Université de PARIS SUD
Master de Physique M1
Parcours Physique Appliquée

Année Universitaire 2004-2005

Cours/TD/TP : E. Trizac

Physique numérique et analyse statistique

Table des matières

1 Exercices de cours

1.1 Différences finies

Exercice I

On note f_k les valeurs prises par un signal $f(x)$ ($\mathbb{R} \to \mathbb{R}$) en des points régulièrement échantillonnés. On introduit l'opérateur "différence vers l'avant" : $\Delta f_k \equiv f_{k+1} - f_k$. Quelle est l'expression de $\Delta^r f$? Même question avec la différence centrée $\delta f_k \equiv f_{k+1/2} - f_{k-1/2}$.

Exercice II

Comment peut-on construire un polynôme d'interpolation de degré m pour le signal précédent, connaissant les valeurs f_k, f_{k+1},..., f_{k+m} prises par f aux points x_k, x_{k+1},..., x_{k+m} ? Exprimer alors le résultat à l'aide de l'opérateur de différence vers l'avant.

Exercice III

Soit $f_{i,j} \equiv f(x_i, y_j)$ la valeur prise par une fonction f sur le nœud (i,j) d'un réseau carré ($x_i = x_0 + i\Delta x$ et $y_j = y_0 + j\Delta y$). Proposer une approximation simple pour la dérivée seconde $[\partial_x \partial_y f]_{i,j}$. L'égalité des dérivées partielles croisées est-elle bien vérifiée ?

En déduire deux schémas d'approximation du laplacien $\nabla^2 f$ au nœud (i,j).

Exercice IV : Différences finies et diffusion de la chaleur

On considère un métal de conductivité thermique λ, constituant un mince barreau rectiligne de longueur L. Proposer un algorithme permettant de calculer la température $T(x_i, t_n)$ au point x_i à l'instant t_n, connaissant le profil initial $T(x, 0)$ et les conditions aux limites en $x = 0$ et $x = L$. Peut-on qualifier ce schéma d'explicite ?

1.2 Algèbre linéaire

Exercice I : Principe et intérêt de la décomposition LU

Pour résoudre le système linéaire $\mathbb{A}\vec{x} = \vec{b}$, une méthode plus moderne et plus efficace que celle du pivot de Gauss consiste à factoriser $\mathbb{A}$ sous la forme "$\mathbb{L}$ower-$\mathbb{U}$pper" : on cherche une représentation de $\mathbb{A}$ comme produit de deux matrices triangulaires (inférieure pour $\mathbb{L}$ et supérieure pour $\mathbb{U}$) :

$$\mathbb{A} = \mathbb{L}\mathbb{U} \quad \text{avec} \quad \mathbb{L} = \begin{pmatrix} l_{11} & 0 & \dots & \dots & 0 \\ l_{21} & l_{22} & 0 & \dots & 0 \\ \vdots & & & & \vdots \\ \vdots & & & & 0 \\ l_{N1} & \dots & \dots & \dots & l_{NN} \end{pmatrix} \quad \text{et} \quad \mathbb{U} = \begin{pmatrix} u_{11} & u_{12} & \dots & \dots & u_{1N} \\ 0 & u_{22} & \dots & \dots & u_{2N} \\ \vdots & & & & \vdots \\ 0 & \dots & 0 & & \vdots \\ 0 & \dots & \dots & 0 & u_{NN} \end{pmatrix}$$

Etant donné que $\mathbb{A}\vec{x} = \vec{b} \Leftrightarrow \mathbb{L}(\mathbb{U}\vec{x}) = \vec{b}$, on cherche d'abord $\vec{y}$ tel que $\mathbb{L}\vec{y} = \vec{b}$ puis $\vec{x}$ solution de $\mathbb{U}\vec{x} = \vec{y}$. En raison de la forme triangulaire de $\mathbb{L}$ et $\mathbb{U}$, ces équations sont simples à résoudre.

1) Combien la donnée de $\mathbb{A} = \mathbb{L}\mathbb{U}$ fournit-elle d'équations scalaires ? Dénombrer les inconnues du problème. Conclusion ?

2) Une manière de lever l'ambiguïté précédente est de remplir la diagonale principale de $\mathbb{L}$ avec des 1 : $l_{ii} = 1$ pour $1 \leqslant i \leqslant N$. Dans ces conditions, établir qu'une procédure *opérationnelle* d'obtention des éléments l_{ij} et u_{ij} est la suivante :

$$\text{pour } j = 1, 2 \dots, N; \qquad u_{1j} = a_{1j}$$

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}\, u_{kj}; \qquad 2 \leqslant i \leqslant j$$

$$l_{ij} = \frac{1}{u_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj} \right); \qquad j+1 \leqslant i \leqslant N$$

3) Donner la décomposition $L\,U$ des matrices

$$\mathbb{A} = \begin{pmatrix} 2 & 3 \\ 0 & 1 \end{pmatrix} \quad \text{et} \quad \mathbb{B} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

4) La connaissance de la décomposition $L\,U$ d'une matrice $\mathbb{M}$ fournit-elle sans calcul supplémentaire la décomposition $L\,U$ de la matrice ${}^t\mathbb{M}$ (transposée de $\mathbb{M}$) ?

5) On considère une matrice carrée $N * N$.

a) Proposer deux méthodes de calcul du déterminant.

b) Exprimer en fonction de N l'ordre de grandeur des nombres de multiplications requis dans les deux cas. Conclure.

6) Discuter *succinctement* l'intérêt que présente la décomposition $L\,U$ pour le calcul matriciel.

Exercice II : Méthode de récursion

Résoudre, par la méthode de récursion propre aux matrices tridiagonales, le système linéaire $\mathbb{A}\,\vec{x} = \vec{b}$ avec :

$$\mathbb{A} = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 2 & 3 & 1 & 0 \\ 0 & 1 & 4 & 2 \\ 0 & 0 & 1 & 3 \end{pmatrix} \qquad \text{et} \qquad \vec{b} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

Exercice III : Retour sur la conduction thermique

On reprend l'exercice de la partie 1.1 sur la diffusion de la chaleur. La température $T(x,t)$ en un point x à l'instant t obéit à l'équation

$$\frac{\partial T(x,t)}{\partial t} = \lambda \frac{\partial^2 T(x,t)}{\partial x^2}. \tag{1}$$

On note $T_i^n \equiv T(x_i, t_n)$. Pour des raisons de stabilité, on choisit de discrétiser l'équation aux dérivées partielles (1) en faisant les approximations

$$\frac{\partial T}{\partial t}(x_i, t_n) \simeq \frac{1}{\Delta t}\left[T_i^{n+1} - T_i^n\right] \tag{2}$$

$$\text{et} \qquad \frac{\partial^2 T}{\partial x^2}(x_i, t_n) \simeq \frac{\delta^2 T_i^{n+1}}{(\Delta x)^2} = \frac{1}{(\Delta x)^2}\left[T_{i+1}^{n+1} - 2T_i^{n+1} + T_{i-1}^{n+1}\right] \tag{3}$$

En désignant $\vec{T}^n$ le vecteur colonne formé des éléments $(T_i^n)_{1 \leqslant i \leqslant N}$, mettre l'équation d'évolution de la température sous la forme $\mathbb{A}\vec{T}^{n+1} = \vec{T}^n$, où $\mathbb{A}$ est une matrice tridiagonale (attention aux conditions aux limites, supposées telles que la température est fixée aux points extrêmes du barreau). Pourquoi cette méthode est-elle implicite ?

1.3 Variables aléatoires

Exercice I

Les grains (supposés sphériques) constituant une poudre sont caractérisés par une distribution de volume

$$f(v) = \frac{1}{v_0} \exp(-v/v_0).$$

Suivant quelle loi leur diamètre est-il distribué ? Le diamètre moyen est-il égal au diamètre le plus probable ? Même question en ce qui concerne le volume.

Exercice II

Soient x_1 et x_2 deux variables aléatoires de densité de probabilité jointe $f(x_1, x_2)$. On effectue le changement de variables $y_1 = x_1 + x_2$ et $y_2 = x_1 - x_2$. Quelle est la densité jointe du couple (y_1, y_2) ?

Exercice III

1) A partir d'une variable aléatoire y uniformément répartie dans l'intervalle $[0, 1]$, proposer une méthode pour engendrer une variable aléatoire x de densité de probabilité $f(x) = \alpha/x^2$ pour $x \in [1, +\infty[$. On précisera la valeur du coefficient α.

2) Même question ensuite dans le cas où $x \in [2, +\infty[$ (et où l'on suppose donc que x ne peut pas prendre de valeurs comprises entre $-\infty$ et 2).

3) Généraliser la procédure précédente aux distributions de la forme $g(x) = \beta/x^n$ sur $[1, +\infty[$.

Exercice IV

On dispose d'un générateur de nombres aléatoires de densité uniforme dans $[0, 1]$.

1) Proposer deux méthodes pour engendrer un vecteur aléatoire uniformément réparti dans le disque de rayon 1 (disque unité), dont l'une se passe entièrement du calcul de fonctions trigonométriques. Laquelle de ces deux approches est *a priori* la plus efficace d'un point de vue informatique ?

2) Même question pour engendrer un vecteur aléatoire de module 1 (qui se trouve donc sur le cercle unité).

3) On s'intéresse désormais à la géométrie sphérique. On pose

$$\begin{cases} x = \sin(\theta)\,\cos(\varphi) \\ y = \sin(\theta)\,\sin(\varphi) \\ z = \cos\theta \end{cases}$$

Quelles conditions doivent vérifier θ et φ pour que le vecteur de coordonnées (x, y, z) soit un vecteur aléatoire "isotrope" sur la sphère de rayon 1 ? *Indication : contrairement à φ, θ ne doit pas être choisi de densité de probabilité uniforme.*

Exercice V

En fonction de la pulsation ω, la densité spectrale d'énergie électromagnétique d'un corps noir à la température T se met sous la forme :

$$I(\omega) = \frac{\hbar}{\pi^2 c^3} \frac{\omega^3}{e^{\hbar\omega/(kT)} - 1},$$

où k est la constante de BOLTZMANN, $2\pi\hbar$ la constante de PLANCK, et c la vitesse de la lumière.

1) La relation donnant la longueur d'onde λ en fonction de la pulsation est $\omega\lambda = 2\pi c$. Donner l'expression de la fonction J telle que la densité d'énergie comprise dans la bande de longueur d'onde $[\lambda_0, \lambda_0 + \delta\lambda_0]$ s'écrive $J(\lambda_0)\,\delta\lambda_0$.

2) Donner un équivalent de $J(\lambda)$ pour les grandes longueurs d'onde λ.

3) Calculer la densité totale d'énergie du corps noir (énergie comprise dans la bande de pulsation $[0;+\infty[$). On donne :

$$\int_0^\infty \frac{x^3}{e^x - 1}\,dx \;=\; \frac{\pi^4}{15}$$

4) Expliquer *succinctement* comment engendrer numériquement une variable aléatoire de même densité que ω. Le cas échéant, quels sont les problèmes rencontrés ?

Exercice VI

On dispose d'un générateur de nombres pseudo-aléatoires y uniformément distribués dans [0,1].

1) Proposer une procédure –directement utilisable dans un programme informatique– qui permette de créer une variable aléatoire Lorentzienne x de densité de probabilité

$$f(x) = \frac{1}{\gamma}\,\frac{1}{1+x^2},$$

où x varie dans l'intervalle $]-\infty;+\infty[$. Préciser également la valeur de la constante γ.

2) Généraliser la méthode au cas de la densité

$$f_\alpha(t) = \frac{\beta}{\pi}\,\frac{1}{\alpha^2 + t^2},$$

où α est un paramètre donné. De nouveau, la variable aléatoire en question a pour domaine de variation $]-\infty;+\infty[$. Quelle doit être la valeur de β ? Comment peut-on interpréter le coefficient α ? Peut-on considérer qu'il s'agit de l'écart-type de la distribution ?

Exercice VII : Méthode de transformation généralisée

On souhaite engendrer un couple (x_1, x_2) de variables aléatoires gaussiennes indépendantes, de densité de probabilité

$$p(x_1, x_2) \;=\; \frac{1}{2\pi}\,\exp\left(-\frac{x_1^2}{2} - \frac{x_2^2}{2}\right). \qquad (\mathcal{G})$$

1) On introduit les coordonnées polaires (r, θ) : $x_1 = r\cos\theta$ et $x_2 = r\sin\theta$. Quelle est la densité de probabilité jointe $q(r, \theta)$ pour le couple (r, θ) ?

