f_3=f(6)=41$.
by a cubic spline $g(x)$ s.t $k_0=0$, $k_3=-12$

<Solution>

$n=3$, $h=2$

$$\begin{cases} k_0+4k_1+k_2=\frac{3}{2}(f_2-f_0)=60 \\ k_1+4k_2+k_3=\frac{3}{2}(f_3-f_1)=48 \end{cases} \Rightarrow \begin{matrix} k_1=12 \\ k_2=12 \end{matrix}$$

$$\Rightarrow \begin{cases} p_0(x)=1+x^3, & (0\leq x\leq 2) \\ p_1(x)=25-36x+18x^2-2x^3, & (2\leq x\leq 4) \\ p_2(x)=-103+60x-6x^2, & (4\leq x\leq 6) \end{cases}$$

19.5 Numerical Integration

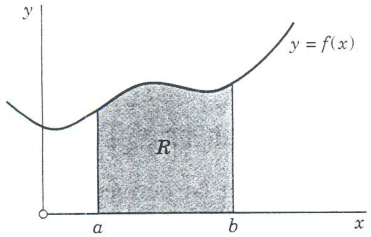


Fig. 437. Geometric interpretation of a definite integral

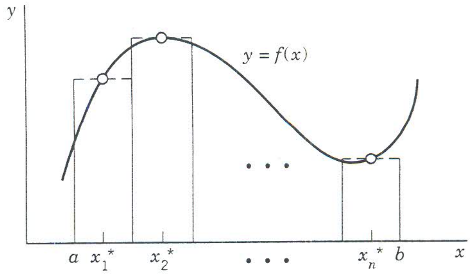


Fig. 438. Rectangular rule

Rectangular Rule

$$J = \int_a^b f(x)\,dx \approx h\left[f(x_1^*) + f(x_2^*) + \cdots + f(x_n^*)\right]$$

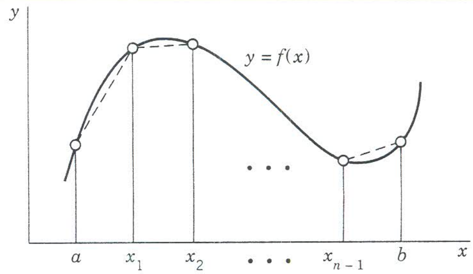


Fig. 439. Trapezoidal rule

Trapezoidal Rule

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

$$\approx \sum_{i=1}^{n} \frac{1}{2}\left[f(x_{i-1}) + f(x_i)\right]\cdot h$$

$$= h\left[\frac{1}{2}f(\underset{x_0}{\underset{\shortparallel}{a}}) + f(x_1) + \cdots + f(x_{n-1}) + \frac{1}{2}f(\underset{x_n}{\underset{\shortparallel}{b}})\right]$$

Error Bounds for the Trapezoidal Rule

$$\varepsilon = -\frac{(b-a)}{12}h^2 f''(\hat{t}) \quad \text{for some } \hat{t} \in (a,b)$$

$$\Rightarrow KM_2 \leq \varepsilon \leq KM_2^*, \quad \text{where } K = -\frac{(b-a)}{12}h^2$$

$$M_2^* \leq f''(\hat{t}) \leq M_2$$

Error Estimation

$$J = J_h + \varepsilon_h = J_{h/2} + \varepsilon_{h/2}, \quad \text{where } \varepsilon_{h/2} \approx \frac{1}{4}\varepsilon_h$$

$$\Rightarrow \varepsilon_{h/2} \approx \frac{1}{3}\left(J_{h/2} - J_h\right)$$

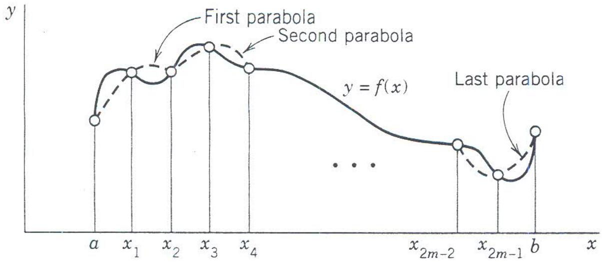


Fig. 440. Simpson's rule

Simpson's Rule

$$J = \int_a^b f(x)\,dx \approx \frac{h}{3}\left[f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots \right.$$
$$\left. \cdots + 2f_{2m-2} + 4f_{2m-1} + f_{2m}\right]$$

Error Bounds

$$\varepsilon = -\frac{(b-a)}{180} h^4 f^{(4)}(\hat{t}) \quad \text{for some } \hat{t} \in (a,b)$$

$\Rightarrow$ Simpson's Rule is exact for polynomials of up to degree 3 !!!
($\because$ $f^{(4)}(\hat{t}) \equiv 0$ for them)

Numerical Stability

$$\frac{h}{3}\left|\varepsilon_0 + 4\varepsilon_1 + 2\varepsilon_2 + \cdots + \varepsilon_{2m}\right| \leq \frac{(b-a)}{3\cdot 2m}\cdot 6mu = (b-a)u$$

round-off unit

Error Estimation

$$\varepsilon_{h/2} \approx \frac{1}{15}\left(J_{h/2} - J_h\right)$$

Adaptive Integration

adaptive step size h to the variability of $f(x)$

Gauss Quadrature Formula

$$\int_{-1}^{1} f(t)\,dt \approx \sum_{j=1}^{n} A_j f_j, \quad \text{where } f_j = f(t_j)$$

Gauss has shown that the above formula is exact for polynomials of degree up to $2n-1$, where t_j is the jth zero of the Legendre polynomial P_n and A_j depends on n but not on $f(t)$!

Table 19.7 Gauss Integration: Nodes t_j and Coefficients A_j

n	Nodes t_j	Coefficients A_j
2	−0.57735 02692	1
	0.57735 02692	1
3	−0.77459 66692	0.55555 55556
	0	0.88888 88889
	0.77459 66692	0.55555 55556
4	−0.86113 63116	0.34785 48451
	−0.33998 10436	0.65214 51549
	0.33998 10436	0.65214 51549
	0.86113 63116	0.34785 48451
5	−0.90617 98459	0.23692 68851
	−0.53846 93101	0.47862 86705
	0	0.56888 88889
	0.53846 93101	0.47862 86705
	0.90617 98459	0.23692 68851

Ex 8

$$\int_0^2 \frac{1}{4}\pi x^4 \cos\frac{1}{4}\pi x\,dx$$

$$= \int_{-1}^{1} \frac{1}{4}\pi (t+1)^4 \cos\frac{1}{4}\pi(t+1)\,dt$$

$$= A_1 f_1 + A_2 f_2 + A_3 f_3 + A_4 f_4$$

$$= A_1(f_1 + f_4) + A_2(f_2 + f_3)$$

$$= 1.25950$$

The error is 0.00003!

The error is impressive compared with the amount of work!

Ex 7

$$\int_0^1 \exp(-x^2)\,dx = \frac{1}{2}\int_{-1}^{1} \exp\left(-\frac{1}{4}(t+1)^2\right)dt$$

$$\approx \frac{1}{2}\left[\frac{5}{9}\exp\left(-\frac{1}{4}\left(1-\sqrt{\frac{3}{5}}\right)^2\right) + \frac{8}{9}\exp\left(-\frac{1}{4}\right) + \frac{5}{9}\exp\left(-\frac{1}{4}\left(1+\sqrt{\frac{3}{5}}\right)^2\right)\right]$$

$$= 0.746815$$