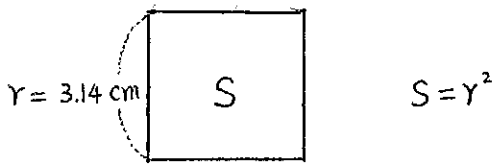
Chap. 4 Source of Errors

(1) Errors in given input data

round-off error in measurement



(2) Round-off errors during the computation

:

S = 3.14 x 3.14 (9.8596)

= 9.86

:

(3) Truncation error

Taylor series

(4) Simplification in the mathematical model

Ex. Method to determine the value of g (gravity)

$\sin\theta \approx \theta$ (θ is very small)

(5) Propagated error

Iteration methods

(6) Computer-caused loss of significance

$a - b$ when $a \approx b$

Question: $(a - b + c) = (\underline{a + c} - b)$? overflow

Absolute error (true error)

$$Et = |\text{true value} - \text{approximation}|$$

Relative error

$$\varepsilon_t = \left| \frac{\text{true value - approximation}}{\text{true value}} \right| \quad \text{when true value} \neq 0$$

Note. Usually, we do not know the true value a priori.

$$\varepsilon_a = \frac{\text{approximate error}}{\text{approximation}}$$

Iteration Case:

$$\varepsilon_a = \frac{\text{present approximation - previous approximation}}{\text{present approximation}}$$

- **Stopping criterion:** $|\varepsilon_a| < \varepsilon_s$

Note. At least n significant digits, if

$$\varepsilon_s = (0.5 \times 10^{2-n})\ \%$$

Ex 4.1 Find $e^{0.5}$ with at least 3 significant digits. ($e^{0.5} = 1.648721\ldots$)

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!} \qquad \textbf{(MacLaurin series)}$$

Stopping criterion: $\varepsilon_s = (0.5 \times 10^{2-3})\% = 0.05\%$

1st: $e^x = 1$ $\quad \varepsilon_t = \left|\frac{1.648721 - 1}{1.648721}\right| \times 100\% = 39.3469\%$

2nd: $e^x = 1 + 0.5 = 1.5$

$$\varepsilon_t = \left|\frac{1.648721 - 1.5}{1.648721}\right| = 9.02\% \qquad \varepsilon_a = \left|\frac{1.5 - 1.0}{1.5}\right| = 33.3\%$$

3rd: $e^x = 1 + 0.5 + (0.5)^2/2 = 1.625$

$$\varepsilon_t = \left|\frac{1.648721 - 1.625}{1.648721}\right| = 1.43875\% \qquad \varepsilon_a = \left|\frac{1.625 - 1.5}{1.625}\right| = 7.6923\%$$

4th: $e^x = 1 + 0.5 + (0.5)^2/2 + (0.5)^3/6 = 1.6458333$

$$\varepsilon_t = \left|\frac{1.648721 - 1.6458333}{1.648721}\right| = 0.17514\% \qquad \varepsilon_a = \left|\frac{1.6458333 - 1.625}{1.6458333}\right| = 1.26582\%$$

5th: $e^x = 1 + 0.5 + (0.5)^2/2 + (0.5)^3/6 + (0.5)^4/24 = 1.6484374$

$$\varepsilon_t = \left|\frac{1.648721 - 1.6484374}{1.648721}\right| = 0.0172\% \qquad \varepsilon_a = \left|\frac{1.6484374 - 1.6458333}{1.6484374}\right| = 0.15799\%$$

6th: $e^x = 1 + 0.5 + (0.5)^2/2 + (0.5)^3/6 + (0.5)^4/24 + (0.5)^5/120 = 1.6486978$

$$\varepsilon_t = \left|\frac{1.648721 - 1.6486978}{1.648721}\right| = 0.0014\% \qquad \varepsilon_a = \left|\frac{1.6486978 - 1.6484374}{1.6486978}\right| = 0.01579\%$$

Terminate because $0.01579\% < 0.05\%$

Rounding and Chopping

Ex. 3 decimal

			error
0.2397	rounding:	0.240	0.0003
	chopping:	0.239	0.0007
0.2375	rounding:	0.238	0.0005
	chopping:	0.237	0.0005
0.2372	rounding:	0.237	0.0002
	chopping;	0.237	0.0002

