Université "François Rabelais" de Tours
U. F. R. Sciences et Techniques
DEUG Sciences de la Matière 2ème Niveau 2000–2001

Introduction aux Techniques du Calcul Numérique

Stam NICOLIS
E-mail: nicolis@celfi.phys.univ-tours.fr
Web: http://www.phys.univ-tours.fr/∼nicolis
CNRS–Laboratoire de Mathématiques et Physique Théorique (UPRES A 6083)
et
Département de Physique, Université de Tours
Parc de Grandmont, 37200 Tours

Résumé

Le but de ce polycopié est de présenter les méthodes de base du calcul numérique, qui servent à résoudre des problèmes actuels en physique et en chimie.

1 Calcul des Zéros d'une Fonction

On commence par le problème numérique, certainement le plus commun–trouver les zéros d'une fonction.

Il est bien connu que calculer analytiquement les racines d'une fonction donnée est, en général, impossible. En ce qui concerne les polynômes, qui sont les fonctions les plus "simples", il a fallu plusieurs siècles pour qu'on puisse avancer des expressions bien-connues pour les racines de l'équation quadratique, aux expressions pour celles de la cubique et quartique (la Renaissance). La recherche pour les expressions des racines des équations *génériques* de degré plus grand que quatre dura plusieurs siècles, avant qu'on ne commence à soupçonner–à la fin du dix-huitième siècle et vers les début du dix-neuvième grâce à Lagrange et au travail culminant de Henrik Abel et Evariste Galois qu'il était *impossible* d'exprimer les zéros d'une équation polynomiale *générique*, de degré plus grand que quatre, à l'aide d'un nombre *fini* des racines carrées, cubiques, quartiques, etc. d'expressions rationnelles des coefficients du polynôme[1]

∗ Il existe, bien sûr des classes entières de polynômes de degré plus grand que quatre, dont les zéros peuvent s'exprimer de cette manière-e.g. $a_6x^6 + a_5x^5 + a_4x^4 + a_3x^3 + a_4x^2 + a_5x + a_6 = 0$ se laisse factoriser à un produit de $x^2 + 1$ et d'un polynôme de degré quatre pour n'importe quel choix des coefficients a_6, a_5, a_4, a_3. Mais cette classe est une classe *particulière*-l'équation générique du sixième degré n'admettant pas une telle factorisation.

Pour les fonctions plus compliquées, bien sûr, il n'existe pas d'expressions génériques. Mais il y a un autre point, qui est le fil conducteur de ce cours:

Malheureusement, d'un point de vue purement utilitaire, les expressions intéressantes, jolies et merveilleuses des racines des équations cubiques et quartiques sont très mal-adaptées pour une évaluation *numérique*. Ce qui veut dire que, si on cherche à calculer les racines, bien spécifiques, d'une équation à coefficients donnés, l'évaluation des expressions résultantes est tellement sujette à l'erreur numérique, que cette approche n'est pas compétitive, face aux méthodes qu'on exposera dans ce chapitre[2] En plus, ces méthodes sont capables de traiter sur le même plan fonctions "simples" et compliquées.

Le problème peut être posé, alors, de la manière suivante: on cherche les racines de

$$f(x) = 0 \tag{1}$$

où $x \in \mathbb{R}^n$ et $f : \mathbb{R}^n \to \mathbb{R}^n$. On va surtout s'intéresser au cas $n = 1$–mais il est clair que le cas $n > 1$ fera une apparition lorsqu'on aura à traiter les racines *complexes* d'une fonction réelle.

[1] La chose merveilleuse est que ce résultat n'a été obtenu que comme un cas particulier dans une étude beacoup plus vaste, à savoir celles des symétries des familles d'objets sous permutations-ici les objets sont les coefficients du polynôme. Cette étude signait l'acte de naissance de la *théorie des groupes* qui allait jouer un rôle fondamental tant en mathématiques qu'en physique, chimie et biologie.

[2] Même pour l'équation quadratique, les expressions habituellement données souffrent d'erreur d'arrondissement.

1.1 Les Zéros sont Réels et Simples

En ce qui concerne les racines *réelles* les choses–en théorie– sont assez immédiates: on sait que, si la fonction est *continue* dans un intervalle $[a, b]$ et $f(a)f(b) < 0$, il y aura, au moins, un point intermédiaire, x^*, pour lequel $f(x^*) = 0$–cf. fig 1. Si, par contre, la fonction ne change pas de

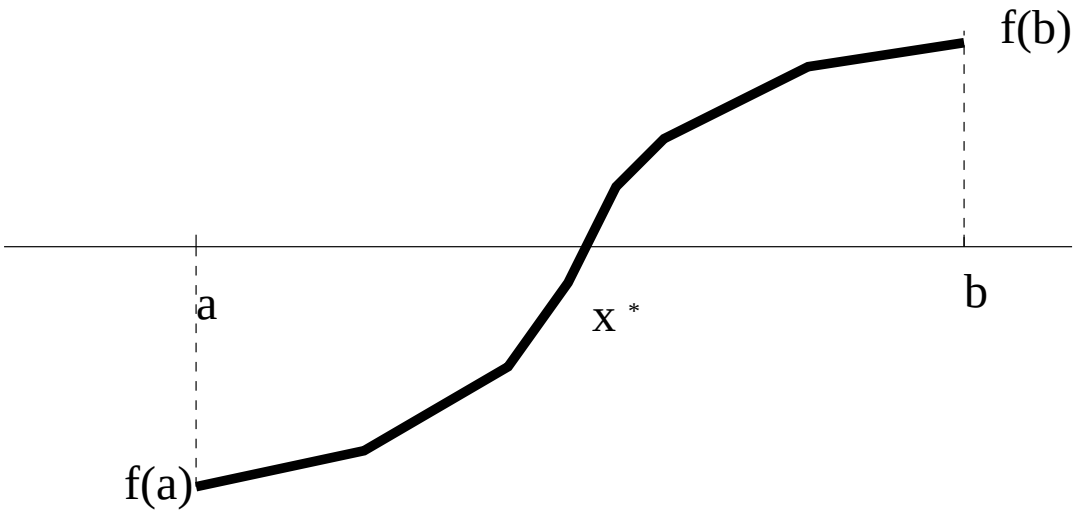


Figure 1: Exemple du théorème des valeurs intermédiaires: une fonction continue dans un intervalle, qui prend des valeurs de signe opposé au bout de cet intervalle, doit s'annuler au moins une fois dans cet intervalle.

signe, dans aucun intervalle, alors ses racines sont complexes-et la situation...se complique.

La puissance de ce théorème réside dans son aspect positif-il nous garantit des renseignements sur l'existence d'une racine contre un prix typiquement faible (la connaissance des valeurs d'une fonction). Il peut ainsi devenir le point de départ d'une méthode d'approximation de la racine, ainsi qu'un moyen de contrôle de ladite approximation. La stratégie générale se dessine, alors, clairement: une fois qu'on a trouvé un intervalle quelconque, qui contienne la racine (les complications des racines multiples sont momentanement négligées-on y revient pour les polynômes plus tard), on essaiera de le réduire, jusqu'à ce que sa taille soit comparable à la précision numérique, auquel cas on ne pourra pas distinguer, numériquement, l'un bout de l'autre.

L'algorithme prend la forme suivante:
$k = 0; a_0 = a, \;\; b_0 = b$
tant que $|a_k - b_k| > \epsilon\{$
$m_k = (a_k + b_k)/2.$
si $f(m_k)f(a_k) < 0$ **alors**
$a_k = a_k; b_k = m_k$
sinon $a_k = m_k; b_k = b_k; k = k + 1\}$
$x^* = m_k$

La constante ϵ est la précision du calcul-sur un ordinateur typique elle est de l'ordre de 10^{-8} en *simple précision* et 10^{-16} en *double précision*. On note que, dans cette méthode, la

taille de l'intervalle se divise par un facteur 2 à chaque itération-d'où son nom-*la méthode de la dichotomie* (ou *bissection*).

On doit maintenant addresser un point très important , du point de vue pratique: à précision fixe, combien d'itérations doit-on effectuer? La simplicité de l'algorithme nous permet d'y répondre tout de suite[3]:

$$k_{\min} = \frac{\ln \left| \frac{b_0 - a_0}{\epsilon} \right|}{\ln 2} \tag{2}$$

Pour mieux comprendre la portée de ce résultat, étudierons un problème concret.

Points de rebroussement. Soit un point matériel, de masse m, relié à un ressort idéal, horizontal, de constante k, déplacé de x de sa position d'équilibre. Son énergie potentielle est $V(x) = (1/2)kx^2$. Si son énergie totale est E, son déplacement maximale sera la racine de l'équation $E = V(x)$. Un dessin peut convaincre que il y a deux points qui vérifient cette équation, pour n'importe quelle valeur de l'énergie-le mouvement de la masse est *borné* et *périodique*-et ces points ne dépendent pas de la masse du point. Si on choisit $k = 2J/m^2$ et $E = 0.7J$ on se ramène à l'équation

$$x^2 = 0.7 \tag{3}$$

Pour $x > 0$ on peut facilement se rendre compte que $0.64 = 0.8^2 < 0.7 < 0.81 = 0.9^2$, ce qui conduit au choix $[0.8, 0.9]$ pour l'intervalle initial et une précision 10^{-1} déjà–on sait que $0.8 < x^* = 0.8\ldots < 0.9$. L'équation (2) prédit que 3 itérations supplémentaires suffiront pour qu'on soit sûr du chiffre suivant et qu'à 5 itérations la précision sera de 10^{-4}, i.e. on sera sûr de trois chiffres après la décimale. En fait, on trouve $a_1 = 0.8, b_1 = 0.85$; $a_2 = 0.825, b_2 = 0.85$; $a_3 = 0.825, b_3 = 0.8375$; $a_4 = 0.83125, b_4 = 0.8375$–et l'assurance que $x^* = 0.83\ldots$.

Pour un cas moins trivial, on pourra prendre un modèle plus réaliste du ressort, dans lequel on tient compte des effets non-linéaires, qui entrent en jeu lorsque les déplacements du point d'équilibre ne peuvent plus être considérés comme "petits" et l'approximation linéaire devient insuffisante. Une expression du type $V(x) = (1/2)kx^2 + ax^3$, par exemple, avec $a < 0$, pourra servir pour comprendre ce qui se passe lorsque le ressort se casse – et la masse se libère– quand l'énergie E dépasse une valeur critique $\hat{V}$. Alors l'équation devient

$$ax^3 + (1/2)kx^2 - E = 0 \tag{4}$$

Prenons la même valeur pour l'énergie, $E = 0.7$ et pour la constante $k = 2J/m^2$ et essayons de comprendre l'effet d'un coefficient a petit–$a \simeq -0.1$ par exemple. On peut considérer le terme cubique comme une *perturbation* de l'équation (3); mais il s'agit d'une perturbation bien particulière car, pour $a = 0$, l'équation (3) n'a que deux solutions , tandis que pour $a \neq 0$, mais aussi petit qu'on veut, l'équation (4) a trois solutions; trois solutions réelles lorsque $E < \hat{V}$, deux solutions réelles pour $E = \hat{V}$ et une solution réelle lorsque $E > \hat{V}$

L'équation prend alors la forme suivante

$$0.7 = x^2 - 0.1x^3 = x^2(1 - (1/10)x) \equiv g(x) \tag{5}$$

et il est facile de se rendre compte que les racines se trouvent dans les intervalles suivants:[4] $x_1 \in (-1, 0)$, $x_2 \in (0, 1)$, $x_3 \in (9, 10)$–i.e. $x_1 = -0\ldots$, $x_2 = 0\ldots$ et $x_3 = 9\ldots$.

Notre méthode prédit que avec 5 itérations on sera sûr des deux chiffres après la virgule[5]

Après cette expérience pratique on peut résumer la situation en disant que l'on dispose d'une méthode sûre–et on peut se poser la question de l'amélioration de sa performance, i.e. on voudrait pouvoir *accélerer* la convergence vers la racine. Une méthode couramment utilisée est

[3] Démontrer ce résultat.

[4] **Exercice:** Démontrer ce résultat.

[5] **Exercice:** Vérifier cette conjecture.

la *méthode Newton-Raphson*. Ses avantages sont qu'elle converge vers la racine plus rapidement que la méthode de la dichotomie et qu'on peut la généraliser à plus d'une dimension. Son désavantage est qu'elle n'offre aucune garantie de convergence vers la racine–si on commence "près de la racine", alors elle marche très bien-mais si on est "loin" elle diverge de manière assez spectaculaire; alors on peut considérer ceci comme un avantage si on veut. Les notions de "proche" et "loin" dépendent fortement de la fonction étudiée!! Notre intuition est un guide précieux et le calcul de quelques valeurs de la fonction peut nous donner une idée assez précise de son comportement.

