Chapitre 1 : Analyse d'erreurs et arithmétique de l'ordinateur

Ibrahima Dione (Université Laval)

24 janvier 2017

1. André Fortin: *Analyse numérique pour ingénieurs*, Presses Internationales Polytechnique, Montréal, Québec (2016).

Erreur absolue et erreur relative

Définition 1.1

Soit x un nombre et x^* son approximation. L'erreur absolue, notée $E_a(x^*)$, est définie par

$$E_a(x^*) = |x - x^*|. \tag{1.1}$$

Définition 1.2

Soit x un nombre et x^* son approximation. L'erreur relative, notée $E_r(x^*)$, est définie par

$$E_r(x^*) = \frac{|x - x^*|}{|x|} = \frac{E_a(x^*)}{|x|}, \ x \neq 0. \tag{1.2}$$

Remarque(s) :

- L'erreur absolue $E_a(x^*)$ donne une *mesure quantitative* de l'erreur commise en estimant la valeur exacte x, à partir de x^*.
- L'erreur relative $E_r(x^*)$ mesure l'importance (une mesure qualitative) de l'erreur commise en estimant x, à partir de x^*. Elle est souvent exprimée en pourcentage en la multipliant par 100, pour la rendre plus significative.

En pratique, on ne connait que la valeur approximée x^* et une borne supérieure de l'erreur absolue, qu'on note Δx, qui est quand même considérée comme étant l'erreur absolue

$$\underbrace{|x - x^*|}_{E_a(x^*)} \leq \Delta x.$$

Note: On a les équivalences suivantes :

$$|x - x^*| \leq \Delta x \iff -\Delta x \leq x - x^* \leq \Delta x \iff -\Delta x + x^* \leq x \leq \Delta x + x^*.$$

On note parfois $x = x^* \pm \Delta x$ et on interpréte ce résultat en disant que l'on a estimé la valeur exacte x à partir de x^* avec une incertitude de Δx de part et d'autre.

L'erreur relative $E_r(x^*)$ est ainsi remplacée par le rapport $\frac{\Delta x}{|x^*|}$: $E_r(x^*) \simeq \frac{\Delta x}{|x^*|}$.

Définition 1.3

Si l'erreur absolue vérifie $\Delta x \leq 0,5 \times 10^m$, alors le chiffre dans x^* correspondant à la m^e puissance de 10 est dit *significatif* et tous ceux à sa gauche, correspondant aux puissances de 10 supérieures à m, le sont aussi. On arrête le compte au dernier chiffre non nul.

Note: Il existe une exception à la règle. Si le chiffre correspondant à la m^e puissance de 10 est nul ainsi que tous ceux à sa gauche, on dit qu'il n'y a aucun chiffre significatif.

Définition 1.4

Inversement, si une approximation x^* est donnée avec un nombre n de chiffres significatifs, on commence à compter à partir du premier chiffre non nul à gauche et l'erreur absolue est inférieure à $0,5$ fois la puissance de 10 correspondant au dernier chiffre significatif.

Remarque(s) : En pratique, on cherchera à déterminer une borne pour Δx aussi petite que possible et donc la valeur de m la plus petite possible.

Exemple 1

On approxime le nombre $x = \pi$ par la quantité $x^* = \frac{22}{7} = 3,142857\cdots$.

- erreur absolue : $E_a(x^*) = |\pi - \frac{22}{7}| = 0,001264489\cdots \simeq 0,1264489 \times 10^{-2}$,
- erreur relative : $E_r(x^*) = \frac{|\pi - \frac{22}{7}|}{|\pi|} = 0,000402499\cdots \simeq 0,402499 \times 10^{-3}$.

Voir l'exemple suivant, pour la suite de celui-ci.

Exemple 2

À l'Exemple précédent, on a l'erreur absolue dans l'approximation de la valeur exacte $x = \pi$ par $x^* = \frac{22}{7} = 3,142857\cdots$ qui est donnée par

$$E_a(x^*) = |\pi - \frac{22}{7}| = 0,001264489\cdots \simeq 0,1264489 \times 10^{-2}.$$

Puisque l'erreur absolue est plus petite que $0,5 \times 10^{-2}$, alors le chiffre des centièmes est significatif et on a en tout 3 chiffres significatifs (en l'occurrence $3,14$).
Si l'on retient que $x^* = 3,1416$ comme approximation de $x = \pi$, alors on a l'erreur absolue

$$E_a(x^*) = |\pi - 3,1416| \simeq 0,73 \times 10^{-5}.$$

Ainsi, on a l'estimation suivante de l'erreur absolue

$$E_a(x^*) \simeq 0,73 \times 10^{-5} = 0,073 \times 10^{-4} \leq 0,5 \times 10^{-4}.$$

Le chiffre (sur $x^* = 3,1416$) correspondant à la $4^{ième}$ position après la virgule (ici 6), est significatif ainsi que tous les chiffres situés à sa gauche : $x^* = 3,1416$ a 5 chiffres significatifs.

Note: Il est à noter que le chiffre 6 dans $x^* = 3,1416$ est significatif même si la quatrième décimale de π est un 5 ($\pi = 3,14159\cdots$). Significatif ne veut pas dire exact, mais dont on contrôle l'erreur, qui ont du sens.

