

Comparació de sistemes numèrics de punt fixe i punt flotant per a un aprenentatge significatiu

Proposta didàctica

Samuel S. Pérez

Curs 2025/26

UNIVERSITAT
DE VALÈNCIA  Facultat de Ciències Matemàtiques

DEPARTAMENT DE MATEMÀTIQUES

Atenció

La majoria de les diapositives han sigut extretes de les utilitzades a classe.

Motivació

- És clar que l'emmagatzematge i la manipulació dels nombres enters o reals en un ordinador ha d'involucrar només una quantitat finita de dígitos significatius.
- És per aquesta raó que els ordinadors utilitzen sistemes numèrics finits per a les representacions numèriques.
- Actualment s'utilitzen representacions anomenades de punt flotant.
- Altres dispositius utilitzen representacions de punt fixe.

Definició

Un sistema numèric de punt fixe $\mathcal{X} \subset \mathbb{R}$ és un subconjunt dels nombres reals format per nombres de la forma

$$y = \pm(s_1 s_2 \dots s_i . d_1 d_2 \dots d_f)_\beta. \quad (1)$$

El sistema $\mathcal{X}$ està caracteritzat per tres paràmetres enters:

- *la base β ,*
- *la precisió o longitud de paraula per a la part entera i ,*
- *la precisió o longitud de paraula per a la part fraccionaria f .*

Utilitzarem la notació $\mathcal{X}(\beta, i, f)$

Definició

Un sistema numèric de punt flotant $\mathcal{F} \subset \mathbb{R}$ és un subconjunt dels nombres reals format per nombres de la forma

$$y = \pm (.d_1 d_2 \dots d_t)_\beta \times \beta^e. \quad (2)$$

El sistema $\mathcal{F}$ està caracteritzat per quatre paràmetres enters:

- la base β ,
- la precisió o longitud de paraula t ,
- el rang dels exponents $e_{\min} \leq e \leq e_{\max}$.

La mantissa $m = (.d_1 d_2 \dots d_t)_\beta$ compleix $0 \leq m < 1$ i $0 \leq d_i < \beta$. Per assegurar una representació única, d_1 , ha de complir $d_1 \neq 0$. En aquest cas, la representació (2) és diu **normalitzada**, i t és la longitud de paraula de la representació.

Utilitzarem la notació $\mathcal{F}(\beta, t, e_{\min}, e_{\max})$

Explicant la semàntica

Sea 0.003741

$\mathcal{X}(10, 4, 4)$	$\mathcal{F}(10, 3, -2, 2)$
↓	↓
0000.0037	$0.374 \cdot 10^{-2}$
0	
.	
0	0
0	.
3	3
7	7
	4
	×
	10^{-2}

Sea 37.41

$\mathcal{X}(10, 4, 4)$	$\mathcal{F}(10, 3, -2, 2)$
↓	↓
0037.4100	$0.374 \cdot 10^2$
0	0
0	.
3	3
7	7
.	4
4	×
1	10^2

Noteu com, en la representació en el sistema numèric de punt flotant, el punt "flota"; com si els números foren líquid.

Exemple: Punt fixe

Per exemple, si $\beta = 2$, $i = 2$ i $f = 2$, aleshores els nombres en punt flotant positius són $(y = (s_1 s_2 . d_1 d_2)_2)$

m_f	m_i	m_i	m_i	m_i
00	.00	.01	.10	.11
01	.00	.01	.10	.11
10	.00	.01	.10	.11
11	.00	.01	.10	.11

0, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75,
2.00, 2.25, 2.50, 2.75, 3.00, 3.25, 3.50, 3.75

Taula $\beta = 2, i = 2$ i $f = 2$

m_i	00	00	00	00	01	01	01	01
m_f	.00	.01	.10	.11	.00	.01	.10	.11
y	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75
m_i	10	10	10	10	11	11	11	11
m_f	.00	.01	.10	.11	.00	.01	.10	.11
y	2.00	2.25	2.50	2.75	3.00	3.25	3.50	3.75

