DIRO
IFT 2425

DÉMONSTRATION N° 1

Max Mignotte

DIRO, Département d'Informatique et de Recherche Opérationelle, local 2384.
http : //www.iro.umontreal.ca/∼mignotte/IFT2425/
E-mail : mignotte@iro.umontreal.ca

I. Erreur relative, absolue, cse et propagation d'erreurs

Soit les valeurs suivantes

$$\begin{aligned} A &= 549.12 \\ B &= 1327.5 \\ C &= 10250.65 \\ D &= 47.2 \\ E &= 5.1278 \\ F &= 0.00371 \end{aligned}$$

où tout les chiffres sont significatifs.

1. Classer ces valeurs par ordre croissant d'erreur absolue.
2. Donner leur erreur relative et classer ces valeurs par ordre croissant d'erreur relative.
3. Calculer, en arrondissant au nombre de cse adéquat
   (a) $f(A, B) = A + B$,
   (b) $f(A, B, C) = (A + C)/B$.
4. Calculer $g((A+C)/B)$ avec $g(x) = x^2 - 2\sqrt{x}$. Estimer l'erreur absolu commise sur ce calcul et arrondir au nombre de cse adéquat.

Réponse

1. Par définition (cf. cours), un chiffre significatif d'une valeur Q^* (Q^* est l'approximation de Q) est exact si l'erreur absolue ($\Delta Q = |Q - Q^*|$) sur cette valeur est inférieur ou égale à une demi fois l'unité du rang du chiffre. i.e., si

$$\Delta Q \leq 0.5 \times 10^r \quad \text{où } r \text{ est le rang du chiffre.}$$

Pour A, on a 5 chiffres significatif et le rang du dernier chiffre est $r = -2$ (2 décimales après la virgule). Donc, on a

$$\Delta A = 0.5 \times 10^{-2}$$

De même, nous avons

$$\begin{aligned} \Delta B &= 0.5 \times 10^{-1} \\ \Delta C &= 0.5 \times 10^{-2} \\ \Delta D &= 0.5 \times 10^{-1} \\ \Delta E &= 0.5 \times 10^{-4} \\ \Delta F &= 0.5 \times 10^{-5} \end{aligned}$$

De ce fait, nous avons le classement par ordre croissant suivant

$$\Delta F < \Delta E < \Delta A \leq \Delta C < \Delta D \leq \Delta B$$

2. L'erreur relative d'une valeur Q est, par définition, $\Delta_r(Q) = |Q - Q^*| / |Q^*| = |\Delta Q| / |Q^*|$. On a donc les erreurs relatives suivantes

$$\begin{aligned}
\Delta_r(A) &= \frac{\Delta A}{A} = \frac{0.005}{549.12} \approx 9.11 \times 10^{-6} \\
\Delta_r(B) &= \frac{\Delta B}{B} = \frac{0.05}{1327.5} \approx 3.77 \times 10^{-5} \\
\Delta_r(C) &= \frac{\Delta C}{C} = \frac{0.005}{10250.65} \approx 4.88 \times 10^{-7} \\
\Delta_r(D) &= \frac{\Delta D}{D} = \frac{0.05}{47.2} \approx 1.06 \times 10^{-3} \\
\Delta_r(E) &= \frac{\Delta E}{E} = \frac{0.00005}{5.1278} \approx 9.75 \times 10^{-6} \\
\Delta_r(F) &= \frac{\Delta F}{F} = \frac{0.000005}{0.00371} \approx 1.35 \times 10^{-3}
\end{aligned}$$

De ce fait, nous avons le classement par ordre croissant d'erreur relative suivant

$$\Delta_r(C) < \Delta_r(A) < \Delta_r(E) < \Delta_r(B) < \Delta_r(D) < \Delta_r(F)$$

3.

(a) On a $f(A, B) = f(A^*, B^*) \pm \Delta f$, avec $A = A^* \pm \Delta A$ et $B = B^* \pm \Delta B$. L'erreur absolu sur Δf est alors

$$\Delta f = |f(A^*, B^*) - f(A, B)|$$

et peut être approchée, à l'aide de la différentielle par

$$\Delta f = \left| \frac{\partial f}{\partial A}(A^*, B^*) \right| \Delta A + \left| \frac{\partial f}{\partial B}(A^*, B^*) \right| \Delta B$$

Ce qui revient à négliger les termes d'ordres supérieur dans la formule de Taylor. On a donc puisque $\frac{\partial f}{\partial A}(A^\star, B^\star) = \frac{\partial f}{\partial B}(A^\star, B^\star) = 1$

$$\begin{aligned}
\Delta f &= \Delta A + \Delta B \\
&= 0.005 + 0.05 \\
&= 0.055
\end{aligned}$$