En définitif, la méthode se base sur la figure suivante (cf. fig. 2). On déduit

$$0 \equiv f^{'}(x_0)x_1 + (f(x_0) - f^{'}(x_0)x_0) \Rightarrow x_1 = x_0 - \frac{f(x_0)}{f^{'}(x_0)} \tag{6}$$

relation qu'on peut facilement généraliser en

$$x_{k+1} = x_k - \frac{f(x_k)}{f^{'}(x_k)}, \;\; k = 0, \ldots \tag{7}$$

On peut déjà se rendre compte d'une source d'ennuis: si la dérivée s'annulle quelque part, alors la méthode s'arrête à ce point-là (cf. fig. 2 en bas). Mais encore pire est la situation, lorsque $f^{'}(x_l) \approx 0$–alors l'erreur devient incontrôlable, mais on n'a pas, *a priori*, un message d'erreur franc–si $|f^{'}(x)| \approx \epsilon$ dans un intervalle $\delta >> \epsilon$, alors on ne peut pas avoir une bonne estimation de la racine (qui est – presque – une racine multiple, de multiplicité assez élevée). En pratique (i.e. pour les applications en physique) la première situation est, toutefois, plus fréquemment rencontrée que la deuxième.

Mais il y a un autre problème qui est lié à la convergence de l'algorithme: cette fois-ci il n'y a aucune garantie que, même lorsque $f^{'}(x) \neq 0$, la suite $\{x_k\}$ convergera, d'abord vers une limite-et que cette limite soit une racine de la fonction $f(x)$. Et il est amusant de constater que l'algorithme de Newton-Raphson, de nos jours, est intéressant autant pour les propriétés des suites qu'il génère, que pour son efficacité à trouver les racines des fonctions.

Car, maintenant qu'on a mis en évidence tous les défauts de cette méthode, on peut exposer l'autre face de la médaille: sa vitesse de convergence, *quand elle converge*. Si on note la racine par x^*, soit $\epsilon_k \equiv x_k - x^*$. On utilise, maintenant, l'hypothèse de base de cette méthode, i.e. qu'on est "suffisamment près" de la racine, pour effectuer un développement limité autour de x^* en ϵ_k; de l'éq. (7) on tire

$$\varepsilon_{k+1} = \varepsilon_k - \frac{f(x^* + \varepsilon_k)}{f^{'}(x^* + \varepsilon_k)} \tag{8}$$

$$\begin{aligned} f(x^* + \varepsilon_k) &= 0 + \varepsilon_k f^{'}(x^*) + (1/2)\varepsilon_k^2 f^{''}(x^*) + O(\varepsilon_k^3) \\ f^{'}(x^* + \varepsilon_k) &= f^{'}(x^*) + \varepsilon_k f^{''}(x^*) + (1/2)\varepsilon_k^2 f^{'''}(x^*) + O(\varepsilon_k^3) \end{aligned}$$

En remplaçant dans eq. (8) on trouve

$$\varepsilon_{k+1} = \frac{(1/2)\varepsilon_k^2 f^{''}(x^*) + O(\varepsilon_k^3)}{f^{'}(x^*) + \varepsilon_k f^{''}(x^*) + (1/2)\varepsilon_k^2 f^{'''}(x^*) + O(\varepsilon_k^3)} =$$

$$= \varepsilon_k^2 \frac{f^{''}(x^*)}{2f^{'}(x^*)} \frac{1 + O(\varepsilon_k)}{1 + O(\varepsilon_k)} = \varepsilon_k^2 \frac{f^{''}(x^*)}{2f^{'}(x^*)} \left(1 + O(\varepsilon_k)\right) \tag{9}$$

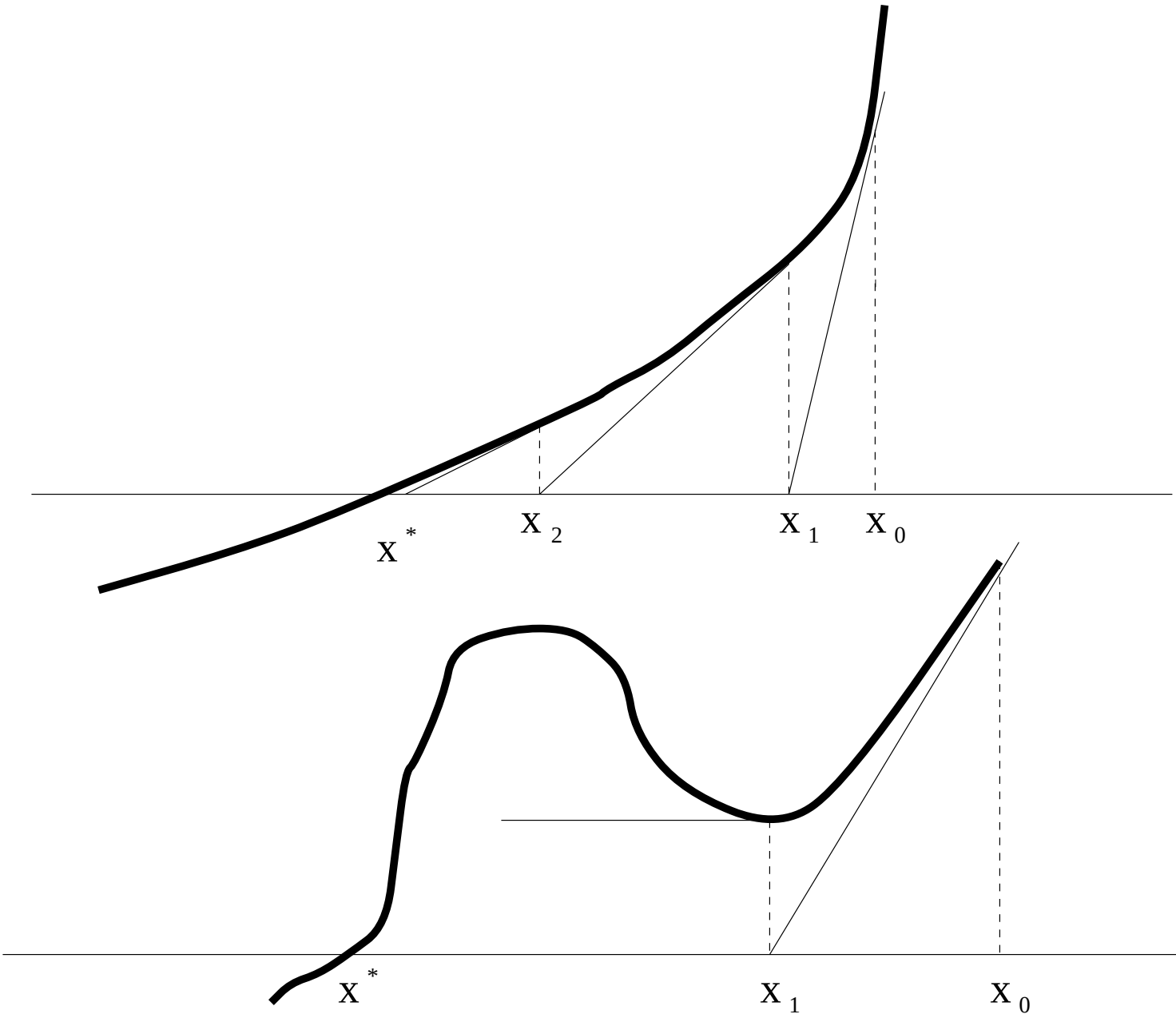


Figure 2: La méthode itérative de Newton-Raphson. x_0 est le point du départ, qui doit–déjà–être proche de la racine.

On peut comparer cette erreur avec celle de la méthode de la dichotomie, où on a trouvé que $\varepsilon_{k+1} = \varepsilon_k \times C$, C =constante et on se rend compte tout de suite que la nouvelle méthode est en effet très efficace: on peut gagner 2 chiffres par itération! La condition préalable étant qu'on soit suffisamment près de x^* pour que le développement limité ait un sens et que le coefficient numérique $f''(x^*)/(2f'(x^*))$ soit petit (i.e. $O(1)$). Bien sûr, la deuxième condition ne peut être vérifiée qu'*a posteriori*-quant à la première elle présuppose une intuition "physique", acquise, en partie, par évaluation numérique préalable. On peut, à ce moment, écrire l'algorithme de Newton-Raphson de la manière suivante:

$k = 0; x_0 = x_{initial}$
tant que $|f(x_k)| > \varepsilon$ **et** $(|x_k - x_{k-1}| > \varepsilon$ **quand** $k > 0)\{$
$\quad x_{k+1} = x_k - f(x_k)/f'(x_k);\ \ k = k + 1\}$
$x^* = x_k$

1.2 Les Zéros sont Complexes

Qu'est-ce qui se passe si on *veut* trouver des racines complexes? Aucune des deux méthodes ne semble pouvoir nous aider. Mais la méthode de Newton–Raphson se laisse généraliser de la manière suivante: on écrit

$$f(x + \varepsilon) = f(x) + \varepsilon f'(x)$$

et si on fait l'hyopthèse que $f(x + \varepsilon) \approx 0$ on obtient une équation pour "l'amélioration" ε

$$\varepsilon = -\frac{f(x)}{f'(x)}$$

et une nouvelle approximation pour la racine, *viz.*

$$x_{\text{nouvelle}} = x_{\text{ancienne}} + \varepsilon = x_{\text{ancienne}} - \frac{f(x_{\text{ancienne}})}{f'(x_{\text{ancienne}})}$$

Cette approche, algébrique, à la méthode de Newton–Raphson, se laisse facilement généraliser au cas de la recherche des zéros des fonctions de plus d'une variable réelle. Un exemple servira d'illustration.

Exemple:
Imaginons qu'on cherche les zéros (complexes) de la fonction $f(z) = z^2 + z + 1$, qui est l'équation caractéristique de l'équation différentielle

$$m\frac{d^2x}{dt^2} + \Gamma\frac{dx}{dt} + kx = 0$$

qui est l'équation du mouvement d'un oscillateur harmonique amorti (e.g. masse attachée à un ressort qui bouge dans un fluide visqueux) et on est dans le régime "sous-amorti". On écrit $z = \lambda_R + \mathrm{i}\lambda_I$ et on trouve que $f(z) = 0$ est équivalente à

$$(\lambda_R^2 - \lambda_I^2 + \lambda_R + 1) + \mathrm{i}(2\lambda_R\lambda_I + \lambda_I) = 0$$

Par construction $\lambda_R, \lambda_I \in \mathbb{R}$ et on sait que cette dernière équation implique que les parties réelle et imaginaire doivent s'annuler toutes les deux,

$$f(z) = 0 \Leftrightarrow \begin{cases} f_R(\lambda_R, \lambda_I) = 0 \\ \text{et} \\ f_I(\lambda_R, \lambda_I) = 0 \end{cases}$$

où $f_R(\lambda_R, \lambda_I) \equiv \lambda_R^2 - \lambda_I^2 + \lambda_R + 1$ et $f_I(\lambda_R, \lambda_I) \equiv 2\lambda_R\lambda_I + \lambda_I$.

Suivant l'approche algébrique à la méthode Newton-Raphson, on cherche les "corrections", $\delta\lambda_R$ et $\delta\lambda_I$ en faisant un développement limité:

$$f_R(\lambda_R + \delta\lambda_R, \lambda_I + \delta\lambda_I) = f_R(\lambda_R, \lambda_I) + \delta\lambda_R \left.\frac{\partial f_R}{\partial \lambda_R}\right|_{\lambda_R,\lambda_I} + \delta\lambda_I \left.\frac{\partial f_R}{\partial \lambda_I}\right|_{\lambda_R,\lambda_I}$$
$$f_I(\lambda_R + \delta\lambda_R, \lambda_I + \delta\lambda_I) = f_I(\lambda_R, \lambda_I) + \delta\lambda_R \left.\frac{\partial f_I}{\partial \lambda_R}\right|_{\lambda_R,\lambda_I} + \delta\lambda_I \left.\frac{\partial f_I}{\partial \lambda_I}\right|_{\lambda_R,\lambda_I}$$

On pose les membres de gauche égaux à zéro et on se trouve devant un système linéaire de deux équations pour les deux inconnus, $\delta\lambda_R$ et $\delta\lambda_I$. Il est facile de résoudre ce système–surtout si on l'écrit sous forme matricielle

$$\begin{pmatrix} -f_R(\lambda_R, \lambda_I) \\ -f_I(\lambda_R, \lambda_I) \end{pmatrix} = \begin{pmatrix} \partial f_R/\partial\lambda_R & \partial f_R/\partial\lambda_I \\ \partial f_I/\partial\lambda_R & \partial f_I/\partial\lambda_I \end{pmatrix} \begin{pmatrix} \delta\lambda_R \\ \delta\lambda_I \end{pmatrix}$$

Il est facile de trouver l'inverse d'une matrice 2×2,

$$\frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

ce qui conduit à la solution pour $\delta\lambda_R$ et $\delta\lambda_I$

$$\begin{pmatrix} \delta\lambda_R \\ \delta\lambda_I \end{pmatrix} = \frac{1}{(2\lambda_R + 1)^2 + 4\lambda_I^2} \begin{pmatrix} 2\lambda_R + 1 & -2\lambda_I \\ 2\lambda_I & 2\lambda_R + 1 \end{pmatrix} \begin{pmatrix} \lambda_I^2 - \lambda_R^2 - \lambda_R - 1 \\ -2\lambda_R\lambda_I - \lambda_I \end{pmatrix}$$

On ajoute ce vecteur au vecteur (λ_R, λ_I) et on obtient la nouvelle approximation pour la partie réelle et la partie imaginaire de la racine. On continue ainsi jusqu'à ce que

- $\sqrt{(\delta\lambda_R)^2 + (\delta\lambda_I)^2} < \varepsilon$
- $|f_R(\lambda_R, \lambda_I)| < \varepsilon$
- $|f_I(\lambda_R, \lambda_I)| < \varepsilon$

Toutes les trois conditions doivent être remplies–comme pour le cas d'un zéro réel, il faut que la correction soit plus petite que la précision, et que la valeur des fonctions, qui sont sensées s'annuler à ce point, soit effectivement zéro dans la précision, sous laquelle on travaille.