Exemple: Punt flotant

Per exemple, si $\beta = 2$, $t = 2$, $e_{min} = -1$ i $e_{max} = 3$, aleshores els nombres en punt flotant positius són ($y = (.d_1 d_2 d_3)_2 2^e$)

m	e	e	e	e	e
.00	-1	0	1	2	3
.01	-1	0	1	2	3
.10	-1	0	1	2	3
.11	-1	0	1	2	3

0, 0.125

0.25, 0.375, 0.50, 0.75, 1, 1.5, 2, 3, 4, 6

Taula $\beta = 2, t = 2, e_{min} = -1$ i $e_{max} = 3$

m	.00	.01	.10	.11	.00	.01	.10	.11
e	- 1	- 1	-1	- 1	0	0	0	0
y	0	0.125	0.25	0.375	0	0.25	0.50	0.75
m	.00	.01	.10	.11	.00	.01	.10	.11
e	1	1	1	1	2	2	2	2
y	0	0.50	1	1.50	0	1	2	3
m	.00	.01	.10	.11				
e	3	3	3	3				
y	0	2	4	6				

Exercici

Expressa en base 10 tots els nombres estrictament positius que es poden representar en els següents sistemes de punt fixe:

1 $\mathcal{X}(3, 1, 1)$.

2 $\mathcal{X}(2, 1, 2)$.

3 $\mathcal{X}(3, 2, 2)$.

4 $\mathcal{X}(2, 3, 2)$.

En els últims dos casos ajudat de funcions de Matlab, si fa falta

Exercici

Expressa en base 10 tots els nombres estrictament positius que es poden representar en els següents sistemes de punt flotant:

❶ $\mathcal{F}(3, 2, -1, 1)$.

❷ $\mathcal{F}(2, 3, -1, -1)$.

❸ $\mathcal{F}(3, 4, -2, 2)$.

❹ $\mathcal{F}(2, 5, -2, -2)$.

En els últims dos casos ajudat de funcions de Matlab, si fa falta

Màxim i mínim en un sistema de punt fixe i flotant

- Només hi ha una quantitat finita de nombres en un sistema numèric de punt fixe o flotant.
- Tots els nombres estrictament positius d'un sistema numèric de punt fixe, $y = (s_1 s_2 \dots s_i . d_1 d_2 \dots d_f)_\beta$, compleixen

$$\beta^{-f} \leq y \leq \beta^i - \beta^{-f}.$$

- Tots els nombres estrictament positius d'un sistema numèric de punt flotant, $y = (.d_1 d_2 \dots d_t)_\beta \times \beta^e$, compleixen

$$\beta^{e_{\min}-t} \leq y \leq \beta^{e_{\max}} - \beta^{e_{\max}-t}.$$

No és difícil aconseguir el mateix rang. Una condició suficient:

$$i = e_{\max}, \quad f = t - e_{\min}, \quad t = i + f$$

Representació en punt fixe i flotant

- La representació en punt fixe o flotant d'un nombre és una aplicació.

$$\begin{array}{rcl} fl & : & \mathbb{R} \rightarrow \mathcal{X}(\beta, i, f), \mathcal{F}(\beta, t, e_{min}, e_{max}) \\ & & x \mapsto fl(x) \end{array}$$

- En general la aplicació serà no injectiva
- Hi ha dues formes principals de traduir un nombre real x a la seua representació de punt flotant de longitud t i base β :
 - per **truncament** ($fl^T(x)$)
 - per **arrodoniment** ($fl^R(x)$)

Representació en punt fixe

Siga $x = \pm(s_1 \dots s_i . d_1 d_2 \dots d_f d_{f+1} \dots)_\beta$ on $d_f \neq 0$ i $0 \leq s_j, d_j < \beta$.