2) Pour la suite, on fait le choix de prendre r et θ indépendantes. Quelles sont alors les expressions des densités de probabilité $p_R(r)$ et $p_\Theta(\theta)$?

3) Vérifier que les densités $p_R(r)$ et $p_\Theta(\theta)$ sont convenablement normalisées.

4) Comment peut-on engendrer θ suivant la distribution p_Θ voulue ?

5) Même question pour la distance à l'origine r.

6) Justifier la procédure suivante pour créer x_1 et x_2 indépendantes et distribuées suivant la loi $(\mathcal{G})$:

- échantillonner y_1 et y_2 uniformément dans $[0, 1]$
- calculer $\quad x_1 = \sqrt{-2\ln(y_1)}\,\cos(2\pi y_2)$
  $\quad\quad\;\; x_2 = \sqrt{-2\ln(y_1)}\,\sin(2\pi y_2).$

7) Quel peut-être l'intérêt informatique de cette méthode ?

8) Généraliser la procédure au cas où l'on souhaite obtenir x_1 et x_2 gaussiennes de moyennes ($\langle x_1 \rangle$ et $\langle x_2 \rangle$) et d'écart type (σ_1 et σ_2) quelconques.

9) Proposer succinctement une méthode permettant d'engendrer un couple de variables gaussiennes corrélées.

Exercice VIII

On dispose d'un générateur de nombres aléatoires uniformément répartis dans l'intervalle [0,1]. Soit y la variable aléatoire correspondante. A partir de y, on souhaite engendrer une variable aléatoire $x \in [0, \infty[$ de densité de probabilité f exponentielle :

$$f(x) = C \exp(-\lambda x) \quad \text{où} \quad \begin{cases} \lambda \text{ est le paramètre de la distribution} \\ C \text{ est une constante.} \end{cases}$$

1) Rappeler *brièvement* les propriétés attendues pour le générateur de la variable y.

2) Préciser la valeur de C.

3) Quelle doit être la relation $x(y)$ pour que x ait la densité de probabilité f ?

4) Dans le cas où x varie dans l'intervalle $[0, A]$, quelle est la valeur de C ?

5) Reprendre la question 3) dans le cas précédent.

6) Pour engendrer x de densité exponentielle dans $[0, A]$, on pourrait avoir recours à la méthode du rejet dont le préalable consisterait à échantillonner x uniformément dans $[0, A]$, et z uniformément dans un intervalle I. Plusieurs intervalles I conviennent *a priori*.

a) Quelle propriété doit vérifier chacun de ces intervalles ?

b) Quel est l'intervalle le plus avantageux (*i.e.* celui dont la longueur est minimale) ?

c) Exprimer en fonction de A et λ le taux de rejet correspondant (rapport entre le nombre de couples (x, z) "convenables" et le nombre total de couples échantillonnés). Conclusion ?

7) Proposer une méthode permettant d'engendrer, à partir de la variable y uniformément répartie dans [0,1], une variable aléatoire $\alpha \in [0, \infty[$, de densité de probabilité

$$g(\alpha) = 2\,\alpha\, e^{-\alpha^2}.$$

8) Justifier l'assertion suivant laquelle "un processus exponentiel est dépourvu de mémoire". On pourra calculer la *probabilité conditionnelle* de l'événement $x \in [x_1, x_2]$, *sachant* que $x \geqslant x_1$.

Exercice IX : Distribution du produit de nombres aléatoires

La question 2 fait appel au théorème de la limite centrale, dont l'énoncé est rappelé page 31

1) Préliminaires

On vous propose le jeu suivant : si un dé à six faces, non truqué, tombe sur les valeurs 5 ou 6, votre mise initiale est multipliée par 6. Lorsque le dé tombe sur les valeurs 1, 2, 3 ou 4, votre mise est divisée par 3. La somme obtenue est alors remise en jeu pour une seconde partie, et ainsi de suite.

a) Dressez l'inventaire des situations possibles après une partie. Quelle est la probabilité d'avoir augmenté sa mise (situation de gain) après une partie ?

b) De même, quelle est la probabilité de gain après 2 parties ? Après 3 parties ? Il peut être utile de recenser les cas possibles et les probabilités associées.

c) Pensez-vous avoir intérêt à jouer un grand nombre de parties ? La justification rigoureuse de votre choix fait l'objet de la suite du problème.

2) On considère un ensemble de N variables aléatoires indépendantes strictement positives $x_1, x_2 \ldots, x_N$. Toutes ces variables ont même loi de probabilité, et l'on suppose l'existence de $\langle \ln x_1 \rangle$ et de $\langle (\ln x_1)^2 \rangle$. On s'intéresse à la distribution du produit

$$p = \prod_{i=1}^{N} x_i, \quad \text{et l'on note} \quad \begin{cases} m = \langle \ln x_1 \rangle = \langle \ln x_2 \rangle = \ldots = \langle \ln x_N \rangle \\ \sigma^2 = \langle (\ln x_1)^2 \rangle - \langle \ln x_1 \rangle^2 \end{cases}$$

a) Quelle est la valeur moyenne $\langle p \rangle$?

b) Donner l'expression de l'espérance et de la variance de la variable aléatoire $s = (\ln p)/N$.

c) Lorsque N devient grand, quelle est l'expression de la densité de probabilité de s, que l'on notera $g(s)$?

d) Soit $\widetilde{p}$ la variable aléatoire définie par $\widetilde{p} = p^{1/N}$. Quelle est la relation entre la densité de probabilité de $\widetilde{p}$, notée $f(\widetilde{p})$, et $g(s)$?

e) Donner l'expression de la densité $f(\widetilde{p})$ lorsque N devient grand.

f) On note $\widetilde{p}^*$ la valeur de $\widetilde{p}$ la plus probable dans la limite $N \to \infty$. Montrer que

$$\widetilde{p}^* = e^{\langle \ln x \rangle}.$$

g) Lorsque N croît, est-il possible que $\langle p \rangle$ tende vers l'infini tout en ayant $\widetilde{p}^* < 1$?

h) Donner l'expression de $\langle p \rangle$ et $\widetilde{p}^*$ lorsque les différentes variables x_i n'ont pas toutes la même densité de probabilité.

3) Comment s'interprètent ces résultats dans le contexte du jeu introduit dans la partie préliminaire ? Quel est le gain moyen après N tirages ? Comparer avec le gain le plus probable que l'on calculera. Finalement, doit-on accepter de jouer ? On pourra considérer ici que p^*, la valeur la plus probable de p, se comporte pour les grandes valeurs de N comme $(\widetilde{p}^*)^N$.

4) On peut reprendre le problème du jeu de dés sans faire appel au théorème de la limite centrale.

a) Quelles sont les différentes valeurs possibles pour le gain après N parties ? Préciser les probabilités associées.

b) En déduire la valeur du gain le plus probable lorsque N devient grand. Retrouve-t-on les résultats de la question **3)** ?

Exercice X : corrélation et causalité

Les deux affirmations suivantes sont extraites d'articles récents parus dans la presse quotidienne "du soir". Quels commentaires vous inspirent-elles ?

1) Les compagnies d'assurance ont constaté que 80% des accidents de voiture se produisaient à moins de 30 km du domicile du conducteur. On en conclut à un relâchement de la vigilance sur les trajets de proximité.

2) Il a été prouvé que l'espérance de vie des personnes pratiquant le jogging à 60 ans était significativement supérieure à celle de la population générale du même âge. Ce qui démontre le bénéfice de cette activité.

1.4 Calcul d'intégrales

Exercice I

Pour évaluer l'intégrale $\mathcal{I}$ d'une fonction f entre a et b, on subdivise $[a,b]$ en N segments de même longueur, et on emploie une méthode donnant le résultat $\mathcal{I}_N$, qui diffère de $\mathcal{I}$ d'une erreur en $1/N^3$. Proposer à partir de $\mathcal{I}_N$ et $\mathcal{I}_{2N}$, une approximation de $\mathcal{I}$ d'erreur en $1/N^4$ *a priori*.

Exercice II

On souhaite calculer les intégrales du type $\int_{-\infty}^{\infty} f(x)\,\exp(-x^2)\,dx$ par la méthode de quadrature, en se contentant de $N=2$ termes dans l'approximation

$$\int_{-\infty}^{\infty} f(x)\,e^{-x^2}\,dx \;\simeq\; \sum_{i=1}^{2} w_i\,f(x_i). \tag{4}$$

1) Si f est une fonction polynomiale, préciser le degré maximal pour lequel l'approximation (4) est exacte.

2) Donner l'expression des trois premiers polynômes de HERMITTE $P_0(x)$, $P_1(x)$ et $P_2(x)$.

3) En déduire les poids w_1 et w_2 associés à P_2. Quelles sont les valeurs x_1 et x_2 correspondantes ?

4) Vérifier explicitement le résultat énoncé à la question 1.

Exercice III

On rappelle la relation de récurrence entre les polynômes de HERMITTE :

$$P_{j+1}(x) \;=\; 2xP_j(x) \;-\; 2j\,P_{j-1}(x).$$

1) Lorsque a est un réel positif, donner l'expression de $\displaystyle\int_{-\infty}^{+\infty} x^p\,\exp(-ax^2)\,dx$, pour $p=0$, 2, 4 et 6.

2) Proposer une méthode de quadrature qui permette de calculer *exactement* les intégrales du type

$$\int_{-\infty}^{+\infty} f(x)\,\exp(-x^2)\,dx$$

pour toutes les fonctions f polynomiales de degré inférieur ou égal à 7. On pourra remarquer que le polynôme dérivé $P_j'(x)$ est proportionnel à $P_{j-1}(x)$.

3) Vérifier explicitement le résultat précédent dans le cas où f est un polynôme quelconque de degré 7.

1.5 Equations différentielles ordinaires

Exercice I : Saute-mouton et oscillateur harmonique

On applique la méthode du "saute-mouton" à deux pas à l'oscillateur harmonique. On note $\vec{y}$ le vecteur colonne formé de la position x et de la vitesse $v=\dot{x}$ de l'oscillateur, dont la pulsation propre est ω_0.

1) Etablir que $\dfrac{d\vec{y}}{dt} = \mathbb{L}\,\vec{y}$ où $\mathbb{L}$ est une matrice que l'on précisera.

2) Quelle est la relation entre $\vec{y}_{n+1}$, $\vec{y}_n$, $\vec{y}_{n-1}$ et $\mathbb{L}$?

3) De même, quelle est l'équation d'évolution du vecteur erreur, que l'on note $\vec{e}_n$ au pas n ?

4) Afin d'étudier la stabilité de la méthode, on définit $\vec{\eta}_n \equiv \begin{pmatrix} \vec{e}_n \\ \vec{e}_{n-1} \end{pmatrix}$. Déterminer la matrice de gain $\mathbb{G}$ telle que

$$\vec{\eta}_{n+1} \;=\; \mathbb{G}\,\vec{\eta}_n . \tag{5}$$

5) Montrer que les valeurs propres g de $\mathbb{G}$ sont solutions de

$$g^4 \;+\; g^2\left[4(\omega_0\Delta t)^2 \;-\; 2\right] \;+\; 1 \;=\; 0. \tag{6}$$

On pourra utiliser que pour des matrices carrées $\mathbb{A}, \mathbb{B}, \mathbb{C}, \mathbb{D}$ de même dimension

$$\det \begin{pmatrix} \mathbb{A} & \mathbb{B} \\ \mathbb{C} & \mathbb{D} \end{pmatrix} \;=\; \det\left(\mathbb{A}\mathbb{D} \;-\; \mathbb{B}\mathbb{C}\right) \quad \text{lorsque } \mathbb{C} \text{ et } \mathbb{D} \text{ commutent.} \tag{7}$$

6) Que peut-on en conclure sur la stabilité de l'algorithme du saute-mouton appliqué à l'oscillateur harmonique ?

1.6 Equations aux dérivées partielles

Exercice I

Considérons l'équation d'onde

$$\frac{\partial^2 u}{\partial t^2} \;=\; c^2\,\frac{\partial^2 u}{\partial x^2} \tag{8}$$

pour un champ $u(x,t)$. En posant

$$\left\{ \begin{array}{l} r = \partial_t u \\ s = c\,\partial_x u \end{array} \right. \qquad \text{et} \qquad \vec{u} = \begin{pmatrix} r \\ s \end{pmatrix} \tag{9}$$

mettre l'équation d'onde sous forme conservative :

$$\frac{\partial \vec{u}}{\partial t} + \operatorname{div} \vec{\jmath} = \vec{0}, \tag{10}$$

où le courant $\vec{\jmath}$ est relié à $\vec{u}$ par une matrice $\mathbb{C}$ dont on précisera les éléments ($\vec{\jmath} = \mathbb{C}\,\vec{u}$).

