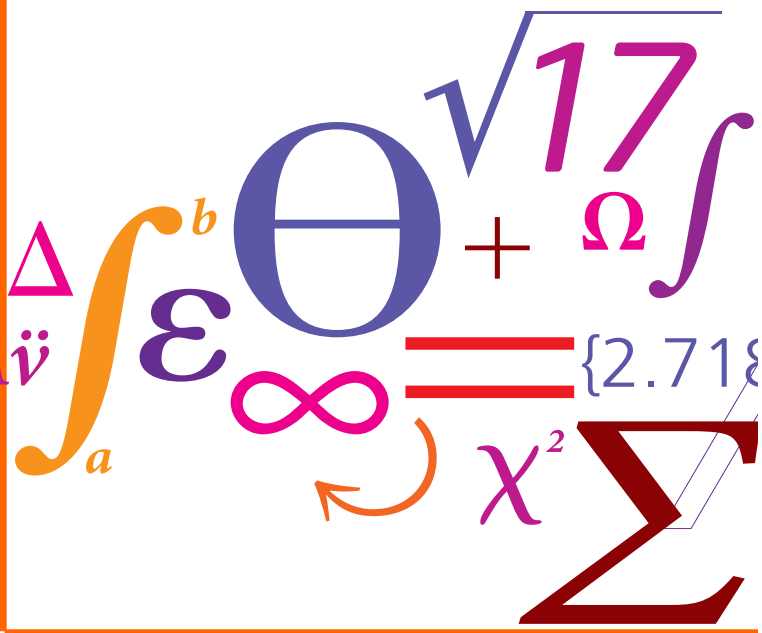
République Algérienne Démocratique et Populaire – الجمهورية الجزائرية الديمقراطية الشعبية
Ministère de l'enseignement supérieur et de la recherche scientifique
Université Mohamed Khider – Biskra

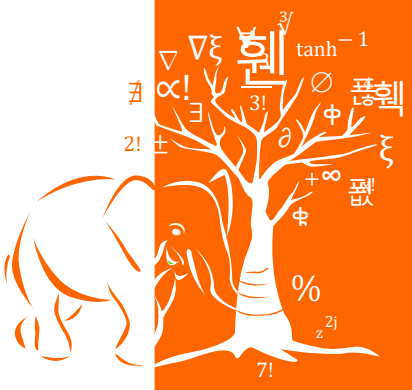
Méthodes mathématiques & algorithmes pour la physique

Niveau Master Physique
∀ l'option

$$(EIv'')''=q-\rho A\ddot{v}$$



Dr. Samir KENOUCHE



Chapitre[1] Concavité et convexité des fonctions de plusieurs variables

Samir Kenouche - Département des Sciences de la Matière - UMKB

Module : Méthodes Mathématiques et Algorithmes pour la Physique
Version corrigée, améliorée et actualisée le 10/10/2020

Résumé

Ce chapitre débute par un bref rappel sur des notions élémentaires portant notamment sur la caractérisation de la convexité, de la concavité ainsi que les points stationnaires. Ces notions fondamentales sont indispensables et nécessaires afin d'appréhender le fonctionnement et le fondement des algorithmes d'optimisation traités dans le deuxième chapitre. Les problèmes rencontrés en physique, sont très souvent complexes nécessitant des modèles mathématiques multidimensionnels, il est donc fréquent de recourir à des fonctions de plusieurs variables. Ainsi, l'étude de la concavité et de convexité de ces fonctions repose pleinement sur la détermination de la matrice *Hessienne*. Par ailleurs, ce chapitre se veut avant tout introductif afin de tester et de consolider les connaissances mathématiques des étudiants (es).

Table des matières

I. Introduction

L'optimisation numérique de problèmes complexes, issus de la physique, a connu un essor considérable ces dernières années grâce notamment au développement de l'informatique et des moyens de calcul très puissants. Une opération d'optimisation se présente systématiquement comme l'ultime étape après avoir mis en équation ou modéliser le problème physique en question. Nous comprenons ainsi aisément que la compréhension du formalisme inhérent à la détermination ainsi qu'à l'analyse des points critiques (encore appelés points stationnaires) est une étape cruciale et indispensable afin de mener à bien cette opération d'optimisation. Il convient de préciser que ce premier chapitre se veut introductif pour le deuxième, dans lequel une étude avancée et détaillée des algorithmes d'optimisation sera conduite.

II. Notations différentielles

Nous entamons ces notes de cours en rappelant succinctement la notion de la dérivée. Cette quantité mathématique joue un rôle central dans la définition et la détermination des points stationnaires. Soit la fonction continue d'une seule variable $f : [a, b] \subset \mathbb{R} \longrightarrow \mathbb{R}$. On définit sa fonction dérivée par l'expression :

$$\frac{df(x)}{dx} = \left[\frac{\Delta f(x)}{\Delta x}\right]_{\Delta x \approx 0} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} = f^{'}(x) \tag{1}$$

S. Kenouche est docteur en Physique de l'Université des Sciences et Techniques de Montpellier et docteur en Chimie de l'Université A. Mira de Béjaia.
page web personnelle : http://www.sites.univ-biskra.dz/kenouche

La variation totale absolue de f engendrée par un accroissement de $x+\Delta x$ s'écrit $\Delta f = f(x+\Delta x) - f(x)$. Par conséquent, la variation totale relative est la quantité :

$$\frac{\Delta f(x)}{\Delta x} \tag{2}$$

Rappelons que pour une variation infinitésimale ($\Delta x \longrightarrow 0$), nous pouvons écrire :

$$\Delta x \longrightarrow dx$$

$$\Delta f \longrightarrow df$$

Ayant un point x_0 appartenant à l'intervalle de définition de f, le nombre $f^{'}(x_0)$ exprime la pente, au point $(x_0, f(x_0))$, de la droite tangente à la courbe $y = f(x)$. En effet, le rapport :

$$f^{'}(x_0) = \lim_{\Delta x \to 0} \frac{f(x_0+\Delta x) - f(x_0)}{\Delta x} \tag{3}$$

vaut la pente de la droite passant par les points $(x_0, f(x_0))$ et $(x_0+\Delta x, f(x_0+\Delta x))$. Cette pente exprime le taux de changement de $f(x)$ causé par la variation Δx. Il en ressort que l'équation de la droite tangente à la courbe $y = f_T(x)$ au point (x_0, y_0) s'écrit simplement :

$$y - y_0 = f^{'}(x_0)(x - x_0) \Longrightarrow f_T(x) = y_0 + f^{'}(x_0)(x - x_0) \tag{4}$$

Cette approximation est d'autant plus vraie que $\Delta x \longrightarrow 0$. Lorsque la limite de l'Eq. (3) existe, on dira que f est dérivable en x_0. Dans le cas où la fonction f est dérivable pour tout $x \in [a, b] \subset \mathbb{R}$, on dira qu'elle est dérivable sur $[a, b] \subset \mathbb{R}$ et on aura par conséquent la fonction $f^{'}(x)$. Cette dernière peut également être dérivable et on aura $f^{''}(x)$ ... etc. Notons que la notion de la *dérivée directionnelle* sera définie dans le prochain chapitre.

Pour une fonction de deux variables $f(x, y) \in \mathbb{R}$, lorsque les deux variables sont modifiées de $x + \Delta x$ et $y + \Delta y$, la variation totale de f s'obtient selon :

$$\Delta f = f(x + dx, y + dy) - f(x, y) \tag{5}$$

En développant le premier terme du second membre en *séries de Taylor*, il vient :

$$f(x + dx, y + dy) = f(x, y) + \frac{\partial f(x, y)}{\partial x}\, dx + \frac{\partial f(x, y)}{\partial y}\, dy + O(dx^2, dy^2) \tag{6}$$

En substituant (6) dans (5), nous obtenons :

$$\Delta f = \frac{\partial f(x, y)}{\partial x}\, dx + \frac{\partial f(x, y)}{\partial y}\, dy \tag{7}$$

La variation totale de la fonction $f(x, y)$ s'écrit ainsi en fonction des variations partielles de f par rapport aux deux variables.

III. Convexité et concavité

Nous rappelons que la dérivée traduit les variations, croissance ($f^{'}(x) > 0$) et décroissance ($f^{'}(x) < 0$), d'une fonction. Une dérivée nulle, en un point x_0, est caractérisée par une tangente horizontale et donc la fonction admet un point stationnaire : un minimum, un maximum ou un point selle. Cette condition est nécessaire pour l'existence d'un extremum, mais elle n'est pas suffisante. A titre d'exemple, la fonction $f(x) = x^3$ n'admet pas d'extremum local en $x_0 = 0$ bien que sa dérivée $f^{'}(x) = 3\,x^2$ s'y annule, ce point est *singulier*.

Définition 1 *: La courbe de $f(x)$ est dite* **convexe** *si tous les points de la courbe $y = f(x)$ se trouvant au dessous de la tangente*[1] *en un point quelconque de l'intervalle de définition de la fonction, autrement dit : $\forall\, x \in [a,b]\, f(x) < f_T(x)$. Cette condition est satisfaite si $f''(x) < 0$.*

Définition 1' *: La courbe de $f(x)$ est dite* **concave** *si tous les points de la courbe $y = f(x)$ se trouvant au dessus de la tangente en un point quelconque de l'intervalle de définition de la fonction, autrement dit : $\forall\, x \in [a,b]\, f(x) > f_T(x)$. Cette condition est satisfaite si $f''(x) > 0$.*

Ces résultats se démontrent aisément en développant la fonction f en série de *Taylor* au voisinage du point critique x_0, soit :

$$f(x_0 + \delta x) = f(x_0) + f'(x_0)\,\delta x + f''(x_0)\,\delta x^2 + \epsilon(\delta x^3) \tag{8}$$

Avec x_0 est un point critique par conséquent $f'(x_0) = 0$, il vient :

$$\Rightarrow \Delta f = f''(x_0)\,\delta x^2 + \epsilon(\delta x^3) \tag{9}$$

Ainsi la nature du point critique est déterminée par le signe de $f''(x_0)$. Si $f''(x_0) > 0 \Rightarrow \Delta f > 0$ alors $f(x_0 + \delta x) > f(x_0)$. Nous en déduisons que la fonction f prend des valeurs supérieures au voisinage $(x_0 + \delta x)$ alors il s'agit nécessairement d'un minimum. Nous concluons pour les fonctions d'une seule variable, les conditions $f'(x_0) = 0$ et $f''(x_0) > 0$ sont suffisantes pour que la fonction $f(x)$ admette un minimum local (concavité). Les conditions $f'(x_0) = 0$ et $f''(x_0) < 0$ prouvent l'existence d'un maximum local (convexité).

Propriétés *: Si x_0 est un optimum local alors c'est un optimum global. Ainsi, le plus petit des minima locaux est un minimum global. De la même façon, le plus grand des maxima locaux est un maximum global.*

A. *Généralisation aux fonctions de plusieurs variables*

Réécrivons les résultats obtenus pour le cas monodimensionnel au cas multidimensionnel. Soit f une fonction de classe $\mathcal{C}^2$ ($[\mathcal{D} \subset \mathbb{R}^n]$). La détermination de la nature des extremums sera conduite en analysant la série de *Taylor* généralisée au cas des fonctions de plusieurs variables. Afin de simplifier le formalisme, prenons $n = 2$ et travaillons avec $X \in \mathbb{R}^2$ et $f(X) \in \mathbb{R}$.

a) ***Théorème****: Soit $X^* = (x_0, y_0)^T \in \mathbb{R}^2$ un point critique de $f(X) \Rightarrow \nabla f(X^*) = 0$. Nous posons $D = (\delta a, \delta b)^T$, c'est l'accroissement infinitésimal de f alors $\exists\, \eta > 0$ tel que pour tous $(\delta a, \delta b)^T \in \mathcal{D} \subset \mathbb{R}^2$ vérifiant $\parallel (\delta a, \delta b)^T \parallel_2 < \eta$, nous avons :*

$$f(X + D) = \sum_{n=0}^{\infty} \frac{(D \cdot \nabla)^n\, f(X)}{n!} \tag{10}$$

Autrement dit, trouver la nature d'un point critique (maximum, minimum, selle, singulier) revient à étudier *localement le comportement* de la fonction f dans son voisinage $\parallel (\delta a, \delta b)^T \parallel_2 < \eta$, c'est-à-dire : $x_0 + \delta a \in [x_0 - \eta\,,\, x_0 + \eta]$ et $y_0 + \delta b \in [y_0 - \eta\,,\, y_0 + \eta]$. Tenant compte de (10), la série de *Taylor* au voisinage du point critique $(x_0 + \delta a, y_0 + \delta b)^T$ se développe comme suit :

$$f(x_0 + \delta a, y_0 + \delta b) = f(x_0, y_0) + \frac{(D \cdot \nabla)\, f(X)}{1!} f(x_0, y_0) + \frac{(D \cdot \nabla)^2\, f(X)}{2!} f(x_0, y_0) + \cdots\cdots$$
$$+ \frac{(D \cdot \nabla)^p\, f(X)}{p!} f(x_0, y_0) + \cdots\cdots + \frac{(D \cdot \nabla)^{(n+1)}\, f(X)}{(n+1)!} f(x_0, y_0)$$

1. Nous rappelons que l'équation de la tangente au point x_0 vaut : $f_T(x) = f(x_0) + f'(x_0)\underbrace{(x - x_0)}_{\delta x}$.

Tronquons la série de *Taylor* à l'ordre deux, il vient :

$$f(x_0+\delta a, y_0+\delta b) = f(x_0,y_0) + \frac{(D\cdot\nabla)\,f(X)}{1!}\,f(x_0,y_0) + \frac{(D\cdot\nabla)^2\,f(X)}{2!}\,f(x_0,y_0) + \epsilon(\delta a^3, \delta b^3) \tag{11}$$

Avec $\nabla = \left(\dfrac{\partial}{\partial x}, \dfrac{\partial}{\partial y}\right)$ est l'opérateur différentiel. Calculons les termes linéaire et quadratique de la série (11) :

$$(D\cdot\nabla)\,f = \left((\delta a,\delta b)\cdot\left(\frac{\partial}{\partial x},\frac{\partial}{\partial y}\right)\right) f \tag{12}$$

$$= \left(\delta a\,\frac{\partial f}{\partial x} + \delta b\,\frac{\partial f}{\partial y}\right) \tag{13}$$

$$\tag{14}$$

De façon analogue pour l'ordre quadratique,

$$(D\cdot\nabla)^2\,f = \left((\delta a,\delta b)\cdot\left(\frac{\partial}{\partial x},\frac{\partial}{\partial y}\right)\right)^2 f \tag{15}$$

$$= \left(\delta a\,\frac{\partial f}{\partial x} + \delta b\,\frac{\partial f}{\partial y}\right)^2 \tag{16}$$

$$= \delta a^2\,\frac{\partial^2 f}{\partial x^2} + +2\,\delta a\,\delta b\,\frac{\partial^2 f}{\partial x\,\partial y} + \delta b^2\,\frac{\partial^2 f}{\partial y^2} \tag{17}$$

$$\tag{18}$$

Substituons (14) et (18) dans la série (11), nous obtenons :

$$f(x_0+\delta a, y_0+\delta b) = f(x_0,y_0) + \left(\delta a\,\frac{\partial f}{\partial x} + \delta b\,\frac{\partial f}{\partial y}\right) + \frac{1}{2}\left[\delta a^2\,\frac{\partial^2 f}{\partial x^2} + 2\,\delta a\,\delta b\,\frac{\partial^2 f}{\partial x\,\partial y} + \delta b^2\,\frac{\partial^2 f}{\partial y^2}\right] + \epsilon(\delta a^3,\delta b^3) \tag{19}$$

$$\Rightarrow \Delta f = \left(\delta a\,\frac{\partial f}{\partial x} + \delta b\,\frac{\partial f}{\partial y}\right) + \frac{1}{2}\left[\delta a^2\,\frac{\partial^2 f}{\partial x^2} + 2\,\delta a\,\delta b\,\frac{\partial^2 f}{\partial x\,\partial y} + \delta b^2\,\frac{\partial^2 f}{\partial y^2}\right] + \epsilon(\delta a^3,\delta b^3) \tag{20}$$

Le vecteur $X^* = (x_0,y_0)^T$ étant un point critique alors :

$$\frac{\partial f(x_0,y_0)}{\partial x} = \frac{\partial f(x_0,y_0)}{\partial y} = 0 \tag{21}$$

Substituons (21) dans (20), il en découle :

$$\Delta f = \frac{1}{2}\left[\delta a^2\,\frac{\partial^2 f}{\partial x^2} + 2\,\delta a\,\delta b\,\frac{\partial^2 f}{\partial x\,\partial y} + \delta b^2\,\frac{\partial^2 f}{\partial y^2}\right] + \epsilon(\delta a^3,\delta b^3) \tag{22}$$

Sous forme matricielle cette dernière relation s'écrit :

$$\Delta f = \frac{1}{2}\,(\delta a,\delta b)^T \times \begin{pmatrix} \dfrac{\partial^2 f}{\partial x^2} & \dfrac{\partial^2 f}{\partial x\,\partial y} \\ \dfrac{\partial^2 f}{\partial y\,\partial x} & \dfrac{\partial^2 f}{\partial y^2} \end{pmatrix} \times \begin{pmatrix} \delta a \\ \delta b \end{pmatrix} + \epsilon(\delta a^3,\delta b^3) \tag{23}$$

Cette formulation met en exergue la matrice carrée des dérivées partielles d'ordre 2, dite matrice *Hessienne*. Le *déterminant* et la *trace* de cette matrice jouent un rôle fondamental dans la détermination de la nature des

points critiques. La généralisation de cette matrice à $\mathbb{R}^{n\times n}$ est immédiate :

$$H_{ij}(f)=\frac{\partial^2 f}{\partial x_i\, x_j}=\begin{pmatrix}\dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\,\partial x_n}\\ \dfrac{\partial^2 f}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\,\partial x_n}\\ \vdots & \vdots & \ddots & \vdots\\ \dfrac{\partial^2 f}{\partial x_n\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_n\, x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2}\end{pmatrix} \tag{24}$$

Bien entendu cela suppose l'existence des dérivées partielles secondes de la fonction. Nous remarquons que la matrice *Hessienne* de f est symétrique du fait des dérivées secondes mixtes $\partial_{xy}f$ et $\partial_{yx}f$. On rappellera que pour une fonction de deux variables $f(x_1,x_2)$, nous avons la possibilité de constituer quatre dérivées partielles.

$$\frac{\partial^2 f}{\partial x_1^2}=\frac{\partial}{\partial x_1}\left(\frac{\partial f}{\partial x_1}\right)\quad;\quad\frac{\partial^2 f}{\partial x_1\,\partial x_2}=\frac{\partial}{\partial x_1}\left(\frac{\partial f}{\partial x_2}\right)$$

$$\frac{\partial^2 f}{\partial x_2^2}=\frac{\partial}{\partial x_2}\left(\frac{\partial f}{\partial x_2}\right)\quad;\quad\frac{\partial^2 f}{\partial x_2\,\partial x_1}=\frac{\partial}{\partial x_2}\left(\frac{\partial f}{\partial x_1}\right)$$

Cependant, seulement trois de ces dérivées sont distinctes. Ainsi, on démontre que

$$\frac{\partial^2 f}{\partial x_1\,\partial x_2}=\frac{\partial^2 f}{\partial x_2\,\partial x_1}$$

On saisit à ce propos que pour une dérivée partielle du second ordre mixte, l'ordre dans lequel on calcule les dérivées n'a pas d'impact sur le résultat. Notons aussi que le graphe d'une fonction de plusieurs variables, par exemple en deux dimensions, est défini formellement :

$$\mathcal{G}_f\equiv\{(x,y,z)=f(x,y))\mid(x,y)\in\mathcal{D}\}$$

Ainsi, à tout point $(x,y)\in\mathcal{D}$ ayant l'image $f(x,y)\in\mathbb{R}$ correspond un point du graphe $(x,y,f(x,y))\in\mathbb{R}^3$. Ce champ de points formera le relief de la fonction en question.

b) ***Théorème****: Soit f une fonction de classe $\mathcal{C}^2$ ($[\mathcal{D}\subset\mathbb{R}^n]$). Soit $X^*=(x_0,y_0,z_0,\cdots)\in\mathbb{R}^n$ un point critique de $f(X)\Rightarrow\nabla f(X^*)=0$. La matrice Hessienne $H_{ij}(f)$ au point $X^*=(x_0,y_0,z_0,\cdots)$ est symétrique dont les valeurs propres $\{\lambda_i\}_{i=1,n}\in\mathbb{R}$ et $\lambda_1\leq\lambda_2\leq\cdots\leq\lambda_p\leq\cdots\leq\lambda_n$.*

- *si $\{\lambda_i\}_{i=1,n}>0\Rightarrow f$ admet un minimum local au point critique $X^*=(x_0,y_0,z_0,\cdots)$. Dans ce cas de figure, la matrice $H_{ij}(f)$ est dite définie positive.*
- *si $\{\lambda_i\}_{i=1,n}<0\Rightarrow f$ admet un maximum local au point critique $X^*=(x_0,y_0,z_0,\cdots)$. Dans ce cas de figure, la matrice $H_{ij}(f)$ est dite définie négative.*
- *si $\{\lambda_i\}_{i=1,n}=0\Rightarrow$ il faudra analyser la série de Taylor a des ordres ≥ 3 pour pouvoir conclure sur le comportement de f au voisinage du point critique.*
- *si $\forall j\neq i$ tel que $\lambda_j=0$, pour conclure sur la nature du point critique, il faudra déterminer le signe de la trace soit : $T_{H(f)}$.*

Comme il a été mentionné précédemment, la matrice *Hessienne* $H_{ij}(f)$, est symétrique donc elle est diagonalisable. Le signe est conservé par changement de base. Cela signifie concrètement que les signes de $\{\lambda_i\}_{i=1,n}$ de la matrice diagonale sont identiques à ceux des éléments diagonaux de la *Hessienne*. Par conséquent, il

n'est pas nécessaire de calculer les valeurs propres pour conclure sur la nature d'un point critique. Du fait de la conservation du signe, après changement de base, nous nous baserons uniquement sur la matrice $H_{ij}(f)$. Nous en déduisons les cas de figure ci-dessous :

A- si $\mid H_{ij}(f) \mid > 0 \Rightarrow$ les $\{\lambda_i\}_{i=1,n}$ ont le même signe.

A-1 si la trace $T_{H(f)} > 0$ alors les $\{\lambda_i\}_{i=1,n} > 0$. De la formule de *Taylor* (23), il en découle $\Delta f > 0 \Rightarrow f(x_0 + \delta a, y_0 + \delta b) > f(x_0, y_0)$ alors la fonction f prend des valeurs supérieures au voisinage $(x_0 + \delta a, y_0 + \delta b)$. La fonction croît, nous concluons que le point critique (x_0, y_0) est nécessairement un minimum local.

A-2 si la trace $T_{H(f)} < 0$ alors les $\{\lambda_i\}_{i=1,n} < 0$. De la formule de *Taylor* (23), il en découle $\Delta f < 0 \Rightarrow f(x_0 + \delta a, y_0 + \delta b) < f(x_0, y_0)$ alors la fonction f prend des valeurs inférieures au voisinage $(x_0 + \delta a, y_0 + \delta b)$. La fonction diminue, nous concluons que le point critique (x_0, y_0) est nécessairement un maximum local.

B- si $\mid H_{ij}(f) \mid < 0 \Rightarrow$ les $\{\lambda_i\}_{i=1,n}$ sont de signes contraires alors la fonction f admet un point *selle*.

C- si $\mid H_{ij}(f) \mid = 0 \Rightarrow \forall j \neq i,\ \lambda_j = 0$. Pour conclure sur la nature du point critique, il faut déterminer le signe de la trace de la *Hessienne*.

C-1 si la trace $T_{H(f)} > 0$ alors f admet un minimum local.

C-2 si la trace $T_{H(f)} < 0$ alors f admet un maximum local.

Pour le cas d'une fonction de trois variables $X \in \mathbb{R}^3$, d'après la théorie développée précédemment, un déterminant $\mid H_{ij}(f) \mid > 0$ signifie un produit positif des valeurs propres $\Pi_{i=1}^{3} \lambda_i > 0$. Par conséquent nous pouvons dégager deux séquences de signes (+, +, +) ou (-, -, +) sachant que d'après le théorème ci-dessus nous obtenons toujours $\lambda_1 \leq \lambda_2 \leq \lambda_3$. Afin de conclure sur la nature des points critiques nous devons calculer la trace de la *Hessienne*.

D-1 si la trace $T_{H(f)} < 0$ alors le point critique est un *point-selle*.

D-2 si la trace $T_{H(f)} > 0$ c'est un cas ambigu. Il faudra calculer les valeurs propres.

D-1-1 si $\Pi_{i=1}^{3} \lambda_i > 0$ alors f admet un minimum local.

D-1-2 si $\Pi_{i=1}^{3} \lambda_i > 0$ alors si toutes les valeurs propres sont positives, f admet un minimum local. Si au moins une des valeurs propres est négative alors il s'agit d'un *point-selle*.

Une autre manière de vérifier si $H_{ij}(f)$ est définie positive (existence d'un minimum local), ou définie négative (existence d'un maximum local) est de considérer :

$$H_{ij}(f) \text{ est définie positive} \Longrightarrow \forall X \neq 0 \in \mathbb{R}^n : X^T H_{ij}(f) X > 0 \tag{25}$$

$$H_{ij}(f) \text{ est définie négative} \Longrightarrow \forall X \neq 0 \in \mathbb{R}^n : X^T H_{ij}(f) X < 0 \tag{26}$$

Une matrice $H_{ij}(f) \in \mathbb{R}^{n \times n}$ ne vérifiant aucunes de ces propriétés est dite *indéfinie*[2]. Un moyen pratique de vérifier si une matrice est définie positive est de calculer les n-déterminants[3] des sous-matrices de $H_{ij}(f)$,

2. Faudra-il encore vérifier que $H_{ij}(f)$ est semi-définie positive $\Longrightarrow \forall X \in \mathbb{R}^n : X^T H_{ij}(f) X \geq 0$ et $H_{ij}(f)$ est semi-définie négative $\Longrightarrow \forall X \in \mathbb{R}^n : X^T H_{ij}(f) X \leq 0$. Par ailleurs, une matrice symétrique vérifie $A = A^T$, ses valeurs propres sont réelles. De plus si A est définie positive alors elle est inversible et ses valeurs propres $\{\lambda_i\}_{i=1,n}$ sont positives

3. Ces n-déterminants sont appelés les *mineurs principaux dominants*.

soit :

$$\triangle_1 = \mid a_{11} \mid \quad \triangle_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \quad \triangle_3 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \quad \cdots\cdots \quad \triangle_n = \begin{vmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{vmatrix}$$

1) Si $\{\triangle_i\}_{i=1,n} > 0 \Rightarrow H_{ij}(f)$ est définie positive.
2) Si $\{\triangle_i\}_{i=1,n} < 0 \Rightarrow H_{ij}(f)$ est définie négative.

Concluons cette analyse sur la nature d'un point critique par le théorème ci-dessous.

Théorème *(conditions suffisantes d'optimalité) : Soit f une application $\mathbb{R}^n \longrightarrow \mathbb{R}$ de classe $\mathcal{C}^2$ sur $\mathbb{R}^n$. Si $X^* \in \mathbb{R}^n$ est un minimum local (respectivement un maximum local) de f alors :*

$$\nabla f(X^*) = 0 \qquad \textit{Condition d'optimalité du premier ordre}$$

et de plus,

$$H_{ij}(f(X^*)) = \nabla^2 f(X^*) \qquad \textit{Condition d'optimalité du second ordre}$$

est définie positive (respectivement définie négative)

La condition du premier ordre sur le gradient de la fonction, $\nabla f(X^*) = 0$, permet d'identifier un point stationnaire sans que l'on sache s'il s'agit d'un minimum, d'un maximum ou d'un point de selle. La condition du second ordre portant sur le *Hessien* de f, $\nabla^2 f(X^*)$, détermine la nature de ce point critique.

B. ***Définition du point-selle***

Soient $\mathcal{D}_1$ et $\mathcal{D}_2$ deux sous-ensembles et $f : \mathcal{D}_1 \times \mathcal{D}_2 \longmapsto \mathbb{R}$. Nous disons que le point critique $(x_s, y_s) \in \mathcal{D}_1 \times \mathcal{D}_2$ est un point-selle de f sur $\mathcal{D}_1 \times \mathcal{D}_2$ si :

$$\forall (x, y) \in \mathcal{D}_1 \times \mathcal{D}_2 : f(x_s, y) \leq f(x_s, y_s) \leq f(x, y_s) \tag{27}$$

Dans ce cas $f(x_s, y_s)$ est appelée la valeur-selle de $f(x, y)$. Autrement dit, $y \longmapsto f(x_s, y)$ atteint un maximum en y_s sur $\mathcal{D}_2$ et $x \longmapsto f(x, y_s)$ atteint un minimum en x_s sur $\mathcal{D}_1$.

Comme exemple d'application, considérons :

$$\begin{aligned} f : \underbrace{[-2\ ;\ 2]}_{\mathcal{D}_1} \times \underbrace{[-2\ ;\ 2]}_{\mathcal{D}_2} &\longmapsto \mathbb{R} \\ (x, y) &\longmapsto f(x, y) = x^2 - y^2 \end{aligned}$$

Ci-dessous le script Matlab® correspondant.

```
clear all ; clc ; close all ;

% Samir KENOUCHE Le 08/08/2019
% VISUALISATION DU POINT-SELLE DE F(X,Y) = X^2 - Y^2
lB = -2 ; uB = 2 ; n = 50 ; h = (uB - lB)/n ;

% POINT CRITIQUE => GRADIENT NUL AU POINT X* = (xs,ys) = (0,0)
xs = 0 ; ys = 0 ;  % POINT CRITIQUE OU STATIONNAIRE

[x,y] = meshgrid(lB :h: uB) ; fun = x.^2 - y.^2 ;
figure('color',[1 1 1]) ; surf(x,y,fun) ; hold on ;
plot(xs,ys,'ro','MarkerSize',22,'LineWidth',2) ;
plot(xs,ys,'rx','MarkerSize',22,'LineWidth',2) ; xlabel('x') ; ylabel('y') ;
title('POINT-SELLE DE F(X,Y) = X^2 - Y^2 AVEC (Xs,Ys) = (0,0)') ;
view(-32,14) ;
```

Exercice ❶ ☞ Ⓡ

Soit la fonction définie sur un ouvert $\mathcal{D} \subset \mathbb{R}^2$ et deux fois continument dérivable.

$$f(X) = x^4 + y^4 - 4\,x\,y + 1 \tag{28}$$

1) Déterminer les points critiques de f.
2) Identifier la nature de ces points critiques.

Solution: *(a) les points critiques sont déterminés en annulant le gradient* $\nabla f(X) = 0$ *:*

$$\begin{cases} \dfrac{\partial f}{\partial x} = 4\,x^3 - 4\,y = 4\,(x^3 - y) = 0 \\ \dfrac{\partial f}{\partial y} = 4\,y^3 - 4\,x = 4\,(y^3 - x) = 0 \end{cases} \tag{29}$$

Ce qui donne,

$$\frac{\partial f}{\partial x} = (x^3 - y) = 0 \tag{30}$$

$$\frac{\partial f}{\partial y} = (y^3 - x) = 0 \tag{31}$$

De l'équation (30)*, nous déduisons*

$$y = x^3 \tag{32}$$

Par substitution dans l'équation (31) *nous obtenons*[4] *:*

$$x^9 - x = 0 \Leftrightarrow x\,(x^8 - 1) = x\,(x^4 - 1)\,(x^4 + 1) = 0 \tag{33}$$

$$\Rightarrow x\,(x^2 - 1)\,(x^2 + 1)\,(x^4 + 1) = 0 \tag{34}$$

4. Nous utilisons la propriété suivante : $(a^2 - b^2) = (a - b) \times (a + b)$

$$\Rightarrow x = 0 \quad ; \quad x = 1 \quad ; \quad x = -1 \tag{35}$$

La solution pour laquelle $x^2 + 1 = 0 \Rightarrow x = \sqrt{j^2\,1} = \pm j \in \mathbb{C}$ *est refusée car* $X = (x, y) \in \mathbb{R}^2$. *A partir de* (32) *nous en déduisons les points critiques suivants :*

$$x = 0 \quad \Rightarrow \quad y = 0 \quad \Rightarrow \quad X_c^{(1)} = (0\ ;\ 0)^T$$

$$x = 1 \quad \Rightarrow \quad y = 1 \quad \Rightarrow \quad X_c^{(2)} = (1\ ;\ 1)^T$$

$$x = -1 \quad \Rightarrow \quad y = -1 \quad \Rightarrow \quad X_c^{(3)} = (-1\ ;\ -1)^T$$

(b) Identifions désormais la nature des ces points critiques. Nous devons donc déterminer la matrice Hessienne $\in \mathbb{R}^{2\times 2}$. *Calculons les éléments de cette matrice :*

$$\begin{cases} \dfrac{\partial f}{\partial x} = 4\,x^3 - 4\,y \quad \Rightarrow \quad \dfrac{\partial^2 f}{\partial x^2} = 12\,x^2 \\ \dfrac{\partial f}{\partial y} = 4\,y^3 - 4\,x \quad \Rightarrow \quad \dfrac{\partial^2 f}{\partial y^2} = 12\,y^2 \end{cases} \tag{36}$$

Pour les éléments hors de la diagonale principale :

$$\frac{\partial^2 f}{\partial x\,\partial y} = \frac{\partial^2 f}{\partial y\,\partial x} = -4 \tag{37}$$

D'où la matrice Hessienne,

$$\begin{pmatrix} 12\,x^2 & -4 \\ -4 & 12\,y^2 \end{pmatrix} \tag{38}$$

Calculons son déterminant,

$$|H_{2\times 2}(f)| = \begin{vmatrix} 12\,x^2 & -4 \\ -4 & 12\,y^2 \end{vmatrix} = 12\,x^2 \times 12\,y^2 - 16 \tag{39}$$

$$- X_c^{(1)} = (0\ ;\ 0)^T \quad \Rightarrow \quad |H_{2\times 2}(f)| = -16 < 0 \quad \Rightarrow \quad X_c^{(1)} \quad \textit{est un point selle}$$

$$- X_c^{(2)} = (1\ ;\ 1)^T \quad \Rightarrow \quad |H_{2\times 2}(f)| = 128 > 0 \quad \textit{et} \quad T_{H(f)} = 12\,x^2 + 12\,y^2 = 24 > 0$$

$$\Rightarrow \quad X_c^{(2)} \quad \textit{est un minimum local}$$

$$- X_c^{(3)} = (-1\ ;\ -1)^T \quad \Rightarrow \quad |H_{2\times 2}(f)| = 128 > 0 \quad \textit{et} \quad T_{H(f)} = 12\,x^2 + 12\,y^2 = 24 > 0$$

$$\Rightarrow \quad X_c^{(3)} \quad \textit{est un minimum local}$$

Exercice ❷ ☞ Ⓢ

Considérons les fonctions de trois variables, $f(X)$ tel que $X \in \mathbb{R}^3$, suivantes :

$$f_1(X) = x_1^2 + 2\,x_2^2 + 2\,x_3^2 + x_1\,x_2 + 2\,x_2\,x_3 + 4\,x_1 + 6\,x_2 + 12\,x_3$$

$$f_2(X) = 6\,x_1^2 + x_2^3 + 3\,x_2^2 + 6\,x_1\,x_2 + \frac{1}{4}\,x_3^4 - \frac{1}{3}\,x_3^3$$

$$f_3(X) = x_1^2 - 2\,x_2^2 - 3\,x_3^2 + 4\,x_1\,x_2 + 6\,x_1\,x_3 - 8\,x_2\,x_3$$

1) Identifier analytiquement les points stationnaires.
2) Déterminer la nature de ces points stationnaires.
3) Écrire chaque fonction sous la forme quadratique :

$$f(X) = \frac{1}{2}\,X^T\,A\,X - B^T\,X \quad \text{tel que} \quad \mathrm{A} \in \mathbb{R}^{n\times n} \quad \text{et} \quad B^T \in \mathbb{R}^n \tag{40}$$

4) Trouver la forme quadratique associée à la matrice :

$$\begin{pmatrix} 1 & 2 & 4 \\ 2 & 2 & 6 \\ 4 & 6 & 4 \end{pmatrix}$$

5) Tenant compte de (40), montrer que la forme quadratique algébrique s'écrit sous la forme :

$$f(X) = \sum_i a_i\,x_i^2 + 2\sum_i \sum_{j\neq i} b_{ij}\,x_i\,x_j \tag{41}$$

Chapitre[2] Algorithmes d'optimisation

Samir Kenouche - Département des Sciences de la Matière - UMKB

Module : Méthodes Mathématiques et Algorithmes pour la Physique

Résumé

Un processus d'optimisation consiste à choisir entre plusieurs solutions possibles, celle qui est la meilleure. Il s'agit ainsi de minimiser ou de maximiser un critère sur l'ensemble de toutes les solutions admissibles. Nous aborderons les algorithmes d'optimisation usuels à savoir : *Newton*, *quasi-Newton*, la famille des méthodes de *descente de gradient* et les algorithmes génétiques. La programmation des algorithmes abordés dans ce chapitre sera conduite au moyen de scripts Matlab®. Pendant les séances de cours, nous traiterons uniquement des problèmes d'optimisation non contraints. La résolution de problèmes contraints se fera lors des séances de travaux pratiques. La dernière section de ce chapitre est donnée à cet effet.

Table des matières

I. Introduction

Les algorithmes d'optimisation sans ou sous contraintes (unconstrained and constrained problem en anglais), de fonctions mathématiques uni et multidimensionnelles, ont pour objectif de chercher un vecteur $\hat{X}$ tel que $f(\hat{X})$ soit un extremum (minimum ou maximum) de la fonction en question. L'optimisation s'apparente systématiquement à une minimisation car trouver le maximum d'une fonction f revient à minimiser la fonction $-f$. Dans cette topologie, la fonction à optimiser est appelée *fonction-objectif* ou encore *critère d'optimisation*. Ainsi, l'opération d'optimisation est formalisée selon l'écriture :

$$\hat{X} \in \underset{X\in\mathbb{R}^n}{\operatorname{argmin}} f(X) \tag{1}$$

S. Kenouche est docteur en Physique de l'Université des Sciences et Techniques de Montpellier et docteur en Chimie de l'Université A. Mira de Béjaia.
page web personnelle : http://www.sites.univ-biskra.dz/kenouche
Version corrigée, améliorée et actualisée le 10.10.2020.

Avec $\hat{X} = (\hat{x}_1, \hat{x}_2, ..., \hat{x}_n)^T$ sont les coordonnées du point critique. On fait appel aux algorithmes d'optimisation afin de résoudre des problèmes de différentes natures, comme par exemple, trouver les zéros de fonctions non-linéaires, ajustement de données expérimentales selon le critère des *moindres carrés linéaire et non-linéaire*, résolution de systèmes d'équations de plusieurs variables ... etc. En général, la recherche des extremums est atteinte en procédant au calcul des dérivées premières (gradient de la fonction) et des dérivées secondes (Hessien de la fonction). Toutefois, trouver un minimum global n'est pas toujours chose facile et il n'y a pas d'algorithme d'optimisation parfait. En effet, pour résoudre un problème d'optimisation convenablement : (1) il faut bien poser le problème (2) choisir le bon algorithme (3) savoir interpréter les résultats qui en découlent.

A. *Convergence*

L'étude de la convergence d'une méthode numérique est atteinte à travers la suite des itérés $\left\{X^{(k)}\right\}_{k\in\mathbb{N}}$ générés par l'algorithme. Ce dernier est dit convergent si pour tout $X^{(0)} \in \mathbb{R}^n$ nous avons :

$$\lim_{k\to+\infty} \parallel X^{(k)} - X^* \parallel = 0 \tag{2}$$

Avec $X^* \in \mathbb{R}^n$ est la solution approchée, de la valeur exacte, déterminée avec une tolérance fixée préalablement. La condition (2) garantit, à partir d'une certaine itération, la satisfaction du critère d'arrêt pour la tolérance exigée. Nous rappelons que le taux de convergence (ou "vitesse" de convergence) et la complexité du problème traité rentrent en ligne de compte lors de l'utilisation d'un algorithme d'optimisation. La convergence ne signifie pas systématiquement l'existence d'un optimum même local. Le schéma numérique adopté doit être à la fois rapide et robuste. Ces deux derniers critères sont contrôlés par l'étude du taux de convergence qui quantifie l'erreur commise à chaque itération, soit :

$$e^{(k)} = \parallel X^{(k)} - X^* \parallel \quad \text{tel que} \quad \nabla f(X^*) = 0 \tag{3}$$

L'estimation de l'erreur servira, entre autre, à comparer le taux de convergence des différentes méthodes numériques. En pratique, l'erreur est représentée graphiquement en traçant la norme de l'erreur, soit $\parallel e^{(k+1)} \parallel$ en fonction de $\parallel e^{(k)} \parallel$ avec une échelle logarithmique. Ainsi, *l'ordre* noté p, de la méthode numérique s'obtient à partir de :

$$\parallel e^{(k+1)} \parallel \approx A \parallel e^{(k)} \parallel^p \Longrightarrow \log \parallel e^{(k+1)} \parallel \approx p \log \parallel e^{(k)} \parallel + \text{cste} \tag{4}$$

L'ordre, p, est quantifié via la pente de l'équation ci-dessus. Il en ressort les conclusions suivantes :

1) Si $p = 1 \Longrightarrow X^{(k)}$ converge linéairement vers la solution approchée. Dans ce cas on gagne la même quantité de précision à chaque itération. Il en résulte :

$$\lim_{k\to+\infty} = \frac{\parallel X^{(k+1)} - X^* \parallel}{\parallel X^{(k)} - X^* \parallel} = \tau \quad \text{avec} \quad 0 < \tau < 1 \tag{5}$$

   Elle est dite superlinéaire si $\tau = 0$.
2) Si $p = 2 \Longrightarrow X^{(k)}$ converge quadratiquement vers la solution approchée. Dans ce cas on gagne le double de précision à chaque itération.
3) Si $p = 3 \Longrightarrow X^{(k)}$ converge cubiquement vers la solution approchée. Dans ce cas on gagne le triple de précision à chaque itération. Il en résulte :

$$\lim_{k\to+\infty} = \frac{\parallel X^{(k+1)} - X^* \parallel}{\parallel X^{(k)} - X^* \parallel^p} = \tau \quad \text{avec} \quad \tau \geq 0 \tag{6}$$

D'un point de vue pratique et pour un k suffisamment élevé, la vitesse de convergence d'une méthode itérative est évaluée "empiriquement" au moyen de la relation :

$$K_p(X,k) = \frac{\| X^{(k+2)} - X^{(k+1)} \|}{\| X^{(k+1)} - X^{(k)} \|^p} \tag{7}$$

Bien évidemment, nous visons à ce que la convergence de l'algorithme soit la plus élevée possible afin de tendre vers la solution en un minimum d'itérations pour la tolérance exigée. La convergence quadratique est plus rapide que celle superlinéaire. A son tour cette dernière, est plus rapide que la convergence linéaire. Tenant compte de l'équation (7), il vient que plus $K_p(X,k)$ tend vers zéro plus le taux de convergence de la méthode est élevé.

B. ***Critères d'arrêt***

Les méthodes numériques, dites locales, procèdent de façon itérative. Typiquement, étant donné une valeur initiale, un nouvel itéré est mis à jour afin de converger vers un point stationnaire. Ce processus est réitéré jusqu'à la satisfaction d'un ou plusieurs critères d'arrêt de l'algorithme. D'un point de vue purement théorique, le schéma itératif est infini. En pratique, la suite d'approximations successives est tronquée dés que l'on considère avoir atteint la précision requise. Étant donné une tolérance ϵ, les critères d'arrêt pouvant êtres envisagés sont :

$$\| X^{(k+1)} - X^{(k)} \| < \epsilon \tag{8}$$

$$\| \nabla f(X^{(k)}) \| < \epsilon \tag{9}$$

$$\frac{\| f(X^{(k+1)}) - f(X^{(k)}) \|}{\| f(X^{(k)}) \|} < \epsilon \tag{10}$$

Il est possible d'envisager également un quatrième critère, celui du nombre d'itérations dépassant un seuil fixé préalablement. Il se peut que le critère (23) ne soit pas satisfait même si l'algorithme converge. Les erreurs d'arrondis dues à l'accumulation des opérations arithmétiques peuvent êtres du même ordre de grandeur que le gain de précision obtenu à l'itération en cours. Le critère (8) est recommandé, le schéma itératif est interrompu lorsqu'il ne produit plus de gain significatif en terme de précision. En outre, dans certaines situations la divergence d'un algorithme ne signifie pas forcément l'inexistence de la solution, il faudra juste adapter le nombre d'itérations et/ou la tolérance considérés. D'un point de vue purement pratique, une combinaison de ces critères est prise en compte.

II. Algorithme de Newton

Comme il a été signalé au début de ce chapitre, l'opération d'optimisation s'apparente systématiquement à une minimisation, car trouver le maximum d'une fonction f revient à minimiser la fonction $-f$ selon l'équivalence :

$$\text{Arg}\max_x f(x) \Leftrightarrow \text{Arg}\min_x(-f(x)) \tag{11}$$

Les deux opérations d'optimisation sont pleinement équivalentes. La méthode de Newton est efficiente, notamment parce qu'elle prend en compte la courbure de la fonction-objectif à travers le calcul de sa seconde dérivée. Cette méthode est praticable si à chaque itération, la dérivée seconde est définie ou la matrice hessienne est définie positive pour le cas $X^{(k)} \in \mathbb{R}^n$. Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^2$. Le développement en série de *Taylor* d'ordre deux (modèle quadratique) est une fonction $P_{X^{(k)}}(X) : \mathbb{R}^n \longrightarrow \mathbb{R}$, avec :

$$P_{X^{(k)}}(X) = f(X^{(k)}) + (X - X^{(k)})^T \nabla f(X^{(k)}) + \frac{1}{2!}(X - X^{(k)})^T \nabla^2 f(X^{(k)})(X - X^{(k)}) \tag{12}$$

Avec, $\nabla^2 f(X^{(k)})$ est la matrice Hessienne de f en $X = X^{(k)}$ soit $H_f(X^{(k)}) = \nabla^2 f(X^{(k)})$. Posons $U = X - X^{(k)}$, il vient :

$$P_{X^{(k)}}(X+U) = f(X^{(k)}) + U^T \nabla f(X^{(k)}) + \frac{1}{2!} U^T \nabla^2 f(X^{(k)}) U \tag{13}$$

$$\min\{P_{X^{(k)}}(X+U)\} \Longleftrightarrow \nabla P_{X^{(k)}}(X+U) = 0 \tag{14}$$

Une condition suffisante d'optimalité :

$$\nabla P_{X^{(k)}}(X+U) = \nabla f(X^{(k)}) + \nabla^2 f(X^{(k)}) U = 0 \tag{15}$$

$$\nabla f(X^{(k)}) = -U \nabla^2 f(X^{(k)}) \Rightarrow U = -\frac{\nabla f(X^{(k)})}{\nabla^2 f(X^{(k)})} \tag{16}$$

$$X = X^{(k)} - \frac{\nabla f(X^{(k)})}{\nabla^2 f(X^{(k)})} \quad \text{avec} \quad U = X - X^{(k)} \tag{17}$$

On calcule alors un nouveau point, $X^{(k+1)}$, qui minimise $P_{X^{(k)}}$ soit :

$$X^{(k+1)} = X^{(k)} - \frac{\nabla f(X^{(k)})}{\nabla^2 f(X^{(k)})} \tag{18}$$

Ou bien avec une écriture équivalente :

$$X^{(k+1)} = X^{(k)} - H_f(X^{(k)})^{-1} f(X^{(k)}) \tag{19}$$

Ci-dessous l'écriture algorithmique de la méthode de *Newton*

Algorithm 1 Algorithme de Newton

Input : $f(X) \in \mathcal{C}^2$, $X^{(0)} \in \mathbb{R}^n$
$k \longleftarrow 0$

1. **Tant que** critère d'arrêt n'est pas vérifié faire :
2. Calcul du gradient $G_k \longleftarrow -\nabla f(X^{(k)})$
3. Calcul de H_f^{-1} $H_k \longleftarrow [H_f(X^{(k)})]^{-1}$
4. Nouvelle itération $X^{(k+1)} \longleftarrow X^{(k)} + G_k H_k$
5. Mise à jour : $k \longleftarrow k+1$
6. **Fin**

Appliquons le schéma numérique (19) à la fonction-objectif ci-dessous.