Comment choisir le point du départ? Ici on voit que les calculs se simplifient énormement si on choisit $\lambda_R^{(0)} = -1/2$. Pour $\lambda_I^{(0)}$ on peut choisir la valeur $\pm 1/2$ aussi, car on note la combinaison $2\lambda_I$ dans les éléments de la matrice. Il est vivement conseillé de faire un dessin du plan (λ_R, λ_I) et de suivre le mouvement des points. Qu'est-ce qui se passe si on choisit d'autres points de départ?

Remarque 1: Du point de vue de la géométrie, les racines qu'on cherche sont les points d'intersection des courbes de niveau 0

$$f_R(\lambda_R, \lambda_I) = 0, \quad f_I(\lambda_R, \lambda_I) = 0$$

Dessiner ces courbes.

Remarque 2: On note que la recherche des zéros complexes est un exemple particulièrement simple de la recherche de la solution d'un système d'équations non-linéaires.

Un exemple plus concret (et plus pratique) est fourni par la recherche du point d'équilibre d'un point matériel, qui bouge sur un plan. Soit son énergie potentielle

$$\mathcal{V}(x, y) = \frac{1}{2}k(x^2 + y^2) - a(x^2 + y^2)^2 - F_x \cdot x - F_y \cdot y \tag{10}$$

Calculer son point d'équuilibre numériquement, en prenant des valeurs numériques de votre choix pour les paramètres.

1.3 Conclusions

Dans cette section on a exposé les méthodes couramment utilisées pour calculer numériquement les zéros d'une fonction. Chacune a des points forts, qu'on doit exploiter le plus possible, ainsi que des points faibles, dont on doit tenir compte, pour que les résultats obtenus aient un sens. La méthode de la dichotomie (ou bissection) est sûre, mais lente; Newton-Raphson est plus rapide, mais dangéreuse, lorsque nos connaissances sur la fonction sont trop limitées; en outre elle n'est compétitive que si l'on peut calculer la dérivée de la fonction avec autant de précision que la fonction elle-même; mais elle est la seule à être généralisable aux racines complexes, ainsi qu'à la recherche des solutions d'un système d'équations non-linéaires.

2 Intégration Numérique

Une différence assez amusante entre le calcul des dérivées et celui des intégrales est qu'il y a des intégrales qu'on ne peut pas calculer analytiquement–tandis qu'il n'y a pas de barrière *théorique* qui nous empêche de calculer la dérivée d'une fonction (ou de décider qu'elle n'existe pas).

Quelques exemples:

$$\int e^{-x^2} dx, \quad \int \frac{e^{-x}}{x} dx$$

Puis, comme on sait, la réprésentation *analytique* n'est pas nécessairement la mieux adaptée pour les calculs *numériques*.

L'exemple physique le plus simple est, encore une fois, celui du mouvement d'une particule dans un potentiel, à une dimension d'espace. Comme on a vu dans la section précédente, une fois l'énergie donnée on dispose de toutes les informations nécessaires pour calculer les bornes de la trajectoire–maintenant on va aborder le problème dynamique du calcul du temps d'aller d'un point à un autre.

La solution de ce problème, qui est un problème au coeur de la physique, consiste, en principe, à intégrer les équations différentielles du mouvement. Les méthodes numériques pour

ça sont l'objet de la prochaine section. A une dimension on peut faire mieux, lorsque l'énergie est conservée– on peut se ramener à une équation différentielle ordinaire du premier ordre, alors au calcul d'une intégrale. Cette intégrale ne peut se calculer analytiquement que dans très peu de cas, par conséquent une approche numérique nous est imposée.

On se souvient que la conservation de l'énergie, à une dimension, avec un degré de liberté,

$$E = \frac{1}{2}m\left(\frac{dx}{dt}\right)^2 + V(x) = const. \tag{11}$$

conduit à l'équation de Newton,

$$m\frac{d^2x}{dx^2} = -\frac{dV}{dx} \tag{12}$$

i.e. une équation différentielle ordinaire du deuxième ordre–on doit préciser vitesse initiale et position initiale. Mais on voit immédiatement qu'il est possible de résoudre eq. (11) pour la vitesse, i.e. de se ramener à l'équation différentielle du *premier ordre*

$$\frac{dx}{dt} = \sqrt{\frac{2}{m}\left(E - V(x)\right)} \tag{13}$$

avec comme condition initiale $x(0) = x_0$-effectivement on a remplacé la condition sur la vitesse par l'énergie. Mais cette équation est immédiatement soluble

$$t = \int_{x_0}^{x} \frac{dz}{\sqrt{\frac{2}{m}\left(E - V(z)\right)}} \tag{14}$$

qui donne la trajectoire $x(t)$ par inversion. On remarque, alors, qu'on peut traiter l'évaluation des intégrales et la résolution des équations différentielles ordinaires d'une manière unifiée; et c'est une question de tactique numérique laquelle des deux approches est mieux adaptée au problème particulier.

En ce qui concerne l'évaluation de l'intégrale, un point de départ est, bien sûr, la définition mathématique[6]

$$\int_a^b f(x)dx \approx \sum_{i=0}^{N+1} f(x_i)\delta x_i \tag{15}$$

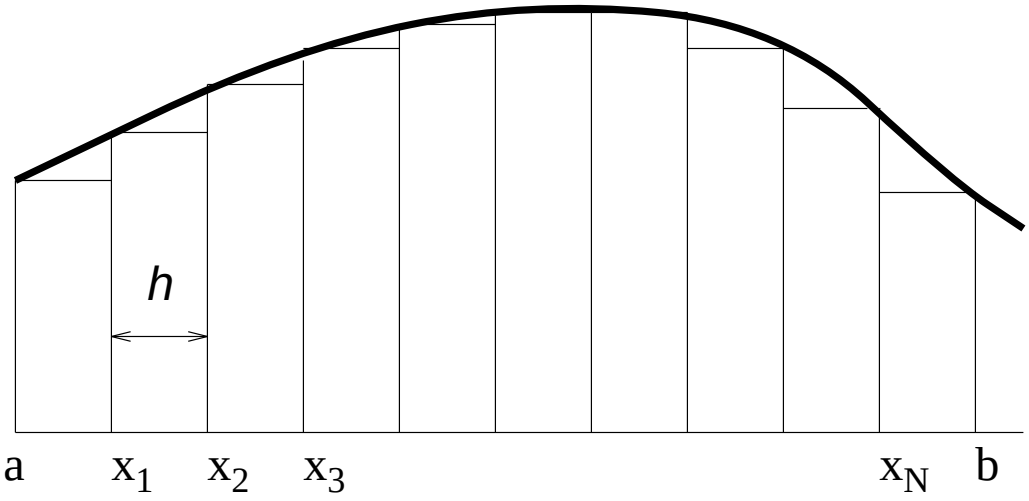
cf. fig. 2. Le problème est que la convergence de la somme vers l'intégrale est très lente et donne lieu à une erreur appréciable–cf. fig. 2 en bas. Une meilleure approximation consiste alors de prendre des trapèzes au lieu des parallélogrammes droits

$$\int_a^b f(x)dx \approx h\left(\frac{1}{2}f(a) + f(x_1) + \cdots + f(x_N) + \frac{1}{2}f(b)\right) \tag{16}$$

Exemple:
Un exemple très simple: soit la fonction $f(x) = Ax + B$. Si on essaie de calculer son intégrale entre $x = a$ et $x = b$ d'après la définition de Riemann, en utilisant un pas uniforme h on trouve

$$\int_a^b f(x)dx \equiv \lim_{N\to\infty} h\sum_{k=0}^{N-1} f(x_k)$$

[6] On se limite à l'intégrale d'après B. Riemann ici.



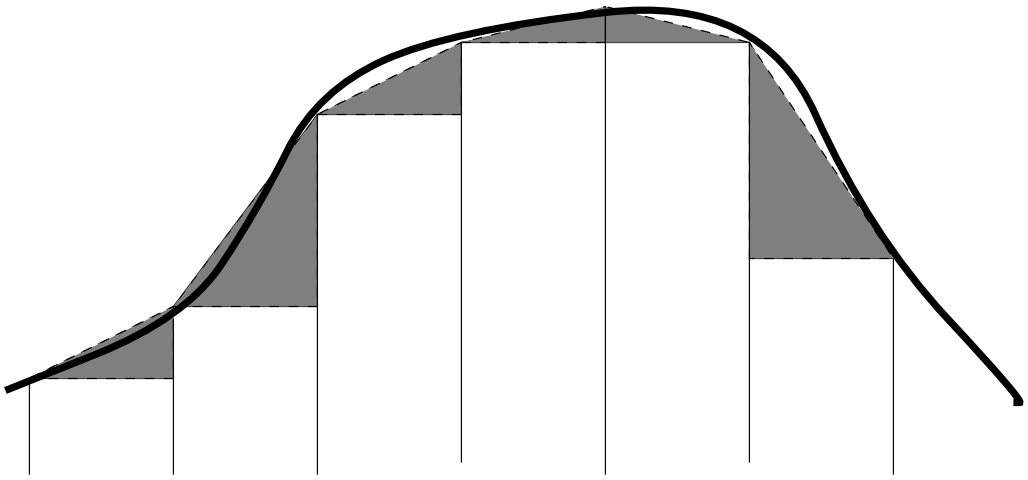


Figure 3:

où $x_k = a + kh$ et $h = (b-a)/N$. Pour N fini on trouve

$$h\sum_{k=0}^{N-1} Ax + B = h\left(BN + Aa(N-1) + Ah\sum_{k=0}^{N-1} k\right) =$$
$$h\left((B+Aa)(N-1) + B + AhN(N-1)/2\right) =$$
$$B(b-a) + (A/2)(b^2-a^2) + (A/2N)(b-a)^2$$

On réconnaît immédiatement le résultat exact– et on se rend compte que l'erreur qu'on commet est $O(1/N)$. Si on utilise la méthode des trapèzes on trouve

$$h\left[\tfrac{1}{2}(Aa+B) + Ax_1 + B + Ax_2 + B + \cdots + Ax_{N-1} + B + \tfrac{1}{2}(Ab+B)\right] =$$
$$h\left[(A/2)(a+b) + NB + Aa(N-1) + Ah\sum_{k=0}^{N-1} k\right] =$$
$$h\left[(A/2)(a+b) + NB + Aa(N-1) + AhN(N-1)/2\right] =$$
$$\tfrac{b-a}{N}\left[(A/2)(b+a) + NB + Aa(N-1) + A(b-a)(N-1)/2\right] =$$
$$\tfrac{b-a}{N}\left[(A/2)(b+a)N + BN\right]$$

et on réconnaît le résultat *exact*–mais on voit qu'il n'y a pas de terme d'erreur–même pour N fini! Comment peut-on comprendre ça? En fait, les deux formules précédentes peuvent s'unifier par le raisonnement suivant: la première (la "définition") fait l'hypothèse que la fonction soit *constante* par morceaux, tandis que l'approximation des trapèzes fait l'hypothèse que la fonction soit *linéaire* par morceaux–or la fonction, qu'on a choisie, est bien linéaire et n'est pas constante dans la mesure où $A \neq 0$; d'ailleurs on voit que le terme d'erreur est indépendent de B et s'annule si $A = 0$.

Evidemment on peut continuer dans cette direction et arriver, en particulier, à la *formule de Simpson*, qui fait l'hypothèse que la fonction est quadratique par morceaux (i.e. peut être approchée par un polynôme de degré 2 dans un sous-intervalle donné).

