Université Paris Descartes
UFR de Mathématiques et Informatique
45, rue des Saints-Pères 75270 Paris cedex 06

Numération et Logique
TD 4 - Représentation des nombres réels en machine

Objectif : Manipulation des nombres en virgule flottante

Exercice 1 : codage des réels en base 10

On se place dans la base décimale (base dix), en virgule flottante normalisée. On représente les nombres réels dans un registre de 7 cases : l'exposant signé est placé le plus à gauche et comporte 2 cases (le signe et 1 chiffre décimal) et la mantisse signée comporte 5 cases (le signe et 4 chiffres décimaux).

1. Quel réel est représenté par | - | + | 2 | 3 | 4 | 5 | 6 | ?

   $-0,3456 * 10^2 = -34,56$

2. Donner le plus petit (puis le plus grand) nombre strictement positif exprimable. Le plus petit (puis le plus grand) nombre strictement négatif exprimable.

   Le plus petit positif : -9+1000= $0,1 * 10^{-9}$, le plus grand positif : +9+9999=$0,9999 * 10^9$
   Le plus grand négatif : -9-1000=$-0,1 * 10^{-9}$, le plus petit négatif : +9-9999=$-0,9999 * 10^9$

3. Combien de nombres réels différents sont représentables ? Expliquer le résultat.

   9 chiffres pour la 1iere case, 10 pour les trois autres ! $9 * 10^3 = 9000$
   Pour l'exposant, on peut aller de -9 à +9 soit 19 possibilités.
   Résultat : 19x9000=$171 * 10^3$
   De même pour les strict. négatifs ce qui fait un total de 342000 réels représentables plus 1 pour le 0.

4. Quel est le successeur de + $0,9999 * 10^{-8}$? Quelle est la différence entre ces 2 nombres ?

   Successeur de + $0,9999 * 10^{-8}$ est $0,1 * 10^{-7} = 10^{-8}$
   Il y a $0,0001 * 10^{-8}$ de différence entre les deux soit $1 * 10^{-12}$

5. Donner un exemple de deux autres nombres, l'un étant successeur de l'autre et tels que leur différence n'ait pas la même valeur que la différence de la question précédente.

   $0,1 * 10^0$ et son successeur $0,1001 * 10^0$ ont $1 * 10^{-4}$ de différence.

   Autre exemple : $0,8888 * 10^8$ et $0,8889 * 10^8$, la différence est $0,0001 * 10^8 = 10^4$

6. Donner la précision machine ε, définie par le fait que $1 + \varepsilon$ est le successeur de 1.

   Par les cours $\varepsilon = b^{1_p} = 10^{1-4}$, p étant la taille de la mantisse. On peut aussi refaire le calcul direct, on a d'une part $1 = 0,1000 * 10^1$ et son successeur $1 + \varepsilon = 0,1001 * 10^1$, ces nombre sont distant de $\varepsilon = 0,0001 * 10^1 = 10^{-4} * 10^1 = 10^{-3}$.

Exercice 2 : codage des réels en base 2

1. On considère des flottants de la forme $\pm\ (0,d_{-1}d_{-2}d_{-3})_2 * (2^e)_{10}$. La mantisse est normalisée donc $d_{-1} = 1$ et $d_{-2}, d_{-3} \in \{0,1\}$ tandis que $e_{10} \in \{-1,0,1,2\}$.

(a) Comment représenter 0 ?

(b) Combien de nombres peut-on représenter ?

(c) Tracer les nombres positifs représentables sur la droite réelle.

(d) Illustrer des cas de overflow, underflow et erreurs d'arrondi dans ce système.

(e) Que se passe-t-il si l'on ajoute un bit caché ?

(f) Donner la précision machine sans et avec la technique du bit caché.

(a) 0 peut être représenté par que des zéros (ou une autre convention).

(b) Il y a $2*1*2*2*4+1=33$ nombres représentables.

