

Esercitazione n. 1
Sistemi aritmetici Floating Point a Precisione Finita
Corso di Calcolo Numerico
Corso di Laurea in Informatica
prof. Almerico Murli
a.a. 2004/2005

• **Esercizio 1**

Si consideri il seguente sistema aritmetico floating-point a precisione finita:

$$F(\beta = 10; t = 3; emin = -3; emax = +3) \quad .$$

Si rappresentino in F i numeri seguenti e si indichi l'errore relativo che si commette nella loro rappresentazione:

$$\begin{array}{llll} x = 12.13 & , & \tilde{x} = 0.121 \times 10^3 & , \text{ Non rappresentabile esattamente in } F; \\ y = 1.2 & , & \tilde{y} = 0.120 \times 10^2 & , \text{ Rappresentabile esattamente in } F; \\ z = 3467.864 & , & \text{—————} & , \text{ Overflow.} \end{array}$$

(I corrispondenti errori relativi sono $E_x = 0.363 \times 10^{-4}$ e $E_y = 0$).

• **Esercizio 2**

Si consideri il seguente sistema aritmetico floating-point a precisione finita:

$$F(\beta = 10; t = 5; emin = -9; emax = +9) \quad .$$

Nel sistema F , utilizzando l'arrotondamento e $t_{reg} = 2t$, si calcoli:

- l' ϵ macchina ($\epsilon_{mac} = 0.5 \times 10^{1-5}$);
- il massimo ed il minimo numero rappresentabile ($rmin = 0.1 \times 10^{-9}$, $rmax = 0.99999 \times 10^9$);
- il più piccolo numero y che non dá contributo alla somma con $x = 0.3211 \times 10^2$;
- $(a + b) \times c$ e $a \times c + b \times c$, dove $a = 0.3923 \times 10^4$, $b = 0.45 \times 10^0$, $c = 0.3 \times 10^2$, (si osserva che in F non vale la proprietà distributiva del prodotto rispetto alla somma);

– $(a + b) + c$ e $a + (b + c)$ dove $a = 0.2 \times 10^2$, $b = 0.323 \times 10^{-3}$, $c = 0.477 \times 10^{-3}$ (si osserva che in F non vale la proprietà associativa della somma);

• **Esercizio 3**

Si scriva un algoritmo che assegnato x valuti e^x .

(L'algoritmo può essere scritto sulla base della seguente formula ricorrente:

$$\begin{aligned} S_0 &= 1 \\ S_n &= S_{n-1} + \frac{x^n}{n!} \quad . \end{aligned}$$

Considerazioni sulla determinazione del più piccolo n dopo il quale $\frac{x^n}{n!}$ non dà più contributo alla somma con S_{n-1} .)