Exercice ❶ ☞ Ⓡ

Soit la fonction-objectif de trois variables, $f(X)$ tel que $X \in \mathbb{R}^3$, suivante :

$$f(X) = \frac{1}{4}\,x_3^4 + x_2^3 - \frac{1}{3}\,x_3^3 + 6\,x_1^2 + 3\,x_2^2 + 6\,x_1\,x_2 \tag{20}$$

La tolérance vaut $\epsilon = 10^{-5}$ et pour l'évaluation numérique prendre $X^{(0)} = (-0.6637, 0.1758, 1.5778)^T$.

1) Trouver analytiquement l'optimum de la fonction-objectif.
2) Évaluer numériquement la fonction-objectif en utilisant l'algorithme de Newton.
3) Écrire un script Matlab® de l'algorithme de Newton permettant sa minimisation.

Les résolutions analytique et numérique seront menées pendant la séance de cours. Le script Matlab® est donné ci-dessous

```
clear all ; clc ; close all ;

% Samir KENOUCHE - LE 08/08/2019
% ALGORITHME DE NEWTON

fx = @(x) (1/4)*x(3)^4 + x(2)^3 - (1/3)*x(3)^3 + 6*x(1)^2 + ...
    3*x(2)^2 + 6*x(1)*x(2) ;

rng(1110,'twister')  ; % INITIALISATION DU GENERATEUR DES NOMBRES ALEATOIRES
xinit = randn(1,3)' ;  % INITIALISATION ALEATOIRE

it = 0 ; itmax = 20 ;  % NOMBRE D'ITERATIONS
err = 10.^(-5) ;       % TOLERANCE

while it < itmax

    Hx(1,1) = 12 ; Hx(1,2) = 6 ; Hx(1,3) = 0 ;
    Hx(2,1) = 6 ; Hx(2,2) = 6*xinit(2) + 6 ; Hx(2,3) = 0 ;
    Hx(3,1) = 0 ; Hx(3,2) = 0 ; Hx(3,3) = 3*xinit(3)^2 - 2*xinit(3) ;

    dfx = [12*xinit(1) + 6*xinit(2) ; ...
         3*xinit(2)^2+ 6*xinit(1) + 6*xinit(2) ; xinit(3)^3 - xinit(3)^2] ;

    xk = xinit - inv(Hx)*dfx  ;

         if  norm(inv(Hx)*dfx) <= err   % TEST D'ARRET
               solution = xk ;          % OPTIMUM OBTENU
               f_min = fx(solution) ;   % MINIMUM DE LA FONCTION-OBJECTIF
               iteration_max = it ;     % NOMBRE D'ITERATIONS PERMETTANT LA CV
          break
         end

    xinit = xk ; it = it + 1 ;          % MISE A JOUR
end

%% SOLUTION : X* =(0, 0, 1) EN 04 ITERATIONS
```

Il est possible de voir la méthode de *Newton* comme un algorithme de descente de gradient à pas fixe valant l'unité. La famille des algorithmes de descente de gradient fera l'objet des sections suivantes. Nonobstant, la simplicité et la convergence quadratique de l'algorithme de Newton, ce dernier souffre d'un certain nombres d'inconvénients : il peut diverger si le point initial est trop éloigné de la solution recherchée. En outre, l'algorithme ne peut être utilisé si la fonction-objectif n'est pas deux fois dérivables. Dans certaines situations, même si la dérivée seconde ou le Hessien pour la cas multidimensionnel existe, son calcul peut se révéler fastidieux ou très laborieux ou encore impossible à atteindre.

***Nota Bene**: Toutes les méthodes numériques ont une structure d'une suite numérique. Ce qui les distinguent, c'est la formule du schéma numérique de la méthode en question. Toute suite de nombres réels (ou de nombre complexe $\mathbb{C}$) est convergente si et seulement si elle vérifie le critère de Cauchy.*

***Définition d'une suite de Cauchy**: Une suite $\left\{X^{(k)}\right\}_{k\in\mathbb{N}}$ est de Cauchy (ou vérifie le critère de Cauchy) si pour un rang $k \in \mathbb{N}$ suffisamment grand nous avons :*

$$\lim_{k\longrightarrow+\infty} \| X^{(k)} - X^{*} \| = 0 \tag{21}$$

Ce critère s'énonce aussi,

$$\forall \epsilon > 0, \exists\, \eta \quad tel\ que \quad k > \eta \quad (ou \quad \eta(\epsilon)) \Rightarrow \| X^{(k)} - X^{*} \| \leq \epsilon$$

Une façon de démontrer le théorème consiste à démontrer que $\left\{X^{(k)}\right\}_{k\in\mathbb{N}}$ est une suite de Cauchy[1]*. Soit $q \in [0\,;\,1[$ et $\forall k \in \mathbb{N}$, nous avons :*

$$\| f(X^{(k)}) - f(X^{(k-1)}) \| \leq q \ \| X^{(k)} - X^{(k-1)} \|$$

$$\Rightarrow \quad \| X^{(k+1)} - X^{(k)} \| \leq q \ \| X^{(k)} - X^{(k-1)} \| \tag{22}$$

D'un autre côté, nous pouvons écrire

$$\| X^{(k)} - X^{(k-1)} \| \leq q \ \| X^{(k-1)} - X^{(k-2)} \| \tag{23}$$

En substituant (23) dans (24) nous obtenons :

$$\Rightarrow \quad \| X^{(k+1)} - X^{(k)} \| \leq q^2 \ \| X^{(k-1)} - X^{(k-2)} \| \tag{24}$$

En procédant par récurrence nous démontrons[2] *:*

$$\Rightarrow \quad \| X^{(k+1)} - X^{(k)} \| \leq q^k \ \| X^{(1)} - X^{(0)} \| \tag{25}$$

Avec $q \in [0\,,\,1[$, la distance $\| X^{(k+1)} - X^{(k)} \|$ tend vers zéro pour k assez grand alors $\left\{X^{(k)}\right\}_{k\in\mathbb{N}}$ est une suite de Cauchy donc elle est convergente.

1. Rappelons que toute suite convergente dans $\mathbb{R}$ est une suite de Cauchy.
2. Pour les puristes, nous opterons plutôt pour $\| X^{(k+1)} - X^{(k)} \| \leq \dfrac{q^k}{1-q} \ \| X^{(1)} - X^{(0)} \|$

A. *Algorithme de quasi-Newton*

L'intérêt d'utiliser l'algorithme de quasi-Newton par rapport à celui de Newton tient aux éléments suivants : (1) le Hessien n'est pas calculé explicitement, ce dernier exige un temps de calcul souvent très élevé (2) la détermination de la direction de descente d_k ne nécessite qu'une multiplication matrice-vecteur, sans se soucier du calcul de gradient (3) les approximations du Hessien $H^{(k+1)}$ sont définies positives, garantissant ainsi que les directions de recherche sont systématiquement descendantes. Le schéma numérique de Broyden-Fletcher-Goldforb-Shanno, communément appelé BFGS est :

$$H^{(k+1)} = H^{(k)} + \left(1 + \frac{y_k^T\, H^{(k)}\, y_k}{y^T\, d_k}\right) \frac{d_k\, d_k^T}{y_k^T\, d_k} - \frac{H^{(k)}\, y_k\, d_k^T + d_k\, y_k^T\, H^{(k)}}{y_k^T\, d_k} \tag{26}$$

Avec,

$$y_k = \nabla f(X^{(k+1)}) - \nabla f(X^{(k)})$$

D'autres versions de cet algorithme existent aussi, nous citerons notamment la formule de Davidon-Fletcher-Powell, communément appelée DFP. L'inconvénient des méthodes de quai-Newton est que l'on exploite pas toute l'information portant sur la forme de la fonction-objectif. Par ailleurs, l'initialisation de $H_0 \in \mathbb{R}^{n\times n}$ se fait très souvent avec la matrice identité $I \in \mathbb{R}^{n\times n}$. La pratique montre que ce choix semble donner des résultats très probants. Ci-dessous, l'écriture algorithmique de la méthode de *Quasi-Newton*

Algorithm 2 Algorithme de Quasi-Newton

Input : $f(X) \in \mathcal{C}^1$, $X^{(0)} \in \mathbb{R}^n$, $H_0 = I_{n\times n}$
$k \longleftarrow 0$

1. **Tant que** critère d'arrêt n'est pas vérifié faire :
2. Calcul du gradient $G_k \longleftarrow -\nabla f(X^{(k)})$
3. Pas optimal $\lambda_k \longleftarrow \text{argmin}(X^{(k)} + \lambda\, G_k)$
4. La descente $d_k \longleftarrow \lambda_k\, G_k\, H_0$
5. Nouvelle itération $X^{(k+1)} \longleftarrow X^{(k)} + d_k$
6. Calcul intermédiaire $y_k \longleftarrow \nabla f(X^{(k+1)}) - \nabla f(X^{(k)})$
7. Approx. de la Hess. $H^{(k+1)} \longleftarrow H^{(k)} + \left(1 + \frac{y_k^T\, H^{(k)}\, y_k}{y^T\, d_k}\right) \frac{d_k\, d_k^T}{y_k^T\, d_k} - \frac{H^{(k)}\, y_k\, d_k^T + d_k\, y_k^T\, H^{(k)}}{y_k^T\, d_k}$
8. Mise à jour : $k \longleftarrow k+1$
9. **Fin**

Exercice ❷ ☞ ®

Nous appliquerons cette méthode pour chercher le minimum de la fonction-objectif à deux dimensions suivante :

$$\begin{cases} f(X) = \dfrac{1}{2}\, x^2 + \dfrac{7}{2}\, y^2 \\ \\ \text{Avec} \quad X^{(0)} = (7.5, 2.2) \end{cases} \tag{27}$$

La tolérance considérée est $\epsilon = 10^{-5}$. Ci-dessous, le script Matlab®

```
close all ; clc ; clear all ;

% Le 12.08.2019 - Samir KENOUCHE
% Algorithme de BFGS - quasi-Newton

n = 2 ; Hinit = eye(n,n) ;
fxy = @(x,y) x.^2*(1/2) + y.^2*(7/2) ; dfx = @(x) x ; dfy = @(y) 7.*y ;
it = 0 ; itmax = 10 ; tol =  1e-5 ; xinit = [7 ; 2] ;

while it < itmax

   d = - Hinit*[dfx(xinit(1)) ; dfy(xinit(2))] ;
   lambdak = -(xinit(1)*d(1) + 7*xinit(2)*d(2))/(d(1).^2 + 7*d(2).^2) ;

   dk = lambdak*d ; xk = xinit + dk ;

   if norm(dk) < tol
     solution = xk   ; % VECTEUR SOLUTION
     nbr_it = it ;     % NOMBRE D'ITERATION PERMETTANT LA CV
       break
   end

   yk = [dfx(xk(1)) ; dfy(xk(2))] - [dfx(xinit(1)) ; dfy(xinit(2))] ;
   Hk = Hinit + (1 + (yk'*Hinit*yk)/(yk'*dk))*(dk*dk')/(yk'*dk) - ...
       (Hinit*yk*dk' + dk*yk'*Hinit)/(yk'*dk) ;

   Hinit = Hk ; xinit = xk ;  it = it + 1 ; % MISE A JOUR
end

solution
```

Les méthodes de Newton et de quasi-Newton peuvent rencontrer des problèmes tels que le calcul du Hessien soit trop complexe ou qu'il n'existe pas. La nécessité d'appliquer une inversion matricielle à chaque itération, ceci peut se révéler rédhibitoire pour des problèmes d'optimisation mobilisant beaucoup de variables. Ces méthodes peuvent donc devenir impraticables. Une alternative consiste à utiliser la famille des algorithmes de descente de gradient. Ces méthodes ne requièrent pas le calcul explicite ou l'approximation du Hessien. Elles n'exigent pas *de facto* le stockage du Hessien ($\mathbb{R}^{n\times n}$), mais seulement un ou quelques vecteurs ($\in \mathbb{R}^n$).

III. Algorithmes de descente de gradient

Un algorithme de descente de gradient est mis en œuvre suivant les modes de choix des *directions* de descente successives, ensuite par l'amplitude du *pas* effectué dans la direction choisie. La famille des algorithmes de descente de gradient est à la base des processus d'optimisation de problèmes plus au moins complexes. Le terme *descente* vient du fait que cet algorithme cherche l'extremum suivant une direction opposée à celle du gradient de la *fonction-objectif*. Le pas de descente λ_k est soit constant (*méthode de gradient à pas fixe*) ou bien incrémenté selon les méthodes de *Wolfe*, de *gradient à pas optimal* et de *gradient conjugué*. Typiquement, le pas de descente est obtenu au moyen d'une recherche linéaire vérifiant : $f(x^{(k)} + \lambda_k \nabla f(x^{(k)})) < f(x^{(k)})$. Dans cette section, nous rappellerons succinctement la notion de la *dérivée directionnelle*. Cette étape est importante afin de rendre compte du principe de fonctionnement des algorithmes de descente de gradient. Plus de détails sur cette notion sont disponibles *en annexe* (A).

Par définition, la dérivée directionnelle de $f(x,y)$ dans la direction du vecteur unitaire $\overrightarrow{d} = a\,\vec{i} + b\,\vec{j}$ au point $f(x_0,y_0)$ est :

$$f_{\overrightarrow{d}}(x_0,y_0) = \lim_{\lambda\to 0} \frac{f(x_0+\lambda\,a, y_0+\lambda\,b) - f(x_0,y_0)}{\lambda} = \nabla f(x_0,y_0)\ \overrightarrow{d} \tag{28}$$

Théorème *: La dérivée directionnelle est maximale lorsque* $\overrightarrow{d}$ *a la même direction que* $\nabla f(x_0,y_0)$ *de plus le taux de variation maximal de* $f(x,y)$ *en* (x_0,y_0) *est* $||\nabla f(x_0,y_0)||$.

Ce théorème peut être prouvé en considérant que $f_{\overrightarrow{d}}(x_0,y_0) = \nabla f(x_0,y_0)\ \overrightarrow{d} = ||\nabla f(x_0,y_0)||\,||\overrightarrow{d}||\cos(\theta) = ||\nabla f(x_0,y_0)||\cos(\theta)$. *Ainsi* $f_{\overrightarrow{d}}(x_0,y_0)$ *est maximale si* $\cos(\theta) = \pm 1$ *autrement dit si la condition* $\nabla f(x_0,y_0)//\overrightarrow{d}$ *est satisfaite. Dans le cas où* $\theta = \pi/2 \Rightarrow \nabla f(x_0,y_0) \perp \overrightarrow{d}$. *Ce résultat indique que si je me déplace dans une direction perpendiculaire au* ∇f, *le taux de variation de la fonction* $f(x,y)$ *est nul.*

- *Si les deux vecteurs* ∇f *et* $\overrightarrow{d}$ *ont la même direction et le même sens, dans ce cas le vecteur unitaire* $\overrightarrow{d}$ *désigne une direction de croissance maximale de* $f(x,y)$.
- *Si les deux vecteurs* ∇f *et* $\overrightarrow{d}$ *ont la même direction et de sens opposé, dans ce cas le vecteur unitaire* $\overrightarrow{d}$ *désigne une direction de décroissance maximale de* $f(x,y)$. *Cette condition est prouvée selon :*

$$\begin{aligned}
\nabla f(x_0,y_0) \times \overrightarrow{d} &= \nabla f(x_0,y_0) \times (-\nabla f(x_0,y_0)) \\
&= -\nabla f(x_0,y_0) \times \nabla f(x_0,y_0) \\
&= \underbrace{-\underbrace{||\nabla(x_0,y_0)||^2}_{>0}}_{<0} \\
&\Rightarrow \overrightarrow{d} = -\nabla f \quad \textit{est une direction de descente}
\end{aligned}$$

Théorème*: Soit* f *une fonction* $\mathcal{C}^1$ *sur un ouvert* $\mathcal{D}$ *de* $\mathbb{R}^2$ *et* P_0 *un point d* $e\mathcal{D}$ *tel que* $f(P_0) = k$. *Nous supposons que le* ∇f *est non nul au point* $P_0 = (x_0,y_0)$. *La tangente en* P_0 *à la courbe d'équation* $f(x,y) = c$ *a comme équation :*

$$\nabla f(x_0,y_0) \cdot \vec{P_0P} = 0 \tag{29}$$

Ou de façon équivalente :

$$\frac{\partial f(x_0,y_0)}{\partial x}\,(x-x_0) + \frac{\partial f(x_0,y_0)}{\partial y}\,(y-y_0) = 0 \tag{30}$$

Démonstration*: Nous travaillerons sur la courbe de niveau* $f(x,y) = 0$. *Le point* $P_0(x_0,y_0)$ *peut se déplacer sur la courbe de niveau de sorte que des coordonnées changent en fonction du temps. Cela nous permet d'écrire* $P_0(x_0 + a\,t,\ y_0 + b\,t)$. *Ainsi nous obtenons l'équation paramétrique (de paramètre* t*) :*

$$f(x_0 + a\,t,\ y_0 + b\,t) = f(x(t), y(t)) = 0 \tag{31}$$

En dérivant (31) *nous obtenons :*

$$\frac{df}{dt} = \frac{\partial f}{\partial x}\,\frac{dx(t)}{dt} + \frac{\partial f}{\partial y}\,\frac{dy(t)}{dt} = 0 \tag{32}$$

Au point $P_0(x_0,y_0)$ *correspondant à* $t = 0$, *il vient :*

$$\frac{df(x_0,y_0)}{dt} = \frac{\partial f(x_0,y_0)}{\partial x}\,\frac{dx(t)}{dt} + \frac{\partial f(x_0,y_0)}{\partial y}\,\frac{dy(t)}{dt} = 0 \tag{33}$$

$$\Rightarrow \quad \frac{df(x_0,y_0)}{dt} = \frac{\partial f(x_0,y_0)}{\partial x}\, a + \frac{\partial f(x_0,y_0)}{\partial y}\, b = 0 \tag{34}$$

$$\Rightarrow \quad \frac{df(x_0,y_0)}{dt} = \nabla f(x_0,y_0)\cdot(a,\, b) \tag{35}$$

Avec a et b sont les coordonnées de la pente de l'équation tangente au point (x_0, y_0). Nous concluons ainsi que le gradient de f est toujours orthogonal (produit scalaire (35) est nul) aux courbes de niveau.

A. *Méthode de gradient à pas fixe*

L'idée de base de cette approche consiste à chercher une suite $\{x_k\}_{k\in\mathbb{N}} \in \mathbb{R}^n$ dont le successeur de x_k doit satisfaire la condition :

$$f(x_{k+1}) < f(x_k) \tag{36}$$

Afin d'établir cette suite, on exploitera la dérivée directionnelle définie précédemment. Considérons la figure ci-dessous :

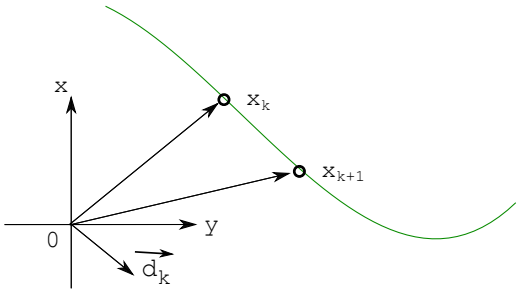


FIGURE 1: Dérivée directionnelle dans la direction $\vec{d_k}$.

Compte tenu de l'Eq. (92) (voir l'annexe), à une dimension on obtient :

$$f_{\overrightarrow{d}}(x_k) = \lim_{\lambda\to 0} \frac{f(x_{k+1}) - f(x_k)}{\lambda} \Rightarrow f_{\overrightarrow{d}}(x_k) = \lim_{\lambda\to 0} \frac{f(x_k + \lambda\, a) - f(x_k)}{\lambda} \tag{37}$$

Comme précédemment, nous avons

$$\overrightarrow{x_k x_{k+1}} // \overrightarrow{d_k} \Rightarrow \overrightarrow{x_k x_{k+1}} = \lambda\, \overrightarrow{d_k} \quad \text{avec} \quad \lambda \in \mathbb{R}_+^*$$

D'un autre côté,

$$\overrightarrow{x_k x_{k+1}} = \overrightarrow{O x_{k+1}} - \overrightarrow{O x_k} \quad \Rightarrow \quad \overrightarrow{O x_{k+1}} = \overrightarrow{O x_k} + \overrightarrow{x_k x_{k+1}}$$

$$\Rightarrow \quad \overrightarrow{O x_{k+1}} = \overrightarrow{O x_k} + \lambda\, \overrightarrow{d_k} \tag{38}$$

D'où le schéma numérique de l'algorithme de descente de gradient :

$$x^{(k+1)} = x^{(k)} + \lambda\, d_k = x^{(k)} - \lambda \nabla f(x^{(k)}), \qquad \lambda > 0 \tag{39}$$

Ainsi,

$$f(x_{k+1}) < f(x_k) \Leftrightarrow f(x^{(k)} - \lambda \nabla f(x^{(k)}) < f(x_k) \tag{40}$$

Avec, $d_k = -\nabla f(x^{(k)})$ est la direction de plus forte descente de f au point x_k. Le paramètre λ est le pas de descente qui est constant dans cette version de l'algorithme. Notons que la vitesse de convergence de cet algorithme est proportionnelle à la valeur de λ. Une valeur initiale pour incrémenter l'algorithme est $\lambda = 0.01$. Cette méthode est itérative, elle a donc besoin bien évidemment d'une valeur initiale pour démarrer l'incrémentation de l'algorithme. Ci-dessous, l'écriture algorithmique de la méthode de gradient à pas fixe.

Algorithm 3 Algorithme de descente de gradient à pas fixe

Input : $f(X) \in \mathcal{C}^1$, $X^{(0)} \in \mathbb{R}^n$, $\lambda \in \mathbb{R}_+^*$
$k \longleftarrow 0$

1. **Tant que** critère d'arrêt n'est pas vérifié faire :
   2. Direction de descente $d_k \longleftarrow -\nabla f(X^{(k)})$
   3. Nouvelle itération $X^{(k+1)} \longleftarrow X^{(k)} + \lambda_k\, d_k$
   4. Mise à jour : $k \longleftarrow k + 1$
5. **Fin**

Afin d'illustrer le principe de fonctionnement de cet algorithme, nous chercherons à optimiser une fonction-objectif d'une seule variable.

Exercice ❸ ☞ ®

Cherchons la minimisation de la fonction-objectif :

$$f(x) = 4\,x^2 + e^x \quad \text{avec} \quad \nabla f(x) = 8\,x + e^x$$

Cette fonction est au moins de classe $\mathcal{C}^1$ sur $\mathbb{R}$. Exécutons le schéma numérique de l'algorithme de descente de gradient à pas fixe pour $x_0 = 1.50000$, $\lambda = 0.02$ et $\epsilon = 10^{-3}$, soit :

$$x^{(k+1)} = x^{(k)} + \lambda\, d_k = x^{(k)} - \lambda \nabla f(x^{(k)}) = x^{(k)} - \lambda \left[8\,x^{(k)} + e^{x^{(k)}}\right]$$

Commençons par appliquer le schéma numérique ci-dessus,

$$\begin{aligned}
x^{(1)} &= 1.500000 - 0.02\,\nabla f(1.500000) = 1.170366 \\
x^{(2)} &= 1.170366 - 0.02\,\nabla f(1.170366) = 0.918644 \\
x^{(3)} &= 0.918644 - 0.02\,\nabla f(0.918644) = 0.721543 \\
x^{(4)} &= 0.721543 - 0.02\,\nabla f(0.721543) = 0.564944 \\
x^{(5)} &= 0.564944 - 0.02\,\nabla f(0.564944) = 0.439366 \\
&\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \\
&\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \\
x^{(30)} &= -0.10695 - 0.02\,\nabla f(-0.10695) = -0.10781
\end{aligned}$$

Le test d'arrêt $|x^{(k+1)} - x^{(k)}| < \epsilon$ est positif au bout de la $30^{\text{ème}}$ itération, soit :

$$|x^{(30)} - x^{(29)}| = 8.5000 \times 10^{-4} < \epsilon$$

L'algorithme converge vers la solution $\underline{x^* = -0.10781}$ correspondant au minimum de la fonction-objectif considérée.

Exercice ❹ ☞ Ⓡ

Soit la fonction-objectif ci-dessous :

$$\begin{cases} f(x) = \cos(2\,x)\,\sqrt{x^2+1} \\ \text{Avec} \quad x_0 = -2,\ x \in [-4, 0] \end{cases} \tag{41}$$

1) Écrire un script Matlab® de l'algorithme de descente de gradient permettant la minimisation de la fonction-objectif.

Script Matlab®

```
clear all ; close all ; clc ;
% Le 05.08.2019 - Samir KENOUCHE
%%%%%%%%%%%%%%%%%%%%%% DSCENTE DE GRADIENT A PAS FIXE %%%%%%%%%%%%%%%%%%%%%%
syms x
fun = cos(2*x) + sqrt(x^2 + 1) ; dfun = diff(fun,x) ;
xi = -4:0.004: 0 ; fx = subs(fun, x, xi) ; dfx = subs(dfun, x, xi) ;

xinit = -3 ; lambda = 0.1/2 ; itmax = 100 ; it = 0 ; echelle = 0.5 ;
tol =  1e-3 ; plot(xi,fx,'LineWidth',1) ; hold on ;

 while it < itmax

    J = subs(dfun, x, xinit) ;
    xn = xinit - lambda*J ; xinit = xn ;

    it = it + 1;

    fxn = subs(fun, x, xn) ; dfxn = subs(dfun , x, xn) ;

    if abs(lambda*J) < tol
         sol = xn
         break
    end

    dk = - dfxn ; % DIRECTION DE DESCENTE
    plot(xn, fxn,'xr','MarkerSize',7) ; hold on ;
    plot(xn, fxn,'or','MarkerSize',7) ; hold on ;
   quiver(xn,fxn,dk,0, echelle,'Color', 'k','LineWidth',1) ; hold on;
 end
h = gca ;
str(1) = {'Plot of the function :'};
str(2) = {'$$y = \cos(2\,x) + \sqrt{x^2 + 1}$$'};
str(3) = {'With the minimum:'};
str(4) = {['$$x_{min} = $$', num2str(sol)]};
set(gcf,'CurrentAxes',h) ;

text('Interpreter','latex', 'String',str,'Position',[-3.9 1.5],'FontSize',12)
```

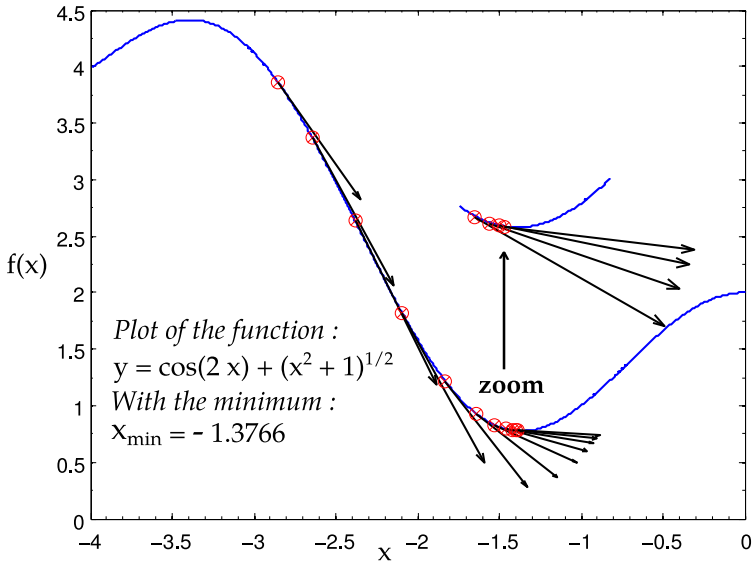


FIGURE 2: Minimisation avec la *descente de gradient à pas fixe*

Le minimum renvoyé par ce script est : `sol = -1.3668`. L'inconvénient de cette méthode est qu'elle réclame une valeur initiale très proche de la solution approchée. Dans le cas contraire, l'algorithme ne convergera pas vers le véritable minimum de la fonction. Nous faisons remarquer que le critère d'arrêt $\| X^{(k+1)} - X^{(k)} \|$ est totalement équivalent à celui de $\| \lambda_k \nabla f X^{(k)} \|$.

B. *Méthode de gradient à pas optimal*

L'inconvénient d'utiliser un pas de descente constant ($\lambda = cst$) est que l'algorithme converge très lentement si le pas est trop petit. De plus, l'algorithme peut devenir instable dans le cas où le pas est trop grand. On comprend donc que le choix optimal de la valeur de λ devient délicat dans ce cas de figure. Il faudra par conséquent ajuster la valeur de λ à chaque itération. Cette procédure est appelée *recherche en ligne* (line search) [3]. Nous appliquerons à cet effet la méthode de *gradient à pas optimal*. L'idée de base de cette méthode est de chercher λ_k diminuant d'avantage $f(x^{(k)})$ dans la direction $d_k = -\nabla f(x^{(k)})$, selon :

$$\varphi(\lambda_k) = \underset{\lambda>0}{\operatorname{argmin}}\, f(x^{(k)} - \lambda \nabla f(x^{(k)})) \Leftrightarrow \varphi^{'}(\lambda_k) = 0 \Leftrightarrow \varphi(\lambda_{opt}) \leq \varphi(\lambda_k) \tag{42}$$

Il convient de noter, qu'en toute rigueur le pas de descente n'est pas λ mais $\lambda\, d_k$. L'opération d'optimisation du pas de descente permet de répondre à la question : quelle distance doit-on parcourir ? Cherchons la valeur du pas λ_k minimisant une fonction-objectif $f(X^{(k+1)})$, avec $X \in \mathbb{R}^n$, dans la direction de descente d_k. Nous avons :

$$f(X^{(k+1)}) = f(X^{(k)} + \lambda_k\, d_k) \tag{43}$$

Dans ce chapitre, il est question que de fonctions quadratiques. Ainsi la fonction $f(X)$ peut se mettre sous la forme :

$$f(X) = \frac{1}{2} X^T A X - b^T X \Rightarrow \nabla f(X) = A X - b = g \tag{44}$$

3. Notons qu'il existe d'autres méthodes de recherche linéaire moins restrictives. Nous citerons les méthodes de Armijo, de Goldstein et de Wolf. Ces méthodes autorisent n'ont pas un choix unique du pas optimal, mais un intervalle de valeurs.

Avec A une matrice symétrique définie positive. Le schéma numérique de l'algorithme de descente de gradient impose :

$$X^{(k+1)} = X^{(k)} - \lambda_k \nabla f(X^{(k)}) \tag{45}$$

Nous cherchons un pas optimal tel que :

$$\lambda_k = \underset{\lambda \geq 0}{\operatorname{argmin}} f(X^{(k)} - \lambda \nabla f(X^{(k)})) \tag{46}$$

Ainsi,

$$\frac{\partial f(X^{(k+1)})}{\partial \lambda}(\lambda = \lambda_k) = 0 \Leftrightarrow \nabla f(X^{(k+1)})\ \nabla f(X^{(k)}) = 0 \tag{47}$$

Tenant compte de (44) il vient

$$\begin{aligned}
\nabla f(X^{(k+1)})\ \nabla f(X^{(k)}) &= \Big[A\,X^{(k+1)} - b\Big]^T\ \nabla f(X^{(k)}) = 0 \\
&= \Big[A\,\Big[X^{(k)} - \lambda_k \nabla f(X^{(k)})\Big] - b\Big]^T\ \nabla f(X^{(k)}) = 0 \\
&= \Big[A\,X^{(k)} - b - A\,\lambda_k \nabla f(X^{(k)})\Big]^T\ \nabla f(X^{(k)}) = 0 \\
&= (A\,X^{(k)} - b)^T \nabla f(X^{(k)}) - \lambda_k \nabla f(X^{(k)})^T A \nabla f(X^{(k)}) = 0
\end{aligned}$$

$$\Rightarrow \lambda_k = \frac{(A\,X^{(k)} - b)^T \nabla f(X^{(k)})}{\nabla f(X^{(k)})^T A \nabla f(X^{(k)})} = \frac{\nabla f(X^{(k)})^T \nabla f(X^{(k)})}{\nabla f(X^{(k)})^T A \nabla f(X^{(k)})} \tag{48}$$

La formule générale de la méthode de descente de gradient à pas optimal dans la direction d_k est :

$$X^{(k+1)} = X^{(k)} + \frac{d_k^T\, g_k}{d_k^T\, A\, d_k}\, d_k \quad \text{avec} \quad g_k = \nabla f(X^{(k)}) \quad \text{et} \quad d_k^T = \nabla f(X^{(k)})^T \tag{49}$$

Algorithm 4 Algorithme de descente de gradient à pas optimal

Input : $f(X) \in \mathcal{C}^1$, $X^{(0)} \in \mathbb{R}^n$
$k \longleftarrow 0$

1. **Tant que** critère d'arrêt n'est pas vérifié faire :
   2. Direction de descente $d_k \longleftarrow -\nabla f(X^{(k)})$
   3. Trouver un pas tel que $\lambda_k \longleftarrow \min\limits_{\lambda_k \geq 0} f(X^{(k)} - \lambda \nabla f(X^{(k)}))$
   4. Nouvelle itération $X^{(k+1)} \longleftarrow X^{(k)} + \lambda_k\, d_k$
   5. Mise à jour : $k \longleftarrow k + 1$
6. **Fin**

Exercice ❺ ☞ ®

Illustrons la méthode de gradient à pas optimal pour la fonction-objectif de deux variables suivante :

$$f(X) = 4\,(x^2 + y^2) - 2\,x\,y - 6\,(x + y) \quad \text{avec} \quad X = (x, y)^T \tag{50}$$

Le gradient s'écrit :

$$\nabla f(X) = \begin{pmatrix} 8\,x - 2\,y - 6 \\ -2\,x + 8\,y - 6 \end{pmatrix} \tag{51}$$

Et le Hessien,

$$H(f) = \nabla^2 f(X) = \begin{pmatrix} 8 & -2 \\ -2 & 8 \end{pmatrix} \tag{52}$$

La Matrice Hessienne est symétrique définie positive. La fonction $f(X)$ est strictement convexe et unimodale. Désormais notons $d_k = (d_1, d_2)^T$ la direction de gradient, soit :

$$\nabla f(X) = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \begin{pmatrix} 8\,x - 2\,y - 6 \\ -2\,x + 8\,y - 6 \end{pmatrix} \tag{53}$$

Nous avons aussi :

$$\varphi(X + \lambda_k \nabla f(X)) = 4\,(x + \lambda_k\,d_1)^2 + 4\,(y + \lambda_k\,d_2)^2 - 2\,(x + \lambda_k\,d_1)\,(y + \lambda_k\,d_2) - 6\,(x + \lambda_k\,d_1 + y + \lambda_k\,d_2)$$

calculons la dérivée :

$$\begin{aligned} &\varphi^{'}(\lambda_k) = 0 \\ &\varphi^{'}(\lambda_k) = 8\,(x + \lambda_k\,d_1)\,d_1 + 8\,(y + \lambda_k\,d_2)\,d_2 - 2\,(x + \lambda_k\,d_1)\,d_2 - 2\,(y + \lambda_k\,d_2)\,d_1 - 6\,(d_1 + d_2) = 0 \\ &\Longrightarrow \lambda_k = \frac{6\,(d_1 + d_2) + 2\,(x\,d_2 + y\,d_1) - 8\,(x\,d_1 + y\,d_2)}{8\,(d_1^2 + d_2^2) - 4\,d_1\,d_2} \end{aligned}$$

Il en résulte le schéma numérique de l'algorithme de descente de gradient à pas optimal :

$$(x^{(k+1)}, y^{(k+1)}) = (x^{(k)}, y^{(k)}) - \frac{6\,(d_1 + d_2) + 2\,(x\,d_2 + y\,d_1) - 8\,(x\,d_1 + y\,d_2)}{8\,(d_1^2 + d_2^2) - 4\,d_1\,d_2}\,(\underbrace{8\,x^{(k)} - 2\,y^{(k)} - 6}_{d_1}, \underbrace{-2\,x^{(k)} + 8\,y^{(k)} - 6}_{d_2})$$

Script Matlab® :

```
clc ; clear all ; close all ;

% Le 08/08/2019 - Samir KENOUCHE
% METHODE DE GRADIENT A PAS OPTIMAL

 fxy = @(x,y) 4*(x.^2 + y.^2) - 2*x*y - 6*(x+y) ;  itmax = 5 ; it = 0 ;
 dfx = @(x,y) 8*x - 2*y - 6 ; dfy = @(x,y) -2*x + 8*y - 6 ;

while it < itmax

   xinit = [.3 .4] ;
   df = - [dfx(xinit(1),xinit(2)) dfy(xinit(1),xinit(2))] ;

   lambdak = (6*(df(1) + df(2)) + 2*(xinit(1)*df(2) + xinit(2)*df(1)) - ...
         8*(xinit(1)*df(1)+ xinit(2)*df(2)))/(8*(df(1).^2 + df(2).^2) - ...
         4*df(1)*df(2)) ;      % PAS OPTIMAL EXACTE

   xk = xinit + lambdak*df     % NOUVEL ITERE

   xinit = xk ; it = it + 1 ;  % MISE A JOUR

end
```

```
f_min = fxy(xk(1),xk(2)) ; % MIN DE LA FONCTION-OBJECTIF

%%%%%%%%%%%%%%%%%%%%%%%%%% AFFICHAGE GRAPHIQUE %%%%%%%%%%%%%%%%%%%%%%%%%%
x = -10:0.3: 8 ; y = -10:0.3:8 ; [xgrid, ygrid] = meshgrid(x,y) ;
zgrid = fxy(xgrid,ygrid) ; figure('color',[1 1 1]) ;
contourf(xgrid,ygrid,zgrid) ; xlabel('x') ; ylabel('y') ; hold on ;
plot(xk(1),xk(2),'rx') ; plot(xk(1),xk(2),'ro') ;
```

Exercice ❻ ☞ Ⓡ

Nous appliquerons cette méthode pour chercher le minimum de la fonction :

$$\begin{cases} f(x,y) = \dfrac{1}{2}\,x^2 + \dfrac{7}{2}\,y^2 \\ \text{Avec} \quad x_0 = (7.5, 2.2) \end{cases} \tag{54}$$

Avec un raisonnement analogue que précédemment, nous obtenons l'expression analytique du pas optimal

$$\lambda_k = \frac{-(x^{(k)}\,d_1 + 7\,y^{(k)}\,d_2)}{d_1^2 + 7\,d_2^2} \tag{55}$$

Il en résulte le schéma numérique :

$$(x^{(k+1)}, y^{(k+1)}) = (x^{(k)}, y^{(k)}) - \frac{(x^{(k)}\,d_1 + 7\,y^{(k)}\,d_2)}{d_1^2 + 7\,d_2^2}\,(\underbrace{x^{(k)}}_{d_1}, \underbrace{7\,y^{(k)}}_{d_2}) \tag{56}$$

Script Matlab® :

```
clear all ; close all ; clc ;

% Le 09/08/2019 - Samir KENOUCHE
% GRADIENT A PAS OPTIMAL

fxy = @(x,y) x.^2*(1/2) + y.^2*(7/2) ; dfx = @(x) x ; dfy = @(y) 7.*y ;
it = 0 ; itmax = 40 ; echelle = 0.6 ;  tol =  1e-6 ; xinit = [7.5 2.2] ;

x = -2:0.2: 8 ; y = -6:0.2:6 ; [xgrid, ygrid] = meshgrid(x,y) ;
zgrid =  fxy(xgrid, ygrid) ; figure('color',[1 1 1]) ;
contourf(xgrid,ygrid,zgrid) ; hold on ; xlabel('x') ; ylabel('y') ;

 while it < itmax

   dk =  - [dfx(xinit(1)) dfy(xinit(2))] ;
   lambdak = -(xinit(1)*dk(1) + 7*xinit(2)*dk(2))/(dk(1).^2 + 7*dk(2).^2) ;

   xk = xinit + lambdak*dk ;

  %%%%%%%%%%%%%%%%%%%%%%%%%%%% TEST D'ARRET %%%%%%%%%%%%%%%%%%%%%%%%%%%%
      if norm(lambdak*dk) <= tol
         solution = xk ;           % SOLUTION OBTENUE
         nbre_it = it ;            % NOMBRE D'ITERATION PERMETTANT LA CV
```

```
        break
      end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

 plot(xk(1), xk(2),'xw','MarkerSize',7) ; hold on ;
 plot(xk(1), xk(2),'ow','MarkerSize',7) ; hold on ;
 quiver(xk(1), xk(2), -dfx(xk(1)),- dfy(xk(2)),...
     echelle,'Color', 'r', 'LineWidth',1) ;

 xinit = xk ;  it = it + 1 ;      % MISE A JOUR
end

f_min = fxy(solution(1),solution(2)) ; % MIN DE LA FONCTION-OBJECTIF
```

L'affichage graphique généré par ce script est :

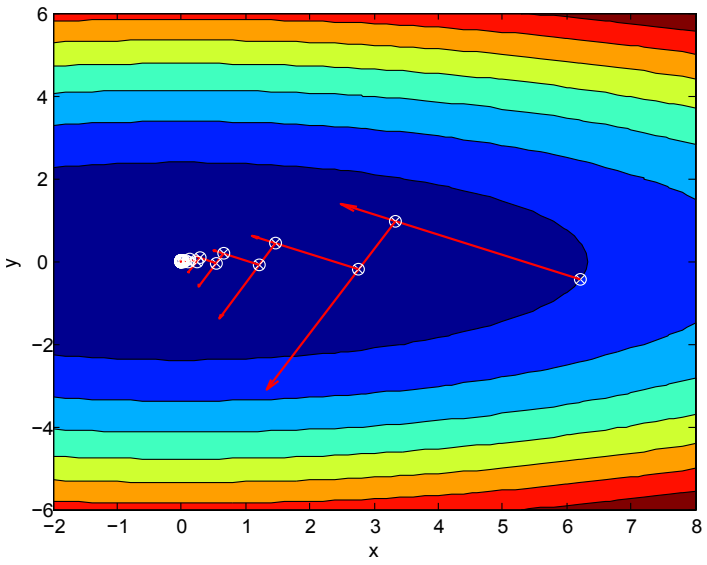


FIGURE 3: Minimisation par la méthode de la *descente de gradient à pas optimal*

Dans le direction $\overrightarrow{d_k}$, à l'itération $x^{(k+1)}$ l'algorithme calcule le minimum de $\nabla f(x^{(k+1)})$, soit

$$\varphi(\lambda) = d_k \nabla f(x^{(k+1)}) \tag{57}$$

La recherche du minimum impose de calculer la dérivée :

$$\varphi^{'}(\lambda) = 0 \Leftrightarrow < d_k | \nabla f(x^{(k+1)}) >= 0 \Leftrightarrow - < \nabla f(x^{(k)}) | \nabla f(x^{(k+1)}) >= 0 \tag{58}$$

Ainsi, le produit scalaire des deux gradients est nul, par conséquent les directions de descente successives calculées par l'algorithme sont orthogonales. Ceci explique pourquoi la convergence suit une trajectoire en zigzag à angles droits. L'avantage d'optimiser le pas de descente de gradient rend l'algorithme moins sensible, par rapport au pas fixe, au choix de la valeur initiale. La vitesse de convergence est améliorée également. Néanmoins, cette méthode peut se révéler moins efficace dans le cas où la fonction est caractérisée par des pentes peu marquées. L'autre inconvénient tient au fait qu'il est difficile, voir impossible, de trouver une expression analytique $\lambda_k = f(x_k)$ pour des fonctions plus complexes. Chaque méthode est adaptée à un problème spécifique.

C. *Méthode de gradient conjugué*

Cette méthode procède par la détermination successive de directions de recherche et de longueurs de descente. Contrairement aux autres méthodes de descente qui rendent compte uniquement du comportement local de la fonction-objectif qui se matérialise par une structure en zigzag. La méthodes du gradient conjugué permet de s'affranchir de cette structure en déterminant des directions de recherche différentes des directions précédentes.

Définition: Soit une matrice $A \in \mathbb{R}^{n\times n}$ définie positive. Deux directions d_{k+1} et d_k de $\mathbb{R}^n$ sont conjuguées par rapport à la matrice A si

$$d_{k+1}^T\, A\, d_k = 0 \qquad \forall k \tag{59}$$

En outre, si $A = I$ (matrice identité), les directions conjuguées d_{k+1} et d_k sont orthogonales. Nous cherchons ainsi d_{k+1} dans le plan formé par les directions orthogonales d_k et g_{k+1} soit :

$$d_{k+1} = -g_{k+1} + \beta_k\, d_k \qquad \beta_k \in \mathbb{R} \quad et \quad g_{k+1}, d_k \in \mathbb{R}^n \tag{60}$$

Cherchons l'expression de β_k en combinant les relations (59) et (60), soit :

$$(-g_{k+1} + \beta_k\, d_k)^T\, A\, dk = 0 \Rightarrow -g_{k+1}^T\, A\, d_k + \beta_k\, d_k^T\, A\, d_k = 0$$

$$\Rightarrow \beta_k = \frac{g_{k+1}^T\, A\, d_k}{d_k^T\, A\, d_k} \tag{61}$$

L'expression du pas optimal a déjà été déterminée précédemment (48) :

$$\lambda_k = \frac{\nabla f(X^{(k)})^T\, \nabla f(X^{(k)})}{\nabla f(X^{(k)})^T\, A\, \nabla f(X^{(k)})} \tag{62}$$

Algorithm 5 Algorithme de descente de gradient conjugué

Input : $f(X) \in \mathcal{C}^1$, $X^{(0)} \in \mathbb{R}^n$, $\nabla f(X^{(0)}) \in \mathbb{R}^n$, A tel que $\forall M \in \mathbb{R}^n, M \neq 0 : M^T\, A\, M > 0$

$k \longleftarrow 0$

1. **Tant que** critère d'arrêt n'est pas vérifié faire :
   2. Pas optimal $\lambda_k = \dfrac{d_k^T\, d_k}{d_k^T\, A\, d_k}$
   3. Nouvelle itération $X^{(k+1)} \longleftarrow X^{(k)} + \lambda_k\, d_k$
   4. Calcul du scalaire $\beta_k = \dfrac{g_{k+1}^T\, A\, d_k}{d_k^T\, A\, d_k}$
   5. Direction conjuguée $d_{k+1} = -g_{k+1} + \beta_k\, d_k$
   6. Mise à jour : $k \longleftarrow k+1$
7. **Fin**

Soulignons qu'il existe plusieurs méthodes pour calculer le paramètre β_k. Dans l'équation (61), nous avons utilisé la méthode de *Fletcher-Reeves*. Nous citerons également les méthodes de *Polack-Ribière* et de *Hestenes-Stiefel*.