Représentation des entiers positifs en binaire

Pour transformer un entier positif N dans sa représentation binaire habituelle, il faut déterminer les a_i tels que :

$$(N)_{10} = (a_{n-1}a_{n-2}a_{n-3}\cdots a_2 a_1 a_0)_2$$

où l'indice inférieur indique la base utilisée. Une autre représentation est

$$N = a_{n-1}\times 2^{n-1} + a_{n-2}\times 2^{n-2} + a_{n-3}\times 2^{n-3} + \cdots + a_2\times 2^2 + a_1\times 2^1 + a_0\times 2^0$$

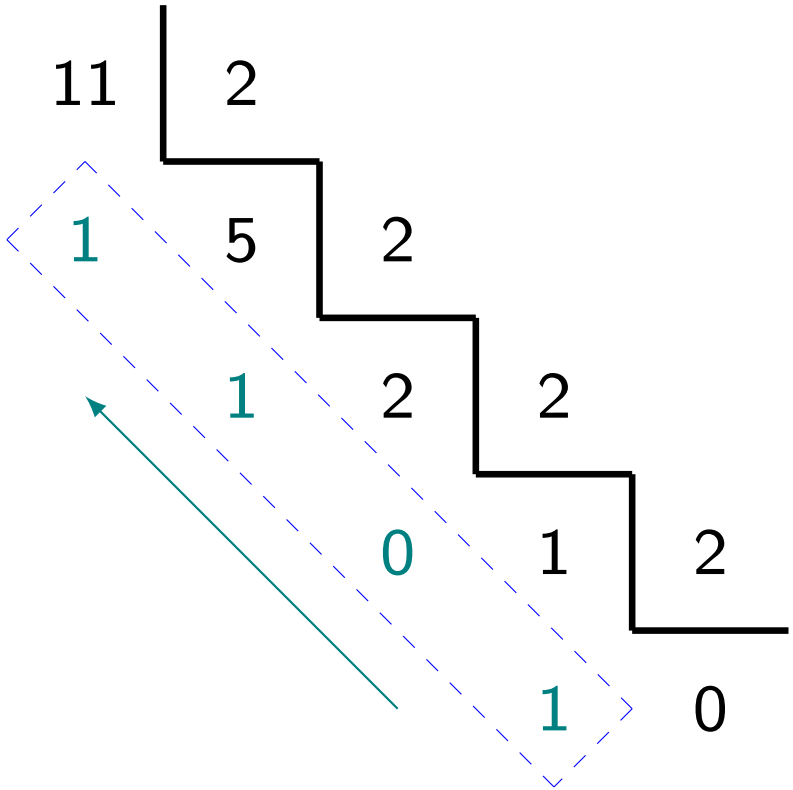
Question: Comment déterminer les bits a_i permettant la représentation binaire de l'entier N ?

Méthode : On obtient la valeur des a_i par la démarche suivante :

- on divise N par 2 pour obtenir a_0 (le reste de la division) plus un entier ;
- on refait le même raisonnement avec la partie entière de $\frac{N}{2}$ (en négligeant la partie fractionnaire ou reste) pour obtenir a_1 ;
- on continue ainsi jusqu'à ce que la partie entière soit nulle.

Exemple 3

Pour convertir le nombre décimal $N = (11)_{10}$ en binaire, on procède comme suit :



Ainsi, l'entier décimal 11 s'écrit 1011 en binaire ($(11)_{10} = (1011)_2$).

Conversion d'une fraction décimale en valeur binaire

Si f est une fraction décimale comprise entre 0 et 1, sa conversion en binaire consiste à trouver d_i

$$(f)_{10} = (0, d_1 d_2 d_3 \cdots)_2 = d_1 \times 2^{-1} + d_2 \times 2^{-2} + d_3 \times 2^{-3} + \cdots$$

Note: La suite $d_1 d_2 d_3 \cdots$ sera appelée la mantisse et sera détaillée à la Sous-section 4.

Méthode : Pour obtenir les bits d_i, on procède comme suit :

- on multiplie f par 2, pour obtenir d_1 plus une fraction ;
- on applique le même raisonnement à $(2f - d_1)$, on obtient d_2 plus une fraction ;
- on poursuit ainsi jusqu'à ce que la partie fractionnaire soit nulle ou que l'on ait atteint le nombre maximal de chiffres prévu à la mantisse.

Remarque(s) : Un ordinateur ou une calculatrice ne peuvent stocker qu'un nombre fini de bits pour représenter un nombre. Un recours est fait à la troncature ou à l'arrondi :

- Si l'on travaille en notation décimale et si l'on souhaite utiliser la troncature avec 4 chiffres dans la mantisse (n dans le cas général), on coupe les décimales restantes à partir de la cinquième ($(n+1)^e$).
- Pour arrondir, on ajoute 5 unités au cinquième ($(n+1)^e$) chiffre de la mantisse et l'on tronque le résultat. Le même procédé s'applique en binaire (ou base 2).