- Representació per truncament:

$$fl^T(x) = \pm(s_1 \dots s_i . d_1 d_2 \dots d_t)_\beta.$$

- Representació per arredoniment:

$$fl^R(x) = \pm(s_1 \beta^{i-1} + \dots + s_i + d_1 \beta^{-1} + d_2 \beta^{-2} + \dots + d_{t-1} \beta^{-(t-1)} + \tilde{d}_t \beta^{-t}),$$

on

$$\tilde{d}_t = \begin{cases} d_t, & 0 \leq (.d_{t+1} d_{t+2} \dots)_\beta < 1/2 \\ d_t + 1, & 1/2 \leq (.d_{t+1} d_{t+2} \dots)_\beta < 1. \end{cases}$$

$$\text{sent } (.d_{t+1} d_{t+2} \dots)_\beta = d_{t+1} \beta^{-1} + d_{t+2} \beta^{-2} + \dots$$

Representació en punt flotant

Siga $x = \pm(.d_1 d_2 \dots d_t d_{t+1} \dots)_\beta \cdot \beta^e$ on $d_1 \neq 0$ i $0 \leq d_i < \beta$.

- Representació per truncament:

$$fl^T(x) = \pm(.d_1 d_2 \dots d_t)_\beta \cdot \beta^e.$$

- Representació per arredoniment:

$$fl^R(x) = \pm(d_1 \beta^{-1} + d_2 \beta^{-2} + \dots + d_{t-1} \beta^{-(t-1)} + \tilde{d}_t \beta^{-t}) \cdot \beta^e,$$

on

$$\tilde{d}_t = \begin{cases} d_t, & 0 \leq (.d_{t+1} d_{t+2} \dots)_\beta < 1/2 \\ d_t + 1, & 1/2 \leq (.d_{t+1} d_{t+2} \dots)_\beta < 1. \end{cases}$$

$$\text{sent } (.d_{t+1} d_{t+2} \dots)_\beta = d_{t+1} \beta^{-1} + d_{t+2} \beta^{-2} + \dots$$

Exemples

Errors

- La representació en punt fixe o flotant implica un error de representació.
- Cal estimar estos errors i controlar quan siga possible la seua propagació.

Error absolut i relatiu

Si $x \in \mathbb{R}$ i $\tilde{x}$ és una aproximació a x , definim:

- **L'error absolut** de l'aproximació: $E_a(\tilde{x}) = |x - \tilde{x}|$.
- **L'error relatiu** de l'aproximació: $E_r(\tilde{x}) = \frac{|x - \tilde{x}|}{|x|}$.

Exemples

Proposició

Si $x = (s_1 \dots s_i . d_1 \dots d_f d_{f+1} d_{f+2} \dots)_\beta \in \mathbb{R}$, amb $d_j \neq 0$ per a algún $1 \leq j \leq f$, i $fl^T(x)$ i $fl^R(x)$ són les representacions de punt fixe amb i dígits sencers, f dígits fraccionaris i base β , obtingudes respectivament per truncament i arrodoniment, aleshores:

$$E_a(fl^T(x)) \leq \beta^{-f}, \quad E_a(fl^R(x)) \leq \frac{1}{2} \cdot \beta^{-f},$$

$$E_r(fl^T(x)) \leq \beta^{j-i}, \quad E_r(fl^R(x)) \leq \frac{1}{2} \cdot \beta^{j-i}.$$

Tenim $j = 1$ si $d_1 \neq 0$

Proposició

Si $x = (.d_1 \dots d_t d_{t+1} d_{t+2} \dots)_\beta \cdot \beta^e \in \mathbb{R}$, amb $d_j \neq 0$ per a algún $1 \leq j \leq t$, i $fl^T(x)$ i $fl^R(x)$ són les representacions de punt flotant (normalitzades si $e > e_{\min}$) amb t dígit i base β , obtingudes respectivament per truncament i arrodoniment, aleshores:

$$E_a(fl^T(x)) \leq \beta^{e-t}, \quad E_a(fl^R(x)) \leq \frac{1}{2} \cdot \beta^{e-t},$$

$$E_r(fl^T(x)) \leq \beta^{j-t}, \quad E_r(fl^R(x)) \leq \frac{1}{2} \cdot \beta^{j-t}.$$