Exercice ❼ ☞ ®

Illustrons la méthode de gradient conjugué pour la fonction-objectif de deux variables suivante :

$$f(X) = 5\, x^2 + \frac{y^2}{2} - 3\,(x+y) \quad \text{avec} \quad X = (x,y)^T \tag{63}$$

Prendre comme vecteur initial $X^{(0)} = [-2, 1]^T$ et une tolérance $\epsilon = 10^{-4}$. Le critère d'arrêt est :

$$\| \frac{f(X^{(k+1)}) - f(X^{(k)})}{f(X^{(k)})} \| < \epsilon$$

Script Matlab® :

```
clc ; clear all ; close all ;

% Le 10/08/2019 - Samir KENOUCHE
% METHODE DU GRADIENT CONJUGUE

 fxy = @(x,y) 5*x.^2 + y.^2*(1/2) - 3*(x+y) ;  itmax = 15 ; it = 0;
 dfx = @(x) 10*x - 3 ; dfy = @(y) y - 3 ; tol = 1e-04 ;

 A = [10 0 ; 0 1] ;  b = [3 ; 3] ; xinit = [-2 ; 1] ;
   dkinit =  - [dfx(xinit(1)) ; dfy(xinit(2))] ; xtest = zeros(2,itmax);

while it < itmax

   lambdak = (dkinit'*dkinit)/(dkinit'*A*dkinit) ;

   xmin = xinit + lambdak*dkinit ;

   if abs((fxy(xmin(1), xmin(2)) - fxy(xinit(1),...
           xinit(2)))/fxy(xinit(1), xinit(2))) < tol ;

         solution = xmin   ; % VECTEUR SOLUTION
         nbr_it = it ;       % NOMBRE D'ITERATION PERMETTANT LA CV
       break
   end

   betak = ([dfx(xmin(1)) ; dfy(xmin(2))]'*A*dkinit)/(dkinit'*A*dkinit)  ;
   dk =  - [dfx(xmin(1)) ; dfy(xmin(2))] + betak*dkinit
   xinit = xmin ; dkinit = dk ; it = it + 1 ; % MISE A JOUR
end

f_min = fxy(solution(1),solution(2)) ; % MIN DE LA FONCTION-OBJECTIF

%%%%%%%%%%%%%%%%%%%%%%%%%% AFFICHAGE GRAPHIQUE %%%%%%%%%%%%%%%%%%%%%%%%%%
x = -20:0.3: 18 ; y = -20:0.3:18 ; [xgrid, ygrid] = meshgrid(x,y) ;
zgrid = fxy(xgrid,ygrid) ; figure('color',[1 1 1]) ;
contourf(xgrid,ygrid,zgrid) ;  xlabel('x') ; ylabel('y') ; hold on ;
plot(solution(1),solution(2),'rx') ; plot(solution(1),solution(2),'ro') ;

% VERIFICATION : A * solution - b = 0 Avec XMIN = (0.30 ; 3.0)
% ATTENTION AUX ERREURS D'ARRONDI
```

La méthode de gradient conjugué appliquée à une fonction-objectif d'ordre n converge au plus en n itérations. Notons toutefois que les erreurs d'arrondi dans le calcul des directions conjuguées font que nous n'obtenons pas toujours la solution exacte en n itérations. Cette méthode réclame un nombre d'itérations proportionnel

à $\sqrt{\text{cond}(A)}$. Contrairement aux autres méthodes de gradient, exigeant un nombre d'itérations proportionnel à $\text{cond}(A)$.

IV. Algorithmes génétiques

Cette classe[4] d'algorithmes est dite évolutionniste, car fondée sur les mécanismes Darwiniens de la sélection naturelle. Le fonctionnement de ces algorithmes est comme suit : partant d'une population initiale d'individus, choisis aléatoirement, la performance relative de chaque individu est mesurée au moyen d'une fonction *fitness*. Les valeurs de cette fonction sont classées selon un ordre décroissant. Nous créons ainsi une nouvelle population qui a hérité du bagage génétique des meilleurs individus de la population initiale. Cette nouvelle population est crée en se servant des opérations : *sélection*, *croisement* et *mutation*. Ce processus est réitéré jusqu'à la sélection du meilleur individu, constituant la solution du problème d'optimisation.

Avant de décrire les différents mécanismes du fonctionnement de ces algorithmes, nous donnerons, chemin faisant, quelques définitions élémentaires. Un individu I (ou chromosome) est codé sous forme d'une chaine binaire constituée de m gènes, $I = \{g_1, g_2, g_3, \cdots, g_m\}$ tel que $\forall i \in [1, m], g_i \in V = \{0, 1\}$. Important ! il existe également le codage réel tel que $g_i \in \mathbb{R}$. Toutefois, dans cette section nous nous focaliserons exclusivement sur le codage binaire. Pour une population constituée de n individus $\mathcal{P} = \{I_1, I_2, I_3, \cdots, I_n\}$ et chaque individu est codé en m gènes, nous obtenons :

$$\mathcal{P} = \begin{cases} I_1 = (g_1^1, g_2^1, g_3^1, \cdots, g_m^1\} \\ I_2 = (g_1^2, g_2^2, g_3^2, \cdots, g_m^2\} \\ I_3 = (g_1^3, g_2^3, g_3^3, \cdots, g_m^3\} \\ \vdots \\ I_n = (g_1^n, g_2^n, g_3^n, \cdots, g_m^n\} \end{cases} \tag{64}$$

Chaque individu I_i (appelé aussi chromosome ou encore génome) de la population $\mathcal{P}$ est formé de m gènes, choisis de manière totalement aléatoire. Dans le cas du codage binaire, la fonction *fitness* (ou efficacité) $f \in \mathbb{R}_+^n$. Cette fonction sert à classer les individus en fonction de leur performance. Elle peut être vue également comme une mesure d'adaptation des individus à leur environnement. En pratique, on utilise une fonction de décodage δ permettant le passage du système binaire au système décimal, soit :

$$\delta : \{0, 1\}^m \longmapsto \mathbb{R}$$
$$f : \delta\,\{0, 1\}^m \longmapsto \mathbb{R}_+^*$$

Dans sa formulation la plus simple, la fonction δ est définie selon :

$$\delta(I_i) = \sum_{j=1}^{m} g_j\, 2^{m-j} \quad \text{tel que} \quad \delta(I_i) \in \mathbb{R}_+^* \tag{65}$$

Plus de détails sur la conversion des systèmes de numération sont donnés en annexe (B). Dans ce qui suit, nous décrirons succinctement les trois mécanismes de base de la sélection naturelle. Ensuite, nous prendrons un exemple d'application numérique afin de faciliter la compréhension des différentes notions inhérentes aux algorithmes génétiques.

4. Les algorithmes génétiques font partie des méthodes dites de recherche globale : cela signifie la possibilité d'atteindre un optimum, ou du moins un meilleur point, même si ce dernier n'est pas dans le voisinage immédiat de l'itéré précédent. Toutes les autres méthodes d'optimisation vues précédemment sont dites de recherche locale : Le nouvel itéré est recherché dans le voisinage immédiat (d'où la notion de localité) de l'itéré précédent.

a) ***Sélection****:* Le processus de sélection naturelle permet aux gènes les mieux adaptés de se reproduire plus souvent et de contribuer davantage aux générations futures. Dans une population $\mathcal{P} = \{I_1, I_2, I_3, \cdots, I_n\}$, chaque individu est évalué par la fonction *fitness* $R_w^i = f(\delta(I_i))$. En pratique, nous associons à chaque individu, une probabilité P_s^i, d'être sélectionné.

$$P_s^i = \frac{R_w^i}{\sum\limits_{i \in \mathcal{P}} R_w^i} \tag{66}$$

Il existe une panoplie de processus de sélection. Nous avons pris la méthode la plus intuitive. Autrement dit, la probabilité de reproduction d'un individu dépend de sa valeur au regard de l'ensemble des valeurs de la population.

b) ***Croisement****:* Nous vérifions d'abord s'il y a croisement (ou reproduction) selon une probabilité de croisement typiquement $P_c \in V(0.85, 1)$. Il existe plusieurs types de croisement, nous considérons ici le croisement dit arithmétique. Soient une population de deux individus $\{I_1, I_2\}$ et un nombre aléatoire $\alpha \in V(0, 1)$. S'il y a croisement les rejetons $r_i = \{r_1, r_2, r_3\}$ (enfants) de la nouvelle génération s'obtiennent selon :

$$r_i = \begin{cases} r_1 = \alpha\, I_1 + \alpha\, I_2 \\ r_2 = (1-\alpha)\, I_1 + \alpha\, I_2 \\ r_3 = \alpha\, I_1 + (1-\alpha)\, I_2 \end{cases} \tag{67}$$

L'opération de croisement atteint chaque gène de chaque individu, selon le processus ci-dessous.

$$r_1 = \begin{pmatrix} \alpha\, g_1^1 + \alpha\, g_1^2 \\ \alpha\, g_2^1 + \alpha\, g_2^2 \\ \alpha\, g_3^1 + \alpha\, g_3^2 \\ \vdots \\ \alpha\, g_m^1 + \alpha\, g_m^2 \end{pmatrix} \tag{68}$$

$$r_2 = \begin{pmatrix} (1-\alpha)\, g_1^1 + \alpha\, g_1^2 \\ (1-\alpha)\, g_2^1 + \alpha\, g_2^2 \\ (1-\alpha)\, g_3^1 + \alpha\, g_3^2 \\ \vdots \\ (1-\alpha)\, g_m^1 + \alpha\, g_m^2 \end{pmatrix} \tag{69}$$

$$r_3 = \begin{pmatrix} \alpha\, g_1^1 + (1-\alpha)\, g_1^2 \\ \alpha\, g_2^1 + (1-\alpha)\, g_2^2 \\ \alpha\, g_3^1 + (1-\alpha)\, g_3^2 \\ \vdots \\ \alpha\, g_m^1 + (1-\alpha)\, g_m^2 \end{pmatrix} \tag{70}$$

Les individus de la nouvelle génération $\{r_1, r_2, r_3\}$ ont hérité en théorie les meilleurs gènes des individus de la génération précédente.

*c) **Mutation**:* Nous vérifions d'abord s'il y a mutation selon une probabilité de mutation typiquement $P_m \in V(10^{-4}, 10^{-1})$. Pour un gène ayant une probabilité de mutation P_m, nous obtenons de façon totalement aléatoire :

$$g_i^{'} = \begin{cases} g_i + \psi[\max(g_i) - g_i] & \text{1ère possibilité} \\ \\ g_i - \psi[\max(g_i) - g_i] & \text{2ème possibilité} \end{cases} \tag{71}$$

La fonction ψ est définie selon l'expression :

$$\psi(x) = x\,\beta\left(\frac{gT - gt}{gT}\right)^b \tag{72}$$

Avec β est un nombre aléatoire $\in V(0,1)$, gt est la génération courante, gT est la génération maximale et b est le degré d'extinction.

Exercice ❽ ☞ Ⓡ

En guise d'application numérique, soit à maximiser la fonction-objectif à une seule variable :

$$x^* \in \underset{x\in\mathbb{R}^n}{\text{argmax}}\ 17\,(x - x^2) \tag{73}$$

Si l'on souhaite la minimiser on prendra tout simplement :

$$x^* \in \underset{x\in\mathbb{R}^n}{\text{argmin}}\ -\ 17\,(x - x^2) \tag{74}$$

Prendre une probabilité de croisement $P_c = 0.75$ et une probabilité de mutation $P_m = 0.001$. La population initiale est formée de six individus. Ces derniers sont constitués de quatre gènes. Cet exemple simple vise à illustrer concrètement le fonctionnement des opérations de sélection, de croisement, de mutation et l'utilisation des différents paramètres des algorithmes génétiques permettant d'atteindre l'optimum. Cet exercice sera intégralement résolu lors de la séance de cours.

Exercice ❾ ☞ Ⓡ

Pour des problèmes d'optimisation plus complexes, nous utiliserons la fonction Matlab® prédéfinie `ga`. En guise d'application de cette fonction, nous souhaitons optimiser la fonction-objectif précédente (63) dont nous rappelons la formule :

$$f(X) = 5\,x^2 + \frac{y^2}{2} - 3\,(x + y) \quad \text{avec} \quad X = (x, y)^T$$

La syntaxe de la fonction Matlab® prédéfinie `ga` est donnée dans le script Matlab® ci-dessous :

```
clc ; clear all ;

% Samir KENOUCHE - Unconstrained optomization using algorithm genetic
% Le 14/09/2019

addpath('C:\Users\kenouche') ;

myfitness = @(x) 5*x(1).^2 + x(2).^2*(1/2) - 3*(x(1)+x(2)) ;

cineq = [] ; ceq = [] ;   % unconstrained problem
```

```
rng(1037,'twister') ;      % Control random number generation
xinit = randn([1 2]) ;     % Initialisation
nbrevars = 2 ;             % Number of variables
LB = [-inf -inf] ;         % Lower bound
UB = [+inf +inf] ;         % Upper bound

options = gaoptimset('CreationFcn', @gacreationlinearfeasible, ...
    'PlotFcns', @gaplotbestf, 'CrossoverFraction', 0.75, ...
    'InitialPopulation',xinit,'MutationFcn',@mutationadaptfeasible, ...
    'EliteCount', 2,'PopulationSize',50, 'TolFun', 1e-03,'Display','iter') ;

[xoptimal,fval,exitflag,output,population,scores] = ga(myfitness, ...
    nbrevars,[],[],[],[],LB,UB, cineq, ceq, options) ;

disp(output.message) ;
str = ['LES POINTS STATIONNAIRES OBTENUS : ' num2str(xoptimal)]
% LES POINTS STATIONNAIRES OBTENUS : 0.3  3
```

En conclusion, nous soulignons que les trois processus d'un algorithme génétique sont régit par un certain nombre de paramètres fixés préalablement. Ces derniers conditionnent la réussite de l'optimisation du problème étudié. Parmi ces paramètres, nous citerons (1) La taille de la population $\mathcal{P}$, et la taille de la chaine de codage des individus. Si cette population est trop grande, le temps de calcul de l'algorithme peut se révéler très contraignant. En revanche, si la taille de la population est petite, l'algorithme convergera vraisemblablement vers un mauvais individu. (2) Le choix de la valeur de la probabilité de croisement P_c, celle-ci dépend de la forme de la fonction *fitness*. En pratique, le choix de cette probabilité se fait de façon heuristique. Plus P_c est élevée, plus la population subit des transformations profondes. (3) Le choix de la valeur de la probabilité de mutation P_m. Elle est généralement faible puisqu'une valeur élevée risque de conduire à une solution divergente.

Une dernière notion inhérente aux algorithmes génétiques est celle de *l'élitisme*. Cette opération consiste à conserver le meilleur individu dans la génération ultérieure. Après le croisement, on compare le meilleur individu de la génération courante avec le meilleur individu de la génération antérieure. A l'issue de cette comparaison, le meilleur individu est conservé au détriment de l'autre qui est éliminé.

Exercice ⓫ ☞ Ⓡ

1) Minimiser en utilisant l'algorithme génétique la fonction-objectif définie par :

$$f(X) = 100\,(x_1^2 - x_2)^2 + (1 - x_1)^2$$

$$\hat{X} \in \underset{X\in\mathbb{R}^n}{\operatorname{argmin}} \quad f(X) \text{tel que} \begin{cases} x_1x_2 + x_1 - x_2 + 1.5 \leq 0 \\ 10 - x_1\,x_2 \leq 0 \\ 0 \leq x_1 \leq 1,\ 0 \leq x_2 \leq 13 \end{cases}$$

D'abord nous commençons par écrire les contraintes dans un fichier `M-file`, qu'on appellera ensuite depuis le script principal.

```
function [cineq, ceq] = myconstraint(x)
  cineq = [1.5 + x(1)*x(2) + x(1) - x(2) ;
  -x(1)*x(2) + 10] ;
  ceq = [] ;
return
```

Ce fichier est sauvegardé obligatoirement sous le nom `myconstraint.m` et doit être dans le même répertoire que le script principal ci-dessous :

```
clc ; clear all ;

% Samir KENOUCHE - Constrained optomization using algorithm genetic
% Le 15/09/2019
addpath('C:\Users\kenouche') ;

myfitness = @(x) 100*(x(1)^2 - x(2))^2 + (1 - x(1))^2 ;

rng(1403,'twister') ;   % Control random number generation
xinit = randn([1 2]) ;  % initialisation
nbrevars = 2 ;          % Number of variables
LB = [0 0] ;            % Lower bound
UB = [1 13] ;           % Upper bound

options = gaoptimset('CreationFcn', @gacreationlinearfeasible, ...
    'PlotFcns', @gaplotbestf, 'CrossoverFraction', 0.75, ...
    'InitialPopulation',xinit,'MutationFcn',@mutationadaptfeasible, ...
    'EliteCount', 1,'PopulationSize',50, 'TolFun', 1e-03, ...
    'Display','iter') ;

[xopti,fval,exitflag,output,population,scores] = ga(myfitness, nbrevars, ...
    [],[],[],[],LB,UB, @myconstraint, options) ;

disp(output.message)
str = ['Les points statonnaires obtenus : ' num2str(xopti)]
%% Les points stationnaires obtenus : 0.812202  12.3122
```

V. Travaux pratiques avec des fonctions Matlab prédéfinies

A. *Optimisation sans contraintes*

La formulation mathématique générale d'un problème d'optimisation sans contraintes s'écrit selon :

$$\hat{X} = \underset{X \in \mathcal{D}}{argmin}\, f(X) \tag{75}$$

Avec, $\mathcal{D}$ est un sous-ensemble de $\mathbb{R}^n$. Les variables $X = (x_1, x_2, ..., x_n)^T$ sont appelées *variables d'optimisation* ou *variables de décision*. La fonction f, à valeurs réelles, définie par $f : \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}$ est la *fonction-objectif* ou *fonction de coût*. Dans cette section, on abordera les fonctions Matlab® prédéfinies, dédiées à la résolution de problèmes d'optimisation sans contraintes. Toutes ces fonctions sont disponibles dans la boite à outil *Optimization Toolbox* du logiciel. Les fonctions Matlab® prédéfinies destinées à la minimisation de fonctions d'une seule variable sont `fminbnd` et `fminunc`. Notons que cette dernière peut également être utilisée pour les fonctions de plusieurs variables. Ces commandes présentent une syntaxe très similaires.

```
options = optimset('param 1', value 1, 'param 2', value 2, ...)
[x, fval, exitflag, output, grad, hessian]= fminunc(fun, x0, options) % pour fminunc
[x, fval, exitflag, output] = fminbnd(fun ,lB, uB, options)           % pour fminbnd
```

La fonction `fminunc` accepte comme arguments en entrée, la fonction à minimiser `fun`, les valeurs initiales `x0` pour initialiser la recherche des minimums et en dernier lieu l'argument `options` spécifiant les différents champs d'optimisation. Ces derniers sont modifiés en appelant la fonction `optimset`, dont ses différents paramètres seront décrits dans l'exercice ci-dessous. L'argument `fun` peut être définie en tant que *objet inline*, *fonction anonyme* ou bien une *fonction M-file*. Les sorties renvoyées sont : `x` représentant le minimum trouvé après convergence et `fval` est le nombre d'évaluation de la fonction `fun`. Une valeur de la sortie `exitflag = 1` signifie que l'algorithme a bel et bien convergé vers la solution approchée. Plus généralement, une valeur de `exitflag > 1` signifie que l'algorithme a convergé vers la solution. Une valeur de `exitflag < 1` signifie que l'algorithme n'a pas convergé. Dans le cas où `exitflag = 0`, cela veut dire que le nombre d'itérations ou le nombre d'évaluation de la fonction est atteint. La sortie `output` renvoie des champs relatifs au type d'algorithme utilisé, le nombre d'itérations conduisant à la solution approchée, un message sur l'état de l'optimisation ... etc. Les sorties `grad` et `hessian` renvoient respectivement le *Jacobien* (première dérivée) et le *Hessien* (second dérivée) de la fonction à minimiser.

La différence entre les fonctions `fminunc` et `fminbnd` se situe au niveau de l'initialisation de la recherche de la solution. En effet, `fminunc` démarre la recherche à partir d'une valeur initiale apportée par `x0`. En revanche, `fminbnd` effectue sa recherche à partir d'un intervalle dont les bornes inférieure et supérieure sont indiquées respectivement par les entrées `lB` et `uB`.

Exercice ❶ ☞ ®

1) Minimiser la fonction-objectif définie par :

$$\begin{cases} f(x) = (x-1) \times \exp(-x^2 + 2\,x + 1) \\ x \in [-4, 4] \end{cases} \tag{76}$$

en utilisant les fonctions prédéfinies `fminbnd` et `fminunc`

Script Matlab®

```
clear all; clc ;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
lB = -4 ; uB = 4 ; xinit = 1 ;
fx = @(x) (x-1).*exp(-x.^2 + 2.*x + 1) ;

opts = optimset('Display','iter','FunValCheck','on','TolX',1e-8) ;
% parametres d'optimisation
[xMin1, funEval1, exitTest1, output1, grad1, hessian1] = fminunc(fx, xinit, opts) ;
% premiere possibilite
[xMin2, funEval2, exitTest2, output2] = fminbnd(fx, lB, uB, opts) ;
% Deuxieme posibilite
```

Les arguments de sortie renvoyés par `fminunc` sont :

```
%%%%%%%%%%%%%%%%%%%% AFFICHAGE PAR DEFAUT 1ER CAS %%%%%%%%%%%%%%%%%%%%
 Iteration  Func-count       f(x)
     0          2               0
     1          4         -2.71828
     2          6         -3.16544
     3          8         -3.16849
```

```
     4          10         -3.16903
     5          12         -3.16903
     6          14         -3.16903
Optimization terminated: relative infinity-norm of gradient less than options.TolFun.
%%%%%%%%%%%%%%%%%%% AFFICHAGE PAR DEFAUT 2EME CAS %%%%%%%%%%%%%%%%%%
 Func-count     x          f(x)
    1      -0.944272    -0.327815
    2       0.944272    -0.410501
    3        2.11146      2.38771
    4      0.0274024     -2.79063
    5     0.00805821     -2.74001
    6       0.252738     -3.15901
    7        0.51688     -2.82668
    8       0.278372     -3.16771
    9        0.29087     -3.16901
   10       0.293071     -3.16903
   11         0.2929     -3.16903
   12       0.292893     -3.16903
   13       0.292893     -3.16903
   14       0.292893     -3.16903
   15       0.292893     -3.16903

Optimization terminated:
 the current x satisfies the termination criteria using OPTIONS.TolX of 1.000000e-08
```

Les autres sorties sont affichées comme suit :

```
>> xMin1 =
            0.2929
             % MINIMUM OBTENU

>> funEval1 =
            -3.1690
             % VALEUR DE LA FONCTION A LA DERNIERE ITERATION

>> exitTest1 =
             1
             % TEST DE CONVERGENCE POSITIF

>> grad1 =

         -5.9605e-08
             % VALEUR DU GRADIENT DE LA FONCTION-OBJECTIF

>> hessian1 =

             12.6783
             % VALEUR DU HESSIEN DE LA FONCTION-OBJECTIF

>> xMin2 =
             0.2929
```

```
              % MINIMUM OBTENU AVEC fminbnd

>> funEval2 =
             -3.1690

>> exitTest2 =
              1
```

L'argument `opts`, de type *srtucture*, compte les options d'optimisation spécifiées dans `optimset`. Le champ de la structure `opts` indiqué par `optimset('Display','iter', ...)` affiche des détails pour chaque itération. Si l'on désire afficher uniquement les détails de la dernière itération, on remplacera `'iter'` par `'final'`. Dans le cas où on ne veut afficher aucun détails, on mettra la valeur `'off'`. Le champ indiqué par `optimset(...., 'FunValCheck','on',...)` contrôle si les valeurs de la fonction sont réelles et affiche dans le cas contraire un avertissement quand la fonction en question renvoie une valeur complexe ou NaN. On peut suspendre cette vérification, en remplaçant la valeur `'on'` par `'off'`. Le champ d'optimisation `optimset(...., 'TolX', 1e-08)` correspond à la tolérance admise pour la solution approchée. D'autres champs d'optimisation existent comme `MaxFunEvals` qui fixe le nombre maximum d'évaluation de la fonction à optimiser et `MaxIter` fixant également le nombre maximum d'itération. Voir aussi `GradObj`, `OutputFcn`, `PlotFcns`, ... etc dont la description est disponible dans le *help* de Matlab®.

La sortie `xMin1 = 0.2929` est le minimum trouvé. Ce dernier est cherché autour de la valeur initiale `xinit`. La sortie `funEval1 = -3.1690` exprime l'évaluation de la fonction à la dernière itération, c'est-à-dire pour `fx(x = xMin1)`. L'argument `exitTest = 1` signifie que l'algorithme a convergé vers la solution, une valeur négative indiquerai le contraire. L'argument de sortie `output1`, de type *structure*, renvoie les champs suivants :

```
>> output1 =                          % pour la commande fminunc
            iterations: 6
             funcCount: 14
              stepsize: 1
         firstorderopt: 5.9605e-08
             algorithm: [1x38 char]  % 'Quasi-Newton line search'
               message: [1x85 char]  %  Optimization terminated ...

>> output2                            %  pour la commande fminbnd
            iterations: 14
             funcCount: 15
             algorithm: [1x46 char]  % 'golden section search'
               message: [1x111 char] % Optimization terminated ...
```

La fonction a été évaluée 14 fois et l'algorithme converge vers la solution approchée au bout de la $6^{\text{ième}}$ itération. La sortie `stepsize: 1` indique le pas final de l'algorithme moyenne dimension de `Quasi-Newton`. Le mot `Optimization` affiché comme `message`, fait référence au faite que l'algorithme de minimisation fonctionne selon le critère *des moindres carrés*. `fminbnd` présente les mêmes propriétés que celles de `fminunc`, à la différence que `fminbnd` cherche le minimum dans un intervalle, donné en argument d'entrée avec la borne inférieure `lB` (lower Bound) et la borne supérieure `uB` (upper Bound).

Nous allons désormais tester la fonction `fminsearch` qui s'utilise pour la minimisation de fonctions multi-dimensionnelles. Sa syntaxe usuelle est analogue à celle de `fminunc`, à la différence près que `fminsearch` ne renvoie ni le *Jacobien* ni le *Hessien* de la fonction à optimiser. Ceci provient du fait que cette fonction est

basée sur l'algorithme *Simplex*. Nous allons procéder à son implémentation dans l'exercice ci-dessous.

Exercice ❷ ☞ ®

1) Minimiser la fonction-objectif de deux variables définie par :

$$f(x_1, x_2) = x_1^2 + (x_2 - 2)^2 \tag{77}$$

analytiquement puis en utilisant la fonction prédéfinie `fminsearch`

Commençons par déterminer, analytiquement, les extremums de la fonction $f(x_1, x_2)$. Le gradient de la fonction s'écrit :

$$\begin{cases} \dfrac{\partial f}{\partial x_1} = 2\,x_1 = 0 \\ \dfrac{\partial f}{\partial x_2} = 2\,(x_2 - 2) = 0 \end{cases} \tag{78}$$

Le vecteur du point critique est donné donc par $\hat{X} = (\hat{x}_1 = 0\,;\, \hat{x}_2 = 2)$. Cherchons désormais la nature de ce point, s'agit-il d'un minimum ou d'un maximum ?. Calculons le déterminant du *Hessien* de f.

$$\begin{vmatrix} \dfrac{\partial^2 f(\hat{X})}{\partial x_1^2} & \dfrac{\partial^2 f(\hat{X})}{\partial x_1\,\partial x_2} \\ \dfrac{\partial^2 f(\hat{X})}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 f(\hat{X})}{\partial x_2^2} \end{vmatrix} \Longrightarrow \begin{vmatrix} 2 & 0 \\ 0 & 2 \end{vmatrix} = 4 > 0 \tag{79}$$

À partir de ces résultats, il en découle que le point $\hat{X} = (\hat{x}_1 = 0\,;\, \hat{x}_2 = 2)$ est le minimum recherché. Nous résoudrons le même système de façon algorithmique en se servant de la fonction `fminsearch`. Voici le script Matlab®

```
clear all; clc ;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
xinit = [1 1] ; fun = @(x) x(1).^2 + (x(2) - 2).^2;
opts = optimset('Display','iter','FunValCheck','on','TolX',1e-8) ;

[xMin, funEval, exitTest, output] = fminsearch(fun,  xinit, opts) ;
```

Ci-dessous les différentes sorties renvoyées par le script.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%% SORTIE PAR DEFAUT %%%%%%%%%%%%%%%%%%%%%%%%%%%%
 Iteration   Func-count     Min f(x)
     0           1               2
     1           3          1.9025
     2           5         1.66563
     3           7         1.38266
     4           9        0.889414
     5          11        0.353447
     6          13       0.0265259
     7          14       0.0265259
     8          16       0.0265259
     9          18       0.0265259
```

```
    ...          ...        .........

     63          125       2.15781e-17
     64          127       2.15781e-17

Optimization terminated:
 the current x satisfies the termination criteria using OPTIONS.TolX of 1.000000e-08 and F(
   X) satisfies the convergence criteria using OPTIONS.TolFun of 1.000000e-04
```

```
>> xMin =

              0.0000    2.0000

>> funEval =

              2.1578e-17

>> exitTest =

              1
```

```
>> output

        iterations: 64
         funcCount: 127
         algorithm: [1x33 char] % 'Nelder-Mead simplex direct search'
           message: [1x194 char] % Optimization terminated: the current x satisfies the
   termination
```

L'algorithme converge vers la solution approchée `Xmin` au bout de 64 itérations. Notons que cette convergence est atteinte car les conditions d'arrêt de l'algorithme sont satisfaites, soit une tolérance `TolX = 1e-8`. En choisissant par exemple une tolérance de `TolX = 1e-3`, l'algorithme converge vers la solution approchée (`xMin = [-0.0002 ; 2.0004]`) au bout de 28 itérations seulement. Dans le cas où cette dernière n'est pas spécifiée, la valeur par défaut est `TolX = 1e-6`. On comprend alors que le choix de la tolérance est conditionné par la précision recherchée. Il est important de rappeler aussi que tous ces algorithmes d'optimisation sont basés sur des processus itératifs, donc fortement dépendant du choix de la valeur initiale. Autrement dit, plus la valeur initiale est proche de la solution approchée plus l'algorithme converge rapidement.

Comme il a été mentionné dans la section précédente, la fonction `fminunc` s'utilise aussi pour l'optimisation de fonctions à plusieurs variables. Nous allons l'utiliser pour optimiser la fonction $f(x_1, x_2) = 2\,x_1^2 + x_1\,x_2 + 2\,x_2^2 - 6\,x_1 - 5\,x_2 + 3$

Script Matlab®

```
clear all; clc ;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
xinit = [-1 1] ;
fx = @(x) 2*x(1).^2 + x(1)*x(2) + 2*x(2).^2 - 6*x(1) - 5*x(2) + 3 ;
```

```
opts = optimset('LargeScale','off','Display','iter','FunValCheck',...
    'on','MaxIter', 20,'TolX', 1e-5) ;
[x, funEval, exitTest, output, grad, hessian] = fminunc(fx, xinit, opts) ;
```

Il est possible de définir explicitement le *gradient* et le *Hessien* de la *fonction-objectif*. Le premier intérêt de cette procédure est de fournir les expressions analytiques, sans approximation, des dérivées. Si ces dernières ne sont pas explicitées, Matlab® calcule une approximation selon la méthode des *différences finies*. L'autre intérêt tient à l'accroissement, notamment pour des systèmes plus complexes, de la vitesse de convergence. Ainsi, le *gradient* et le *Hessien* sont indiqués via la syntaxe `options = optimset('GradObj','on','Hessian','on')`. On commence d'abord par définir la fonction *M-file* suivante.

```
function [fun, Jacobien, Hessien] = myfun(x)

fun = 2*x(1).^2 + x(1)*x(2) + 2*x(2).^2 - 6*x(1) - 5*x(2) + 3 ;

% Compute the objective function value at x
if nargout > 1
% fun called with two output arguments
grad(1) = 4*x(1) + x(2) - 6 ; % Gradient of the function evaluated at x
grad(2) = x(1) + 4*x(2) - 5 ;

Jacobien = grad ;
end

if nargout > 2

hessian(1,1) = 4 ;  hessian(2,1) = 1 ; % Hessian evaluated at x
hessian(2,1) = 1 ;  hessian(2,2) = 4 ;
Hessien = hessian ;

end
return
```

Cette fonction est sauvegardée sous le nom `myfun.m`. L'appel de cette dernière se fait avec le script suivant :

```
clear all; clc ;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
xinit = [1 1] ;
opts = optimset('GradObj','on','Hessian','on','Display','iter') ;
[x, fval, exitflag, output] = fminunc(@myfun, xinit, opts)
```

En écrivant `opts = optimset('GradObj','off','Hessian','off')`, Matlab® évalue le gradient et le Hessien de la fonction-objectif par la méthode des *différences finies* (DF). Le champ d'optimisation `opts = optimset('DerivativeCheck','on')` compare le gradient fourni par l'utilisateur à celui évalué par *DF*. Le champ d'optimisation `opts = optimset('FinDiffType', 'forward')` (par défaut) stipule que le gradient sera estimé par *différences finies progressives*. En indiquant la valeur `'central'`, dans ce cas, le gradient sera estimé par *différences finies centrées*. Le paramètre d'optimisation `FinDiffType` (type de différences finies) n'est valable que si le paramètre `DerivativeCheck` est activé. D'autres paramètres d'optimisation existent, à

l'instar de `DiffMaxChange`, `DiffMinChange`, `FinDiffRelStep`, ... etc. Consulter le `help` de Matlab® pour de amples informations.

Exercice ❸ ☞ Ⓢ

– Justifier de l'existence d'un extremum des fonctions-objectif suivantes :

$$\begin{cases} f_1(x_1,x_2) = (1-x_1)^2 + 100\,(x_2 - x_1^2)^2 \\ f_2(x_1,x_2) = 20 + x_1^2 + x_2^2 - 10\,(\cos(2\pi x_1) + \cos(2\pi x_2)) \\ f_3(x_1,x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2 \\ f_4(x_1,x_2) = \dfrac{1}{2} - \sin(x_1^2 + x_2^2) \\ f_5(x_1,x_2) = x_1^3 + x_2^3 - 3\,x_1\,x_2 \\ f_6(x_1,x_2) = x_1^2 - x_1\,x_2 + x_2^2 + 3\,x_1 - 2\,x_2 + 1 \end{cases} \tag{80}$$

– Trouver les extremums des fonctions ci-dessus, analytiquement, ensuite en se servant de la fonction prédéfinie `fminsearch`

*B. **Optimisation avec contraintes***

De nombreux problèmes en physique, en chimie, en ingénierie et en économie nécessitent de minimiser une fonction-objectif soumise à plusieurs contraintes. Dans ce qui suit, nous nous intéresserons à la résolution de problèmes d'optimisation sous contraintes dont la formulation mathématique générale est donnée par :

$$\hat{X} = \underset{X\in\mathbb{R}^n}{argmin}\ f(X) \tag{81}$$

$$\text{tel que} \begin{cases} h_i(x) = 0, & \{i = 1,2,...,n\} \\ g_j(x) \leq 0, & \{j = 1,2,...,m\} \end{cases}$$

Avec, P est un sous-ensemble non vide de $\mathbb{R}^n$ défini par des contraintes d'égalité et/ou d'inégalité de fonctions :

$$P = \{x \in \mathbb{R}^n : h_i(x) = 0, g_j(x) \leq 0\} \tag{82}$$

Ainsi, l'ensemble P est appelé domaine des contraintes, $g = (g_1, g_2, ..., g_m)$ sont les contraintes d'inégalité et $h = (h_1, h_2, ..., h_n)$ sont les contraintes d'égalité. Dans cette section, il sera question de présenter l'ensemble des fonctions Matlab® prédéfinies, dédiées à la résolution de problèmes d'optimisation avec contraintes. Toutes ces fonctions sont disponibles dans la boite à outil `Optimization Toolbox` de Matlab®.

La fonction `linprog` (Linear programming) solutionne un processus d'optimisation, écrit sous une formulation linéaire minimisant la quantité :

$$\underset{X\in\mathbb{R}^n}{min}\ f^T\,x \text{ tel que} \begin{cases} A \times x \leq B \\ Aeq \times x = Beq \\ lB \leq x \leq uB \end{cases} \tag{83}$$

`linprog` s'utilise avec deux types d'algorithmes *Large-échelle* (Large-Scale Optimization) et *Moyenne-échelle* (Medium-Scale Optimization). Le premier type est utilisé pour des systèmes complexes en terme de taille et sous certaines conditions qu'on ne va pas détailler ici. On se contentera d'utiliser le deuxième type et plus précisément la méthode *Simplex*. La syntaxe usuelle de `linprog` est :

```
[x, fval, exitflag, output, lambda] = linprog(f, A, B, Aeq, Beq, lB , uB, x0, options)
```

Les quantités `x, f, B, Beq, lB` et `uB` sont des vecteurs. Les arguments `A` et `Aeq` sont des matrices. L'argument `f` désigne le vecteur des coefficients des variables, tel que $f^T\,x = f(1)\,x(1) + f(2)\,x(2) \ldots f(n)\,x(n)$. Les contraintes linéaires de type équation et inéquation sont écrites respectivement selon $Aeq \times x = Beq$ et $A \times x \leq B$. L'argument de sortie `lambda` désigne le *multiplicateur de Lagrange*. Les arguments `lB` et `uB` délimitent l'intervalle de définition des variables en question. Pour les variables non bornées, on mettra `lB = -inf` et `uB = inf`.

Exercice ❹ ☞ ®

1) Minimiser la fonction-objectif définie par :

$$f(x) = -5\,x_1 - 4\,x_2 - 6\,x_3$$

$$\hat{X} = \underset{X\in\mathbb{R}^3}{argmin}\, f(X) \quad \text{tel que} \begin{cases} x_1 - x_2 + x_3 \leq 20 \\ 3\,x_1 + 2\,x_2 + 4\,x_3 \leq 42 \\ 3\,x_1 + 2\,x_2 \leq 30 \\ 0 \leq x_1,\ 0 \leq x_2,\ 0 \leq x_3 \end{cases}$$

2) Maximiser la fonction-objectif définie par :

$$f(x) = 14\,x_1 + 6\,x_2$$

$$\hat{X} = \underset{X\in\mathbb{R}^2}{argmin}\, f(X) \quad \text{tel que} \begin{cases} x_1 + x_2 \leq 7.50 \\ 11\,x_1 + 3\,x_2 \leq 0.40 \\ 12\,x_1 + 21\,x_2 \leq 1.50 \\ x_1 \geq 0,\ x_2 \geq 0 \end{cases}$$

Script Matlab®, concernant la minimisation

```
clear all ; clc ;

opts = optimset('LargeScale', 'off','Simplex', 'on') ;
options = optimset(opts, 'Display', 'iter', 'TolFun', 1e-5) ;

f = [-5 ; -4 ; -6] ; A = [1 -1 1 ; 3 2 4 ; 3 2 0] ;
B = [20; 42; 30] ; lB = [0 ; 0 ; 0] ; uB = [inf ; inf ; inf] ;

[x, fval, exitflag, output,lambda] = linprog(f, A, B, [], ...
[],lB, uB, [], options) ;

point_critique = sprintf('%6.4f \n', x)
```

Le champ d'optimisation `opts = optimset('LargeScale', 'off' ...)` signifie qu'on utilisera l'algorithme *Moyenne-échelle* pour résoudre ce problème. Le critère d'arrêt est considéré pour la fonction à travers `options = optimset... 'TolFun', 1e-5)`, autrement dit l'algorithme s'arrête une fois que la condition $|f(x_{k+1}) - f(x_k)| \leq 1e-5$ est satisfaite. Ci-dessous, l'affichage généré par le script.

```
The default starting point is feasible, skipping Phase 1.

Phase 2: Minimize using simplex.
      Iter              Objective              Dual Infeasibility
```

```
                        f'*x                    A'*y+z-w-f
       0                   0                       8.77496
       1                 -63                       1.11803
       2                 -78                             0
Optimization terminated.

>> point_critique =       % solution
                    0.0000
                   15.0000
                    3.0000
>> fval =
                   -78
>> exitflag =
                    1
```

On constate que l'algorithme *Simplex* converge vers la solution approchée au bout de trois itérations. Ci-dessous, le script Matlab® relatif à la maximisation.

```
clear all ; clc ;

opts = optimset('LargeScale', 'off','Simplex', 'on') ;
options = optimset(opts, 'Display', 'iter', 'TolFun', 1e-5) ;

f = [-14 ; -6] ; A = [1 1 ; 11 3 ; 12 21] ;
B = [7.50 ; 0.40 ; 1.50] ; lB = [0 ; 0] ; uB = [inf ; inf] ;

[x, fval, exitflag, output,lambda] = linprog(f, A, B, [], ...
[],lB, uB, [], options) ;

point_critique = sprintf('%6.4f \n', x)
```

Maximiser la fonction $14\,x_1 + 6\,x_2$ revient à minimiser $-14\,x_1 - 6\,x_2$. Ci-dessous, l'affichage généré par ce script.

```
The default starting point is feasible, skipping Phase 1.

Phase 2: Minimize using simplex.
      Iter            Objective              Dual Infeasibility
                        f'*x                    A'*y+z-w-f
       0                   0                       15.2315
       1            -0.509091                      2.18182
       2                -0.64                            0
Optimization terminated.

>> point_critique =       % point critique du maximum
                   0.0200
                   0.0600
>> fval =
                  -0.6400
```

```
>> exitflag =
                    1
```

Désormais on se servira de la fonction `fmincon`. La topologie générale d'un processus d'optimisation avec contraintes sur les variables, peut s'écrire suivant la notation compacte suivante :

$$\hat{X} \in \underset{X\in\mathbb{R}^n}{argmin} f(X) \quad \text{tel que} \quad \begin{cases} A \times x \leq B \\ Aeq \times x = Beq \\ C(x) \leq x \\ Ceq(x) = x \\ lB \leq x \leq uB \end{cases} \tag{84}$$

Les quantités x, B, Beq, lB, et uB sont des vecteurs. A et Aeq sont des matrices, $f(x)$, $C(x)$ et $Ceq(x)$ sont des fonctions pouvant êtres également des fonctions non-linéaires. Afin de résoudre des problèmes d'optimisation sous contraintes, on se servira de `fmincon`. Cette fonction sert à optimiser des fonctions-objectifs multidimensionnelles non-linéaires. Par défaut, l'algorithme d'optimisation est basé sur la méthode **SQP** (Sequential Quadratic Programming). La syntaxe usuelle de `fmincon` s'écrit selon :

```
[x, fval, exitflag, output, lambda, grad, hessian] = fmincon(fun, x0, A, B, Aeq, Beq, lB,
   uB, @nonlcon, options)
```

Les contraintes linéaires de type équation et inéquation sont écrites respectivement selon $Aeq \times x = Beq$ et $A \times x \leq B$. Les contraintes non-linéaires de type équation et inéquation sont données respectivement par $Ceq(x) = x$ et $C(x) \leq x$. La fonction *M-file* `@nonlcon` contient les contraintes non-linéaires de type équation et inéquation. Les autres arguments en entrée et en sortie ont la même signification que ceux des fonctions vues précédemment. Il est très important de souligner que si un ou plusieurs arguments ci-dessus sont manquants, on doit les remplacer par un ensemble vide `[]`. L'ordre d'apparition des arguments en entrée est important, on doit toujours commencer par les contraintes de type inégalité même si ce type de contrainte est vide. Nous commencerons dans un premier temps par résoudre un problème d'optimisation sous contraintes linéaires.

Exercice ❺ ☞ ®

– On se propose de minimiser la fonction-objectif définie par :

$$f(x_1, x_2) = 2\,x_1^2 + x_1\,x_2 + 2\,x_2^2 - 6\,x_1 - 6\,x_2 + 15$$

$$\hat{X} = \underset{X\in\mathbb{R}^2}{argmin}\, f(X) \quad \text{tel que} \quad \begin{cases} x_1 + 2\,x_2 \leq 5 \\ 4\,x_1 \leq 7 \\ x_2 \leq 2 \\ -2\,x_1 + 2\,x_2 = -1 \end{cases}$$

Script Matlab®

```
clc ; clear all ;
%%%%%%%%%%%%%%%%%%% Optimisation sous contraintes %%%%%%%%%%%%%%%%%%
fun = @(x) 2*x(1).^2 + x(1)*x(2) + 2*x(2).^2 - 6*x(1) - 6*x(2) + 15 ;

A = [1 2 ; 4 0 ; 0 1] ; B = [5 7 2] ;
Aeq = [-2 2] ; Beq = -1 ; xinit = [-1 1/2] ;
```

```
options = optimset('LevenbergMarquardt','on','Display','iter', ...
    'TolX', 1e-4) ;
[x, fval, exitflag, output, lambda, grad, hessian] = fmincon(fun, ...
    xinit, A, B, Aeq, Beq, [], [], [], options)
```

Le champ d'optimisation `optimset('LevenbergMarquardt','on', ...)` indique que l'opération d'optimisation sera menée par le biais de l'algorithme de *Levenberg-Marquardt*. Le choix de cet algorithme peut se faire également avec la syntaxe `optimset('NonlEqnAlgorithm','lm', ...)`. On peut aussi faire appel à l'algorithme de *Gauss-Newton*, en utilisant la syntaxe `optimset('NonlEqnAlgorithm','gn', ...)`. Les différentes sorties renvoyées par ce script sont énumérées ci-dessous.

```
%%%%%%%%%%%%%%%%%%%%%%% affichage par defaut %%%%%%%%%%%%%%%%%%%%%%%
Iter F-count        f(x)   constraint
    0       3          20            4
    1       6      8.4375    1.332e-15
    2       9     8.05206            0
    3      12      7.9875     2.22e-16

Optimization terminated: first-order optimality measure less
 than options.TolFun and maximum constraint violation is less
 than options.TolCon.
No active inequalities.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
>> x =                  % coordonnees du point critique

      1.4500    0.9500

>> fval =

      7.9875

>> exitflag =

      1

>> output =

         iterations: 3
          funcCount: 12
       lssteplength: 1
           stepsize: 0.1607
          algorithm: [1x44 char]
      firstorderopt: [1x1 double]
    constrviolation: [1x1 double]
            message: [1x144 char]

>> lambda =

         lower: [2x1 double]
         upper: [2x1 double]
         eqlin: 0.3750
```

```
      eqnonlin: [0x1 double]
       ineqlin: [3x1 double]
    ineqnonlin: [0x1 double]

>> grad =

      0.7500
     -0.7500

>> hessian =

      2.6500    2.3500
      2.3500    2.6500
```

Nous résolvons dans l'exercice ci-dessous, un problème d'optimisation avec contraintes non-linéaires.

Exercice ❻ ☞ Ⓡ

– On se propose de minimiser le fonction-objectif définie par :

$$f(x_1, x_2) = \exp(x_1)\,(4\,x_1^2 + 2\,x_2^2 + 4\,x_1\,x_2 + 2\,x_2 + 1)$$

$$\hat{X} = \underset{X \in \mathbb{R}^2}{argmin}\, f(X) \quad \text{tel que} \begin{cases} 2 + x_1\,x_2 - x_1 - x_2 \leq 0 \\ -x_1\,x_2 \leq 10 \end{cases}$$

On commence d'abord par écrire le fichier `M-file` correspondant aux contraintes non-linéaires. Ce fichier bien entendu est sauvegardé sous le nom `mycontr.m`. Voici le script :

```
function [C, Ceq] = mycontr(x)
% fonction definissant les contraintes non lineaires

C = [2 + x(1)*x(2) - x(1) - x(2) ; -x(1)*x(2) - 10] ; % inequation
Ceq = [] ; % pas de contraintes non lineaires en equation
return
```

Ci-dessous le script Matlab® du programme appelant.