De plus

$$\begin{aligned}
f(A^*, B^*) &= A^* + B^* \\
&= 549.12 + 1327.5 \\
&= 1876.62
\end{aligned}$$

Nous avons l'erreur absolu Δf tel que $\Delta f = 0.055 < 0.5 \times 10^r$ avec $r = 0$. Le rang du dernier chiffre significatif est 0 et le calcul s'écrit donc en arrondissant

$$\begin{aligned}
f(A, B) &= A + B \\
&= 1877
\end{aligned}$$

(b) Même raisonement que précédement, on obtient

$$\begin{aligned}\Delta f &= \left|\frac{\partial f}{\partial A}(A^*,B^*,C^*)\right|\Delta A + \left|\frac{\partial f}{\partial B}(A^*,B^*,C^*)\right|\Delta B + \left|\frac{\partial f}{\partial C}(A^*,B^*,C^*)\right|\Delta C\\ &= \left|\frac{1}{B}\right|\Delta A + \left|-\frac{(A+C)}{B^2}\right|\Delta B + \left|\frac{1}{B}\right|\Delta C\end{aligned}$$

Numériquement, on obtient

$$\begin{aligned}\Delta f &= \frac{0.005}{1327.5} + \frac{549.12+10250.65}{(1327.5)^2}\times 0.05 + \frac{1}{1327.5}\times 0.005\\ &= 0.000313\end{aligned}$$

Le rang du dernier cse est -3 car $\Delta f = 0.000313 < 0.5\times 10^r$ avec $r=-3$. De plus on a

$$\begin{aligned}f(A^*,B^*,C^*) &= \frac{A^*+C^*}{B^*}\\ &= \frac{549.12+10250.65}{1327.5}\\ &= 8.13543\end{aligned}$$

Donc en arrondissant au rang $r=-3$

$$f(A^*,B^*,C^*) = 8.135$$

4.

En posant $g(A,B,C) = g(f(A,B,C))$ avec $f(A,B,C) = (A+C)/B$, nous avons

$$\begin{aligned}\Delta g &= \left|\frac{\partial g}{\partial A}(A^*,B^*,C^*)\right|\Delta A + \left|\frac{\partial g}{\partial B}(A^*,B^*,C^*)\right|\Delta B + \left|\frac{\partial g}{\partial C}(A^*,B^*,C^*)\right|\Delta C\\ &= \left|\frac{\partial g}{\partial f}\frac{\partial f}{\partial A}(A^*,B^*,C^*)\right|\Delta A + \left|\frac{\partial g}{\partial f}\frac{\partial f}{\partial B}(A^*,B^*,C^*)\right|\Delta B + \left|\frac{\partial g}{\partial f}\frac{\partial f}{\partial C}(A^*,B^*,C^*)\right|\Delta C\end{aligned}$$

avec

$$\frac{\partial g}{\partial f}(A^*,B^*,C^*) = 2f(A^*,B^*,C^*) - \frac{1}{\sqrt{f(A^*,B^*,C^*)}}$$

Donc

$$\begin{aligned}\Delta g &= \left|2\Big(\frac{A^*+C^*}{B^*}\Big) - \frac{1}{\sqrt{\frac{A^*+C^*}{B^*}}}\right| \left(\left|\frac{\partial f}{\partial A}(A^*,B^*,C^*)\right|\Delta A + \left|\frac{\partial f}{\partial B}(A^*,B^*,C^*)\right|\Delta B + \left|\frac{\partial f}{\partial C}(A^*,B^*,C^*)\right|\Delta C\right)\\ &= 15.92024145\times 0.000313 = 0.004983\\ &\leqq 0.5\times 10^r \quad \text{avec} \quad r=-2\end{aligned}$$

Le rang du dernier cse est -2 donc on a

$$g\left(\frac{A+C}{B}\right) = 60.48$$

II. Mesure d'incertitude

Dans un circuit électrique, la résistance R est équivalente à deux résistances R_1 et R_2 placées en parallèle, est donnée par la relation

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}$$

R_1 et R_2 sont connues avec une incertitude maximum de 2%. Utiliser la notion de diférentielle pour estimer l'incertitude maximum sur la valeur de R.