L'équation (8) est-elle de nature elliptique ? parabolique ? hyperbolique ? Donner enfin la forme générale de ses solutions.

Exercice II : Méthode FTCS et analyse de stabilité

Soit l'équation d'advection

$$\frac{\partial u}{\partial t} \;=\; -c\,\frac{\partial u}{\partial x} \tag{11}$$

que l'on cherche à résoudre par la méthode FTCS.

1) Pourquoi l'équation (11) est-elle une équation hyperbolique cachée ?

2) Quelle forme prend-elle, écrite en terme des variables discrétisées $u_j^n \equiv u(x_j, t_n)$?

3) En décomposant u_j^n en série de FOURIER

$$u_j^n \;=\; \sum_k \widehat{\mathcal{U}}_k^n \, e^{ikx_j}, \tag{12}$$

exprimer le facteur de gain $g(k)$ défini par

$$\widehat{\mathcal{U}}_k^{n+1} \;=\; g(k)\; \widehat{\mathcal{U}}_k^n. \tag{13}$$

4) Calculer le module de $g(k)$ et conclure sur la stabilité de la méthode proposée.

Exercice III : résolution d'une EDP elliptique par transformation de Fourier

On souhaite résoudre numériquement l'équation aux dérivées partielles

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -\rho(x,y) \tag{14}$$

où le terme de source $\rho(x,y)$ est une donnée du problème. On s'intéresse au cas où u vérifie les conditions de NEUMANN sur la frontière carrée du domaine $[0,L]\times[0,L]$. Les axes des x et y sont discrétisés avec le même pas Δl, pour donner un maillage $N \times N$. Dans ces conditions, et avec les notations usuelles, on rappelle la décomposition en série de FOURIER du signal $u(x,y)$:

$$u_{kl} \;=\; \frac{1}{2}\,\widehat{U}_{00} + \frac{4}{N^2}\sum_{m=1}^{N-1}\sum_{n=1}^{N-1} \widehat{U}_{mn}\, \cos\left[\frac{\pi km}{N}\right]\, \cos\left[\frac{\pi ln}{N}\right],$$

qui s'inverse en

$$\widehat{U}_{mn} \;=\; \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} u_{kl}\, \cos\left[\frac{\pi km}{N}\right]\, \cos\left[\frac{\pi ln}{N}\right].$$

1) Préciser la relation entre les composantes de FOURIER de u et celles de ρ.

2) Proposer une méthode de résolution de l'équation (14). On prendra soin de la formuler de manière à ce qu'elle soit directement utilisable dans un programme informatique.

On rappelle que $\quad \cos p + \cos q \;=\; 2\,\cos\left[\dfrac{p+q}{2}\right]\cos\left[\dfrac{p-q}{2}\right]$

Exercice IV

Le but de cet exercice est de montrer que deux équations différentielles admettant la même solution peuvent être l'une stable et l'autre instable vis-à-vis d'une méthode de résolution numérique donnée.

On considère les deux équations différentielles suivantes :

$$\begin{aligned} (\mathcal{E}1)\;:\quad & \frac{dy}{dt} \;=\; -y^2 \\ (\mathcal{E}2)\;:\quad & \frac{dy}{dt} \;=\; \frac{5y}{t+1} - \frac{6}{(t+1)^2}. \end{aligned}$$

1) Donner les solutions générales de $(\mathcal{E}1)$ et $(\mathcal{E}2)$.

2) Montrer que dans le cas de $(\mathcal{E}1)$, si la condition initiale $y(0)$ est modifiée, cela ne change pas le comportement de la solution $y(t)$ aux temps longs. Qu'en est-il pour $(\mathcal{E}2)$? Quelle indication peut-on en tirer concernant la résolution numérique des problèmes $(\mathcal{E}1)$ et $(\mathcal{E}2)$?

3) Montrer que lorsque $y(0) = 1$, les solutions de $(\mathcal{E}1)$ et $(\mathcal{E}2)$ coïncident pour $t \geqslant 0$.

4) On souhaite intégrer numériquement $(\mathcal{E}1)$ avec l'algorithme d'EULER. La méthode est-elle stable ?

5) De même, l'algorithme d'EULER appliqué à $(\mathcal{E}2)$ est-il stable ?

Exercice V : Méthode de Lax

Bien que l'on sache en trouver une solution analytique, on s'intéresse à la résolution numérique de l'équation d'advection

$$\frac{\partial u}{\partial t} = v \frac{\partial u}{\partial x}, \tag{15}$$

où v est une constante positive.

1) Pour quelle raison "algorithmique" ?

2) Justifier l'appartenance de cette équation à la famille des équations aux dérivées partielles hyperboliques.

3) On discrétise la variable d'espace x en introduisant le pas Δx (de même avec le temps en introduisant le pas Δt). On note $u_j^n \equiv u(x_j, t_n)$ avec $x_j = x_0 + j\Delta x$ et $t_n = t_0 + n\Delta t$. Rappeler brièvement les différentes approximations mises en jeu pour obtenir la méthode discrétisée dite "FTCS" :

$$u_j^{n+1} = u_j^n + \frac{v\Delta t}{2\Delta x} \left(u_{j+1}^n - u_{j-1}^n\right). \tag{16}$$

Le schéma FTCS est-il stable (on se contentera d'une brève justification) ?

4) La méthode de LAX consiste à modifier (16) en remplaçant le terme u_j^n dans le membre de droite par la valeur moyenne

$$\frac{1}{2}\left(u_{j+1}^n + u_{j-1}^n\right). \tag{17}$$

Cette modification n'est pas sans conséquences sur la stabilité de l'algorithme. Pour mettre en place l'analyse de VON NEUMANN, on décompose formellement le signal $u(x,t)$ en série de FOURIER :

$$u_j^n = \sum_k \widehat{\mathcal{U}}_k^n \, e^{ikx_j}. \tag{18}$$

a) Quelle est la relation entre $\widehat{\mathcal{U}}_k^{n+1}$ et $\widehat{\mathcal{U}}_k^n$?

b) En déduire les conditions pour lesquelles la méthode de LAX est stable. Cette dernière représente-t-elle un gain par rapport au schéma FTCS ?

c) À l'aide de la solution générale de (15) que l'on rappellera, comment peut-on interpréter physiquement la condition de stabilité obtenue précédemment ?

d) Donner un développement limité à l'ordre k^2 de $|\widehat{\mathcal{U}}_k^{n+1}/\widehat{\mathcal{U}}_k^n|$, au voisinage de $k = 0$. Comment peut-on interpréter cette quantité ?

5) On cherche dans cette question à établir dans quelle mesure la forme discrétisée de LAX est une approximation convenable de (15) [en d'autres termes, la limite du continu $\Delta t \to 0$ et $\Delta x \to 0$ redonne-t-elle bien (15) ?]

a) Réécrire la forme de LAX pour (15) [obtenue avec la substitution proposée à la question 4)], sous la forme

$$\frac{1}{\Delta t}\left(u_j^{n+1} - u_j^n\right) = \alpha \frac{1}{\Delta x}\left(u_{j+1}^n - u_{j-1}^n\right) + \beta \frac{1}{(\Delta x)^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right). \qquad (19)$$

On déterminera l'expression des coefficients α et β.

b) En déduire l'équation aux dérivées partielles obtenue par passage à la limite continue. Peut-on en proposer une solution analytique pour des conditions initiales particulières ? S'agit-il de l'équation d'advection de départ ?

c) Quelle condition doit vérifier β pour que la méthode de LAX donne de bons résultats ? Montrer que cette condition impose une borne inférieure pour Δt (lorsque Δx est supposé fixé) alors qu'inversement, la condition de stabilité imposait une borne supérieure pour Δt.

1.7 Exercices divers

Exercice I

1) Pour quelle raison l'évaluation numérique de la différence de deux réels proches pose t-elle problème ?

2) On veut mettre en place une méthode de résolution numérique d'équations du second degré en X, de la forme

$$a\,X^2 + b\,X + c = 0 \quad \text{où} \quad \begin{cases} (a,b,c) \in \mathbb{R}^3 \\ a \neq 0,\, b \neq 0. \end{cases}$$

Ecrire l'expression des deux racines X_1 et X_2 de l'expression précédente. Quel problème numérique poserait un calcul direct de ces deux racines, dans le cas où $|ac| \ll b^2$?

3) Lorsque $b > 0$, laquelle de ces deux racines peut-on calculer numériquement avec une bonne précision ?

4) Pour obtenir l'autre racine, on remarque que l'évaluation du produit P des racines n'est *a priori* pas source d'ennui. Pour $b > 0$ et $|ac| \ll b^2$, proposer alors une méthode explicite fournissant les deux racines en question, sans nécessiter le calcul de la différence de deux réels proches.

5) Généraliser la méthode au cas où le signe de b est quelconque.

Exercice II : Elimination de GAUSS et erreurs d'arrondi

Certains systèmes mathématiquement équivalents ne le sont pas nécéssairement d'un point de vue numérique. L'inexactitude d'un résultat peut provenir de la nature du problème lui-même, ou de la manière de le traiter. Nous allons nous intéresser à ce second cas de figure en considérant le système linéaire

$$\begin{cases} 10^{-9}x + y = 1 \\ x + y = 2. \end{cases} \qquad (S)$$

1) Donner la solution exacte du problème.

2) Avec une précision $p = 10^{-8}$, comment ce résultat "s'arrondit"-il ?

3) On considère une méthode d'élimination de GAUSS "du pauvre", sans pivot, c'est-à-dire en s'interdisant d'interchanger les lignes ou les colonnes.

a) Effectuer pas à pas les différentes étapes que suivrait une hypothétique machine de précision $p = 10^{-8}$.

b) Quelle solution obtient-on alors pour le système (S) ? Comparer au résultat exact.

4) Reprendre la question 3) avec la méthode d'élimination vue en cours. Conclusion ?

Exercice III : Estimation de π et accélération de convergence

On s'intéresse à la série de terme général

$$a_n \;=\; 2^n \sin\left(\frac{\pi}{2^n}\right).$$

1) Compte tenu de la relation $\sin^2 x = \dfrac{1}{2} - \dfrac{1}{2}\sqrt{1 - \sin^2 2x}$, valable pour $x \in [-\dfrac{\pi}{4}, \dfrac{\pi}{4}]$, donner les valeurs exactes de a_1, a_2 et a_3.

2) Lorsque n devient grand, calculer a_n avec une précision d'ordre 4^{-2n} (inclus).

3) On souhaite construire une suite $(b_n)_{n\geqslant 1}$ dont la limite soit celle de $(a_n)_{n\geqslant 1}$, mais de telle sorte de cette limite soit atteinte plus rapidement (c'est-à-dire que pour une valeur de n donnée –et *a priori* élevée–, l'écart à la limite soit plus faible en valeur absolue). On définit b_n par

$$b_n \;=\; \alpha\, a_{n+1} \;+\; \beta\, a_n.$$

Calculer les valeurs de α et β.

4) On construit une nouvelle suite $(c_n)_{n\geqslant 1}$ dans un même souci d'accélération de convergence. Pour ce faire, on considère

$$c_n \;=\; \gamma\, b_{n+1} \;+\; \delta\, b_n$$

Calculer γ et δ en imposant de nouveau que $(c_n)_{n\geqslant 1}$ ait la même limite que $(a_n)_{n\geqslant 1}$.

5) Donner les valeurs exactes de b_1, b_2 et c_1. Evaluer ensuite numériquement ces quantités et en particulier, calculer l'écart à leur limite commune. Quel est l'intérêt informatique de la suite $(c_n)_{n\geqslant 1}$?

Exercice IV : Format double précision

Avec le format "double précision", les réels sont codés sur 64 bits, répartis en un bit de signe, 11 bits pour l'exposant, et 52 pour la mantisse.

1) On se place dans l'hypothèse d'une machine fonctionnant suivant le même principe que la machine 32 bits étudiée en cours (ce qui signifie en particulier que la longueur effective de la mantisse est de 53 bits). Quelle doit être la valeur du biais e_0 pour que le plus petit réel positif codable soit (*grosso modo*) l'inverse du plus grand réel ? Préciser les valeurs de ces nombres.

2) Dans la norme *IEEE*, les réels sont écrits sous la forme

$$x \;=\; (-1)^s\,(1+m)\,2^{\,e-e_0}.$$

Le signe s est codé sur 1 bit, et la mantisse m sur 52, sans hypothèse supplémentaire. Il est en particulier possible ici d'avoir $m = 0$. Reprendre les questions précédentes avec ce nouveau format, où l'exposant est de nouveau stocké sur 11 bits.