Tenim $j = 1$ si $d_1 \neq 0$

Proves

Errors: comparació de fites

En principi la precisió d'ambdós sistemes son comparables:

$$\frac{\text{fita} (E_a^{\mathcal{F}} (fl^{T,R}(x)))}{\text{fita} (E_a^{\mathcal{X}} (fl^{T,R}(x)))} = \frac{\beta^{e-t}}{\beta^{-f}} = \beta^{e-t+f}$$

$$\frac{\text{fita} (E_r^{\mathcal{F}} (fl^{T,R}(x)))}{\text{fita} (E_r^{\mathcal{X}} (fl^{T,R}(x)))} = \frac{\beta^{j-t}}{\beta^{j-f}} = \beta^{f-t}$$

Ara bé, recordem que l'elecció $i = e_{\max}$, $f = t - e_{\min}$ i $t = i + f$ fa que tinguem el **mateix rang i també mateix valor màxim i mínim**, en ambdós sistemes de punt fixe i flotant. Però en aquest cas:

$$\frac{\text{fita} (E_a^{\mathcal{F}} (fl^{T,R}(x)))}{\text{fita} (E_a^{\mathcal{X}} (fl^{T,R}(x)))} = \beta^{e-t+f} = \beta^{e-i} = \beta^{e-e_{\max}} < 1$$

$$\frac{\text{fita} (E_r^{\mathcal{F}} (fl^{T,R}(x)))}{\text{fita} (E_r^{\mathcal{X}} (fl^{T,R}(x)))} = \beta^{f-t} = \beta^{-i} < 1$$

Per tant, **guanyem precisió amb el sistema de punt flotant**.

Comparació punt fixe i punt flotant

Un podria pensar que, fixat un rang, estem reservant més espai en memòria en la representació de punt flotant a expenses de guanyar precisió.

Veiem un exemple, més a prop de la manera de treballar de les computadores actuals, on es reserva la mateixa quantitat de memòria per emmagatzemar un número.

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 4 bits per a la part entera.
- 3 bits per a la part fraccionaria.

Punt flotant:

- 4 bits per a la longitud de paraula.
- 3 bits per a l'exponent:

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 4 bits per a la part entera.
- 3 bits per a la part fraccionaria.

$$x_{\max}^{\mathcal{X}} = (1111.111)_2 = 15.875$$

Punt flotant:

- 4 bits per a la longitud de paraula.
- 3 bits per a l'exponent:

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 4 bits per a la part entera.
- 3 bits per a la part fraccionaria.

$$x_{\max}^{\mathcal{X}} = (1111.111)_2 = 15.875$$

$$x_{\min}^{\mathcal{X}} = (0000.001)_2 = 0.125$$

Punt flotant:

- 4 bits per a la longitud de paraula.
- 3 bits per a l'exponent:

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 4 bits per a la part entera.
- 3 bits per a la part fraccionaria.

$$x_{\max}^{\mathcal{X}} = (1111.111)_2 = 15.875$$

$$x_{\min}^{\mathcal{X}} = (0000.001)_2 = 0.125$$

Punt flotant:

- 4 bits per a la longitud de paraula.
- 3 bits per a l'exponent:
 - ▶ $e_{\max} = (111)_2 = 7.$

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 4 bits per a la part entera.
- 3 bits per a la part fraccionaria.

$$x_{\max}^{\mathcal{X}} = (1111.111)_2 = 15.875$$

$$x_{\min}^{\mathcal{X}} = (0000.001)_2 = 0.125$$

Punt flotant:

- 4 bits per a la longitud de paraula.
- 3 bits per a l'exponent:
 - ▶ $e_{\max} = (111)_2 = 7.$
 - ▶ $e_{\min} = (000)_2 = 0.$

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 4 bits per a la part entera.
- 3 bits per a la part fraccionaria.