Exemple 4

Pour convertir la fraction décimale $f = 0,0625$, on procède comme suit :

$$\begin{aligned} 0,0625 \times 2 = 0,1250 &\longrightarrow d_1 = 0 \\ 0,1250 \times 2 = 0,2500 &\longrightarrow d_2 = 0 \\ 0,2500 \times 2 = 0,5000 &\longrightarrow d_3 = 0 \\ 0,5000 \times 2 = 1,0000 &\longrightarrow d_4 = 1 \end{aligned}$$

signifiant ainsi $(0,0625)_{10} = (0,0001)_2$.
Pour la fraction décimale $f = \frac{1}{3}$, on procède de la même façon :

$$\begin{aligned} \frac{1}{3} \times 2 = \frac{2}{3} + 0 &\longrightarrow d_1 = 0 \\ \frac{2}{3} \times 2 = \frac{1}{3} + 1 &\longrightarrow d_2 = 1 \\ \frac{1}{3} \times 2 = \frac{2}{3} + 0 &\longrightarrow d_3 = 0 \\ \frac{2}{3} \times 2 = \frac{1}{3} + 1 &\longrightarrow d_4 = 1 \end{aligned}$$

On peut poursuivre la conversion à l'infini et obtenir $\left(\frac{1}{3}\right)_{10} = (0,010101\cdots)_2$

Représentation en complément à 2 des entiers signés

Si l'on dispose de n bits à la mantisse pour exprimer l'entier N en binaire , alors celui-ci s'écrirait

$$N = -a_{n-1} \times 2^{n-1} + a_{n-2} \times 2^{n-2} + a_{n-3} \times 2^{n-3} + \cdots + a_2 \times 2^2 + a_1 \times 2^1 + a_0 \times 2^0,$$

où il est important de remarquer le signe négatif devant le terme a_{n-1}.

Méthode :

- Un entier positif est représenté par 0 suivi de son expression binaire en $(n-1)$ bits.
- Pour obtenir la représentation d'un nombre négatif N, il suffit de lui ajouter 2^{n-1} et de transformer le résultat en forme binaire. Le résultat final est donc 1 suivi de la représentation sur $(n-1)$ bits de $N + 2^{n-1}$.

Exemple 5

La représentation en complément à 2 sur 4 bits de 0101 vaut :

$$-0 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0.$$

En forme décimale, cela correspond au nombre 5. Suite de l'exemple (voir notes de cours).

Représentation par excès des entiers signés

Méthode :

- Si l'on veut exprimer un entier décimal N avec un excès d, il suffit de lui ajouter l'excès et de représenter le résultat sous forme binaire.
- Inversement, si l'on a la représentation binaire par excès d'un entier, il suffit de calculer sa valeur en base 10 et de soustraire l'excès d pour obtenir l'entier recherché.

Note: En général, la valeur de l'excès d est 2^{n-1} ou $2^{n-1}-1$ (soit 2^3 ou 2^3-1 pour $n=4$), où n est le nombre de bits à la mantisse avec lequel la représentation est faite.

Exemple 6

Soit un mot de 8 bits et un excès $d = 2^7 - 1 = 127$. Pour représenter $(-100)_{10}$, il suffit de lui ajouter 127, ce qui donne 27, et d'exprimer le résultat sur 8 bits, soit 00011011.

Représentation des nombres réels

Définition 2.1

Soit x un nombre réel. On appelle représentation en notation flottante ou notation en point flottant ou encore en virgule flottante du nombre x, s'il est écrit sous la forme

$$x = \pm m \times b^{\ell}, \tag{2.1}$$

où ℓ est l'exposant, b est la base et m est la mantisse dont la forme générale est

$$m = 0, d_1 d_2 d_3 \cdots d_n \cdots = d_1 \times b^{-1} + d_2 \times b^{-2} + d_3 \times b^{-3} + \cdots + d_n \times b^{-n} + \cdots$$

où les d_i vérifient

$$1 \leq d_1 \leq (b-1) \quad \text{et} \quad 0 \leq d_i \leq (b-1), \text{ pour } i = 2, 3, \cdots, n, \cdots .$$

L'inégalité $d_1 \geq 1$ signifie que la mantisse est normalisée et garantit l'unicité de la représentation (2.1).

Note: La mantisse satisfait $\frac{1}{b} \leq m < 1$ et la représentation de 0 devient une exception puisque la mantisse ne s'annule jamais. Dans le système décimal, la représentation (2.1) devient $x = \pm m \times 10^{\ell}$.

Exemple 7

Pour une mantisse de longueur infinie, la troncature et l'arrondi sont également envisagés :

- Troncature à n chiffres : on coupe la mantisse m au-delà de la position n

$$m = d_1 \times b^{-1} + d_2 \times b^{-2} + d_3 \times b^{-3} + \cdots + d_n \times b^{-n}.$$

- Arrondi à n chiffres : On ajoute un 5 (en décimal) ou un 1 (en binaire) à la position $n+1$, ensuite on fait une troncature à n chiffres

$$m = d_1 \times b^{-1} + d_2 \times b^{-2} + d_3 \times b^{-3} + \cdots + \tilde{d}_n \times b^{-n}.$$

Pour représenter $10,2$ en notation flottante à 4 chiffres dans la mantisse et en base $b = 10$, on a $0,1020 \times 10^2$ au lieu de $0,0102 \times 10^3$, grâce à la normalisation.