$\diamond$

2 TD/Cours

2.1 Introduction aux réseaux de neurones : le perceptron

Bien que le fonctionnement du cerveau humain soit de mieux en mieux connu, les mécanismes du transfert d'informations entre quelque 10^{11} cellules (neurones) au travers de 10^{14} jonctions (les synapses) sont extrêmement complexes et mal compris. De nombreux scientifiques travaillent à l'élaboration de programmes informatiques s'inspirant du fonctionnement et de l'architecture du cerveau. Les algorithmes ainsi obtenus sont appelés *réseaux de neurones* et leurs implémentations "hardware" constituent des *ordinateurs neuronaux* (de caractéristiques très différentes de celles des ordinateurs traditionnels).

Un réseau de neurones ne fonctionne pas comme un programme ordinaire, mais "apprend" à partir d'exemples, en ajustant le poids de ses synapses. Après la phase d'apprentissage, il peut "généraliser", c'est-à-dire déduire une règle des exemples qui lui ont été fournis. Il ne fonctionne pas suivant la logique stricte du "oui ou non" mais suivant celle du "plus ou moins".

Les réseaux de neurones trouvent des applications dans des domaines très variés : reconnaissance vocale, identification des défauts d'un moteur à partir du bruit émis par celui-ci, tentative de prédiction des cours de la bourse, élaboration d'automates pouvant jouer aux échecs ou au backgammon contre les humains... Le réseau le plus simple est le perceptron, qui a beaucoup été étudié depuis les années 1960.

Le perceptron

Il est constitué de N neurones S_i, $1 \leqslant i \leqslant N$, dits neurones d'entrée, N poids synaptiques w_i, et un neurone de sortie S_0 directement connecté à tous les neurones d'entrée par l'intermédiaire des synapses. Comme dans les cellules nerveuses, S_0 réagit à la somme des activités de chacun des S_i, pondérées par les coefficients w_i. Ces derniers modélisent la "force" avec laquelle le signal arrivant de S_i est converti en potentiel électrique dans le noyau de S_0. Les w_i sont des nombres réels décrivant un processus biochimique complexe et une jonction synaptique peut avoir un effet stimulant ($w_i > 0$) ou inhibant ($w_i < 0$).

Modélisation

Si la somme des $w_i S_i$ est positive, le neurone de sortie est pris actif ($S_0 = 1$), et inactif dans le cas contraire ($S_0 = -1$) :

$$S_0 \;=\; \text{signe}\left(\sum_{i=1}^{N} w_i S_i\right) \qquad (\mathcal{R}_{\vec{w}})$$

S_0 est ainsi une fonction booléenne de $\{-1,1\}^N \to \{-1,1\}$. On adoptera la notation vectorielle : $\overrightarrow{S} = (S_1, S_2, \dots\ S_N)$, pour écrire la relation $(\mathcal{R})$ sous la forme $S_0 = \text{signe}(\overrightarrow{w} \cdot \overrightarrow{S})$. Géométriquement parlant, le perceptron fonctionnant suivant la règle $(\mathcal{R}_{\vec{w}})$ sépare les sorties $S_0 = 1$ et $S_0 = -1$ par l'hyperplan d'équation $\overrightarrow{w} \cdot \overrightarrow{S} = 0$.

Pour que le perceptron puisse "apprendre" et "généraliser", il faut lui fournir des exemples. Dans notre cas, il s'agit d'un jeu de M entrées $\overrightarrow{S}^{\alpha}$ et M sorties S_0^{α} avec $1 \leqslant \alpha \leqslant M$. Considérons un perceptron "professeur" fonctionnant suivant la règle $\mathcal{R}$ avec un vecteur de poids synaptiques $\overrightarrow{w}_*$ (règle $\mathcal{R}_{\vec{w}_*}$). On souhaite qu'après la phase d'apprentissage, un perceptron "élève" de vecteur $\overrightarrow{w}$ initialement arbitraire (*e.g.* $\overrightarrow{w} = \overrightarrow{0}$) puisse se comporter comme le professeur, c'est-à-dire avoir un neurone de sortie S_0 dans l'état $\text{signe}(\overrightarrow{w}_* \cdot \overrightarrow{S})$ pour toute entrée $\overrightarrow{S}$. Pour ce faire, il est nécessaire de corriger l'élève lorsqu'il fournit une mauvaise réponse, ce qui est réalisé en modifiant légèrement ses poids synaptiques.

Règle d'apprentissage

Soit $(\overrightarrow{S}^{\alpha}, S^{\alpha}_{0,*})$ le $\alpha^{\text{ième}}$ exemple proposé au perceptron. Par définition,

$$S^{\alpha}_{0,*} = \text{signe}(\overrightarrow{w}_* \cdot \overrightarrow{S}^{\alpha}). \tag{1}$$

Une mauvaise réponse est donnée lorsque $S^{\alpha}_{0,*}\overrightarrow{w} \cdot \overrightarrow{S}^{\alpha} < 0$ et on choisit dans ce cas de modifier $\overrightarrow{w}$ d'une quantité

$$\Delta\overrightarrow{w} = \frac{1}{N} S^{\alpha}_{0,*} \overrightarrow{S}^{\alpha}. \tag{$\mathcal{A}$}$$

Si la réponse donnée est correcte, les poids des synapses ne sont pas modifiés. Lorsque le $\alpha^{\text{ième}}$ exemple a été appris, on passe au suivant. Cette règle d'apprentissage a été proposée dans les années 1960 par ROSENBLATT.

1) Lorsqu'il s'agit d'apprendre un seul exemple $\overrightarrow{S}^{1}$, montrer que le perceptron élève se trompe une fois au plus. On suppose qu'initialement, $\overrightarrow{w} = \overrightarrow{0}$.

2) Dans le cas général où l'on souhaite que l'élève apprenne la règle $\mathcal{R}_{\vec{w}_*}$, nous allons montrer que seul un nombre borné d'exemples est nécessaire, après quoi le perceptron donne toujours la bonne réponse.

a) Notons $\vec{z}^{\alpha} \equiv S^{\alpha}_{0,*}\overrightarrow{S}^{\alpha}$. Etablir l'existence d'une constante c strictement positive telle que $\forall \vec{z}^{\alpha}$, $\overrightarrow{w}_* \cdot \vec{z}^{\alpha} \geqslant c$

b) La variable t compte le nombre d'erreurs commises lors de l'apprentissage, c'est-à-dire le nombre de fois où les poids synaptiques ont été changés. L'algorithme commence avec $\overrightarrow{w}(t=0) = \overrightarrow{0}$. Quelle est la relation entre $\overrightarrow{w}(t+1)$ et $\overrightarrow{w}(t)$?

c) En déduire la relation

$$\overrightarrow{w}^2(t) \leqslant \frac{t}{N}. \tag{2}$$

d) Montrer de même que l'on peut écrire

$$\overrightarrow{w}_* \cdot \overrightarrow{w}(t) \geqslant \frac{ct}{N}. \tag{3}$$

e) A l'aide des inégalités (2), (3), et du théorème de SCHWARZ, établir finalement la majoration

$$t \leqslant N\frac{|\overrightarrow{w}_*|^2}{c^2}. \tag{4}$$

Quelle conclusion peut-on en tirer ?

Efficacité du perceptron : voir le TP de langage C pour l'apprentissage d'un "rythme" (suite périodique de -1 et 1), la "prédiction" d'une série pseudo-aléatoire...

2.2 Erreurs machine et précision des calculs

Les quelques rappels qui suivent concernant la représentation interne des nombres par les ordinateurs font partie des détails qu'il est utile de garder à l'esprit pour maintenir arbitrairement petites les erreurs d'arrondis.

Dans une machine 32-bits, un nombre réel x est codé en base 2, écrit sous la forme

$$x = \pm m\, 2^{\,e-e_0},$$

et stocké comme suit :

$\pm$	$\boxed{e \text{ (exposant)}}$	$\boxed{m \text{ (mantisse)}}$		
1 bit	sur 8 bits	sur 23 bits	$\longrightarrow$	32 bits au total

Le biais e_0 est un entier positif fixé propre à chaque machine, qu'il faut ajouter à l'exposant "réel" $e-e_0$ pour que e soit positif. De plus, la mantisse m est normalisée, c'est-à-dire translatée vers la gauche autant que possible pour avoir toujours un 1 en première position. Chaque "saut" vers la gauche fait baisser e d'une unité, de telle sorte que le produit $m\, 2^{\,e}$ reste constant. Le bit de gauche dans l'écriture binaire de la mantisse étant nécessairement 1, il n'est pas utile de le garder en mémoire, ce qui permet un saut supplémentaire vers la gauche (et donc un gain en précision). La longueur effective de la mantisse est ainsi de 24 bits.

Exercice

Pour une machine 32-bits de biais $e_0 = 151$:

- ◇ quel est le plus petit réel positif non nul ?
- ◇ quel est le plus grand entier ?
- ◇ quelle est la représentation interne de $x = 0.25$?

◇

Avant toute addition ou soustraction, les exposants des deux arguments sont égalés. Pour cela, le plus petit exposant est augmenté et sa mantisse translatée par conséquent vers la droite. Les bits ainsi "expulsés" sont perdus, d'où la perte de précision[1].

Exercice

On définit la précision d'un ordinateur comme le plus petit réel p positif tel que $1 + p \neq 1$.

- ◇ Quelle est la précision d'une machine 32-bits ?
- ◇ De quelle caractéristique interne dépend-elle ?
- ◇ Comparer la précision au plus petit réel positif. Conclusion ?

[1] De plus, la mantisse obtenue après une soustraction n'est en général pas normalisée (le premier bit n'est pas 1). La normalisation consiste à translater les bits en question vers la gauche. Durant cette opération, apparaissent à droite de la mantisse des bits "inconnus" auquel on donne arbitrairement la valeur 0

2.3 Optimisation et méthode du gradient conjugué

La résolution du système linéaire $N \times N$ $\mathbf{A}\,\vec{x} = \vec{b}$ peut être interprétée comme la recherche du vecteur $\vec{x}$ qui minimise la fonction

$$f(\vec{x}) = \frac{1}{2}\left|\mathbf{A}\,\vec{x} - \vec{b}\right|^2, \tag{1}$$

avec la valeur minimale $f = 0$. Différents algorithmes de minimisation peuvent être implémentés pour une fonction scalaire de N variables. Dans le cas présent où f est une forme quadratique en $\vec{x}$, la méthode du *gradient conjugué* est particulièrement efficace. Contrairement aux autres méthodes itératives vues en cours (relaxation de JACOBI, de GAUSS-SEIDEL...), elle ne nécessite aucune inversion de matrice.

Considérons dans un premier temps la méthode moins efficace de *descente la plus raide* (introduite par CAUCHY). En partant d'un point P_0 de coordonnées $\vec{x}_0$, on se déplace dans la direction de la pente la plus raide, c'est-à-dire suivant le gradient local

$$\vec{g}_0 = -\overrightarrow{\text{grad}}\, f(P_0). \tag{2}$$

Comme l'indique la figure 1 (correspondant au cas $N = 2$ où les lignes de niveau de la fonction f sont des ellipses concentriques), cette direction ne mène en général pas au point O qui minimise f. Au mieux peut-on rechercher le point P_1 qui minimise f sur le chemin précédent (droite passant par P_0 et de vecteur directeur $\vec{g}_0$. Arrivé en P_1, on se déplace de nouveau suivant le gradient local $\vec{g}_1$, qui est par construction orthogonal à $\vec{g}_0$ (cf figure 1). En itérant la procédure, on suit un trajet en zigzag qui mène au minimum mais où tous les virages sont à angle droit, ce qui est assez inefficace. Arrivé en P_1, il serait plus judicieux de se déplacer directement suivant $\vec{h}_1$, vers le point O (voir la figure 1). Comment construire $\vec{h}_1$ (appelé gradient conjugué de $\vec{g}_0$ et $\vec{g}_1$) ?

1) Etablir qu'en se déplaçant suivant $\vec{h}_1$, le gradient local $\vec{g} = -\overrightarrow{\text{grad}}\, f$ doit faire un angle constant avec $\vec{h}_1$ (voir la figure 2).

2) En déduire que suivant $\vec{h}_1$, les variations du gradient local sont orthogonales à $\vec{g}_0$. Connaissant $\vec{g}_0$ et $\vec{g}_1$, cette propriété fournit une procédure de construction de $\vec{h}_1$.

3) Donner l'expression de $\vec{g}$ en un point P de coordonnées $\vec{x}$.

