Pert-2: Representasi Bilangan Real Pada Komputer

Sebelum masuk materi, tugas pekanan ke-1 dikumpul dan boleh dibahas sekilas. Kalau waktu masih sisa setelah semua materi pertemuan ini dibahas, silakan kembali membahas tugas-1.

- Why does the computer or calculator represent $\sqrt{3}$ as 1.73205080756888?
- Is this value exact for $\sqrt{3}$.
- What is the result of $\sqrt{3} \times \sqrt{3}$, really?
- Try to implement it on calculator.

Computer or calculator might produce an error when doing calculation on real number even very tiny. Such an error is called **the rounding error**. This error can not be avoided because of the inherent system of computer in representing real numbers.

Floating-Number System

In the floating-number system, a number is characterized by 4 integers:

β	base or radic
t	precision
$[L, U]$	exponent range

A real number x in the floating-number system is represented as

$$x = \pm \left(d_0 + \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_{t-1}}{\beta^{t-1}} \right) \beta^e \tag{1}$$

where $0 \leq d_i \leq \beta - 1$, $i = 0, 1, \cdots, t-1$ and $L \leq e \leq U$. The part $d_0 d_1 d_2 \cdots d_{t-1}$ is called **the mantisa** or significand, the porsion of mantisa $d_1 d_2 \cdots d_{t-1}$ called **the fraction**.

Example

Suppose β is the common base, i.e. $\beta = 10$. The number $x = 1.56789$ is represented as

$$1.56789 = 1 + \frac{5}{10} + \frac{6}{10^2} + \frac{7}{10^3} + \frac{8}{10^4} + \frac{9}{10^5}.$$

We have $t = 6$, and this representation is called has six digit precision.

System of representaion IEEE

Most computer today use base $\beta = 2$ (binary) and adopt the system IEEE DP in representing the real numbers. This system with $\beta = 2$ such that $d_i = 0$ atau 1, $t = 53$, $L = -1022$ dan $U = 1023$. For a 64-bit computer, the number in the format IEEE DP[1] forms an *array*:

$$\underbrace{\times \times \times \times \cdots \times}_{64\,\text{bit}}$$

consisting of 0 and 1. The meaning of the number with this representation as follows

$$(-1)^s \times 2^{c-1023} \times (1+f). \tag{2}$$

where parameters , **c** and **f** is understood as follows

- s have value 0 or 1 indicating the sign, positive or negative,
- c is obtained from next 11 bit berikutnya indicating exponent.
- f is 52 last bit indicating mantisa and fraction.

By this format, c runs within interval $0 < \mathbf{c} < 2047$ and exponent range in $(-1023, 1024)$.

[1] *Institute for Electrical and Electronic Engineers Double Precision*

Example: Convert decimal to 64-bit system

Let a number is represented in 64-bit as

$$\underbrace{0}_{\textit{sign}}\ \underbrace{1000000011}_{\textit{exponent}}\ \underbrace{1010100100010000000000000000000000000000000000000000}_{\textit{mantisa}}$$

What is the real number relating to this representation

- Most left bit is $s = 0$. So, this number is positive.
- Next 11 bit 10000000011 give the exponent which is equivalent to

$$c = 1 \times 2^{10} + 0 \times 2^{9} + \cdots + 0 \times 2^{2} + 1 \times 2^{1} + 1 \times 2^{0} = 1027$$

- Last 52 bit as the mantisa meaning

$$f = 1 \times \left(\frac{1}{2}\right)^{1} + 1 \times \left(\frac{1}{2}\right)^{3} + 1 \times \left(\frac{1}{2}\right)^{5} + 1 \times \left(\frac{1}{2}\right)^{8} + 1 \times \left(\frac{1}{2}\right)^{12}$$

- Hence, this notation is to represent the number

$$\begin{aligned}(-1)^{s} \times 2^{c-1023} \times (1+f) &= (-1)^{0} \times 2^{1027-1023}(1+f) \\ &= 27.56640625\end{aligned}$$

Example: Convert to the 64-bit number system of decimal number 32.75.