$$x_{\max}^{\mathcal{X}} = (1111.111)_2 = 15.875$$

$$x_{\min}^{\mathcal{X}} = (0000.001)_2 = 0.125$$

Punt flotant:

- 4 bits per a la longitud de paraula.
- 3 bits per a l'exponent:
 - ▶ $e_{\max} = (111)_2 = 7.$
 - ▶ $e_{\min} = (000)_2 = 0.$
 - ▶ 8 exponents disponibles que es traslladen de manera que $-3 \leq e \leq 4.$

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 4 bits per a la part entera.
- 3 bits per a la part fraccionaria.

$$x_{\max}^{\mathcal{X}} = (1111.111)_2 = 15.875$$

$$x_{\min}^{\mathcal{X}} = (0000.001)_2 = 0.125$$

Punt flotant:

- 4 bits per a la longitud de paraula.
- 3 bits per a l'exponent:
 - ▶ $e_{\max} = (111)_2 = 7$.
 - ▶ $e_{\min} = (000)_2 = 0$.
 - ▶ 8 exponents disponibles que es traslladen de manera que $-3 \leq e \leq 4$.

$$x_{\max}^{\mathcal{F}} = (.1111)_2 \cdot 2^4 = 15$$

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 4 bits per a la part entera.
- 3 bits per a la part fraccionaria.

$$x_{\max}^{\mathcal{X}} = (1111.111)_2 = 15.875$$

$$x_{\min}^{\mathcal{X}} = (0000.001)_2 = 0.125$$

Punt flotant:

- 4 bits per a la longitud de paraula.
- 3 bits per a l'exponent:
 - ▶ $e_{\max} = (111)_2 = 7$.
 - ▶ $e_{\min} = (000)_2 = 0$.
 - ▶ 8 exponents disponibles que es traslladen de manera que $-3 \leq e \leq 4$.

$$x_{\max}^{\mathcal{F}} = (.1111)_2 \cdot 2^4 = 15$$

$$x_{\min}^{\mathcal{F}} = (.0001)_2 \cdot 2^{-3} \approx 10^{-2}$$

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 3 bits per a la part entera.
- 4 bits per a la part fraccionaria.

Punt flotant:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 3 bits per a la part entera.
- 4 bits per a la part fraccionaria.

$$x_{\max}^{\mathcal{X}} = (111.1111)_2 = 7.9375$$

Punt flotant:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 3 bits per a la part entera.
- 4 bits per a la part fraccionaria.

$$x_{\max}^{\mathcal{X}} = (111.1111)_2 = 7.9375$$

$$x_{\min}^{\mathcal{X}} = (000.0001)_2 = 0.0625$$

Punt flotant:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 3 bits per a la part entera.
- 4 bits per a la part fraccionaria.

$$x_{\max}^{\mathcal{X}} = (111.1111)_2 = 7.9375$$

$$x_{\min}^{\mathcal{X}} = (000.0001)_2 = 0.0625$$

Punt flotant:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:
 - ▶ $e_{\max} = (1111)_2 = 15$.

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 3 bits per a la part entera.
- 4 bits per a la part fraccionaria.

$$x_{\max}^{\mathcal{X}} = (111.1111)_2 = 7.9375$$

$$x_{\min}^{\mathcal{X}} = (000.0001)_2 = 0.0625$$

Punt flotant:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:
 - ▶ $e_{\max} = (1111)_2 = 15.$
 - ▶ $e_{\min} = (0000)_2 = 0.$

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 3 bits per a la part entera.
- 4 bits per a la part fraccionaria.

$$x_{\max}^{\mathcal{X}} = (111.1111)_2 = 7.9375$$

$$x_{\min}^{\mathcal{X}} = (000.0001)_2 = 0.0625$$

Punt flotant:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:
 - ▶ $e_{\max} = (1111)_2 = 15$.
 - ▶ $e_{\min} = (0000)_2 = 0$.
 - ▶ 16 exponents disponibles que es traslladen de manera que $-7 \leq e \leq 8$.