```
clc ; clear all ;

fun = @(x) exp(x(1))*(4*x(1)^2 + 2*x(2)^2 + 4*x(1)*x(2) + 2*x(2) + 1)

xinit = [-1 1] ;

options = optimset('LevenbergMarquardt','on','Display','iter', ...
    'TolX', 1e-4) ;

[x, fval, exitflag, output, lambda, grad, hessian] = fmincon(fun, ...
    xinit, [], [], [], [], [], [], @mycontr, options)
```

Les arguments de sorties renvoyés sont :

```
%%%%%%%%%%%%%%%%%%%%%%% affichage par defaut %%%%%%%%%%%%%%%%%%%%%%%
Iter F-count        f(x)   constraint
    0       3      1.8394            1
    1       6     1.98041      -0.1839
    2       9     1.63636       0.2596
    3      12     0.34132        4.524
    4      15    0.530719       0.3344
    5      18    0.447582     -0.08164
    6      21    0.123147        2.352
    7      24   0.0575464       0.3957
    8      27   0.0335016       0.1153
    9      30   0.0331726    0.0001285
   10      33    0.033173    1.589e-10

Optimization terminated: first-order optimality measure less
 than options.TolFun and maximum constraint violation is less
 than options.TolCon.
Active inequalities (to within options.TolCon = 1e-06):
  lower      upper     ineqlin   ineqnonlin
                                        1
                                        2
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
>> x =

    -9.0990    1.0990

>> fval =

     0.0332

>> exitflag =

      1

>> output =

         iterations: 10
          funcCount: 33
      lssteplength: 1
           stepsize: [1x1 double]
          algorithm: [1x44 char]
      firstorderopt: [1x1 double]
    constrviolation: [1x1 double]
            message: [1x144 char]

>> lambda =

        lower: [2x1 double]
        upper: [2x1 double]
```

```
         eqlin: [0x1 double]
      eqnonlin: [0x1 double]
        ineqlin: [0x1 double]
    ineqnonlin: [2x1 double]

>> grad =

    0.0255
   -0.0034

>> hessian =

    0.0280    0.0035
    0.0035    0.0096
```

Exercice ❼ ☞ Ⓢ

– On se propose de minimiser les fonctions-objectif définies par :

$$f(x_1, x_2) = \frac{1}{2}(x_1 - 3)^2 + \frac{1}{2}(x_2 - 1)^2$$

$$\hat{X} = \underset{X \in \mathbb{R}^2}{argmin}\, f(X) \quad \text{tel que} \begin{cases} x_1 + x_2 - 1 \leq 0 \\ x_1 - x_2 - 1 \leq 0 \\ -x_1 + x_2 - 1 \leq 0 \\ -x_1 - x_2 - 1 \leq 0 \end{cases}$$

La fonction `quadprog` (Quadratic programming) solutionne un processus d'optimisation, écrit sous une formulation quadratique qui minimise la quantité :

$$\underset{X \in \mathbb{R}^n}{min}\, f(X) = \frac{1}{2}\, x^T\, H\, x + f^T\, x \text{ tel que } \begin{cases} A \times x \leq B \\ Aeq \times x = Beq \\ lB \leq x \leq uB \end{cases} \tag{85}$$

La syntaxe usuelle de cette fonction est :

```
[x, fval, exitflag, output, lambda] = quadprog(H, f, A, B, Aeq, Beq, lb, ub, x0, options)
```

L'argument `H` est le Hessien de la fonction-objectif et `f` représente le vecteur des coefficients de la partie linéaire de la fonction-objectif. Ces deux arguments sont obligatoires tandis que les autres sont optionnels. Ces derniers ont la même signification que ceux de la fonction `fmincon`, pareil également pour les arguments de sortie. L'ordre d'apparition des arguments doit être respecté, si un argument optionnel n'est pas utilisé, il faudra le remplacer par l'ensemble vide `[]`.

Exercice ❽ ☞ Ⓡ

– On se propose de minimiser les fonctions-objectif définies par :

$$f(x_1, x_2) = x_1^2 + 4\, x_1 + 5\, x_2$$

$$\hat{X} = \underset{X \in \mathbb{R}^2}{argmin}\, f(X) \quad \text{tel que} \begin{cases} 2\, x_1 + x_2 \geq 10 \\ 3\, x_1 + 6 x_2 \leq 80 \\ 5\, x_1 + 7\, x_2 \leq 50 \\ x_1, x_2 \geq 0 \end{cases}$$

$$f(x_1, x_2, x_3) = x_1^2 + x_1\,x_2 + 2\,x_2^2 + 2\,x_3^2 + 2\,x_2\,x_3 + 4\,x_1 + 6\,x_2 + 12\,x_3$$

$$\hat{X} = \underset{X\in\mathbb{R}^3}{argmin}\, f(X) \quad \text{tel que} \begin{cases} x_1 + x_2 + x_3 \geq 6 \\ -x_1 - x_2 + 2\,x_3 \geq 2 \\ 0 \leq x_1, x_2, x_3 \leq 100 \end{cases}$$

- Réécrire la fonction-objectif selon la notation générale décrite par l'Eq. (85).
- Trouver les coordonnées du point critique en utilisant `quadprog`.

Réécrivons d'abord ce système selon la notation générale décrite par l'Eq. (85).

$$\underset{X\in\mathbb{R}^2}{min}\ \frac{1}{2}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 4 \\ 5 \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\begin{bmatrix} -2 & -1 \\ 3 & 6 \\ 5 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} 10 \\ 80 \\ 50 \end{bmatrix}$$

D'après la définition sur les contraintes d'inégalités, elles doivent être écrites, inférieure ou égale à une constante. La contrainte $2\,x_1 + x_2 \geq 10$ est réécrite sous la forme $-2\,x_1 - x_2 \leq 10$. Voici le script Matlab®, pour la fonction à deux variables

```
clc ; clear all ;

H = [2 0 ; 0 0] ; f = [4 ; 5] ;
A = [-2 -1 ; 3 6 ; 5 7] ; B = [10 ; 80 ; 50] ; lB = [0 ; 0] ;
uB = [inf ; inf] ;

options = optimset('LargeScale','off','TolX', 1e-7) ;
[x, fval, exitflag, output, lambda] = quadprog(H, f, A, B, [],[], ...
    lB, uB, [], options) ;

if exitflag > 0

   disp('L''algorithme a converge vers la solution :')
   point_critique = x
else
    disp('L''algorithme n''a pas converge !')
end
```

Ci-dessous les sorties correspondantes :

```
Optimization terminated.
L'algorithme a converge vers la solution :

point_critique =

     0
     0
```

Notons qu'on peut réécrire ce script en considérant l'inégalité $x_1, x_2 \geq 0$ comme deux contraintes séparées selon $-x_1 \leq 0$ et $-x_2 \leq 0$ ce qui revient à écrire deux nouvelles lignes `[-1 0 ; 0 -1]` dans la matrice A et `[0 ; 0]` dans le vecteur B. Dans ce cas `lB = []` et `uB = []`. Ci-dessous le script.

```
clc ; clear all ;

H = [2 0 ; 0 0] ; f = [4 ; 5] ;
A = [-2 -1 ; 3 6 ; 5 7 ; -1 0 ; 0 -1] ; B = [10 ; 80 ; 50 ; 0 ; 0] ;

options = optimset('LargeScale','off','TolX', 1e-7) ;
[x, fval, exitflag, output, lambda] = quadprog(H, f, A, B, [],[], ...
    [], [], [], options) ;

if exitflag > 0

   disp('L''algorithme a converge vers la solution :')
   point_critique = x
else

    disp('L''algorithme n''a pas converge !')

end
```

Script Matlab® pour la fonction de trois variables

```
clc ; clear all ;

H = [2 1 0 ; 1 4 2 ; 0 2 4] ; f = [4 ; 6 ; 12] ;
A = [-1 -1 -1 ; 1 1 -2] ; B = [-6 ; -2] ; lB = [0 ; 0 ; 0] ;
uB = [100 ; 100 ; 100] ; xinit = [1 ; 1 ; 1] ;

options = optimset('Diagnostics','on','TolX', 1e-7) ;
[x, fval, exitflag, output, lambda] = quadprog(H, f, A, B, [],[], ...
    lB, uB, xinit,options) ;

if exitflag > 0

   disp('L''algorithme a converge vers la solution :')
   point_critique = x
else

    disp('L''algorithme n''a pas converge !')

end
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Diagnostic Information
Number of variables: 3
 Number of linear inequality constraints:    2
```

```
 Number of linear equality constraints:      0
 Number of lower bound constraints:          3
 Number of upper bound constraints:          3

Algorithm selected
   medium-scale: active-set
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
End diagnostic information
```

Comme on peut le constater, le champ `optimset('Diagnostics','on', ...)` renvoie des informations sur la fonction-objectif, comme le nombre de variables, le nombre d'équation et d'inéquation ... etc.

```
Optimization terminated.
L'algorithme a converge vers la solution :

>> point_critique =
                   3.3333
                   0.0000
                   2.6667
```

Afin de maximiser la fonction-objectif au moyen de la fonction `quadprog`, on prendra `-H` et `-f`.

Exercice ❾ ☞ Ⓢ

– Minimiser les fonctions-objectif définies par :

$$f(x_1, x_2) = (x_1 - 2)^2 + (x_2 - 2)^2$$

$$\hat{X} = \underset{X \in \mathbb{R}^2}{argmin}\, f(X) \quad \text{tel que} \begin{cases} x_1 + 2\,x_2 \leq 3 \\ 3\,x_1 + 2x_2 \geq 3 \\ x_1 - 2\,x_2 \leq 2 \\ x_1, x_2 \geq 0 \end{cases}$$

$$f(x_1, x_2, x_3) = x_1^3 + x_2^3 + x_3^3$$

$$\hat{X} = \underset{X \in \mathbb{R}^3}{argmin}\, f(X) \quad \text{tel que} \begin{cases} x_1^3 + x_2^3 + x_3^3 = 1 \\ 2\,x_3^3 - x_2^2 \leq 0 \\ x_1 \geq 0 \\ x_3 \leq 0 \end{cases}$$

– Réécrire la fonction-objectif selon la notation générale décrite par l'Eq. (85).
– Trouver les coordonnées du point critique en utilisant `quadprog`. Pour la fonction à trois variables, prenez le vecteur des valeurs initiales $(1, 0, -1)$.

Pour conclure cette section, il recommandé également dans le même sillage de consulter les fonctions d'optimisation prédéfinies `fseminf`, `fminimax` et `fgoalattain`.

Annexe A
Dérivée directionnelle

Nous souhaitons quantifier le taux de variation de la fonction $f : \mathbb{R}^2 \longmapsto \mathbb{R}$ lorsqu'elle passe d'un point $f(x_0, y_0)$ au point $f(x, y)$. Nous travaillerons sur le plan de projection xoy, ce taux de variation est évalué par le segment de droite $\overline{P_0P}$. Soit $\overrightarrow{d} = a\,\vec{i} + b\,\vec{j}$ un vecteur unitaire ayant la même direction que $\overrightarrow{P_0P}$.

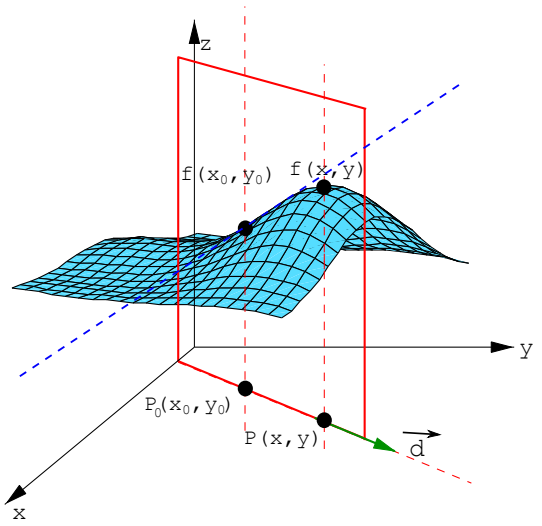


Figure 4: Dérivée directionnelle au point P_0 dans la direction $\vec{d}$.

Sur la figure ci-dessus :

$$\begin{aligned} \overrightarrow{P_0P} // \overrightarrow{d} \Rightarrow \overrightarrow{P_0P} &= \lambda\,\overrightarrow{d} \quad \text{avec} \quad \lambda \in \mathbb{R}_+^* \\ \Rightarrow \overrightarrow{P_0P} &= \lambda\,(a\,\vec{i} + b\,\vec{j}) = \lambda\,a\,\vec{i} + \lambda\,b\,\vec{j} \end{aligned} \tag{86}$$

D'un autre côté on peut définir le vecteur $\overrightarrow{P_0P}$ par rapport à l'origine O selon :

$$\overrightarrow{P_0P} = \overrightarrow{OP} - \overrightarrow{OP_0} \tag{87}$$
$$= (x\,\vec{i} + y\,\vec{j}) - (x_0\,\vec{i} + y_0\,\vec{j}) \tag{88}$$
$$= x\,\vec{i} + y\,\vec{j} - x_0\,\vec{i} - y_0\,\vec{j} \tag{89}$$
$$= (x - x_0)\,\vec{i} + (y - y_0)\,\vec{j} \tag{90}$$

Par identification des équations (86) et (90), il vient :

$$\begin{cases} x - x_0 = \lambda\,a \\ y - y_0 = \lambda\,b \end{cases} \Rightarrow \begin{cases} x = x_0 + \lambda\,a \\ y = y_0 + \lambda\,b \end{cases} \tag{91}$$

Par voie de conséquence, la dérivée directionnelle de $f(x, y)$ dans la direction du vecteur unitaire $\overrightarrow{d} = a\,\vec{i} + b\,\vec{j}$ au point $f(x_0, y_0)$ est :

$$f_{\overrightarrow{d}}(x_0, y_0) = \lim_{\lambda \to 0} \frac{f(x_0 + \lambda\,a, y_0 + \lambda\,b) - f(x_0, y_0)}{\lambda} \tag{92}$$

Théorème *: La dérivée directionnelle est maximale lorsque* $\overrightarrow{d}$ *a la même direction que* $\nabla f(x_0, y_0)$ *de plus le taux de variation maximal de* $f(x, y)$ *en* (x_0, y_0) *est* $||\nabla f(x_0, y_0)||$.

Ce théorème peut être prouvé en considérant que $f_{\overrightarrow{d}}(x_0, y_0) = \nabla f(x_0, y_0)\ \overrightarrow{d} = ||\nabla f(x_0, y_0)||\,||\overrightarrow{d}||\cos(\theta) = ||\nabla f(x_0, y_0)||\cos(\theta)$. Ainsi $f_{\overrightarrow{d}}(x_0, y_0)$ est maximale si $\cos(\theta) = \pm 1$ autrement dit si la condition $\nabla f(x_0, y_0) // \overrightarrow{d}$ est satisfaite. Dans le cas où $\theta = \pi/2 \Rightarrow \nabla f(x_0, y_0) \perp \overrightarrow{d}$. Ce résultat indique que si je me déplace dans une direction perpendiculaire au ∇f, le taux de variation de la fonction $f(x, y)$ est nul.

- Si les deux vecteurs ∇f et $\overrightarrow{d}$ ont la même direction et le même sens, dans ce cas le vecteur unitaire $\overrightarrow{d}$ désigne une direction de croissance maximale de $f(x, y)$.
- Si les deux vecteurs ∇f et $\overrightarrow{d}$ ont la même direction et de sens opposé, dans ce cas le vecteur unitaire $\overrightarrow{d}$ désigne une direction de décroissance maximale de $f(x, y)$. Cette condition est prouvée selon :

$$\begin{aligned}
\nabla f(x_0, y_0) \times \overrightarrow{d} &= \nabla f(x_0, y_0) \times (-\nabla f(x_0, y_0)) \\
&= -\nabla f(x_0, y_0) \times \nabla f(x_0, y_0) \\
&= \underbrace{-\underbrace{||\nabla(x_0, y_0)||^2}_{>0}}_{<0} \\
&\Rightarrow \overrightarrow{d} = -\nabla f \quad \text{est une direction de descente}
\end{aligned}$$

Comme exemple d'application, calculons la dérivée directionnelle de la fonction $f(x, y) = 2\,e^{(x^2\,y)}$ au point (2, 3) dans la direction formant un angle de 75° avec l'axe des x positif. La solution est donnée ci-dessous :

$$\begin{aligned}
f_{\overrightarrow{d}}(2, 3) &= \nabla f(2, 3)\ \overrightarrow{d} \\
&= \frac{\partial f}{\partial x}(2, 3) \times \cos(75^\circ) + \frac{\partial f}{\partial y}(2, 3) \times \sin(75^\circ) \\
&= 4\,x\,y\,e^{(x^2\,y)} \times \cos(75^\circ) + 2\,x^2\,e^{(x^2\,y)} \times \sin(75^\circ) \\
&= 24\,e^{(4\times 3)} \times 0.25 + 8\,e^{(4\times 3)} \times 0.96 \\
&= 6\,e^{(12)} + 7.70\,e^{(12)} \\
&= 13.70\,e^{(12)}
\end{aligned}$$

Annexe B
Codage binaire

Dans ce système de numération binaire, on utilise la base 2. En prenant par exemple le nombre 55, celui-ci se décompose en : $55 = 32 + 16 + 4 + 2 + 1$.

$55 = 1 \times 32 + 1 \times 16 + 1 \times 4 + 1 \times 2 + 1 \times 1$
$55 = 1 \times 2^5 + 1 \times 2^4 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0$
$55 = 1 \times 2^5 + 1 \times 2^4 + \mathbf{0} \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0$
$55 = 110111$

On utilise également la notation $55_{(10)} = 110111_{(2)}$. Il est possible d'arriver au même résultat, en procédant par des divisions successives du nombre 55 suivant la base 2. La condition d'arrêt correspond à un quotient nul, ensuite on remonte les restes des divisions, selon :

$$\begin{array}{r|l} 55 & 2 \\ \hline 4 & 27 \\ \hline 15 & \\ 14 & \\ \hline 1 & \end{array} \qquad \begin{array}{r|l} 27 & 2 \\ \hline 2 & 13 \\ \hline 07 & \\ 6 & \\ \hline 1 & \end{array} \qquad \begin{array}{r|l} 13 & 2 \\ \hline 12 & 6 \\ \hline 1 & \end{array} \qquad \begin{array}{r|l} 6 & 2 \\ \hline 6 & 3 \\ \hline 0 & \end{array}$$

$$\begin{array}{r|l} 3 & 2 \\ \hline 2 & 1 \\ \hline 1 & \end{array} \qquad \begin{array}{r|l} 1 & 2 \\ \hline 0 & 0 \\ \hline 1 & \end{array}$$

On obtient $55_{(10)} = 110111_{(2)}$ (de la droite vers la gauche). On peut proposer l'algorithme ci-dessous pour transformer un nombre de la base décimale, par exemple le nombre 125, vers la base binaire.

```
%%%%%%%%%%%%%%%%% CONVERSION DECIMAL VERS LE BINAIRE %%%%%%%%%%%%%%%%%
clc ; clear all ;
% Samir KENOUCHE - le 09/08/2019

decNumber = 125 ; base = 2 ; binNumber = [];
quot = 1 ;

while quot ~= 0

    quot = floor(decNumber/base) ; reste = decNumber - quot*base  ;
    decNumber = floor(decNumber/base);
    binNumber = [reste binNumber] ;

end

binNumber
```

Ce script renvoie le résultat suivant `binNumber = 1111101`.

$$\begin{array}{r|l} 125 & 2 \\ \hline 12 & 62 \\ \hline 05 & \\ 4 & \\ \hline 1 & \end{array} \quad \begin{array}{r|l} 62 & 2 \\ \hline 6 & 31 \\ \hline 02 & \\ 2 & \\ \hline 0 & \end{array} \quad \begin{array}{r|l} 31 & 2 \\ \hline 2 & 15 \\ \hline 11 & \\ 10 & \\ \hline 1 & \end{array} \quad \begin{array}{r|l} 15 & 2 \\ \hline 14 & 7 \\ \hline 1 & \end{array}$$

$$\begin{array}{r|l} 7 & 2 \\ \hline 6 & 3 \\ \hline 1 & \end{array} \quad \begin{array}{r|l} 3 & 2 \\ \hline 2 & 1 \\ \hline 1 & \end{array} \quad \begin{array}{r|l} 1 & 2 \\ \hline 0 & 0 \\ \hline 1 & \end{array}$$

On retrouve ainsi le même résultat $1255_{(10)} = 1111101_{(2)}$. À l'inverse, la conversion du système binaire vers le système décimal se fera, suivant :

$110111_{(2)} = 1 \times 2^{(6-1)} + 1 \times 2^{(6-2)} + 0 \times 2^{(6-3)} + 1 \times 2^{(6-4)} + 1 \times 2^{(6-5)} + 1 \times 2^{(6-6)}$.
$110111_{(2)} = 1 \times 2^5 + 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0$.
$110111_{(2)} = 32 + 16 + 0 + 4 + 2 + 1 = 55_{(10)}$

Cette conversion est implémentée dans le script Matlab® ci-dessous :

```
%%%%%%%%%%%%%%%%%% CONVERSION BINAIRE VERS LE DECIMAL %%%%%%%%%%%%%%%%%
clc ; clear all ;
% Samir KENOUCHE - le 09/08/2019

binNumber = [1 1 0 1 1 1] ; base = 2 ; decNumber = 0 ;

for i = 1 : numel(binNumber)

