§ Machine numbers

~ represented by a finite number of binary digits (bits)

e.g. a single-precision real number is usually represented by

a word = 4 bytes = 32 bits

e.g. 16-base system: $\pm a \times 16^b$

X	XXXXXXX	XXX⋯XXX
1 bit	7 bits	24 bits
sign	exponential part	fractional part

• exponential part (7 bits): $\pm a \times 16^b$

\# of numbers that can be composed = $2^7 = 128$

$$\begin{cases} 64 \text{ for zero and positive exponents: } 0,1,2,\cdots 63 \\ 64 \text{ for negative exponents: } -1,-2,\cdots,-64 \end{cases}$$

0000000 ⇨ 0 ⇨ −64

0000001 ⇨ 1 ⇨ −63

...

1000000 ⇨ 64 ⇨ 0

…

1111111 ⇨ 127 ⇨ 63

- fractional part: (24 bits) $\pm a \times 16^b$

$$X_1X_2X_3\cdots X_{22}X_{23}X_{24} \equiv X_1\cdot 2^{-1} + X_2\cdot 2^{-2} + \cdots + X_{24}\cdot 2^{-24}$$

$$\text{maximum}\quad 111\cdots 111 \equiv 2^{-1}+2^{-2}+2^{-3}+\cdots+2^{-24} \approx \frac{2^{-1}}{1-2^{-1}} = 1$$

e.g. $\underline{0}\ \underline{1000010}\ \underline{101100\ldots000}$

$$1000010 = 2^1 + 2^6 = 66 \quad \Rightarrow b = 66 - 64 = 2$$

$$101100\ldots000 = 2^{-1} + 2^{-3} + 2^{-4} = 0.6875 = a$$

$$\underline{0}\ \underline{1000010}\ \underline{101100\ldots000} \equiv +0.6875\times 16^2 = 176 \quad (10\,\text{base})$$

§ Machine numbers --- 32 bits

~ # of machine numbers = $2^{32} = 1024^{3.2} = 4\times 1024^3 = 4G$

the maximum one = 0(1111111)(11……1) $\approx 16^{63} \approx 10^{76}$

the minimum nonzero one = 0(0000000)(00……01)

$$= 2^{-24}\times 16^{-64} \approx 10^{-84}$$

WARMING MESSAGE:

OVERFLOW ~ appears a number which absolute value > 10^{76}

UNDERFLOW ~ appears a nonzero number which absolute value < 10^{-84}

~ a finite set of real numbers

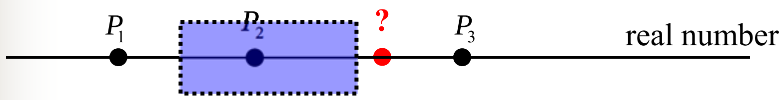
§ Machine numbers --- discrete number system

$$\underline{0}\ \underline{10000000}\ \underline{101100\ldots000} = 0.6875 \quad (P_2)$$

The two nearby machine numbers are:

$$\underline{0}\ \underline{10000000}\ \underline{101100\ldots001} = 0.6875 + 2^{-24} \ (P_3)$$

$$\underline{0}\ \underline{10000000}\ \underline{101011\ldots11} = 0.6875 - 2^{-24} \ (P_1)$$



~ all represented by P_2

rounding error $\equiv |P - P_2|$

§ Rounding Errors

Suppose a machine can represent a number up to k digits in the following form: $\pm 0.d_1 d_2 \cdots d_k \times 10^n$, $1 \le d_1 \le 9$ and $0 \le d_i \le 9, i = 2,3,\cdots,k$

How to present π=3.141592653589793…..? Machine-dependent!

e.g. k = 7

chopping method : $fl(\pi) = 0.3141592 \times 10^1$

rounding method : $fl(\pi) = 0.3141593 \times 10^1$

$$error = |\pi - fl(\pi)|$$

§ Rounding Errors

Round-off errors are unavoidable

and accumulate as computations go on.

$E_n \equiv$ magnitude of rounding error after n subsequent operations

* linear growth : $E_n \approx CnE_0$ for some constant C

Usually unavoidable but acceptable as long as C and E_0 are sufficiently small.

