- Floating point representation of numbers

Representing numbers.

The real number system. $x = \pm (d_0.d_1 d_2 d_3 \cdots)_\beta \times \beta^E$ ← base β rep.

Ex: $x = 3.14159265358\ldots$

$= + \left(\frac{3}{10^0} + \frac{1}{10^1} + \frac{4}{10^2} + \cdots\right) \times 10^{+0}$

$= + (3.14159265\ldots)_{10} \times 10^{+0}$

$= + S \times 10^{+0}$, with $1 \le S < 10$.

The floating-point number system.

A floating-point number x has the form

$$x = \pm S_\beta \times \beta^e = \pm \left(\frac{d_0}{\beta^0} + \frac{d_1}{\beta^1} + \cdots + \frac{d_{p-1}}{\beta^{p-1}}\right) \times \beta^e, \text{ with } e_{min} \le e \le e_{max}.$$

(p digits)

The set F of floating-point numbers is defined by

base β

floating-point precision p

exponent range $e_{min} \le e \le e_{max}$

Binary system: $\beta = 2$

$$x = \pm (1.b_1 b_2 \cdots b_{p-1})_2 \times 2^e, \text{ where digits } \{b_i\} \text{ satisfy } b_i = \{0, 1\}.$$

"normalized rep" — for uniqueness of representation

$m = 1 + \frac{b_1}{2} + \frac{b_2}{2^2} + \cdots + \frac{b_{p-1}}{2^{p-1}}$ is called the mantissa.

Spacing of numbers in $\mathbb{F}$:

Consider the numbers on either side of 1.

$1 = (1.00\ldots0)\times 2^0$

the next number in $F(p,M)$ is

$(1.0\ldots01)_2 \times 2^0 = 1 + \frac{1}{2^{p-1}}$. let $\varepsilon = \frac{1}{2^{p-1}}$

$\uparrow$ $1/2^{p-1}$

the next number is

$(1.00\ldots010)_2 \times 2^0 = 1 + \frac{1}{2^{p-1}} + \frac{1}{2^{p-1}} = 1 + 2\varepsilon$

$\uparrow$ $1/2^{p-2}$

so the numbers following 1 are

$1,\ 1+\varepsilon,\ 1+2\varepsilon,\ \ldots,\ 1+k\varepsilon.$

Notice: $1+k\varepsilon = (1.1\ldots1)_2\times 2^0 = 2 - \frac{1}{2^{p-1}}$

$k\varepsilon = 1 - \frac{1}{2^{p-1}} = \underbrace{\frac{1}{2^{p-1}}}_{\varepsilon}\underbrace{(2^{p-1}-1)}_{k} \Rightarrow k = 2^{p-1}-1$

The numbers preceding 1 are:

$(1,\ 1+\varepsilon,\ 1+2\varepsilon,\ \ldots,\ 1+k\varepsilon)\times 2^{-1}$

In particular, $1-(1+k\varepsilon)\times\frac{1}{2} = 1-\left(\frac{1}{2}+\left(\frac{1}{2}-\frac{1}{2^p}\right)\right) = \frac{1}{2^p}$

The spacing between numbers $(1,\ 1+\varepsilon,\ \ldots,\ 1+k\varepsilon)\times 2^e$ is $2^e\times\varepsilon = \frac{2^e}{2^{p-1}}$

So, some problems arise when small numbers are added to big ones.

double precision: $\varepsilon_p = \frac{1}{2^{52}}$, called "machine ε"

⇒ matlab command: eps

$e_{max} = 1023$, so the spacing of numbers from

$1 \times 2^{53}, \ldots, (2 - \frac{1}{2^{52}}) \times 2^{1023}$ is greater than one

⇒ If you need to represent integers (and only integers) exactly, you should use the integer class.

⇒ There are also extended arithmetic packages available.

⇒ common uses: cryptography algorithms, very large primes + factorizations.

Rounding

The absolute error between x and $x_f = fl(x)$, its floating-pt representation is $|x - x_f|$

The relative error is $\frac{|x - x_f|}{|x|}$.

Using the "round to nearest" rule: $\frac{|x - x_f|}{|x|} \leq \frac{\varepsilon_p}{2}$.

Other "rounding standards" include

"round up"

"round down"

Notice: zero is not represented in $F(p, M)$:

it is added to the set of floating-pt numbers.

Other additional numbers:

Inf: overflow: $x > z_{max}$

NaN: not a number: If $|x| < z_{min}$, $fl(x) = 0$

and any ratio of the form 0/0 is a NaN.

Ex:

$$fl\left[\frac{(2^{54}-1)-2^{54}}{(2^{54}-0.5)-2^{54}}\right] = \text{NaN} \qquad \text{exact: } 2$$

What if you replace 2^{54} with 2^{53} ?

Ex:

$$fl\left[\frac{(2^{54}-2)-2^{54}}{(2^{54}-1)-2^{54}}\right] = -\text{Inf} \qquad \text{exact: } 2$$

Ex:

$$fl\left[\frac{(2^{54}-1)-2^{54}}{(2^{54}-2)-2^{54}}\right] = 0 \qquad \text{exact: } \frac{1}{2}$$

If you are not aware of these issues, they can cause problems...

Ex: $\sin(k\pi) = 0$ for all integers k.

BUT: try $\sin(2^{53})$ in Matlab $(-0.8489\ldots)$

Try and predict what Matlab gives for:

$(2^{54}-2)-2^{54}$

$(2^{54}+2)-2^{54}$

Why??