Virgule flottante : zéro, écart et rôle du biais

1. La représentation du zéro

En virgule flottante, la formule générale est

$$x = \pm(1.mantisse) \times 2^{e-\text{biais}}.$$

Mais on fait une exception quand l'exposant est nul :

$$x = \pm(0.mantisse) \times 2^{1-\text{biais}}.$$

Ainsi, le zéro est obtenu quand

$$\text{exposant} = 0, \quad \text{mantisse} = 0.$$

2. Le problème sans biais : un écart trop grand

Supposons que l'on n'utilise pas de biais et que $e = e_{\text{codé}}$.

— Pour $e = 0$ (subnormaux) :

$$x = \pm(0.mantisse) \times 2^{1-0} = \pm(0.mantisse) \times 2,$$

ce qui couvre des valeurs dans $[0, 2)$.

— Pour $e = 1$ (normalisés) :

$$x = \pm(1.mantisse) \times 2^1 \in [2, 4).$$

Exemple avec 4 bits de mantisse :

$$x_{\text{max subnorm}} = 0.1111_2 \times 2 = 1.875, \qquad x_{\text{min normalisé}} = 1.0000_2 \times 2 = 2.$$

L'écart relatif près de 2 devient ainsi important.

Lien avec la représentation simplifiée vue auparavant

$$x = (1.mantisse) \times 2^e \ :$$

les codes qui représentaient l'intervalle $[1, 2)$ dans cette version sont, sans biais, étendus pour couvrir $[0, 2)$. La densité de représentations baisse alors fortement dans $[1, 2)$, ce qui explique l'écart observé.

3. Introduction du biais

Pour réduire cet écart et équilibrer la représentation :

$$e_{\text{effectif}} = e_{\text{codé}} - \text{biais}.$$

Avec ce décalage :

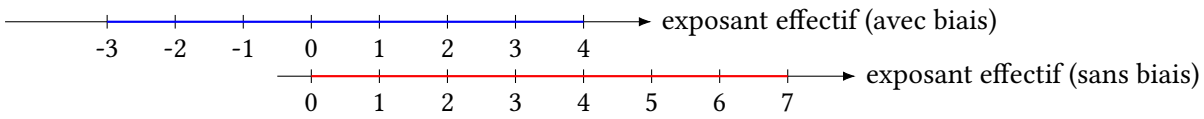
— On obtient des exposants négatifs, ce qui permet des plages $[0.5, 1)$, $[1, 2)$, etc.

— Les subnormaux se connectent sous $2^{1-\text{biais}}$, avec une granularité plus fine.

— L'erreur relative reste plus régulière sur toute la plage.

Exemple avec 3 bits d'exposant

- Sans biais : $e_{\text{codé}} \in [0, 7] \implies e_{\text{effectif}} \in [0, 7]$.
- Avec biais = 3 : $e_{\text{codé}} \in [0, 7] \implies e_{\text{effectif}} \in [-3, 4]$.



4. Pourquoi pas le complément à deux ?

On pourrait imaginer coder l'exposant en complément à deux. Cependant, le biais est préféré pour deux raisons principales :

- Circuits plus simples : l'exposant effectif s'obtient par une simple soustraction, pas besoin d'interpréter un entier signé.
- Comparaison naturelle : les exposants codés restent dans l'ordre des entiers non signés. Comparer deux flottants revient donc à comparer directement les champs binaires (utile pour tri, minimum/maximum).

Avec complément à deux, l'ordre des codes binaires ne correspondrait plus à l'ordre numérique.

5. Tableau récapitulatif

Exposant	Mantisse	Valeur représentée	Cas
0	0	0	zéro
0	$\neq 0$	$\pm(0.\text{mantisse}) \times 2^{1-\text{biais}}$	subnormaux
$\neq 0$	$--$	$\pm(1.\text{mantisse}) \times 2^{e-\text{biais}}$	normalisés

Appendix A : Analyse de plage pour le format $1/3/4$

Avec biais = 3

$$x^{\text{sub}}_{\text{min}>0} = 2^{-(p+2)} = 2^{-6} = 1/64, \qquad x^{\text{sub}}_{\text{max}} = 15/64.$$

$$x^{\text{norm}}_{\text{min}} = 1.0 \times 2^{-2} = 1/4, \qquad x^{\text{norm}}_{\text{max}} = (31/16) \cdot 8 = 15.5.$$

Granularité près de $[1, 2) : \Delta = 2^{-p} = 1/16$.

Sans biais

$$x^{\text{sub}}_{\text{min}>0} = 2^{1-p} = 1/8, \qquad x^{\text{sub}}_{\text{max}} = 1.875.$$

$$x^{\text{norm}}_{\text{min}} = 2, \qquad x^{\text{norm}}_{\text{max}} = 248.$$

Granularité près de $[1, 2) : \Delta = 1/8$.

Synthèse Dans $[1, 2)$, avec biais, les pas valent $1/16$, contre $1/8$ sans biais. Le biais améliore donc la précision relative dans cette zone critique.