First, represent 32.75 using base 2:

$$\begin{aligned} 32.75 &= (-1)^0 \times 2^5 \times \left(1 + \frac{0.75}{2^5}\right) \\ &= (-1)^0 \times 2^5 \times \left(1 + \left(\frac{(1/2) + (1/4)}{2^5}\right)\right) \\ &= (-1)^0 \times 2^5 \times \left(1 + \left(\frac{1}{2}\right)^6 + \left(\frac{1}{2}\right)^7\right) \end{aligned}$$

We have $s = 0$, $c - 1023 = 5 \rightarrow c = 1028$ and

$$\begin{aligned} f &= \left(\frac{1}{2}\right)^6 + \left(\frac{1}{2}\right)^7 \\ &= 0 \times \left(\frac{1}{2}\right)^1 + \cdots + 1 \times \left(\frac{1}{2}\right)^6 + 1 \times \left(\frac{1}{2}\right)^7 + 0 \times \left(\frac{1}{2}\right)^8 \cdots + 0 \times \left(\frac{1}{2}\right)^{52}. \end{aligned}$$

Bit sequence for c is obtained as

$$c = 1028 = 1024 + 4 = 2^{10} + 2^2$$

so that in base 2 we get next 11-bit 10000000100. Mantisa f is obtained as

$$00000110000000000000000000000000000000000000000000.$$

Consider the only 6th and 7th bit are 1, else are 0. Hence, the 64-bit

Metode Pembulatan dan Dampaknya

Aproksimasi titik mengambang suatu bilangan real x ditulis $\text{fl}(x)$, yaitu pembulatan (*floating*) dari x. Dua metode pembulatan:

- Pemotongan (*chooping*): ekspansi x dalam basis β dibuang setelah digit ke $t-1$. Misalkan $x = d_1 d_2 d_3 d_4 d_5 \cdots$ akan dibulatkan menjadi 4 digit signifikan (di sini $t = 4$) maka digit setelah $t - 1 = 3$ dibuang sehingga diperoleh $\text{fl}(x) = d_0, d_1 d_2 d_3$. Aturan ini disebut juga **pembulatan menuju nol** sebab $\text{fl}(x)$ adalah bilangan titik mengambang berikutnya yang menuju nol dari x.
- Pembulatan terdekat (*rounding*): $\text{fl}(x)$ adalah bilangan titik mengambang terdekat dengan x. Dalam kasus seimbang (*tie*), $\text{fl}(x)$ diambil bilangan titik mengambang yang digit terakhirnya genap. Oleh karena itu aturan ini juga disebut **pembulatan ke genap**.

Contoh Pembulatan

Bilangan	truncating	rounding	Bilangan	truncating	rounding
1.649	1.6	1.6	1.749	1.7	1.7
1.650	1.6	1.6	1.750	1.7	1.8
1.651	1.6	1.7	1.751	1.7	1.8
1.699	1.6	1.7	1.799	1.7	1.8

Amati!
1.650 dibulatkan ke 1.6, sedangkan 1.750 dibulatkan ke 1.8.

- Selisih sebelum pembulatan $|1.750 - 1.650| = 0.1$
- Seleish setelah pembulatan $|1.8 - 1.6| = 0.2$ naik 2 kali lipat dari sebelumnya.

Order of Convergence

Definition

For a sequence $(\alpha_n)_{n=1}^{\infty}$ which converges to a number α for n get large, and for two positive p and K such that

$$|\alpha - \alpha_n| \leq \frac{K}{n^p} \text{ for all } n \text{ big enough,} \tag{3}$$

the we call $(\alpha_n)_{n=1}^{\infty}$ is convergent to α with order $\mathcal{O}\left(\frac{1}{n^p}\right)$ [read: big-oh dari $\frac{1}{n^p}$].