Computer Number Reprsentation

Base β system: $\underbrace{2,\ 8,\ 16,}_{\text{computer}}\ \underbrace{10}_{\text{human}}\ \ 12,\ 60$

Ex. $(10.1)_8 = 1 \times 8^1 + 0 \times 8^0 + 1 \times 8^{-1}$
$= 8 + 0 + 0.125$
$= 8.125$

Binary: 0, 1
Octal: 0,1,2,3,4,5,6,7
Hexadecimal: 0,1,2,3,4,5,6,7,8,9, A, B, C, D, E, F

Ex. $(0.276)_8 = 2 \times 8^{-1} + 7 \times 8^{-2} + 6 \times 8^{-3}$

$= 2/8 + 7/64 + 6/512$
$= 190/512$
$= 0.371\cdots$

Ex. $(2.125)_{10} = (\quad)_2$

Ex. Binary $\Longleftrightarrow$ Hexadecimal

$(\underbrace{0}\underbrace{1}\underbrace{1}\underbrace{0}\underbrace{0}\underbrace{0}\underbrace{1}\underbrace{1})_2 \Rightarrow (63)_{16}$

Integer Representation

(1) Sign-Magnitude Method

+18 = 0 0 0 1 0 0 1 0
-18 = 1 0 0 1 0 0 1 0

Drawbacks –

(2) Two's Complement Method

Ex. (-3) in 8-bit number system

+3	0 0 0 0 0 0 1 1
complement	1 1 1 1 1 1 0 0
add 1	1 1 1 1 1 1 0 1 (= -3)

Ex. 5 – 3 = 5 + (-3)

5	0 0 0 0 0 1 0 1
-3	1 1 1 1 1 1 0 1 +

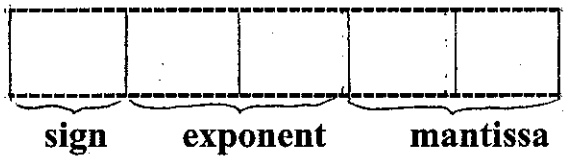
1	0 0 0 0 0 0 1 0 (= 2)
ignore	

Note. No subtractor is necessary.

Note: If two numbers are added and they are both positive or both negative, then overflow occurs iff the result has the opposite sign.

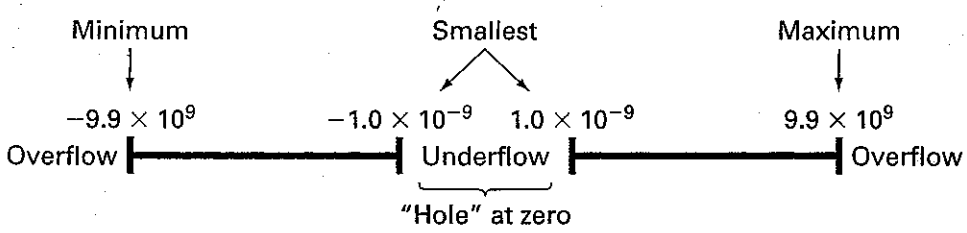
Ex. 4.2

Hypothetical computer – base-10 with 5-digit word size



Normalized decimal number: $S_1 d_1 . d_2 \times 10^{S_0 d_0}$

largest positive value: $+9.9 \times 10^{+9}$ **(10 billion)**
smallest positive value: $+1.0 \times 10^{-9}$



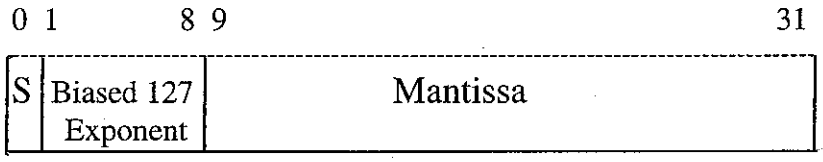
Note. Exponent mantissa are finite.

