Méthodes itératives de résolution d'équation

Exercice 1 Résoudre l'équation $f(x)=0$ avec les trois méthodes – Newton, sécante, bissectrice – pour les fonctions suivantes :
1) $f(x)=2^x-10$; 2) $f(x)=x^2+4x-1$, $r_0=5$ our $r_0=4$; 3) $f(x)=x^3-x+1$, $r_0=-1$;
4) $f(x)=\cos x-2x$, $r_0=0$. 5) $f(x)=(x-1)^{10}$, avec $r_0=1{,}5$.

1) en partant de 3 on a successivement 3,360 6, 3,322 443, 3,321 928 187 0, 3,321 928 09
2) en partant de 4 on a successivement 1,416 67, 0,440 040 65, 0,244 59, 0,236 084 1, 0,236 067 977
4) en partant de 0 on a successivement 0,5, 0,450 6, 0,450 183 647 5, 0,450 183 611
Pour le dernier : on arrive vite avec la méthode de Newton à un $r_n=1{,}0054$, pour lequel $f(r_n)=2{,}06\times10^{-23}$: on a $f(r_n)$ très proche de 0, alors que r_n est encore loin de la solution exacte 1.

Exercice 2 *(D'après : M. A. Cornea-Hasegan, R. A. Golliver, and P. W. Markstein. Correctness proofs outline for newton-raphson based floating-point divide and square root algorithms. In* 14th IEEE Symposium on Computer Arithmetic (Arith-14 '99)*, pages 96–105. IEEE Computer Society, 1999.)*
L'architecture IA-64 de Intel, fournit une opération de division entre nombres flottants qui utilise la méthode de Newton-Raphson de la manière suivante: pour calculer a/b, on calcule d'abord $1/b$ par résolution d'une équation avec la méthode de Newton-Raphson, puis on multiplie le résultat par a.

Question 2.1 $1/b$ est solution de $f(x)=0$, où $f(x)=b-1/x$. Donner une définition de la suite $(x_n)_{n\geq0}$ correspondant à la méthode de Newton pour résoudre $f(x)=0$.

On suppose maintenant que b est un nombre flottant normalisé. Une première approximation à 2 chiffres significatifs de $1/b$ est donnée par une table d'après les deux premiers chiffres de la partie fractionnaire de b.

Question 2.2 Quelle est la taille de cette table?

(L'architecture IA-64 fournit une opération `frcpa`, qui donne une approximation avec 8 bits significatifs de l'inverse d'un flottant, tirée d'une table d'après les 8 premiers bits de ce nombre).

Question 2.3 Calculer une valeur approchée de $1/21{,}34$ à 10^{-5} près. Combien faut-il d'itérations? (L'inversion est réalisée avec 4 itérations sur l'architecture IA-64).

Question 2.4 Donner une méthode permettant de calculer de manière similaire $1/\sqrt{a}$ pour tout flottant positif a, sans faire de division. Comment peut-on ensuite calculer $\sqrt{a}$, toujours sans faire de division. Calculer ainsi $\sqrt{1{,}3456}$ et $\sqrt{4}$.

On cherche un zéro de f définie par $f(r)=a-1/r^2$.

Exercice 3 Dans de nombreuses applications graphiques (notamment les jeux vidéos) il est nécessaire de normaliser des vecteurs, c'est à dire de les diviser par leur norme euclidienne, qui est une racine carrée[1]. Il faut donc pouvoir calculer très rapidement l'inverse d'une racine carrée, et les fabricants de cartes graphiques ont étudié ce calcul dans les années 70, en utilisant – entre autres - la méthode de Newton de résolution d'équations : on peut calculer une approximation de $1/\sqrt{a}$ en résolvant l'équation $f(x)=0$ pour f définie par $f(x)=a-1/x^2$.

Question 3.1 Après avoir calculé une première approximation de $1/\sqrt{5}$ à 10^{-1} avec par exemple la méthode de la bissectrice, calculez une approximation à 10^{-8} près avec la méthode de Newton.