Concrètement:

Soit l'intervalle $[a,b]$ et son milieu, $m \equiv (a+b)/2$. Il existe un, et un seul, polynôme de degré 2, $P(x)$, qui prend les valeurs $f(a)$, $f(m)$ et $f(b)$ aux points a, m et b respectivement. Pour le trouver on pose $P(x) \equiv Ax^2 + Bx + C$ et on impose les conditions $P(a) = f(a)$, $P(m) = f(m)$ et $P(b) = f(b)$. On obtient un système linéaire de trois équations pour les trois variables A, B et C. Pour ce polynôme on a

$$\int_a^b P(x)dx = h\left(\frac{1}{3}P(a) + \frac{4}{3}P(m) + \frac{1}{3}P(b)\right) \tag{17}$$

où $h \equiv (b-a)/2$. La démonstration consiste à remplacer $P(x)$ par $Ax^2 + Bx + C$ et à vérifier qu'on obtient une identité. Une méthode beaucoup plus rapide (et générale) est la suivante:

On sait que

- Tout polynôme $P(x)$, de degré 2 peut être écrit comme combinaison linéaire des monômes x^2, x et 1.

  Démonstration: Immédiate–$P(x) = A \cdot x^2 + B \cdot x + C \cdot 1$ pour A, B, C constantes arbitraires.

- L'intégration est une opération *linéaire*:

  Démonstration:

$$\int_a^b \left(c_1 f(x) + c_2 g(x)\right) dx = c_1 \int_a^b f(x)dx + c_2 \int_a^b g(x)dx \tag{18}$$

- Il suffit de savoir intégrer dans l'intervalle $[0,1]$.

  Démonstration:

$$\int_a^b f(x)dx = \int_0^{b-a} f(x+a)d(x+a) = (b-a)\int_0^1 f(\frac{x+a}{b-a}(b-a))d\frac{x+a}{b-a} \tag{19}$$

Par conséquent, on peut écrire

$$\begin{aligned}
\int_0^1 x^2dx &= \frac{1}{2}\left(\alpha\cdot 0^2 + \beta\cdot\left(\frac{1}{2}\right)^2 + \gamma\cdot 1^2\right) \\
\int_0^1 xdx &= \frac{1}{2}\left(\alpha\cdot 0^1 + \beta\cdot\left(\frac{1}{2}\right)^1 + \gamma\cdot 1^1\right) \\
\int_0^1 dx &= \frac{1}{2}\left(\alpha\cdot 1 + \beta\cdot 1 + \gamma\cdot 1\right) \qquad (20) \\
& \qquad (21)
\end{aligned}$$

et il est facile de résoudre ce système pour les coefficients $\alpha = 1/3$, $\beta = 4/3$ et $\gamma = 1/3$. La propriété fondamentale de ces coefficients est qu'ils sont *indépendants* du polynôme du deuxième degré qu'on est en train d'intégrer ainsi que de l'intervalle $[a,b]$ qui nous intéresse–ils dépendent *uniquement* du fait qu'on s'est limité à un polynôme de degré au plus 2.

Conclusion:

Pour approximer l'intégrale d'une fonction par la méthode de Simpson, on remplace la fonction par le polynôme de degré 2 qui passe par les mêmes valeurs que la fonction aux deux bouts de l'intervalle, ainsi qu'au milieu et on calcule l'intégrale de ce polynôme exactement.

Exemple:

$$I = \int_0^{\pi/2} \sin x dx$$

On va calculer cette intégrale par trois méthodes: (a) exactement, (b) en calculant explicitement le polynôme $P(x)$ et intégrant ce polynôme et (c) en applicant la formule de Simpson.

(a) Il est facile de voir que $I = -\cos(\pi/2) + \cos(0) = 1$.

(b) On pose $P(x) = Ax^2 + Bx + C$. Les A, B et C seront solution du système

$$\begin{gathered}
C = f(0) = 0 \\
A\pi^2/16 + B\pi/4 + C = 1/\sqrt{2} \\
A\pi^2/4 + B\pi/2 + C = 1
\end{gathered}$$

On trouve $A = -(8/pi^2)(\sqrt{2}-1)$,$B = (8/\pi)(1/\sqrt{2} - 1/4)$. Si on remplace $\sin x$ par $P(x)$ dans I et on fait l'intégrale on trouve

$$I_{\text{poly}} = \int_0^{\pi/2} P(x)dx = A\frac{\pi^3}{24} + B\frac{\pi^2}{8} = \frac{\pi}{4}\frac{2\sqrt{2}+1}{3} = 1.00228$$

(c) On calcule I par la formule de Simpson

$$I_{\text{Simpson}} = \frac{1}{2}\frac{\pi}{2}\left[\frac{1}{3}\sin(0) + \frac{4}{3}\sin(\pi/4) + \frac{1}{3}\sin(\pi/2)\right] = \frac{\pi}{4}\left[\frac{4}{3\sqrt{2}} + \frac{1}{3}\right] = \frac{\pi}{4}\frac{2\sqrt{2}+1}{3} = 1.00228$$

On trouve effectivement le même résultat. Qu'est ce qu'il serait arrivé si, au lieu de prendre comme polynôme celui qui passe exactement par 0, $1/\sqrt{2}$ et 1 aux points 0, $\pi/4$ et ${}_pi/2$, on avait pris comme polynôme le développement limité de $\sin x$ autour du point $x = 0$ jusqu'à l'ordre 2? On sait bien que $\sin x = x + O(x^3)$ et le polynôme est du premier ordre (car sinon on n'utilise plus la méthode de Simpson!). On trouve

$$I_{\text{Taylor}} = \int_0^{\pi/2} x dx = \frac{\pi^2}{8} = \frac{\pi}{4}\frac{\pi}{2} = 1.2337$$

qui est un résultat nettement inférieur en précision. Mais ici on connaît le résultat exact. Peut-on affirmer que le polynôme $P(x)$ donnera *toujours* un résultat meilleur que le développement limité de Taylor à l'ordre 2? Intuitivement on pourrait penser que oui, car $P(x)$ est en accord non seulement au point $x = a$ (comme est le développement limité) mais aussi aux points $x = m$ et $x = b$. Mais on pourrait avancer qu'il n'est pas exclu que l'erreur commise par $P(x)$ ailleurs "compense" l'accord obtenu et donne un résultat pire...Ce qui est clair est que pour une fonction *quadratique* $P(x)$ coincide avec le développement de Taylor, alors on aura le résultat *exact*.

Exemple:

Essayons alors de voir ce qui se passe si on intègre la fonction $f(x) = \cos x$ car on sait que le développement limité à l'ordre 2 de $\cos x$ autour de $x = 0$ est $\cos x = 1 - x^2/2 + O(x^4)$.

(a) Le résultat exact est

$$I = \int_0^{\pi/2} \cos x dx = \sin(\pi/2) = 1$$

(b) Le polynôme $P(x) = Ax^2 + Bx + C$ a comme coefficients $A = (8/\pi^2)(\sqrt{2} - 1)$, $B = (2/\pi)(4/\sqrt{2} - 3)$ et $C = 1$. Si on l'intègre on trouve

$$I_{\text{poly}} = \int_0^{\pi/2} P(x) dx = A(\pi^3/24) + B(\pi^2/8) + C(\pi/2) = (\pi/12)(1 + 2\sqrt{2}) = 1.00228$$

(c) La méthode de Simpson donne directement

$$I_{\text{Simpson}} = \frac{\pi}{4}\left[\frac{1}{3} + \frac{4}{3}\frac{1}{\sqrt{2}}\right] = (\pi/12)(1 + 2\sqrt{2}) = 1.00228$$

(d) Enfin, si on remplace la fonction $\cos x$ par son développement limité autour de $x = 0$ on trouve

$$I_{\text{Taylor}} = \int_0^{\pi/2} \left(1 - \frac{x^2}{2}\right) dx = (\pi/2)\left(1 - \frac{\pi^2}{24}\right) = 0.9248$$

Ces résultats indiquent que, même pour une fonction comme le cosinus, le polynôme $P(x)$ offre une meilleure approximation sur un intervalle de valeurs plus large que le développement de Taylor. Evidemment ces exemples ne remplacent pas une preuve! Mais elles suggèrent que les contrexemples seront vraiment non-triviaux.

On peut généraliser cette formule pour N points intermédiaires (au lieu d'un point–attention que le nombre des points intermédiaires doit être impaire pour que la formule ait un sens)en

l'appliquant dans chaque sous-intervalle séparemment–on trouve alors

$$\int_a^b f(x)dx = h\left(\frac{1}{3}f(a) + \frac{4}{3}\left(f(x_1) + f(x_3) + \cdots + f(x_N)\right) + \frac{2}{3}\left(f(x_2) + f(x_4) + \cdots + f(x_{N-1})\right) + \frac{1}{3}f(b)\right) \tag{22}$$

où, cette fois-ci, $h \equiv (b-a)/(N+1)$.

Il est clair qu'on peut continuer dans cette voie et établir des formules analogues, qui soient exactes pour des polynômes de degré supérieur à 2.

Enfin il faut dire que l'utilité pratique de la méthode et de Simpson ne réside pas moins dans le fait qu'on peut trouver des transformations entre l'intégrale originale et l'intégrale d'une fonction qui peut se calculer exactement, ou avec une erreur plus petite que l'originale. Quelques exemples serviront d'illustration:

1. $$I = \int_0^{\pi/2} \sin 3x\, dx$$

   L'intégrande $(\sin 3x)$ n'est pas un polynôme en x–par conséquent ni la méthode de trapèzes, ni celle de Simpson peut être appliquée. Néanmoins, on peut trouver un changement de variables qui ramenera l'intégrale à une forme, pour laquelle une (au moins) de ces méthodes nous donnera le résultat exact.

   Démonstration:

   (a) $\sin 3x\, dx = -\frac{1}{3} d\cos 3x$; posons, alors, $u = \cos 3x$. On obtient

   $$I = \int_1^0 -du = \int_0^1 du = 1$$

   Une constante est une fonction intégrable exactement soit par la méthode des trapèzes, soit par celle de Simpson,

   $$I = 1 \times \left[\frac{1}{2} + \frac{1}{2}\right] = \frac{1}{2} \times \left[\frac{1}{3} + \frac{4}{3} + \frac{1}{3}\right] \tag{23}$$

   (b) $\sin 3x = \sin(2x + x) = \sin 2x \cos x + \cos 2x \sin x = \sin x(4\cos^2 x - 1)$; l'intégrande peut, donc, s'écrire comme $\sin 3x$, $dx = -(4\cos^2 x - 1)\, d\cos x$. On pose alors $u = \cos x$ et on obtient

   $$I = \int_1^0 -(4u^2 - 1)\, du = \int_0^1 (4u^2 - 1)\, du$$

   L'intégrande est manifestement un polynôme de deuxième degré dans la variable u–par conséquence la méthode de Simpson nous donnera le résultat éxact, à savoir

   $$I = \frac{1}{2}\left[\frac{1}{3}(-1) + \frac{4}{3}(0) + \frac{1}{3}(3)\right] = 1/3$$

2. $$I = \int_{100}^{100+a} \frac{dx}{1+x^6}$$

avec $0 < a << 100$. On va voir que la méthode des trapèzes peut être utilisée ici.

Démonstration:

$$\frac{1}{1+100^6} - \frac{1}{1+(100+a)^6} = \frac{6}{1+100^6}\frac{a}{100} + O((\frac{a}{100})^2)$$

obtenu en utilisant

$$(1+y)^n - 1 = ny + O(y^2), \quad \frac{1}{1+y} = 1 - y + O(y^2)$$

Notre résultat implique que la fonction $f(x) = 1/(1+x^6)$ peut être approchée par une fonction linéaire lorsque $x >> 1$ et la largeur de l'intervalle d'intégration est $<< x$ (i.e. l'intervalle peut être plus grand que 1).

Par contre, très près de l'origine les choses changent. La raison est que l'origine est un maximum de $f(x)$, par consequent, près de l'origine

$$f(x) = 1 - x^6 + O(x^{12})$$

ce qui implique que *très* près de l'origine la fonction est plate et la méthode de trapèzes ou Simpson sont satisfaisantes–mais, un peu plus loin, on commence à ressentir de la courbure; on passe par un point d'inflection–ce qui implique que, dans l'intervalle [0.8,1] la fonction peut être approchée par une droite-plus loin on ressent de nouveau une forte courbure, jusqu'à ce que la courbe s'aplatisse, cf. fig 2.

En ce qui concerne l'implémentation algorithmique les choses sont très simples–la méthode des trapèzes ne demande que l'évaluation d'une somme, sur tous les points intermédiaires, avec le même coefficient, i.e. 1, tandis que la formule de Simpson a deux coefficients–pour les points paires et impaires. Le problème de la qualité de l'approximation est assez non-trivial, mais, à une dimension, on peut contrôler la situation assez bien. Aussi on a pu se permettre de prendre le pas d'intégration, h, uniforme. En plus d'une dimensions on ne peut plus se permettre cette luxe et on est obligé de prendre en compte des renseignements sur les régions, où l'intégrand varie le plus vite (i.e. pôles) et d'autres techniques, telles *l'approche multi-grille* et *la méthode Monte Carlo* deviennent compétitives.