(c) Les quatre mantisses possibles sont (en binaire) :

$$0,100=(1/2)_{10}\,;0,101=(1/2+1/8)_{10}\,;0,110=(1/2+1/4)_{10}\,;0,111=(1/2+1/4+1/8)_{10}$$

On a quatre exposants possibles :
Pour $e=-1$, on multiplie par $(2^{-1})_{10}$ et l'on a les nombres :

$$(1/4)_{10}\,;(1/4+1/16)_{10}\,;(1/4+1/8)_{10}\,;(1/4+1/8+1/16=7/16)_{10}$$

Pour $e=0$, on multiplie par $(2^0=1)_{10}$ et l'on a les nombres :

$$(1/2)_{10}\,;(1/2+1/8)_{10}\,;(1/2+1/4)_{10}\,;(1/2+1/4+1/8)_{10}$$

Pour $e=+1$, on multiplie par $(2^1)_{10}$ et l'on a les nombres :

$$(1)_{10}\,;(1+1/4)_{10}\,;(1+1/2)_{10}\,;(1+1/2+1/4)_{10}$$

Pour $e=+2$, on multiplie par $(2^2=4)_{10}$ et l'on a les nombres :

$$(2)_{10}\,;(2+1/2)_{10}\,;(2+1=3)_{10}\,;(2+1+1/2=3,5)_{10}$$

Plus le 0, donc facile à tracer. On (re)constate que les nombres ne sont pas espacés uniformément.

(d) Overflow avec 2+3.
Comme 1/4 est le plus nb. strict. positif, on obtint un underflow avec
$1/2-7/16=1/16=(0,0001)_2*(2^0)_{10}=(0,001)_2*(2^{-1})_{10}$ nombre sous-normal.
Erreur d'arrondi avec $a=-3$, $b=+3$ et $c=1/4$:
On a $(a+b)+c=c=1/4$. mais $b+c=3,25=2+1+1/4=(11,01)_2=(0,1101)_2*(2^2)$ n'est pas exactement représentable car la mantisse est trop grande ! Le résultat sera arrondi.
Si $b+c=(0,110\,)_2*(2^2)=3=b$, alors $a+(b+c)=a+b=0$.
Si $b+c=(0,111\,)_2*(2^2)=3,5$, alors $a+(b+c)=0,5$.
Dans aucun des cas on a l'égalité $(a+b)+c=a+(b+c)$.

(e) Avec un bit caché on multiplie les mantisses possibles par 2, en binaire :

$$0,1000\,;0,1001\,;0,1010\,;0,1011\,;0,1100\,;0,1101\,;0,1110\,;0,1111$$

ce qui donne en base 10 : $1/2, 1/2+1/16, 1/2+1/8, 1/2+1/8+1/16, 1/2+1/4, 1/2+1/4+1/8, 1/2+1/4+1/8+1/16$

(f) Sans bit caché, on a $p=3$ d'où $\varepsilon=2^{1-3}=1/4$, avec on a $p=4$ et donc $\varepsilon=2^{1-4}=1/8$.

2. Quelles sont les valeurs des exposants dont les représentations biaisées sur 4 bits sont :
0000 ; 0100 ; 1010 ?

Sur 4 bits le biais est de $2^{4-1} = 8$ qui est le chiffre que l'on a ajouté pour obtenir ces exposants
0000 est la notation en binaire naturel de 0 et donc la représentation de l'exposant 0-8=-8
0100 est la notation en binaire naturel de 4 et la représentation de l'exposant 4-8=-4
1010 est la notation en binaire naturel de 10 et la représentation de l'exposant 10-8=2

3. Quelles sont les représentations (en notation biaisée sur 8 bits) des exposants suivants : - 4 , -125 , 120 ?

Sur 8 bits, il faut donc ajouter 2^{8-1}=128 pour obtenir la représentation de l'exposant.
-4 : -4+128=124=127-2-1 $\rightarrow$0111 1100
-125 : 125+128=3=$\rightarrow$0000 0011
120 : 120+128=228=248=(256-1)-4-2-1$\rightarrow$1111 1000

4. Soit un ordinateur fictif à mots de 8 bits où les réels sont représentés de la façon normalisée, en utilisant la technique du bit caché.
Les 8 bits B1 | B2 B3 B4 | B5 B6 B7 B8 ont la signification suivante :
- B1 représente le signe de la mantisse, codé sur 1 bit ;
- B2 B3 B4 sont 3 bits qui codent l'exposant, en notation biaisée ;
- B5 B6 B7 B8 sont 4 bits qui codent la mantisse.
Quels sont les nombres réels que l'on peut représenter ?
Donner quelques réels et leur représentation.