Taking $h := \frac{1}{n}$, the expression (3) is equivalent to

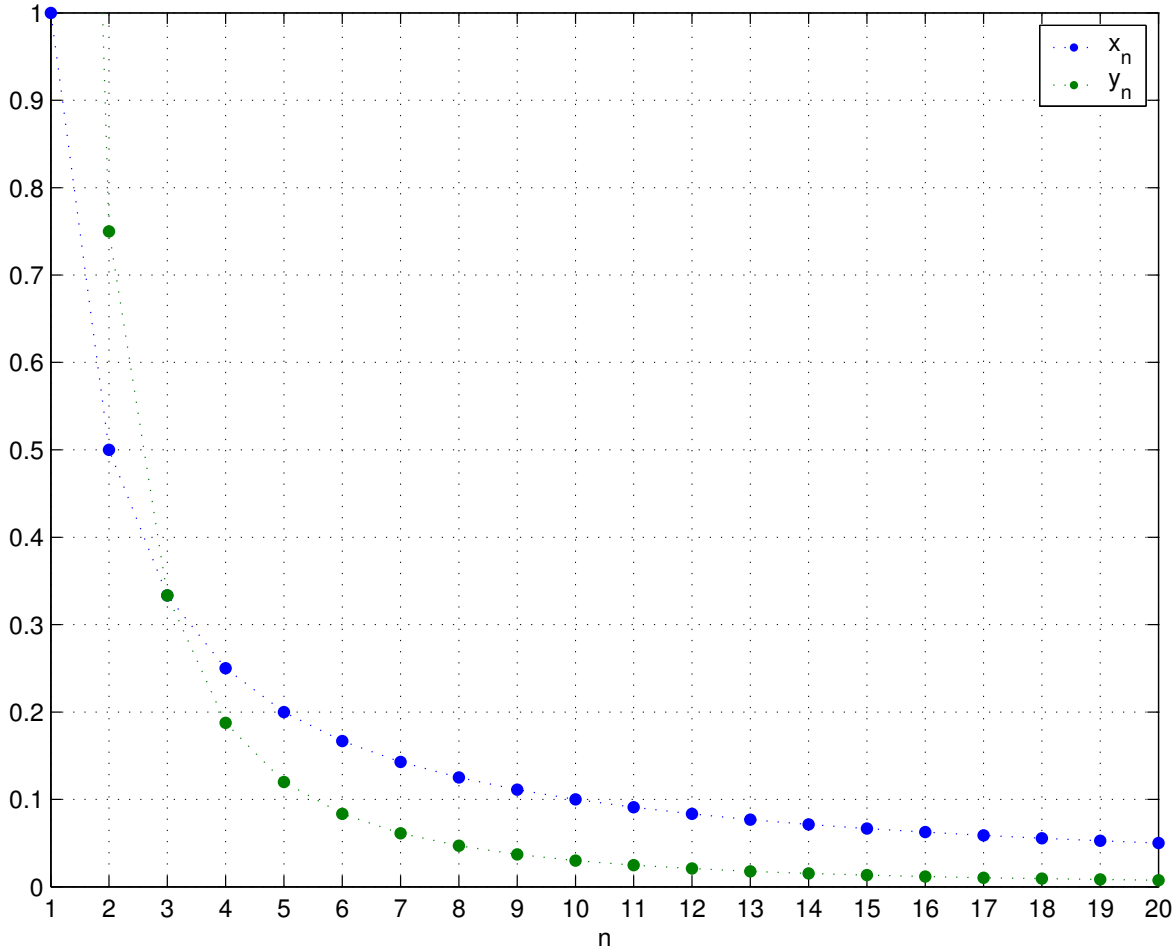
$$|x_h - x| \leq Kh^p \text{ for } h \to 0, \tag{4}$$

written as $\mathcal{O}(h^p)$. For $p = 1$ (first order), $p = 2$ (second order), etc. The larger p the faster convergence.

Example of Convergence Order

1. Both the following sequences (x_n) dan (y_n) as $x_n := \frac{1}{n}$, $y_n := \frac{3}{n^2}$ convergence to 0. Which sequence is faster?
2. Misalkan $x_h := \frac{\sin h}{h}$ dan $y_h := \frac{e^h - 1}{h}$. Buktikan kedua barisan ini konvergen ke 1 untuk $h \to 0$. Misalkan diambil $h := \frac{1}{n}$, tentukan $n \in \mathbb{N}$ terkecil masing-masing agar $|x_h - 1| < 10^{-4}$ dan $|y_h - 1| < 10^{-4}$. Barisan mana yang lebih cepat kekonvergenannya menuju 1? Dapatkah order konvergensi masing-masing barisan ditentukan? Visualisasikan pola kekonvergenan ini secara grafis.

Ilustrasi Grafis Contoh 1



Tugas Pekanan ke-2

1. Diberikan masalah komputasi $(\frac{1}{5} + \frac{3}{11}) - \frac{3}{20}$. Hitunglah kesalahan relatifnya jika setiap hasil perhitungan dibulatkan ke dalam bentuk tiga digit signifikan dengan metode:
   1. pemotongan (chooping)
   2. pembulatan (rounding).
2. Gunakan metode pembulatan ke tiga digit terdekat untuk perhitungan $5\pi - 7e + \frac{3}{67}$. Hitunglah galat mutlak dan galat relatifnya jika nilai eksaknya dihitung akurat sampai dengan lima digit.
3. Seandainya dalam konversi nilai sebuah mata kuliah adalah sebagai berikut: jika skor lebih dari atau sama dengan 80 maka mendapat A sedangkan skor lebih dari 75 dan kurang dari 80 mendapat nilai A-. Selanjutnya nilai A dikonversi lagi ke angka 4 dan A- dikonversi ke nilai angka 3.5. Misalkan dua orang mahasiswa masing-masing mendapat nilai 79.9 dan 80. Tentukan perbedaan relatif kedua mahasiswa ini sebelum dan sesudah pembulatan (konversi). Berikan ulasan.
4. Soal no 11 pada buku teks hal 29.
5. Soal no 12 pada buku teks hal 29.
6. Soal no 16 pada buku teks hal 30.
7. Mengapa kita perlu mengetahui order konvergesi barisan aproksimasi yang konvergen ke solusi eksak.