increase exponent → increase range
increase mantissa → increase precision

tradeoff

Floating-Point Numbers

$$\pm S \times B^{\pm E}$$



-127 ~ +128

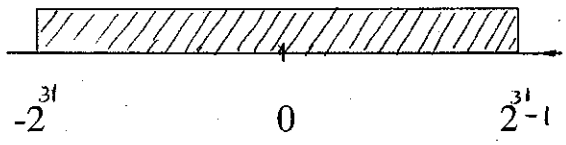
Ex. $1.1010001 \times 2^{10100} = 0\ 10010011\ 10100010\ldots$

$1.1010001 \times 2^{-10100} = 1\ 01101011\ 10100010\ldots$

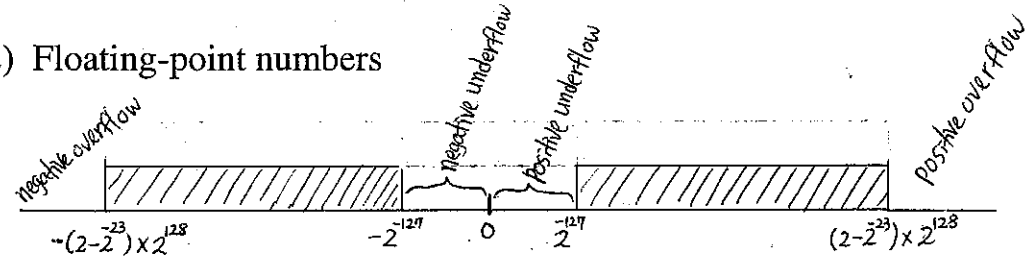
Note. Normalized: $+ 1.bbb..b * 2^{\pm E}$
Base: Either 2 or 16 (IBM S/370)

Expressible numbers in 32-bit word

(1) 2's complement integers

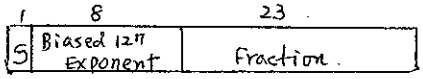


(2) Floating-point numbers



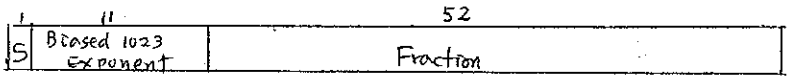
IEEE Standard

Single form



$10^{-38} - 10^{38}$

Double form



$10^{-308} - 10^{308}$

Rounding Policy

- Round to nearest number (default)

Example.

(1)	....xxx $\overset{23}{\mid}$ 1 0 0 0 1 0 ...	roundup
(2)	....xxx $\mid$ 0 1 1 0 1 0 ...	truncate

Case of tie: .. xxx | 1 0 0 0 0 ...

Solution.

..xx1 | 1 0 0 0 0 ...	roundup
..xx0 | 1 0 0 0 0 ...	truncate

Loss of Significance

Reason. $x \ominus y \rightarrow fl(x \ominus y)$ but $x \ominus y \rightarrow fl[fl(x) \ominus fl(y)]$

Ex.

```
sum = 0
for i = 1 : 10000
   sum = sum + 0.0001
end
print, sum      // 0.99999999999991 //
```

Ex. 4-digit mantissa and 1-digit exponent

$4000 + 0.001$

$$\begin{array}{rl} 0.4000 & \times 10^4 \\ 0.0000001 & \times 10^4 \\ \hline 0.4000001 & \times 10^4 \end{array} \rightarrow 0.4000 \times 10^4$$

Suggestions

Ex. Infinite series

$S = a_1 + a_2 + a_3 + \ldots + a_n \cdots$ where $|a_i| > |a_{i+1}|$

Sum the series in reverse order.

Ex. $f(x) = x - \sin x$, where $x = 1/15$ on 10-digit machine

$$\begin{array}{rl} x = & 0.6666666667 \times 10^{-1} \\ \sin x = & 0.6661729492 \times 10^{-1} \\ \hline x - \sin x = & 0.4937175000 \times 10^{-4} \end{array}$$

Use Taylor series

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

$$x - \sin x = \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

$$= 0.4937174327 \times 10^{-4}$$

Taylor Series

(1) $$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \qquad -\infty < x < \infty$$

(2) $$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \qquad -\infty < x < \infty$$

(3) $$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \qquad -\infty < x < \infty$$

(4) $$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots$$ rhs. ratio is x. To converge, $|x|<1$ $\therefore$ $-1 < x < 1$