   decNumber = decNumber + binNumber(i)*base^(numel(binNumber) - i) ;
end

decNumber
```

Pour la conversion de nombres fractionnaires on écrira :

$01101.010_{(2)} = 0 \times 2^4 + 1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3}$
$01101,010_{(2)} = 0 \times 16 + 1 \times 8 + 1 \times 4 + 0 \times 2 + 1 \times 1 + 0 \times 0.5 + 1 \times 0.25 + 0 \times 0.12$
$01101.010_{(2)} = 13.25_{(10)}$

Pour la conversion d'un nombre fractionnaire d'un système décimal vers le binaire, on procèdera comme suit :

$55.8625_{(10)} \stackrel{?}{=} xxxxx_{(2)}$

On commence d'abord par écrire la partie entière en binaire : $55_{(10)} = 110111_{(2)}$. Ensuite on multiplie par 2 la partie fractionnaire :

$0.8625 \times 2 = 1.725 = 1 + 0.725$
$0.7250 \times 2 = 1.450 = 1 + 0.450$

$0.4500 \times 2 = 0.900 = 0 + 0.900$
$0.9000 \times 2 = 1.800 = 1 + 0.800$
$0.8000 \times 2 = 1.600 = 1 + 0.600$ ...

À partir de cet exemple, on note que la partie entière est codée, par des divisions successives par 2, sur un nombre donné de bits. La partie fractionnaire est codée sur un nombre donné de bits en multipliant successivement par 2 jusqu'à ce que la partie fractionnaire soit nulle ou le nombre de bits considéré est atteint. Cette conversion donne :

$$55.8625_{(10)} = 110111.11011_{(2)} = 11011.111011_{(2)}\, 2^1$$

Chapitre[3] Équations aux dérivées partielles et fonctions de *Green*

Samir Kenouche - Département des Sciences de la Matière - UMKB

Méthodes Mathématiques et Algorithmes pour la Physique

Résumé

Ce chapitre débute par un rappel succinct de quelques notions élémentaires sur les équations différentielles. Les méthodes numériques *d'Euler*, de *Heun*, de *Crank-Nicolson* et de *Runge-Kutta* ont été étudiées en détail en deuxième années L2, module : *Méthodes numériques et programmation*. A cet effet, je renvoie les lecteurs intéressés à mon site web personnel. Ainsi, il a été jugé inutile de les rappeler ici, en revanche, l'accent a été mis particulièrement sur les résolutions analytique et numérique des équations aux dérivées partielles. A cet égard, de nombreux problèmes issus de la physique ont été résolus, à l'instar de : l'équation de Schrödinger, la corde élastique, diffusion de la chaleur, convection-diffusion de la chaleur et l'équation de la flexion. L'équation de Schrödinger est l'équation maitresse de la mécanique quantique, décrivant les propriétés de la matière à des échelles atomiques. Cette équation a été résolue en détail sans occultation des interprétations physiques des différentes étapes. Les autres problèmes (linéaire et non-linéaire) ont été résolus numériquement, au moyen de scripts Matlab®, en considérant la méthode des *différences finies*. Finalement, soulignons que les fondements mathématiques des méthodes numériques sont délibérément occultés afin de privilégier d'avantage les aspects opérationnels, qui sont les plus importants pour un non-mathématicien. L'objectif est de conférer aux étudiants (es) issus des sciences expérimentales, un certain nombre d'outils mathématiques, afin qu'ils puissent savoir se servir d'un ordinateur pour résoudre différents problèmes de la physique.

Table des matières

S. Kenouche est docteur en Physique de l'Université des Sciences et Techniques de Montpellier (France) et docteur en Chimie de l'Université de Béjaia (Algérie).
Site web : voir http://www.sites.univ-biskra.dz/kenouche
Version corrigée, améliorée et actualisée le 10.10.2020.

I. Introduction

UNe équation différentielle est une relation fonctionnelle dont l'inconnue est une fonction $y(t)$, avec $t \in [a,\, b]$. La forme générale d'une telle équation s'écrit :

$$f(y^{(n)}, y^{(n-1)}, y^{(n-2)}, ..., y^{(1)}, y, t) = \phi(t) \tag{1}$$

Avec, $y^{(n)}$ est la nième dérivée de la fonction y et $\varphi(t)$ désigne le second membre de l'équation différentielle. Dans le cas où $\phi(t) = 0$, on dira que l'équation différentielle est homogène. L'existence d'une solution unique de l'équation différentielle est tributaire de l'imposition de certaines conditions initiales sur $y(t)$ et ses dérivées. Dans l'équation ci-dessus, les conditions initiales sont les valeurs de $y(a), y^{(1)}(a), y^{(2)}(a), ..., y^{(n)}(a)$. Cependant, il faut noter que très souvent la solution analytique n'existe pas, et on doit par conséquent approcher la solution exacte $y(t)$ par des méthodes numériques en calculant des approximations successives à chaque instant.

A. *Problème de Cauchy*

Le problème de Cauchy consiste à trouver une fonction continûment dérivable $y : t \in \mathbb{R}^+ \longrightarrow y(t) \in \mathbb{R}$ vérifiant :

$$\mathcal{P} : \begin{cases} y^{'}(t) = f(t, y(t)), \quad t > 0 \\ \\ y(0) = y_0 \end{cases} \tag{2}$$

La première équation est une équation différentielle et la deuxième relation exprime une *condition de Cauchy* ou *condition initiale*. Une solution $y(t)$ au problème (2) est appelée *intégrale de l'équation différentielle*. En effet, ce problème est équivalent à l'équation intégrale :

$$y(t) = y_0 + \int_{t_0}^{t} f(u, y(u))\, du \tag{3}$$

a) ***Définition****: soit $f : \mathbb{I} \times \mathbb{R} \longmapsto \mathbb{R}$ une fonction donnée, s'il existe une constante $L > 0$ telle que :*

$$|f(t, u) - f(t, v)| \leq L\,|u - v| \tag{4}$$

$\forall\, u, v \in \mathbb{R}$ et $\forall\, t \in \mathbb{I}$ alors f est dite lipschitzienne de rapport L sur $\mathbb{I} \times \mathbb{R}$ ou simplement L-lipschitzienne.

b) ***Théorème****: si f est continue sur $\mathbb{I} \times \mathbb{R}$ et L-lipschitzienne par rapport à sa deuxième variable $y(t)$ alors le problème de Cauchy admet une solution unique sur $\mathbb{I}$, $\forall\, u(0) \in \mathbb{R}$.*

c) ***Démonstration****:* pour démontrer ce théorème nous considérons l'application $\varphi : (\mathbb{R} \to \mathbb{R}^n) \longmapsto (\mathbb{R} \to \mathbb{R}^n)$ qui est définie par :

$$\varphi(y)(t) = y_0 + \int_{t_0}^{t} f(u, y(u))\, du \tag{5}$$

De sorte que,

$$\varphi(y)(t_0) = y_0 + \underbrace{\int_{t_0}^{t_0} f(u, y(u))\, du}_{=0} = y_0 \tag{6}$$

Ainsi $\varphi(y)(t_0)$ satisfait toujours la condition initiale. En outre, si elle admet un point fixe, ie :

$$\varphi(y) = y \quad \Rightarrow \quad y(t) = y_0 + \int_{t_0}^{t} f(u, y(u))\, du \tag{7}$$

$$\Rightarrow \quad \frac{dy(t)}{dt} = 0 + \frac{d}{dt}\left[\int_{t_0}^{t} f(u, y(u))\, du\right] \quad \Rightarrow \quad y^{'}(t) = f(t, y(t)) \tag{8}$$

Donc un point fixe de l'application φ sera forcément une solution de l'équation différentielle avec la même condition initiale. La démonstration du théorème revient désormais à prouver que l'application φ est *contractante*. Cela signifie qu'elle va contracter l'espace des fonctions de sorte à rapprocher toute fonction d'une fonction qui est solution du problème. Ci-dessous la démonstration :

$$\begin{aligned}
\| \varphi^{(1)}(y_1)(t) - \varphi^{(1)}(y_2)(t) \| &= \| y_0 + \int_{t_0}^{t} f(u, y_1(u))\, du - y_0 - \int_{t_0}^{t} f(u, y_2(u))\, du \| \quad \text{inég. triang.} \\
&\leq \int_{t_0}^{t} \| f(u, y_1(u)) - f(u, y_2(u)) \|\, du \quad \text{avec } f \text{ est L-lipschitzienne} \\
&\leq L \int_{t_0}^{t} \underbrace{\| y_1(u) - y_2(u) \|}_{\leq \|y_1 - y_2\|_\infty} du \\
&\leq L\,(t - t_0)\ \| y_1 - y_2 \|_\infty
\end{aligned}$$

En itérant une deuxième fois :

$$\begin{aligned}
\| \varphi^{(2)}(y_1)(t) - \varphi^{(2)}(y_2)(t) \| &= \| y_0 + \int_{t_0}^{t} f(u, \varphi(y_1)(u))\, du - y_0 - \int_{t_0}^{t} f(u, \varphi(y_2)(u))\, du \| \\
&\leq \int_{t_0}^{t} \| f(u, \varphi(y_1)(u)) - f(u, \varphi(y_2)(u)) \|\, du \\
&\leq \int_{t_0}^{t} \| \varphi(y_1)(u) - \varphi(y_2)(u) \|\, du \\
&\leq L^2 \int_{t_0}^{t} \| y_1 - y_2 \|_\infty\, du \\
&\leq L^2\, \frac{(t - t_0)^2}{2}\ \| y_1 - y_2 \|_\infty
\end{aligned}$$

Par récurrence à la nième itération, nous obtenons :

$$\| \varphi^{(n)}(y_1)(t) - \varphi^{(n)}(y_2)(t) \| \leq L^n\, \frac{(t - t_0)^n}{n}\ \| y_1 - y_2 \|_\infty \tag{9}$$

Le terme $L^n \dfrac{(t - t_0)^n}{n}$ *tend vers zéro quand n tend vers l'infinie* $\Rightarrow$ *l'application* φ *est contractante. Ce théorème garantit l'unicité des solutions des équations différentielles pour une condition initiale donnée. Autrement dit, à deux conditions initiales différentes correspondent deux solutions différentes. Ceci est d'une importance majeure pour pouvoir prédire l'état d'un système à un instant ultérieur. Comme par exemple la prédiction de la trajectoire d'une particule à partir de l'instant courant.*

Nous rappelons que pour des fonctions continues $f : \mathcal{D} \subset \mathbb{R} \longmapsto \mathbb{R}\ \forall\ p \geq 1$, *les normes p sont définies par :*

$$\| f \|_p = \left[\int_{t \in \mathcal{D}} |f(t)|^p\, dt \right]^{1/p} \tag{10}$$

En l'occurrence la norme Euclidienne $p = 2$ *est définie par :*

$$\| f \|_2 = \left[\int_{t \in \mathcal{D}} |f(t)|^2\, dt \right]^{1/2} \tag{11}$$

La norme infinie s'écrit :

$$\| f \|_\infty = \sup_{t \in \mathcal{D}} |f(t)| = \lim_{p \longrightarrow +\infty} |f(t)|_p \tag{12}$$

II. Équation aux dérivées partielles

Une équation aux dérivées partielles (EDP en abrégé) exprime une relation fonctionnelle entre les variables indépendantes $X = (x, x, z, \cdots)$, la fonction inconnue $u(x, y, z, \cdots) \in \Omega \subset \mathbb{R}$ et ses dérivées partielles. La formulation générale d'une telle relation s'écrit suivant :

$$\sum_{k=0}^{n} a_k(x, y, z, \cdots) \cdot D^k u(x, y, z, \cdots) = \phi(x, y, z, \cdots), \qquad D^k \equiv \frac{\partial^k}{\partial X^k} \tag{13}$$

Ou de façon équivalente,

$$\mathcal{F}\left[D^n u(x, y, z, \cdots), D^{n-1} u(x, y, z, \cdots), \cdots, D^1 u(x, y, z, \cdots), u(x, y, z, \cdots), x, y, z, \cdots\right] = \phi(x, y, z, \cdots) \tag{14}$$

Où $\mathcal{F}$ exprime une relation fonctionnelle entre les variables indépendantes $X = (x, x, z, \cdots)$, la fonction inconnue $u(x, y, z, \cdots)$ et ses dérivées partielles. On appelle *ordre* d'une EDP l'ordre de la plus grande dérivée présente dans l'équation. On appelle *problème aux limites*, une EDP munie de conditions aux limites [1] sur la totalité de la frontière du domaine sur lequel elle est posée. En se limitant aux EDPs de l'ordre deux ($k = 2$), l'EDP exprimée par (13) est dite *linéaire* car les coefficients $a_k(x, y) \neq f(u)$ et également la source $\phi(x, y) \neq f(u)$. Cette linéarité peut s'exprimer sous la forme : si $u_1(x, y)$ et $u_2(x, y)$ sont solutions de (13) alors :

$$\mathcal{F}\left[a_1\, u_1(x, y) + a_2\, u_2(x, y)\right] = a_1 \mathcal{F}\left[u_1(x, y)\right] + a_2 \mathcal{F}\left[u_2(x, y)\right] \tag{15}$$

De la même façon, $\forall \lambda \in \mathbb{R}$,

$$\mathcal{F}\left[\lambda\, u(x, y)\right] = \lambda\, \mathcal{F}\left[u(x, y)\right] \tag{16}$$

Une EDP est dite *semi-linéaire* si elle est écrite sous la forme :

$$\sum_{k \leq 2} a_k(x, y) \cdot D^{(2)} u(x, y) = \phi(D^{(1)} u(x, y), x, y) \tag{17}$$

Une EDP est dite *quasi-linéaire* si elle est écrite sous la forme :

$$\sum_{k \leq 2} a_k(D^{(1)} u(x, y), u(x, y), x, y) \cdot D^{(2)} u(x, y) = \phi(D^{(1)} u(x, y), x, y) \tag{18}$$

Elle est linéaire seulement par rapport aux dérivées partielles de $u(x, y)$ d'ordre deux. Une EDP linéaire d'ordre deux est explicitée sous la forme :

$$\begin{aligned} A(x, y)\frac{\partial^2 u(x, y)}{\partial x^2} + B(x, y)\frac{\partial^2 u(x, y)}{\partial x\, y} + C(x, y)\frac{\partial^2 u(x, y)}{\partial y^2} + \\ D(x, y)\frac{\partial u(x, y)}{\partial x} + E(x, y)\frac{\partial u(x, y)}{\partial y} + F(x, y) u(x, y) = \phi(x, y) \end{aligned} \tag{19}$$

Les coefficients peuvent êtres égaux à des constantes et A, B, C sont obligatoirement différents de zéro, sinon on retombe sur une EDP d'ordre un. Notons que cette EDP est *inhomogène* car le second membre est non nul, elle est dite *homogène* dans le cas contraire. Les EDPs linéaires d'ordre deux sont classées selon la valeur du discriminant Δ, soit :

$$\Delta = B(x, y)^2 - 4\, A(x, y)\, C(x, y) : \begin{cases} \Delta = 0 \Rightarrow \text{EDP parabolique} \\ \Delta > 0 \Rightarrow \text{EDP hyperbolique} \\ \Delta < 0 \Rightarrow \text{EDP elliptique} \end{cases} \tag{20}$$

1. Le lecteur est invité à distinguer les conditions dites de *Dirichlet* (ocndition imposée sur la valeur de la solution à la frontière), de *Neuman* (condition imposée sur la valeur de la dérivée de la solution) et de *Dirichle-Neuman* (condition mixte). La précision globale du schéma dépend de la précision sur la discrétisation des conditions aux limites.

Il est possible de démontrer que cette classification est invariante par changement de bases dans le plan. A titre informatif, les équations de Poisson et de Laplace sont *elliptiques*, celle de la chaleur est *parabolique*, tandis que les équations des cordes vibrantes sont *hyperboliques*.

A. *Résolution analytique de quelques EDPs*

Dans ce qui suit, nous nous proposons de résoudre analytiquement *l'équation des ondes* ainsi que l'équation maitresse de la théorie quantique, à savoir *l'équation de Schrödinger*. *L'équation des ondes* est résolue en utilisant la méthode de *séparation de variables* (ou méthode de Fourier) et *l'équation de Schrödinger* est résolue en se servant des *séries entières*.

1) ***Établissement de l'équation de Schrödinger****:* cette équation joue un rôle fondamental en mécanique quantique car sa résolution permet la description des propriétés de la matière à des échelles atomiques. L'atome d'hydrogène est l'un des rares systèmes réalistes (communément appelé système à deux corps) qui peuvent réellement être résolus de manière analytique. La géométrie sphérique de ce système suggère l'utilisation des coordonnées sphériques avec le noyau à l'origine. Erwin Schrödinger et Werner Heisenberg ont ébauché séparément l'équation régissant la description et l'évolution des systèmes quantiques. Schrödinger a opté pour un formalisme mathématique utilisant les équations aux dérivées partielles, alors que Heisenberg a utilisé un formalisme matriciel. Bien que les deux approches se sont révélées mathématiquement équivalentes. La plupart des ouvrages débutent par l'équation de Schrödinger, qui semble avoir une meilleure interprétation physique par le biais de l'équation des ondes classique. En effet, l'équation de Schrödinger peut être vue comme une forme de l'équation des ondes appliquée aux ondes de matière. L'équation des ondes classique unidimensionnelle est donnée par :

$$\frac{\partial^2 u(x,t)}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 u(x,t)}{\partial t^2} \tag{21}$$

Avec les conditions aux limites :

$$u(0,t) = 0 \quad \text{et} \quad u(l,t) = 0 \tag{22}$$

Ces conditions stipulent que l'amplitude de vibration est nulle aux extrémités $x = 0$ et $x = l$. En séparant les variables spatiale et temporelle :

$$u(x,t) = \psi(x) f(t)$$

Nous obtenons,

$$\frac{1}{\psi(x)}\frac{d^2\psi(x)}{dx^2} = \frac{1}{v^2\, f(t)}\frac{d^2 f(t)}{dt^2} \tag{23}$$

Les deux termes sont égaux si et seulement s'ils valent la même constante, notée k. Cette dernière est arbitraire dans le sens où son signe est inconnu. Elle peut être négative, positive ou même nulle :

$$\begin{cases} \dfrac{1}{\psi(x)}\dfrac{d^2\psi(x)}{dx^2} = k \\[2ex] \dfrac{1}{v^2\, f(t)}\dfrac{d^2 f(t)}{dt^2} = k \end{cases} \quad \text{que nous écrirons sous la forme} \quad \begin{cases} \psi(x)'' - k\,\psi(x) = 0 \\[2ex] f(t)'' - k\,v^2\, f(t) = 0 \end{cases} \tag{24}$$

En utilisant la méthode de séparation des variables, nous sommes passés d'une équation aux dérivées partielles à deux équations différentielles ordinaires de second ordre sans second membre. La constante k

est appelée *constante de séparation*. Dans ce qui suit, nous résolvons l'équation différentielle dépendante de la variable spatiale :

$$\psi(x)^{''} - k\,\psi(x) = 0 \tag{25}$$

Afin d'atteindre cet objectif, nous envisageons trois cas de figure pour la constante de séparation k.

o Cas où $k = 0$

$$k = 0 \Rightarrow \psi(x)^{''} = 0 \Rightarrow \psi(x)^{'} = a_1 \Rightarrow \psi(x) = a_1\,x + b_1 \tag{26}$$

Avec a_1 et b_1 sont des constantes d'intégration pouvant êtres déterminées tenant compte des conditions aux limites (22) :

$$\left\{\begin{array}{l} u(0,t) = \psi(0)\,f(t) = 0 \\ \\ u(l,t) = \psi(l)\,f(t) = 0 \end{array}\right. \qquad f(t) \neq 0 \;\Rightarrow \left\{\begin{array}{l} \psi(0) = 0 \\ \\ \psi(l) = 0 \end{array}\right. \tag{27}$$

A partir de (26) on obtient :

$$\psi(x=0) = b_1 = 0 \quad \text{et} \quad \psi(x=l) = a_1\,l + 0 = 0 \Rightarrow a_1 = b_1 = 0 \Rightarrow \psi(x) = 0$$

Autrement dit,

$$\psi(x) = 0 \Rightarrow u(x,t) = 0, \quad \forall x \in [0 \;\; l]$$

Cette solution est mathématiquement juste mais physiquement inacceptable dans le sens où elle n'apporte aucune information pertinente sur le mouvement ondulatoire de l'équation (21). La solution $u(x,t) = 0$ signifie qu'il n'existe aucun mouvement ondulatoire ! c'est une solution dite triviale. Ce qui nous amène à dire :

Toutes les solutions physiquement acceptables sont solutions de l'équation (21)*, mais toutes les solutions de* (21) *ne sont pas physiquement acceptables.*

o Cas où $k > 0$, posons $k = \beta^2,\ \beta \in \mathbb{R}$

Posons $\psi^{''} = \lambda^2$ (donc $\psi^{'} = \lambda^1$ et $\psi = \lambda^0 = 1$) et écrivons le polynôme caractéristique de l'équation (25) qui devient :

$$\lambda^2 - \beta^2 = 0 \Rightarrow \lambda_{1,2} = \pm\beta \tag{28}$$

La solution générale de l'équation (25), pour k positif, prend la forme :

$$\psi(x) = c_1\,e^{\lambda_1\,x} + c_2\,e^{\lambda_2\,x} = c_1\,e^{\beta\,x} + c_2\,e^{-\beta\,x} \tag{29}$$

Appliquons les conditions aux limites de (22), il vient :

$$\psi(x=0) = c_1 + c_2 = 0 \Rightarrow c_2 = -c_1$$

$$\psi(x=l) = c_1\,e^{\beta\,l} + c_2\,e^{-\beta\,l} = c_1\,(e^{\beta\,l} + e^{-\beta\,l}) = 0 \Rightarrow c_1 = 0 \Rightarrow c_2 = 0$$

De façon analogue que précédemment, nous obtenons une solution triviale !. Intéressons-nous désormais au dernier cas.

o Cas où $k < 0$, posons $k = -\beta^2,\ \beta \in \mathbb{R}$

Écrivons le polynôme caractéristique de l'équation (25) qui devient :

$$\lambda^2 + \beta^2 = 0 \Rightarrow \lambda_{1,2}^2 = -\beta^2 \Rightarrow \lambda_{1,2}^2 = j^2\,\beta^2 \Rightarrow \lambda_{1,2} = \pm j\,\sqrt{\beta} \tag{30}$$

La solution générale de l'équation (25), pour k négatif, prend la forme :

$$\psi(x) = c_1\,e^{\lambda_1\,x} + c_2\,e^{\lambda_2\,x} = c_1\,e^{j\,\beta\,x} + c_2\,e^{-j\,\beta\,x} \tag{31}$$

Afin de simplifier les calculs, écrivons l'équation (31) sous forme d'une combinaison de fonctions sinusoïdales. Utilisons pour cela la formule d'Euler :

$$e^{\pm j\,\theta} = \cos(\theta) \pm j\,\sin(\theta) \tag{32}$$

A partir de l'équation (31) :

$$\begin{aligned}\psi(x) &= c_1\,\cos(\beta\,x) + c_1\,j\,\sin(\beta\,x) + c_2\,\cos(\beta\,x) - c_2\,j\,\sin(\beta\,x)\\ &= \underbrace{(c_1 + c_2)}_{c_\alpha \in \mathbb{R}}\,\cos(\beta\,x) + \underbrace{(c_1\,j - j\,c_2)}_{c_\beta \in \mathbb{C}}\,\sin(\beta\,x)\end{aligned}$$

$$\Rightarrow \psi(x) = c_\alpha\,\cos(\beta\,x) + c_\beta\,\sin(\beta\,x) \tag{33}$$

Comme précédemment, les constantes c_α et c_β sont déterminées à partir des conditions aux limites (22), soit :

$$\psi(x = 0) = c_\alpha = 0$$

$$\psi(x = l) = 0 = c_\alpha\,\cos(\beta\,l) + c_\beta\,\sin(\beta\,l)$$

$$\Rightarrow \psi(x = l) = c_\beta\,\sin(\beta\,l) = 0 \tag{34}$$

Cette équation est nulle dans deux cas de figure. D'abord si $c_\beta = 0$ dans ce cas il en résulte $c_\alpha = c_\beta = 0 \Rightarrow \psi(x) = 0$ c'est une solution triviale qui n'est pas intéressante d'un point de vue physique. Ensuite, la deuxième condition si :

$$c_\beta \neq 0 \Rightarrow \sin(\beta\,l) = 0 \Rightarrow \beta\,l = n\,\pi, \quad \text{avec} \quad n \in \mathbb{N}^* \tag{35}$$

Ainsi, la solution de l'équation (25) pour $k < 0$ prend la forme :

$$\psi(x) = c_\beta\,\sin\left(\frac{n\,\pi\,x}{l}\right) \quad \text{avec} \quad n \in \mathbb{N}^* \tag{36}$$

Cette solution décrit l'amplitude spatiale de l'onde de matière en fonction de la position. Résolvons désormais $f(t)'' - k\,v^2\,f(t) = 0$ pour $k = -\beta^2$, avec $\beta \in \mathbb{R}$ soit :

$$f(t)'' + \beta^2\,v^2\,f(t) = 0 \tag{37}$$

De manière analogue que précédemment, au moyen du polynôme caractéristique nous obtenons la solution générale :

$$f(t) = c_1\,e^{\lambda_1\,t} + c_2\,e^{\lambda_2\,t} = c_1\,e^{j\,\beta\,v\,t} + c_2\,e^{-j\,\beta\,v\,t} \tag{38}$$

Les termes de la solution (38) sont oscillatoires, par conséquent la quantité $v\,\beta$ doit forcément valoir les dimensions d'une pulsation w soit :

$$\Rightarrow f(t) = c_1\, e^{j\,\omega\, t} + c_2\, e^{-j\,\omega\, t} \tag{39}$$

Qui peut s'écrire également sous la forme équivalente :

$$f(t) = A\, \cos(\omega\, t + \phi) \tag{40}$$

Tenant compte des solutions (36) et (40), la solution de l'équation (21) devient :

$$u(x,t) = \sum_{n=1}^{\infty} A_n\, \cos(\omega_n\, t + \phi_n)\, \sin\left(\frac{n\,\pi\, x}{l}\right) \tag{41}$$

Ce n'est pas cette solution qui nous intéresse dans ce cas précis. Elle est donnée à titre informatif. Le but est d'obtenir une solution générale de $f(t)$ une fois la nature (positive, négative ou nulle) de la constate de séparation k est connue. Désormais nous pouvons écrire :

$$u(x,t) = \psi(x)\, A\, \cos(\omega\, t + \phi) \tag{42}$$

En substituant (42) dans (23) :

$$\frac{1}{\psi(x)}\, \frac{d^2\psi(x)}{dx^2} = \frac{-A\, w^2\, \cos(\omega\, t + \phi)}{v^2\, A\, \cos(\omega\, t + \phi)} \tag{43}$$

$$\Rightarrow \psi''(x) + \frac{w^2}{v^2}\, \psi(x) = 0 \tag{44}$$

Par ailleurs, l'énergie totale d'une particule est la somme des parties cinétique et potentielle soit :

$$E = \frac{p^2}{2m} + V(x) \tag{45}$$

En tirant la quantité de mouvement p :

$$p = \{2m[E - V(x)]\}^{1/2} \tag{46}$$

En utilisant la formule de de Broglie pour la longueur d'onde :

$$\lambda = \frac{h}{p} = \frac{h}{\{2m[E - V(x)]\}^{1/2}} \tag{47}$$

Le terme ω^2/v^2 peut être réécrit en fonction de λ, nous rappelons que $\omega = 2\pi\nu$ et $\nu\lambda = v$.

$$\frac{\omega^2}{v^2} = \frac{4\pi^2\nu^2}{v^2} = \frac{4\pi^2}{\lambda^2} = \frac{2m[E - V(x)]}{\hbar^2} \tag{48}$$

En substituant ce dernier résultat dans l'équation (44), nous obtenons la fameuse équation de Schrödinger indépendante du temps :

$$\frac{d^2\psi(x)}{dx^2} + \frac{2m}{\hbar^2}[E - V(x)]\psi(x) = 0 \tag{49}$$

qui est presque toujours écrite sous la forme :

$$-\frac{\hbar^2}{2m}\, \frac{d^2\psi(x)}{dx^2} + V(x)\psi(x) = E\psi(x) \tag{50}$$

Cette équation unidimensionnelle à une seule particule peut facilement être étendue au cas tridimensionnel :

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r})+V(\mathbf{r})\psi(\mathbf{r})=E\psi(\mathbf{r}) \tag{51}$$

Cette équation peut également traiter un problème à deux corps en remplaçant m par une masse réduite $\mu=\dfrac{m_2\,m_1}{m_1+m_2}$. Soulignons que Schrödinger a d'abord présenté son équation indépendante du temps, ensuite il a postulé l'équation plus générale dépendante du temps.

2) ***Équation dépendante du temps****:* examinons désormais l'équation de Schrödinger dépendante du temps. Dans la section précédente, l'équation de Schrödinger indépendante du temps d'une particule a été déterminée à partir de l'équation des ondes classique et de la relation de de Broglie. En revanche, l'équation de Schrödinger dépendante du temps ne peut être obtenue au moyen de méthodes élémentaires et est généralement donnée comme postulat de la mécanique quantique. L'équation de Schrödinger dépendante du temps à une seule particule est la suivante :

$$j\hbar\frac{\partial\psi(\mathbf{r},t)}{\partial t}=-\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r},t)+V(\mathbf{r})\psi(\mathbf{r},t) \tag{52}$$

Où V est supposé être une fonction réelle et représente l'énergie potentielle à laquelle est soumise la particule. Notons que l'équation (52) ne tient pas encore compte des effets de spin ou relativistes. Bien entendu, l'équation dépendante du temps peut être utilisée afin d'établir l'équation indépendante du temps. Si nous écrivons la fonction d'onde comme un produit de deux fonctions spatiale et temporelle, $\psi(\mathbf{r},t)=\psi(\mathbf{r})f(t)$, alors l'équation (52) devient :

$$\frac{j\hbar}{f(t)}\frac{df}{dt}=\frac{1}{\psi(\mathbf{r})}\left[\frac{\hbar^2}{2m}\nabla^2+V(\mathbf{r})\right]\psi(\mathbf{r}) \tag{53}$$

Puisque le terme de gauche de l'équation est dépendant uniquement du temps et le terme de droite dépend uniquement de l'espace, l'égalité de l'équation (53) est satisfaite dans le cas où les deux termes sont égaux à la même constante. Si nous désignons cette constante E (puisque le côté droit doit clairement avoir les dimensions de l'énergie), nous en obtenons deux équations différentielles ordinaires, à savoir :

$$\frac{1}{f(t)}\frac{df(t)}{dt}=-\frac{jE}{\hbar} \tag{54}$$

et

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r})+V(\mathbf{r})\psi(\mathbf{r})=E\psi(\mathbf{r}) \tag{55}$$

Cette dernière équation est celle de Schrödinger indépendante du temps. La solution de (54) est :

$$f(t)=e^{-j\,E\,t/\hbar}\quad\text{avec}\quad Re[e^{-j\,E\,t/\hbar}]=\cos(\omega\,t) \tag{56}$$

Nous retrouvons le résultat de $f(t)$ écrit pour le cas de l'équation de Schrödinger indépendante du temps. L'Hamiltonien de l'équation (55) est un opérateur hermitien et les valeurs propres d'un opérateur hermitien doivent être réelles, donc la constante E est réelle. Cela signifie que les solutions $f(t)$ sont purement oscillatoires, rappelons la formule d'Euler $e^{\pm i\theta}=cos(\theta)\pm i\ sin(\theta)$. Par voie de conséquence si :

$$\psi(\mathbf{r},t)=\psi(\mathbf{r})\,e^{-j\,E\,t/\hbar} \tag{57}$$

alors la fonction d'onde totale $\psi(\mathbf{r},t)$ diffère de $\psi(\mathbf{r})$ uniquement par un facteur de phase d'amplitude constante. Cela a des conséquences intéressantes. Tout d'abord, la quantité $|\psi(\mathbf{r},t)|^2$ est indépendante du temps, car nous pouvons

$$|\psi(\mathbf{r},t)|^2=\psi^*(\mathbf{r},t)\,\psi(\mathbf{r},t)=e^{jEt/\hbar}\,\psi^*(\mathbf{r},t)\,e^{-jEt/\hbar}\,\psi(\mathbf{r},t)=\psi^*(\mathbf{r})\psi(\mathbf{r})$$

Deuxièmement, la valeur attendue pour tout opérateur indépendant du temps est également indépendante du temps, si $\psi(\mathbf{r},t)$ satisfait l'équation (57). Par le même raisonnement :

$$\langle A\rangle = \int \psi^*(\mathbf{r},t)\,\hat{A}\,\psi(\mathbf{r},t) = \int \psi^*(\mathbf{r})\,\hat{A}\,\psi(\mathbf{r})$$

Pour ces raisons, les fonctions d'onde de la forme (57) sont appelées états stationnaires. L'équation (57) représente une solution particulière de l'équation (52). La solution générale de l'équation (52) serait une combinaison linéaire de ces solutions particulières :

$$\Psi(\mathbf{r},t) = \sum_i c_i\,\psi_i(\mathbf{r})\,e^{-jE_it/\hbar}$$

*B. **Solution exacte de l'équation de Schrödinger***

La résolution exacte de l'équation de Schrödinger en coordonnées cartésiennes est inextricable pour l'atome d'hydrogène ou les ions hydrogénoïdes [2] (He^+, Li^{2+}, $\cdots$ etc) à cause de la non séparabilité des variables. Cette difficulté est levée si l'on considère un système de coordonnées sphériques dont les variables sont séparables. Les coordonnées sphériques facilitent grandement la résolution exacte de l'équation de Schrödinger pour l'atome d'hydrogène et les ions hydrogénoïdes. Ainsi avant de rentrer dans le vif du sujet, nous commencerons par résumer les principaux résultats d'un tel système de coordonnées.

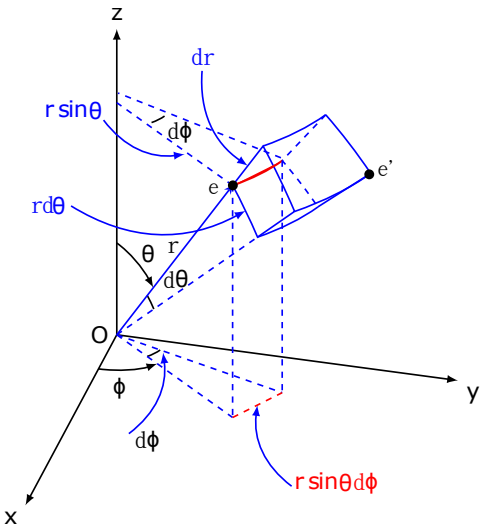


FIGURE 1: Représentation des coordonnées sphériques

Il faut bien garder à l'esprit que lors de l'intégration volumique, il est judicieux de savoir que l'élément de volume en question dépend du système de coordonnées. L'élément de volume forme un parallélépipède dont les arêtes quantifient les déplacements élémentaires obtenus lorsque l'on fait varier une seule des trois coordonnées. Dans le système de coordonnées sphériques, les déplacements élémentaires s'écrivent selon $dv = du\,dv\,dw$. Le déplacement de l'électron dans la direction radiale (r) conduit à une variation élémentaire $du = dr$. En gardant la distance radiale et l'angle azimutal (ϕ) fixes, l'abscisse curviligne générée par la variation de l'angle θ devient $v = r\,\theta \Rightarrow\ dv = r\,d\theta$. Avec un raisonnement analogue, nous obtenons pour le dernier déplacement élémentaire $dw = r\,sin(\theta)\,d\phi$. Ce qui donne comme élément de volume :

$$dv = r^2\,sin(\theta)\,dr\,d\theta\,d\phi \tag{58}$$

2. Atomes ayant une structure électronique semblable à celle de l'atome d'hydrogène.

Il convient de remarquer également que cet élément de volume n'est pas constant, car il dépend de la distance radiale et de l'angle θ. L'écriture de l'équation de Schrödinger d'un tel système donne :

$$\left(-\frac{\hbar^2}{2\,m}\,\nabla^2 + V\right)\,\psi = E\,\psi \tag{59}$$

Où ∇^2 est l'opérateur Laplacien et V est l'énergie potentielle électrostatique ou *Coulombienne* qui est donnée par :

$$V(r) = -\frac{1}{4\,\pi\,\epsilon_0}\,\frac{Z\,e^2}{r} \tag{60}$$

Où ϵ_0 est la permittivité du vide (pas besoin d'une permittivité relative car l'espace à l'intérieur de l'atome est "vide"), les charges e et Ze sont respectivement celles de l'électron et du noyau, pour l'hydrogène et les ions hydrogénoïdes le nombre d'électron $Z = 1$. La distance radiale, r, décrit l'éloignement de l'électron par rapport au noyau. L'énergie potentielle *Coulombienne* est inversement proportionnelle à la distance entre l'électron et le noyau et ne dépend d'aucun angle. Un tel potentiel est appelé *potentiel central*. La solution exacte de l'équation de Schrödinger pour l'atome d'hydrogène et les ions hydrogénoïdes est obtenue sous la forme :

$$\psi(r,\theta,\phi) = \underbrace{R_{n,l}(r)}_{\text{taille de l'orbitale}} \times \underbrace{Y_l^m(\theta,\phi)}_{\text{forme de l'orbitale}} \tag{61}$$

En utilisant le Laplacien en coordonnées sphériques, l'équation de Schrödinger devient :

$$-\frac{\hbar^2}{2\,m}\underbrace{\left[\frac{1}{r^2}\,\frac{\partial}{\partial r}\left(r^2\,\frac{\partial}{\partial r}\right) + \frac{1}{r^2\sin\theta}\,\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\,\frac{\partial^2}{\partial\phi^2}\right]}_{\nabla^2}\psi + V(r)\psi = E\,\psi \tag{62}$$

$$\Rightarrow \frac{1}{r^2}\,\frac{\partial}{\partial r}\left(r^2\,\frac{\partial\psi}{\partial r}\right) + \frac{1}{r^2\sin\theta}\,\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\,\frac{\partial^2\psi}{\partial\phi^2} + \frac{2\,m}{\hbar^2}$$

$$\times\left(E + \frac{Z\,e^2}{4\pi\,\epsilon_0\,r}\right)\psi = 0 \tag{63}$$

En utilisant la méthode de séparation des variables, nous considérons une solution $\psi_{n,l,m}(r,\theta,\phi)$ s'écrivant comme un produit d'une fonction radiale $R_{n,l}(r)$ et d'une fonction angulaire $Y_{l,m}(\theta,\phi)$:

$$\psi_{n,l,m}(r,\theta,\phi) = R_{n,l}(r)\times Y_{l,m}(\theta,\phi) \tag{64}$$

Ce qui donne :

$$\Rightarrow \frac{Y}{r^2}\,\frac{\partial R}{\partial r}\left(r^2\,\frac{\partial R}{\partial r}\right) + \frac{R}{r^2\sin\theta}\,\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial Y}{\partial\theta}\right) + \frac{R}{r^2\sin^2\theta}\,\frac{\partial^2 Y}{\partial\phi^2} + \frac{2\,m}{\hbar^2}$$

$$\times\left(E + \frac{Z\,e^2}{4\pi\,\epsilon_0\,r}\right)RY = 0 \tag{65}$$

Désormais nous multiplions par r^2 et divisons par RY afin de séparer la variable radiale et les variables angulaires :

$$\Rightarrow \frac{1}{R}\,\frac{d}{dr}\left(r^2\,\frac{dR}{dr}\right) + \frac{1}{Y\sin\theta}\,\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial Y}{\partial\theta}\right) + \frac{1}{Y\sin^2\theta}\,\frac{\partial^2 Y}{\partial\phi^2} + \frac{2\,m\,r^2}{\hbar^2}$$

$$\times\left(E + \frac{Z\,e^2}{4\pi\,\epsilon_0\,r}\right) = 0 \tag{66}$$

$$\Rightarrow \underbrace{\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right)+\frac{2\,m\,r^2}{\hbar^2}\left(E+\frac{Z\,e^2}{4\pi\,\epsilon_0\,r}\right)}_{\text{Partie radiale}}+$$

$$\underbrace{\frac{1}{Y\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial Y}{\partial\theta}\right)+\frac{1}{Y\sin^2\theta}\frac{\partial^2 Y}{\partial\phi^2}}_{\text{Partie angulaire}}=0 \tag{67}$$

Les deux parties s'annulent dans le cas où les deux termes (radial et angulaire) sont égaux à la même constante mais de singe opposé. La constante choisie est connue sous le nom de *constante de séparation*, notons cette constante K. Ainsi nous obtenons les deux équations différentielles suivantes :

$$\Rightarrow \underbrace{\frac{d}{dr}\left(r^2\frac{dR}{dr}\right)+\frac{2\,m\,r^2}{\hbar^2}\left(E+\frac{Z\,e^2}{4\pi\,\epsilon_0\,r}\right)R-K\,R}_{\text{Partie radiale}}=0 \tag{68}$$

$$\Rightarrow \underbrace{\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial Y}{\partial\theta}\right)+\frac{1}{\sin^2\theta}\frac{\partial^2 Y}{\partial\phi^2}+K\,Y}_{\text{Partie angulaire}}=0 \tag{69}$$

1) ***Résolution de la partie angulaire****:* La partie angulaire contient encore des termes dépendant à la fois de θ et ϕ. Une autre séparation des variables est nécessaire. Remplaçons la fonction angulaire $Y(\theta,\phi)$ par le produit :

$$Y(\theta,\phi)=f(\theta)\times g(\phi) \tag{70}$$

$$\Rightarrow \frac{g}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{df}{d\theta}\right)+\frac{f}{\sin^2\theta}\frac{d^2g}{d\phi^2}+K\,f\,g=0 \tag{71}$$

En isolant les deux variables :

$$\Rightarrow \frac{\sin\theta}{f}\frac{d}{d\theta}\left(\sin\theta\frac{df}{d\theta}\right)+K\,\sin^2\theta+\frac{1}{g}\frac{d^2g}{d\phi^2}=0 \tag{72}$$

De la même façon que précédemment notons B la constante de séparation, nous obtenons les deux équations différentielles suivantes :

$$\Rightarrow \underbrace{\frac{\sin\theta}{f}\frac{d}{d\theta}\left(\sin\theta\frac{df}{d\theta}\right)+K\,\sin^2\theta}_{\text{Partie polaire}}-B=0 \tag{73}$$

$$\Rightarrow \underbrace{\frac{1}{g}\frac{d^2g}{d\phi^2}}_{\text{Partie azimutale}}+B=0 \tag{74}$$

La solution générale de la partie azimutale est donnée par :

$$g(\phi)=c_1\,e^{\lambda_1\,\phi}+c_2\,e^{\lambda_2\,\phi} \tag{75}$$

La solution de la partie azimutale s'obtient en écrivant le polynôme caractéristique de l'équation différentielle en question :

$$\lambda^2+B=0\Leftrightarrow\lambda_{1,2}^2=j^2\,B\Rightarrow\lambda_{1,2}=\pm\,j\,\sqrt{B} \tag{76}$$

Il est clair que la constante B doit être positive. Notons cette constante m^2 (donc $B = m^2$) $\Rightarrow \lambda_1 = +j\,m$ et $\lambda_2 = -j\,m$. Ainsi la solution générale prend la forme :

$$g(\phi) = c_1\, e^{j\,m\,\phi} + c_2\, e^{-j\,m\,\phi} \tag{77}$$

L'angle ϕ est l'azimut, c'est-à-dire que si nous considérons l'atome comme un globe, alors c'est la longitude de la position de l'électron. Nous pouvons choisir le *méridien de Greenwich* de l'atome d'une manière mathématiquement commode en fixant $c_2 = 0$. En terminologie quantique, m est appelé un nombre quantique [3] car il limite les valeurs possibles de la fonction d'onde (et donc des observables) à des multiples entiers.

$$g_m(\phi) = c_1\, e^{j\,m\,\phi} \tag{78}$$

L'indice m est ajouté à $g_m(\phi)$ car il est désormais clair qu'il existe autant de solutions qu'il existe des valeurs autorisées de m. La condition de périodicité de l'angle ϕ impose :

$$g_m(\phi) = c_1\, e^{j\,m\,\phi} = c_1\, e^{j\,m\,(\phi+2\pi)} = c_1\, e^{j\,m\,\phi}\, e^{j\,m\,2\pi} \Rightarrow\ e^{j\,m\,2\pi} = 1 \tag{79}$$

$$\Rightarrow \quad m = 0, \pm 1, \pm 2, \pm 3, \cdots \tag{80}$$

Considérons désormais la partie polaire dont l'équation différentielle s'écrit :

$$\frac{\sin\theta}{f}\,\frac{d}{d\theta}\left(\sin\theta\frac{df}{d\theta}\right) + K\,\sin^2\theta - m^2 = 0 \tag{81}$$

Qui se réarrange :

$$\frac{1}{\sin\theta}\,\frac{d}{d\theta}\left(\sin\theta\frac{df}{d\theta}\right) + \left(K\,-\frac{m^2}{\sin^2\theta}\right)\,f = 0 \tag{82}$$

Posons $x = \cos\theta \Rightarrow \dfrac{d}{d\theta} = \dfrac{dx}{d\theta}\,\dfrac{d}{dx} = -\sin\theta\,\dfrac{d}{dx}$ alors l'équation précédente devient :

$$\frac{1}{\sin\theta}\,(-\sin\theta)\,\frac{d}{dx}\left(\sin\theta(-\sin\theta)\frac{df}{dx}\right) + \left(K\,-\frac{m^2}{\sin^2\theta}\right)\,f = 0 \tag{83}$$

En exploitant la relation trigonométrique $\sin^2\theta + \cos^2\theta = 1 \Rightarrow \sin^2\theta = 1 - \cos^2\theta = 1 - x^2$. Il en résulte :

$$\frac{d}{dx}\left((1-x^2)\,\frac{df}{dx}\right) + \left(K\,-\frac{m^2}{1-x^2}\right)\,f = 0 \tag{84}$$

En différentiant le premier terme de l'équation, nous obtenons l'expression finale :

$$(1-x^2)\,\frac{d^2f}{dx^2} - 2\,x\frac{df}{dx} + \left(K\,-\frac{m^2}{1-x^2}\right)\,f = 0 \tag{85}$$

La formule finale de cette équation différentielle (n'oublions pas que nous avons déjà posé $x = \cos\theta$) est bien connue dans la littérature mathématique. Les solutions de cette équation sont obtenues pour $K = l(l+1)$ et sont connues sous le nom de polynômes de Legendre :

$$f_{l,m}(\theta) = (-1)^m\,\sqrt{\frac{(2\,l+1)!\,(l-m)!}{4\,\pi\,(l+m)!}} \times P_{l,m}(\cos\theta) \quad \text{pour } m \geq 0 \quad \text{et } n \geq l+1 \tag{86}$$

Avec,

$$P_{l,m}(\cos\theta) = (-1)^m\,\sqrt{(1-\cos^2\theta)^m}\,\frac{d^m}{d\cos^m\theta}\,P_l(\cos\theta) \tag{87}$$

3. C'est le nombre quantique azimutal

Où,

$$P_l(\cos\theta) = \frac{(-1)^l}{2^l\,!}\,\frac{d^l}{d\cos^l\theta}(1-\cos^2\theta)^l \tag{88}$$

Les polynômes associés de Legendre $P_{l,m}(\cos\theta)$ sont construits à partir des polynômes de Legendre $P_l(\cos\theta)$. La solution finale [4] de la partie angulaire s'écrit comme le produit des solutions des parties polaire et azimutale soit :

$$Y_{l,m}(\theta,\phi) = f_{l,m}(\theta)\times e^{j\,m\,\phi} \tag{89}$$

Dans la littérature mathématique les solutions $Y_{l,m}(\theta,\phi)$ sont appelées *harmoniques sphériques* [5]. Ces fonctions sont particulièrement utiles pour résoudre des problèmes invariants par rotation. Pour résumer cette section, en utilisant la méthode de séparation des variables grâce aux coordonnées sphériques, nous avons séparé l'équation angulaire en deux parties azimutale (dépendant de ϕ) et polaire (dépendant de θ). Les solutions de l'équation de l'angle azimutal sont des exponentielles incluant le nombre quantique magnétique m comme argument. Les solutions de l'équation de l'angle polaire sont les polynômes associés de Legendre, qui sont différents pour chaque choix du nombre quantique azimutal l et de nombre quantique magnétique m. Les deux nombres quantiques sont introduits dans les équations différentielles respectives en tant que constantes de séparation.

2) ***Résolution de la partie radiale****:* dans ce qui suit, nous développerons une approche étape par étape afin de résoudre la partie radiale de l'équation de Schrödinger pour l'atome d'hydrogène et les hydrogénoïdes. Les énergies propres négatives de l'Hamiltonien sont recherchées comme solution, car elles représentent les états liés de l'atome. Nous avons déjà obtenu pour la partie radiale l'expression suivante :

$$\frac{d}{dr}\left(r^2\,\frac{dR}{dr}\right) - \frac{2\,m\,r^2}{\hbar^2}\,(V(r)-E)\;R = l(l+1)\,R \tag{90}$$

Où la constante de séparation $K = l(l+1)$ a été obtenue pour la partie angulaire. Ci-dessous les étapes détaillées de la résolution de la partie radiale.

❶ ☞ : Nous devons d'abord simplifier l'équation radiale pour faciliter la résolution de l'équation différentielle. Les sous-étapes suivantes utilisent la technique des substitutions pour créer une équation différentielle résoluble.

$$\text{Posons } u(r) = r\,R(r) \Rightarrow R(r) = \frac{u(r)}{r} \Rightarrow \frac{dR(r)}{dr} = \left(r\frac{du(r)}{dr} - u\right)\times\frac{1}{r^2}$$

$$\frac{-\hbar^2}{2\,m}\frac{d^2u(r)}{dr^2} + \left[-\frac{e^2}{4\pi\,\epsilon_0\,r} + \frac{\hbar^2}{2\,m}\,\frac{l(l+1)}{r^2}\right]u = E\,u \tag{91}$$

$$\text{Posons } \gamma^2 = \frac{-2\,m\,E}{\hbar} \text{ il vient :}$$

$$\frac{1}{\gamma^2}\frac{d^2u(r)}{dr^2} = \left[1 - \frac{m\,e^2}{2\pi\,\epsilon_0\,\hbar^2\gamma}\times\frac{1}{\gamma\,r} + \frac{l(l+1)}{\gamma^2\,r^2}\right]u \tag{92}$$

$$\text{Posons } \rho = \gamma\,r \text{ et } \rho_0 = \frac{m\,e^2}{2\pi\,\epsilon_0\,\hbar^2\gamma} \text{ nous obtenons :}$$

$$\frac{d^2u(r)}{d\rho^2} = \left[1 - \frac{\rho_0}{\rho} + \frac{l(l+1)}{\rho^2}\right]u \tag{93}$$

❷ ☞ : Maintenant que l'équation est sous une forme appropriée pour la solution. Cette étape consiste à identifier les points singuliers. Il existe des points singuliers où la fonction d'onde doit tendre vers zéro. Dans

4. Pour $m<0$ nous avons $Y_{l,-m}(\theta,\phi) = Y^*_{l,m}(\theta,\phi)$

5. Ces harmoniques sphériques sont normalisées, ce qui explique la disparition de la constante d'intégration c_1 obtenue pour la partie azimutale

ce cas, la fonction d'onde doit s'annuler au centre de l'atome, donc pour $r = 0$. Elle doit également s'annuler à une distance relativement "grande" du noyau, prise comme $r \longrightarrow \infty$. Chaque point singulier doit être considéré individuellement. A ce stade nous étudierons donc le comportement asymptotique de $u(\rho)$:

– Pour $\rho \longrightarrow \infty \Rightarrow \dfrac{\rho_0}{\rho} \longrightarrow 0$ et $\dfrac{l(l+1)}{\rho^2} \longrightarrow 0$. Comme la distance r de l'atome va à l'infini, ces termes tendent vers zéro et sont donc sans importance pour cette partie de la solution. Ainsi, l'équation différentielle à résoudre, sous la condition d'une distance infinie, devient :

$$\Rightarrow \quad \frac{d^2u(r)}{d\rho^2} = u \tag{94}$$

La solution générale de l'équation différentielle ainsi obtenue est : $u(\rho) = A_1\, e^{-\rho} + A_2\, e^{\rho}$. Le deuxième terme de la solution e^{ρ} est refusé d'un point de vue physique. Car $\longrightarrow \infty \Rightarrow e^{\rho} \longrightarrow \infty$ or l'électron possède une limite spatiale par rapport au noyau. Il en ressort :

$$\Longrightarrow u(\rho) \simeq e^{-\rho} \tag{95}$$

– Le deuxième point singulier c'est quand $r \longrightarrow 0$ (donc au centre de l'atome) $\Rightarrow \dfrac{l(l+1)}{\rho^2} >> \dfrac{\rho_0}{\rho}$ et $\dfrac{l(l+1)}{\rho^2} >> 1$. Par conséquent, à ce deuxième point singulier, où r tend vers zéro, l'équation différentielle devient :

$$\Rightarrow \quad \frac{d^2u(r)}{d\rho^2} = \frac{l(l+1)}{\rho^2}\, u \tag{96}$$

La solution générale de cette équation différentielle est de la forme :

$$\Longrightarrow u(\rho) \simeq \rho^{l+1} \tag{97}$$

En combinant les solutions obtenues pour les deux points singuliers (95) et (97), la solution complète de l'équation (93) prend la forme :

$$\Longrightarrow u(\rho) = \rho^{l+1}\, e^{-\rho}\, y(\rho) \tag{98}$$

Où $y(\rho)$ est un polynôme exprimé sous forme :

$$y(\rho) = \sum_{j=0}^{\infty} c_j\, \rho^j \tag{99}$$

Nous nous sommes servi du comportement asymptotique de $u(\rho)$ pour trouver l'expression de $u(\rho)$ pour $0 < \rho < \infty$.

❸ ☞ : Après avoir obtenu une forme générale de la solution complète. Nous devons maintenant trouver une équation pour la partie polynomiale $y(\rho)$ de cette solution complète. Pour cela, calculons les dérivées première et second de l'équation (98) :

$$\frac{du(\rho)}{d\rho} = \rho^l\, e^{-\rho} \left[(\rho + 1 - 1)\, y + \rho \frac{dy(\rho)}{d\rho}\right] \tag{100}$$

et,

$$\frac{du(\rho)}{d\rho} = \rho^l\, e^{-\rho} \left[(-2\rho - 2 + l + \frac{l(l+1)}{l})\right] y + 2\,(l + 1 - \rho)\, \frac{dy(\rho)}{d\rho} + \rho\, \frac{d^2y(\rho)}{d\rho^2} \tag{101}$$

En substituant (100) et (101) dans (93) nous obtenons :

$$\rho\, \frac{d^2y(\rho)}{d\rho^2} + 2\,(l + 1 - \rho)\, \frac{dy(\rho)}{d\rho} + [\rho_0 - 2\,(l + 1)]\, y = 0 \tag{102}$$

En utilisant les séries entières [6], nous chercherons des solutions de $y(\rho)$ de la forme :

$$y(\rho) = \sum_{j=0}^{\infty} c_j \, \rho^j \tag{103}$$

Où les inconnus sont les coefficients c_j. Les dérivées de (103) se calculent selon :

$$\frac{dy(\rho)}{d\rho} = \sum_{j=0}^{\infty} j \, c_j \, \rho^{j-1} = \sum_{j=0}^{\infty} j + 1 \, c_{j+1} \, \rho^j \tag{104}$$

et,

$$\frac{d^2y(\rho)}{d\rho^2} = \sum_{j=0}^{\infty} j \, (j+1) \, c_{j+1} \, \rho^{j-1} \tag{105}$$

Substituons (104) et (105) dans l'équation (102) :

$$\rho \sum_{j=0}^{\infty} j \, (j+1) \, c_{j+1} \, \rho^{j-1} + 2(l+1) \sum_{j=0}^{\infty} (j+1) \, c_{j+1} \, \rho^j - 2\rho \sum_{j=0}^{\infty} j \, c_j \, \rho^{j-1}, +$$

$$[\rho_0 - 2(l+1)] \sum_{j=0}^{\infty} c_j \, \rho^j = 0 \tag{106}$$

Qui se simplifie :

$$\sum_{j=0}^{\infty} j \, (j+1) \, c_{j+1} \, \rho^j + 2(l+1) \sum_{j=0}^{\infty} (j+1) \, c_{j+1} \, \rho^j - 2 \sum_{j=0}^{\infty} j \, c_j \, \rho^j + [\rho_0 - 2(l+1)]$$

$$\times \sum_{j=0}^{\infty} c_j \, \rho^j = 0 \tag{107}$$

Nous n'avons pas encore fini avec le polynôme en question, car nous devons déterminer la relation de récurrence pour ses coefficients. Nous devons déterminer cette relation non seulement pour savoir comment le polynôme sera généré, mais aussi pour déterminer les limites de la sommation de la série. Le polynôme (107) vaut zéro si et seulement si $c_j = 0$, soit :

$$j \, (j+1) \, c_{j+1} \, + 2(l+1) \, (j+1) \, c_{j+1} \, - 2 \, j \, c_j \, + [\rho_0 - 2(l+1)] \, c_j \, = 0 \tag{108}$$

Nous obtenons la formule de récurrence suivante :

$$c_{j+1} = \frac{2 \, (j+l+1) - \rho_0}{j \, (j+1) + 2 \, (l+1) \, (j+1)} \, c_j = \frac{2 \, (j+l+1) - \rho_0}{(j+1) \, [j + 2 \, (l+1)]} \, c_j \tag{109}$$

Cette formule de récurrence décrit le comportement des coefficients de la série (du polynôme). Le polynôme $y(\rho)$ doit tendre vers zéro, donc le comportement du polynôme doit être examiné lorsque $j \longrightarrow \infty$. Cela fera l'objet de la prochaine étape.

❹ ☞ : Dans cette dernière étape de résolution de l'équation radiale, nous examinons le polynôme pour déterminer s'il est fini. Sinon, nous déterminons quelle condition est nécessaire pour le rendre fini. Cette condition de finitude produit un nombre quantique qui caractérise l'état du système et sert à quantifier

6. Le lecteur est renvoyé aux références :
Griffiths, Introduction to Quantum Mechanics, Prentice Hall, Englewood Cliffs, New Jersey, (1995), pp. 134-141. ET Cohen-Tannoudji, Diu, and Laloe, Quantum Mechanics, John Wiley & Sons, New York, (1977), pp. 794-797.

les énergies d'état lié de l'atome. En théorie, la formule (109) peut se développer à l'infinie, étudions son comportement quand $j \longrightarrow \infty$ (c'est-à-dire pour $j >> 1$) :

$$j >> 1 \Rightarrow c_{j+1} \simeq \frac{2\,j}{j\,(j+1)}\,c_j \Rightarrow c_{j+1} \simeq \frac{2}{(j+1)}\,c_j \tag{110}$$

$$\Rightarrow c_{j+2} \simeq \frac{2^2}{(j+1)\,(j+2)}\,c_j, \quad \cdots\cdots, \quad c_j \simeq \frac{2^j}{j!}c_0 \tag{111}$$

En substituant (111) dans (103) nous obtenons :

$$y(\rho) = c_0 \underbrace{\sum_{j=0}^{\infty} \frac{2^j}{j!}\,\rho^j}_{\text{Devlop. Taylor de } e^{2\rho}} \tag{112}$$

En substituant (112) dans (98) nous obtenons :

$$\Longrightarrow u(\rho) = c_0\,\rho^{l+1}\,e^{-\rho}e^{2\rho} \quad \Rightarrow \quad \lim_{\rho\to\infty} u(\rho) \longrightarrow \infty \tag{113}$$

La solution (113) diverge lorsque r est très grand. Ce comportement de la solution n'est pas acceptable car l'électron possède une limite spatiale finie par rapport à sa distance du noyau. Par conséquent, la série doit être tronquée à un nombre particulier, afin de forcer le polynôme d'avoir un comportement correct. Afin de déterminer le rang où doit se produire la troncature, nous fixons le coefficient du polynôme $y(\rho)$ égal à zéro à un nombre maximum, forçant la terminaison de la solution à de grandes distances du noyau. A partir de la formule de récurrence (109), nous avons :

$$c_{j_{max}+1} = 0 \quad \Rightarrow \quad 2\underbrace{(j_{max}+l+1)}_{n} - \rho_0 = 0 \quad \Rightarrow \quad \rho_0 = 2\,n \tag{114}$$

Où n est le nombre quantique principal. La nouvelle formule de récurrence s'obtient :

$$c_{j+1} = \frac{2\,(j+l+1) - 2\,n}{(j+1)\,[j+2\,(l+1)]}\,c_j \tag{115}$$

Appliquons cette relation par exemple pour $n = 3$ et $l = 1 \Rightarrow j_{max} = 1 \Rightarrow c_1 = -\frac{1}{2}\,c_0$:

$$\Rightarrow \quad y_n^l(\rho) = y_3^1(\rho) = c_0 - \frac{1}{2}\,c_0\,\rho = c_0\,(1 - \frac{1}{2}\,\rho) \tag{116}$$

On peut continuer ce processus à l'infini, mais on cherche une solution analytique à cette équation. La forme asymptotiquement suggérée nous donne un point de départ pour chercher la solution finale. Tenant compte de cette forme, nous soupçonnons une solution de la forme :

$$w(\rho) = \rho^{l+1} \times e^{-\rho} \tag{117}$$

$$w^{'} = -\rho^{l+1}\,e^{-\rho} + (l+1)\,\rho^l\,e^{-\rho} \tag{118}$$

$$w^{''} = e^{-\rho}\,\rho^{l+1} - (l+1)\,\rho^l\,e^{-\rho} + l\,(l+1)\,\rho^{l-1} - (l+1)\,\rho^l\,e^{-\rho} \tag{119}$$

Pour $l = 1$, nous obtenons :

$$w^{''} = e^{-\rho}\,\rho^2 - 2\,\rho\,e^{-\rho} + 2 - 2\,\rho\,e^{-\rho} \tag{120}$$

$$w^{''} = e^{-\rho}\,[\rho^2 - 2\,\rho + 2 - 2\,\rho] \tag{121}$$

$$w^{''} = e^{-\rho}\,[\rho^2 - 2\,\rho + 2 - 2\,\rho] \tag{122}$$

$$e^{\rho}w^{''} = [\rho^2 - 4\,\rho + 2] \tag{123}$$

En dérivant la dernière équation nous obtenons :

$$\frac{d[e^{\rho}\,w^{''}]}{d\rho} = 2\,\rho - 4 = \frac{1}{2}\,\rho - 1 \tag{124}$$

En comparant (116) à (124) nous déduisons :

$$y_3^1(\rho) = (-1)^3\ \frac{d[e^{\rho}\,w^{''}(\rho)]}{d\rho} \tag{125}$$

Cette dernière relation s'écrit également sous la forme :

$$y_3^1(\rho) = (-1)^3\ \left[\frac{d}{d\rho}\right]^3\ \left[e^{\rho}\left[\frac{d}{d\rho}\right]^2\ e^{-\rho}\,\rho^{l+1}\right] \tag{126}$$

Le deuxième terme de l'équation (126) n'est autre que le *polynômes associès de Laguerre* d'ordre $n = 3$, notée $L_3^2(2\,\rho)$. En définitif, la généralisation de ce résultat est immédiate et nous obtenons :

$$y(\rho) = L_{n-l-1}^{2l+1}(2\,\rho) \quad \forall\, n \geq l+1 \tag{127}$$

Nous rappelons que les polynômes associés de Laguerre sont définis par les relations suivantes :

$$L_p^q = (-1)^p\ \frac{d^p}{d\rho^p}\,e^{\rho}\ \left[\frac{d}{d\rho}\right]^q\ [e^{-\rho}\,\rho^q] \tag{128}$$

En combinant (98) et (127) nous obtenons la solution finale [7] :

$$u(\rho) = (2\,\rho)^{l+1}\,e^{-\rho}\,L_{n-l-1}^{2l+1}(2\,\rho) \quad \forall\, n \geq l+1 \tag{129}$$

Au début de la résolution de la partie radiale, nous avons déjà posé : $\gamma^2 = \dfrac{-2\,m\,E}{\hbar}$, $\rho = \gamma\,r$ et $\rho_0 = \dfrac{m\,e^2}{2\,\pi\,\epsilon_0\,\hbar^2\,\gamma} = 2n$.

$$\Rightarrow \quad E = \frac{\hbar^2\,\gamma^2}{2\,m} = -\frac{m\,e^4}{8\,\pi^2\,\epsilon_0^2\,\hbar^2\,\gamma_0^2} = -\frac{m\,e^4}{8\,\pi^2\,\epsilon_0^2\,\hbar^2\,(2\,n)^2} \tag{130}$$

$$\Rightarrow \quad E_n = -\frac{m}{2\,\hbar^2}\left[\frac{e^2}{4\,\pi\,\epsilon_0}\right]^2 \times \frac{1}{n^2} \tag{131}$$

D'un autre côté :

$$\Rightarrow \quad \gamma^2 = \frac{-m\,E}{\hbar^2} = \frac{2\,m^2\,e^4}{8\,\pi^2\,\epsilon_0^2\,\hbar^2\,(2n)^2} \quad \Rightarrow \quad \gamma = \underbrace{\frac{m\,e^2}{4\,\pi\,\epsilon_0\,\hbar}}_{a_0} \times \frac{1}{n} \quad \Rightarrow \quad \gamma = \frac{1}{a_0\,n}$$

Avec $a_0 = 0.53\,10^{-10}\,m$ est le rayon de Bohr. Revenons maintenant à la fonction radiale initiale $R(r)$ par le changement de variable que nous avons réalisé au début de notre résolution soit :

$$u(r) = r\,R(r) \quad \Rightarrow \quad R(r) = \frac{u(r)}{r} \quad \text{Où} \quad \rho = \gamma\,r = \frac{r}{a_0\,n} \tag{132}$$

$$R_{n,l}(r) = \frac{1}{r}\left(\frac{2\,r}{a_0\,n}\right)^{l+1}\ e^{-\left(\frac{r}{a_0\,n}\right)}\ L_{n-l-1}^{2l+1}\left(\frac{2\,r}{a_0\,n}\right) \tag{133}$$

7. Les polynômes associés de Laguerre, produits par la troncature de la série, sont identifiés par deux indices ou nombres quantiques, n et l. Les solutions physiquement acceptables exigent que n soit supérieur ou égal à $l+1$.

$$R_{n,l}(r) = \left(\frac{1}{r}\right)\left(\frac{2\,r}{n\,a_0}\right)\left(\frac{2\,r}{a_0\,n}\right)^{l} e^{-\left(\frac{r}{a_0\,n}\right)} L_{n-l-1}^{2l+1}\left(\frac{2\,r}{a_0\,n}\right) \tag{134}$$

$$\Rightarrow \quad R_{n,l}(r) = \left(\frac{2}{n\,a_0}\right)\left(\frac{2\,r}{a_0\,n}\right)^{l} e^{-\left(\frac{r}{a_0\,n}\right)} L_{n-l-1}^{2l+1}\left(\frac{2\,r}{a_0\,n}\right) \tag{135}$$

L'équation (135) est la solution finale non normalisée de l'équation radiale. Pour tenir compte de la normalisation :

$$\Rightarrow \quad R_{n,l}(r) = N_{n,l}\left(\frac{2\,r}{a_0\,n}\right)^{l} e^{-\left(\frac{r}{a_0\,n}\right)} L_{n-l-1}^{2l+1}\left(\frac{2\,r}{a_0\,n}\right) \tag{136}$$

Où le facteur $\left(\frac{2}{n\,a_0}\right)$ est adossé à la constante de normalisation $N_{n,l}$ qui s'obtient en calculant l'intégrale :

$$N_{n,l}^2 \int_0^{\infty} R_{n,l}(r)^*\, R_{n,l}(r)\, r^2\, dr = 1 \tag{137}$$

Où r^2 provient de l'élément de volume exprimé en coordonnées sphériques. Le calcul de cette constante étant très laborieux, nous donnons sa valeur :

$$N_{n,l} = \left[\left(\frac{2}{n\,a_0}\right)^3 \frac{(n-l-1)!}{2n\,[(n+1)!]^3}\right]^{1/2} \tag{138}$$

La solution normalisée de la partie radiale s'écrit alors :

$$\Rightarrow R_{n,l}(r) = \left[\left(\frac{2}{n\,a_0}\right)^3 \frac{(n-l-1)!}{2n\,[(n+1)!]^3}\right]^{1/2}\left(\frac{2\,r}{a_0\,n}\right)^{l} e^{-\left(\frac{r}{a_0\,n}\right)} L_{n-l-1}^{2l+1}\left(\frac{2\,r}{a_0\,n}\right) \tag{139}$$

La solution exacte (valeurs et fonctions propres) de l'équation de schrodinger pour l'atome d'hydrogène (et les ions hydrogénoïdes He^{+}, Li^{2+}, $\cdots$ etc) s'obtient en multipliant les solutions des parties angulaires (89) et radiale (139) :

$$\begin{aligned}\psi_{n,l,m}(r,\theta,\phi) = {} & \left[\left(\frac{2}{n\,a_0}\right)^3 \frac{(n-l-1)!}{2n\,[(n+1)!]^3}\right]^{1/2}\left(\frac{2\,r}{a_0\,n}\right)^{l} e^{-\left(\frac{r}{a_0\,n}\right)} L_{n-l-1}^{2l+1}\left(\frac{2\,r}{a_0\,n}\right) \\ & \times \underbrace{(-1)^m \sqrt{\frac{(2\,l+1)!\,(l-m)!}{4\,\pi\,(l+m)!}} \times P_{l,m}(\cos\theta) \times e^{j\,m\,\phi}}_{Y_l^m(\theta,\phi)}\end{aligned} \tag{140}$$

Ou simplement en occultant la partie angulaire :

$$\begin{aligned}\Rightarrow \quad \psi_{n,l,m}(r,\theta,\phi) = {} & \left[\left(\frac{2}{n\,a_0}\right)^3 \frac{(n-l-1)!}{2n\,[(n+1)!]^3}\right]^{1/2}\left(\frac{2\,r}{a_0\,n}\right)^{l} e^{-\left(\frac{r}{a_0\,n}\right)} \\ & \times L_{n-l-1}^{2l+1}\left(\frac{2\,r}{a_0\,n}\right) \times Y_l^m(\theta,\phi)\end{aligned} \tag{141}$$

Le terme exponentiel décroissant supplante le terme polynomial croissant de sorte que la fonction d'onde globale $\psi_{n,l,m}(r,\theta,\phi)$ tend vers zéro pour les grandes valeurs de r (loin du noyau), c'est ce qui est attendu. Les valeurs propres sont obtenues avec l'équation (131) :

$$\Rightarrow \quad E_n = -\frac{m}{2\,\hbar^2}\left[\frac{e^2}{4\,\pi\,\epsilon_0}\right]^2 \times \frac{1}{n^2} \tag{142}$$

Les harmoniques sphériques $Y_l^m(\theta,\phi)$, fournissent des informations sur la position de l'électron autour du noyau, et la fonction radiale $R_{n,l}(r)$ décrit l'éloignement de l'électron par rapport au noyau. Comme on peut le constater, l'équation de Schrödinger requiert trois nombres quantiques $(n,\, l,\, m)$ afin de spécifier une fonction d'onde pour l'électron. Les nombres quantiques fournissent des informations sur la distribution spatiale d'un électron. Bien que n puisse prendre n'importe quel nombre entier positif non nul, seules certaines valeurs de l et de m sont autorisées pour une valeur donnée de n. Le nombre quantique principal n indique l'énergie de l'électron et la distance moyenne d'un électron par rapport au noyau. Plus un électron est proche du noyau, chargé positivement, plus l'électron est fortement attiré par le noyau comparativement à un électron plus éloigné dans l'espace. Cela signifie que les électrons ayant une valeur de n plus élevée sont plus faciles à éliminer d'un atome.

Le deuxième nombre quantique l est appelé nombre quantique azimutal. Ce dernier décrit la forme de la région de l'espace occupée par un électron, donc la sous-couche considérée. Les valeurs de ce nombre quantique sont données par $n \geq l+1$. Le troisième nombre quantique, est le nombre quantique magnétique m. Ce nombre quantique décrit l'orientation de la région dans l'espace occupé par un électron par rapport à un champ magnétique appliqué. Les valeurs autorisées de m dépendent de la valeur de l selon $-l \leq m \leq +l$. Chaque combinaison autorisée des trois nombres quantiques fournit une distribution spatiale particulière à l'électron.

Il est intéressant de comparer les résultats obtenus en résolvant l'équation de Schrödinger avec le modèle de l'atome d'hydrogène de Bohr. Les valeurs propres (spectre énergétique) sont quasiment identiques. Toutefois, les modèles de Schrödinger et de Bohr sont différents à bien des égards, notamment en ce qui concerne les deux points énumérés ci-dessous :

1) *Le modèle de Schrödinger n'associe pas d'orbites bien définies pour l'électron. Les fonctions d'onde donnent seulement la probabilité de trouver l'électron dans l'élément de volume dv à différentes directions (θ et ϕ) et distances du noyau (r).*
2) *Les nombres quantiques apparaissent spontanément lors de la résolution de l'équation de Schrödinger alors que Bohr a dû postuler l'existence d'états énergétiques quantifiés. Bien que plus complexe, le modèle de Schrödinger conduit à une meilleure correspondance entre la théorie et l'expérience.*

$\cdots$ *Ouuf c'est terminé* $\cdots$ *dire que l'hydrogène est l'atome le plus simple !*

C. ***Résolution numérique de quelques EDPs***

Mis à part certaines EDPs particulières, la grande majorité des EDPs issues de la physique et de la chimie n'admette pas de solution explicite ou analytique. Il est donc impératif de recourir à la résolution numérique sur ordinateur pour évaluer qualitativement et quantitativement les solutions. Le principe de base de ces méthodes de résolution numérique des EDPs, consiste à chercher des valeurs numériques discrètes approchant au mieux la solution exacte. Le concept le plus important dans cette résolution est celui de *discrétisation*, marquant la passage du continu au discret. Dans cette section, on se propose de résoudre numériquement quelques EDPs régissant des phénomènes bien connus de la physique. La résolution numérique sera conduite en considérant la méthode des *différences finies*.

Commençons par chercher les approximations des dérivées première et seconde. Soit $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^2(\mathbb{R}^2)$ et écrivons le développement en *séries de Taylor* de $u(x_i+\Delta x, t_j)$ autour du point (x_i, t_j), soit :

$$u(x_i+\Delta x,t_j) = u(x_i,t_j) + \frac{\Delta x}{1!}\frac{\partial u}{\partial x}\bigg|_{(x_i,t_j)} + \frac{\Delta x^2}{2!}\frac{\partial^2 u}{\partial x^2}\bigg|_{(x_i,t_j)} + \cdots + \frac{\Delta x^n}{n!}\frac{\partial^n u}{\partial x^n}\bigg|_{(x_i,t_j)} + O(\Delta x^{n+1}) \qquad (143)$$

L'approximation par les différences finies de la dérivée première est obtenue en tronquant la série de Taylor à l'ordre deux, soit :

$$u(x_i+\Delta x,t_j)=u(x_i,t_j)+\frac{\Delta x}{1!}\frac{\partial u}{\partial x}\bigg|_{(x_i,t_j)}+\frac{\Delta x^2}{2!}\frac{\partial^2 u}{\partial x^2}\bigg|_{(x_i,t_j)}+O(\Delta x^3) \tag{144}$$

Après réarrangement, il vient :

$$\frac{\partial u}{\partial x}\bigg|_{(x_i,t_j)}=\underbrace{\frac{u(x_i+\Delta x,t_j)-u(x_i,t_j)}{\Delta x}}_{\text{Approx. par DF}}-\underbrace{\frac{\Delta x^2}{2!}\frac{\partial^2 u}{\partial x^2}\bigg|_{(x_i,t_j)}+O(\Delta x^3)}_{\text{Erreur de troncature}} \tag{145}$$

Cette dernière relation est présentée sous la forme :

$$\frac{\partial u}{\partial x}\bigg|_{(x_i,t_j)}=\underbrace{\frac{u(x_i+\Delta x,t_j)-u(x_i,t_j)}{\Delta x}}_{\text{Approx. par DF}}+\underbrace{O(\Delta x)}_{\text{Erreur de troncature}} \tag{146}$$

L'écriture de $O(\Delta x)$ indique que l'approximation de la première dérivée par la formule des différences finie est d'ordre un par rapport au pas de discrétisation Δx. Cela signifie concrètement que lorsqu'on divise le pas de discrétisation Δx par une constante arbitraire $a>0$ implique que l'erreur d'approximation, entre dérivée exacte et approchée, est divisée par a. Avec un raisonnement analogue, il est possible aussi de calculer une approximation de la dérivée seconde. Afin d'atteindre cette approximation, nous devons réaliser deux développement en séries de Taylor (à droite et à gauche), soit :

$$u(x_i+\Delta x,t_j)=u(x_i,t_j)+\frac{\Delta x}{1!}\frac{\partial u}{\partial x}\bigg|_{(x_i,t_j)}+\frac{\Delta x^2}{2!}\frac{\partial^2 u}{\partial x^2}\bigg|_{(x_i,t_j)}+\frac{\Delta x^3}{3!}\frac{\partial^3 u}{\partial x^3}\bigg|_{(x_i,t_j)}+O(\Delta x^4) \tag{147}$$

$$u(x_i-\Delta x,t_j)=u(x_i,t_j)-\frac{\Delta x}{1!}\frac{\partial u}{\partial x}\bigg|_{(x_i,t_j)}+\frac{\Delta x^2}{2!}\frac{\partial^2 u}{\partial x^2}\bigg|_{(x_i,t_j)}-\frac{\Delta x^3}{3!}\frac{\partial^3 u}{\partial x^3}\bigg|_{(x_i,t_j)}+O(\Delta x^4) \tag{148}$$

En sommant membre par membre les deux dernières relations, nous obtenons l'approximation recherchée :

$$\frac{\partial^2 u}{\partial x^2}\bigg|_{(x_i,t_j)}=\underbrace{\frac{u(x_i+\Delta x,t_j)-2\,u(x_i,t_j)+u(x_i-\Delta x,t_j)}{\Delta x^2}}_{\text{Approx. par DF}}+\underbrace{O(\Delta x^2)}_{\text{Erreur de troncature}} \tag{149}$$

Théorème *: la solution numérique d'un schéma itératif aux différences finies, d'un problème linéaire aux valeurs initiales, converge vers la solution exacte si le schéma est consistant et stable.*

Définition 1 *: une approximation est dite consistante d'ordre p s'il existe une constante arbitraire $c>0$ indépendante du pas de discrétisation telle que cette erreur soit majorée par la quantité $c\,\Delta x^p$. Soit $u(x,t)\in\Omega\subset\mathbb{R}$ une fonction de classe $\mathcal{C}^4(\Omega)$, pour la dérivée d'ordre un, nous avons :*

$$\underbrace{\left|u'(x,t)-\underbrace{\frac{u(x_i+\Delta x,t_j)-u(x_i-\Delta x,t_j)}{2\Delta x}}_{Approx.\ par\ DF}\right|}_{\epsilon(u)}\leq\underbrace{\max_{x_1\leq x\leq x_1+\Delta x}\left|\frac{u(x)^{(3)}}{6}\right|}_{c}\Delta x^2 \tag{150}$$

Pour la dérivée seconde, il vient :

$$\left|u''(x,t)-\underbrace{\frac{u(x_i+\Delta x,t_j)-2\,u(x_i,t_j)+u(x_i-\Delta x,t_j)}{\Delta x^2}}_{Approx.\ par\ DF}\right|\leq\underbrace{\max_{x_1\leq x\leq x_1+\Delta x}\left|\frac{u(x)^{(4)}}{2}\right|}_{c}\Delta x^2 \tag{151}$$

De manière générale, le schéma numérique des différences finies est dit consistant à l'équation EDP si cette erreur de troncature tends vers zéro lorsque le pas de discrétisation temporel Δt et le pas de discrétisation spatial Δx tendent indépendamment vers zéro. Autrement dit si,

$$\lim_{\Delta x \approx 0, \Delta t \approx 0} \Big| \epsilon(u) \Big| = 0 \tag{152}$$

Définition 1' *: une solution est dite stable*[8] *si une petite variation des conditions de bord engendre une faible variation de la solution. En terme mathématique simple cela se traduit par :*

$$\forall x \in \mathbb{R}, \forall t \in \mathbb{R}^+ \Rightarrow \Big| u_1(x,t) - u_2(x,t) \Big| \leq \epsilon$$

Il convient de noter également que l'analyse de la stabilité d'un schéma numérique peut être conduite en déterminant le *facteur d'amplification* du schéma itératif.

1) ***En dimension 1****:* dans cette section, nous considérons le problème de la corde élastique fixée aux extrémités $x = 0$ et $x = L$ telle que L est égale à une unité de longueur. La corde subit des déformations selon un mode vertical. L'amplitude des déformations est décrite par la fonction $u(x)$, ainsi le problème est formalisé mathématiquement par :

$$\mathcal{P} : \begin{cases} -u''(x) = f(x), \quad 0 < x < L \\ \\ u(0) = 0 \\ u(L) = 0 \end{cases} \tag{153}$$

Autrement dit, l'Eq. (153) est celle régissant la déformation linéaire de la corde élastique. Le second terme représente la source des déformations. Les conditions aux limites $u(0) = 0$ et $u(L) = 0$ traduisent le fait que la corde ne subit pas de déformations aux extrémités. Il s'agit d'une résolution numérique, donc le calcul d'une approximation de $u(x_i)$ au point x_i. Nous commençons par la discrétisation des abscisses.

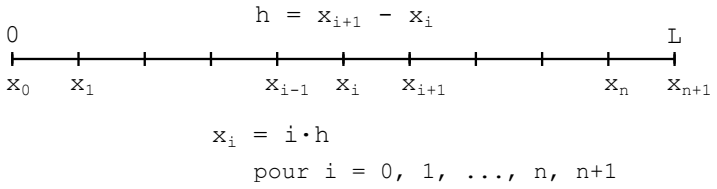


FIGURE 2: Discrétisation du problème aux limites 1 Dim

La discrétisation se fait avec un pas constant $h = x_{i+1} - x_i$ par conséquent tous les points x_i sont équidistants. La solution exacte au point x_i, soit $u(x_i)$ est inconnue. Nous cherchons des solutions approchées u_i qui sont des approximations de la solution exacte $u(x_i)$ au point x_i.

Utilisons une formule des différences finies centrée afin d'approcher $u''(x)$, soit :

$$-\frac{u(x_{i-1}) - 2\,u(x_i) + u(x_{i+1})}{h^2} = f(x_i) + O(h^2) \tag{154}$$

8. Cette notion de stabilité consiste à analyser si les perturbations de la solution numérique ne sont pas amplifiées au cours des itérations. Les calculs sur ordinateur sont déterminés avec une précision finie, sont ainsi sujet à des erreurs d'arrondis. Pendant un calcul itératif, ces erreurs peuvent être amplifiées.

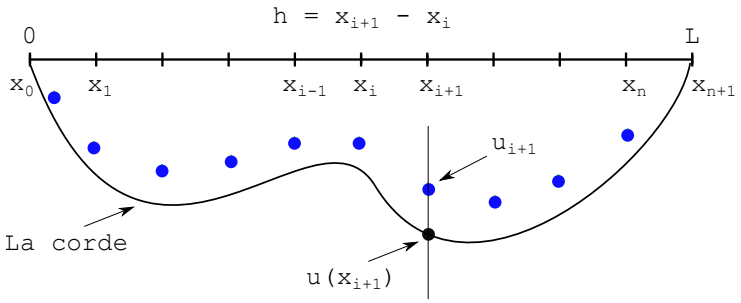


FIGURE 3: Solution exacte $u(x_i)$ versus solution approchée u_i

Le terme $O(h^2)$ stipule que lorsqu'on divise h par une constante a, l'erreur $|u(x_i) - u_i|$ est divisée par a^2. C'est le résultat du théorème suivant :

Si $u(x)$ est de classe $\mathcal{C}^4$ sur $[0\,,\,L]$, alors $\exists\; c \in \mathbb{R}^+$ telle que $\forall\; 0 < h < L$ nous avons :

$$\max_{1\leq i\leq 1} |u(x_i) - u_i| \leq c\, h^2$$
$$\max_{1\leq i\leq 1} |u(x_i) - u_i| \leq \frac{1}{96} \max_{0\leq x\leq L} |u(x)^{(4)}|$$

Ce théorème affirme que l'erreur d'intégration est majorée par la quatrième dérivée de la fonction $u(x)$ et plus h est petit plus on s'approche de la solution exacte. Écrivons maintenant la même formule des différences finies pour les approximations u_i, il vient :

$$\begin{cases} -\dfrac{u_{i-1} - 2\,u_i + u_{i+1}}{h^2} = f(x_i), \quad i = 1, 2, ..., n \\ \\ u(x_0) = u_0 \\ u(L) = u_{n+1} \end{cases} \tag{155}$$

Le terme $O(h^2)$ n'est pas pris en considération dans les calculs. Le schéma (155) correspond à la résolution d'un système linéaire.

$$A_n\,(\mathbb{R}^n \times \mathbb{R}^n)\,u_n\,(\mathbb{R}^n) = f_n\,(\mathbb{R}^n) \tag{156}$$

Explicitons le schéma (155) pour $n = 4$ et pour des conditions aux limites $u_0 = \alpha$ et $u_{n+1} = \beta$. Ces conditions aux limites signifient qu'à $x = 0$ la corde subit une déformation constante égale à la valeur α et à l'autre extrémité $x = L$, la corde subit aussi une déformation constante égale à la valeur β. Nous aurions pu prendre par exemple $u_0 = 0$ et $u_{n+1} = 0$, cela indique que la corde ne subit aucune déformation aux extrémités. Nous préférons prendre un cas général avec les conditions $u_0 = \alpha$ et $u_{n+1} = \beta$.

$$\begin{cases} \dfrac{-u_0+2\,u_1-u_2}{h^2}=f(x_1), & i=1\\ \dfrac{-u_1+2\,u_2-u_3}{h^2}=f(x_2), & i=2\\ \dfrac{-u_2+2\,u_3-u_4}{h^2}=f(x_3), & i=3\\ \dfrac{-u_3+2\,u_4-u_5}{h^2}=f(x_4), & i=4 \end{cases} \Rightarrow \begin{cases} -\alpha+2\,u_1-u_2=h^2\,f(x_1), & i=1\\ -u_1+2\,u_2-u_3=h^2\,f(x_2), & i=2\\ -u_2+2\,u_3-u_4=h^2\,f(x_3), & i=3\\ -u_3+2\,u_4-\beta=h^2\,f(x_4), & i=4 \end{cases}$$

$$\Rightarrow \begin{cases} 2\,u_1-u_2=h^2\,f(x_1)+\alpha, & i=1\\ -u_1+2\,u_2-u_3=h^2\,f(x_2), & i=2\\ -u_2+2\,u_3-u_4=h^2\,f(x_3), & i=3\\ -u_3+2\,u_4=h^2\,f(x_4)+\beta, & i=4 \end{cases} \tag{157}$$

En adoptant une notation matricielle, le problème peut s'écrire :

$$A_n\,u_n=f_n \tag{158}$$

Avec,

$$A_n=\frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & 0\\ -1 & 2 & -1 & 0\\ 0 & -1 & 2 & -1\\ 0 & 0 & -1 & 2 \end{pmatrix}\in\mathbb{R}^4\times\mathbb{R}^4 \tag{159}$$

$$f_n=\begin{pmatrix} h^2\,f(x_1)+\alpha\\ h^2\,f(x_2)\\ h^2\,f(x_3)\\ h^2\,f(x_1)+\beta \end{pmatrix}\in\mathbb{R}^4 \tag{160}$$

La matrice A_n est tridiagonale, symétrique et définie positive. Le vecteur des valeurs de la solution (inconnues) aux points x_i est donné

$$u_h=\begin{pmatrix} u_1\\ u_2\\ u_3\\ u_4 \end{pmatrix}\in\mathbb{R}^4 \tag{161}$$

Ce schéma se généralise pour i $=\{1,2,...,n\}$, selon :

$$A_n = \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R}^n \quad f_n = \begin{pmatrix} h^2 f(x_1) + \alpha \\ f(x_2) \\ \vdots \\ f(x_{n-1}) \\ h^2 f(x_n) + \beta \end{pmatrix} \in \mathbb{R}^n \quad u_n = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \\ u_n \end{pmatrix} \in \mathbb{R}^n$$

Exercice ❶ ☞ Ⓡ

Soit l'équation :

$$\mathcal{P} : \begin{cases} -u''(x) = f(x) \quad 0 \leq x \leq 1 \\ \\ u(0) = 0 \\ u(1) = 0 \end{cases} \tag{162}$$

Le second terme de l'équation vaut :

$$f(x) = e^{3x^2} \times (x + 1) \tag{163}$$

1) Résoudre numériquement pour n = 100 l'équation (162) par la méthode des différences finies.

Voici le script Matlab®