* exponential growth: $E_n \approx C^n E_0$ for some constant $C > 1$

Overflow!

example: compute the series $P_n = \dfrac{1}{3^n}$ with single-precision real numbers

Method 1

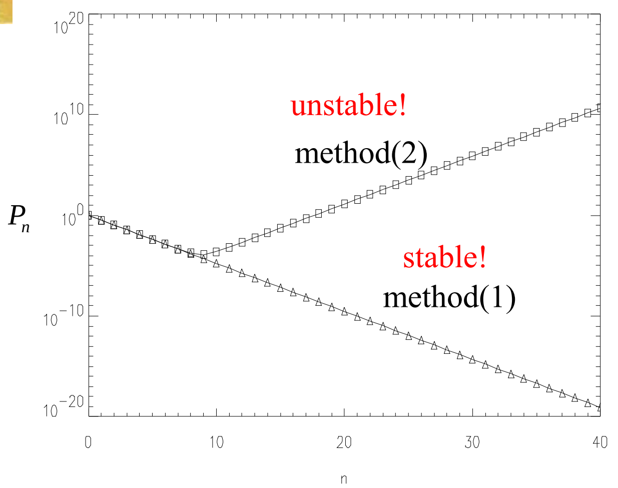
```
P(0)=1
DO n=1,100
    P(n)=1./3.*P(n-1)
END DO
```

Method 2

```
P(0)=1
P(1)=1./3.
DO n=2,100
    P(n)=10./3.*P(n-1)–P(n-2)
END DO
```

Method 1: $fl(P_n) = fl\left(fl\left(\frac{1}{3}\right) * fl(P_{n-1}) \right)$

Method 2: $fl(P_n) = fl\left(fl\left(\frac{10}{3}\right) * fl(P_{n-1}) - fl(P_{n-2}) \right)$



Method 1: $fl(P_n) = fl\left(fl\left(\frac{1}{3}\right) * fl(P_{n-1}) \right)$

Method 2: $fl(P_n) = fl\left(fl\left(\frac{10}{3}\right) * fl(P_{n-1}) - fl(P_{n-2}) \right)$

Method 2

```
P(0)=1
P(1)=1./3.
DO n=2,100
    P(n)=10./3.*P(n-1)–P(n-2)
END DO
```

$$P_n = \frac{1}{3^n}$$

formula correct!

$$P_n = \rho^n$$

$$\rho^n = \frac{10}{3}\rho^{n-1} - \rho^{n-2}$$

$$\rho^2 - \frac{10}{3}\rho + 1 = 0$$

$$(\rho - 3)\left(\rho - \frac{1}{3}\right) = 0$$

$$P_n = A \cdot 3^n + B \cdot \frac{1}{3^n}$$

$$\left.\begin{aligned} P_0 &= 1 = A + B \\ P_1 &= \frac{1}{3} = 3A + \frac{B}{3} \end{aligned}\right\} \Rightarrow \begin{cases} A = 0 \\ B = 1 \end{cases}$$

Ways of Avoiding Rounding Errors:

① Reduce # of computations as many as possible.

$$\pi + e = 3.141592653... + 2.71828182... = 5.85987448...$$

$$\pi * e = 3.141592653... * 2.71828182... = 8.53973422...$$

7 digits + rounding method:

$$fl(fl(\pi) + fl(e)) = fl(3.141593 + 2.718282) = 5.859875$$

$$fl(fl(\pi) * fl(e)) = fl(3.141593 * 2.718282)$$
$$= fl(8.539735703...) = 8.539736$$

② Avoid substraction of two nearly equal numbers.

$$fl(x) - fl(y) = 0.3141593 \times 10^1 - 0.3141291 \times 10^1 = 0.3020000 \times 10^{-3}$$

~ lose 4 digits of significance

(Any further calculations can have only 3, instead of 7, digits of significance.)

③ Avoid dividing by a small number.

original rounding error $= \delta$ exact number $= z = fl(z) + \delta$

divided by a small number $\varepsilon = 10^{-6}$

$$\text{rounding error} = \left|\frac{z}{\varepsilon} - \frac{fl(z)}{\varepsilon}\right| = \left|\frac{\delta}{\varepsilon}\right| = 10^6 |\delta|$$