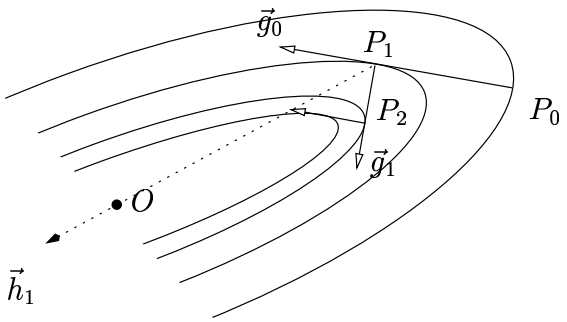

FIG. 1 –

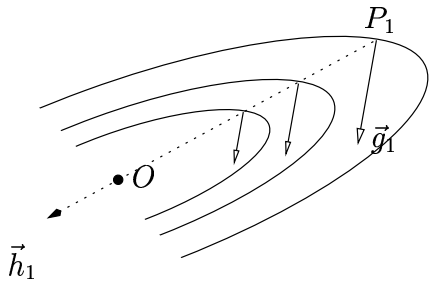

FIG. 2 –

4) Montrer que P_1 a pour coordonnées :

$$\vec{x}_1 \;=\; \vec{x}_0 \;+\; \frac{|\vec{g}_0|^2}{|\mathbf{A}\,\vec{g}_0|^2}\;\vec{g}_0 \tag{3}$$

5) Il est possible de rechercher $\vec{h}_1$ sous la forme d'une combinaison linéaire de $\vec{g}_0$ et $\vec{g}_1$: $\vec{h}_1 = \vec{g}_1 + \alpha\vec{g}_0$. Quelle est alors l'expression de α ?

6) En minimisant f à partir de P_1 suivant $\vec{h}_1$, montrer que l'on obtient le point C_2 de coordonnées

$$\vec{c}_2 = \vec{x}_1 + \frac{\vec{h}_1 \cdot \vec{g}_1}{|\mathbf{A}\,\vec{h}_1|^2}\;\vec{h}_1. \tag{4}$$

Lorsque $\mathbf{A}$ est une matrice 2×2, $\vec{c}_2$ est la solution cherchée. Dans le cas $N > 2$, il faut itérer la procédure en se déplaçant suivant le gradient conjugué de $(\vec{h}_1, \vec{g}_2)$ pour trouver C_3, et ainsi de suite... Le minimum global O est alors trouvé après N pas.

3 Enoncés de problèmes

3.1 Quadrature gaussienne et polynômes de Laguerre

On souhaite calculer numériquement des intégrales du type $\int_0^\infty f(x)\,e^{-x}\,dx$ de telle sorte que le résultat soit exact (modulo la précision de l'ordinateur) pour toutes les fonctions f polynomiales de degré inférieur à une valeur arbitraire. On procède suivant la méthode de quadrature avec une fonction de poids exponentiel, et l'on définit le produit scalaire de deux fonctions f et g par :

$$\langle f|\, g\rangle \equiv \int_0^\infty f(x)\, g(x)\, \exp(-x)\, dx. \tag{1}$$

Il faut ensuite construire une famille de polynômes $\{P_j(x)\}_{j=0,1,2\ldots}$, deux à deux orthogonaux (c'est-à-dire $\langle P_l|\, P_m\rangle = 0$ pour tous les entiers l et m positifs), et tels que $P_j(x)$ soit de degré j exactement. On choisit $P_0(x) = 1$ comme polynôme constant. L'ensemble des polynômes précédents constitue la famille des polynômes de Laguerre.

1) Justifier succinctement que l'équation (1) définit bien un produit scalaire.

2) Soit α un réel positif. Calculer $\int_0^\infty e^{-\alpha x}\,dx$. En déduire que

$$\int_0^\infty x^n\, e^{-x}\, dx \;=\; n! \tag{2}$$

où n est un entier positif quelconque.

3) On cherche la relation entre P_0 et P_1 sous la forme

$$P_1(x) \;=\; (\beta_1 - x)P_0(x). \tag{3}$$

En exprimant la condition d'orthogonalité entre P_0 et P_1, calculer β_1. Donner alors l'expression de $P_1(x)$.

4) Soit x_1^1 la racine de P_1. L'approximation

$$\int_0^\infty f(x)\, e^{-x}\, dx \;\simeq\; w_1^1 f(x_1^1) \tag{4}$$

doit être exacte pour tous les polynômes f de degré inférieur ou égal à N_1.

a) Préciser la valeur de N_1.

b) Comment le coefficient w_1^1 est-il relié à P_0 et P_1 ? Calculer w_1^1.

c) Vérifier que la relation (4) est bien exacte pour les polynômes de degré inférieur à N_1. Qu'en est-il pour un polynôme de degré $N_1 + 1$?

5) On cherche P_2 sous la forme :

$$P_2(x) \;=\; (\beta_2 - x)P_1(x) \;-\; \gamma_2 P_0(x). \tag{5}$$

a) En exploitant la condition $\langle P_0|\, P_2\rangle = 0$, calculer γ_2.

b) Utiliser de même la relation $\langle P_1|\, P_2\rangle = 0$ pour montrer que P_2 s'écrit :

$$P_2(x) \;=\; x^2 \;-\; 4x + 2. \tag{6}$$

c) Donner les deux racines x_1^2 et x_2^2 de $P_2(x)$.

6) La méthode de quadrature fournit l'approximation

$$\int_0^\infty f(x)\, e^{-x}\, dx \;\simeq\; w_1^2 f(x_1^2) \;+\; w_2^2 f(x_2^2). \tag{7}$$

a) Calculer w_1^2 et w_2^2.

b) Pour quels polynômes l'approximation (7) est-elle exacte ? Vérifier explicitement cette affirmation.

7) Comment généraliser la relation (7) au calcul de $\int_0^\infty f(x)\, e^{-\alpha x}\, dx$, avec $\alpha > 0$?

8) Les résultats obtenus permettent-ils d'approximer numériquement des intégrales du type $\int_0^1 h[\ln(t)]\, dt$?

9) Déduire de ce qui précède un schéma d'approximation numérique pour le calcul de $\int_0^\infty g(x)\, dx$. Pour quel type de fonctions g cette méthode est-elle susceptible de donner de bons résultats ?

10) Question subsidiaire : donner l'expression de $P_3(x)$. Plus généralement, quelle est la relation de récurrence entre P_{j+1}, P_j et P_{j-1} ?

$\diamond$

3.2 Recherche de racines d'équations

1) On s'intéresse à une formule itérative $x_{n+1} = g(x_n)$ où $g(x)$ est une fonction quelconque et x_0 une valeur initiale arbitraire. On suppose pour le moment que l'équation $x = g(x)$ admet une solution unique, que l'on note x^*. Si la suite $\{x_n\}$ converge, c'est donc nécessairement vers x^*.

a) Si x_n est proche de x^* (i.e. $x_n = x^* + \epsilon$ où $|\epsilon| \ll |x^*|$), calculer x_{n+1}.

b) En déduire une condition [portant sur la valeur de la dérivée $g'(x^*)$] pour que la suite converge.

c) On suppose la condition précédente remplie. La vitesse de convergence de la suite est d'autant meilleure que le rapport $|x_{n+1} - x^*|/|x_n - x^*|$ est faible. Dès lors, quelle est la valeur optimale de $g'(x^*)$?

2) On souhaite appliquer les considérations précédentes à la recherche de x^*, la racine supposée unique de l'équation $f(x) = 0$. Pour ce faire, on définit la fonction $g(x) = x + \theta f(x)$ où θ est un réel qu'il s'agit de bien choisir.

a) Peut-on toujours trouver une gamme de valeurs possibles pour θ, qui soit telle que la série $x_{n+1} = g(x_n)$ converge dès lors que x_0 est suffisamment proche de x^* ?

b) Exprimer en fonction de x^* la valeur la plus avantageuse pour θ.

3) En déduire un algorithme de recherche numérique de la solution de $f(x) = 0$. On suppose ici que la fonction dérivée f' est connue. Cette hypothèse peu réaliste sera levée par la suite (cf question 6).

4) En supposant x_n proche de x^* et $|h| \ll |x_n|$, calculer $f(x_n + h)$. On suppose également que $f'(x_n)$ est différent de 0. En déduire une interprétation graphique de la valeur choisie à la question 3 pour x_{n+1}.

5) Pour quelle raison la méthode précédente –dite de NEWTON– est-elle inopérante dans le cas où x^* est une racine multiple de f ? En quoi la fonction $u(x) = f(x)/f'(x)$ permet-elle de lever la difficulté ?

6) En pratique, dans la plupart des cas, la fonction dérivée f' est inconnue, et la méthode de Newton doit être modifiée. Une variante possible est fournie par le schéma itératif suivant :

$$x_{n+1} = \frac{x_{n-1}\, f_n - x_n\, f_{n-1}}{f_n - f_{n-1}},$$

où $f_n \equiv f(x_n)$. Justifier cette relation, sur laquelle repose la méthode dite de la sécante.

7) Soit une fonction f dont les valeurs sont connues sur un ensemble discret donné de points x_0, x_1,..., x_N. Proposer *sans calcul* le principe d'une méthode de recherche des solutions de $f(x) = 0$. Cette procédure devra mériter le qualificatif de "méthode d'interpolation inverse".

3.3 Interpolation de Bernstein

On cherche à interpoler une fonction f de la variable réelle x sur l'intervalle $[0,1]$, connaissant les $n+1$ valeurs de f prises aux points régulièrement espacés $x_k = k/n$ (où $k = 0, 1, \ldots n$). A cette fin, on étudie le polynôme interpolateur P_n défini par

$$P_n(x) \;=\; \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\, C_n^k\, x^k\, (1-x)^{n-k}, \tag{1}$$

où les C_n^k sont les coefficients standard du binôme.

1) L'expression (1) permet-elle d'interpoler une fonction définie sur un intervalle quelconque ?

2) Connaissant $f(0)$, $f(1/2)$ et $f(1)$, calculer $P_1(x)$ et $P_2(x)$ pour une fonction f quelconque. Comparer $f(1/2)$ à $P_1(1/2)$ et $P_2(1/2)$. Quel commentaire cela vous inspire t-il ?

3) Dans le cas où la fonction f est constante sur $[0,1]$, donner l'expression prise par $P_n(x)$ pour un degré n arbitraire ($n \geqslant 0$).

4) Etablir l'identité

$$\sum_{k=0}^{n} k\, C_n^k\, x^k\, (1-x)^{n-k} \;=\; n\, x. \tag{2}$$

Quelle conclusion peut-on en tirer sur la qualité de l'interpolation fournie par P_n pour la fonction $f(x) = x$ (avec $n \geqslant 0$ quelconque).

5) On s'intéresse dans cette question à la fonction parabolique $f(x) = x^2$.

a) Calculer P_1 et P_2.

b) A l'aide de la relation (2), exprimer $\sum_{k=0}^{n} k^2\, C_n^k\, x^k\, (1-x)^{n-k}$ sous la forme d'un polynôme du second degré en x.

c) En déduire que sur tout l'intervalle $[0,1]$,

$$|f(x) - P_n(x)| \leqslant \frac{1}{4n}.$$

Cette borne supérieure est-elle atteinte ? Le cas échéant, pour quelle valeur de x ? Conclusion ?

d) A l'aide de $P_{2n}(x)$ et $P_n(x)$, quelle est la meilleure formule d'interpolation que l'on peut proposer pour f ?

6) Pour une fonction f donnée, on approxime $\int_0^1 f$ par $\int_0^1 P_n$. Quelle expression approchée obtient-on pour $\int_0^1 f$? Retrouve t-on une formule classique ? On donne :

$$\int_0^1 x^n\, (1-x)^m\, dx \;=\; \frac{n!\; g(m)}{(n+m+1)!},$$

où g est une fonction que l'on déterminera.

7) Quelle serait l'erreur commise en approximant un polynôme de degré 3 par son interpolation (1) ?

3.4 Intégration numérique de l'équation du déclin

Pour une fonction $y(t)$ et une constante λ positive, on considère l'équation du déclin :

$$\frac{dy}{dt} = -\lambda\, y, \tag{3}$$

Partie A

1) Citer une situation physique dans laquelle cette équation intervient. Préciser alors la signification des variables y, t et λ.

2) Pour résoudre numériquement ce problème, on dicrétise l'axe des t avec un pas Δt. Dans toute la suite, on note $t_n = n\,\Delta t$ et $y_n = y(t_n)$ où $n \in \mathbb{N}$. Proposer une approximation de la dérivée première dy/dt en t_n, d'erreur $\mathcal{O}[(\Delta t)^2]$, faisant intervenir les quantités y_n, y_{n+1} et y_{n+2}.