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 3 bits per a la part entera.
- 4 bits per a la part fraccionaria.

$$x_{\max}^{\mathcal{X}} = (111.1111)_2 = 7.9375$$

$$x_{\min}^{\mathcal{X}} = (000.0001)_2 = 0.0625$$

Punt flotant:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:
 - ▶ $e_{\max} = (1111)_2 = 15$.
 - ▶ $e_{\min} = (0000)_2 = 0$.
 - ▶ 16 exponents disponibles que es traslladen de manera que $-7 \leq e \leq 8$.

$$x_{\max}^{\mathcal{F}} = (.111)_2 \cdot 2^8 = 224$$

Comparació punt fixe i punt flotant

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt fixe:

- 3 bits per a la part entera.
- 4 bits per a la part fraccionaria.

$$x_{\max}^{\mathcal{X}} = (111.1111)_2 = 7.9375$$

$$x_{\min}^{\mathcal{X}} = (000.0001)_2 = 0.0625$$

Punt flotant:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:
 - ▶ $e_{\max} = (1111)_2 = 15$.
 - ▶ $e_{\min} = (0000)_2 = 0$.
 - ▶ 16 exponents disponibles que es traslladen de manera que $-7 \leq e \leq 8$.

$$x_{\max}^{\mathcal{F}} = (.111)_2 \cdot 2^8 = 224$$

$$x_{\min}^{\mathcal{F}} = (.001)_2 \cdot 2^{-7} \approx 10^{-3}$$

En punt fixe es sacrifica prou en el rang a expenses de guanyar precisió.

Mentre que en punt flotant, dedicar més bits per a l'exponent, augmenta prou el rang i la precisió queda del mateix ordre.

Errors: comparació de rangs

Imaginem ara que volem la mateixa precisió relativa. Això s'aconsegueix amb $t = f$. Així el mateix nombre de bits per a la part sencera i en punt fixe s'han dedicat al exponent e en punt flotant. Es a dir,

$$e_{\max} = 2^i - 1, \quad e_{\min} = 0,$$

i el número d'exponents és $e_{\max} - e_{\min} + 1 = 2^i$. Però, com es fa la translació:

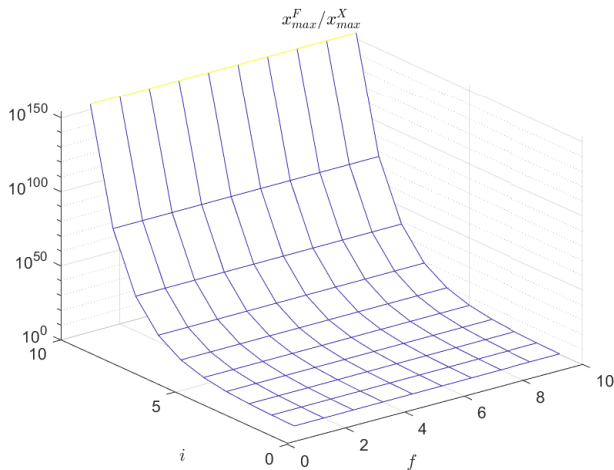
$$e_{\max} = 2^i - 2^{i-1} = 2^{i-1}, \quad e_{\min} = -2^{i-1} + 1.$$

Recordant el màxim i el mínim en cada sistema numèric ($\beta = 2$):

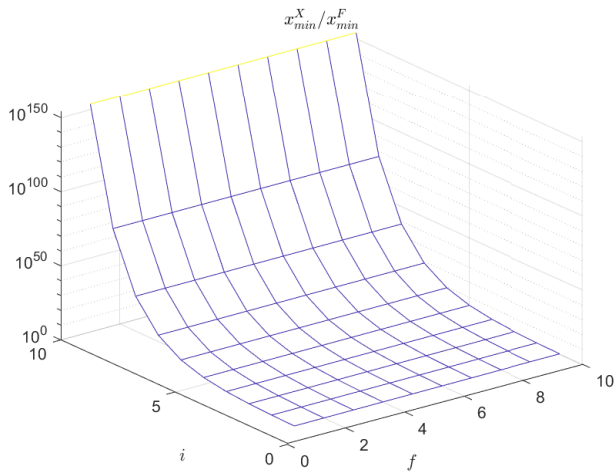
$$x_{\max}^{\mathcal{X}} = 2^i - 2^{-f}, \quad x_{\max}^{\mathcal{F}} = 2^{e_{\max}}(1 - 2^{-t}) = 2^{2^{i-1}}(1 - 2^{-f})$$

