Facultat de Matemàtiques i Estadística

Universitat Politècnica de Catalunya - BarcelonaTech

Àlgebra Lineal Numèrica (Q2)

Àlex Batlle Casellas

March 26, 2019

Índex

0 Aritmètica finita i control d'errors

Representem els nombres en un sistema de numeració posicional dependent d'una certa base b. Per representar un nombre, seguim el següent esquema:

$$(d_p d_{p-1} \dots d_0 . d_{-1} \dots d_{-q}) = d_p b^p + \dots + d_1 b^1 + d_0 + d_{-1} b^{-1} + \dots + d_{-q} b^{-q} = \sum_{i=-q}^{p} d_i b^i.$$

En un ordinador, representem els nombres en binari; anem a veure com representem els enters i els reals:

Enters Els enters els representem de la forma següent:

d_{s-1}	d_{s-2}	$\cdots$	d_1	d_0

Utilitzem s bits, i el primer és pel signe ($d_{s-1} = 1 \to -, d_{s-1} = 0 \to +$). Per tant, els enters en un ordinador es representen com una suma així

$$(-1)^{s-1} \sum_{i=0}^{s-2} d_i 2^i.$$

D'aquí, deduïm que el nombre màxim que podem representar és

$$|N_{\text{màx}}| = \sum_{i=0}^{s-2} 2^i = 2^{s-1} - 1.$$

En C/C++, que és el llenguatge que utilitzarem, tenim els següent tipus de dades

Type	**Bytes**	**Bits**	$N_{\text{màx}}$
char	1	8	$2^7 - 1 = 127$
short	2	16	$2^{15} - 1 = 32767$
int	4	32	$2^{31} - 1 = 2147483647$
unsigned int	4	32	$2^{32} - 1 = 4294967295$
float	4	32 (M:24,E:8)	$1.7815 \cdot 10^{38}$
double	8	64 (M:53,E:11)	$0.8988 \cdot 10^{308}$

1 Sistemes Lineals.