Remarque(s) : Le choix de la base b est arbitraire, mais les préférences sont :

1. Humains : usuellement décimale ($b = 10$).
2. Ordinateurs : usuellement binaire ($b = 2$) ou hexadécimale ($b = 16$).

Note: Pour la représentation d'un nombre réel sur ordinateur, la base sera généralement $b = 2$ et elle consistera à représenter la mantisse (une fraction), l'exposant (un entier signé) et le signe de ce nombre.

Erreurs dues à la représentation

Utiliser un nombre limité de bits pour représenter un nombre réel a des conséquences :

- On introduit une erreur de représentation qui peut avoir des répercussions significatives sur la précision des résultats.
- Quel que soit le nombre de bits utilisés, il existe un plus petit et un plus grand nombres positifs représentables. Ainsi, on doit recourir à la troncature ou à l'arrondi pour représenter les autres réels en déhors de cet intervalle fini de nombres.

Définition 2.2

La précision machine ε_m est la plus grande erreur relative que l'on puisse commettre en représentant un nombre réel sur ordinateur en utilisant la troncature.

Note: Si on utilise l'arrondi, la précision machine est tout simplement $\frac{\varepsilon_m}{2}$.

Théorème 2.1

Si b est la base utilisée pour réprésenter un nombre avec n nombre de chiffres (bits si $b = 2$) de la mantisse, alors la précision machine vérifie $\varepsilon_m \leq b^{1-n}$.

Arithmétique flottante

Définition 3.1

Soit x, un nombre réel défini sous la forme $x = \pm 0, d_1 d_2 d_3 \cdots d_n d_{n+1} \cdots \times 10^{\ell}$.
On note $fl(x)$, sa représentation en virgule flottante à n chiffres définie par

$$fl(x) = \pm 0, d_1 d_2 d_3 \cdots \tilde{d}_n \times 10^{\ell},$$

où la $n^{ième}$ décimale, $\tilde{d}_n$, dépendra de la méthode de remplissage (troncature/arrondi).

Note: La notation flottante d'un nombre dépend du nombre n de chiffres dans la mantisse, mais aussi du procédé retenu pour éliminer les derniers chiffres à savoir la troncature (qui est biaisée car si $x \geq 0$, on a $fl(x) \leq x$) ou l'arrondi (qui est non biaisé, on a tour à tour $fl(x) \leq x$ ou $fl(x) \geq x$).

La convention IEEE-754 impose l'utilisation de l'arrondi dans la représentation binaire des nombres. Dans les exemples à venir, nous utiliserons l'arrondi.

Exemple 8

Avec $n = 4$ chiffres à la mantisse, on a la représentation en notation flottante de
$x = \frac{1}{3} \longrightarrow fl(\frac{1}{3}) = 0,3333 \times 10^0$ et $x = 12,4551 \longrightarrow fl(12,4551) = 0,1246 \times 10^2$.

Opérations élémentaires

Méthode : Pour effectuer une opération élémentaire (l'addition, la soustraction, la multiplication et la division) en arithmétique flottante, on doit représenter les deux opérandes en notation flottante, effectuer l'opération de la façon habituelle et exprimer le résultat en notation flottante.

$$x + y \longrightarrow fl\Big(fl(x) + fl(y)\Big), \qquad x - y \longrightarrow fl\Big(fl(x) - fl(y)\Big)$$
$$x \div y \longrightarrow fl\Big(fl(x) \div fl(y)\Big), \qquad x \times y \longrightarrow fl\Big(fl(x) \times fl(y)\Big)$$

Exemple 9

Si on a $n = 4$ chiffres à la mantisse, alors on a l'opération en arithmétique flottante

$$\frac{1}{3} \times 3 \longrightarrow fl\Big(fl(\frac{1}{3}) \times fl(3)\Big) = fl\Big((0,3333 \times 10^0) \times (0,3000 \times 10^1)\Big) = 0,9999 \times 10^0$$

Note: On remarque une légère perte de précision par rapport à la valeur exacte 1.

Exemple 10

- Si on a $n = 4$ chiffres à la mantisse, l'opération en arithmétique flottante est

$$\begin{aligned}(0,4035 \times 10^6) \times (0,1978 \times 10^{-1}) &\longrightarrow fl\Big(fl(0,4035 \times 10^6) \times fl(0,1978 \times 10^{-1})\Big) \\ &= fl\Big((0,4035 \times 10^6) \times (0,1978 \times 10^{-1})\Big) \\ &= fl(0,0798123 \times 10^5) \\ &= fl(0,798123 \times 10^4) = 0,7981 \times 10^4\end{aligned}$$

- Toujours avec $n = 4$ chiffres à la mantisse, on effectue l'opération suivante

$$\begin{aligned}(0,4035 \times 10^6) + (0,1978 \times 10^4) &\longrightarrow fl\Big(fl(0,4035 \times 10^6) + fl(0,1978 \times 10^4)\Big) \\ &= fl\Big(0,4035 \times 10^6 + 0,1978 \times 10^4\Big) \\ &= fl(0,4035 \times 10^6 + 0,001978 \times 10^6) \\ &= fl(0,405478 \times 10^6) = 0,4055 \times 10^6\end{aligned}$$

Note: Il est primordial de décaler la mantisse avant d'effectuer l'addition ou la soustraction.

