SN4: séance 1

2018年1月13日 20:21

§1 Représentation approchée d'un nombre réel dans un ordinateur

Def Représentation décimal d'un entier naturel

Soit $b \in \mathbb{N}_{\geq 2}$ une base de numération. Tout entier $n \in \mathbb{N}$ s'écrit de façon unique sous la forme $n = a_0 + a_1 b + \cdots + a_k b^k$, où $a_i \in \{0, \cdots, b-1\}$, $k = \lfloor \log_b n \rfloor$ et $a_k = \lfloor n / b^k \rfloor > 0$. Le nombre n est stocké comme un vecteur $(a_0, \cdots, a_k)$ dans un ordinateur (typiquement avec $b=2$)

Remarque Étant donné une borne N sur le nombre de bits que l'on peut utiliser pour représenter un nombre, on peut stocker n'import quel nombre $n \in \mathbb{N}_{\leq 2^N}$ sans perte d'information. Cependant,

- Un ordinateur a une capacité finie
- Étant donné un intervalle $[a,b]$ avec $a<b$, il n'y a aucune façon systématique de stocker n'import quel nombre réel (ou rationnel) dans $[a,b]$ (car $\mathrm{card}(\mathbb{Q} \cap [a,b]) = \infty$)
- Il est nécessaire de représenter les nombre réels sans forme approchée.

Déf Représentation virgule flottante

Soit $N \in \mathbb{N}_{\geq 1}$ qui désigne le nombre de chiffre significatif de la representation. Si x est un nombre réel, $x \neq 0$, on peut le représenter approximativement comme

$$\tilde{x} := (-1)^s b^k \sum_{i=1}^{N} a_i b^{-i}, \qquad a_i \in \{0, \cdots, b-1\}$$

où k est appelé exposant, $(a_1, \cdots, a_N)$ est appelé la mantisse, et $s \in \{0,1\}$ est appelé signe, définis par les relations suivantes

$$s = \left(1 - \frac{x}{|x|}\right)/2, \quad k = \left\lfloor \frac{\ln |x|}{\ln(b)} \right\rfloor + 1, \quad a_N + a_{N-1} b + \cdots + a_1 b^{N-1} = \lfloor |x| b^{N-k} \rfloor .$$

(On a toujours $|\tilde{x}| \leq |x|$)

On appelle erreur de cette représentation approximative la différence $\Delta_N(x) := |\tilde{x} - x|$

On appelle erreur relative le quotient $\varepsilon_N(x) := \dfrac{\Delta_N(x)}{|x|}$

Si $x = 0$, par convention $\Delta_N(x)$ et $\varepsilon_N(x)$ sont définis comme 0.

S'il n'y pas d'ambiguité sur N, $\Delta_N(x)$ et $\varepsilon_N(x)$ sont notés $\Delta(x)$ et $\varepsilon(x)$

Prop.1 On a $|\varepsilon_N(x)| < b^{1-N}$

Preuve Sans perte de généralité, on suppose $x > 0$. On a $\tilde{x} = \lfloor x b^{N-k} \rfloor b^{k-N}$

Donc $\tilde{x} \leq x < \tilde{x} + b^{k-N}$. En outre, $x \geq b^{k-1}$. Donc $-b^{1-N} < \varepsilon_N(x) \leq 0$.

SN4: séance 1

2018年1月13日 20:21

§2 Phénomène de cumulation

Soient x et y deux nombres réels. On cherche à calculer $x+y$ numériquement par l'ordinateur. Comme x et y sont stocké comme $\tilde{x}$ et $\tilde{y}$ respectivement, l'ordinateur calcule $\tilde{x}+\tilde{y}$ puis stocke $\widetilde{\tilde{x}+\tilde{y}}$. On désigne par $x \tilde{+} y$ ce nombre.

⚠ La loi de composition $\tilde{+}$ (addition approximative) est commutative mais pas associative.

Def On désigne par $\Delta_+(x,y)$ (resp. $\varepsilon_+(x,y)$) l'erreur (resp. l'erreur relative) de $\tilde{+}$

$$\Delta_+(x,y) := |x \tilde{+} y - (x+y)|, \qquad \varepsilon_+(x,y) := \frac{\Delta_+(x,y)}{|x+y|}$$

Si $x+y=0$, par convention $\varepsilon_+(x,y) := 0$.