(5) $$\ln x = (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \frac{(x-1)^4}{4} + \cdots \qquad 0 < x \leqq 2$$

Ex. $$\ln(1.1) \approx 0.1 - \frac{0.01}{2} + \frac{0.001}{3} - \frac{0.0001}{4} + \frac{0.00001}{5}$$
$$= 0.09531\,03333\ldots$$

FORMAL TAYLOR SERIES

$$f(x) \sim f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \frac{f'''(x_0)}{3!}(x - x_0)^3 + \cdots$$

or $$f(x) \sim \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!}(x - x_0)^k$$

When $x_0 = 0$,

$$f(x) \sim f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!}x^k$$

Maclaurin series

Applications.

(1) $f(x) = e^x$

Use Maclaurin Series.

$$f(x) = f(0) + f'(0)\cdot x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \quad (-\infty < x < \infty)$$

(2) $f(x) = \sin x, \quad f(0) = 0$

$f'(x) = \cos x \quad f'(0) = 1$

$f''(x) = -\sin x \quad f''(0) = 0$

$f'''(x) = -\cos x \quad f'''(0) = -1$

$f^{iv}(x) = \sin x \quad f^{iv}(0) = 0$

$f^{v}(x) = \cos x \quad f^{v}(0) = 1$

$\vdots$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} \cdots \quad (-\infty < x < \infty)$$

In practice, we have to truncate the series. For instance,

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!}$$

TAYLOR'S THEOREM

If the function f possesses continuous derivatives of orders $1, 2, \ldots, n+1$ in a closed interval $I = [a, b]$, then for any x_0 in I,

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + \underbrace{\frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}}_{\text{Error term}}$$

where x is any value in I with $\xi = \xi(x)$ being a point lying between x_0 and x.

TAYLOR'S THEOREM

If the function f possesses continuous derivatives of orders $1, 2, \ldots, n+1$ in a closed interval $I = [a, b]$, then for any x in I,

$$f(x + h) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}h^k + E_{n+1}$$

where h is any value such that $x + h$ is in I and where

$$E_{n+1} = \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1} \qquad \Rightarrow \quad E_{n+1} = O(h^{n+1})$$

with ξ between x and $x + h$.

$$\begin{aligned} f(x+h) &= f(x) + f'(\xi_1)h \\ &= f(x) + \mathcal{O}(h) \qquad \tfrac{1}{2}h \longrightarrow \tfrac{1}{2} \text{ error} \\ f(x+h) &= f(x) + f'(x)h + \tfrac{1}{2}f''(\xi_2)h^2 \\ &= f(x) + f'(x)h + \mathcal{O}(h^2) \qquad \tfrac{1}{2}h \longrightarrow \tfrac{1}{4} \text{ error} \\ f(x+h) &= f(x) + f'(x)h + \tfrac{1}{2}f''(x)h^2 + \tfrac{1}{6}f'''(\xi_3)h^3 \\ &= f(x) + f'(x)h + \tfrac{1}{2}f''(x)h^2 + \mathcal{O}(h^3) \\ &\vdots \end{aligned}$$

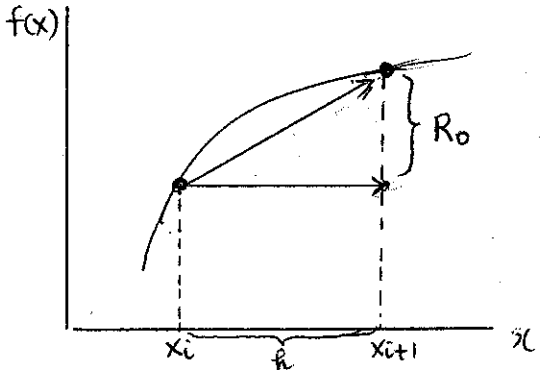
Truncation error on Taylor series

Taylor series: $f(x_{i+1}) = f(x_i) + f'(x_i)\,h + \frac{f''(x_i)}{2!}h^2 + \cdots$