```
clear all ; clc ; close all ;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% SAMIR KENOUCHE - RESOLUTION NUMERIQUE DU PROBLEME AUX LIMITES DE LA CORDE
% ELASTIQUE : - u''(x) = f(x) AVEC LES CONDITIONS u(0) = 0 et u(1) = 0

np = 100 ; pas_x = 1/(np+1) ; xi = 0: pas_x :1 ; % DISCRETISATION
fx = @(xi) exp(3.*xi.^2).*(xi + 1) ;

sur_diag = diag(ones(np - 1, 1) ,1)*(-1) ;  % SUR-DIAGONALE
des_diag = diag(ones(np - 1, 1) ,-1)*(-1) ; % SOUS-DIAGONALE
in_diag = diag(ones(np, 1))*(2) ;           % DIAGONALE PRINCIPALE

An = sur_diag + des_diag + in_diag ;        % MATRICE An
fn = fx(xi(2:end-1)) ;                      % SOURCE DE LA DEFORMATION

un = inv(An)*fn' ;                          % CALCUL DES APPROXIMATIONS
un = [0 un' 0] ;                % ON RAJOUTE LES CONDITIONS AUX LIMITES

fig1 = figure('color',[1 1 1]) ; plot(xi, un,'o') ;
xlabel('x_i') ; ylabel('u_i') ; title('SOLUTION APPROCHEE') ;
```

Considérons désormais des fonctions $c(x)$ et $f(x)$ continues sur l'intervalle $[a, b]$. On se propose de résoudre, par les différences finies, *l'équation de la convection-diffusion*. Soient α et β deux constantes réelles. Le problème

consiste à trouver une fonction $u(x)$ deux fois dérivable sur l'intervalle $[a,b]$ et qui satisfait :

$$\mathcal{P}:\left\{\begin{array}{l} -u^{''}(x)+c(x)\,u^{'}(x)=f(x), \quad x\in[a,b] \\ u(a)=0 \quad \text{et} \quad u(b)=0 \end{array}\right. \tag{164}$$

Comme précédemment, ce type de problème est dénommé *problème aux limites*. Cette dénomination provient du fait que la fonction $u(x)$ doit satisfaire les conditions aux limites, $u(a)=0$ et $u(b)=0$, posées aux bornes de l'intervalle $[a,b]$. Afin de résoudre numériquement (trouver la solution approchée) le système (164), nous recourrons à la méthode des différences finies. Suivant cette méthode numérique, l'intervalle $[a,b]$ sur lequel nous cherchons la solution $u(x)$ est discrétisé en $n+1$ sous-intervalles équidistants de longueur h avec $x_i=x_0+i\,h$ et $i=1,2,3,...,n$. On cherche alors en chacun de ces points une valeur approchée, notée u_i, de $u(xi)$. Ainsi, le système continu initial est substitué par un système discret. L'idée de base de la méthode des différences finies consiste à remplacer l'équation différentielle (164) par un système de n équations algébriques. Ce système d'équations est obtenu en écrivant cette équation différentielle en chaque point de discrétisation x_i, et en substituant également à chaque valeur $u^{''}(x)$ l'approximation de la dérivée seconde :

$$u^{''}(x)\approx\frac{u(x_{i-1})-2\,u(x_i)+u(x_{i+1})}{h^2}+O(h^2) \tag{165}$$

Et la dérivée $u^{'}(x)$ est approchée par

$$u^{'}(x)\approx\frac{u(x_{i+1})-u(x_{i-1})}{2\,h}+O(h) \tag{166}$$

Ainsi, l'équation différentielle (164) est réécrite suivant :

$$\left\{\begin{array}{l} -\dfrac{u(x_{i-1})-2\,u(x_i)+u(x_{i+1})}{h^2}+c(x_i)\,\dfrac{u(x_{i+1})-u(x_{i-1})}{2\,h}=f(x_i) \quad i\in\{1,...,n\} \\ u_0=0 \quad \text{et} \quad u_n=0 \end{array}\right. \tag{167}$$

Comme précédemment en adoptant une notation matricielle après quelques réarrangements, le problème peut s'écrire :

$$A_h\,u_h=b_h \tag{168}$$

Avec $A_h=A_1+A_2$,

$$A_1=\begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix} \tag{169}$$

$$A_2 = \frac{r\,h}{2}\begin{pmatrix} c(x_1) & 0 & \dots & 0 \\ 0 & c(x_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & c(x_n) \end{pmatrix} \times \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ -1 & 0 & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 0 & 1 \\ 0 & \dots & 0 & -1 & 0 \end{pmatrix} \tag{170}$$

$$f_h = h^2 \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{n-1}) \\ f(x_n) \end{pmatrix} \tag{171}$$

Le vecteur des valeurs de la solution (inconnues) aux points x_i est donné

$$u_h = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \\ u_n \end{pmatrix} \in \mathbb{R}^n \tag{172}$$

On peut mettre en évidence le fait que la matrice A_h est inversible (sous l'hypothèse que la fonction $c(x) \geq 0$. La matrice A_h est symétrique définie positive). Les matrices A_1 et A_2 sont tridiagonales. Afin d'obtenir la solution discrète, les valeurs du vecteur u_h, il va falloir résoudre le système linéaire tridiagonal (168).

Exercice ❷ ☞ ®

Soit l'équation de *convection-diffusion* :

$$\mathcal{P} : \begin{cases} -u''(x) + r\,x\,u'(x) = f(x) \\ \\ x \in [0,1] \quad \text{et} \quad u(0) = 0, \quad u(1) = 0 \end{cases} \tag{173}$$

Le second terme de l'équation vaut :

$$f(x) = \frac{(r^2\,e^{(r\,x)}\,(x-1))}{(1-e^r) + r\,x} \tag{174}$$

La solution exacte est donnée par :

$$\text{funex} = \frac{x - (1 - e^{(r\,x)})}{(1-e^r)} \tag{175}$$

1) Résoudre numériquement l'équation (173) par la méthode des différences finies.
2) Tracer, sur la même figure, les solutions exacte et numérique pour `n=64` et `r=1/2`.
3) Étudier l'erreur en fonction du nombre de sous-intervalles de discrétisation et du paramètre `r`.

4) Représenter graphiquement cette erreur en fonction de `n+1` et des valeurs du paramètre `r`.

Voici le script Matlab®

```
clear all ; clc ; close all ;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 19/11/2019 Samir KENOUCHE : ALGORITHME PERMETTANT
% L'IMPLEMENTATION, SOUS MATLAB, DE LA METHODE DES DIFFERENCES FINIES EN
% DIMENSION 1

fx = @(x,r) (r.^2.*exp(r.*x).*(x-1))./(1 - exp(r)) + r.*x ;
n = 64 ; r = 1/2 ; h = 1/(n+1) ; xh = 0 : h: 1 ;
cx = inline('x') ; funex = @(x,r) x-(1-exp(r.*x))./(1- exp(r)) ;

fn = h.^2.*fx(xh(1:n),r) ; cx = cx(xh(1:n)).*r ;

sur_diag = diag(ones(n-1 ,1) ,1) ; sur_diag(sur_diag == 1) = -1 ;
des_diag = diag(ones(n-1 ,1) , -1) ; des_diag(des_diag == 1) = -1 ;
in_diag = diag(ones(n ,1)) ; in_diag(in_diag == 1) = 2 ;

A1 = sur_diag + des_diag + in_diag ; a1 = diag(ones(n ,1)) ;
a1(a1 == 1) = cx ;

sur_diag_a2 = diag(ones(n-1 ,1) ,1) ; sur_diag_a2(sur_diag_a2 == 1) = 1 ;
des_diag_a2 = diag(ones(n-1 ,1) , -1) ; des_diag_a2(des_diag_a2 == 1) = -1 ;
in_diag_a2 = diag(ones(n ,1)) ; in_diag_a2(in_diag_a2 == 1) = 0 ;

a2 = sur_diag_a2 + des_diag_a2 + in_diag_a2 ;
A2 = (r*h/2).*(a1*a2) ; An = A1 + A2 ;

un = fn*inv(An) ;  % RESOLUTION DU SYSTEME LINEAIRE
uh = [0 , un, 0] ; % SOL. FINALE - PRISE EN COMPTE DES CONDITIONS INITIALES

%%%%%%%%%%%%%%%%%%%%%%%%%%%% AFFICHAGE GRAPHIQUE %%%%%%%%%%%%%%%%%%%%%%%%%%%%
figure('color',[1 1 1])

plot(xh,uh,'o','MarkerSize',7,'LineWidth',1) ; hold on ;
xk = 0:0.001:1 ; plot(xk,funex(xk,r),'r','LineWidth',1.2)
axis([-0.1 1.1 -0.005 0.07]) ;

ih =legend('SOLUTION NUMERIQUE','SOLUTION EXACTE') ;
set(ih,'Interpreter','none','Location','South','Box','on',...
    'Color','none') ; xlabel('x','FontSize',12) ; ylabel('u(x)','FontSize',12)

msg1 = strcat('r= ', num2str(r)) ;
gtext(msg1) % cliquer sur la figure pour afficher : msg1
msg2 = strcat('n= ', num2str(n)) ;
gtext(msg2) % cliquer sur la figure pour afficher : msg2
```

Ci-dessous, la visualisation graphique de la solution obtenue.

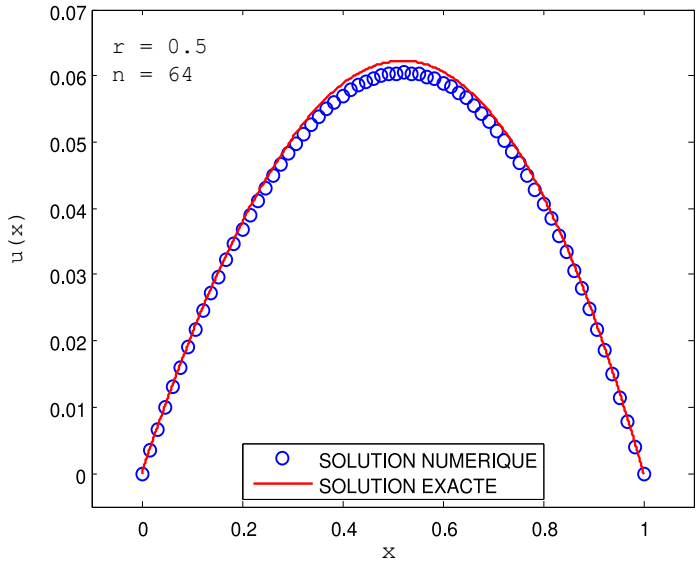


FIGURE 4: Solutions exacte et numérique obtenues par différences finies

```
clear all ; clc ; close all ;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%  18/11/2019 Samir KENOUCHE : ALGORITHME PERMETTANT
% L'ANALYSE DE L'ERREUR EN FONCTION DU NOMBRES DE SOUS-INTERVALLES DE
% DISCRETISATION
fx = @(x,r) (r.^2.*exp(r.*x).*(x-1))./(1 - exp(r)) + r.*x ;  r = 1/2 ;
cx1 = inline('x') ; funex = @(x,r) x-(1-exp(r.*x))./(1- exp(r)) ;
n = 5 : 5 : 120 ;

for ik = 1:length(n)

  h = 1/(n(ik)+1) ; xh = 0:h: 1 ;
  fn = h.^2.*fx(xh(1:n(ik)),r) ; cx = cx1(xh(1:n(ik))).*r ;

 sur_diag = diag(ones(n(ik)-1 ,1) ,1) ; sur_diag(sur_diag == 1) = -1 ;
 des_diag = diag(ones(n(ik)-1 ,1) , -1) ; des_diag(des_diag == 1) = -1 ;
 in_diag = diag(ones(n(ik), 1)) ; in_diag(in_diag == 1) = 2 ;

 A1 = sur_diag + des_diag + in_diag ; a1 = diag(ones(n(ik) ,1)) ;
 a1(a1 == 1) = cx ;

 sur_diag_a2 = diag(ones(n(ik)-1 ,1) ,1) ; sur_diag_a2(sur_diag_a2 == 1) = 1 ;
 des_diag_a2 = diag(ones(n(ik)-1 ,1) , -1) ;
 des_diag_a2(des_diag_a2 == 1) = -1 ;
 in_diag_a2 = diag(ones(n(ik) ,1)) ; in_diag_a2(in_diag_a2 == 1) = 0 ;

 a2 = sur_diag_a2 + des_diag_a2 + in_diag_a2 ;
 A2 = (r*h/2).*(a1*a2) ; An = A1 + A2 ;

 un = fn*inv(An) ; % RESOLUTION DU SYSTEME LINEAIRE
```

```
 uh = [0 , un, 0] ; % SOL. FINALE - PRISE EN COMPTE DES CONDITIONS INITIALES

 err(ik) = max(abs(uh - funex(xh,r))) ;

end

figure('color',[1 1 1]) ; hold on ; box on ;
loglog(n+1,err,'o','MarkerSize',7,'LineWidth',1) ; % ECHELLE LOGARITHMIQUE
xlabel('n + 1','FontSize',12) ; ylabel('Erreur absolue','FontSize',12) ;
```

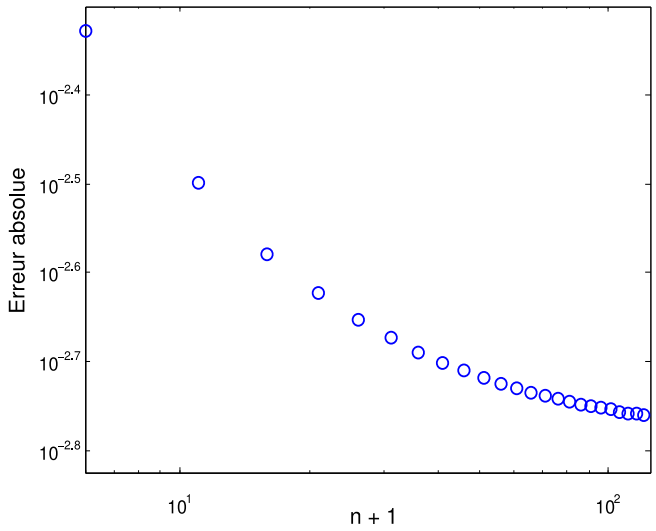


FIGURE 5: Erreur absolue versus le nombre d'intervalles de discrétisation.

```
clear all ; clc ; close all ;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%  18/11/2019 Samir KENOUCHE : ALGORITHME PERMETTANT
% L'ANALYSE DE L'ERREUR EN FONCTION DU NOMBRES DE SOUS-INTERVALLES DE
% DISCRETISATION
fx = @(x,r) (r.^2.*exp(r.*x).*(x-1))./(1 - exp(r)) + r.*x ;
cx1 = inline('x') ; funex = @(x,r) x-(1-exp(r.*x))./(1- exp(r)) ;
n = 64 ;

for r = 1:40

  h = 1/(n + 1) ; xh = 0:h: 1 ;
  fn = (h.^2).*fx(xh(1:n),r) ; cx = cx1(xh(1:n)).*r ;

 sur_diag = diag(ones(n - 1 ,1) ,1) ; sur_diag(sur_diag == 1) = -1 ;
 des_diag = diag(ones(n - 1 ,1) , -1) ; des_diag(des_diag == 1) = -1 ;
 in_diag = diag(ones(n, 1)) ; in_diag(in_diag == 1) = 2 ;

 A1 = sur_diag + des_diag + in_diag ; a1 = diag(ones(n, 1)) ;
 a1(a1 == 1) = cx ;

 sur_diag_a2 = diag(ones(n - 1 ,1) ,1) ; sur_diag_a2(sur_diag_a2 == 1) = 1 ;
 des_diag_a2 = diag(ones(n - 1 ,1) , -1) ;
```

```
des_diag_a2(des_diag_a2 == 1) = -1 ;
in_diag_a2 = diag(ones(n ,1)) ; in_diag_a2(in_diag_a2 == 1) = 0 ;

a2 = sur_diag_a2 + des_diag_a2 + in_diag_a2 ;
A2 = (r*h/2).*(a1*a2) ; An = A1 + A2 ;

un = fn*inv(An) ; % RESOLUTION DU SYSTEME LINEAIRE
uh = [0 , un, 0] ; % SOL. FINALE - PRISE EN COMPTE DES CONDITIONS INITIALES

err_r = max(abs(uh - funex(xh, r))) ;

figure(1) ; hold on ; box on ;
semilogy(r,err_r,'o','MarkerSize',7,'LineWidth',1)
% ECHELLE SEMI-LOGARITHMIQUE
xlabel('Valeur de r','FontSize',12) ;
ylabel('Erreur absolue','FontSize',12) ;

end
```

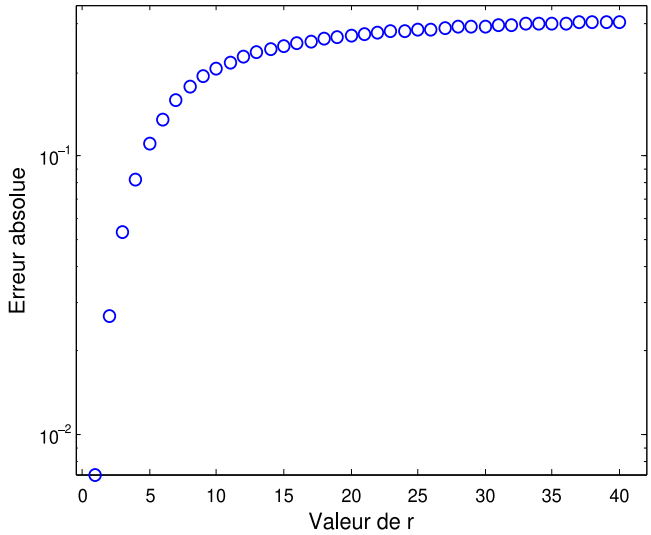


FIGURE 6: Erreur absolue en fonction du paramètre r.

Supplément : avec une démarche analogue on peut résoudre également le problème de la *flexion simple* dont la formulation mathématique est donnée par :

$$\mathcal{P} : \begin{cases} -u''(x) + c(x)\,u(x) = f(x), \quad x \in [a,b] \\ u(a) = \alpha \quad \text{et} \quad u(b) = \beta \end{cases} \tag{176}$$

En utilisant la formule des différences finies centrées, le problème devient :

$$\begin{cases} -\dfrac{u(x_{i-1}) - 2\,u(x_i) + u(x_{i+1})}{h^2} + c(x_i)\,u_i = f(x_i) \quad i \in \{1, ..., n\} \\ u_0 = \alpha \quad \text{et} \quad u_n = \beta \end{cases} \tag{177}$$

En adoptant une notation matricielle :

$$A_h\, u_h = b_h \tag{178}$$

Avec,

$$A_h = A_h^{(0)} + \begin{pmatrix} c(x_1) & 0 & \dots & 0 \\ 0 & c(x_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & c(x_n) \end{pmatrix} \quad \text{et} \quad A_h^{(0)} = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix} \tag{179}$$

Les deux matrices ci-dessus peuvent se combiner pour donner :

$$A_h = \frac{1}{h^2} \begin{pmatrix} 2 + c(x_1)\, h^2 & -1 & 0 & \dots & 0 \\ -1 & 2 + c(x_2)\, h^2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 + c(x_{n-1})\, h^2 & -1 \\ 0 & \dots & 0 & -1 & 2 + c(x_n)\, h^2 \end{pmatrix} \tag{180}$$

$$b_h = \begin{pmatrix} f(x_1) + \alpha\, h^{-2} \\ f(x_2) \\ \vdots \\ f(x_{n-1}) \\ f(x_n) + \beta\, h^{-2} \end{pmatrix} \tag{181}$$

Le vecteur des solution (inconnues) aux points x_i est donné par :

$$u_h = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \\ u_n \end{pmatrix} \in \mathbb{R}^n \tag{182}$$

2) ***Problème non-linéaire****:* considérons le problème non-linéaire suivant :

$$\mathcal{P} : \begin{cases} -u''(x) + x\, u(x)^3 = f(x), \quad 0 < x < L \\ \\ u(0) = \alpha \\ u(L) = \beta \end{cases} \tag{183}$$

Contrairement au cas linéaire (problème (153)) où le terme $x\, u(x)^3$ n'existe pas, pour le problème (183) nous avons une relation non-linéaire entre la source de la déformation $f(x)$ et l'amplitude de la déformation $u(x)$. Autrement dit, si j'applique par exemple une force $f(x)$ deux fois plus grande, l'amplitude de la déformation

$u(x)$ n'est pas doublée. En appliquant la formule des différences finies centrées il vient :

$$\begin{cases} -\dfrac{u_{i-1}-2\,u_i+u_{i+1}}{h^2}+x_i\,u_i^3=f(x_i) \quad i\in\{1,...,n\} \\ \\ u_0=\alpha \quad \text{et} \quad u_{n+1}=\beta \end{cases} \tag{184}$$

Contrairement au problème (153), ici nous cherchons à résoudre un système de n équations non-linaires. Il existe dans la littérature plusieurs méthodes itératives pour effectuer ce calcul. On citera les méthodes de Newton, de fausse position ou de la sécante. Je renvoie, les lecteurs intéressés par ces méthodes, à mon cours d'analyse numérique que je dispense aux deuxièmes années des filières physique et chimie. Ce cours est disponible en version pdf. Pour n = 4, on obtient le système d'équations :

$$\begin{cases} \dfrac{-u_0+2\,u_1-u_2}{h^2}+x_1\,u_1^3=f(x_1), \quad i=1 \\ \\ \dfrac{-u_1+2\,u_2-u_3}{h^2}+x_2\,u_2^3=f(x_2), \quad i=2 \\ \\ \dfrac{-u_2+2\,u_3-u_4}{h^2}+x_3\,u_3^3=f(x_3), \quad i=3 \\ \\ \dfrac{-u_3+2\,u_4-u_5}{h^2}+x_4\,u_4^3=f(x_4), \quad i=4 \end{cases} \Rightarrow \begin{cases} -\alpha+2\,u_1-u_2+h^2\,x_1\,u_1^3-h^2\,f(x_1)=0 \\ \\ -u_1+2\,u_2-u_3+h^2\,x_2\,u_2^3-h^2\,f(x_2)=0 \\ \\ -u_2+2\,u_3-u_4+h^2\,x_3\,u_3^3-h^2\,f(x_3)=0 \\ \\ -u_3+2\,u_4-\beta+h^2\,x_4\,u_4^3-h^2\,f(x_4)=0 \end{cases}$$

En posant $u_0=0$ et $u_{n+1}=0$ (pas de déformations aux extrémités) on obtient :

$$\begin{cases} 2\,u_1-u_2+h^2\,x_1\,u_1^3-h^2\,f(x_1)=0 \\ \\ -u_1+2\,u_2-u_3+h^2\,x_2\,u_2^3-h^2\,f(x_2)=0 \\ \\ -u_2+2\,u_3-u_4+h^2\,x_3\,u_3^3-h^2\,f(x_3)=0 \\ \\ -u_3+2\,u_4+h^2\,x_4\,u_4^3-h^2\,f(x_4)=0 \end{cases}$$

On peut par exemple résoudre ce système par la méthode de Newton, c'est une méthode itérative d'ordre deux donc elle converge rapidement. Nous allons illustrer cette méthode à l'aide d'un exemple. Soit à résoudre le système d'équations non-linéaires suivant :

$$f(U)=\begin{cases} f_1(u_1,u_2)=2\,u_1-u_2+e^{u_1}=0 \\ \\ f_2(u_1,u_2)=-u_1+2\,u_2+e^{u_2}=0 \end{cases} \tag{185}$$

Nous cherchons u_1 et u_2 telle que $f(U)=0$, avec le vecteur $U=(u_1,u_2)^T$. Le schéma numérique de la méthode de Newton pour résoudre le système (185) est :

$$U_{k+1}=U_k-\frac{f(U)}{\nabla f(U)}$$

$$[u_1^{k+1},u_2^{k+1}]=[u_1^k,u_2^k]-\frac{f(u_1,u_2)}{\nabla f(u_1,u_2)}$$

Avec $\nabla f(u_1, u_2)$ est la matrice Jacobienne :

$$\nabla f(u_1, u_2) = \begin{pmatrix} \dfrac{\partial f_1}{\partial u_1} & \dfrac{\partial f_1}{\partial u_2} \\ \dfrac{\partial f_2}{\partial u_1} & \dfrac{\partial f_2}{\partial u_2} \end{pmatrix} \tag{186}$$

On donne une valeur initiale au vecteur U_k ensuite on calcule les successeurs U_{k+1}. On cesse les itérations une fois le test d'arrêt est positif pour une tolérance donnée, par exemple $|-\frac{f(U)}{\nabla f(U)}| < \epsilon$.

3) ***En dimension 2****:* il existe une myriade de problèmes en physique et en chimie admettant comme formulation mathématique, une équation aux dérivées partielles. Rappelons que cette dernière exprime une relation fonctionnelle dont l'inconnue est une fonction de plusieurs variables. Dans l'équation même apparait la fonction de plusieurs variables recherchée ainsi que ses dérivées partielles. Soit $\rho : (x,t) \in [0,1]\times\mathbb{R}^+ \longmapsto \rho(x,t) \in \mathbb{R}$ une fonction continue, nous considérons un problème parabolique consistant à déterminer $u : (x,t) \in [0,1]\times\mathbb{R}^+ \longmapsto u(x,t) \in \mathbb{R}$ qui satisfait :

$$\mathcal{P} : \begin{cases} \tau\, c_p \dfrac{\partial u(x,t)}{\partial t} - k \dfrac{\partial^2 u(x,t)}{\partial x^2} = \rho(x,t) \\ u(x,0) = u_0(x) \quad 0 \le x \le 1 \\ u(0,t) = u(a,t) = 0 \quad 0 \le t \end{cases} \tag{187}$$

C'est l'équation de la diffusion de la chaleur, avec $\rho(x,t)$ est la source de chaleur. Les constantes positives τ, c_p et k, caractéristiques du matériau en question, représentent respectivement la densité volumique, la chaleur spécifique massique et la conductivité thermique. Afin d'alléger les écritures ces coefficients sont pris égaux à l'unité. A partir de ce problème, nous cherchons à déterminer la quantité de chaleur fournie au point x à l'instant t. A partir d'un développement de Taylor on démontre ces approximations des dérivées partielles :

$$\frac{\partial u(x,t)}{\partial t} \approx \frac{u(x_i, t_{j+1}) - u(x_i, t_j)}{\Delta t} + O(\Delta t) \tag{188}$$

$$\frac{\partial u(x,t)}{\partial t} \approx \frac{u(x_i, t_j) - u(x_i, t_{j-1})}{\Delta t} - O(\Delta t) \tag{189}$$

$$\frac{\partial^2 u(x,t)}{\partial x^2} \approx \frac{u(x_{i-1}, t_j) - 2\,u(x_i, t_j) + u(x_{i+1}, t_j)}{\Delta x^2} + O(\Delta x^2) \tag{190}$$

Nous utiliserons ces schémas des différences finies afin d'approcher $u(x,t)$ du problème ci-dessus. Nous cherchons des approximations $u(i,j)$ de la solution exacte $u(x_i,t_j)$ aux nœuds $(x_i,t_j) = (i\times\Delta x, j\times\Delta t)_{i,j=\{0,1,\ldots,n+1\}}$. Cette discrétisation définie un maillage ou une grille selon le domaine $\Omega = [i\times\Delta x\,,\, j\times\Delta t]^2$. La condition initiale $u(x,0) \simeq u(i\times\Delta x, 0) = 0$ signifie que la quantité de chaleur $u(i\times\Delta x, 0)$ est connue sur chaque nœud $x_i = i\times\Delta x$ à l'instant initial $(t=0)$. D'un autre côté, les conditions aux limites $u(0, j\times\Delta t) = u(a, j\times\Delta t) = 0$ signifient que la quantité de chaleur, apportée aux limites (ou aux bords) du domaine Ω, est nulle. En effet, à partir des équations (189), (190) et en posant $u(x_i,t_j) \simeq u(i,j)$ on obtient :

$$\frac{u(i,j) - u(i,j-1)}{\Delta t} - \frac{u(i-1,j) - 2u(i,j) + u(i+1,j)}{\Delta x^2} = \rho(i,j)$$

$$\Delta x^2\, u(i,j) - \Delta x^2\, u(i,j-1) - \Delta t\, u(i-1,j) + 2\,\Delta t\, u(i,j) - \Delta t\, u(i+1,j) = \rho(i,j)\,\Delta t\,\Delta x^2$$

Tenant compte des conditions aux limites $u(0, j \times \Delta t) = u(a, j \times \Delta t) = 0$ il vient :

$$\Delta x^2\, u(i,j) - \Delta x^2\, u(i,j-1) - \cancel{\Delta t\, u(i-1,j)} + 2\,\Delta t\, u(i,j) - \cancel{\Delta t\, u(i+1,j)} = \rho(i,j)\,\Delta t\,\Delta x^2$$

$$(\Delta x^2 + 2\,\Delta t)\, u(i,j) = \Delta x^2\, u(i,j-1) + \rho(i,j)\,\Delta t\,\Delta x^2$$

$$(I + \frac{2\,\Delta t}{\Delta x^2})\, u(i,j) = u(i,j-1) + \rho(i,j)\,\Delta t$$

$$(I + \frac{A_n\,\Delta t}{\Delta x^2})\, u(i,j) = u(i,j-1) + \rho(i,j)\,\Delta t$$

Tenant compte de la condition initiale $u(x,0) \simeq u(i \times \Delta x, 0) = 0$ le schéma numérique final devient :

$$(I + \frac{A_n\,\Delta t}{\Delta x^2})\, u(i,j+1) = u(i,j) + \rho(i,j)\,\Delta t$$

Avec I est la matrice identité. La matrice A_n est tridiagonale, symétrique et définie positive valant :

$$A_n = \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R}^n$$

Voici le script Matlab®

```
clear all ; clc ; close all ;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% LE 04/12/2019 Samir KENOUCHE : RESOLUTION DE L'EQUATION DE LA CHALEUR PAR
% DIFFERENCES FINIES EN DIMENSION 2

nb = 10 ; pas_temps = 0.01 ; pas_x = 1/(nb+1) ; T = 0.7 ;
xi = 0: pas_x :1 ; ti = 0: pas_temps :T ; ux = sin(pi.*xi) ; % Condition
                                                            % Initiale
u = zeros([numel(xi)-2 numel(ti)]) ; u(:, 1) = ux(2:end-1)' ;
fx = -2 + 6.*xi ; fn = zeros(size(u)) ;
mat_diag = 2*diag(ones(nb,1))-diag(ones(nb-1 ,1),1)-diag(ones(nb-1,1),-1) ;
my_mat = (eye(nb)+(pas_temps/pas_x.^2).*mat_diag) ; it = 1 ;

while  it < numel(ti)

     fn(:, it) = fx(2:end-1)' ;
     u(:, it+1) = my_mat \(u(:,it) + pas_temps*fn(:,it)) ; % DIVISION A GAUCHE
     it = it + 1 ;
end

ui = cat(1,zeros([1 numel(ti)]),u,zeros([1 numel(ti)])); % AJOUT DES CONDITION
                                                         % INITIALE ET AU LIMITE
fig1 = figure('color',[1 1 1]) ; [xn, tn] = meshgrid(xi, ti) ;
```