Prop.2 $$\varepsilon_+(x,y) \le \varepsilon(\tilde{x}+\tilde{y}) + \left|\frac{x}{x+y}\right| \varepsilon(x)(1+\varepsilon(\tilde{x}+\tilde{y})) + \left|\frac{y}{x+y}\right| \varepsilon(y)(1+\varepsilon(\tilde{x}+\tilde{y}))$$

En particulier, si x et y sont de même signe, alors on a

$$\varepsilon_+(x,y) < 2b^{1-N} + b^{2-2N}$$

Rem. Si $x=\tilde{x}$, $\Delta_+(x,y) \le \Delta(\tilde{x}+\tilde{y}) + \Delta(y)$

Si $x=\tilde{x}$ et $y=\tilde{y}$, $\Delta_+(x,y) = \Delta(x+y)$

Preuve On a $\Delta_+(x,y) = |\widetilde{\tilde{x}+\tilde{y}} - (\tilde{x}+\tilde{y}) + (\tilde{x}-x) + (\tilde{y}-y)|$

$$\le \Delta(\tilde{x}+\tilde{y}) + \Delta(x) + \Delta(y)$$

Donc $$\varepsilon_+(x,y) \le \frac{|\tilde{x}+\tilde{y}|}{|x+y|}\varepsilon(\tilde{x}+\tilde{y}) + \frac{|x|}{|x+y|}\varepsilon(x) + \frac{|y|}{|x+y|}\varepsilon(y)$$

$$\le \frac{|x+y| + |x|\varepsilon(x) + |y|\varepsilon(y)}{|x+y|}\varepsilon(\tilde{x}+\tilde{y}) + \frac{|x|}{|x+y|}\varepsilon(x) + \frac{|y|}{|x+y|}\varepsilon(y),$$

d'où la première inégalité.

Si x et y sont de même signe, alors $\left|\frac{x}{x+y}\right| + \left|\frac{y}{x+y}\right| = 1$.

D'après la proposition 1, on a $\max(\varepsilon(x), \varepsilon(y), \varepsilon(\tilde{x}+\tilde{y})) < b^{1-N}$. Donc

$$\varepsilon_+(x,y) \le b^{1-N} + b^{1-N}(1+b^{1-N}) = 2b^{1-N} + b^{2-2N}.$$ ∎

⚠ Si x et y ne sont pas de même signe et si $|x+y|$ est petit, alors l'erreur de calcul peut être très significatif

Il est donc d'éviter les sommations dans lesquelles des termes de signes opposés se compensent.

Déf Si x et y sont des nombres réels, on définit $x \tilde{\times} y := \widetilde{\tilde{x}\tilde{y}}$

C'est le produit approximatif de x et y donné par l'ordinateur.

On désigne par $\Delta_x(x,y)$ (resp. $\mathcal{E}_x(x,y)$) l'erreur (resp. l'erreur relative) de $\tilde{x}$

$$\Delta_x(x,y) := |x\tilde{x}y - x\times y|, \qquad \mathcal{E}_x(x,y) = \frac{\Delta_x(x,y)}{|xy|} \text{ si } xy\neq 0$$

Si $xy=0$, $\mathcal{E}_x(x,y)$ est définie comme 0 par convention.

Prop 3 $\mathcal{E}_x(x,y) \leq \mathcal{E}(\tilde{x}\tilde{y}) + \mathcal{E}(x) + \mathcal{E}(y) \leq 3\,b^{1-N}$

Rem Si $x=\tilde{x}$, alors $\Delta_x(x,y) \leq \Delta(x\tilde{y}) + |x|\,\Delta(y)$

Si $x=\tilde{x}$ et $y=\tilde{y}$, alors $\Delta_x(x,y) = \Delta(xy)$

Preuve
$$\begin{aligned}\Delta_x(x,y) &= |\tilde{x}\tilde{y} - \tilde{x}\tilde{y} + \tilde{x}\tilde{y} - x\tilde{y} + x\tilde{y} - xy| \\ &\leq \Delta(\tilde{x}\tilde{y}) + \Delta(x)|\tilde{y}| + |x|\Delta(y)\end{aligned}$$

Donc $\mathcal{E}_x(x,y) \leq \left|\frac{\tilde{x}\tilde{y}}{xy}\right| \mathcal{E}(\tilde{x}\tilde{y}) + \left|\frac{\tilde{y}}{y}\right| \mathcal{E}(x) + \mathcal{E}(y)$

Comme $|\tilde{x}| \leq |x|$ et $|\tilde{y}| \leq |y|$, on obtient la première inégalité.