$r_{n+1}=r_n-f(r_n)/f'(r_n)=r_n-(a-1/r_n^2)/(2/r_n^3)=r_n-\frac{1}{2}ar_n^3+\frac{1}{2}r^n=r_n(\frac{3}{2}-\frac{a}{2}r_n^2)$. *En partant de $r_0=0{,}425$: $r_1=0.4456$, $r_2=0.4472047$, $r_3=0.4472135952$, $r_4=0.4472135954$; une itération de moins en partant de $0{,}45$.*

Si a est une puissance paire de 2, de la forme $a=2^{2n}$, alors $1/\sqrt{a}=2^{-n}$. Plus généralement, si a est un nombre proche de 2^{2n}, alors $1/\sqrt{a}$ est proche de 2^{-n} ; en posant $n=\lfloor\log_2(a)/2\rfloor$, on peut prendre 2^{-n-1} comme première approximation de $1/\sqrt{a}$.

Question 3.2 Calculez une première approximation de $1/\sqrt{5}$ avec cette méthode, puis écrivez le code en C ou C++ d'une fonction permettant de calculer l'inverse de la racine d'un nombre en calculant cette première approximation, puis en itérant la méthode de Newton jusqu'à une précision donnée en paramètre.

Le code suivant est devenu célèbre :

```
float FastInvSqrt(float x) { float xhalf = 0.5f * x;
  int i = *(int*)&x;          // evil floating point bit level hacking
  i = 0x5f3759df - (i >> 1);  // what the fuck?
  x = *(float*)&i; x = x*(1.5f-(xhalf*x*x)); return x; }
```

[1] Si x est un vecteur en dimension 3, sa norme euclidienne / longueur est $\sqrt{x_1^2+x_2^2+x_3^2}$.

Nombres flottants

Exercice 4 On suppose un système de représentation en base 10, avec 2 chiffres pour l'exposant et 3 pour la mantisse, et un décalage de 49 : $(s;e;m)=(-1)^s\times 0{,}m\times 10^{e-49}$

Question 4.1 Soient $x=(0;99;900)$, $y=(0;96;400)$ et $z=(0;96;342)$. Dans quel ordre faut-il additionner ces trois nombres afin d'avoir un résultat le plus exact possible ?

Question 4.2 Calculer en arithmétique à 3 chiffres $f(4{,}71)$, où f est définie par $f(x)=x^3-6{,}10x^2+3{,}20x+1{,}50$, une première fois sans factoriser le polynôme ; puis une nouvelle fois en factorisant.

Exercice 5 Les nombres de bits alloués à la mantisse et à l'exposant du type long double ne sont pas standard, le but de cet exercice est de les déterminer sur l'ordinateur sur lequel on travaille.

Question 5.1 Expliquer comment sont représentés, selon la norme IEEE 754, et pour des nombres donnés $|m|$ de bits pour la mantisse, $|e|$ de bits pour l'exposant, et pour un excédant q donné, les nombres 1 ; $1{,}5$; $1+1/2^3$. Implémenter un algorithme permettant de calculer le nombre de bits $|m|$ de la mantisse.

Question 5.2 Comment sont représentés les nombres $1/2$; $1/4$; $1/2^3$? Quels sont, en fonction de l'excédant q, le plus petit réel normalisé >0 et le plus petit réel >0 représentables ? Implémenter un algorithme qui permet de calculer l'excédant du type long double. *+ petit réel >0 normalisé : 2^{1-q} ; plus petit réel représentable : $2^{1-q-|m|}$; donc si N est la plus petite puissance de 2 représentable, $N=q+|m|+1$.*

Question 5.3 Expliquer comment sont représentés, en fonction de l'excédant q, les réels 2 ; 4 ; 2^3. Quel est, en fonction du nombre de bits $|e|$ alloués à l'exposant, le plus grand N tel que 2^N peut être représenté ? En déduire un algorithme qui permet de calculer le nombre de bits de l'exposant. *+ grande puissance de 2 représentable : $2^{2^{|e|}-2-q}$, donc $|e|=\log_2(N+2+q)$.*

Question 5.4 Comparer vos résultats aux renseignements fournis par la librairie <limits> (documentée par exemple à l'adresse `www.cplusplus.com/reference`), par exemple :

numeric_limits<double>::digits **:** le nombre de chiffres binaires qui peuvent être représentés sans modification dans le type long double ;

numeric_limits<double>::max() **:** retourne le plus grand réel qui peut être représenté dans le type double.