Un dernier problème pratique concerne les *singularités*. Il y en a de deux types–intégrables et non-intégrables. Aucune astuce numérique ne peut aider dans le second cas. En ce qui concerne le premier cas, un changement de variables opportun peut toujours être trouvé.

Un exemple typique est le calcul de la période du mouvement d'un point matériel dans un potentiel–alors on a

$$T \equiv 2t(x_1 \to x_2) = \sqrt{\frac{m}{2}} \int_{x_1}^{x_2} \frac{dx}{\sqrt{E - V(x)}} \tag{24}$$

où $E = V(x_1)$, $E = V(x_2)$. On voit immédiatement qu'on ne peut pas appliquer la méthode de Simpson (ni celle des trapèzes) immédiatement, car l'intégrande est infinie aux points x_1 et x_2. Pourtant, on sait que, si ces points ne sont pas des extréma du potentiel $V(x)$, l'intégrale *doit converger*–la période doit être finie; par contre si un (au moins) de ces points est aussi un extrémum du potentiel $V(x)$ on s'attend à ce que l'intégrale diverge. Concrétement, si on prend pour $V(x)$ une forme polynômiale, on peut réaliser la construction suivante:

$$E - V(x) \equiv \mathcal{P}(x) = (x - x_1)P(x) = (x_2 - x)Q(x) \tag{25}$$

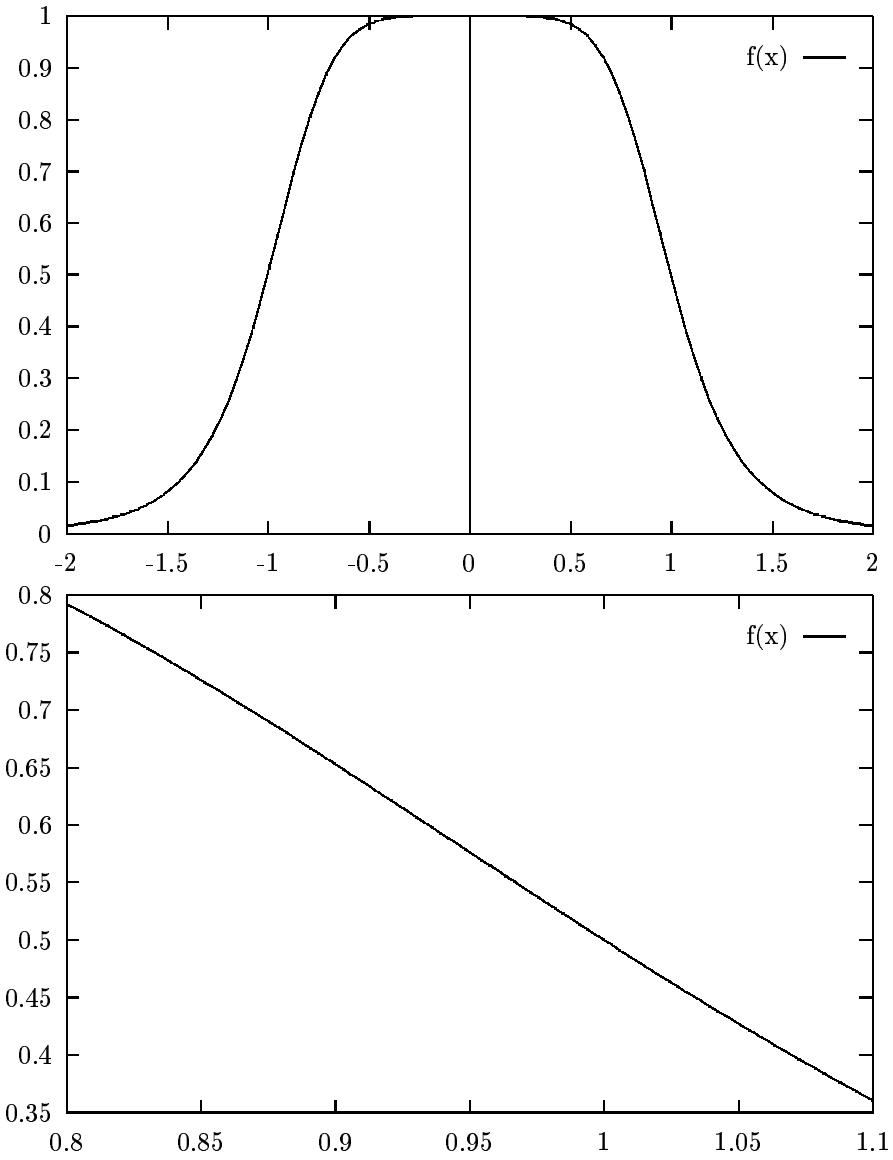


Figure 4: La courbe $f(x) = 1/(1+x^6)$. Détail (en bas).

où $P(x)$ et $Q(x)$, par hypothèse, satisfont à $P(x_1) \neq 0$ et $Q(x_2) \neq 0$. Soit, maintenant, y un point de l'intervalle (x_1, x_2) avec la propriété[7] $P(y) \neq 0$ *et* $Q(y) \neq 0$. Alors on peut écrire

$$T = 2\sqrt{\frac{m}{2}} \left(\int_{x_1}^{y} \frac{dx}{\sqrt{\mathcal{P}(x)}} + \int_{y}^{x_2} \frac{dx}{\sqrt{\mathcal{P}(x)}} \right) \tag{26}$$

et on utilise la décomposition $\mathcal{P}(x) = (x - x_1)P(x)$ dans l'intégrale sur l'intervalle $[x_1, y]$ et $\mathcal{P}(x) = (x_2 - x)Q(x)$ dans celle sur l'intervalle $[y, x_2]$. Alors on note qu'on peut écrire

$$\int_{x_1}^{y} \frac{dx}{\sqrt{(x - x_1)P(x)}} = \int_{x_1}^{y} \frac{d\left[2\sqrt{x - x_1}\right]}{\sqrt{P(x)}} = \int_{0}^{2\sqrt{y - x_1}} \frac{du}{\sqrt{P(x_1 + \frac{u^2}{4})}} \tag{27}$$

(28)

et la dernière intégrale ne présente aucun problème désormais. De la même manière on trouve

$$\int_{y}^{x_2} \frac{dx}{\sqrt{(x_2 - x)Q(x)}} = \int_{y}^{x_2} \frac{d\left[-2\sqrt{x_2 - x}\right]}{\sqrt{Q(x)}} = \int_{-2\sqrt{x_2 - y}}^{0} \frac{dv}{\sqrt{Q(x_2 - \frac{v^2}{4})}} \tag{29}$$

Un point très intéressant est le suivant: si on pouvait faire les deux intégrales analytiquement le résultat ne dépendrait pas du choix du point intermédiaire y. Mais si on calcule ces intégrales par une méthode approchée, telle la méthode des trapèzes ou la méthode de Simpson, le résultat dépendra, en général, du point y–alors on peut se poser la question si c'est possible de choisir y de manière à récupérer certaines propriétés qu'on sait d'avance qu'elles sont vraies. Un exemple montrera de quoi il s'agit.

Exemple:
Soit une particule de masse m qui bouge sous l'influence du potentiel $V(x) = kx^2$ et a énergie totale E. On veut calculer la période par la méthode des trapèzes. Montrer que, si on fait l'intégrale analytiquement, on trouve que la période ne dépend pas de l'énergie totale, ni de si on a coupé l'intervalle d'intégration en morceaux.

Monrtrer aussi que, si on fait le calcul par la méthode des trapèzes, on trouve un résultat qui dépend de l'énergie totale ainsi que du point intermédiaire y mais qu'on peut *toujours* choisir y de la sorte que le résultat ne dépende pas de l'énergie totale.

[7] **Exercice:** Pourquoi peut-on être assuré qu'un tel point doit toujours exister?

2.1 Conclusions

L'évaluation numérique d'une intégrale a deux degrés de liberté: l'ensemble des points, sur lequel on évalue l'intégrand et les poids associés aux points. Dans la formule inspirée de la définition classique, la distribution des points intermédiaires est uniforme et les coefficients sont tous égaux à 1. Dans la méthode des trapèzes et celle de Simpson, la distribution des points reste toujours uniforme mais les coefficients changent. On peut imaginer qu'on puisse changer aussi la distribution des points-cette idée est à la base de la méthode de Gaußet de Legendre, qui s'inspire des propriétés des polynômes orthogonaux.

Dans des problèmes concrets le choix des points intermédiaires ne peut pas être innocent. On doit les choisir de manière à ce que le résultat numérique conserve, le plus possible, les propriétés *qualitatives* du résultat "exact" qu'on cherche à atteindre.

3 Résolution Numérique des Equations Différentielles Ordinaires

Les équations différentielles jouent un rôle fondamental dans la déscription de la nature depuis Newton. Mais écrire une équation différentielle et la résoudre sont deux exercices assez distincts. Pendant longtemps on ne disposait que d'outils analytiques, ce qui limitait de manière assez marquant le champ de recherche. L'invention de méthodes numériques a libéré la physique, la chimie, l abiologie, etc... de la contrainte de formuler des modèles trop simplifiés pour pouvoir vraiment décrire la complexité des phénomènes naturelles. Le but de cette section est d'exposer les méthodes couramment utilisées de nos jours pour résoudre les équations différentielles ordinaires, qu'on rencontre dans des applications en physique (ou chimie, biologie...on rencontre souvent les mêmes équations dans des contextes totalement différents). Ainsi on pourra affirmer qu'avec ces méthodes on pourra comprendre la Mécanique classique d'un degré de liberté à 1,2 ou 3 dimensions d'espace.

On exposera les méthodes dans un cadre général–soit x la variable dépendente, t la variable indépendente. Alors on cherche la solution du problème suivant

$$\frac{dx}{dt} = f(x(t); t) \tag{30}$$

avec comme condition "initiale" $x(0) = x_0$. Si l'équation est de degré supérieur à un (comme sont les équations de mouvement en Mécanique) , on se ramène à un *système* d'équations, où l'algèbre linéaire sera utile, surtout lorsque la dimension de l'espace est plus grande que 1; à une dimension on n'a pas besoin d'outils si puissants-on écrit tout simplement

$$\begin{array}{rcl} \frac{dx}{dt} & = & v \\ \frac{dv}{dt} & = & \frac{1}{m} f(x(t), t) \end{array} \tag{31}$$

avec comme conditions initiales $x(0) = x_0$ et $v(0) = v_0$.

La idée derrière toutes les méthodes numériques est de discrétiser le temps (la variable indépendente) et remplacer les dérivées par des différences finies

$$\frac{dx}{dt} \to \frac{\Delta x}{\Delta t} \equiv \frac{x(t+h) - x(t)}{h} \tag{32}$$

Cette approche conduit tout naturellement à la *méthode d'Euler*

$$x(t+h) = x(t) + hf(x(t),t) \tag{33}$$

pour l'eq. (30) et

$$\begin{array}{c} x(t+h) = x(t) + hv(t) \\ v(t+h) = v(t) + hf(x(t),t) \end{array} \tag{34}$$

pour l'éq. (31). On voit tout de suite que ces équations ne représentent que le premier terme des développements limités dans le "temps" autour du point $t = 0$; i.e. la précision de la méthode d'Euler est $O(h^2)$. On peut maintenant poser la question, s'il est possible d'éliminer le terme d'$O(h^2)$ et avoir ainsi une méthode dont l'erreur serait $O(h^3)$ au moins. Ceci est obtenu de la manière suivante:

$$\begin{array}{rcl} x(t+h) & = & x(t) + hf(x(t),t) + (h^2/2)f'(x(t),t) + (h^3/6)f''(x(t),t) + \cdots \\ x(t-h) & = & x(t) - hf(x(t),t) + (h^2/2)f'(x(t),t) - (h^3/6)f''(x(t),t) + \cdots \end{array} \tag{35}$$

et on voit tout de suite qu'on a

$$x(t+h) = x(t-h) + 2hf(x(t),t) + O(h^3) \tag{36}$$

i.e. on fait un pas en arrière à $t = 0$, pour calculer $x(-h)$ avec la méthode d'Euler standard, après quoi on calcule $x(h), x(2h), \ldots$ avec la nouvelle méthode.

Exemple: $x' = -x$, $x(0) = 1$, $h = 0.01$.

t	Euler	Pt. du milieu	Résultat exact
1.00000×10^{-2}	0.990000	0.990000	0.990050
2.00000×10^{-2}	0.980100	0.980200	0.980199
3.00000×10^{-2}	0.970299	0.970396	0.970446
4.00000×10^{-2}	0.960596	0.960792	0.960789
5.00000×10^{-2}	0.950990	0.951180	0.951229
6.00000×10^{-2}	0.941480	0.941768	0.941765
7.00000×10^{-2}	0.932065	0.932345	0.932394
8.00000×10^{-2}	0.922745	0.923122	0.923116
9.00000×10^{-2}	0.913517	0.913882	0.913931
1.00000×10^{-1}	0.904382	0.904844	0.904837

L'algorithme pour Euler est immédiat; pour le point du milieu, il faut faire attention au fait qu'il s'agit d'une récurrence de deuxième ordre.