La deuxième inégalité provient de la proposition 1 ※

§3 Règle de Hörner

On s'intéresse au calcul numérique d'un polynôme $P(x) = \sum_{k=0}^{n} a_k x^k$

Méthode naïve : $S_0 = \tilde{a}_0$, $x_0 = 1$ et par récurrence

$$x_k = x_{k-1} \tilde{\times} x, \quad \mu_k = a_k \tilde{\times} x_k, \quad S_k = S_{k-1} \tilde{+} \mu_k \qquad (k \in \{1,\dots,n\})$$

$$P(x) \approx S_n$$

Méthode de Hörner (par récurrence descendante) :

$$p_n = \tilde{a}_n$$

$$p_{k-1} = a_{k-1} \tilde{+} (x \tilde{\times} p_k) \qquad P(x) \approx p_0$$

Cette méthode est basée sur l'égalité

$$P(x) = a_0 + x\left(a_1 + x\left(a_2 + \cdots + x\left(a_{n-1} + x\,a_n\right)\cdots\right)\right)$$

Estimation d'erreur

Pour simplicité on suppose que $a_k = \tilde{a}_k$ et $x = \tilde{x}$. Soit $\delta = b^{1-N}$

- Méthode naïve :

$$|x_k - x_{k-1}x| = \Delta(x_{k-1}x) \leq |x|^k \delta \quad \rightsquigarrow \quad |x_k - x^k| \leq (k-1)|x|^k \delta$$

$$|\mu_k - a_k x_k| = \Delta(a_k x_k) \leq |a_k x^k| \delta$$

$$\rightsquigarrow \quad |\mu_k - a_k x^k| \leq |\mu_k - a_k x_k| + |a_k| \cdot |x_k - x^k| \leq k|a_k x^k| \delta$$

$$|S_k - (S_{k-1} + \mu_k)| = \Delta(S_{k-1} + \mu_k) \leq \delta\left(\sum_{i=0}^{k} |a_i x^i|\right)$$

SN4: séance 1

2018年1月13日 20:21

$$\Rightarrow |S_k - (u_0+\dots+u_k)| \le \delta \sum_{i=0}^{k} \sum_{j=0}^{i} |a_j x^j| = \delta \sum_{j=0}^{k} (k-j+1)|a_j x^j|$$

$$\Rightarrow |S_k - \sum_{j=0}^{k} a_j x^j| \le \delta \sum_{j=0}^{k} (k-j+1)|a_j x^j| + \delta \sum_{j=0}^{k} j|a_j x^j| = \delta(k+1) \sum_{j=0}^{k} |a_j x^j|$$

- Méthode de Hörner. Soient $q_n = a_n$ et $q_{k-1} = a_{k-1} + x q_k$

$$|x\tilde{x} p_k - x p_k| \le \delta |x p_k|$$

$$|p_{k-1} - (a_{k-1} + x\tilde{x} p_k)| \le (|a_{k-1}| + |x p_k|)\delta$$

Donc $|p_{k-1} - q_{k-1}| \le |x| \cdot |p_k - q_k| + (|a_{k-1}| + 2|x p_k|)\delta$

$$\Rightarrow |p_0 - q_0| \le (|a_0| + 2|x p_1|)\delta + |x|\left((|a_1| + 2|x p_2|)\delta + |x|(\dots)\right)$$

$$= \delta \sum_{i=0}^{n} |a_i x^i| + 2\delta \sum_{i=1}^{n} |x^i p_i|$$

$$\le \delta \sum_{i=0}^{n} |a_i x^i| + 2\delta \sum_{i=1}^{n} \sum_{j=i}^{n} |a_j x^j|$$

$$= \delta \sum_{j=0}^{n} |a_j x^j| + 2\delta \sum_{j=1}^{n} j|a_j x^j| \le \delta \sum_{j=0}^{n} (2j+1)|a_j x^j|.$$