Lecture 4 (hrs. 5,6) - September 30, 2025, 16:30 - 18:30 F3

(1.16) Definition (finite precision floating point numbers).

Let β be an integer greater than or equal to two and let m be an integer greater than or equal to 1. The set

$$F(\beta,m) = \{0\} \cup \{x \text{ in } R \text{ s.t. } x = (-1)^s \beta^b\, 0.c_1 \ldots c_m \text{ where } s \in \{0,1\},\ b \in Z,\ c_1,\ldots,c_m \text{ radix } \beta \text{ digits},\ c_1 \neq 0\}$$

is called the ‘set of radix β *floating point* numbers with *precision* m’.

(1.17) Example.

Consider the set F(10,1).

- $1/100 \in F(10,1)$: $1/100 = 10^{-2} = 10^{-1}\ 0.1$
- $11/100 \notin F(10,1)$: $11/100 = 0.11 = 10^{0}\ 0.11$ and the fraction 0.11 is *not compatible* with precision m = 1
- all the positive elements of F(10,1) with zero exponent are:

  $$B = \{0.1 ; 0.2 ; \ldots ; 0.9\}$$

  all those with exponent $b \in Z$:

  $$10^b\ B \text{ (positive)} \quad , \quad -10^b\ B \text{ (negative)}$$

- $F(10,1) = \bigcup_{b \in Z} (-1)\ 10^b\ B \cup \{0\} \cup \bigcup_{b \in Z} 10^b\ B$

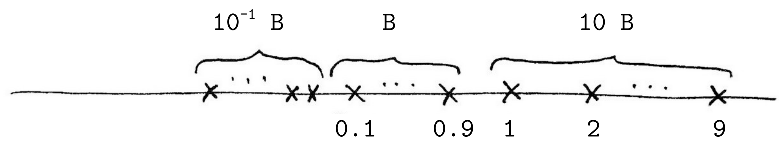
(1.18) Remark (properties of $F(\beta,m)$).

(1) it is a *proper subset* of Q (hence, it is countable and ordered)
(2) it is *symmetric* with respect to zero
(3) zero is its *unique* accumulation point
(4) $\sup F(\beta,m) = +\infty$, $\inf F(\beta,m) = -\infty$

(1.19) Remark (distance between consecutive elements).

In F(10,1):

(positive elements)



Distance between consecutive elements: $10^{-1}\ 0.1$ (b = -1), $0.1 = 10^{0}\ 0.1$ (b = 0), $1 = 10^{1}\ 0.1$ (b = 1).

- exponent = b $\Rightarrow$ distance between consecutive elements in F(10,1): $10^b\ 0.1 = 10^b\ 10^{-1}$

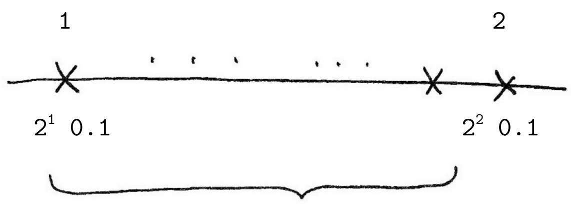
- in $F(\beta,m)$: given $\xi = \beta^b\, g$, let $\sigma(\xi)$ the *successor* of ξ, it is:

$$\sigma(\xi) - \xi = \beta^{b-m}$$

- the distance is greater the larger the exponent is (‘the farther ξ is from zero’).

(1.20) Remark.

In Example (1.10), Lecture 3, we have:



* $\alpha \in (1,2)$

* in *Scilab* (*Octave, Matlab*):

F(2,53)

$b = 1 \Rightarrow$ dist. between consec. elements $= 2^{1-53} = 2^{-52} \approx 2.22\ 10^{-16}$

- When $E = 10^{-16}$ the *bisezione function* found the smallest possible (non-degenerate) interval containing the zero α and with endpoints in F(2,53), but this interval has measure > E.
- *There's no point* in choosing $E < \beta^{b-m}$.

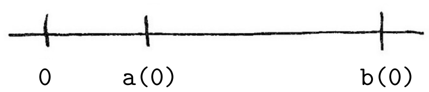
(1.21) Halt condition (with relative error bound).

Given a *positive* real number E...

$$\text{if } \frac{\text{measure } I(k)}{\min\{|a(k)|,|b(k)|\}} < E \text{ then STOP}$$

Properties of the halt condition:

(1) the condition is *computable*
(2) if $0 \notin I(0)$ we have: for every k, $0 \notin I(k)$ and

- 

$\Rightarrow \quad \min\{|a(k)|,|b(k)|\} = a(k) > 0$

since $a(0) \leqslant a(k) < b(0)$ then when $k \to \infty$ it is measure $I(k) / a(k) \to 0$

- 

$\Rightarrow \quad \min\{|a(k)|,|b(k)|\} = |b(k)| > 0$

since $|b(0)| \leqslant b(k) < |a(0)|$ then when $k \to \infty$ it is measure $I(k) / |b(k)| \to 0$

hence: the condition is *certainly* satisfied after a *finite* number of iterations (the criterion is *effective*).

(3) <u>if</u> f is a <u>continuous</u> function, then:

- there exists $\alpha \in I(k)$ zero of f

- $$\frac{|x(k) - \alpha|}{|\alpha|} \leqslant \frac{\text{measure } I(k) / 2}{|\alpha|} < 1/2 \, \frac{\text{measure } I(k)}{\min\{|a(k)|,|b(k)|\}} < E/2 < E$$

- x(k) approximates α with *relative error* < E: 'the procedure returns an approximation *as accurate as required by the user'*

- *there's no point* in choosing $E < \beta^{1-m}$