En fait il existe une manière plus intuitive de formuler la méthode du point du milieu, en travaillant dans l'intervalle $[t, t+h]$, au lieu de l'intervalle $[t-h, t+h]$, comme on a fait plus haut, i.e.[8]

$$\begin{array}{rcl} k_1 & = & hf(x(t),t) \\ k_2 & = & hf(x(t) + \frac{k_1}{2}, t + h/2) \\ x(t+h) & = & x(t) + k_2 \end{array} \tag{37}$$

L'idée est que la méthode d'Euler considère que la pente dans l'intervalle $[t, t+h]$ est la pente à t, c.à.d. $f(x(t);t)$. La méthode du point du milieu considère que la pente dans l'intervalle

[8] **Exercice:** Démontrez que l'éq. (37) est, effectivement, équivalente à l'éq. (36).

$[t, t+h]$ est la pente...au point du milieu, c.à.d. $f(x(t+(h/2)); t+(h/2))$. Dans l'exemple ci-dessus il est possible de voir par un calcul explicite que

$$x(t+h) = x(t) + hx'(t) = x(t) - hx(t) = (1-h)x(t)$$

par la méthode d'Euler et

$$x(t+h) = x(t) + hx'(t+(h/2)) = x(t) - h(x(t) + (h/2)x'(t)) = \\ x(t) - h(x(t) - (h/2)x(t)) = (1 - h + h^2/2)x(t)$$

par la méthode du point dyu milieu.

Il est possible d'aller encore plus loin et obtenir une erreur d'$O(h^5)$–c'est *la méthode de Runge et Kutta d'ordre 4* qui prend la forme suivante[9]

$$\begin{array}{rcc} k_1 & = & hf(x(t), t) \\ k_2 & = & hf(x(t) + k_1/2, t + h/2) \\ k_3 & = & hf(x(t) + k_2/2, t + h/2) \\ k_4 & = & hf(x(t) + k_3, t + h) \\ x(t+h) & = & x(t) + k_1/6 + k_2/3 + k_3/3 + k_4/6 + O(h^5) \end{array} \tag{38}$$

Pour un problème de mécanique, où l'équation différentielle est du second degré, on définit la vitesse par $dx/dt = v$, $dv/dt = f(x(t), t)/m$ et on applique deux fois les formules (38) par pas de temps h.

Exemple: Ressort harmonique. On discute brièvement un exemple classique–un ressort linéaire, horizontal, avec une masse, m, attachée au bout. L'équation du mouvement est

$$m\frac{d^2x}{dt^2} = -kx \tag{39}$$

où x est le déplacement par rapport à la position de l'équilibre. On impose les conditions initiales $x(0)$ et $v(0) \equiv \frac{dx}{dt}|_{t=0}$; et on transforme cette équation du deuxième degré en deux équations du premier ordre

$$\begin{array}{rcc} v & = & \frac{dx}{dt} \\ -\frac{k}{m}x & = & \frac{dv}{dt} \end{array} \tag{40}$$

La question pratique qui se soulève maintenant est comment peut-on contrôler la méthode d'Euler, du point du milieu, de Runge et Kutta, dans des cas concrets? Une méthode très puissante se base sur les lois de conservation. En effet, en mécanique, en absence de frottement, l'énergie totale est conservée. On peut se poser la question si les méthodes approchées conservent elles aussi l'énergie. Un exemple servira à clarifier les choses:

Exemple:

Soit une masse m, attachée à un ressort idéal de constante k. Les équations du mouvement sont

$$dx/dt = v$$
$$dv/dt = -(k/m)x$$

[9] **Exercice:** Démontrer que la méthode Runge-Kutta, en effet, élimine tous les termes, jusqu'à l'$O(h^5)$.

et elles conservent l'énergie totale, $E = mv^2/2 + kx^2/2$. Les équations discrètes, obtenues par l'approximation d'Euler, sont

$$\begin{aligned} x_{n+1} &= x_n + v_n h \\ v_{n+1} &= v_n + (-(k/m)x_n)h \end{aligned}$$

et un calcul direct montre que $E_{n+1} = mv_{n+1}^2/2 + kx_{n+1}^2/2 = (1+\delta)E_n$ avec

$$\delta = \frac{h^2 k}{m}$$

On peut résoudre la récurrence complètement et trouver

$$E_{n+1} = (1+\delta)^n E_0 = (1 + n\delta + O(n^2\delta^2))E_0$$

Cette relation indique que l'approximation d'Euler ne conserve pas l'énergie, mais cette différence peut être considérée "petite" d'une itération à la suivante, car de même ordre, $O(h^2)$ que l'erreur commise. Mais l'erreur s'accumule et, après n itérations on a une erreur de $n\delta$ en premier ordre en δ. Si $n\delta$ devient comparable à 1 alors notre approximation ne fait plus de sens. Conclusion: l'erreur accumulée fait si que le nombre total d'itérations, pendant lequel on peut utiliser la méthode d'Euler est

$$n_{\max} = 1/\delta = \frac{m}{h^2 k}$$

On note qu'il est proportionnel à la masse, inversement proportionnel à la constante du ressort ainsi qu'au pas. On note aussi que $n_{\max}h$ est un *temps*–et puisque la période est proportionnelle à $\sqrt{m/k}$, on a intérêt à ce que $n_{\max}h$ soit au moins égal à la période! Ces conditions imposent des contraintes non-triviales pour le choix du pas h, données les paramètres physiques m et k!

Exercice: Calculer le nombre limite d'itérations pour la méthode du point du milieu ainsi que la méthode de Runge et Kutta pour ce système.

3.1 Conclusions

On n'a présenté ici que les aspects les plus immédiats et nécessaires. Un problème très important, du point de vue pratique, est le choix du pas h optimal ainsi que sa variation: dans la plupart des application réelles il faut varier ce pas, afin qu'il soit plus petit, quand les forces (par exemple) varient rapidement dans le temps et plus grand quand elles varient peu–ainsi on peut allier la précision avec la vitesse.

Mais, avec les techniques de cette section, il est possible de résoudre tous les problèmes de la Mécanique classique à un degré de liberté. C'est déjà quelque chose! Dans la section qui suit on exposera les techniques qui étendront ce pouvoir sur la mécanique à plusieurs degrés de liberté.

Les Matrices et leurs applications

Les matrices sont les structures, que l'ordinateur a le plus grand mal à gérer de manière efficace-c'est seulement les dernières années avec le standard Fortran 90 que les manipulations des vecteurs et des matrices sont devenues vraiment performantes.

Ici on n'a à notre disposition que le standard Fortran 77-mais les techniques qu'on va décrire seront indépendantes du standard.

La physique, qui sera notre point de départ, concerne la déscription du comportement d'un système à plusieurs degrés de liberté, qui interagissent entre eux. On va étudier deux exemples représentatifs: la resolution d'un système d'équations linéaires et le calcul des valeurs propres d'une matrice.

4 Résolution d'un système d'équations linéaires

Un problème *mathématique* qu'on rencontre assez souvent est la résolution d'un système d'équations linéaires, i.e.

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b} \tag{41}$$

où $\mathbf{A} \equiv A_{ij}$ est une matrice $n \times n$, $\mathbf{x}$ est le vecteur inconnu et $\mathbf{b}$ le vecteur connu. Le but est de calculer la valeur des composantes de x, une fois les éléments de la matrice A et du vecteur b sont donnés. Comme il est bien connu, ce problème a une solution unique si, et seulement si, $\det A \neq 0$ et la solution peut être exprimée, formellement, par

$$\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{b} \tag{42}$$

L'ennui est que le calcul de l'inverse d'une matrice, directement, est une opération assez coûteuse et, surtout, vulnérable à des instabilités numériques. Pour cette raison, on décrira deux méthodes qui conduisent à la solution des éq. (41) de manière plus stable–ce sont la méthode *Gauß-Jordan* et la méthode *LU*.

4.1 La méthode Gauß-Jordan

L'idée de cette méthode est de réduire la matrice A en forme *triangulaire* et calculer, par la suite, les composantes x_i. La raison, pour laquelle cette approche est efficace, est que, si A est triangulaire, on peut calculer les x_i très facilement:

$$A = \begin{pmatrix} a_{11} & a_{12} \cdots & a_{1n} \\ 0 & a_{22} \cdots & a_{2n} \\ \cdots & \cdots & \cdots \\ 0 \cdots 0 \cdots & \cdots 0 & a_{nn} \end{pmatrix} \tag{43}$$

On trouve immédiatement,

$$x_n = b_n / a_{nn} \tag{44}$$

$$x_{n-1} = (b_{n-1} - a_{n-1,n} x_n) / a_{n-1,n-1} \tag{45}$$

et, en général[10],

$$x_{n-k} = (b_{n-k} - \sum_{l=n}^{n-k+1} a_{n-k,l}x_l)/a_{n-k,n-k},\ k = 1, \ldots, n-1 \tag{46}$$

L'exercice à résoudre, alors, consiste à transformer A en matrice triangulaire. On procède de la manière suivante:

1. On remplace la dernière équation par une combinaison linéaire d'elle-même et de la première équation, de telle sorte que l'élément a_{n1} soit égal à zéro, i.e. on multiplie la ligne $(a_{11}, \ldots, a_{1n})$ par $-a_{n1}/a_{11}$ et on remplace la ligne $(a_{n1}, a_{n2}, \ldots, a_{nj}, \ldots, a_{nn})$ par

$$\left(0, a_{n2} - a_{12}\frac{a_{n1}}{a_{11}}, \ldots, a_{nj} - a_{1j}\frac{a_{n1}}{a_{11}}, \ldots, a_{nn} - a1n\frac{a_{n1}}{a_{11}}\right)$$

   Evidemment, on remplace b_n par

$$b_n - b_1\frac{a_{n1}}{a_{11}}$$

   .

2. On continue, de cette manière, à annuller les éléments de la première colonne, à l'exception du a_{11}, bien sûr et on se ramène à la forme suivante

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ 0 & a_{32} & \cdots & a_{3n} \\ \cdots & a_{kj} & \cdots & a_{kn} \\ 0 & a_{n2} & \cdots & a_{nn} \end{pmatrix} \tag{47}$$

   où on a utilisé le même symbole pour les éléments de la dernière ligne.

3. On procède, de façon analogue, à annuller les éléments $a_{n2}, a_{n-1,2}, \ldots, a_{32}$; on considère, en effet, la matrice

$$A^{'} = \begin{pmatrix} a_{22} & \cdots & a_{2n} \\ a_{32} & \cdots & a_{3n} \\ \cdots & \cdots & \cdots \\ a_{k2} & \cdots & a_{kn} \\ a_{n2} & \cdots & a_{nn} \end{pmatrix}; \tag{48}$$

   on multiplie les éléments $(a_{22}, \ldots, a_{2n})$ par $-a_{n2}/a_{22}$ et on remplace la dernière ligne par la résultante, qui a $a_{n2} = 0$ et ainsi de suite jusqu'à la troisième ligne.

4. Il est facile, alors, de se convaincre que,au début de l'étape $m(= 1, \ldots, n-1)$ on aura annulé les éléments $a_{n1}, a_{n-1,1}, \ldots, a_{21}$;$a_{n2}, a_{n-1,2}, \ldots, a_{32}$;$\ldots a_{n,m-1}, \ldots, a_{n-1,m-1}, \ldots, a_{m,m-1}$.

   A la fin de l'étape $n-1$ la matrice aura alors une forme triangulaire.

[10] **Exercice:** Démontrer cette formule.

Exemple:

$$\begin{pmatrix} 1 & 2 & 1 \\ 1 & 1 & 1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \tag{49}$$

On multiplie la première ligne par 1 et on remplace la troisième par $(0, 3, 2)$. On trouve

$$\begin{pmatrix} 1 & 2 & 1 \\ 1 & 1 & 1 \\ 0 & 3 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}$$

Ensuite, on multiplie la première ligne par 1 et on la soustrait de la deuxième (pour annuler a_{21}); le système devient, alors

$$\begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & 0 \\ 0 & 3 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix}$$

On multiplie la deuxième ligne par 3 et on la retranche de la troisième, pour annuller a_{32}–et on a fini. On obtient

$$\begin{pmatrix} 1 & 2 & 1 \\ 0 & 3 & 0 \\ 0 & 0 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -3 \\ -1 \end{pmatrix}$$

et on trouve immédiatement $z = 1/2$; $y = -1$;$x = 5/2$. Si on remplace dans le système *original*, éq. (49), on trouve, effectivement, une identité.