Interpolation polynomiale et intégration

Exercice 6 Pour chacune des fonctions suivantes, calculez une approximation de la valeur de la fonction en $0{,}45$ grâce au polynôme d'interpolation de la fonction en 0, $0{,}6$ et $0{,}9$ (en utilisant les polynômes de Lagrange), et donnez une majoration de l'erreur commise : $f(x)=\cos(x)$ $g(x)=\sqrt{1+x}$ $h(x)=\ln 1+x$

Exercice 7 *(D'après : S. Story and P. T. P. Tang. New algorithms for improved transcendental functions on ia-64. In* 14th IEEE Symposium on Computer Arithmetic (Arith-14 '99)*, pages 4–11. IEEE Computer Society, 1999.)*
On s'intéresse au calcul de la valeur de $\cos(x)$ pour un réel quelconque x.

Question 7.1 Calculer les coefficients du polynôme d'interpolation de la fonction $\cos$ en $-\frac{\pi}{2}$, 0 et $\frac{\pi}{2}$. En déduire une valeur approchée de $\cos(1)$.

Question 7.2 Comment peut-on calculer une valeur approchée de $\cos(x)$ pour $x\in[\frac{\pi}{2},\frac{3\pi}{2}]$?

Question 7.3 Comment peut-on ramener, pour tout $x\in\mathbf{R}$, le calcul de $\cos(x)$ au calcul de $\cos(y_x)$ pour $y_x\in[-\frac{\pi}{2},\frac{3\pi}{2}]$?

Question 7.4 Calculer un majorant de l'erreur commise avec cette approximation de $\cos(x)$, en admettant que pour tout $x\in[a,b]$, $|(x-a)(x-\frac{a+b}{2})(x-b)|\leq|b-a|^3/12\sqrt{3}$.

Exercice 8 Calculez chacune des intégrales suivantes avec les méthodes des trapèzes et de Simpson en utilisant la valeur de la fonction en 7 points , éventuellement après avoir fait un changement de variable approprié, et donnez une majoration de l'erreur commise : 1) $\int_0^1 \exp(-t^2/2)dt$ 2) $\int_1^2 \frac{\ln(t)}{(t-1)^{1/5}}dt$; 3) $\int_1^{+\infty}\frac{1}{1+t^4}dt$.
Pour chacune de ces intégralles, combien faudrait-il de sous-intervalles pour garantir une précision de 10^{-10} ?

Exercice 9 On peut calculer une valeur approchée de $\ln(x)$ en calculant une valeur approchée de $\int_1^x \frac{dt}{t}$.

Question 9.1 Calculer $\ln(2)$ avec la formule de Simpson, sans découper l'intervalle. Quelle majoration peut-on donner de l'erreur?

Question 9.2 En considérant suffisamment de sous-intervalles, calculez une valeur approchée de $\ln(2)$ à 10^{-5} près. En combien de sous-intervalles faudrait-il découper l'intervalle $[1;2]$ pour obtenir cette précision avec la formule des trapèzes par intervalles ?

Équations linéaires et vecteur propres

Exercice 10

Question 10.1 Montrer le déroulement de l'algorithme d'élimination, avec échange de lignes si nécessaire, sur les problèmes suivants :

$$\left\{\begin{array}{rrrcr} 4x_0 & -x_1 & +x_2 & = & 8 \\ 2x_0 & +5x_1 & +2x_2 & = & 3 \\ x_0 & +2x_1 & +4x_2 & = & 11 \end{array}\right. ; \left\{\begin{array}{rrrcr} x_0 & -x_1 & +3x_2 & = & 2 \\ 3x_0 & -3x_1 & +x_2 & = & -1 \\ x_0 & +x_1 & & = & 3 \end{array}\right. ; \left\{\begin{array}{rrrcr} x_0 & +x_1 & +x_2 & = & 4 \\ 2x_0 & +2x_1 & +x_2 & = & 6 \\ x_0 & +x_1 & 2x_2 & = & 6 \end{array}\right.$$

Exercice 11 Écrire un algorithme qui implémente la méthode de Jacobi sans utiliser d'opération sur les matrices.