$$x_{\min}^{\mathcal{X}} = 2^{-f}, \quad x_{\min}^{\mathcal{F}} = 2^{e_{\min}-t} = 2^{-2^{i-1}+1-f}$$

Errors: comparació de rangs



Errors: comparació de rangs



Avatantages i desavantages

- Punt fixe
 - ▶ Gran simplicitat en l'artimètica
 - ▶ És molt fàcil perdre rang per guanyar precisió.
- Punt flotant
 - ▶ Aritmètica més complexa. En sumes i restes s'han d'alinear les mantisses primer.
 - ▶ Rang molt més ampli de representació a mateixa precisió.

Aperitiu de l'estàndard IEEE

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt flotant amb notació normalitzada $d_1 \neq 0$:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:

Aperitiu de l'estàndard IEEE

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt flotant amb notació normalitzada $d_1 \neq 0$:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:

Aperitiu de l'estàndard IEEE

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt flotant amb notació normalitzada $d_1 \neq 0$:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:

Aperitiu de l'estàndard IEEE

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt flotant amb notació normalitzada $d_1 \neq 0$:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:
 - ▶ $e_{\max} = (1111)_2 = 15$.

Aperitiu de l'estàndard IEEE

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt flotant amb notació normalitzada $d_1 \neq 0$:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:
 - ▶ $e_{\max} = (1111)_2 = 15$.
 - ▶ 0 i 15 es reserven per al 0 i ∞ del sistema.

Aperitiu de l'estàndard IEEE

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt flotant amb notació normalitzada $d_1 \neq 0$:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:
 - ▶ $e_{\max} = (1111)_2 = 15$.
 - ▶ 0 i 15 es reserven per al 0 i ∞ del sistema.
 - ▶ 14 exponents disponibles que es traslladen de manera que $-6 \leq e \leq 7$.

Aperitiu de l'estàndard IEEE

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt flotant amb notació normalitzada $d_1 \neq 0$:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:
 - ▶ $e_{\max} = (1111)_2 = 15$.
 - ▶ 0 i 15 es reserven per al 0 i ∞ del sistema.
 - ▶ 14 exponents disponibles que es traslladen de manera que $-6 \leq e \leq 7$.

$$x_{\max}^{\mathcal{F}} = (.111)_2 \cdot 2^7 \approx 10^2$$

Aperitiu de l'estàndard IEEE

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt flotant amb notació normalitzada $d_1 \neq 0$:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:
 - ▶ $e_{\max} = (1111)_2 = 15$.
 - ▶ 0 i 15 es reserven per al 0 i ∞ del sistema.
 - ▶ 14 exponents disponibles que es traslladen de manera que $-6 \leq e \leq 7$.

$$x_{\max}^{\mathcal{F}} = (.111)_2 \cdot 2^7 \approx 10^2$$

$$x_{\min}^{\mathcal{F}} = (.100)_2 \cdot 2^{-6} \approx 10^{-2}$$

Aperitiu de l'estàndard IEEE

Fixem $\beta = 2$ i un espai de 8 bits (1 bit es dedica per al signe):

Punt flotant amb notació normalitzada $d_1 \neq 0$:

- 3 bits per a la longitud de paraula.
- 4 bits per a l'exponent:
 - ▶ $e_{\max} = (1111)_2 = 15$.
 - ▶ 0 i 15 es reserven per al 0 i ∞ del sistema.
 - ▶ 14 exponents disponibles que es traslladen de manera que $-6 \leq e \leq 7$.

$$x_{\max}^{\mathcal{F}} = (.111)_2 \cdot 2^7 \approx 10^2$$

$$x_{\min}^{\mathcal{F}} = (.100)_2 \cdot 2^{-6} \approx 10^{-2}$$

Noteu que $d_1 = 1$ sempre!