3) Expliquer succinctement comment construire une approximation de dy/dt en t_n, d'erreur $\mathcal{O}[(\Delta t)^p]$, à l'aide des quantités y_n, y_{n+1},..., y_{n+p}.

Partie B

On cherche à mettre en place une méthode du "saute-mouton" pour l'intégration numérique de l'équation (1), en faisant l'approximation

$$\left.\frac{dy}{dt}\right|_{t=t_n} \simeq \frac{1}{2\Delta t}\,(y_{n+1} - y_{n-1})\,.$$

1) Quelle est la relation entre l'erreur commise lors de l'approximation précédente et le pas Δt ? En déduire l'ordre de la méthode.

2) Déduire de ce qui précède la version discrétisée de (1).

3) La méthode obtenue est-elle implicite ou explicite ? Justifier votre réponse.

4) Soit $y^*(t)$ la solution exacte de (1), associée à une condition initiale donnée. Au pas n, l'algorithme du saute-mouton fournit la solution $y_n = y^*(t_n) + e_n$, où e_n est l'erreur commise. Quelle est la relation entre e_{n-1}, e_n et e_{n+1} ?

5) Exprimer la relation précédente sous la forme matricielle : $\begin{pmatrix} e_{n+1} \\ e_n \end{pmatrix} = \overleftrightarrow{\mathbf{G}} \begin{pmatrix} e_n \\ e_{n-1} \end{pmatrix}$.

6) Montrer que les valeurs propres de $\overleftrightarrow{\mathbf{G}}$ s'écrivent $g_\pm = -\lambda\Delta t \pm \sqrt{(\lambda\Delta t)^2 + 1}$.

7) Conclure quant à la stabilité de l'algorithme appliqué à l'équation du déclin.

Partie C

Dans cette partie, l'analyse de stabilité est reprise dans le cas où l'on emploie la méthode d'intégration de Runge-Kutta d'ordre 2.

1) Quelle est alors la forme discrétisée de (1) ?

2) En déduire la relation entre e_n et e_{n+1}.

3) L'algorithme en question est-il stable ? Le cas échéant, préciser la valeur maximale "autorisée" pour le pas Δt.

3.5 Méthode de prédiction/correction

Partie A : principe

On met en place une méthode de résolution en deux temps pour l'équation différentielle ordinaire

$$\frac{dy}{dt} = f(y). \tag{1}$$

On suppose connues les valeurs y_0, y_1,..., y_n prises par y sur l'axe temporel discrétisé, avec $y_i = y(t_i)$ où $t_i = t_0 + i\Delta t$. On note $f_i = f(y_i)$.

Tout d'abord, on détermine une valeur y^P_{n+1} pour y_{n+1} en intégrant (1) après avoir extrapolé f à partir des valeurs f_n, f_{n-1}, f_{n-2}... (étape de prédiction). Connaissant y^P_{n+1}, on calcule dans un second temps $f^P_{n+1} = f(y^P_{n+1})$ et on réintègre (1) en interpolant f à l'aide des valeurs f^P_{n+1}, f_n, f_{n-1},..., pour en déduire une valeur modifiée de y_{n+1}, plus précise que y^P_{n+1} (étape de correction). On itère ensuite la méthode pour calculer y_{n+2}, y_{n+3},...

1) Connaissant f_n, f_{n-1}, f_{n-2}..., une formule de Newton-Gregory permet d'extrapoler f suivant

$$f(t_n + \tau) = f_n + u\,\nabla f_n + \frac{u(u+1)}{2}\,\nabla^2 f_n + \ldots \tag{2}$$

avec $u = \tau/\Delta t$. On considère ici que $f(y)$ dépend du temps via $y(t)$.

a) Justifier l'appellation "NGB" pour l'approximation précédente.

b) Quelles sont les expressions de ∇f_n et $\nabla^2 f_n$?

c) Comment s'écrit le terme suivant dans le développement (2) ? Donner l'expression de $\nabla^3 f_n$.

2) *Prédiction du second ordre* : on tronque (2) après le terme $u\,\nabla f_n$.

a) Quel est l'ordre en Δt de l'erreur commise ?

b) L'expression correspondante pour f est utilisée pour intégrer l'équation différentielle (1) entre t_n et t_{n+1} et obtenir ainsi une prédiction y^P_{n+1} :

$$y^P_{n+1} = y_n + \Delta t \int_0^1 f(t_n + u\Delta t)\,du. \tag{3}$$

Déterminer y^P_{n+1} en fonction de y_n, f_n et f_{n-1}.

c) Quel est l'ordre en Δt de l'erreur commise pour la prédiction y^P_{n+1} ?

3) *Correction du second ordre*

La valeur y^P_{n+1} sert à déterminer la quantité $f^P_{n+1} = f(y^P_{n+1})$. On peut ensuite s'attendre à obtenir une meilleure approximation pour f sur l'intervalle $[t_n, t_{n+1}]$ en interpolant f autour de t_{n+1} à l'aide de f^P_{n+1} et f_n, plutôt qu'avec l'extrapolation (2) qui faisait intervenir f_n et f_{n-1}.

a) Comment s'écrit (3) dans le cas présent ?

b) Donner l'expression de ∇f_{n+1}.

c) En déduire finalement

$$y_{n+1} = y_n + \frac{\Delta t}{2}\left[f^P_{n+1} + f_n\right] + \mathcal{O}\left[(\Delta t)^3\right]. \tag{4}$$

4) *Prédiction/correction du troisième ordre*

a) On souhaite pousser à l'ordre suivant en Δt la méthode exposée aux questions précédentes. Dans ces conditions, exprimer y^P_{n+1} en fonction de y_n, f_n, f_{n-1} et f_{n-2}.

b) Que devient alors la relation (4) ? Indication : le préfacteur de f_{n-1} est $-\Delta t/12$. Montrer que la somme des différents préfacteurs des f_i mis en jeu est Δt.

c) Proposer succinctement un algorithme de prédiction/correction du troisième ordre pour la résolution numérique d'équations différentielles ordinaires du type $dy/dt = f(y,t)$.

Partie B : application

On souhaite tester l'efficacité de la méthode de prédiction/correction du second ordre sur l'exemple de l'équation $dy/dt = -\lambda y$ où $\lambda > 0$.

1) En posant $\alpha = \lambda \Delta t$, montrer que l'on a

$$y_{n+1} = y_n \left(1 - \alpha + \frac{3\,\alpha^2}{4}\right) - \frac{\alpha^2}{4}\, y_{n-1}.$$

2) Ecrire l'équation de propagation de l'erreur e_n commise sur y_n sous la forme

$$\begin{pmatrix} e_{n+1} \\ e_n \end{pmatrix} = \overleftrightarrow{\mathbf{G}} \begin{pmatrix} e_n \\ e_{n-1} \end{pmatrix}. \tag{5}$$

3) En déduire que le schéma numérique en question est stable pour $\alpha \leqslant 2$. On ne cherchera pas à calculer explicitement les valeurs propres de $\overleftrightarrow{\mathbf{G}}$.

4) Si l'on se contente de la valeur de y_{n+1} donnée par l'étape de prédiction :

$$y^P_{n+1} \;=\; y_n \;+ \frac{\Delta t}{2}\,[3 f_n - f_{n-1}]\,, \tag{6}$$

a) quelle est l'équation vérifiée par les valeurs propres de la matrice de gain $\overleftrightarrow{\mathbf{G}}$?

b) pour quelles valeurs de α l'algorithme correspondant est-il stable ?

c) conclusion ?

$\diamond$

3.6 Un problème de convection-diffusion

Soit l'équation aux dérivées partielles (où C et D sont des constantes)

$$\left.\frac{\partial u}{\partial t}\right|_x = D \left.\frac{\partial^2 u}{\partial x^2}\right|_t + C \left.\frac{\partial u}{\partial x}\right|_t, \qquad (\mathcal{E})$$

1) Dans quel contexte physique une telle relation peut-elle intervenir ? Quelles contraintes pèsent-elles sur les signes de C et D ?

2) L'équation $(\mathcal{E})$ est-elle hyperbolique ? parabolique ? elliptique ? Justifier votre réponse.

3) Afin de la résoudre, on discrétise les axes des positions x et du temps t en introduisant les pas respectifs Δx et Δt. Proposer la discrétisation la plus simple possible qui soit explicite et de type FTCS ("*Forward Time Centered Space*").

4) On s'intéresse désormais à la stabilité de l'algorithme obtenu, en suivant la procédure de VON NEUMANN. Montrer que la fonction d'amplification $g(k)$ vérifie

$$|g(k)|^2 = \left[1 - 4\,a \sin^2\left(\frac{k\Delta x}{2}\right)\right]^2 + 4\,b^2 \sin^2(k\Delta x) \qquad (1)$$

où l'on a posé $\quad a = \dfrac{D\Delta t}{(\Delta x)^2} \quad$ et $\quad b = \dfrac{C\Delta t}{2\Delta x}$.

5) En déduire que l'algorithme est stable pour $2\,(b^2 - a^2)\,\cos^2\left(\dfrac{k\Delta x}{2}\right) \leqslant a - 2a^2$.

6) En étudiant séparément les cas $|b| < a$ et $|b| \geqslant a$, donner la condition sous laquelle l'algorithme est stable. On pourra faire une représentation graphique comme celle proposée sur la figure ci-dessous.

7) La méthode qui précède permet-elle de traiter les cas où C et D sont fonctions de x ?

8) Montrer que modulo un changement de référentiel $(x,t) \to (\widetilde{x},t)$ que l'on précisera, l'équation $(\mathcal{E})$ se met sous la forme d'une équation de diffusion. On résout cette nouvelle équation par la même méthode FTCS que précédemment. Quelle est la condition de stabilité correspondante ? Conclusion.

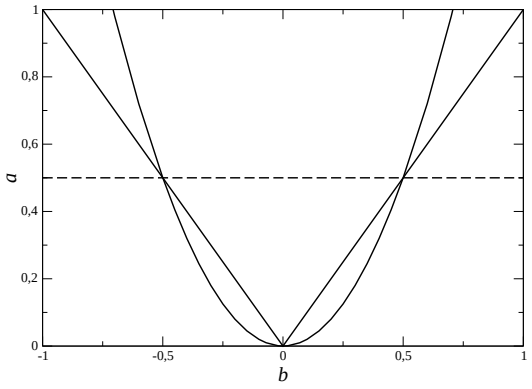

FIG. 3 – Étude graphique (incomplète !) des cas de stabilité

3.7 Quelques variations autour du thème de l'équation de la chaleur

I Problème stationnaire et méthode de relaxation

On cherche à mettre en place une méthode de résolution des équations du second ordre de la forme

$$\frac{d^2y}{dx^2} = s(x,y), \tag{1}$$

qui constituerait alors une alternative à la méthode du tir. La fonction source $s(x,y)$ est donnée, et l'on cherche la solution $y(x)$ associée. La position x varie dans l'intervalle $[0,L]$, qui est discrétisé en M valeurs équiréparties x_1, x_2... x_M. On note y_i la valeur $y(x_i)$ prise en $i = 1...M$. De même, $s_i = s(x_i, y_i)$. Les conditions aux limites sont données [$y(0)$ et $y(L)$ ont des valeurs prescrites qu'il n'est pas utile de préciser].

1) En utilisant l'approximation de NEWTON-GREGORY pour discrétiser la dérivée seconde, quelle forme prend l'équation (1) ? Préciser l'erreur commise *a priori*.

2) Même question avec l'approximation de STIRLING. C'est cette relation que l'on considérera dans la suite

3) L'idée de la méthode de relaxation est d'améliorer par itération une solution d'essai $y^1(x)$, que l'on note vectoriellement sur l'axe des positions discrétisées : $\vec{y}^1 = (y_1^1, y_2^1, ...y_M^1)$. Pour ce faire, on introduit un vecteur $\vec{e}^1$ de composantes

$$e_i^1 = y_{i+1}^1 - 2y_i^1 + y_{i-1}^1 - (\Delta x)^2 s_i, \qquad \text{où} \qquad \Delta x = L/(M-1).$$

Pourquoi est-il légitime de qualifier $\vec{e}^1$ de vecteur d'erreur ? Quels sont les indices i pour lesquels la relation précédente a un sens ? Préciser les valeurs de e_1^1 et e_M^1.