Exposant sur 3 bits en notation biaisée : biais = 2^2. 8 possibilités de -4 à +3
Mantisse en bit caché. donc sont représentables les mantisses 1 B5 B6 B7 B8 soit 2^4 mantisses possibles.
Nombres négatifs : plus petit : 1 111 1111 = $((-0,11111)10^{111})_2$=$((-0,96875)*2^3)_{10}$=$(-7,75)_{10}$
plus grand : 1 000 0000=$((-0,1)*10^{000})_2$=$(-0,5*2^{-4})_{10}$=-0,03125
Nombres positifs : plus petit : 0 000 0000 par convention c'est zéro
plus grand : 0 111 1111 = $(0,11111*10^{111})_2 = (0,96875*2^3)_{10} = (7,75)_{10}$

5. Quel est le réel x ayant 0100 1111 pour représentation ?
Donner la représentation du réel immédiatement supérieur à x ainsi que sa valeur.

Le premier bit est le bit de signe : Il s'agit d'un réel positif
100 pour l'exposant biaisé (biais = 2^2) on enlève donc 4 pour avoir l'exposant réel. L'exposant réel est ici 0
1111 correspond à la mantisse en base 2 0,11111
Le nombre en base 2 est donc $((0,11111)*100)_2 = (2^{-1}+2^{-2}+2^{-3}+2^{-4}+2^{-5})*2^0 = (0,94625)_{10}$
Le nombre suivant dans la représentation est 01010000.

Exercice 3 : codage des réels en base 2 (cas réaliste)

Soit une machine où les réels sont représentés sur 32 bits (numérotés de droite à gauche de 0 à 31) avec les bits 0 à 22 qui contiennent la mantisse m normalisée, sans bit caché ; les bits 23 à 30 (8 bits) l'exposant biaisé et enfin le bit 31 qui est le bit de signe pour la mantisse.

1. Justifiez pourquoi on a $0,5 \leq m < 1$, en base 10.

Pour un exposant nul, la plus petite mantisse normalisée est $(0,10\ldots0)_2 = (0,5)_{10}$.
La plus grande est $(0,11\ldots1)_2 = (2^{-1}+\cdots 2^{-}-23)_{10} < 1$.

2. Donnez le réel qui correspond aux 32 bits ci-contre : 1 01111100 11010000000000000000000.

Bit de signe : 1 donc nombre négatif
mantisse : (0,1101)=$2^{-1}+2^{-2}+2^{-4}=(0,8125)_{10}$
Exposant biaisé : 01111100 donc exposant codé = $124-2^7$=124-128=-4
Nombre représenté : $-0,8125*2^{-4}=-0,05078125$

3. Quelle est la représentation du nombre réel $26,75$ dans cette machine ?

$(26,75)_{10}$=$(2^1+2^3+2^4+2^{-1}+2^{-2})_{10}$=$(11010,11)_2=0,1101011_2*10_2^{101}$
Mantisse : 1101011
Exposant biaisé = 5+2^7=5+128=133=$(10000101)_2$
Représentation : 0 10000101 11010110000000000000000

4. Donnez l'ordre de grandeur du plus petit nombre réel positif représentable dans cette machine et celui du plus grand nombre réel strictement positif représentable.

Plus petit :
exposant = 0000 0000, exposant codé $0-2^7=-128$.
avec la mantisse $0,100_2=0,5_{10}$ on a $1/2*2^{-128}=2^{-129}\sim 1,5*10^{-39}$ Plus grand :
Exposant=1111 1111, exposant codé = 255-$2^7=(127)_{10}$
la mantisse la plus grande est $0,11111111111111111111111=(\sum_{i=1}^{23}2^{-i}=1-2^{-23})_{10}$
d'où un nombre de l'ordre de $2^{127}\sim 1,7*10^{28}$

5. Pour l'exposant $+23$, soit un nombre (en base 10) compris entre $0,5*2^{23}$ et 2^{23} quelle sera la précision des nombres réels représentés (c'est-à-dire, l'écart entre deux nombres successifs d'exposant 23) ?