Réponse

On peut écrire $R = f(R_1, R_2) = R_1R_2/(R_1 + R_2)$. Si R_1 et R_2 sont les valeurs exactes inconnues, ΔR_1 et ΔR_2 les erreurs de mesure (inconnues mais petites), on a par hypothèse

$$\begin{array}{rrcl} & \Delta_r(R_1) & = & \left|\dfrac{\Delta R_1}{R_1}\right| \leq 0.02 \\ \text{et} & \Delta_r(R_2) & = & \left|\dfrac{\Delta R_2}{R_2}\right| \leq 0.02 \end{array}$$

L'erreur sur R est alors $\Delta R = |f(R_1^*, R_2^*) - f(R_1, R_2)|$ avec R_1^* et R_2^* valeur approximée de R_1 et R_2, et peut être approchée, à l'aide de la différentielle, par

$$\Delta R = \Delta f = \left|\frac{\partial f}{\partial R_1}(R_1^*, R_2^*)\right| \Delta R_1 + \left|\frac{\partial f}{\partial R_2}(R_1^*, R_2^*)\right| \Delta R_2$$

Si on néglige les termes d'ordre supérieur dans la formule de Taylor. On a donc

$$\Delta R = \Delta f = \left|\frac{R_2^2}{(R_1 + R_2)^2}\right| \Delta R_1 + \left|\frac{R_1^2}{(R_1 + R_2)^2}\right| \Delta R_2$$

et pour $\frac{\Delta R}{R}$ (en rappelant que $R = R_1R_2/(R_1 + R_2)$)

$$\frac{\Delta R}{R} = \left|\frac{R_2}{R_1 + R_2}\right| \left|\frac{\Delta R_1}{R_1}\right| + \left|\frac{R_1}{R_1 + R_2}\right| \left|\frac{\Delta R_2}{R_2}\right|$$

On a une estimation de $\frac{\Delta R}{R}$ en majorant cette dernière expression à l'aide de l'inégalité suivante

$$\begin{array}{rcl} \left|\dfrac{R_2}{R_1 + R_2}\right| \left|\dfrac{\Delta R_1}{R_1}\right| + \left|\dfrac{R_1}{R_1 + R_2}\right| \left|\dfrac{\Delta R_2}{R_2}\right| & \leq & \dfrac{R_2}{R_1 + R_2} \times 0.02 + \dfrac{R_1}{R_1 + R_2} \times 0.02 \\ \left|\dfrac{\Delta R}{R}\right| & \leq & 0.02 \end{array}$$

L'incertitude sur R est donc aussi de 2%

III. Arithmétique des ordinateurs et erreurs en notation flottante

Déterminer et exprimer les quantitées ci dessous après troncature ou arrondissement en notation flottante de la forme

$$0.\, x_1\, x_2\, x_3 \ldots x_{s-1}\, x_s \times b^e$$

où le premier *digit* après la décimale est différent de zéro pour le système de représentation simplifié suivant

$$b = 10, \quad s = 3, \quad -9 \leq e \leq 9$$

où b, s et e représente la base, le nombre de chiffre de la mantisse et e l'exposant (cf. cours)

1. $12.3 + 0.0234$
2. $-0.0321 + 0.000136$
3. $12.3 - 0.0234$
4. $-321 + 32.1$
5. 132×0.987
6. $-2.14/0.000137$
7. $(-0.111 + 0.222) \times 0.00111/999$ (de gauche à droite)

Réponse

1.

$$12.3 + 0.0234 = 0.123 \times 10^2 + 0.000234 \times 10^2 = 0.123234 \times 10^2$$

$$\begin{array}{rcl} \text{Tronqué} & : & 0.123\ E2 \\ \text{Arrondi} & : & 0.123\ E2 \end{array}$$

2.

$$-0.0321 + 0.000136 = -0.321 \times 10^{-1} + 0.00136 \times 10^{-1} = -0.31964 \times 10^{-1}$$

$$\begin{array}{rcl} \text{Tronqué} & : & -0.319\ E-1 \\ \text{Arrondi} & : & -0.320\ E-1 \end{array}$$

3.

$$12.3 - 0.0234 = 0.123 \times 10^2 - 0.000234 \times 10^2 = 0.122766 \times 10^2$$

$$\begin{array}{rcl} \text{Tronqué} & : & 0.122\ E2 \\ \text{Arrondi} & : & 0.123\ E2 \end{array}$$

4.

$$-321 + 32.1 = -0.0321 \times 10^3 + 0.321 \times 10^3 = -0.2889 \times 10^3$$

$$\begin{array}{rcl} \text{Tronqué} & : & -0.288\ E3 \\ \text{Arrondi} & : & -0.289\ E3 \end{array}$$

5.

$$132 \times 0.987 = 0.132 \times 10^3 \times 0.987 \times 10^0 = 0.130284 \times 10^3$$

$$\begin{array}{rcl} \text{Tronqué} & : & 0.130\ E3 \\ \text{Arrondi} & : & 0.130\ E3 \end{array}$$

6.