4) On suppose que $\vec{y}^1$ est très légèrement perturbé d'une quantité $\delta\vec{y}$: $\vec{y}^2 = \vec{y}^1 + \delta\vec{y}$. Montrer que la variation correspondante de l'erreur commise se met sous la forme

$$\delta e_i = \alpha_i\,\delta y_{i-1} + \beta_i\,\delta y_i + \gamma_i\,\delta y_{i+1} \qquad \text{pour} \qquad 2 \leqslant i \leqslant M-1.$$

On donnera les expression des coefficients α_i, β_i et γ_i. Quelles sont en particulier les valeurs de α_1 et γ_M ? Quel choix peut-on faire pour β_1 et β_M ?

5) Mettre la relation précédente entre $\delta\vec{e} = \vec{e}^2 - \vec{e}^1$ et $\delta\vec{y}$ sous une forme matricielle. Par quelle méthode la matrice en question peut-elle s'inverser ? En déduire une procédure itérative de résolution de l'équation (1).

II Cas instationnaire : étude de quelques schémas simples

On s'intéresse désormais à la résolution de l'équation aux dérivées partielles

$$\frac{\partial u}{\partial t} = \lambda \frac{\partial^2 u}{\partial x^2}. \tag{2}$$

De nouveau, il n'est pas nécessaire de préciser les conditions aux limites ni la condition initiale.

1) Dans le cas de l'équation de la chaleur, quelles quantités représentent les grandeurs $u(x,t)$ et λ ? Citer un autre contexte dans lequel l'équation (2) intervient, en précisant de nouveau le sens de u et λ. Quel est le signe de λ ?

2) Le problème discrétisé associé le plus simple est, en notant $u_j^n = u(x_j, t_n)$:

$$\frac{1}{\Delta t}\left(u_j^{n+1} - u_j^n\right) \;=\; \lambda \frac{(\delta^2 u)_j^n}{(\Delta x)^2} \;=\; \frac{\lambda}{(\Delta x)^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right). \qquad (3)$$

Afin d'en étudier la stabilité, on décompose u_j^n sous la forme $u_j^n = \sum_k \widehat{\mathcal{U}}_k^n \exp(\imath k x_j)$.

a) Calculer le facteur d'amplification $g(k)$ défini par $\widehat{\mathcal{U}}_k^{n+1} = g(k)\,\widehat{\mathcal{U}}_k^n$.

b) En déduire la condition de stabilité de l'algorithme en question.

c) Cet algorithme est-il explicite ou implicite ?

3) On cherche à améliorer la méthode en remplaçant $(\delta^2 u)_j^n$ dans (3) par une moyenne pondérée de $(\delta^2 u)_j^n$ et de $(\delta^2 u)_j^{n+1}$:

$$\frac{1}{\Delta t}\left(u_j^{n+1} - u_j^n\right) \;=\; \frac{\lambda}{(\Delta x)^2}\left[(1-\theta)(\delta^2 u)_j^n + \theta(\delta^2 u)_j^{n+1}\right], \quad \text{avec} \quad 0 \leqslant \theta \leqslant 1.$$

a) Montrer que $g(k)$ s'écrit $g(k) = \dfrac{1 - y(1-\theta)}{1 + \theta y}$ avec $y = \dfrac{4\lambda\Delta t}{(\Delta x)^2}\sin^2\left(\dfrac{k\Delta x}{2}\right)$.

b) En déduire que pour $\theta_0 \leqslant \theta \leqslant 1$, la méthode est toujours stable. Quelle est la valeur de θ_0 ?

c) Pour $\theta \leqslant \theta_0$, quelle est la condition de stabilité ?

4) Une troisième méthode de résolution consiste à écrire

$$\frac{3}{2}\frac{u_j^{n+1} - u_j^n}{\Delta t} \;-\; \frac{\alpha}{\Delta t}\left(u_j^n - u_j^{n-1}\right) \;=\; \frac{\lambda}{(\Delta x)^2}\left(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}\right)$$

a) Justifier que la seule valeur acceptable pour α est $\alpha = 1/2$.

b) Montrer que $g(k)$ est solution de

$$g^2\left[3 + 8\frac{\lambda\Delta t}{(\Delta x)^2}\sin^2\left(\frac{k\Delta x}{2}\right)\right] \;-\; 4g \;+\; 1 \;=\; 0.$$

c) Tracer le module des solutions $g_\pm(k)$ en fonction de la grandeur y introduite à la question **3)** et conclure quant à la stabilité de la méthode.

5) Avec le schéma discrétisé considéré à la question **2)**, en quoi le choix de paramètres vérifiant $\lambda\Delta t = (\Delta x)^2/6$ est-il optimal ? On pourra considérer l'erreur en $\mathcal{O}[(\Delta t)^n] + \mathcal{O}[(\Delta x)^m]$ liée à la discrétisation de (2), et remarquer que si $u(x,t)$ est solution de (2), alors u vérifie également l'équation

$$\frac{\partial^2 u}{\partial t^2} \;=\; \lambda^2 \frac{\partial^4 u}{\partial x^4}.$$

3.8 Méthodes de Runge-Kutta d'ordre 2 et 4

Pour résoudre l'équation différentielle ordinaire $dy/dt = f(y)$, la méthode d'ordre 2 (RK2) proposée par RUNGE et KUTTA consiste à écrire –avec les notations habituelles–

$$y_{n+1} = y_n + \Delta t\, f\left(y_n + \frac{\Delta t}{2} f(y_n)\right). \qquad (1)$$

On rappelle également sa généralisation à l'ordre 4 (RK4)

$$\left.\begin{array}{ll} k_1 = \Delta t\, f(y_n) & ;\ k_2 = \Delta t\, f\left(y_n + \dfrac{k_1}{2}\right) \\ k_3 = \Delta t f\left(y_n + \dfrac{k_2}{2}\right) & ;\ k_4 = \Delta t f(y_n + k_3) \end{array}\right| \quad \text{puis} \quad y_{n+1} = y_n + a\,[k_1 + 2k_2 + 2k_3 + k_4]\,.$$

Le coefficient a est pour le moment supposé inconnu.

1) L'algorithme RK2 est une méthode à un pas. Qu'en est-il pour RK4 ? Quelle doit être la valeur de a ? Justifier votre réponse.

2) L'erreur associée au schéma RK2 est *a priori* en $\mathcal{O}[(\Delta t)^3]$. En effectuant un développement limité de la relation (1), démontrer explicitement ce résultat. Est-il envisageable que l'assertion précédente soit "pessimiste", et que l'erreur soit en réalité en $\mathcal{O}[(\Delta t)^4]$? On pourra exprimer d^2y/dt^2 et d^3y/dt^3 en fonction de f, df/dy et d^2f/dy^2.

3) De même, pour le schéma RK4, l'erreur commise est en $\mathcal{O}[(\Delta t)^5]$. On se contentera ici de montrer –résultat plus faible– que la méthode RK4 est correcte jusqu'à l'ordre $(\Delta t)^3$ inclus.

4) On s'intéresse désormais au cas particulier $f(y) = -\lambda y$ où λ est un réel positif. Effectuer l'analyse de stabilité des algorithmes RK2 et RK4. On pourra s'aider de la figure 4. Comparer les avantages respectifs de RK2 et RK4.

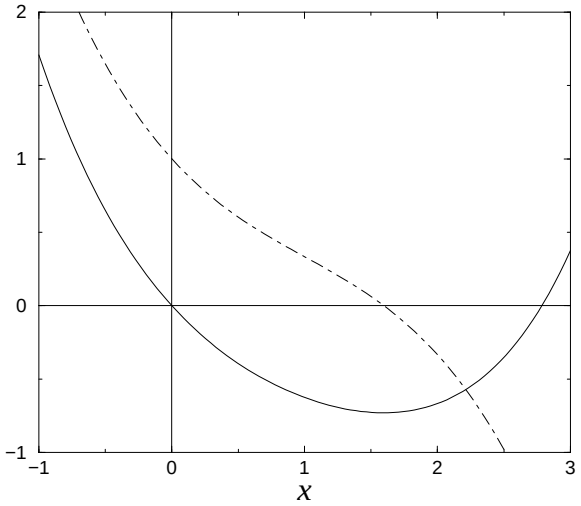

FIG. 4 – Graphe des fonctions $-x + \dfrac{x^2}{2} - \dfrac{x^3}{3!} + \dfrac{x^4}{4!}$ et $1 - x + \dfrac{x^2}{2} - \dfrac{x^3}{3!}$ (pointillés).

5) Si l'on mettait au point une méthode de type RK3, quel serait le domaine de stabilité de l'algorithme associé, toujours dans le cas $f(y) = -\lambda y$, $\lambda > 0$? Même question avec une hypothétique méthode "RK$_\infty$". On pourra de nouveau s'aider de la figure 4.

4 Énoncés de Travaux Pratiques

4.1 Initiation à Mathematica

À l'aide de la documentation fournie, le but de cette première séance avec Mathematica® est d'assimiler les commandes essentielles, ainsi que les possibilités qu'offre le logiciel :

- arithmétique de base,
- fonctions par défaut, symboles et fonctions utilisateur,
- gestion des graphiques,
- listes, gestion des entrées/sorties,
- écriture d'un programme élémentaire, boucles et exécutions conditionnelles,
- différentiation, intégration, calcul symbolique,
- résolution d'équations...

En particulier, on notera que les fonctions "machines" commencent systématiquement par une majuscule et que le caractère _ ("underscore") est indispensable pour définir une fonction "utilisateur" (par exemple, `f[x_]=x*Sin[x]`). Avant de traiter les exercices proposés, il est important de bien différencier les symboles `=` et `:=` (différence entre l'identité et l'égalité). Par ailleurs, on veillera à inhiber le verrouillage numérique de l'écran (pour quelle raison ?) et à valider la commande `Remove["Global‘*"]` d'autant plus fréquemment que les messages d'erreur s'accumulent !

Exercice I

Soit x une variable aléatoire réelle uniformément répartie entre 0 et 1. On définit $y = \sin(\pi x/2)$. Quelle est la densité de probabilité de y ? Utiliser l'aide en ligne pour réaliser/visualiser des histogrammes et vérifier numériquement ce résultat.

Exercice II : Théorème central limite

Énoncé : soient X_1, $X_2 \ldots X_n$, n variables aléatoires de moyenne m et d'écart-type σ. Lorsque n devient grand, la valeur moyenne $\sum_{i=1}^{n} X_i/n$ est une variable aléatoire dont la densité de probabilité tend vers une gaussienne de moyenne m et d'écart-type $\sigma/\sqrt{n}$.

Application : ce résultat fournit une procédure numérique pour engendrer une variable aléatoire de loi gaussienne : si les X_i sont uniformément réparties entre 0 et 1, la somme de 12 termes suffit en pratique (pourquoi a-t-on choisi 12 ?)

$$\left(\sum_{i=1}^{12} X_i\right) - 6 \simeq \mathcal{G}(0,1), \quad \text{gaussienne de moyenne nulle et de variance unité.}$$

Tester cette procédure à l'aide de Mathematica®, et comparer à la méthode de Box-Muller.

Généralisation : le théorème central limite s'étend au cas où tous les X_i ne sont pas de même loi de probabilité. Il suffit d'effectuer la substitution

$$\begin{cases} m \to \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} m_i \\ \sigma^2 \to \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \sigma_i^2, \end{cases}$$

où m_i est la valeur moyenne de X_i, et σ_i son écart-type.

4.2 Calcul vectoriel

TP sous Mathematica®

Soit $\overrightarrow{F}(\vec{r})$ un champ de force. Le chemin $\mathcal{C}$ est une courbe $\vec{r}(t)$ paramétrée par le temps t ($t_i \leqslant t \leqslant t_f$). Le travail correspondant au parcours de $\mathcal{C}$ s'écrit

$$W = \int_{\mathcal{C}} \overrightarrow{F}(\vec{r}) \cdot d\vec{r} = \int_{t_i}^{t_f} \overrightarrow{F}[\vec{r}(t)] \cdot \frac{d\vec{r}(t)}{dt}\, dt. \tag{1}$$

Cette intégrale se calcule simplement lorsque $\overrightarrow{F}(\vec{r})$ dérive d'un gradient (champ conservatif) :

$$\overrightarrow{F} = -\overrightarrow{\text{grad}}\,\Phi \qquad \Rightarrow \qquad W = \Phi[\vec{r}(t_i)] - \Phi[\vec{r}(t_f)]. \tag{2}$$

Le calcul analytique est en général fastidieux. Mathematica® offre la possibilité de le faire.