La mantisse est comprise sur les bits 0 à 22 :$2^{-1}+\cdots+2^{-23}$. Les nombres représentés avec l'exposant 23 sont donc des entiers, la précision est l'unité.

Exercice 4

On considère une machine qui code en base 2 les réels sur 16 bits (numérotés b_{15} à b_0 de gauche à droite) de la façon suivante : le bit de poids fort b_{15} est le bit de signe avec la convention habituelle, les bits b_{14} à b_{10} codent l'exposant en notation biaisée et les bits b_9 à b_0 codent la mantisse en utilisant la technique du bit caché.

1. Donnez la représentation dans cette machine de $(14,75)_{10}$ en expliquant comment vous obtenez l'exposant et la mantisse.

$(14,75)_{10}=2^3+2^2+2^1+2^{-1}+2^{-2}=(1110,11)_2$
Notation normalisée en base 2 : 0,111011*10^{100}
Mantisse (bit caché) : 1101100000
Exposant biaisé : Exposant + biais = 4+2^4 = 20 = $(10100)_2$
Représentation : 0 10100 11 0110 0000

2. Donnez en base 10 le réel codé par 1000 1011 1001 1000.

Signe du nombre : 1, donc réel négatif
Exposant biaisé : 00010 = exposant + biais d'où exposant en base 10 = 2-16=-14
Mantisse en bit caché = $(0,1\,11\,1001\,1000)_2$
D'où $(0,11110011000)_2 * (2^{-14})_{10} = (2^{-1}+2^{-2}+2^{-3}+2^{-4}+2^{-7}+2^{-8}) * 2^{-14}$
$= (0,000057935714722)_{10} = (0,57935714722 * 10^{-4})_{10}$
ce qui est au signe près l enimbre représenté.

Exercice 5 *[Codage des réels en machine, base 2]*
On considère une machine qui code, en base 2, les réels sur 10 bits : | b_9 | b_8 | b_7 | b_6 | b_5 | b_4 | b_3 | b_2 | b_1 | b_0 |
de la façon suivante : le bit de poids fort b_9 est le bit de signe avec la convention habituelle,
les bits b_8 à b_5 codent l'*exposant biaisée*
et les bits b_4 à b_0 codent la *mantisse normalisée* en utilisant la technique du *bit caché*.

1. Déterminez, en base 10, la plus petite et la plus grande valeur (réelle) représentées par la mantisse $b_4b_3b_2b_1b_0$.
   Réponse : *La mantisse normalisée, avec le bit caché, vaut* $m = (0,1b_4b_3b_2b_1b_0)_2$.
   Donc la plus petite valeur est $m_{min} = (0,100000)_2 = (\frac{1}{2})_{10}$ *;*
   la plus grande valeur est
   $m_{max} = (0,111111)_2 = (\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+\frac{1}{2^4}+\frac{1}{2^5}+\frac{1}{2^6})_{10} = (1-\frac{1}{2^6})_{10} = (\frac{2^6-1}{2^6})_{10} = (\frac{63}{64})_{10}$

   Que deviendraient ces valeurs si l'on avait utilisé une représentation normalisée sans la technique du bit caché ?
   Réponse : *Si l'on n'utilise pas le bit caché, mais la mantisse étant normalisée,*
   on a forcément $b_4 = 1$ *d'où* $m = (0,1b_3b_2b_1b_0)_2$.
   Dans ce cas la plus petite valeur est toujours $m_{min} = (0,10000)_2 = (\frac{1}{2})_{10}$ *;*
   tandis que la plus grande valeur est
   $m_{max} = (0,11111)_2 = (\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+\frac{1}{2^4}+\frac{1}{2^5})_{10} = (1-\frac{1}{2^5})_{10} = (\frac{2^5-1}{2^5})_{10} = (\frac{31}{32})_{10}$