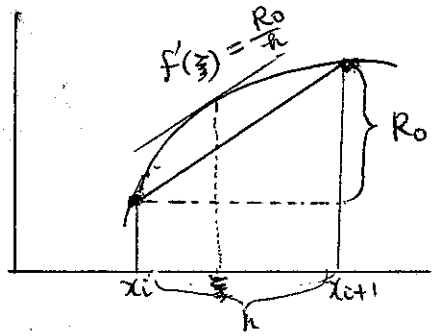
- Zero-order Taylor series

$$f(x_{i+1}) \approx f(x_i)$$

Then $R_0 = f'(x_i)\,h + \frac{f''(x_i)}{2!}h^2 + \frac{f'''(x_i)}{3!}h^3 + \ldots$



- Mean-value theorem



$$f'(\xi) = \frac{R_0}{h}$$

$$R_0 = f'(\xi)\,h \qquad : \quad O(h)$$

- First-order

$$R_1 = \frac{f''(\xi)}{2!}h^2 \qquad : \quad O(h^2)$$

⋮ ⋮

Numerical Differentiation

(1) forward difference

$$f(x_{i+1}) = f(x_i) + f'(x_i)\,h + \frac{f''(x_i)}{2!}h^2 + \dots \quad \cdots \text{(A)}$$

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{h} + O(h)$$

(2) backward difference

$$f(x_{i-1}) = f(x_i) - f'(x_i)\,h + \frac{f''(x_i)}{2!}h^2 - \dots \quad \cdots \text{(B)}$$

$$f'(x_i) = \frac{f(x_i) - f(x_{i-1})}{h} + O(h)$$

(3) centered difference

(A) – (B): $f(x_{i+1}) - f(x_{i-1}) = 2 f'(x_i)\,h + \frac{f^{(3)}(x_i)}{3!}h^3 + \dots$

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2h} - \frac{f^{(3)}(x_i)}{3!}h^2$$

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2h} - O(h^2)$$

Note. Halving the step size → quartering the error

Example 4.4

$$f(x) = -0.1x^4 - 0.15x^3 - 0.5x^2 - 0.25x + 1.2$$

What is $f'(x)$ at $x = 0.5$?

Analytical:

$$f'(x) = -0.4x^3 - 0.45x^2 - 1.0x - 0.25$$
$$f'(0.5) = -0.9125$$

Numerical:

	h = 0.5		h = 0.25	
	f'(0.5)	$\lvert\varepsilon_t\rvert$	f'(0.5)	$\lvert\varepsilon_t\rvert$
Forward O(h)	-1.45	58.9%	-1.155	26.5%
Backward O(h)	-0.55	39.7%	-0.714	21.7%
Centered $O(h^2)$	-1.0	9.6%	-0.934	2.4%

Observations

1. Centered difference method is more accurate than forward or backward method.
2. Halving the step size halves the error of forward/backward method, and quarters the error of the centered method.

fl notation and backward (inverse) error analysis

Hypothetical computer: 5-place decimal machine, 10-place accumulator

(1) $\text{fl}(0.3721871422 \times 10^4) = 0.37219 \times 10^4$

(2) x: 0.37218×10^4
y: 0.71422×10^{-1}

x + y: 0.3721800000×10^4
0.0000071422×10^4

$0.3721871422 \times 10^4 \rightarrow 0.37219 \times 10^4 : z$

$$\text{relative error} : \frac{|x+y-z|}{|x+y|} = \frac{0.0000028578 \times 10^4}{0.3721871422 \times 10^4} \doteq 0.77 \times 10^{-5} \leq 10^{-5}$$

(3) IEEE single precision

$$\frac{|x - \text{fl}(x)|}{x} \leq 2^{-24} \quad \text{or} \quad \text{fl}(x) = x(1+\delta) \quad |\delta| \leq 2^{-24}$$

Def $\text{fl}(x \odot y) = (x \odot y)(1+\delta) \quad |\delta| \leq \underbrace{2^{-24}}_{u}$

$$\text{fl}(x+y) = (x+y)(1+\delta)$$
$$= \underbrace{x(1+\delta)}_{\text{Perturbed } x} + \underbrace{y(1+\delta)}_{\text{Perturbed } y} \rightarrow \textit{backward error analysis}$$

Backward error analysis:

We view fl(x+y) is the exact result of otherwise perturbed $(x+y)(1+\delta)$ for some δ, $|\delta| \leq 2^{-24}$. Output data which the algorithm produces (under the influence of round-off error) is the exact output data of a problem of the same type in which the input data has been changed by a few u. That change, measured with u as a unit, is called the condition number of the algorithm.