Exercice 12 En utilisant la méthode des puissances et la déflation, calculer les valeurs propres et vecteurs propres des matrices suivantes :

$$\begin{pmatrix} 4 & -1 & 1 \\ -1 & 3 & -2 \\ 1 & -2 & 3 \end{pmatrix}, \begin{pmatrix} 5 & -3 \\ 6 & -4 \end{pmatrix}$$

Pour la première matrice : $\lambda^{(1)}=6$, $v^{(1)}=\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$, $\lambda^{(2)}=3$, $v^{(2)}=\begin{pmatrix} -2 \\ -1 \\ 1 \end{pmatrix}$...

Pour la seconde matrice : les valeurs propres sont 2 *et* -1.

Exercice 13 Écrire un algorithme permettant de calculer toutes les valeurs propres d'une matrice en utilisant la déflation.

Résolution d'équations différentielles

Exercice 14 On a $x'(t)=x(t)-t^2+1$ pour $0\leq t\leq 2$ avec $x(0)=0{,}5$.

Question 14.1 Résoudre cette équation avec la méthode d'Euler avec un pas $h=0{,}2$.

Question 14.2 Comparer avec la solution analytique de cette équation, définie par $x(t)=(t+1)^2-0{,}5e^t$.

Question 14.3 Quel pas peut-on prendre pour être certain d'avoir une erreur inférieure à 10^{-15} ?

Exercice 15 Soit l'équation $x'(t)=3x(t)-3t$ pour $0\leq t\leq T$, avec $x(0)=\alpha$.

Question 15.1 En utilisant la méthode d'Euler avec un pas de 1, calculer $x(6)$ pour $\alpha=1/3$.

Question 15.2 Résoudre à nouveau l'équation avec la même méthode, mais avec cette fois-ci la condition initiale $\alpha=0{,}33$.

Question 15.3 Résoudre analytiquement l'équation (une solution est de la forme $x(t)=ae^{bt}+ct+d$), et calculer $x(6)$.

Exercice 16 On s'intéresse à l'évolution, année après année, des populations de deux groupes d'individus, les *proies* et les *prédateurs*: le taux de natalité annuel des proies est de 3 naissances pour une proie, alors que leur taux de mortalité dépend du nombre de prédateurs. Ainsi, s'il y a $y(t)$ prédateurs à un instant t donné, le taux de mortalité annuel des proies à ce moment est estimé à $0{,}002\times y(t)$. A l'inverse, le taux de mortalité annuel des prédateurs est fixe, il vaut $0{,}5$, alors que leur taux de natalité annuel dépend de la quantité de nourriture disponible, et donc du nombre de proies: il vaut $0{,}0006x(t)$, où $x(t)$ est le nombre de proies à l'instant t. Ainsi, l'évolution des nombres de proies et de prédateurs est donnée par les équations suivantes:

$$\left\{\begin{array}{l} x'(t)=3x(t)-0{,}002x(t)y(t) \\ y'(t)=0{,}0006x(t)y(t)-0{,}5y(t) \end{array}\right.$$

En supposant qu'on a initialement 1 000 proies et 500 prédateurs, quelle population disparaît en premier ?

Exercice 17 On a $x''(t)-2x'(t)+2x(t)=e^{2t}\sin(t)$ pour $0\leq t\leq 1$, avec $x(0)=-0{,}4$ et $x'(0)=-0{,}6$. Calculer une approximation de $x(1)$. (Suggestion : on peut introduire une fonction intermédiaire $y(t)\ldots$)

Exercice 18 On s'intéresse à la flexion d'une poutre soumise à une charge uniforme et soutenue à ses deux extrémités. On note:

l: la longueur de la poutre;
q: la charge unitaire à laquelle est soumise la poutre;
E: l'élasticité de la poutre;
S: la traction aux extrémités;
I: le moment d'inertie.

Finalement, $w(t)$ est la mesure de la déformation de la poutre en un point d'abscisse t: c'est la distance entre la droite passant par les deux extrémités et la surface de la poutre. On montre que $w(t)$ est solution de:

$$w''(t)=\frac{S}{EI}w(t)-\frac{q}{2EI}t(l-t)$$

pour $0\leq t\leq l$, avec $w(0)=w(l)=0$.

Question 18.1 Donner le système d'équations permettant de calculer $w(l/2)$ avec la méthode des différences finies vue en cours.