2. Déterminer le biais de l'exposant. En déduire, en base 10, les valeurs que l'exposant peut prendre. Donner en particulier les valeurs des exposants codés par $b_8b_7b_6b_5 = 0000$, 1111 et 1100.
   Réponse : *Le biais sur 4 bits est de* $2^{4-1} = 8$.
   En binaire naturel, les 4 bits $b_8b_7b_6b_5$ *représentent les entiers de 0 à* $2^4 - 1 = 15$,
   en retranchant le biais on a donc l'ensemble d'exposants $\{-8,-7,\ldots,0,\ldots,7\}_{10}$.
   En particulier $(0000)_2 = 0_{10}$*, représente -8 ;* $(1111)_2 = 15_{10}$*, représente 7*
   et $(1100)_2 = 12_{10}$*, représente 4.*

3. Donner, en base 10, le plus petit réel strictement positif et le plus grand réel représentable dans cette machine.

   Réponse : *Le plus petit réel strictement positif représentable est donné par*
   $(0,100000)_2 * (2^{-8})_{10} = (2^{-9})_{10}$ *et le plus grand réel représentable est donné par*
   $(0,111111)_2 * (2^7)_{10} = (\frac{2^6-1}{2^6})_{10} * (2^7)_{10} = 63 * 2 = 126$.

4. Donnez la représentation dans cette machine de $x_1 = (12,75)_{10}$ et $x_2 = (-0,125)_{10}$. Expliquer comment vous obtenez l'exposant et la mantisse.

   Réponse : *On a* $(12,75)_{10} = (1100,11)_2 = (0,\underbrace{1}_{b.c.}10011 * 10^{100})_2$,
   la mantisse (normalisée, bit cachée ! !) est donc $10011 = b_4b_3b_2b_1b_0$

et l'exposant $(100)_2 = 4_{10}$ *dont le code biaisé est* $4+8 = 12_{10} = 1100_2 = b_7b_6b_5b_4$,
comme le nombre est positif $b_7 = 0$ *:*
Donc $x_1 = (12,75)_{10}$ *est représenté par* | *0* | *1* | *1* | *0* | *0* | *1* | *0* | *0* | *1* | *1* |

On a $(0,125)_{10} = (0,001)_2 = (0,\underbrace{1}_{b.c.}00000 * 10^{-10})_2$,
la mantisse (normalisée, bit cachée ! !) est donc $00000 = b_4b_3b_2b_1b_0$
et l'exposant $(-10)_2 = -2_{10}$ *dont le code biaisé est* $-2+8 = 6_{10} = 0110_2 = b_7b_6b_5b_4$,
comme le nombre est négatif $b_9 = 1$ *:*
Donc $x_2 = (-0,125)_{10}$ *est représenté par* | *1* | *0* | *1* | *1* | *0* | *0* | *0* | *0* | *0* | *0* |

5. On effectue la somme $s = x_1 + x_2$ dans cette machine. Quel résultat obtient-on ? Expliquez !

Réponse : *On a*

$$x_1 + x_2 = (12,750)_{10} - (0,125)_{10} = (12,625)_{10} = (1100,101)_2 = (0,\underbrace{1}_{b.c.}100101 * 10^{100})_2$$

On constate que la mantisse de s est trop grande pour être représentée, le dernier 1 sera perdu (cf. question de cours). On est en présence d'un exemple d'erreur d'arrondi.

Le résultat obtenu dans la machine sera a priori un arrondi vers le bas :
$\tilde{s} = (0,\underbrace{1}_{b.c.}10010 * 10^{100})_2 = (1100,10)_2 = (12,5)_{10}$.
Si l'on arrondit vers le haut, l'on obtient $\tilde{\tilde{s}} = (0,\underbrace{1}_{b.c.}10011 * 10^{100})_2 = (1100,11)_2 = (12,75)_{10} =$
x_1*, c'est à dire que l'on n'a pas exécuté la soustraction !*