Opérations risquées

Exemple 11

- En arithmétique flottante $n = 4$, on obtient

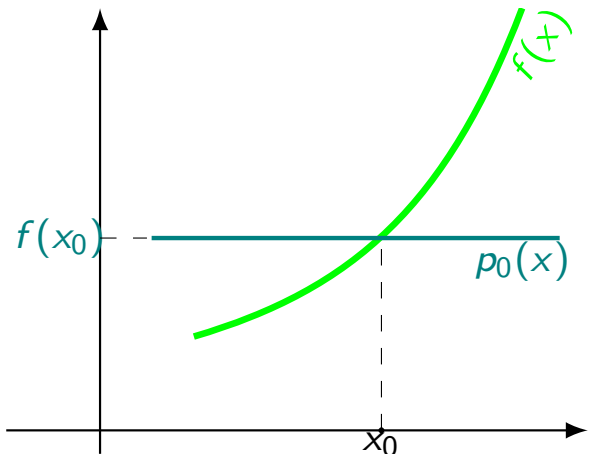
$$\begin{aligned}(0,4000 \times 10^4) + (0,1000 \times 10^{-2}) &\longrightarrow fl\Big(fl(0,4000 \times 10^4) + fl(0,1000 \times 10^{-2})\Big)\\ &= fl\Big(0,4000 \times 10^4 + 0,1000 \times 10^{-2}\Big)\\ &= fl(0,4000 \times 10^4 + 0,0000001 \times 10^4)\\ &= fl(0,4000001 \times 10^4) = 0,4000 \times 10^4\end{aligned}$$

Note: Additionner deux nombres dont les ordres de grandeur sont très différents, est une opération dangereuse ! Le petit nombre disparaît devant le plus grand.

- En arithmétique flottante, soustraire deux nombres presque identiques est dangereux

$$\begin{aligned}(0,5678 \times 10^6) - (0,5677 \times 10^6) &\longrightarrow fl\Big(fl(0,5678 \times 10^6) - fl(0,5677 \times 10^6)\Big)\\ &= fl\Big(0,5678 \times 10^6 - 0,5677 \times 10^6\Big)\\ &= fl(0,0001 \times 10^6) = 0,1000 \times 10^3\end{aligned}$$

Développement de Taylor en une variable

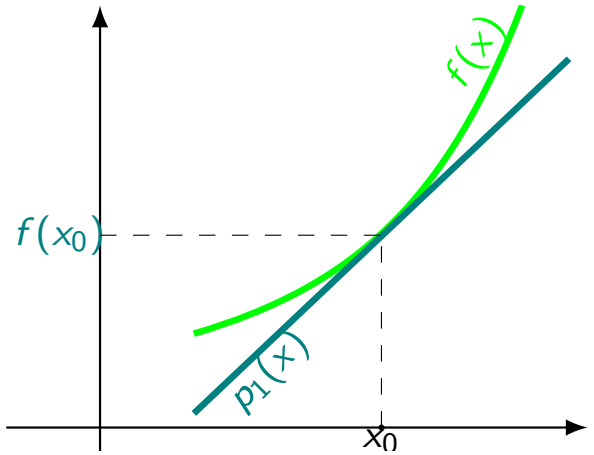


Si $f(x)$ une fonction dont on ne connait que son évaluation au point x_0.

Alors la meilleure approximation polynômiale de $f(x)$, autour de ce point x_0, sera une fonction constante.

C'est à dire un polynôme de degré zero, qu'on va noter p_0 et définir au voisinage de x_0 par

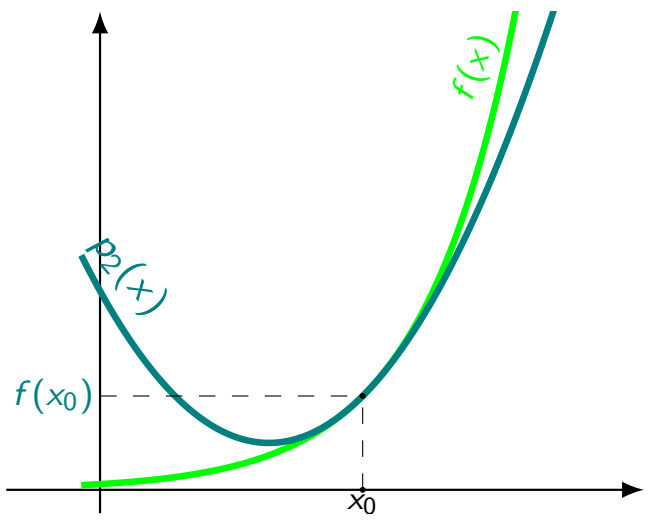
$$p_0(x) = f(x_0).$$



Si on connait la valeur de la fonction $f(x)$ en x_0 ainsi que la pente $f'(x_0)$ en x_0 de cette même fonction.

La meilleur approximation polynômiale de la fonction $f(x)$, au voisinage du point x_0, est un polynôme de degré 1 qui est définie comme suit

$$p_1(x) = f(x_0) + f'(x_0)(x - x_0).$$



Et si connait de la fonction $f(x)$ sa valeur $f(x_0)$, sa pente $f'(x_0)$ et sa dérivée seconde $f''(x_0)$.

Alors on a un polynôme de degré 2, défini par

$$p_2(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)\frac{(x - x_0)^2}{2}$$

qui est la meilleur approximation polynômiale de la fonction $f(x)$ au voisinage du point x_0.