Note. Using backward error analysis, one transfer the problem of estimating the effect of round-off errors during the computation back to the problem of estimating the effect of disturbances in input data.

Alternation Series

$1 - 1/2 + 1/4 - 1/8 + 1/16 - 1/32 \ldots \qquad [= 2/3]$

$= (1 + \frac{1}{4} + \frac{1}{16} + \cdots) - \frac{1}{2}(1 + \frac{1}{4} + \frac{1}{16} + \cdots)$

$= \frac{1}{2}(1 + \frac{1}{4} + \frac{1}{16} + \cdots)$

$= \frac{2}{3}$

ALTERNATING SERIES THEOREM

If $a_1 \geq a_2 \geq \ldots \geq a_n \geq \ldots \geq 0$ for all n and $\lim_{n\to\infty} a_n = 0$, then the alternating series

$$a_1 - a_2 + a_3 - a_4 + \ldots$$

converges; that is,

$$\sum_{k=1}^{\infty} (-1)^{k-1} a_k = \lim_{n\to\infty} \sum_{k=1}^{n} (-1)^{k-1} a_k = \lim_{n\to\infty} S_n = S$$

where S is its sum and S_n is the nth partial sum. Moreover, for all n,

$$|S - S_n| \leq a_{n+1}$$

Ex. How many terms are needed in computing sin 1, with error $\leq 1/2 \cdot 10^{-14}$?

$$\sin 1 = 1 - 1/3! + 1/5! - 1/7! + \ldots$$

Solution.

If we stop at 1/(2n-1)!, then the error does not exceed the first omitted term, 1/(2n+1)!

$$\frac{1}{(2n+1)!} < \frac{1}{2} \times 10^{-14} \quad \rightarrow \quad n > 8$$

EXAMPLE It is known that

$$\frac{\pi^4}{90} = 1^{-4} + 2^{-4} + 3^{-4} + \cdots$$

How many terms should we take in order to compute $\pi^4/90$ with an error of at most $\frac{1}{2} \times 10^{-8}$?

Solution A naive approach is to take

$$1^{-4} + 2^{-4} + \cdots + n^{-4}$$

where n is chosen so that the next term, $(n+1)^{-4}$, is less than $\frac{1}{2} \times 10^{-8}$. This value of n is 118, but this is an erroneous answer, for the partial sum

$$S_{118} = \sum_{k=1}^{118} k^{-4}$$

differs from $\pi^4/90$ by 2×10^{-7}. What we should do, of course, is to select n so that *all* the omitted terms add up to less than $\frac{1}{2} \times 10^{-8}$:

$$\sum_{k=n+1}^{\infty} k^{-4} < \frac{1}{2} \times 10^{-8}$$

By a technique familiar from calculus (see Figure 1.1), we have

$$\sum_{k=n+1}^{\infty} k^{-4} < \int_n^{\infty} x^{-4}\,dx = \frac{x^{-3}}{-3}\bigg|_n^{\infty} = \frac{1}{3n^3}$$

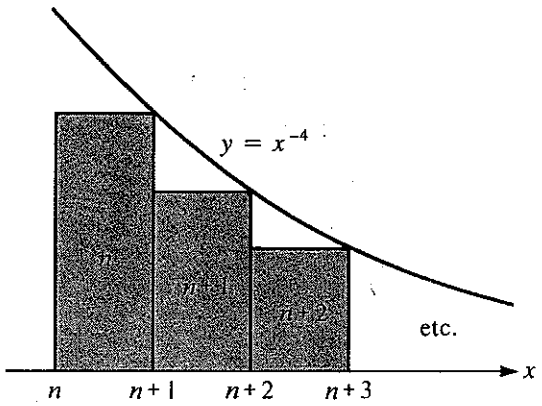


FIGURE 1.1

Thus it suffices to select n so that $1/(3n^3) < \frac{1}{2} \times 10^{-8}$, or $n \geqq 406$.