On peut remarquer que les éléments diagonaux, $a_{m,m}$ jouent un rôle très important, car on divise par eux; et on peut se poser la question de que se passe-t-il si un ou plusieurs de ces éléments s'annulle. L'idée est que, si le système n'est pas singulier, il doit être possible de trouver des éléments diagonaux non-nuls par permutation des équations et/ou des variables. En effet, le système ne doit pas changer si on échange deux équations; par contre les variables vont se mélanger si on échange deux colonnes de la matrice, mais le mélange correspond à une transformation linéaire. Le but de l'exercice est de faire en sorte que les éléments diagonaux soient les plus grands en valeur absolue de tous les éléments d'une ligne ou/et d'une colonne. Cette opération s'appelle *pivotage* et l'élément diagonal *le pivot*. Le *pivotage partiel* est le plus facile à implémenter et suffit pour rendre la procédure stable; par conséquent on l'utilisera par la suite. Dans le pivotage partiel on échange les équations jusqu'à ce qu'on mette sur la diagonale l'élément le plus grand, en valeur absolue, parmi les éléments de la même colonne–tandis que le pivotage complet serait de mettre sur la diagonale l'élément le plus grand, en valeur absolue parmi les éléments de la colonne *et* la ligne. Un exemple fera les choses plus claires.

Exemple:

$$\begin{pmatrix} 2 & 2 & 1 \\ 1 & 1 & 1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \tag{50}$$

On trouve:

$$\begin{pmatrix} 2 & 2 & 1 \\ 1 & 1 & 1 \\ 0 & 4 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ -5 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 2 & 1 \\ 0 & 0 & -1 \\ 0 & 4 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -3 \\ -5 \end{pmatrix}$$

et on voit que $a_2 2 = 0$. Mais on peut échanger les lignes 2 et 3 sans changer les valeurs de x, y, z, i.e. se ramener à

$$\begin{pmatrix} 2 & 2 & 1 \\ 0 & 4 & 3 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -5 \\ -3 \end{pmatrix}$$

qui, dans ce cas, est triangulaire. En général, il faut contrôler $|a_{mm}|$; et si $|a_{mm}| < \epsilon$ il faut échanger la ligne m avec la ligne j, où $|a_{mj}| = \max_{l=1,\ldots,n}\{a_{ml}$. Bien sûr $b_m \leftrightarrow b_j$. **Exercice:** (a). Ecrire l'algorithme Gauß-Jordan avec pivotage partiel. (b). Ecrire un programme Fortran qui réalise cet algorithme.

4.2 La décomposition LU

Cette méthode se base aussi sur l'idée que résoudre un système linéaire, dont la matrice est triangulaire, est trivial–et la pousse jusqu'au bout, i.e. on cherche à exprimer la matrice A comme le produit de deux matrices, dont une est triangulaire inférieure (Lower–triangular en anglais), (i.e. $a_{ij} = 0$ pour $j > i$) et l'autre est triangulaire supérieure (Upper-triangular en aglais), (i.e. $a_{ij} = 0$ pour $j < i$). En équations

$$A \cdot x = (L \cdot U) \cdot x = L \cdot (U \cdot x) = b \tag{51}$$

et, si on pose $y = U \cdot x$ on peut calculer les composantes de y immédiatement, par substitution, ainsi que les composantes de x par la suite.

Alors comment trouver les matrices L et U? Prenons l'exemple d'une matrice 4×4:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{11} & l_{12} & l_{13} & l_{14} \end{pmatrix} \cdot \begin{pmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix} \tag{52}$$

On s'aperçoit que éq. (52) est un système de n^2 équations (autant d'éléments de la matrice A) et $n^2 + n$ inconnus (les variables l_{ij} et u_{ij}). Ce qui veut dire qu'on peut imposer n conditions. Un choix, qui simplifie les choses, est

$$l_{ii} = 1, \ i = 1, \ldots, n$$

et on peut calculer les autres par *l'algorithme de Crout*:

Pour chaque $j = 1, 2, 3, \ldots, n$: (a) pour $i = 1, 2, \ldots, j$

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj}$$

(la somme est zéro lorsque $i = 1$).

(b) pour $i = j+1, \ldots, n$

$$l_{ij} = \frac{1}{u_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj} \right)$$

Exemple: On va résoudre le système de l'exemple précédent par la méthode LU.

$$\begin{pmatrix} 1 & 2 & 1 \\ 1 & 1 & 1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

On pose $l_{11} = l_{22} = l_{33} = 1$.

$j = 1$.

$i = 1. u_{11} = a_{11} = 1$

$i = 2. l_{21} = \frac{1}{u_{11}}(a_{21}) = 1$

$i = 3. l_{31} = \frac{1}{u_{11}}(a_{31}) = 1$

$j = 2$.

$i = 1. u_{12} = a_{12}$

$i = 2. u_{22} = a_{22} - l_{21}u_{12} = 1 - 1 \times 2 = -1$

$i = 3. l_{32} = \frac{1}{-1}(a_{32} - l_{31}u_{12}) = -(-3) = +3.$

$j = 3$.

$i = 1. u_{13} = a_{13} = 1$

$i = 2. u_{23} = a_{23} - l_{21}u_{13} = 1 - 1 \times 1 = 0$

$i = 3. u_{33} = a_{33} - (l_{31}u_{13} + l_{32}l_{23}) = -1 - (1 \times 1 + 3 \times 0) = -2$

Ce qui conduit aux systèmes suivants:

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 3 & 1 \end{pmatrix} \cdot \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix}$$

et on trouve $x' = 1, y' = 1, z' = -1$ et $x = 5/2, y = -1, z = 1/2$.

Une question mathématique pertinente est: quelles sont les conditions, nécessaires et suffisantes, pour qu'une matrice $n \times n$ puisse être mise sous la forme d'un produit $L \cdot U$? Il est possible de démontrer qu'il suffit que tous les *mineurs* de A soient non-nuls.

Mais il se peut qu'on ait une matrice non-singulière, dont un mineur est zéro. Alors que faire? La réponse est qu'on peut décomposer en forme LU non A elle-même, mais une permutation de ligne de A. Il s'agit de l'implémentation du pivotage dans l'algorithme de Crout. L'idée vient de l'étape **(b)** de l'algorithme de Crout. On se rend compte qu'on peut calculer le *numérateur* indépendamment du dénominateur. C'est après avoir calculer *tous* les numérateurs qu'on doit se décider. Alors on choisit, parmi les u de la colonne j, le plus grand, on change la ligne j avec la ligne $i_{\max}$ et on divise *seulement* les éléments de l'étape **(b)** par le nouveau u_{jj}. La raison pour laquelle cette astuce marche est qu'en effet on n'alloue pas de mémoire pour *trois* matrices $n \times n$ (les matrices $\mathbf{A}, \mathbf{L}$ et $\mathbf{U}$) mais on remplace "sur place", vu qu'un élément a_{ij} ne sera utilisé qu'une seule fois.

Exercice: (a). Résoudre le système linéaire précédent, qui avait besoin du pivotage, sous Gauß-Jordan, par la méthode LU.

(b). Ecrire un programme Fortran, qui réalise l'algorithme LU, avec pivotage.

4.3 Calcul du déterminant

Une quantité assez intéressante, tant du point de vue mathématique que du point de vue physique est le *déterminant* d'une matrice. Sa sgnification physique est celle d'un "volume"–à deux dimensions il se réduit à l'aire (orientée), à trois dimensions au vrai volume. En algèbre linéaire on apprend qu'il fait partie des "invariants"–des quantités qui ne changent pas, lorsque la matrice initiale subit une transformation de similarité, $\mathbf{A} \to \mathbf{SAS}^{-1}$. On apprend aussi comment le calculer (le devéloppement en mineurs). Ce calcul coûte $O(n^3)$ opérations pour une matrice générique–mais seulement $O(n)$ opérations pour une matrice triangulaire–car on peut démontrer que le déterminant est égal au produit des éléments diagonaux, dans ce dernier cas. Etant donné que tant la méthode Gauß-Jordan que la méthode LU transforment notre matrice initiale en forme triangulaire, on se rend compte qu'on peut calculer assez facilement le déterminant. Une précaution s'impose: On sait que la valeur du déterminant change de signe lors de léchange de deux lignes ou de deux colonnes–par conséquent on doit introduire une variable auxiliaire, qui change de signe avec chaque changement de ligne lors du pivotage. Cette variable donnera le signe global et le déterminant sera égal à

$$\det \mathbf{A} = \text{parité} \times \prod_{k=1}^{n} a_{kk}$$

Quelles sont les points sensibles? Le produit peut devenir très grand ou très faible (en valeur absolue)–alors il faudrait calculer la somme des logarithmes...mais il faut aussi tenir compte des considérations physiques–si une (ou plusieurs) valeur propre tend vers zéro, i.e. le déterminant tend vers zéro, ceci est l'indication d'une "presque-symétrie" qui vaudrait peut-être la peine d'être étudiée...

5 Valeurs Propres et Vecteurs Propres d'une Matrice Symétrique

Une matrice est la réprésentation d'un opérateur, i.e. d'une application d'un espace à lui-même. Il est bien connu qu'elle reflète le choix d'une *base* dans cet espace et on peut se poser la question de l'existence de quantités, qui caractérisent une réprésentation, *indépendamment* de la base utilisée, i.e. si on peut définir une sorte de "classes d'équivalence"; toutes les matrices qui auraient les mêmes quantités, appartiendraient à la même classe. Il est bien connu que la question, qu'on vient de poser admet une réponse affirmative, à savoir que toutes les matrices qui sont liées par une opération de *similarité*,

$$\mathbf{A} \to \mathbf{S}^{-1}\mathbf{AS} \tag{53}$$

ont les mêmes *valeurs propres*, i.e. les solutions de léquation

$$\det(\mathbf{A} - \lambda \mathbf{1}) = 0 \tag{54}$$

Alors c'est clair que ce sont ces quantités, les valeurs propres, qui sont les choses vraiment intéressantes–et surtout de point de vue physique[11]. Question pratique, maintenant: comment

[11]Une discussion très jolie se trouve dans le livre *The Character of Physical Law* de R. P. Feynman, MIT Press (1965). Il est aussi amusant de constater que l'effort de la *classification des invariants* a commencé, par

les calculer? En général la tâche est très difficile–mais, si on se restreint aux cas relevants pour les sciences naturelles, on se rend compte qu'on a affaire à des matrices symétriques[12] dans la grande majorité des cas. C'est alors à leur étude qu'on va se restreindre, en évoquant les autres cas quand on en aura affaire pratique.

En outre il est bien connu qu'il existe un théorème qui nous assure qu'une matrice symétrique a les valeurs propres *réelles* et les *vecteurs propres* associés peuvent toujours être choisis orthogonaux. Conséquence: on peut restreindre, dans ce cas, nos études aux matrices $\mathbf{A}$ symétriques et aux matrices $\mathbf{S}$ orthogonales, i.e. pour lesquelles $\mathbf{S}^{-1} = \mathbf{S}^{\mathrm{T}}$. L'idée maintenant est d'arriver à une forme diagonale de $\mathbf{A}$, c.à. d. dont les éléments seront les valeurs propres, par opérations orthogonales successives, qui annuleraient les éléments en dehors de la diagonale principale.

$$\mathbf{A} \to \mathbf{P}_{12}^{\mathrm{T}}\mathbf{A}\mathbf{P}_{12} \to \mathbf{P}_{13}^{\mathrm{T}}\mathbf{P}_{12}^{\mathrm{T}}\mathbf{A}\mathbf{P}_{12}\mathbf{P}_{13} \to \cdots \left(\prod \mathbf{P}\right)^{\mathrm{T}} \mathbf{A} \prod \mathbf{P} = \Lambda \tag{55}$$

L'astuce maintenant consiste à trouver une famille de matrices $\mathbf{P}$ pour lesquelles les multiplications matricielles sont les plus simples possible. Jacobi (1846) a proposé alors d'essayer d'annuler les éléments non-diagonaux par rotations *planaires* successives. Ceci veut dire que, pour annuler l'élément (k, l) de la matrice $\mathbf{A}$, a_{kl}, on choisit une matrice P_{kl} qui est égale à la matrice identité, sauf que $p_{kk} = p_{ll} = \cos\theta$ et $p_{kl} = \sin\theta = -p_{lk}$ et on choisit la valeur de l'angle θ de telle sorte que $a'_{kl} = a'_{lk} = \left(\mathbf{P}^{\mathrm{T}}\mathbf{A}\mathbf{P}\right)_{kl} = 0$. On continue de cette manière, pour tous les éléments non-diagonaux. Il est maintenant clair que, pour une matrice $N \times N$, le coût de toutes ces multiplications matricielles devient prohibitif; d'autre part, la grande majorité sont inutiles, car seules 2 lignes et 2 colonnes de la matrice $\mathbf{A}$ changent lors d'une transformation de Jacobi[13] ,à savoir les colonnes k et l ainsi que les lignes k et l. Par conséquence, il est plus efficace d'écrire explicitement les éléments qui changent. Il est plus facile d'effectuer ce calcul dans le cas général. On trouve, en particulier[14]

$$a'_{kl} = \sin\theta\cos\theta\,(a_{k,k} - a_{l,l}) + a_{k,l}\left(\cos^2\theta - \sin^2\theta\right) \tag{56}$$

qui donne pour l'angle qui anulle cet élément

$$\theta^*_{kl} = \frac{1}{2}\tan^{-1}\frac{-2a_{kl}}{a_{kk} - a_{ll}} \tag{57}$$

Cette expression reste encore très théorique. En pratique il faut examiner à part le cas où $|a_{kk} - a_{ll}| < \epsilon$; dans ce cas on prendra $\theta^*_{kl} = \pi/4$. Aussi, si $|a_{kl}| < \epsilon$, il est inutile de faire une rotation–alors on doit aller directement à l'élément suivant. Finalement, après avoir effectué les $n(n-1)/2$ rotations on ne trouvera pas une matrice diagonale, même à la précision de la machine. La raison est que l'algorithme de Jacobi ne converge pas, en général, en un nombre fini d'étapes. *Mais*, en pratique, 4–10 itérations de ces $n(n-1)/2$ rotations suffisent pour réduire la

les mathématiciens, beaucoup plus tôt: formellement avec Sylvester et Cayley au milieu du dix-neuvième siècle, aux travaux duquels s'associent aussi les noms de Klein, Gordan, Hilbert mais cette liste est très partielle. Il ne faut pas non plus oublier les fondateurs de la géométrie projective aux débuts du 19ème siècle, tels Poncelet, Monge.