```
surf(xn, tn, ui') ; xlabel('Distance') ; ylabel('Temps') ; view(114, 18) ;

fig2 = figure('color',[1 1 1]) ; contourf(xn, tn, ui') ;
xlabel('Distance') ; ylabel('Temps') ;
%%%%%%%%%%%%%%%%%%%%%%%%%% PARTIE : ANIMATION %%%%%%%%%%%%%%%%%%%%%%%%%%
figure('Renderer','zbuffer') ; [xn, tn] = meshgrid(xi, ti) ;
colormap(jet) ; xlabel('Distance') ; ylabel('Temps') ;

for in = 1:30
    surf(exp(-0.008*in)*ui'.^2,ui')
    my_animation(in) = getframe ;
end
```

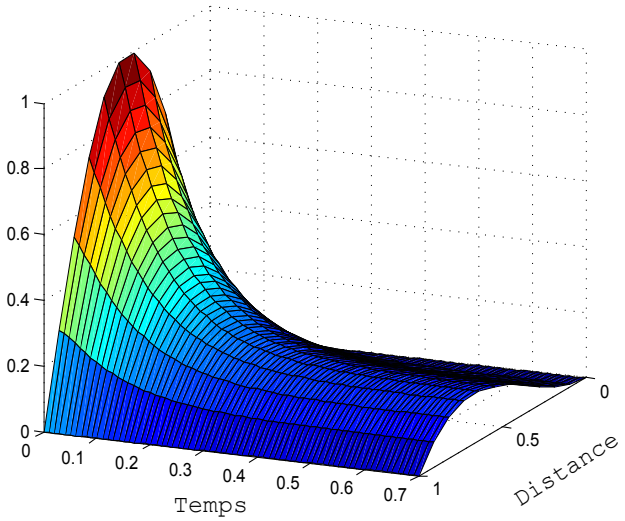


FIGURE 7: Surface de la solution approchée

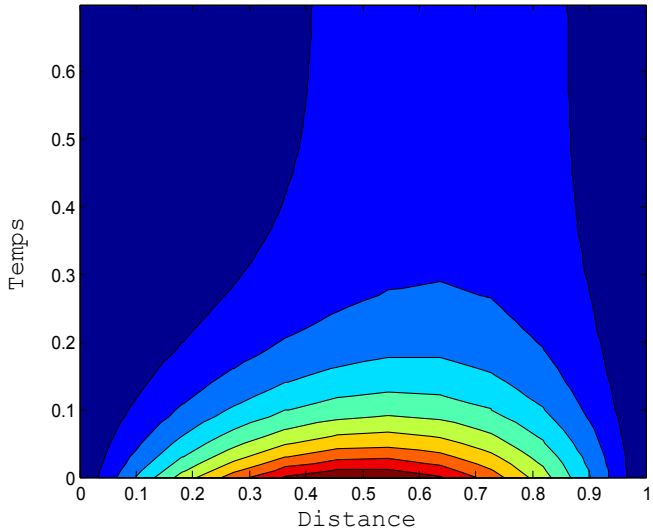


FIGURE 8: Courbe de niveaux de la solution approchée

III. Fonctions de Green

Considérons un système physique ou chimique, par exemple un oscillateur mécanique (peut être une liaison chimique) oscillant autour d'une position d'équilibre. Les oscillations d'amplitude $u(t)$ (sortie, ou réponse du système) sont causées par une excitation $\phi(t)$ (ou source des vibrations de la liaison chimique). D'un point de vue mathématique, la réponse $u(t)$ est une *fonctionnelle* de la source $\phi(t)$ qui s'écrit donc sous la forme : $u[\phi(t)]$. Nous rappelons que la différentielle d'une fonctionnelle quelconque $\mathcal{F}[f(t)]$ s'écrit comme suit :

$$\delta\mathcal{F}[f(t)] = \mathcal{F}[f(t)+\delta f(t)] - \mathcal{F}[f(t)] = \int_a^b \frac{\delta\mathcal{F}[f(t)]}{\delta f(t)}\,\delta f(t)\,dt \tag{191}$$

Appliquons désormais cette formule a notre fonctionnelle $u[\phi(t)]$, il vient :

$$\delta u[\phi(t)] = \int_a^b \underbrace{\frac{\delta u[\phi(t)]}{\delta\phi(\tau)}}_{h_0(t,\tau)}\,\delta\phi(\tau)\,d\tau, \qquad \forall t > \tau \tag{192}$$

L'introduction de la variable temporelle τ est justifiée par le *principe de causalité*[9], qui stipule que la cause $\phi(\tau)$ est systématiquement antérieur à l'effet ou à la réponse $u(t)$, de sorte que nous avons toujours $t > \tau$. Cela traduit une relation chronologique entre l'excitation et la réponse du système en question.

En revanche, la quantité $h_0(t,\tau) = \dfrac{\delta u[\phi(t)]}{\delta\phi(\tau)}$ exprime la *réponse propre ou intrinsèque* du système considéré. La connaissance de cette quantité mathématique est fondamentale, car elle permet de quantifier la réponse (ou la sortie) $u(t)$ quelque soit l'excitation (ou la source) $\phi(t)$. Cela nous amène à écrire le *produit de convolution*[10] suivant :

$$u(t) = \int_a^b \phi(\tau)\,h_0(t,\tau)\,d\tau \tag{193}$$

Dans ce cas de figure, l'excitation $\phi(t)$ est écrite sous forme d'une superposition (ou une somme) d'impulsions de *Dirac*, selon :

$$\phi(t) = \int_a^b \phi(\tau)\,\delta(t,\tau)\,d\tau \tag{194}$$

Il faut noter pour que l'équation (192) soit justifiée, le système considéré doit être *stable*. En effet, un système donné est dit *stable* si, en lui appliquant une excitation bornée quelconque, la réponse reste bornée par une constante, noté m. Traduit en terme mathématique, cela implique :

$$\int_a^b h_0(t,\tau)\,d\tau < m \tag{195}$$

Une autre notion fondamentale est *l'invariance par translation dans le temps*. La plupart des systèmes physiques respecte cette invariance, stipulant que si l'on retarde l'excitation de $\delta\tau$, alors la réponse est également retardée de $\delta\tau$. Cela se traduit par l'écriture :

$$h_0(t,\tau) = h_0(t-\tau) \tag{196}$$

9. La causalité constitue une contrainte majeure et inviolable pour la formalisation de nombreux problèmes en physique et en chimie. Il es résulte ce qui suit : $t < \tau \;\Rightarrow\; h_0(t,\tau) = 0$

10. Cette relation est linéaire, car elle vérifie les propriétés suivantes :

1) $\forall\alpha \in \mathbb{R} : \alpha\,\phi(t) \Rightarrow \alpha\,u(t)$.
2) $\forall\alpha,\beta \in \mathbb{R} : \alpha\,\phi_1(t) + \beta\,\phi_2(t) \Rightarrow \alpha\,u_1(t) + \beta\,u_2(t)$.

En effet, pour une excitation décalée $\phi_2(t-\delta\tau)=\phi_1(t)$ et d'après ce que nous avons vu précédemment, nous pouvons écrire :

$$\int_a^b \phi_1(\tau)\ h_0(t,\tau)\ d\tau = \int_a^b \phi_2(\tau-\delta\tau)\ h_0(t,\tau)\ d\tau \tag{197}$$

En considérant (196) et en opérant le changement de variable suivant $t^{'}=\tau-\delta\tau$, il vient :

$$\int_a^b \phi_1(\tau)\ h_0(t,\tau)\ d\tau = \int_a^b \phi_2(t^{'})\ h_0(t-\tau)\ dt^{'} \quad \Leftrightarrow \quad u_2(t)=u_1(t-\delta\tau) \tag{198}$$

Cela signifie que le système ne change pas ses caractéristiques, à mesure que le temps passe. Compte tenue de toutes les propriétés (causalité, linéarité et invariance) énumérées précédemment, la réponse se réécrit selon :

$$u(t)=\int_a^b \phi(\tau)\ h_0(t-\tau)\ d\tau \tag{199}$$

L'évolution temporelle associée à la grandeur $u(t)$ est ainsi simplement proportionnelle à la fonction $h_0(t-\tau)$. Une fois ces propriétés fondamentales sont rappelées, nous présenterons dans ce qui suit, le cadre théorique des fonctions de *Green*. Un problème aux limites d'ordre n est formulé mathématiquement comme suit :

$$\begin{cases} \hat{\mathcal{L}}(t)\,u(t)=\phi(t), \quad t\in[a,b] \\ \\ U_j(u)=\gamma_j, \quad j=1,m \end{cases} \tag{200}$$

Où,

$$U_j(u)=\sum_{i=0}^{n-1}\alpha_i^j\,u^{(i)}(a)+\beta_i^j\,u^{(i)}(b), \quad \forall\alpha_i^j,\beta_i^j\in\mathbb{R}, \quad n\leq m \tag{201}$$

Avec $\hat{\mathcal{L}}$ est un opérateur de dérivation linéaire. L'équation (201) traduit les conditions aux limites imposées à la solution $u(t)$. Le problème homogène associé à (200) s'écrit selon :

$$\begin{cases} \hat{\mathcal{L}}(t)\,u(t)=0, \quad t\in[a,b] \\ \\ U_j(u)=0, \quad j=1,m \end{cases} \tag{202}$$

Théorème de Fredholm *: Le problème aux limites* (200) *admet une solution unique si et seulement si, le problème homogène* $\hat{\mathcal{L}}(t)\,u(t)=0$, $t\in[a,b]$, $U_j(u)=0$, *admet comme solution triviale* $u(t)=0$.

Définition *:* $G(t,\tau)$ *est une fonction de Green du problème homogène, alors* $\exists$ *une fonction continue* $\phi(t)$ *telle que :*

$$u(t)=\int_a^b G(t,\tau)\,\phi(\tau)\,d\tau$$

soit une solution particulière du problème aux limites (200). *De plus, la fonction* $G(t,\tau)$ *doit satisfaire les propriétés suivantes :*

1) *Pour chaque* $\tau\in[a,b]$, *la fonction* $t\longmapsto G(t,\tau)$ *est solution de l'équation homogène* $\hat{\mathcal{L}}(t)\,u(t)=0$, $U_j(u)=0$, $\forall t\in[a,\tau]$ *et* $t\in[\tau,b]$. *Cela se traduit mathématiquement par :*

$$a_n(t)\,\frac{\partial^n G(t,\tau)}{\partial t^n}+a_{n-1}(t)\,\frac{\partial^{n-1}G(t,\tau)}{\partial t^{n-1}}+\cdots+a_1(t)\,\frac{\partial G(t,\tau)}{\partial t}+a_0(t)\,G(t,\tau)=0 \tag{203}$$

2) *Pour chaque* $\tau \in [a,\, b]$, *la fonction* $t \longmapsto G(t,\tau)$ *vérifie les conditions aux limites homogènes* $U_j(G(t,\tau)) = 0,\ j = 1, m$. *Autrement dit,*

$$\sum_{i=0}^{n-1}\left[\alpha_i^j\,\frac{\partial^i G(a,\tau)}{\partial t^i} + \beta_i^j\,\frac{\partial^i G(b,\tau)}{\partial t^i}\right] = 0, \quad \forall \alpha_i^j, \beta_i^j \in \mathbb{R}, \quad n \leq m \tag{204}$$

Cela se démontre en considérant (201), *selon :*

$$\begin{aligned} U_1 &= \alpha_0^1\,u(a) + \beta_0^1\,u(b) + \alpha_1^1\,u^{'}(a) + \beta_1^1\,u^{'}(b) \\ &= \int_a^b \left[\alpha_0^1\,G(a,\tau) + \beta_0^1\,G(b,\tau) + \alpha_1^1\,G^{'}(a,\tau) + \beta_1^1\,G^{'}(b,\tau)\right]\,\phi(\tau)\,d\tau \\ &= \int_a^b \underbrace{U_1(G(t,\tau))}_{car\ U_1(G(t,\tau))=0}\,\phi(\tau)\,d\tau = 0 \end{aligned}$$

3) *La fonction de Green est continue à* $t = \tau$ *:*

$$\lim_{t\to\underbrace{\tau+\epsilon}_{\tau^+}} G(t,\tau) = \lim_{t\to\underbrace{\tau-\epsilon}_{\tau^-}} G(t,\tau) \quad \Leftrightarrow \quad \lim_{\tau>t} G(t,\tau) = \lim_{\tau<t} G(t,\tau) \tag{205}$$

4) *Une propriété intéressante de la fonction de Green, est la présence d'une discontinuité d'une amplitude valant* $\dfrac{1}{a_n(t)}$ *de sa dérivée d'ordre* $(n-1)$, *à* $t = \tau$ *:*

$$\lim_{t\to\tau^+}\frac{\partial G^{n-1}_{t>\tau}(t,\tau)}{\partial t^{n-1}} - \lim_{t\to\tau^-}\frac{\partial G^{n-1}_{t<\tau}(t,\tau)}{\partial t^{n-1}} = \frac{1}{a_n(t)} \tag{206}$$

Avec $a_n(t)$ *étant le coefficient de la dérivée d'ordre* n *de* $u(t)$.

Les fonctions de *Green* permettent de transformer une équation différentielle (ou aux dérivées partielles) en une équation intégrale, généralement plus simple à résoudre. Ces fonctions sont largement utilisées en Chimie et Physique quantique. L'équation (200) est explicitée sous la forme suivante :

$$\sum_{k=0}^{n} a_k(t)\,u^{(k)}(t) = \phi(t) \Longleftrightarrow \hat{\mathcal{L}}(t)\,u(t) = \phi(t), \qquad \hat{\mathcal{L}}(t) = \sum_{k=0}^{n} a_k(t)\,\frac{d^k}{dt^k} \tag{207}$$

Les coefficients $a_k(t)$ peuvent aussi êtres constants. La fonction de *Green* associée à l'équation ci-dessus, est obtenue en remplaçant la source $\phi(t)$ par une impulsion appliquée à un instant $\tau > t$, soit $\delta(t-\tau)$. La source $\phi(t)$ pouvant être une force agissant sur une particule par exemple, ou encore une source de chaleur ou de vibration ... etc. Quelque soit la nature de cette source, elle peut être vue comme une somme d'impulsions appliquées à différents instants $\tau > t$. Cela se traduit par mathématiquement par l'écriture :

$$\phi(t) = \int_0^{\infty} \phi(\tau)\,\delta(t-\tau)d\tau \tag{208}$$

Où $\phi(\tau)$ est la fonction de poids associée à chaque impulsion. De cette façon, la source $\phi(t)$ est totalement reconstituée par le second membre de l'équation (208). Revenons maintenant à notre équation différentielle (207), si l'on note $G(t,\tau)$ la fonction de *Green*[11] alors :

$$\sum_{k=0}^{n} a_k(t)\,\frac{\partial^k G(t,\tau)}{\partial t^k} = \delta(t-\tau) \Longleftrightarrow \hat{\mathcal{L}}(t)\,G(t,\tau) = \delta(t-\tau) \tag{209}$$

11. Avec τ est une variable factice, ou *dummy variable* en anglais

En développant (209) à l'ordre deux et en optant pour des coefficients constants, il vient :

$$a_2\, G^{''}(t,\tau) + a_1\, G^{'}(t,\tau) + a_0\, G(t,\tau) = \delta(t-\tau) \tag{210}$$

Résoudre l'équation (210), revient à déterminer d'abord la solution générale ($\hat{\mathcal{L}}(t)\, G(t,\tau) = 0$), à laquelle nous ajoutons la solution particulière ($\hat{\mathcal{L}}(t)\, G(t,\tau) = \delta(t-\tau)$). Multiplions (210) par la fonction de poids $\phi(\tau)$, il vient :

$$a_2 \frac{\partial^2}{\partial t^2}\left[\phi(\tau)\ G(t,\tau)\right] + a_1 \frac{\partial}{\partial t}\left[\phi(\tau)\ G(t,\tau)\right] + a_0\ \left[\phi(\tau)\ G(t,\tau)\right] = \phi(\tau)\ \delta(t-\tau) \tag{211}$$

Intégrons les deux membres de cette équation et nous obtenons :

$$a_2 \frac{\partial^2}{\partial t^2}\int_0^\infty \left[\phi(\tau)\ G(t,\tau)\right] d\tau + a_1 \frac{\partial}{\partial t}\int_0^\infty \left[\phi(\tau)\ G(t,\tau)\right] d\tau + a_0 \int_0^\infty \left[\phi(\tau)\ G(t,\tau)\right] d\tau = \underbrace{\int_0^\infty \phi(\tau)\ \delta(t-\tau) d\tau}_{\phi(t)}$$

De cette manière nous avons totalement satisfait l'équation différentielle (207). D'après ce dernier résultat, nous comprenons que l'intégrale :

$$u_p(t) = \int_0^\infty \phi(\tau)\ G(t,\tau)\, d\tau \tag{212}$$

constitue une solution particulière de l'équation différentielle (207). La solution globale (générale + particulière) sera donnée par :

$$u(t) = u_h(t) + u_p(t) = u_h(t) + \int_0^\infty \phi(\tau)\ G(t,\tau)\, d\tau \tag{213}$$

Où $u_h(t)$ dénote la solution générale de l'équation homogène (sans second membre, $\hat{\mathcal{L}}(t)\, u(t) = 0$) associée. Ainsi, si l'on connait la fonction de *Green* $G(t,\tau)$ d'une équation différentielle, une solution particulière s'obtient par l'équation intégrale $u_p(t)$. Toute la problématique consiste donc à déterminer la "bonne" fonction de *Green* du phénomène physique ou chimique étudié. Toutefois, la détermination de $G(t,\tau)$ est totalement tributaire de la connaissance de la solution de l'équation homogène $u_h(t)$.

Par ailleurs, nous aurions pu démontrer l'équation (213), en utilisant une des propriétés de la fonction de *Dirac* à savoir :

$$\int_0^\infty \phi(\tau)\ \delta(t-\tau)\, d\tau = \phi(t) \tag{214}$$

D'après (209), nous avons :

$$\hat{\mathcal{L}}(t)\, G(t,\tau) = \hat{\mathcal{L}}(t) \int_0^\infty \phi(\tau)\, G(t,\tau)\, d\tau = \underbrace{\int_0^\infty \phi(\tau)\, \delta(t-\tau)\, d\tau}_{\phi(t)} \tag{215}$$

L'opérateur différentiel $\hat{\mathcal{L}}(t)$ est indépendant de τ, par conséquent et comme précédemment une solution de (207) s'obtient selon :

$$u_p(t) = \int_0^\infty \phi(\tau)\ G(t,\tau)\, d\tau \tag{216}$$

Il est important de rappeler que la fonction $G(t,\tau)$ satisfait pleinement les conditions aux limites imposées à la solution $u(t)$.

Exercice d'application : on se propose d'étudier un problème aux limites du second ordre dont la formulation mathématique est donnée comme suit :

$$\mathcal{P}:\begin{cases} u''(t)=\phi(t) \quad \text{avec} \quad 0\leq t\leq 2\pi \\ \\ u(0)=u(2\pi)=1 \end{cases}$$

D'après ce que nous avons discuté précédemment, la solution de $(\mathcal{P})$ s'écrira comme :

$$u(t)=u_h(t)+\int_0^{\infty}\phi(\tau)\;G(t,\tau)\,d\tau \tag{217}$$

De plus, à partir de $(\mathcal{P})$, nous écrivons l'équation satisfaite par la fonction de *Green* associée soit :

$$G''(t,\tau)=\delta(t-\tau) \tag{218}$$

Commençons d'abord par résoudre l'équation homogène associée :

$$G''(t,\tau)=0 \quad \Rightarrow \quad G(t,\tau)=\begin{cases} a_1\,t+b_1 & 0\leq t<\tau \\ \\ a_2\,t+b_2 & \tau<t\leq 2\pi \end{cases} \tag{219}$$

Nous allons étudier le cas où $t<\tau$, ou de façon totalement équivalente quand $0\leq t<\tau$ d'après la condition initiale. En effet, tenant compte de la condition $u(0)=1$, exprimée dans $(\mathcal{P})$, il vient :

$$G_{t<\tau}(t,\tau)=a_1\,t+b_1 \quad \Rightarrow \quad G_{t<\tau}(t=0,\tau)=b_1=1 \tag{220}$$

$$\Rightarrow \quad G_{t<\tau}(t,\tau)=a_1 t+1 \tag{221}$$

Étudions également le cas où $\tau<t$, ou de façon totalement équivalente quand $\tau<t\leq 2\pi$ d'après les conditions aux limites. Ainsi, tenant compte de la condition $u(2\pi)=1$, exprimée dans $(\mathcal{P})$, il vient :

$$G_{t>\tau}(t,\tau)=a_2\,t+b_2 \quad \Rightarrow \quad G_{t>\tau}(t=2\pi,\tau)=2\,a_2\pi+b_2=1 \quad \Rightarrow b_2=1-2\,a_2\pi \tag{222}$$

$$\Rightarrow \quad G_{t>\tau}(t,\tau)=a_2\,(t-2\pi)-1 \tag{223}$$

Nous obtenons la fonction de *Green* suivante :

$$G(t,\tau)=\begin{cases} a_1\,t+1 & 0\leq t<\tau \\ \\ a_2\,(t-2\pi)+1 & \tau<t\leq 2\pi \end{cases} \tag{224}$$

Étudions désormais le cas où $t=\tau$, nous savons que dans ce cas de figure la fonction $G(t,\tau)$ est continue, soit :

$$\lim_{\tau>t} G(t,\tau)=\lim_{\tau<t} G(t,\tau) \tag{225}$$

Alors,

$$t=\tau \quad \Rightarrow \quad G_{t<\tau}(t,\tau)=G_{t>\tau}(t,\tau) \quad \Leftrightarrow \quad a_1\tau+1=a_2\,(\tau-2\pi)+1 \tag{226}$$

Pour déterminer les constantes a_1 et a_2, exploitons la discontinuité de $G(t,\tau)$ par rapport à sa dérivée d'ordre un, soit :

$$\int_{\tau-\epsilon}^{\tau+\epsilon}\frac{\partial^2 G(t,\tau)}{\partial t^2}\,dt=\int_{\tau-\epsilon}^{\tau+\epsilon}\delta(t-\tau)\,dt=\lim_{t\to\tau^+}\frac{\partial G_{t>\tau}(t,\tau)}{\partial t}-\lim_{t\to\tau^-}\frac{\partial G_{t<\tau}(t,\tau)}{\partial t}=a_2-a_1=1 \tag{227}$$

$$\Rightarrow \quad a_1 = a_2 - 1 \tag{228}$$

D'après (226) et (228), il en ressort :

$$(a_2 - 1)\,\tau = a_2\,(\tau - 2\pi) \quad \Rightarrow \quad a_2 = \frac{\tau}{2\pi} \quad \text{et donc} \quad a_1 = \frac{\tau - 2\pi}{2\pi} \tag{229}$$

Finalement, la fonction de *Green* associée à $\mathcal{P}$, s'écrit sous la forme :

$$G(t,\tau) = \begin{cases} \dfrac{\tau - 2\pi}{2\pi}\,t & \text{pour} \quad t < \tau \\ \\ \dfrac{\tau}{2\pi}\,(t - 2\pi) & \text{pour} \quad t > \tau \end{cases} \tag{230}$$

Les deux formules apparaissant dans (230) sont les expressions d'une seule et même fonction à savoir $G(t,\tau)$. Chaque expression est valide dans une région donnée, $t < \tau$ ou $t > \tau$. Toutefois, la fonction de *Green* $G(t,\tau) = \dfrac{\tau}{2\pi}\,(t - 2\pi)$ est celle qui répond le plus aux considérations de la physique, sachant que la réponse du système $u(t)$ intervient après une impulsion appliquée à un instant antérieur. Quelque soit la forme mathématique de la fonction source $\phi(t)$ (qui est connue), la solution finale de notre équation différentielle sera obtenue par (217) :

$$u(t) = u_h + \frac{(t - 2\pi)}{2\pi} \int_0^{2\pi} \underbrace{\phi(\tau)}_{\text{fonction donnée}} \tau\, d\tau \tag{231}$$

Cherchons la solution u_h de l'équation homogène $u'' = 0$. L'équation caractéristique s'écrit :

$$\lambda_0^2 = 0 \quad \Rightarrow \quad \lambda_0 = 0 \tag{232}$$

Le discriminant de l'équation caractéristique $\Delta = b^2 - 4\,a\,c = 0$, c'est une racine double et la solution de l'équation homogène est donnée par :

$$u_h(t) = (c_1 + c_2\,t)\,e^{\lambda_0\,t} \quad \forall c_1, c_2 \in \mathbb{R} \tag{233}$$

Il es ressort la solution globale suivante :

$$u(t) = (c_1 + c_2\,t) + \frac{(t - 2\pi)}{2\pi} \int_0^{2\pi} \underbrace{\phi(\tau)}_{\text{fonction donnée}} \tau\, d\tau, \qquad \forall c_1, c_2 \in \mathbb{R} \tag{234}$$

Par exemple, si la source $\phi(t) = \sin(w_0\,t)$, alors :

$$u(t) = (c_1 + c_2\,t) + \frac{(t - 2\pi)}{2\pi} \int_0^{2\pi} \sin(w_0\,\tau)\,\tau\, d\tau, \qquad \forall c_1, c_2 \in \mathbb{R} \tag{235}$$

Un intégration par partie nous donne :

$$\int u\, dv = u\,v - \int v\, du \tag{236}$$

Posons,

$$u = \tau \quad \Rightarrow \quad du = d\tau$$

$$dv = \sin(w_0\,\tau)\, d\tau \quad \Rightarrow \quad v = \frac{-\cos(w_0\,t)}{w_0}$$

$$\int_0^{2\pi} \sin(w_0\,\tau)\,\tau\, d\tau = \frac{\sin(2\pi\,w_0) - 2\pi\,w_0\,\cos(w_0\,2\pi)}{w_0^2} \tag{237}$$

La solution finale est :

$$u(t) = (c_1 + c_2\, t) + \frac{(t - 2\pi)}{2\pi}\left[\frac{\sin(2\pi\, w_0) - 2\pi\, w_0\, \cos(w_0\, 2\pi)}{w_0^2}\right], \qquad \forall c_1, c_2 \in \mathbb{R} \tag{238}$$

A. *Fonction de Green de l'équation de Schrödinger*

En mécanique quantique, les particules élémentaires, apparaissent pour un observateur donné, se comporter comme des ondes appelées ondes *de Broglie*. La mécanique de ces objets devient celle du mouvement des ondes et les fonctions d'onde sont utilisées pour décrire le comportement des ces systèmes quantiques. Le module au carré de la fonction d'onde traduit la probabilité qu'une particule existe en un point donné de l'espace. C'est pourquoi, les fonctions d'ondes sont parfois appelées *ondes de probabilité* pouvant êtres dispersées par un potentiel atomique ou nucléaire $V(r)$. Nous cherchons une solution de l'équation de Schrödinger stationnaire, par exemple, pour le mouvement d'électrons d'énergie E dans un champ d'énergie potentielle $V(r)$ [12]. Si le potentiel est un diffuseur élastique et les ondes de *de Broglie* décrivent des particules non relativistes, l'équation aux dérivées partielles (indépendante du temps) qui décrit le mieux cet effet (de diffusion) est donnée par :

$$-\frac{\hbar^2}{2\, m}\, \nabla^2\psi(r) + V(r)\, \psi(r) = E\, \psi(r) \tag{239}$$

Où k est le nombre d'onde et $V(r)$ est une inhomogénéité qui est responsable de la diffusion de l'onde $\psi(r)$ qui est donc parfois appelé diffuseur. Cette équation est connue sous le nom d'équation de *Schrödinger* d'après le physicien théoricien autrichien *Erwin Schrodinger* qui l'a postulée dans les années 1920. Après réarrangement, nous obtenons :

$$\nabla^2\psi(r) + \underbrace{\frac{2\, m\, E}{\hbar^2}}_{k^2}\, \psi(r) = \underbrace{\frac{2\, m}{\hbar^2}\, V(r)\, \psi(r)}_{Q(r)} \tag{240}$$

$$\Rightarrow \quad \left[\nabla^2 + k^2\right]\, \psi(r) = Q(r) \tag{241}$$

La relation (241) a la forme de l'équation de *Helmhotz*. Notons, cependant, que le terme inhomogène $Q(r)$ dépend explicitement de la solution $\psi(r)$. Nous commencerons par étudier la solution de la fonction de *Green* vérifiant l'équation de *Helmholtz* inhomogène. Comme pour l'exemple d'application précédent, la fonction de *Green* associée à (241) doit être solution de l'équation suivante :

$$\Rightarrow \quad \left[\nabla^2 + k^2\right]\, G(r) = \delta(r) \tag{242}$$

Ainsi, la solution particulière de (241) est :

$$\psi_p(r) = \int_{-\infty}^{+\infty} Q(r_0)\, G(r - r_0)\, dr_0 \tag{243}$$

Nous pouvons vérifier facilement que $\psi_p(r)$ soit solution de (242), autrement dit :

$$\left[\nabla^2 + k^2\right]\, \psi_p(r) = \int_{-\infty}^{+\infty} \left[\nabla^2 + k^2\right]\, G(r - r_0)\, Q(r_0) dr_0 = \int_{-\infty}^{+\infty} \delta(r - r_0)\, Q(r_0)\, dr_0 = Q(r) \tag{244}$$

Toute la problématique consiste donc à déterminer la fonction $G(r)$. D'après ce que nous avons déjà vu, la fonction $G(r)$ pour une équation différentielle donnée représente la réponse à une impulsion $\delta(r)$. Nous allons d'abord résoudre l'équation (242). Pour se faire, nous recourrons à la *Transformée de Fourier* afin de transformer l'équation différentielle (242) en une équation algébrique plus simple à résoudre. Multiplions (242) par le facteur e^{jsr}, il vient :

12. En théorie quantique, il est d'usage d'appeler potentiel tout court, l'énergie potentielle $V(r)$ qui est effectivement le produit d'un potentiel par une charge.

$$\int_{-\infty}^{+\infty}\left[\nabla^2+k^2\right]G(r)\,e^{jsr}dr=\int_{-\infty}^{+\infty}\delta(r)\,e^{jsr}=1\quad\Rightarrow\quad\int_{-\infty}^{+\infty}\nabla^2\,G(r)\,e^{jsr}dr+k^2\int_{-\infty}^{+\infty}G(r)\,e^{jsr}dr=1$$

$$\Rightarrow\quad(js)^2\int_{-\infty}^{+\infty}G(r)\,e^{jsr}dr+k^2\int_{-\infty}^{+\infty}G(r)\,e^{jsr}dr=1\quad\Rightarrow\quad-s^2\int_{-\infty}^{+\infty}G(r)\,e^{jsr}dr+k^2\int_{-\infty}^{+\infty}G(r)\,e^{jsr}dr=1$$

$$\Rightarrow\quad\left[k^2-s^2\right]\underbrace{\int_{-\infty}^{+\infty}G(r)\,e^{jsr}\,dr}_{g(s)}=1$$

Finalement, nous obtenons la *Transformée de Fourier* de la fonction de *Green* dans l'espace réciproque $s\equiv\dfrac{1}{r}$, soit :

$$\Rightarrow\quad g(s)=\frac{1}{k^2-s^2}\tag{245}$$

Pour revenir à l'espace réel, nous devons effectuer une *Transformée de Fourier inverse*. Par définition nous avons :

$$G(r)=\frac{1}{(2\,\pi)^3}\int_{-\infty}^{+\infty}g(s)\,e^{-jsr}\,ds\quad\Rightarrow\quad G(r)=\frac{1}{(2\,\pi)^3}\int_{-\infty}^{+\infty}\frac{e^{-jsr}}{k^2-s^2}\,ds\tag{246}$$

La particule évolue dans un espace en trois dimension, pour représenter cette réalité physique nous devons adopter le système de coordonnées sphériques pour résoudre l'intégrale ci-dessus. L'élément de volume en coordonnées sphériques s'écrit :

$$ds=s^2\,ds\,\sin(\theta)\,d\theta d\phi\tag{247}$$

Et d'un autre côté, nous avons aussi

$$e^{-jsr}=e^{-j\,|s|\,|r|\,\cos(\theta)}\quad\text{C'est un produit scalaire entre deux vecteurs}\tag{248}$$

Afin de ne pas alourdir les écritures mathématiques, nous gardons $e^{-jsr\,\cos(\theta)}$. Par conséquent l'intégrale précédente devient :

$$\Rightarrow\quad G(r)=\frac{1}{(2\,\pi)^3}\int_0^{2\,\pi}d\phi\int_0^{+\infty}\frac{s^2}{k^2-s^2}\,ds\int_0^{\pi}\sin(\theta)\,e^{-jsr\,\cos(\theta)}\,d\theta\tag{249}$$

$$\Rightarrow\quad G(r)=-\frac{1}{(2\,\pi)^2}\int_0^{+\infty}\frac{s^2}{k^2-s^2}\,ds\int_0^{\pi}\sin(\theta)\,e^{-jsr\,\cos(\theta)}\,d\theta\tag{250}$$

Afin de résoudre l'intégrale relative à la variable θ, nous devons procéder au changement de variable suivant :

$$z=\cos(\theta)\quad\Rightarrow\quad dz=-\sin(\theta)\,d\theta$$

Évidemment, les bornes d'intégration changent aussi avec la nouvelle variable d'intégration z, selon :

$$\theta=0\quad\Rightarrow\quad z=+1$$
$$\theta=\pi\quad\Rightarrow\quad z=-1$$

$$\Rightarrow\quad G(r)=+\frac{1}{(2\,\pi)^2}\int_0^{+\infty}\frac{s^2}{k^2-s^2}\,ds\int_{-1}^{1}e^{-jsrz}\,dz\tag{251}$$

$$\Rightarrow\quad G(r)=+\frac{1}{(2\,\pi)^2}\int_0^{+\infty}\frac{s^2}{k^2-s^2}\,ds\left[\frac{e^{-jsrz}}{j\,s\,r}\right]_{z=-1}^{z=1}\tag{252}$$

$$\Rightarrow \quad G(r) = \frac{1}{j\,r\,(2\,\pi)^2} \int_0^{+\infty} \frac{s\left[e^{jsr} - e^{-jsr}\right]}{k^2 - s^2}\, ds \tag{253}$$

Par ailleurs, rappelons que :

$$\int_{-\infty}^{+\infty} f(x)\, dx = 2 \int_0^{+\infty} f(x)\, dx \tag{254}$$

L'intégrale (253) devient :

$$\Rightarrow \quad G(r) = \frac{j}{8\,r\,\pi^2} \int_{-\infty}^{+\infty} \frac{s\left[e^{jsr} - e^{-jsr}\right]}{s^2 - k^2}\, ds \tag{255}$$

Avec,

$$\frac{1}{(s^2 - k^2)} = \frac{1}{(s-k)\,(s+k)} \tag{256}$$

$$\Rightarrow \quad G(r) = \frac{j}{8\,r\,\pi^2} \left[\int_{-\infty}^{+\infty} \frac{s\; e^{jsr}}{(s-k)\,(s+k)}\, ds - \int_{-\infty}^{+\infty} \frac{s\; e^{-jsr}}{(s-k)\,(s+k)}\, ds\right] \tag{257}$$

Les deux intégrales du second membre sont le contour du demi-cercle à $k = \pm s$. Nous aurons une rotation dans le sens des aiguilles d'une montre à $-k$ et le chemin inverse à $+k$. Ces deux intégrales sont résolues en utilisant *la formule de Cauchy* suivante :

$$\oint \frac{f(u)}{u - u_0} = 2\pi\, j\, f(u_0) \tag{258}$$

En appliquant cette formule, nous obtenons :

$$\Rightarrow \quad G(r) = \frac{j}{8\,r\,\pi^2} \left[2\pi\, j \left[\frac{s\; e^{jsr}}{s+k}\right]_{s=k} - (-2\pi\, j) \left[\frac{s\; e^{-jsr}}{s-k}\right]_{s=-k}\right] \tag{259}$$

Nous obtenons finalement la fonction de *Green* pour le problème (242) :

$$\Rightarrow \quad G(r) = -\frac{1}{4\pi\, r}\, e^{jk\,r} \tag{260}$$

Ainsi la solution particulière (243), prend la forme [13] :

$$\psi_p(r) = -\frac{1}{4\pi\, r} \int_{-\infty}^{+\infty} \frac{e^{jk\,|r - r_0|}}{|r - r_0|}\, Q(r_0)\, dr_0 \tag{261}$$

Rappelons que nous avons déjà posé :

$$Q(r) = \frac{2\,m}{\hbar^2}\, V(r)\, \psi(r)$$

D'où,

$$\psi_p(r) = -\frac{m}{2\pi\,\hbar^2} \int_{-\infty}^{+\infty} \frac{e^{jk\,|r - r_0|}}{|r - r_0|}\, V(r_0)\, \psi(r_0)\, dr_0 \tag{262}$$

La solution globale (solution homogène $\psi_h(r)$ + solution particulière $\psi_p(r)$) de l'équation de schrödinger est :

$$\psi(r) = \psi_h(r) - \frac{m}{2\pi\,\hbar^2} \int_{-\infty}^{+\infty} \frac{e^{jk\,|r - r_0|}}{|r - r_0|}\, V(r_0)\, \psi(r_0)\, dr_0 \tag{263}$$

13. Afin de ne pas alourdir les écritures mathématiques, le signe vecteur est délibérément omis. Ainsi, le vecteur r_0 décrit tout les points où règne le potentiel, d'un autre côté, le vecteur r décrit les points à partir desquels nous observons la fonction d'onde.

La solution de l'équation homogène :

$$\left[\nabla^2 + k^2\right] \psi_h(r) = 0 \tag{264}$$

vaut $\psi_h(r) = e^{jk\,r}$, onde plane d'amplitude unitaire. Ainsi, la solution globale (263) devient :

$$\psi(r) = e^{jk\,r} - \frac{m}{2\pi\,\hbar^2} \int_{-\infty}^{+\infty} \frac{e^{jk\,|r-r_0|}}{|r-r_0|}\, V(r_0)\, \psi(r_0)\, dr_0 \tag{265}$$

L'équation (265) est la forme intégrale de l'équation de schrödinger. Elle est totalement équivalente à la forme différentielle, plus familière. Le "paradoxe" de cette équation intégrale réside dans le fait que nous devons d'abord connaitre la solution $\psi(r)$ afin de résoudre cette intégrale, or c'est ce que nous cherchons !. Avant de répondre à cette question, supposons que $r_0 \approx 0$ cela implique que $V(r_0) \approx 0$ pour des régions loin du centre de diffusion du potentiel. Autrement dit, pour de longues distances du centre de diffusion nous avons $|r| > |r_0|$, ce qui conduit à l'approximation suivante :

$$\frac{e^{jk\,|r-r_0|}}{|r-r_0|} \approx \frac{e^{jk\,r}}{r}\, e^{-jk\,r_0} \tag{266}$$

Par voie de conséquence, l'équation (265) devient :

$$\psi(r) = e^{jk\,r} - \frac{m}{2\pi\,\hbar^2}\, \frac{e^{jk\,r}}{r} \int_{-\infty}^{+\infty} e^{-j\,k\,r_0}\, V(r_0)\, \psi(r_0)\, dr_0 \tag{267}$$

Il faut bien rappeler que le phénomène physique rattaché à cette équation, est celui d'une onde plane incidente, $\psi(z) = e^{jk_i\,r}$ voyageant dans la direction $\vec{u}_i$, qui rencontre un potentiel de diffusion $V(r_0)$, produisant une onde sphérique sortante $e^{jk_s\,r_0}$ dans la direction $\vec{u}_s$. Comme une image, nous pouvons imaginer la situation d'une vague d'eau (donc une onde) rencontrant une roche qui va la dévier selon une direction donnée. Le nombre d'onde k est liée à l'énergie de la particule à travers la relation :

$$k \equiv \frac{\sqrt{2\,m\,E}}{\hbar}$$

Tenant compte des directions d'incidence et d'émergence, la solution (267) devient :

$$\psi(r) = e^{jk_i\,r} - \frac{m}{2\pi\,\hbar^2}\, \frac{e^{jk_i\,r}}{r} \int_{-\infty}^{+\infty} e^{-j\,k_s\,r_0}\, V(r_0)\, \psi(r_0)\, dr_0 \tag{268}$$

Afin de résoudre cette équation intégrale, nous devons passer par *l'approximation de Born*[14]. Cela consiste à écrire :

$$\psi(r_0) = e^{jk_i\,r_0} - \frac{m}{2\pi\,\hbar^2}\, \frac{e^{jk_i\,r}}{r'} \int_{-\infty}^{+\infty} e^{-j\,k_s\,r_0'}\, V(r_0')\, \psi(r_0')\, dr_0' \tag{269}$$

Substituons ensuite l'équation (269) dans (268), il vient :

$$\psi(r) = \underbrace{e^{jk_i\,r} - \frac{m}{2\pi\,\hbar^2}\, \frac{e^{jk_i\,r}}{r} \int_{-\infty}^{+\infty} e^{-j\,k_s\,r_0}\, V(r_0)\, e^{jk_i\,r_0}\, dr_0}_{\text{premier approx. de Born}} - \frac{m}{2\pi\,\hbar^2}\, \frac{e^{jk_i\,r}}{r'} \int_{-\infty}^{+\infty} e^{-j\,k_s\,r_0'}\, V(r_0')\, \psi(r_0')\, dr_0' + \cdots$$

Si l'on tronque cette série à la première approximation de *Born*, la solution approximative prend la forme suivante :

$$\psi(r) = e^{jk_i\,r} - \frac{m}{2\pi\,\hbar^2}\, \frac{e^{jk_i\,r}}{r} \int_{-\infty}^{+\infty} e^{-j\,k_s\,r_0}\, V(r_0)\, e^{jk_i\,r_0}\, dr_0 \tag{270}$$

14. Approximation ayant été introduit dans les années 1920 par *Max Born*, physicien théoricien Allemand.

Ou encore :

$$\psi(r) = e^{jk_i\,r} - \frac{e^{jk_i\,r}}{r}\underbrace{\frac{m}{2\pi\,\hbar^2}\int_{-\infty}^{+\infty} e^{j\,(k_i-k_s)\,r_0}\,V(r_0)\,dr_0}_{f(\theta,\phi)} \tag{271}$$

Avec,

$$f(\theta,\phi) = -\frac{m}{2\pi\,\hbar^2}\int_{-\infty}^{+\infty} e^{j\,\kappa\,r_0}\,V(r_0)\,dr_0, \quad \kappa \equiv k_i - k_s \quad \text{et} \quad r > r_0 \tag{272}$$

Désormais la connaissance de $V(r)$ permet la détermination de la solution $\psi(r)$ par le biais de l'équation intégrale (271). La dépendance angulaire de l'amplitude de l'onde émergente f est portée par le nombre d'onde κ. Tout la problématique consiste donc à déterminer l'amplitude de l'onde émergente $f(\theta,\phi)$, ce facteur indique la probabilité de l'émergence dans une direction donnée θ.

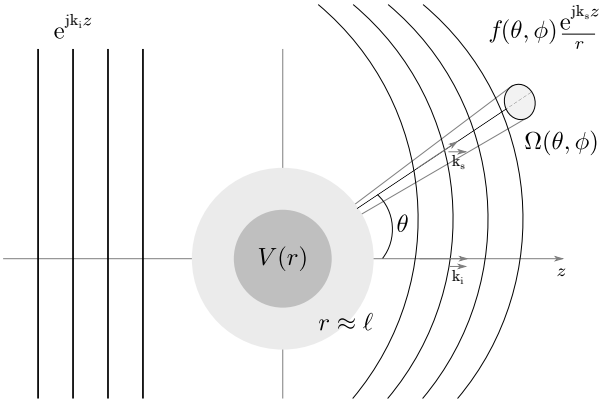


FIGURE 9: Onde plane incidente ($\vec{k}_i$), diffusée par le centre de diffusion $V(r)$ ayant une porté de l'ordre de$l$. Le résultat est une onde sphérique sortante ($\vec{k}_s$) ayant une amplitude $f(\theta,\phi)$ dans l'angle solide Ω.

D'après le schéma, nous pouvons écrire :

$$\sin(\theta/2) = \frac{\frac{\kappa}{2}}{k} \quad \Rightarrow \quad \kappa = k_s - k_i = 2k\,\sin(\theta/2) \tag{273}$$

Où k_i et k_s ont strictement la même amplitude que le nombre d'onde k, mais le premier pointe dans la direction du faisceau incident, tandis que le second pointe vers le détecteur. Cette égalité des amplitudes découle du caractère élastique de l'interaction entre l'onde incidente et le centre de diffusion $V(r)$. Autrement dit, cette interaction ne change pas l'énergie de l'onde incidente, d'oú $|k_i| = |k_s| = |k|$. L'amplitude de l'onde émergente $f(\theta,\phi)$, se réécrit selon :

$$f(\theta,\phi) = -\frac{m}{2\pi\,\hbar^2}\int_{-\infty}^{+\infty} e^{-j\,2\,k\,\sin(\theta/2)\,r_0}\,V(r_0)\,dr_0, \quad r > r_0 \tag{274}$$

Pour une diffusion de *faible énergie*, autrement dit, si l'onde incidente k_i n'est pas trop altérée par le potentiel $V(r_0)$, nous pouvons écrire :

$$E = \underbrace{\frac{2\pi}{\lambda}}_{k}\hbar\,c$$

$$E \to 0 \quad \Rightarrow \quad \lambda \to \infty \quad \Rightarrow \quad k \to 0 \quad \Rightarrow \quad e^{-j\,2\,k\,\sin(\theta/2)\,r_0} \to 1$$

$$\Rightarrow \quad f(\theta) \approx -\frac{m}{2\pi\,\hbar^2}\int_{-\infty}^{+\infty} V(r_0)\,dr_0, \quad r > r_0 \tag{275}$$

En considérant un *potentiel sphérique* et quelque soit l'énergie mobilisée (donc pas forcément dans le cas des faibles énergies), nous exprimons r_0 en coordonnées sphériques. A partir de (272), dans l'argument de l'exponentielle nous avons un produit scalaire entre deux vecteurs :

$$\kappa\, r_0 = ||\kappa||\, ||r_0||\, \cos(\theta_0)$$

Afin de ne pas alourdir les écritures, nous conserverons la notation $\kappa\ r_0$ pour les deux modules. Alors :

$$f(\theta,\phi) = -\frac{m}{2\pi\,\hbar^2}\int_{-\infty}^{+\infty} e^{j\,\kappa\, r_0\,\cos(\theta_0)}\, V(r_0)\, r_0^2\, dr_0\, \sin(\theta_0)\, d\theta_0\, d\phi_0 \tag{276}$$

$$\Rightarrow \quad f(\theta,\phi) = -\frac{m}{2\pi\,\hbar^2}\left[\int_0^{2\pi} d\phi_0 \int_0^{\infty} V(r_0)\, r_0^2\, dr_0 \int_0^{\pi} e^{j\,\kappa\, r_0\,\cos(\theta_0)}\, \sin(\theta_0)\, d\theta_0\right] \tag{277}$$

$$\Rightarrow \quad f(\theta,\phi) = -\frac{m}{2\pi\,\hbar^2}\left[2\pi\int_0^{\infty} V(r_0)\, r_0^2\, dr_0 \left[-\frac{e^{j\,\kappa\, r_0\,\cos(\theta_0)}}{j\ \kappa\, r_0}\right]_0^{\pi}\right] \tag{278}$$

$$\Rightarrow \quad f(\theta,\phi) = -\frac{m}{\hbar^2}\left[\int_0^{\infty} V(r_0)\, r_0^2\, dr_0 \left[\frac{2\,\sin(\kappa\, r_0)}{\kappa\, r_0}\right]\right] \tag{279}$$

$$\Rightarrow \quad f(\theta) = -\frac{2\,m}{\kappa\,\hbar^2}\int_0^{\infty} r_0\, V(r_0)\, \sin(\kappa\, r_0)\, dr_0 \tag{280}$$

Il suffit maintenant de calculer l'intégrale restante sur la variable radiale r_0. Si nous utilisons un potentiel de Coulomb simple où $V(r) \propto 1/r$, alors nous nous heurtons à un problème car l'intégrale ne converge pas comme $r \to \infty$. Pour cette raison, un autre potentiel radialement symétrique est introduit, qui est donné par :

$$V(r) = \frac{e^{-r/l}}{r} \tag{281}$$

Où $l > 0$ est une constante décrivant la portée de l'interaction. Ce type de potentiel est connu sous le nom de *potentiel de Coulomb écranté*[15]. Le paramètre l détermine la gamme sur laquelle le potentiel est influent. Il nous permet d'évaluer analytiquement l'amplitude de diffusion. Ce potentiel est susceptible de décrire l'interaction électromagnétique entre deux particules chargées s'il y a un écrantage de la force lorsque la distance entre les particules est plus grande que l'ordre de la distance caractéristique $\propto l$. Il est possible alors d'observer le comportement de $f(\theta,\phi)^2$ pour un potentiel de Coulomb en laissant l tendre vers zéro. L'amplitude de diffusion devient :

$$\Rightarrow \quad f(\theta) = -\frac{2\,m}{\kappa\,\hbar^2}\int_0^{\infty} e^{-r_0/l}\, \sin(\kappa\, r_0)\, dr_0, \qquad \kappa = 2\,k\,\sin(\theta/2) \tag{282}$$

$$\Rightarrow \quad f(\theta) = -\frac{2\,m}{\kappa\,\hbar^2}\left[\frac{\kappa}{l^2+\kappa^2}\right], \qquad \kappa = 2\,k\,\sin(\theta/2) \tag{283}$$

$$\Rightarrow \quad f(\theta) = -\frac{2\,m}{(2\,k\,\sin(\theta/2))\,\hbar^2}\left[\frac{2\,k\,\sin(\theta/2)}{l^2+(2\,k\,\sin(\theta/2))^2}\right] \tag{284}$$

$$\Rightarrow \quad f(\theta) = -\frac{2\,m}{\hbar^2}\left[\frac{1}{l^2+(2\,k\,\sin(\theta/2))^2}\right] \tag{285}$$

Ainsi, lorsque le paramètre l se rapproche de zéro, on obtient :

$$\Rightarrow \quad f(\theta) \approx -\frac{2\,m}{\hbar^2}\left[\frac{1}{(2\,k\,\sin(\theta/2))^2}\right] \tag{286}$$

15. Ce potentiel s'obtient en superposant une charge ponctuelle positive à l'origine et une charge négative délocalisée au voisinage de l'origine sur une distance de l'ordre de l.

Par conséquent, l'intensité de l'onde émergente (ou dispersée) est :

$$I = |f(\theta)|^2 \propto \frac{1}{\sin^4(\theta/2)} \tag{287}$$