Estàndard IEEE (Institute of Electrical and Electronics Engineers)

- Tots els ordinadors construïts des de 1985 utilitzen l'aritmètica de punt flotant IEEE.
- $\beta = 2$ i expressió normalitzada especial:

$$(1.d_1d_2\dots d_t)_2 \cdot 2^e$$

- L'estàndard de l'IEEE especifica dos formats bàsics simple precisió i doble precisió.
- La finitud de la mantissa és una limitació en la *precisió*.
- La finitud de l'exponent és una limitació en el *rang*.
- Els nombre fora de rang seran 0 (underflow) o inf (overflow). NaN indica valors no numèrics ($0 \cdot \infty$, $0/0$, etc.)

Tipus	Grandària	Mantissa	Exponent	Fita del error relatiu	Rang
Simple	32 bits	23 + 1 bits	8 bits	$2^{-24} \approx 5.96 \times 10^{-8}$	$10^{\pm 38}$
Doble	64 bits	52 + 1 bits	11 bits	$2^{-53} \approx 1.11 \times 10^{-16}$	$10^{\pm 308}$

Alguns exemples en MATLAB

Exemple 1

```
>> a=(2-1/2^52)*2^1023;  
>> b=a+1;  
>> b-a  
ans =  
0
```

Exemple 2

```
>> a=(2-1/2^52)*2^1023;  
>> b=2*a;  
>> b-a  
ans =  
Inf
```

ja que $2*a$ és Inf.

Exemple 3

```
>> a=(2-1/2^52)*2^1023;  
>> b=3*a-2*a  
ans =  
NaN
```

ja que $3*a-2*a$ és Inf-Inf.

Exemple 4

```
>> realmax + 2^1023*2^(-54)  
  
ans =  
  
1.797693134862316e+308  
  
realmax=1.797693134862316e+308
```

Conclusions

Estratègies anteriors

- Explicació de punt flotant directament.
- Explicació de l'estàndardIEEE sense exemples bàsics previs

Conclusions

Estratègies anteriors

- Explicació de punt flotant directament.
- Explicació de l'estàndard IEEE sense exemples bàsics previs

Hem introduït

- Explicació de punt fixe.
- Comparació entre punt fixe i punt flotant en termes precisió i rang.
- Proposta d'activitats de càlcul de conjunts.
- Explicació exemples bàsics previs a l'estàndard IEEE

Conclusions

Estratègies anteriors

- Explicació de punt flotant directament.
- Explicació de l'estàndard IEEE sense exemples bàsics previs

Hem introduït

- Explicació de punt fixe.
- Comparació entre punt fixe i punt flotant en termes precisió i rang.
- Proposta d'activitats de càlcul de conjunts.
- Explicació exemples bàsics previs a l'estàndard IEEE

Resultats

- Alumnes més motivats: comentaris i preguntes a classe.
- Millors resultats: mateixos resultats amb proves de major dificultat.

Conclusions

Estratègies anteriors

- Explicació de punt flotant directament.
- Explicació de l'estàndard IEEE sense exemples bàsics previs

Hem introduït

- Explicació de punt fixe.
- Comparació entre punt fixe i punt flotant en termes precisió i rang.
- Proposta d'activitats de càlcul de conjunts.
- Explicació exemples bàsics previs a l'estàndard IEEE

Resultats

- Alumnes més motivats: comentaris i preguntes a classe.
- Millors resultats: mateixos resultats amb proves de major dificultat.

Conclusions docents

La comparació entre punt flotant i punt fixe contribueix a

- Valorar la utilització del punt flotant en el càlcul computacional.
- Tindre un aprenentatge significatiu en aritmètica de punt flotant...