Note: Autant de fois que $f(x)$ sera dérivable en x_0, le polynôme construit (appelé polynôme de Taylor) donnera une approximation au voisinage de x_0 plus précise.

Définition 4.1

Le polynôme de Taylor de degré n de la fonction $f(x)$ autour de x_0 est défini par

$$p_n(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)\frac{(x - x_0)^2}{2!} + f'''(x_0)\frac{(x - x_0)^3}{3!} + \cdots + f^{(n)}(x_0)\frac{(x - x_0)^n}{n!}, \quad (4.1)$$

où $f^{(n)}(x_0)$ désigne la dérivée d'ordre n de $f(x)$ en x_0.

Théorème 4.1

Soit $f(x)$, une fonction dont les dérivées jusqu'à l'ordre $(n+1)$ existent au voisinage du point x_0. On a l'égalité

$$f(x) = p_n(x) + r_n(x), \tag{4.2}$$

où $p_n(x)$ est le polynôme de Taylor défini en (4.1) et $r_n(x)$ est l'erreur commise en approximant $f(x)$ par $p_n(x)$

$$r_n(x) = f^{(n+1)}(\zeta(x))\frac{(x-x_0)^{n+1}}{(n+1)!}, \tag{4.3}$$

pour un certain $\zeta(x)$ compris entre x_0 et x.

Note: L'erreur $r_n(x)$ est appelée l'erreur de troncature, commise chaque fois que l'on utilise un développement de Taylor. Il n'est possible d'évaluer ce terme d'erreur (car on ne connait pas $\zeta(x)$) mais on déterminera sa borne supérieure.

- L'équation (4.2) est une égalité et ne devient une approximation que lorsque le terme d'erreur est négligé.
- Le terme d'erreur (4.3) devient de plus en plus grand lorsque x s'éloigne de x_0.

- Inversement, pour une valeur de x près de x_0, le terme d'erreur de l'équation (4.3) est de plus en plus petit lorsque n augmente.
- On sait que le point $\zeta(x)$ existe et qu'il varie avec x, mais on ne connaît pas sa valeur exacte. Il n'est donc pas possible d'évaluer le terme d'erreur exactement. On peut tout au plus lui trouver une borne supérieure dans la plupart des cas.

Une forme plus pratique du développement de Taylor est obtenue en posant $h = x - x_0$ ou $x = x_0 + h$:

$$f(x_0 + h) = p_n(h) + r_n(h), \tag{4.4}$$

où le polynôme de Taylor devient

$$p_n(h) = f(x_0) + f'(x_0)h + f''(x_0)\frac{h^2}{2!} + f'''(x_0)\frac{h^3}{3!} + \cdots + f^{(n)}(x_0)\frac{h^n}{n!}, \tag{4.5}$$

et l'erreur de troncature donnée sous la forme

$$r_n(h) = f^{(n+1)}(\zeta(h))\frac{h^{n+1}}{(n+1)!}, \tag{4.6}$$

avec $\zeta(h)$ compris entre x_0 et $x_0 + h$.

Exemple 12

Déterminons une approximation de la fonction $f(x) = e^x$ autour de $x_0 = 0$. Remarquons d'abord que la dérivable à tout ordre n et en tout x de cette fonction est $f^n(x) = e^x$. En particulier en $x_0 = 0$, sachant que $f^{(n)}(0) = 1$, on a l'approximation suivante

$$e^{0+h} = e^h \simeq p_n(h) = 1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + \cdots + \frac{h^n}{n!}, \tag{4.7}$$

où l'expression du terme d'erreur, grâce à l'équation (4.6), est donnée par

$$r_n(h) = e^{\zeta(h)} \frac{h^{n+1}}{(n+1)!},$$

et où $\zeta(h) \in [0, h]$. Il n'est pas possible d'évaluer le terme d'erreur mais on peut déterminer sa borne supérieure. La fonction exponentielle étant croissante et puisque $\zeta(h) \leq h$, alors on a

$$e^{\zeta(h)} \leq e^h.$$

On en conclut la borne suivante de l'erreur $r_n(h)$ sur l'intervalle $[0, h]$

$$r_n(h) = e^{\zeta(h)} \frac{h^{n+1}}{(n+1)!} \leq e^h \frac{h^{n+1}}{(n+1)!}. \tag{4.8}$$

Définition 4.2

Une fonction $E(h)$ est dite grand ordre de h^n au voisinage de 0, s'il existe une constante c telle que

$$E(h) \leq ch^n, \text{ au voisinage de 0 (c'est à dire si } h \longrightarrow 0).$$

Note: On notera la fonction $E(h)$ comme suit $E(h) = \mathcal{O}\left(h^n\right)$.

- Lorsque h est assez petit, la fonction $\mathcal{O}\left(h^n\right)$ décroît suivant ch^n.
- Plus n est grand, plus la décroissance est rapide. Ainsi, une fonction $\mathcal{O}\left(h^3\right)$ décroît plus vite qu'une fonction $\mathcal{O}\left(h^2\right)$, qui elle-même décroît plus vite qu'une fonction $\mathcal{O}\left(h\right)$.
- Une fonction $\mathcal{O}\left(h^n\right)$ est aussi $\mathcal{O}\left(h^m\right)$ si $m < n$.