[12] Pour le cas réel. Pour le cas complexe, qui est aussi relevant pour la physique, on aura affaire à des matrices *hermitiennes*.

[13] **Exercice:** Démontrer cette proposition.

[14] **Exercice:** Démontrer cette formule et établir les autres, énoncées pendant le cours.

valeur absolue d'un élément typique non-diagonale au-delà de la précision de la machine–alors on s'arrête. Si les éléments de la matrice ont une symétrie particulière, alors la procédure peut converger très rapidement.

Exemple: $n = 2$. Soit la matrice 2×2

$$\mathbf{A} = \begin{pmatrix} a & b \\ b & d \end{pmatrix} \tag{58}$$

On doit annuler l'élément $a_{12} = b$ et une rotation suffira. En effet,

$$\mathbf{A}' = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \cdot \begin{pmatrix} a & b \\ b & d \end{pmatrix} \cdot \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{59}$$

et on trouve facilement que

$$\theta_{12}^* = \frac{1}{2}\tan^{-1}\left(\frac{-2b}{a-d}\right) \tag{60}$$

Les éléments $a_{11}^{'}$ et $a_{22}^{'}$ dépendent, en général,de l'angle θ. Pour $\theta = \theta_{12}^*$ ils sont les valeurs propres de la matrice $\mathbf{A}$, tandis que les colonnes de la matrice de rotation sont les *vecteurs propres*, i.e. les vecteurs pour lesquels l'action de la matrice devient une simple *dilatation* (ou contraction)[15].

Exemple: $n = 3$. Soit la matrice 3×3

$$\mathbf{A} = \begin{pmatrix} a & b & b \\ b & a & b \\ b & b & a \end{pmatrix} \tag{61}$$

On a d'abord besoin de 3 itérations, pour annuler les éléments $a_{12} = b$, $a_{13} = b$, $a_{23} = b$. Les trois matrices ont les expressions suivantes

$$\mathbf{P}_{12} = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{62}$$

$$\mathbf{P}_{13} = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \tag{63}$$

$$\mathbf{P}_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix} \tag{64}$$

On commence par le calcul de $\mathbf{P}_{12}^{\mathrm{T}}\mathbf{A}\mathbf{P}_{12}$. Le calcul direct, ou l'application de la formule générale, conduit au résultat $\theta_{12}^* = \pi/4$, qui annulle non seulement a_{12} mais aussi a_{13}–par conséquence

[15] **Exercice:** Finir le calcul. Trouver les valeurs propres et les vecteurs propres lorsque $\mathbf{A}$ est la matrice d'inertie d'une plaque mince, en forme de triangle droit avec cotés (L_1, L_2, L_3) et les axes possibles sont de rotation sont Ox et Oy; Oz est perpendiculaire au plan du triangle et passe par un des ses trois sommets. On a deux moments d'inertie, I_{xx}, I_{yy}. Elles sont les valeurs propres de la matrice d'inertie, ici 2×2. Interpréter le résultat. Autour de quel axe est-il le plus facile de tourner la plaque?

il est inutile de faire la rotation pour annuler a_{13} et on a déjà une valeur propre, $\lambda_1 = a - b$. Il nous reste d'annuler l'élément $a'_{23}(\theta^*_{12}) = b\sqrt{2}$, par une rotation avec la matrice $\mathbf{P}_{23}$. On remarque que, dans *cette* occasion, l'application de cette dernière rotation ne modifiera pas les valeurs des éléments non-diagonaux, déjà nuls, par conséquent on peut isoler la sous-matrice 2×2 et écrire un rotation à deux dimensions. On trouve, après calcul[16], que $\tan 2\theta^*_{23} = -2\sqrt{2}$ et que $a''_{22}(\theta^*_{23}) = a - b$, $a''_{33}(\theta^*_{23}) = a + 2b$. Les vecteurs propres sont les colonnes de la matrice $P_{12}(\theta^*_{12}) \cdot P_{23}(\theta^*_{23})$ et constituent un système orthonormé[17]

References

[1] *Numerical Recipes: The Art of Scientific Computing in Fortran*, W. Press *et al.* Cambridge University Press (1992). Ce livre est disponible sur le Web; son adresse est: `http://cfatab.harvard.edu/nr/nronline.html`.

[2] Pour une discussion très jolie sur la résolution numérique des équations différentielles, ainsi que sur les approximations numériques en général, cf. R. P. Feynman, *The Feynman Lectures in Physics*, vol I. Addison-Wesley (1963) (disponible en édition bilingue ainsi qu'en français à la BU).

A Les particularités de la matrice tridiagonale

Si la matrice $\mathbf{A}$ est *tridiagonale*, i.e. de la forme

$$\mathbf{A}_{kl} = b_k \delta_{k,l-1} + a\delta_{k,k} + b_k \delta_{k,l+1}$$

où on a utilisé le symbole *delta de Kronecker*

$$\delta_{k,l} = \begin{cases} 1 & k = l \\ 0 & k \neq l \end{cases}$$

c.à.d. elle a l'élement a sur la diagonale principale et les éléments b_{k-1} et b_k sur la ligne k.

On peut trouver ses valeurs propres facilement par la méthode suivante: on note que l'équation caractéristique prend la form

$$\begin{array}{rcl} (a_1 - \lambda)v_1 + b_1 v_2 & = & 0 \\ b_{k-1}v_{k-1} + (a_k - \lambda)v_k + b_k v_{k+1} & = & 0, \quad k = 2, \ldots, n-1 \\ b_{n-1}v_{n-1} + (a_n - \lambda)v_n & = & 0 \end{array} \tag{65}$$

alors on se rend compte que les $v_2, \cdots, v_n$ se laissent exprimer en fonction du seul v_n en utilisant les $n-1$ premièeres équations; la dernièere équation exprime v_n en fonction de v_{n-1}, i.e. en fonction de v_1. On obtient ainsi une condition de cohérence, car le v_n doit être le même, indépendamment de l'expression utilisée. Cette condition est justement le polynôme caractéristique. Une fois qu'on a ses racines (i.e. les valeurs propres), on prend v_1 arbitraire et on calcule les composantes du vecteur propre correspondant.

[16] **Exercice:** Finir ce calcul.

[17] **Exercice:** Dessiner le système de ces trois vecteurs propres dans le référentiel canonique, $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$.

Exemple: Soit la matrice

$$\mathbf{A} = \begin{pmatrix} a & b_1 & 0 \\ b_1 & a & b_2 \\ 0 & b_2 & a \end{pmatrix}$$

On trouve facilement que

$$\begin{aligned} (a-\lambda)v_1 + b_1 v_2 &= 0 \\ b_1 v_1 + (a-\lambda)v_2 + b_2 v_3 &= 0 \\ b_2 v_2 + (a-\lambda)v_3 &= 0 \end{aligned}$$

et le polynôme caractéristique s'écrit

$$\frac{b_2}{b_1} = -\frac{b_1}{b_2} + \frac{(a-\lambda)^2}{b_1} b_2 \tag{66}$$

et on peut facilement calculer (ici même analytiquement) les solutions de cette équation. La question intéressante dans cet exemple est la suivante: la matrice est de rang 3–pourtant l'équation à laquelle on a aboutit est de degré 2; qu'est-ce qu'elle est devenue la troisième racine??

On peut court-circuiter ce problème, en raisonnant de la manière suivante: les deux racines de cette équation sont certainement des valeurs propres de la matrice. Mais on sait que la somme des éléments diagonaux d'une matrice est invariante sous transformations de similarité– par conséquent elle est égale à la somme des *toutes* les valeurs propres. Alors on peut trouver cette racine par simple soustraction. En effet on a

$$\mathrm{Tr}\,\mathbf{A} = \lambda_1 + \lambda_2 + \lambda_3 = 3a$$

$$\lambda_{1,2} = a \pm \sqrt{b_1^2 + b_2^2} \Rightarrow \lambda_3 = a$$

On trouve alors qu'une des valeurs propres ne dépend pas de $b_{1,2}$. Coïncidence? On prend une matrice plus générale, où on remplace les éléments diagonaux par a_1, a_2 et a_3. On trouve par inspection que, si $a_1 = a_3 = a$, alors $\lambda_3 = a$ indépendamment des valeurs de a_2 et b_1, b_2. Application: si $a_1 = a+\delta = a_3$, les valeurs propres seront, en général, des fonctions non-triviales de δ. Imaginons, alors, de vouloir faire un devéloppement limité des fonctions $\lambda_i(\delta)$ autour du point $\delta = 0$. On déduit par l'exercice qu'on vient d'étudier que, pour λ_3 ce devéloppement se terminera après la correction linéaire en δ et que cette valeur propre n'aura *aucune autre dépendence* en δ.

Cet argument n'est pas complètement satisfaisant. On aimerait mieux comprendre ce qui se passe, à savoir comment on arrive directement à une équation de deuxième degré dans l'exemple précédent. Ce point doit être mieux approfondi.

B Les récurences

Une autre application du calcul des matrices est fournie par les récurences:

$$x_{n+1} = a x_n + b x_{n-1} \tag{67}$$

avec conditions initiales x_0 et x_1 quelconques. On cherche l'expression de x_n en fonction de n ainsi que son comportement lorsque n devient "grand". Il est claire qu'il s'agit de l'équation caractéristique d'une matrice tridiagonale de rang infini. Les matrices de rang infini sont des objets assez délicates à manipuler; ainsi on essaie de transformer notre problème en une autre, mieux défini. A cette fin écrirons la récurrence précédente comme

$$\begin{array}{rcl} x_{n+1} & = & ax_n + bx_{n-1} \\ x_n & = & x_n \end{array} \tag{68}$$

et introduisons le vecteur $\mathbf{z}_n = (x_n, x_{n-1})^{\mathrm{T}}$, en termes de qui la récuurence devient

$$\mathbf{z}_{n+1} = \mathbf{A}^n \mathbf{z}_n \tag{69}$$

On doit évaluer la n-ième puissance d'une matrice 2×2, *non-symétrique*. On trouve analytiquement valeurs propres et vecteurs propres, $\lambda_{1,2}$ et $\mathbf{X}_{1,2}$. En général, les vecteurs propres seront une base et on peut écrire le vecteur $\mathbf{z}_1$ dans cette base

$$\mathbf{z}_1 = c_1 \mathbf{X}_1 + c_2 \mathbf{X}_2 \tag{70}$$

et

$$\mathbf{A}^{n-1} \cdot \mathbf{z}_1 = \mathbf{z}_n = c_1 \lambda_1^{n-1} \mathbf{X}_1 + c_2 \lambda_2^{n-1} \mathbf{X}_2$$

qui est la solution du problème. Pour n très grand et $|\lambda_1| > |\lambda_2|$ on peut déduire que

$$x_n \approx C_1 \lambda_1^n \left(1 + C_2 \left(\frac{\lambda_2}{\lambda_1} \right)^n \right) \tag{71}$$

qui indique que les x_n suivent une simple exponentielle, à des corrections près, qui deviennent négligeables, dans la limite $n \to \infty$. Ainsi est-il plus intelligent d'écrire

$$\ln x_n = n \ln \lambda_1 + \ln C_1 + \ln \left(1 + C_2 \left(\frac{\lambda_2}{\lambda_1} \right)^n \right) \tag{72}$$

et la question relevante est quand est-ce que le rapport $(\lambda_2/\lambda_1)^n < \epsilon$, avec ϵ la précision de la machine sur laquelle on travaille, par exemple.