$$-2.14/0.000137 = -0.214 \times 10^1/0.137 \times 10^{-3} = -0.1562043796 \times 10^5$$

$$\begin{aligned} \text{Tronqué} &: \quad -0.156\,E5 \\ \text{Arrondi} &: \quad -0.156\,E5 \end{aligned}$$

7.

$$\begin{aligned} (-0.111 + 0.222) \times (0.00111/999 &= 0.111 \times 10^0 \times (0.111 \times 10^{-2})/0.999 \times 10^3 \\ &= 0.123 \times 10^{-3}/0.999 \times 10^3 = 0.123 \times 10^{-6} \end{aligned}$$

$$\begin{aligned} \text{Tronqué} &: \quad 0.123\,E-6 \\ \text{Arrondi} &: \quad 0.123\,E-6 \end{aligned}$$

IV. Méthode de la bissection

Déterminer par la méthode de la bissection (ou méthode de dichotomie) la racine r de l'équation $x^2 - 2 = 0$ contenue dans l'intervalle $[1, 2]$ avec une erreur absolue inférieure à 10^{-2}. Trouver analytiquement combien d'itérations sont nécessaires pour arriver à cette précision.

Réponse

La méthode de la bissection permet de construire à partir de l'intervalle $[a, b]$ contenant r, un nouvel intervalle de longueur moitié contenant r. En appliquant n fois consécutives la méthode, on obtient un intervalle de longueur $(b-a)/2^n$ contenant r. A l'itération n, on a un majorant de l'erreur absolu donné par

$$\Delta r = \frac{|b-a|}{2^n}.$$

On veut $\Delta r < 10^{-2}$, donc

$$\begin{aligned} \frac{|2-1|}{2^n} &< 10^{-2} \\ 2^n &> 10^2 \\ n &> 6.64 \\ n &= 7 \qquad \text{soit 7 itérations.} \end{aligned}$$

On vérifie que $f(1) \times f(2) < 0$. Posons $f(x) = x^2 - 2$, $f(x)$ est croissante et continue sur $[1, 2]$ et $f(1) = -1$ et $f(2) = 2$, ce qui assure l'unicité de la solution. En faisant les calculs avec 3 décimales, on obtient les résultats suivants

x	$f(x)$	encadrement de r	
1	-1	-	-
2	2	1	2
1.5	0.25	1	1.5
1.25	-0.437	1.25	1.5
1.375	-0.109	1.375	1.5
1.437	0.065	1.375	1.437
1.406	-0.023	1.406	1.437
1.421	0.019	1.406	1.421
1.414	-0.00201775	1.414	1.421

On a donc $1.414 < r < 1.421$ où $r = 1.4175 \pm |\Delta_r|$ avec $|\Delta r| = 0.00350 = 0.350 \times 10^{-2}$.

V. Méthode de l'interpolation linéaire

Déterminer en utilisant la méthode d'interpolation linéaire (appelée aussi méthode de la corde ou méhode de Lagrange) la racine r de l'équation $x^2 - 2 = 0$ contenue dans l'intervalle $[1, 2]$ avec une erreur inférieure à 10^{-2} déterminée numériquement.

Réponse

On vérifie que $f(1) \times f(2) < 0$ ($f(1)$ et $f(2)$ sont de signes opposés). Posons $f(x) = x^2 - 2$, $f(x)$ est croissante et continue sur $[1, 2]$ et $f(1) = -1$ et $f(2) = 2$, ce qui assure l'unicité de la solution. On calcule

$$\begin{aligned} x_1 &= 2 - f(2)\Big((2-1)/(f(2) - f(1))\Big) \\ &= 2 - 2 * (1/3) \\ &= \frac{4}{3} \\ &= 1.3333 \end{aligned}$$

Le nouvel intervalle contenant la solution est $[4/3, 2]$. On continue une autre itération

$$\begin{aligned} x_2 &= 2 - f(2)\Big((2-4/3)/(f(2) - f(4/3))\Big) \\ &= 1.4 \end{aligned}$$

Le nouvel intervalle contenant la solution est $[1.4, 2]$. On vérifie que $f(1.4) \times f(2) < 0$. On continue jusqu'a ce que $|x_{n+1} - x_n| < 10^{-2}$.

$$\begin{aligned} x_3 &= 2 - f(2)\Big((2-1.4)/(f(2) - f(1.4))\Big) \\ &= 1.411764 \end{aligned}$$

Le nouvel intervalle est $[1.411764, 2]$.

$$\begin{aligned} x_4 &= 2 - f(2)\Big((2-1.411764)/(f(2) - f(1.411764))\Big) \\ &= 1.413792982 \end{aligned}$$

On peut remarquer que cette méthode converge beaucoup plus rapidement que la méthode de la bissection vue précedemment.