Question: Comment vérifier (ou déterminer) qu'une fonction $E(h)$ est un grand ordre de h^n au voisinage de 0 ? Il suffit que le rapport de $E(h)$ sur $E(\frac{h}{2})$ vérifie

$$\frac{E(h)}{E(\frac{h}{2})} \approx 2^n. \tag{4.9}$$

Exemple 13

Utilisons le développement de la relation (4.7), pour estimer la valeur de $e^{0,1}$ en prenant $h = 0,1$. À partir de l'égalité (4.4), on a l'expression de l'erreur $r_n(h) = f(x_0 + h) - p_n(h)$.

- Pour $h = 0,1$ et $n = 3$, on a l'erreur en valeur absolue au voisinage de $x_0 = 0$ est

$$|r_3(0,1)| = |f(0,1) - p_3(0,1)| = 0,4245 \times 10^{-5}.$$

- Et pour $h = 0,05$ et $n = 3$, l'erreur absolue au voisinage de $x_0 = 0$ est

$$|r_3(0,05)| = |f(0,05) - p_3(0,05)| = 0,263 \times 10^{-6}.$$

- Le rapport des erreurs absolues liées au polynôme d'approximation $p_3(h)$ est donné par

$$\frac{|r_3(0,1)|}{|r_3(0,05)|} = \frac{0,4245 \times 10^{-5}}{0,263 \times 10^{-6}} = 16,14 \approx 2^4.$$

Note: La valeur de ce rapport n'est pas fortuite. En effet, face à un polynôme de Taylor de degré $n = 3$, on obtain à partir de l'inégalité (4.8), un terme d'erreur qui est un grand ordre de h^4 au voisinage de 0 (i.e. une erreur de type $\mathcal{O}\left(h^4\right)$). Et grâce à la relation (4.9), ce rapport se comporte comme 2^4.

Définition 4.3

Une approximation dont le terme d'erreur est un grand ordre de h^n (noté $\mathcal{O}\left(h^n\right)$) est dite d'ordre n.

Remarque(s) : En général le polynôme de Taylor de degré n est une approximation d'ordre $n+1$ car le terme d'erreur est $r_n(h) = \mathcal{O}\left(h^{n+1}\right)$ (ou $r_n(x) = \mathcal{O}\left((x-x_0)^{n+1}\right)$) :

$$f(x_0+h) = p_n(h) + \mathcal{O}\left(h^{n+1}\right) \text{ ou } f(x) = p_n(x) + \mathcal{O}\left((x-x_0)^{n+1}\right). \tag{4.10}$$

Note: Mais il s'agit d'un ordre minimal, on peut être d'ordre plus élevé : calculer le développement de Taylor d'ordre 5 de $\sin(x)$ autour de $x_0 = 0$.

Note: Ne pas confondre le degré du polynôme de Taylor et son ordre d'approximation. Le degré du polynôme de Taylor est la puissance la plus grande de ce polynôme, alors que l'ordre d'approximation (qui donne la qualité de l'approximation) correspond à la puissance de h sur le terme d'erreur.

Développement de Taylor en deux variables

Théorème 4.2

Soit $f(x, y)$ une fonction de deux variables, à valeurs réelles que l'on suppose suffisamment différentiable. Le développement de Taylor d'ordre 2 de $f(x, y)$ au point (x_0, y_0) est

$$f(x_0 + h_1, y_0 + h_2) = f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)h_1 + \frac{\partial f}{\partial y}(x_0, y_0)h_2 + r_1(h_1, h_2), \qquad (4.11)$$

où le terme d'erreur est donné par

$$r_1(h_1, h_2) = \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(\zeta, \eta)h_1^2 + \frac{\partial^2 f}{\partial x \partial y}(\zeta, \eta)h_1 h_2 + \frac{1}{2}\frac{\partial^2 f}{\partial y^2}(\zeta, \eta)h_2^2. \qquad (4.12)$$

pour un certain point (ζ, η) sur le segment droit délimité par (x_0, y_0) et $(x_0 + h_1, y_0 + h_2)$.

Note: Si $f(x, y, z)$ est à 3 variables, son développement de Taylor d'ordre 2 au point (x_0, y_0, z_0) est

$$f(x_0 + h_1, y_0 + h_2, z_0 + h_3) = f(x_0, y_0, z_0) + \frac{\partial f}{\partial x}(x_0, y_0, z_0)h_1 + \frac{\partial f}{\partial y}(x_0, y_0, z_0)h_2 + \frac{\partial f}{\partial z}(x_0, y_0, z_0)h_3 + \mathcal{O}\left(h_1^2\right) + \mathcal{O}\left(h_2^2\right) + \mathcal{O}\left(h_3^2\right),$$

Le développement de Taylor d'ordre 3 de la fonction $f(x, y)$ autour du point (x_0, y_0) est défini par

$$\begin{aligned} f(x_0 + h_1, y_0 + h_2) = f(x_0, y_0) &+ \frac{\partial f}{\partial x}(x_0, y_0)h_1 + \frac{\partial f}{\partial y}(x_0, y_0)h_2 \\ &+ \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(x_0, y_0)h_1^2 + \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0)h_1 h_2 + \frac{1}{2}\frac{\partial^2 f}{\partial y^2}(x_0, y_0)h_2^2 \\ &+ r_2(h_1, h_2) \end{aligned} \tag{4.13}$$

où $r_2(h_1, h_2)$ est le terme d'erreur.