Considérons

$$\overrightarrow{F}(\vec{r}) = \left| \begin{array}{l} 2xy + z^3 \\ x^2 \\ 3xz^2 \end{array} \right. \qquad \text{au point } \vec{r} \text{ de coordonnées } (x, y, z). \tag{3}$$

On définit trois chemins entre les points de coordonnées $(1, 0, 0)$ et $(1, 0, 1)$:

$$\left| \begin{array}{lcl} \mathcal{C}_1 & : & \{\cos(2\pi t), \sin(2\pi t), t\} \\ \mathcal{C}_2 & : & \{1, 0, t\} \\ \mathcal{C}_3 & : & \{1 - \sin(\pi t)/2, 0, [1 - \cos(\pi t)]/2\} \end{array} \right. \qquad 0 \leqslant t \leqslant 1. \tag{4}$$

1) Visualiser les trois courbes $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$.

2) Definir une fonction `travail[c]` qui calcule le travail associé au déplacement sur le chemin `c` dans le champ de force $\overrightarrow{F}$. Quel est le travail correspondant aux trois trajets précédents ? Quelle conclusion est-on tenté d'en tirer ? Vérifier en prenant le rotationnel.

3) Obtenir simplement le potentiel duquel $\overrightarrow{F}$ dérive (en utilisant par exemple la fonction `travail` sur un chemin $\mathcal{C}_4$ judicieusement choisi).

4) Calculer la longueur des courbes $\mathcal{C}_1, \mathcal{C}_2$ et $\mathcal{C}_3$.

5) Reprendre la même analyse pour le champ de force

$$\overrightarrow{G}(\vec{r}) = \left| \begin{array}{l} 2xy^2 + z^3 \\ x^3 \\ 3xz^3 \end{array} \right. . \tag{5}$$

Indication : l'opérateur différentiel `D` *est "listable" (il peut agir sur les éléments d'une liste).*

4.3 Transition gaz/liquide et construction de Maxwell

TP sous Mathematica®

Fréquemment la dépendance mutuelle de quantités physiques est gouvernée par des équations non linéaires que l'on ne peut résoudre que numériquement. Considérons l'équation d'état de gaz modèles, reliant la pression P au volume V et à la température T. Dans un gaz parfait, on ne tient compte ni des interactions entre molécules (ces interactions sont d'autant plus importantes que la densité est élevée), ni du volume des molécules. De nombreuses équations d'état incluant ces deux corrections peuvent être proposées pour rendre compte du comportement des fluides réels, mais aucune n'est en accord quantitatif avec l'expérience à toute température et densité. Toutefois, l'équation de VAN DER WAALS a le mérite de la simplicité et décrit qualitativement l'écart par rapport au comportement des gaz parfaits : pour N molécules de gaz, elle s'écrit

$$\left(P + \frac{aN^2}{V^2}\right)(V - Nb) = Nk_BT, \tag{1}$$

où a et b sont des constantes. Le terme aN^2/V^2 est une pression moléculaire qui traduit l'existence d'une interaction attractive entre les molécules tandis que le terme de répulsion Nb, appelé covolume, provient du volume non nul occupé par celles-ci.

Aux basses températures, les isothermes P-V présentent une "boucle" dont certaines parties correspondent à des états thermodynamiquement instables, associés à une coexistence gaz-liquide. Pour connaître la région de coexistence, il faut résoudre, pour chaque température inférieure à la température critique T_C, des équations non linéaires, qui font en outre intervenir des intégrales (construction de MAXWELL). On obtient ainsi une description de la transition de phase du gaz au liquide.

1) Après avoir visualisé les isothermes, calculer les paramètres critiques (P_C, V_C, T_C) pour lesquels

$$\left.\frac{\partial P}{\partial V}\right|_T = 0 \qquad \text{et} \qquad \left.\frac{\partial^2 P}{\partial V^2}\right|_T = 0.$$

Réponse : $\quad T_C = \dfrac{8a}{27k_Bb}, \qquad V_C = 3Nb, \qquad P_C = \dfrac{a}{27b^2}.$

2) En déduire que l'équation d'état entre les paramètres réduits s'écrit

$$\left(\widetilde{P} + \frac{3}{\widetilde{V}^2}\right)(3\widetilde{V} - 1) \;=\; 8\widetilde{T}, \tag{2}$$

où $\widetilde{X} = X/X_C$. Noter que la relation précédente est "universelle" dans la mesure où elle ne fait plus intervenir les paramètres microscopiques a et b.
Indication : on pourra utiliser la commande `Simplify` *et appliquer des règles de transformation inspirées de la syntaxe*

$$\texttt{Expre}/.\{x \to 1 + a, y \to 1 - b\}$$

qui revient à remplacer x par $1 + a$ et y par $1 - b$ dans `Expre`.

3) Pourquoi les états pour lesquels $\left.\dfrac{\partial P}{\partial V}\right|_T > 0$ sont-ils thermodynamiquement instables ?

4) Vérifier que pour $\widetilde{T} < 1$, les isothermes $\widetilde{P}(\widetilde{V})$ présentent une boucle non-physique de VAN DER WAALS. Celle-ci doit être remplacée par un palier de coexistence à la pression $\widetilde{P}_t$

entre les volumes $\widetilde{V}_1$ et $\widetilde{V}_2$ (voir la figure 5) : pour les volumes compris entre $\widetilde{V}_1$ et $\widetilde{V}_2$, les phases condensée (liquide) et diluée (gaz) existent simultanément, à la pression constante $\widetilde{P}_t$. Pour quelle raison ?

5) En considérant l'énergie libre de différentielle $dF = -SdT - PdV$, justifier la relation

$$\int_{\widetilde{V}_1}^{\widetilde{V}_2} \widetilde{P}(\widetilde{V})\, d\widetilde{V} = \widetilde{P}_t(\widetilde{V}_2 - \widetilde{V}_1).$$

Quelle est son interprétation géométrique ?

6) On cherche à écrire une procédure `iso[T_]` qui trace l'isotherme T (avec son palier de coexistence lorsque $\widetilde{T} < 1$). Dans ce dernier cas, avant de résoudre pour chaque température le système

$$\begin{cases} \mathtt{p[v_1]} == \mathtt{p[v_2]} \\ \mathtt{p[v_1]}(\mathtt{v_2} - \mathtt{v_1}) == \mathtt{Integrate[p[v], \{v, v_1, v_2\}]}, \end{cases} \tag{3}$$

on devra estimer assez précisément autour de quelles valeurs $\mathtt{v}_1^*$ et $\mathtt{v}_2^*$ effectuer la recherche pour le couple $(\mathtt{v_1}, \mathtt{v_2})$. On pourra d'abord rechercher (à l'aide de `FindMinimum`) les valeurs de `v` pour lesquelles `p[v]` atteint respectivement son minimum et son maximum local, puis définir `vtest` comme la moyenne arithmétique des deux grandeurs précédentes, avant de résoudre enfin `p[v]==p[vtest]`. Deux des trois racines de cette équation sont des conditions initiales acceptables pour la résolution de (3).

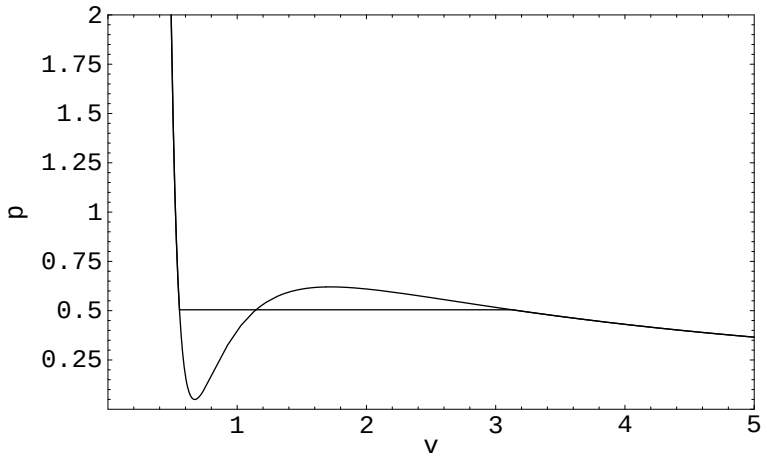

FIG. 5 – Construction de MAXWELL pour l'isotherme $\widetilde{T} = 0.85$. Ici, $\widetilde{V}_1 = 0.553$ et $\widetilde{V}_2 = 3.127$. Pour $\widetilde{V}_1 \leqslant \widetilde{V} \leqslant \widetilde{V}_2$, les phases liquide et gaz coexistent à l'équilibre thermodynamique (la pression réduite du palier est alors $\widetilde{P}_t = 0.507$).

4.4 L'histoire de deux ballons de baudruche

TP sous Mathematica®

Introduction

Deux ballons en caoutchouc identiques, sont partiellement gonflés avec de l'air. Les ballons seront supposés sphériques dans tout ce qui suit. Ils sont reliés par un tuyau comportant une valve. Initialement, leurs tailles diffèrent. Lorsque la valve est ouverte, que se passe t-il ? Le plus petit va t-il se vider dans le gros, ou le contraire va t-il se produire, avant l'obtention d'un état d'équilibre ? Par ailleurs, à l'équilibre, les deux ballons ont-ils la même taille ?

Modélisation

La surpression (par raport au milieu ambiant) à l'intérieur d'un ballon de rayon R s'écrit $\Delta P = 2\sigma/R$ (formule de LAPLACE). Le point clé est ici que la tension de surface σ est une fonction non triviale (et croissante) du rayon. Des arguments thermodynamiques montrent que pour un caoutchouc "idéal", on a

$$\sigma(R) = \kappa \left[1 - \left(\frac{R_0}{R}\right)^6\right].$$

R_0 est le rayon au repos du ballon (i.e. en l'absence de contrainte) et κ une constante sans importance pour la suite. Le cas mentionné ici correspond à la situation où l'énergie interne du système ne dépend pas de la surface occupée, comme c'est le cas pour le gaz parfait. On parlera de "ballon idéal". Attention, cela ne signifie pas que le gaz contenu soit parfait. Nous utiliserons désormais des quantités sans dimension, où le rayon R est exprimé en unité de R_0, et ou la surpression, que nous noterons P dans la suite, devient

$$P = \frac{1}{R}\left[1 - \left(\frac{1}{R}\right)^6\right].$$

Nous nous limiterons au cas où les ballons sont sous contrainte (i.e. plus ou moins gonflés) : $R > 1$.

Etude des ballons idéaux

1) Pour commencer, le tracé de P en fonction de R est instructif... Il permet de comprendre les phénomènes essentiels (et peu intuitifs) que nous allons mettre en évidence.

2) Dans le plan (R_1, R_2), tracer les points pour lesquels un équilibre des pressions est possible. On peut bien entendu se limiter au cas $R_1 < R_2$.

3) La courbe précédente est une séparatrice qui distingue deux types de comportement différents. Lesquels ?

4) En supposant qu'une situation initiale hors équilibre est suivie d'une évolution vers l'équilibre isotherme et que le gaz est parfait, identifier l'ensemble des conditions initiales pour lesquels l'état d'équilibre est constitué par deux ballons de tailles différentes. En d'autres termes, on supposera ici que l'évolution est telle que $P(R_1)R_1^3 + P(R_2)R_2^3$ est constante.

Cas des ballons réels

La formule proposée plus haut pour la tension de surface $\sigma(R)$ est un cas limite difficile à approcher en pratique. De manière plus générale, on a

$$\sigma(R) = \kappa \left[1 - \left(\frac{R_0}{R} \right)^6 \right] f\left(\frac{R}{R^*} \right).$$

où R^* est une nouvelle échelle de longueur. Il s'agit du rayon au delà duquel la membrane ne peut plus être considérée comme idéale. Par ailleurs, la fonction f vérifie

$$f(x) = \begin{cases} 1 & \text{pour} \quad 0 \leqslant x \leqslant 1 \\ 1 + B(x-1)^\alpha & \text{pour} \quad x \geqslant 1 \end{cases}$$

La quantité B (sans réelle pertinence) est positive et le paramètre α est supérieur à 1. On peut considérer que le cas des ballons idéaux est retrouvé en prenant la limite $R^* \to \infty$. Reprendre l'étude précédente. Il est nécessaire de différencier les cas $\alpha \geqslant 2$ et $1 \leqslant \alpha \leqslant 2$. En particulier, montrer que lorsque $\alpha \geqslant 2$, il existe une valeur critique de R^* en dessous de laquelle le gros ballon se dégonfle toujours dans le petit. Que se passe t-il lorsque R^* est supérieur à la valeur critique ? Le cas le plus riche et le plus surprenant est celui où $1 \leqslant \alpha \leqslant 2$. On pourra tenter de discerner les différents comportements génériques dans le plan (R_1, R_2).

Référence : Y. Levin et F.L. da Silveira, "Too rubber ballons : phase diagram of air transfer", *Physical Review E* **69**, 051108 (2004).