Exemple 14

Soit la fonction réelle $f(x, y)$ de deux variables définie par

$$f(x; y) = x^2 + x \sin(y). \tag{4.14}$$

1) Déterminer le polynôme de Taylor de degré 2 au point $(x_0, y_0) = (1, 0)$ de $f(x, y)$.
2) À l'aide de ce polynôme, donner une approximation de $f(1, 1; 0, 1)$

Propagation d'erreurs dans le cas général

Question: Pour une quantité inconnue x mais approchée par une valeur approximative x^*, Que peut-on dire de la précision de l'approximation de la valeur inconnue $f(x)$ par $f(x^*)$? En d'autres termes, si l'on a $x = x^* + \Delta x$, que serait l'erreur absolue $\Delta f = |f(x) - f(x^*)|$?

Pour répondre à cette question, on applique le développement de Taylor autour de x^*

$$f(x) = f(x^* + \Delta x) = f(x^*) \pm f'(x^*)\Delta x + \mathcal{O}\left((\Delta x)^2\right)$$

Note: En négligeant le termes d'erreur, on obtient l'estimation suivante

$$\Delta f = |f(x) - f(x^*)| \simeq f'(x^*)\Delta x \tag{4.15}$$

Remarque(s) : Si $f(x, y)$ est une fonction de deux variables x et y, elles-mêmes approchées par x^* et y^* avec une précision de Δx et Δy, on estime l'erreur absolue Δf par

$$\Delta f = |f(x, y) - f(x^*, y^*)| \simeq \left|\frac{\partial f}{\partial x}(x^*, y^*)\right|\Delta x + \left|\frac{\partial f}{\partial y}(x^*, y^*)\right|\Delta y \tag{4.16}$$

Pour une fonction de trois variables, $f(x, y, z)$, on a l'approximation suivante

$$\Delta f = |f(x, y, z) - f(x^*, y^*, z^*)| \simeq \left|\frac{\partial f}{\partial x}(x^*, y^*, z^*)\right|\Delta x + \left|\frac{\partial f}{\partial y}(x^*, y^*, z^*)\right|\Delta y + \left|\frac{\partial f}{\partial z}(x^*, y^*, z^*)\right|\Delta z$$

Exemple 15

On a mesuré la longueur d'un côté d'une boîte cubique et obtenu $l^* = 10,2\ cm$, avec une précision de l'ordre du millimètre ($\Delta l = 0,1\ cm$). Le volume v de cette boîte est déterminé par l'expression analytique $v(l) = l^3$.
Ainsi, la valeur approximative du volume est $(10,2)^3 = 1061,2\ cm^3$.
Grâce à la formule (4.15), l'erreur liée au calcul de ce volume est estimée comme suit

$$\Delta v \simeq |v'(l^*)|\Delta l = 3(10,2)^2 \times 0,1 = 31,212 \leq 0,5 \times 10^2.$$

Donc seuls les deux premiers chiffres du volume $v = 1061,2\ cm^3$ sont significatifs.

Note: De l'équation (4.16), on déduit la façon dont se propagent les erreurs dans les opérations élémentaires :

$$\begin{aligned}
f(x,y) = x + y \quad &\longrightarrow \quad \Delta f \simeq |\Delta x| + |\Delta y| \\
f(x,y) = x - y \quad &\longrightarrow \quad \Delta f \simeq |\Delta x| + |\Delta y| \\
f(x,y) = x \times y \quad &\longrightarrow \quad \Delta f \simeq |y^*||\Delta x| + |x^*||\Delta y| \\
f(x,y) = x \div y \quad &\longrightarrow \quad \Delta f \simeq \frac{|y^*||\Delta x| + |x^*||\Delta y|}{|y^*|^2}
\end{aligned}$$

À Retenir !

Je dois pouvoir répondre aux questions suivantes :

1. calculer/estimer l'erreur absolue et relative d'un nombre
2. établir les chiffres significatifs partant de l'erreur absolue et réciproquement.
3. faire une représentation en virgule flottante avec mantisse à n chiffres par troncature ou par arrondi
4. normaliser une mantisse et décaler la mantisse le cas échéant
5. faire des opérations élémentaires en virgurle flottante à n chiffres
6. reconnaitre les opérations dangereuses et comprendre l'importance de l'ordre dans les opérations élémentaires
7. utiliser la méthode de Horner
8. construire un polynôme de Taylor de degré n ainsi que son terme d'erreur
9. estimer le reste, approximer l'erreur absolue et obtenir le nombre de chiffres significatifs de mon approximation
10. comprendre et savoir calculer l'ordre d'une méthode
11. savoir calculer la propagation d'erreur pour une fonction d'une ou de plusieurs variables et ainsi calculer l'erreur/chiffres significatifs d'une évaluation de fonction

Exercices suggérés du manuel !

- Exercices [a] suggérés : 1, 8 - 13, 15, 19 - 25, 31, 35, 37
- Exercices fortement suggérés : 1, 9, 20, 24, 31, 37

a. André Fortin : *Analyse numérique pour ingénieurs*, Presses Internationales Polytechnique