Dans cette analyse, nous avons considéré le cas d'un potentiel d'interaction central $V(r)$ qui dépend seulement de la distance radiale par rapport au centre du potentiel. Il est intéressant de noter que l'intensité d'interaction $|f(\theta)|^2$ est indépendante des variables caractérisant l'onde incidente (intensité, densité, ... etc.) et le potentiel en question, elle dépend uniquement de l'angle de déviation.

IV. Annexe : Compléments sur les orbitales des atomes réels

Nous entendons par l'adjectif *réels* les atomes ayant plus d'un électrons, ou simplement les atomes poly-électroniques. En effet, bien que les orbitales atomiques des ions hydrogénoïdes soient des solutions exactes de l'équation Schrödinger donc répondant de manière rigoureuse au formalisme de la mécanique quantique, elle sont néanmoins peu efficace pour d'écrire les propriétés des atomes réels et par extension celles des molécules.

A. *Orbitales de type Slater*

A cet égard, *John Clarke Slater* a proposé de garder la même forme générale que celle des orbitales atomiques hydrogénoïdes mais toutefois avec un terme radial dépendant uniquement cette fois-ci du nombre quantique principal. Dans cette démarche, *John C. Slater* propose également une *constante d'écran* afin de prendre en compte l'effet de la distribution de charge des électrons du cœurs. L'expression mathématique des *orbitales atomiques de Slater* (ces orbitales sont regroupées sous l'acronyme STO, pour **S**later **T**ype **O**rbital) est donnée par :

$$\chi^{STO}_{n,l,m}(r,\theta,\phi) = R_n(r) \times Y^m_l(\theta,\phi) \tag{288}$$

Où la partie radiale est :

$$R_n(r) = \frac{(2\,\zeta)^{\frac{2n+1}{2}}}{(2\,n!)^{(1/2)}}\; r^{n-1}\; e^{-\zeta\, r} \tag{289}$$

Les mêmes harmoniques sphériques $Y^m_l(\theta,\phi)$ des atomes hydrogénoïdes sont utilisées pour décrire la partie angulaire. L'exposant ζ (qui se lit zêta en alphabet latin) est un *paramètre ajustables*, lié à la *charge effective* du noyau. La charge nucléaire étant partiellement "écrantée" par les électrons des couches internes. Cette charge effective est estimée par *les règles de Slater* que détaillerons ci-dessous. En outre, le paramètre ζ controle la largeur de l'orbitale, une valeur élevée de ζ produit une orbitale plus fine et petite valeur produira au contraire une orbitale plus large. A titre d'exemple pour représenter l'orbitale atomique $|1s\rangle$, nous utiliserons l'expression :

$$\chi^{STO}_{1,0,0}(r,\theta,\phi) = \left[\frac{\zeta^3}{\pi}\right]^{1/2} e^{-\zeta\, r} \times Y^0_0(\theta,\phi) \tag{290}$$

Par ailleurs, il est possible utiliser plus d'une STO pour représenter une orbite atomique, comme le montre l'équation :

$$\chi^{STO}_{2,0,0}(r,\theta,\phi) = \left[c_1\, r\, e^{-\zeta_1\, r} \;+\; c_2\, r\, e^{-\zeta_2\, r}\right] \times Y^m_l(\theta,\phi) \tag{291}$$

Dans cette équation, les paramètres ζ_1 et ζ_2 sont ajustés par une procédure de *fitting* (les moindres carrés par exemple). Les coefficients c_1 et c_2 sont déterminés par un calcul variationnel linéaire qui minimise l'énergie du système quantique étudié. Dans cette configuration, la fonction ayant la plus grande valeur de ζ tient compte de la charge près du noyau. Tandis que la fonction ayant la petite valeur de ζ tient compte de la distribution de la charge à des valeurs plus importantes de la distance du noyau. Cette base de fonctions d'onde est appelée ensemble, *base*[16] *à double zêta*. Nous exigeons que ces fonctions de base couvrent tout l'espace de distribution des électrons[17] , ce qui signifie qu'elle doivent former un ensemble complet et doivent décrire la même chose. Par exemple, les harmoniques sphériques ne peuvent pas être utilisées pour décrire une fonction radiale d'un

16. Une base de fonctions d'onde (basis set en anglais) en mécanique quantique est un ensemble de fonctions (appelées fonctions de base) qui sont linéairement combinées pour créer des orbitales moléculaires : $\psi_i = \sum_j c_{ij}\varphi_j$. Par commodité, ces fonctions sont généralement des orbitales atomiques centrées sur les atomes.

17. Cette propriété des fonctions dans l'espace est tout comme la propriété correspondante des vecteurs. Les vecteurs unitaires ($\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$) décrivent des points dans l'espace et forment un ensemble complet puisque toute position dans l'espace peut être spécifiée par une combinaison linéaire de ces trois vecteurs unitaires. Ces vecteurs unitaires sont également appelés vecteurs de base.

atome d'hydrogène car elles n'impliquent pas la variable radiale, mais elles peuvent être utilisées pour décrire les propriétés angulaires de n'importe quel système dans l'espace tridimensionnel.

Par ailleurs, nous pouvons démonter que l'énergie d'un électron occupant une orbitale de Slater à exactement la meme forme que celle obtenu en résolvant l'équation de Schrödinger. En utilisant le Laplacien en coordonnées sphériques, l'équation de Schrödinger devient :

$$-\frac{\hbar^2}{2\,m}\underbrace{\left[\frac{1}{r^2}\,\frac{\partial}{\partial r}\left(r^2\,\frac{\partial}{\partial r}\right)+\frac{1}{r^2\sin\theta}\,\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\,\frac{\partial^2}{\partial\phi^2}\right]}_{\nabla^2}\chi^{STO}+V(r)\,\chi^{STO}=E\,\chi^{STO} \tag{292}$$

En adoptant le système des unités atomiques nous avons : $m=\hbar=e=4\,\pi\,\epsilon_0=1$ et afin de simplifier encore cette expression, posons aussi :

$$\frac{1}{\sin\theta}\,\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\,\frac{\partial^2}{\partial\phi^2}=\Lambda \tag{293}$$

Nous obtenons alors :

$$-\frac{1}{2}\left[\frac{1}{r^2}\,\frac{\partial}{\partial r}\left(r^2\,\frac{\partial}{\partial r}\right)+\frac{\Lambda}{r^2}\right]\chi^{STO}+\underbrace{V(r)}_{-\frac{Z^*}{r}}\chi^{STO}=E\,\chi^{STO} \tag{294}$$

Les orbitales de Slater sont données par :

$$\chi^{STO}_{n,l,m}(r,\theta,\phi)=R_n(r)\times Y^m_l(\theta,\phi) \tag{295}$$

Où la partie radiale dépendant uniquement du nombre quantique n est :

$$R_n(r)=\frac{(2\,\zeta)^{\frac{2n+1}{2}}}{(2\,n!)^{(1/2)}}\;r^{n-1}\;e^{-\zeta\,r}\quad\text{avec}\quad\xi=\frac{Z^*}{n} \tag{296}$$

Il est bien connu que les harmoniques sphériques $Y^m_l(\theta,\phi)$ sont des fonctions propres de l'opérateur Λ :

$$\lambda\,Y^m_l(\theta,\phi)=-l(l+1)\,Y^m_l(\theta,\phi) \tag{297}$$

En substituant (295) dans (294) il vient :

$$-\frac{1}{2}\left[\frac{Y^m_l(\theta,\phi)}{r^2}\,\frac{\partial}{\partial r}\left(r^2\,\frac{\partial\,R_n(r)}{\partial r}\right)+\frac{R_n(r)\,\overbrace{\Lambda\times Y^m_l(\theta,\phi)}^{-l(l+1)\,Y^m_l(\theta,\phi)}}{r^2}\right]-\frac{Z^*}{r}\,R_n(r)\times Y^m_l(\theta,\phi)=E\,R_n(r)\times Y^m_l(\theta,\phi) \tag{298}$$

Calculons d'abord la dérivée :

$$\frac{\partial}{\partial r}\left(r^2\,\frac{\partial\,R_n(r)}{\partial r}\right)=\frac{(2\,\zeta)^{\frac{2n+1}{2}}}{(2\,n!)^{(1/2)}}\;e^{-\xi\,r}\;\left[(n-1)^2\,r^n-2\,n\,\xi\,r^n+\xi^2\,r^{n+1}\right] \tag{299}$$

$$\Rightarrow\quad\frac{\partial}{\partial r}\left(r^2\,\frac{\partial\,R_n(r)}{\partial r}\right)=\underbrace{\frac{(2\,\zeta)^{\frac{2n+1}{2}}}{(2\,n!)^{(1/2)}}\;r^{n-1}\,e^{-\xi\,r}}_{R_n(r)}\;\left[(n-1)^2\,r-2\,n\,\xi\,r+\xi^2\,r^2\right] \tag{300}$$

En substituant (300) dans (298) il vient :

$$-\frac{1}{2}\left[\frac{\chi^{STO}}{r^2}\left[(n-1)^2\,r-2\,n\,\xi\,r+\xi^2\,r^2\right]-\frac{l(l+1)}{r^2}\chi^{STO}\right]-\frac{Z^*}{r}\,\chi^{STO}=E\,\chi^{STO} \tag{301}$$

$$\Rightarrow\quad -\frac{1}{2}\left[\frac{-l(l+1)}{r^2}+\frac{(n-1)^2-2\,n\,\xi-2\,Z^*}{r}+\xi^2\right]\chi^{STO}=E\,\chi^{STO} \tag{302}$$

$$\Rightarrow\quad E_n=-\frac{1}{2}\left[\frac{-l(l+1)}{r^2}+\frac{(n-1)^2-2\,n\,\xi-2\,Z^*}{r}+\xi^2\right] \tag{303}$$

$$\Rightarrow\quad E_n=-\frac{1}{2}\left[\frac{-l(l+1)}{r^2}+\frac{(n-1)^2-2\,n\,\xi-2\,n\,\xi}{r}+\xi^2\right] \tag{304}$$

Pour une distance suffisamment loin du noyau, c'est-à-dire r grand, nous pouvons considérer les approximations :

$$\frac{-l(l+1)}{r^2}\longrightarrow 0\quad\text{et}\quad\frac{(n-1)^2-4\,n\,\xi}{r}\longrightarrow 0 \tag{305}$$

$$\Rightarrow\quad E_n\simeq-\frac{1}{2}\,\xi^2=-\frac{1}{2}\left[\frac{Z^*}{n}\right]^2 \tag{306}$$

Le traitement que nous venons d'effectuer nous informe que les orbitales de Slater décrivent "bien" le comportement de l'électron tant que celui-ci "se tient" loin du noyau. Pour les atomes hydrogénoïdes le paramètre Z exprime la charge du noyau. Pour des atomes non-hydrogénoïdes, ce paramètre est la charge effective Z^* ressentie par l'électron. Selon le modèle de Slater, la charge effective ressentie par l'électron é(i) est déterminée selon :

$$Z_i^*=Z-\sum_{j\leq i}N_j\,\sigma_j\quad\text{avec}\quad 1<j\leq i \tag{307}$$

Le paramètre N_j représente le nombre d'électrons des couches $n_j\leq n_i$. Dans ce modèle de Slater, l'interaction Coloumbienne entre un électron (i) et son noyau est supposée être perturbée ou "écrantée" par les électrons (j) des couches intermédiaires. Cet effet "d'écrantage" de l'intération Coloumbienne électron-noyau est interprétée en terme d'une constante dite *constante d'écran* portant le symbole σ_j. Cette constante dépend de l'emplacement des électrons j ($N-1$ électrons restant) situés entre l'électron i et le noyau. Par conséquent l'électron i ressentira une charge réduite $+Z^*\,|e|$ au lieu de la charge réelle du noyau $+Z\,|e|$.

B. ***Orbitales Gaussienne***

Lorsque des calculs quantiques sur des moléculaires sont menés, il est courant d'utiliser une base composée d'un nombre fini d'orbitales atomiques, centrées sur chaque noyau atomique de la molécule. Ces orbitales atomiques sont bien décrites avec les orbitales de type Slater, car les STO décroissent exponentiellement avec la distance par rapport au noyau, décrivant ainsi précisément le chevauchement à longue distance entre les atomes. Néanmoins, il a été démontré [18] que les intégrales impliquant des fonctions gaussiennes sont plus rapides à calculer que celles impliquant des exponentielles de type STO, ce qui se traduit par un gain en terme de temps de calcul. D'une façon générale, les orbitales gaussiennes s'écriront sous la forme :

$$\chi^{GTO}_{n,l,m}(r,\theta,\phi)=N_{n,l}(\zeta)\;r^l\,r^{2(n-l-1)}\,e^{-\zeta\,r^2}\times Y_l^m(\theta,\phi) \tag{308}$$

Où $N_{n,l}(\zeta)$ est une constante de normalisation. A titre d'exemple pour représenter l'orbitale atomique $|1s\rangle$, nous utiliserons l'expression :

$$\chi^{GTO}_{1,0,0}(r,\theta,\phi)=\left[\frac{2\,\zeta}{\pi}\right]^{3/2}\;e^{-\zeta\,r^2}\times Y_0^0(\theta,\phi) \tag{309}$$

18. Dans les années 1950, par Frank Boys de l'université de Cambridge au Royaume-Uni.

Notons que pour toutes les foncions de base, seule la partie radiale de l'orbitale change. Les fonctions harmoniques sphériques sont utilisées décrire la partie angulaire de l'orbitale. Malheureusement, les fonctions gaussiennes ne reflètent pas précisément la forme d'une orbitale atomique. En particulier, elles sont plutôt plates que raides près du noyau atomique $r \to 0$. En outre, ces orbitales gaussiennes diminuent plus rapidement pour les valeurs élevées $r \to \infty$. Afin de remédier à ces limitations, chaque orbitale STO est remplacée par un certain nombre de fonctions gaussiennes avec des valeurs différentes pour le paramètre ζ. Ces fonctions gaussiennes forment un ensemble de bases gaussiennes primitives. Des combinaisons linéaires des Gaussiennes sont formées pour se rapprocher de la partie radiale d'une orbitale STO. Ces nouvelles fonctions sont définies alors comme des*fonctions gaussiennes contractées (CGTO)* :

$$\chi^{CGTO}(r,\theta,\phi) = \sum_{j}^{N} c_{\zeta_j} r^l \, r^{2(n-l-1)} \, e^{-\zeta_j r^2} \times Y_l^m(\theta,\phi) \tag{310}$$

Où N est la taille de la contraction et c_{ζ_j} sont les coefficients de la contraction. Les fonctions gaussiennes contractées les plus simples sont les bases STO-nG. Cette base tente d'approcher ou de mimer les orbitales STO par la somme de N Gaussiennes.

Chapitre[4] Méthodes de calcul quantique

Samir Kenouche - Département des Sciences de la Matière - UMKB

Méthodes Mathématiques et Algorithmes pour la Physique

Résumé

L'objectif de ce chapitre est d'introduire les méthodes de calcul quantique usuelles afin de quantifier efficacement et précisément l'énergie des systèmes quantiques (atomes, solide, ... etc), car la résolution exacte de l'équation de Schrödinger électronique est impossible. Dans ce chapitre, nous détaillons au prix de quelques répétitions utiles pédagogiquement, le formalisme mathématique de deux méthodes quantiques les plus emblématiques du formalisme quantique, à savoir la théorie de *Hartree-Fock* et la *théorie de la densité de la fonctionnelle*. Ce sont des méthodes de résolution approchée de l'équation de Schrödinger stationnaire à plusieurs corps utilisant le principe variationnel pour approximer la fonction d'onde et l'énergie du niveau fondamental. Toutefois, ce chapitre débute par un bref rappel mathématique inhérent à la théorie quantique.

Table des matières

S. Kenouche est docteur en Physique de l'Université de Montpellier et docteur en Chimie de l'Université de Béjaia.
Site web : voir http://www.sites.univ-biskra.dz/kenouche
Version corrigée, améliorée et actualisée le 10.10.2020.

I. Introduction

LA théorie de *Hartree-Fock* a été développée pour résoudre l'équation de Schrödinger stationnaire. Cette théorie est fondamentale pour une grande partie de la théorie des structures électroniques. Elle constitue l'ossature de la *théorie des orbitales moléculaires*, qui postule que chaque électron peut être décrit comme une fonction à une seule particule (orbitale) qui ne dépend pas explicitement des mouvements instantanés des autres électrons. Pour des systèmes de grande dimension, on utilise la *DFT* (Density Funcional Theory, en anglais) ou la *Théorie de la Densité de la Fonctionnelle*. Alors que la fonction d'onde $\psi(r_1, r_2, \cdots, r_N)$ n'a de sens que par son carré exprimant une mesure directe de la densité électronique $\rho(r)$. L'intérêt de travailler avec la densité électronique, tient au fait que c'est une observable et nous pouvons donc la mesurer expérimentalement, par exemple, par la diffraction des rayons X donnant des cartographies de densité électronique. L'autre intérêt de la *DFT* est que $\rho(r)$ est une fonction de trois variables (x, y, z) calculable en tout point de l'espace. Alors que la fonction d'onde pour un système poly-électroniques dépend des coordonnées de tout les électrons, pour un système à N électrons il en résulte $3\,N$ variables. De ce point de vue, la *DFT* permet une très grande simplification. Le but ultime est une description mathématique de la distribution des électrons dans les systèmes quantiques permettant aux expérimentateurs (chimistes et physiciens) de développer une compréhension approfondie de la liaison et de la réactivité chimiques, de calculer les propriétés physico-chimiques et de faire des prédictions basées sur ces calculs. Par exemple, un domaine de recherche actif dans l'industrie pharmaceutique consiste à calculer les modifications des propriétés chimiques des médicaments à la suite de modifications de la structure chimique. Le choix du modèle de calcul théorique pour un système chimique implique presque toujours un compromis entre la précision et le coût calculatoire. Des méthodes plus précises et des bases plus larges permettent de prolonger la durée des calculs.

Avant d'aborder en détail ces méthodes de calcul quantique, il semble pertinent de rappeler d'abord quelques notions mathématiques fondamentales inhérentes à la physique quantique. Pour les étudiants (es) ayant déjà acquis ces connaissances élémentaires peuvent passer directement à la section (II).

A. *Espace vectoriel des fonctions d'ondes*

L'état spatial d'une particule est décrit par la fonction d'onde $\psi(r)$. Dans un espace à une dimension :

$$\begin{aligned} \psi \;:\; \mathbb{R} &\longmapsto \mathbb{C} \\ x &\longmapsto \psi(x) \end{aligned}$$

Ainsi $\psi(x)$ est une fonction à valeurs complexes. Formellement, l'ensemble des fonctions d'ondes forment un espace vectoriel normé sur le corps $\mathbb{C}$, c'est l'espace de Hilbert $(\mathcal{E}_{\mathcal{H}})$:

$$\begin{aligned} \mathcal{E}_{\mathcal{H}} &\longmapsto \mathcal{E}_{\mathcal{H}} \\ \psi_1, \psi_2 &\longmapsto \psi_1 + \psi_2 \end{aligned}$$

Et,

$$\begin{aligned} \mathbb{C} &\longmapsto \mathcal{E}_{\mathcal{H}} \\ \forall\, \lambda &\longmapsto \lambda\, \psi \end{aligned}$$

Important ! les détails mathématiques sur les espaces vectoriels sur les corps $\mathbb{R}$ et $\mathbb{C}$ sont donnés en annexe (VI). Par ailleurs, selon la *notation de Dirac*, la fonction d'onde $\psi(r)$ est symbolisée par $|\psi\rangle$, appelé *ket*. Nous écrirons :

$$|\psi\rangle = |\psi_1\rangle + |\psi_2\rangle$$

$$|\varphi\rangle = \lambda\,|\psi_1\rangle$$

Dans cette notation, la fonction d'onde devient un point (l'extrémité du vecteur $|\psi\rangle$) de l'espace vectoriel $\mathcal{E}_{\mathcal{H}}$. Le produit scalaire Hermitien de deux fonctions d'ondes $|\psi_1\rangle$ et $|\psi_2\rangle$ est le nombre complexe $\langle\psi_1|\psi_2\rangle$ défini par :

$$\langle\psi_1|\psi_2\rangle = \int_{\mathbb{R}} \psi_1^*\,\psi_2\,dx \tag{1}$$

Avec ψ_1^* est le nombre complexe conjugué de ψ_1. Si $\langle\psi_1|\psi_2\rangle = 0$ alors $|\psi_1\rangle$ et $|\psi_2\rangle$ sont orthogonales. En voici un exemple, soient les fonctions d'ondes $\psi_1(x) = e^{j\beta_1 x}$ et $\psi_2(x) = e^{j\beta_2 x}$ avec $\beta_1 = 2\,\pi/T$ et $\beta_2 = 8\,\pi/T$:

$$\langle\psi_1|\psi_2\rangle = \int_0^T e^{-j\beta_1 x}\,e^{j\beta_2 x}dx = \int_0^T e^{j\,(\beta_2-\beta_1)\,x} = \underbrace{\int_0^T \cos((\beta_2-\beta_1)\,x)}_{=0} + j\,\underbrace{\int_0^T \sin((\beta_2-\beta_1)\,x)}_{=0} = 0 \tag{2}$$

L'intégration d'une fonction sinusoïdale sur sa période T donne systématiquement un résultat nul. Les surfaces sous la courbe des parties négative et positive sont égales en valeur absolue mais elles ont un signe différent. Les fonctions d'ondes $\psi_1(x)$ et $\psi_2(x)$ sont donc orthogonales. Par ailleurs, la norme au carré est définie par :

$$||\psi||^2 = \langle\psi|\psi\rangle = \int_{\mathbb{R}} |\psi(x)|^2\,dx > 0 \tag{3}$$

Cette norme (ou distance) traduit la surface sous la courbe positive de $|\psi(x)|^2$. Dans le cas où $||\psi|| = 1$ ($\langle\psi|\psi\rangle = 1$) on dit que le vecteur $|\psi\rangle$ est normalisé ou de façon équivalente :

$$\int_{\mathbb{R}} |\psi(x)|^2\,dx = 1 \tag{4}$$

Cette normalisation est importante afin d'interpréter $|\psi(x)|^2$ comme une *densité de probabilité de présence*. Les fonctions d'ondes pour lesquelles l'intégrale (4) existe sont appelées des *fonctions de carré sommable* ou de carré intégrable et l'espace de Hilbert sera noté :

$$\mathcal{E}_{\mathcal{H}} = \mathbb{L}^2(\mathbb{R})$$

En voici quelques propriétés du produit scalaire défini plus haut, $\forall\lambda \in \mathbb{C}$:

- $\langle\psi_1|\psi_2\rangle = \langle\psi_2|\psi_1\rangle^*$
- $\langle\lambda\psi_1|\psi_2\rangle = \lambda^*\,\langle\psi_2|\psi_1\rangle$
- $\langle\psi_1+\psi_2|\psi_3\rangle = \langle\psi_1|\psi_3\rangle + \langle\psi_2|\psi_3\rangle$
- $\langle\psi|\lambda_1\,\psi_1 + \lambda_2\,\psi_2\rangle = \lambda_1\,\langle\psi|\psi_1\rangle + \lambda_2\,\langle\psi|\psi_2\rangle$

Notons que le *vecteur dual* $\langle\psi_1|$ est une application, au sens mathématique du terme, qui associe à un vecteur $|\psi_2\rangle$ un nombre complexe résultat de $\langle\psi_1|\psi_2\rangle$. Formellement nous écrivons :

$$\begin{aligned}\langle\psi_1| \ : \ \mathcal{E}_{\mathcal{H}} &\longmapsto \mathbb{C} \\ |\psi_2\rangle &\longmapsto \langle\psi_1|\psi_2\rangle\end{aligned}$$

Ainsi, $\langle\psi_1|$ est une forme linéaire de l'espace vectoriel $\mathcal{E}_{\mathcal{H}}$. Le vecteur dual $\langle\psi_1|$ est appelé *bra* dans la littérature. A partir d'un vecteur $|\psi\rangle \in \mathcal{E}_{\mathcal{H}}$ nous pouvons construire au moyen du produit scalaire, un vecteur dual $\langle\psi| \in \mathcal{E}_{\mathcal{H}}^*$ et inversement, nous avons donc un *isomorphisme*.

$$\forall\,(|\psi_1\rangle, |\psi_2\rangle) \in {\mathcal{E}_{\mathcal{H}}}^2, \quad \forall\,(\lambda_1, \lambda_2) \in \mathbb{C}^2$$

$$\text{Si} \quad |\psi\rangle = \lambda_1\,|\psi_1\rangle + \lambda_2\,|\psi_2\rangle \quad \Rightarrow \quad \langle\psi| = \lambda_1^*\,\langle\psi_1| + \lambda_2^*\,\langle\psi_2|$$

Nous avons : $\psi(x) = |\psi\rangle$ (ket psi) et $\langle\psi_1| = \psi(x)^*$ (bra psi). Dans l'espace vectoriel $\mathcal{E}_{\mathcal{H}}$, on représente les vecteurs d'état sous forme de vecteurs colonnes :

$$|\psi_1\rangle = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{pmatrix} \quad \text{et} \quad |\psi_2\rangle = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_n \end{pmatrix} \tag{5}$$

Les bra (vecteurs duals) associés sont des vecteurs lignes :

$$\langle\psi_1| = (u_1^*, u_2^*, u_3^*, \cdots, u_n^*) \quad \text{et} \quad \langle\psi_2| = (v_1^*, v_2^*, v_3^*, \cdots, v_n^*)$$

De sorte que le produit scalaire s'écrit :

$$\langle\psi_1|\psi_2\rangle = (u_1^*, u_2^*, u_3^*, \cdots, u_n^*) \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_n \end{pmatrix} = \sum_{i=1}^{n} u_i^*\, v_i = \langle\psi_2|\psi_1\rangle^*$$

Un système quantique peut se trouver dans une infinité d'états, en revanche chaque état est unique. L'ensemble des états possibles d'un système forment un espace des états (structure d'un espace vectoriel). Il est toujours possible de combiner ces états pour en former un état possible du système et inversement, il est possible de décomposer un état en combinaison linéaire des états possibles du système. Chaque combinaison est unique pour un état donné, il existe une infinité de combinaison. Ce qui nous amène à définir les bases algébriques :

Tout les espaces vectoriels peuvent êtres étudiés par des vecteurs numériques (les coordonnées). Une base $\{e_j\}_{j\in\mathcal{I}}$ *de* $\mathcal{E}$ *est une famille libre et génératrice si :*

libre : $\forall$ *un sous-ensemble fini* $\mathcal{F} \subset \mathcal{I}, \sum_{j\in\mathcal{F}} c_j\, e_j = 0_{\mathcal{E}}$ *alors* $\forall\, j \in \mathcal{F}, c_j = 0$. *Cela signifie que nous obtenons un vecteur nul* $0_{\mathcal{E}}$ *uniquement si les coefficients de la combinaison* c_j *sont nuls.*

génératrice : $\forall\, v \in \mathcal{E}, \exists\mathcal{H} \subset \mathcal{I}$ *telle que* $\exists!$ *des scalaires* $\{x_h\}_{h\in\mathcal{H}} \Rightarrow v = \sum_{h\in\mathcal{H}} x_h\, e_h$.

Avec les x_h *sont les coordonnées de* $v \in \mathcal{E}$ *dans la base* $\{e_h\}_{h\in\mathcal{H}}$. *Autrement dit, chaque élément de* $\mathcal{E}$ *peut s'écrire comme une combinaison linéaire* ***unique*** *(ou mathématiquement* $\exists!$*) de la base* $\{e_h\}_{h\in\mathcal{H}}$.

Théorème *: $\forall\, v \in \mathcal{E}$ est une combinaison* ***unique*** *des $\{e_h\}_{h\in\mathcal{H}}$ si et seulement si les $\{e_h\}_{h\in\mathcal{H}}$ forment une base dans $\mathcal{E}$. En d'autres mots, si $\{e_h\}_{h\in\mathcal{H}}$ forment une base dans $\mathcal{E}$, alors l'élément $v \in \mathcal{E}$ est identifié par ses coordonnées.*

Exemple d'application 1

Nous souhaitons calculer la matrice canonique, notée $\mathcal{M}_{BC}(f)$ d'une application linéaire f :

$$\begin{aligned} f:\ \mathbb{R}^2 &\longmapsto \mathbb{R}^3 \\ (x,y) &\longmapsto (x+y, 2\,x, y-x) \end{aligned}$$

$$f(x,y) = f(1,0) = (1,2,-1) = 1\,(1,0,0) + 2\,(0,1,0) - 1\,(0,0,1)$$

$$f(x,y) = f(0,1) = (1,0,1) = 1\,\underbrace{(1,0,0)}_{e_1} + 0\,\underbrace{(0,1,0)}_{e_2} + 1\,\underbrace{(0,0,1)}_{e_3}$$

$$\Rightarrow \mathcal{M}_{BC}(f) = \begin{pmatrix} 1 & 1 \\ 2 & 0 \\ -1 & 1 \end{pmatrix}$$

En effet, les vecteurs $\{e_i\}_{i=1,3}$ forment une base dans $\mathbb{R}^3$. Autrement dit, chaque élément de $\mathbb{R}^3$ pouvant s'écrire comme une combinaison linéaire unique des vecteurs de la base $\{e_i\}_{i=1,3}$. Soulignons que les matrices $\mathcal{M}_{BC}$ forment un espace vectoriel car elles vérifient ses propriétés.

Exemple d'application 2

Considérons le $\mathbb{R}^3$-espace vectoriel, $v_1, v_2 \in \mathcal{E}^2$ telle que $v_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$ et $v_2 = \begin{pmatrix} 6 \\ 4 \\ 2 \end{pmatrix}$

– *Montrer que $v = \begin{pmatrix} 9 \\ 2 \\ 7 \end{pmatrix}$ est une combinaison linéaire de v_1 et v_2.*

Cherchons $c_1, c_2 \in \mathbb{R}$ tel que :

$$\begin{pmatrix} 9 \\ 2 \\ 7 \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} + c_2 \begin{pmatrix} 6 \\ 4 \\ 2 \end{pmatrix} \tag{6}$$

$$\Rightarrow \left\{ \begin{array}{l} 9 = c_1 + 6\,c_2 \\ 2 = 2\,c_1 + 4\,c_2 \\ 7 = -c_1 + 2\,c_2 \end{array} \right. \Rightarrow \left\{ \begin{array}{l} c_1 = -3 \\ c_2 = 2 \end{array} \right.$$

B. *Rappels sur les bases orthonormées*

Une base $\{e_1, e_2, e_3, \cdots, e_n\}$ est orthonormée si et seulement si :

$$\begin{cases} e_j \cdot e_i = 1 & \text{si} \quad j = i \\ \\ e_j \cdot e_i = 0 & \text{si} \quad j \neq i \end{cases} \tag{7}$$

C'est le cas par exemple de la base canonique de $\mathbb{R}^n$:

$$\mathbb{R}^n = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \dots \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \tag{8}$$

L'intérêt de ces bases orthonormées c'est qu'on peut calculer facilement un produit scalaire. Prenons deux éléments v_1 et v_2 d'un espace vectoriel, décomposons ces éléments respectivement dans les bases $\{e_i\}_{i=1,n}$ et $\{e_j\}_{j=1,n}$ soit :

$$v_1.v_2 = \left[\sum_{i=1}^{n} x_i \cdot e_i\right] \cdot \left[\sum_{j=1}^{n} y_j \cdot e_j\right]$$

En utilisant la bilinéarité (voir les détails en annexe) du produit scalaire on obtient :

$$v_1.v_2 = \sum_{i=1}^{n}\sum_{j=1}^{n} [x_i \cdot y_j] \cdot \underbrace{[e_i \cdot e_j]}_{=0 \text{ si } i\neq j \text{ et } =1 \text{ sinon}}$$

$$\Rightarrow v_1.v_2 = \sum_{i=1}^{n} x_i \cdot y_i$$

Ainsi le produit scalaire se résume à calculer le produit des coordonnées des deux vecteurs. Un autre avantage majeur des bases orthonormées, est la possibilité de déterminer la ième coordonnée du vecteur v uniquement en connaissant les $\{e_i\}_{i=1,n}$ et $\{e_j\}_{j=1,n}$. Cela n'est pas faisable avec les bases qui ne sont pas orthogonales. En voici la démonstration :

$$v = \sum_{i=1}^{n} [x_i \cdot e_i]$$

$$v \cdot e_j = \left[\sum_{j=1}^{n} x_i \cdot e_i\right] \cdot e_j$$

$$\Rightarrow v \cdot e_j = \sum_{i=1}^{n} x_i \underbrace{[e_i \cdot e_j]}_{=0 \text{ si } i\neq j \text{ et } =1 \text{ sinon}}$$

Tout les termes de la somme s'annulent sauf pour $i = j$, cela donne :

$$v \cdot e_j = x_j$$

Cela signifie que pour connaitre la coordonnée x_j il suffit de calculer $v \cdot e_j$. Ce calcul est indépendant des autres vecteurs de la base. Pour les bases non-orthonormées, chaque coordonnée dépend de tous les vecteurs formant la base. Il est donc impossible d'étudier le vecteur dans une direction donnée, le vecteur ne peut se projeter sur un axe donné.

C. *Représentation matricielle d'un opérateur*

Décomposons la fonction propre $\psi(x)$ sur la base de n fonctions propres $\{|\varphi_i\rangle\}_{i=1,n}$ orthonormées telle que :

$$\begin{cases} \langle\psi_j|\psi_i\rangle = 1 & \text{si} \quad j = i \\ \langle\psi_j|\psi_i\rangle = 0 & \text{si} \quad j \neq i \end{cases} \tag{9}$$

Ainsi,

$$\psi(x) = \sum_{i=1}^{n} c_i\,|\varphi_i\rangle \;\Rightarrow \hat{\mathcal{O}}\psi(x) = \sum_{i=1}^{n} c_i\,\hat{\mathcal{O}}\,|\varphi_i\rangle \tag{10}$$

Pour chaque i, on aura :

$$\hat{\mathcal{O}}\,|\varphi_i\rangle = a_{i1}\,|\varphi_1\rangle + a_{i2}\,|\varphi_2\rangle + \cdots + a_{ij}\,|\varphi_j\rangle + \cdots + a_{im}\,|\varphi_m\rangle \tag{11}$$

Multiplions (11) par le bra $\langle\varphi_j|$ et intégrons :

$$\left\langle\varphi_j|\hat{\mathcal{O}}\,|\varphi_i\right\rangle = \langle\varphi_j|a_{i1}\,|\varphi_1\rangle + \langle\varphi_j|a_{i2}\,|\varphi_2\rangle + \cdots + \langle\varphi_j|a_{ij}\,|\varphi_j\rangle + \cdots + \langle\varphi_j|a_{im}\,|\varphi_m\rangle \tag{12}$$

$$\left\langle\varphi_j|\hat{\mathcal{O}}\,|\varphi_i\right\rangle = a_{i1}\,\langle\varphi_j|\varphi_1\rangle + a_{i2}\,\langle\varphi_j|\varphi_2\rangle + \cdots + a_{ij}\,\langle\varphi_j|\varphi_j\rangle + \cdots + a_{im}\,\langle\varphi_j|\varphi_m\rangle \tag{13}$$

Tenant compte de (9), il vient :

$$\left\langle\varphi_j|\hat{\mathcal{O}}\,|\varphi_i\right\rangle = a_{ij}\,\langle\varphi_j|\varphi_j\rangle \Rightarrow a_{ij} = \frac{\left\langle\varphi_j|\hat{\mathcal{O}}\,|\varphi_i\right\rangle}{\underbrace{\langle\varphi_j|\varphi_j\rangle}_{=1}}$$

$$\Rightarrow a_{ij} = \left\langle\varphi_j|\hat{\mathcal{O}}\,|\varphi_i\right\rangle \tag{14}$$

Si $\hat{\mathcal{O}}\,\varphi_i = h_i\,\varphi_i$, pour les éléments diagonaux $(i = j) \Rightarrow a_{ij} = h_i\,\underbrace{\langle\varphi_j|\varphi_j\rangle}_{=1}$. Les éléments de la diagonale sont les valeurs propres de l'opérateur $\hat{\mathcal{O}}$. D'un autre côté, pour les éléments hors de la diagonale $(i \neq j)$, nous avons $a_{ij} = h_i\,\underbrace{\langle\varphi_j|\varphi_j\rangle}_{=0}$. Par voie de conséquence la matrice de terme a_{ij} est diagonale. Un exercice d'application sur la représentation matricielle d'un opérateur sera résolu pendant les séances de travaux dirigés.

II. Méthodes d'approximation

Il existe très peu de problèmes pour lesquels l'équation de Schrödinger peut être résolue avec exactitude. Cette équation ne peut être résolue exactement pour un atome ou une molécule plus complexe qu'un atome d'hydrogène, dit système à deux corps. Par exemple, il n'existe pas de solution analytique à l'équation de Schrödinger décrivant l'atome d'hélium ayant deux électrons seulement, dit système à trois corps. Pour cet atome, l'équation de Schrödinger est extrêmement compliquée du point de vue mathématique. Par conséquent, des méthodes d'approximation sont nécessaires pour résoudre ce type de problèmes, n'ayant pas de solutions explicites ou analytiques. Les deux méthodes d'approximation les plus couramment utilisées, sont les méthodes des variations et des perturbations. Ces deux méthodes d'approximation, sont des techniques de résolution très puissantes.

Selon la méthode variationnelle, des fonctions d'essai (*trial wavefunctions* en anglais) sont postulées afin d'estimer notamment l'énergie de l'état fondamental du système quantique étudié (atome, molécule, agrégat moléculaire ... etc). La fonction d'essai aura un ou plusieurs paramètres ajustables, qui seront utilisés pour l'optimisation. Cette théorie stipule que l'énergie calculée à partir d'une fonction d'essai est systématiquement supérieure ou égale à la véritable énergie fondamentale du système. En effet, l'égalité se produit uniquement lorsque la fonction d'essai est la véritable fonction d'onde de l'état fondamental. D'un autre côté, l'idée derrière la méthode des perturbations est que le système en question est perturbé ou légèrement modifié par rapport à un système de référence où la solution (fonctions et valeurs propres) est connue. Selon cette méthode, l'Hamiltonien du système quantique est décomposé en deux contributions. Pour la première contribution, la solution est connue, à partir par exemple de problèmes précédemment résolus. La deuxième contribution exprime la perturbation du système quantique par rapport au problème connu. Les fonctions d'onde du terme non perturbatif pour lequel la solution est connue, sont utilisées comme point de départ, puis modifiées pour approximer la vraie fonction d'onde de l'équation de Schrödinger. Dans ce qui suit, nous présenterons le cadre théorique et le formalisme mathématique de la méthode des variations ainsi que celle des perturbations. Des exercices d'application sont donnés à la fin de chaque section.

*A. **Méthode variationnelle***

La valeur attendue de l'Hamiltonien, calculée par le biais d'une fonction d'essai ψ_α est toujours supérieure ou égale à l'énergie fondamentale. Mathématiquement cela se traduit par l'écriture :

$$< \psi_\alpha|\hat{\mathcal{H}}|\psi_\alpha > \geq E_0 \quad \text{avec} \quad \hat{\mathcal{H}}\psi_\alpha = E_0\psi_\alpha \tag{15}$$

*a) **Preuve**:* développons la fonction d'essai ψ_α sur la base des fonctions propres $\{\varphi_j\}_{j=1,n}$ de l'opérateur Hamiltonien.

$$\hat{\mathcal{H}}\psi_\alpha = E_\alpha\psi_\alpha \quad \text{avec} \quad \psi_\alpha = \sum_j c_j\,\varphi_j \tag{16}$$

$$\langle a\rangle = \frac{\langle\psi_\alpha|\hat{\mathcal{H}}|\psi_\alpha\rangle}{\langle\psi_\alpha|\psi_\alpha\rangle} = \frac{\left\langle \sum_{j=1}^{n} c_j\,\varphi_j|\hat{\mathcal{H}}|\sum_{j=1}^{n} c_j\,\varphi_j\right\rangle}{\left\langle \sum_{j=1}^{n} c_j\,\varphi_j|\sum_{j=1}^{n} c_j\,\varphi_j\right\rangle}$$

φ_j étant une fonction propre de $\hat{\mathcal{H}}$, alors $\hat{\mathcal{H}}\varphi_j = E_j\varphi_j$ cela implique également :

$$E_\alpha = \frac{\sum_{i=1}\sum_{j=1} c_i^*\,c_j\,E_j\,\langle\varphi_i|\varphi_j\rangle}{\sum_{i=1}\sum_{j=1} c_i^*\,c_j\,\langle\varphi_i|\varphi_j\rangle}$$

Posons

$$\delta_{ij} = \langle\varphi_i|\varphi_j\rangle = \begin{cases} 1 & \text{si} \quad i = j \\ \\ 0 & \text{si} \quad i \neq j \end{cases} \tag{17}$$

Avec δ_{ij} est le symbôle de Kronecker. Il en découle :

$$E_\alpha = \frac{\sum\limits_{i=1}\sum\limits_{j=1} c_i^* \, c_j \, E_j \, \delta_{ij}}{\sum\limits_{i=1}\sum\limits_{j=1} c_i^* \, c_j \, \delta_{ij}}$$

L'opérateur $\hat{\mathcal{H}}$ étant hermitien donc ses fonctions propres sont orthogonales. Les termes de la somme sont non nuls uniquement si $i = j \Rightarrow \delta_{ij} = 1$. Par conséquent :

$$E_\alpha = \frac{\sum\limits_{j=1} c_j^* \, c_j \, E_j}{\sum\limits_{j=1} c_j^* \, c_j} = \frac{\sum\limits_{j=1} |c_j|^2 \, E_j}{\underbrace{\sum\limits_{j=1} |c_j|^2}_{=1}}$$

Rappelons que $P_j = |c_j|^2$ étant la probabilité d'obtenir l'état quantique j. Ainsi la somme des probabilités de tout les états quantiques possibles vaut 1. Il en résulte :

$$E_\alpha = \sum_{j=1} |c_j|^2 \, E_j \tag{18}$$

Retranchons, pour les termes de cette dernière équation, l'énergie de l'état fondamental E_0. Cela donne :

$$E_\alpha - E_0 = \left[\sum_{j=1} |c_j|^2 \, E_j\right] - E_0$$

Le deuxième terme de cette équation est systématiquement positif ou égale à zéro. Cela implique :

$$E_\alpha - E_0 \geq 0 \Rightarrow E_\alpha \geq E_0 \tag{19}$$

Exercice ❶ ☞ Ⓡ

Estimer l'énergie variationnelle de l'état fondamental de l'atome d'hydrogène en utilisant la fonction d'essai gaussienne suivante :

$$\psi_\alpha = e^{-\alpha \, r^2} \tag{20}$$

Le calcul doit être mené avec le système des unités atomiques et en se servant de la partie radiale du Laplacien exprimé en coordonnées sphériques. On donne :

$$\Delta_r = \frac{1}{r^2}\frac{\partial}{\partial r}\left[r^2 \frac{\partial}{\partial r}\right] \tag{21}$$

$$\int_{-\infty}^{+\infty} x^n \, e^{-\beta x^2} \, dx = \begin{cases} \sqrt{\dfrac{\pi}{\beta}} & \text{si} \quad n = 0 \\ \dfrac{n-1}{\beta^{(n-1)/n}} & \text{si} \quad n \, \text{est pair} \end{cases} \tag{22}$$

et,

$$\int_0^{+\infty} x \, e^{-\beta x^2} \, dx = \frac{1}{2\,\beta} \tag{23}$$

Unité atomique,

$$1\,Ry = \frac{1}{4\pi\epsilon_0}\frac{m_e|e|^4}{2\,\hbar^2} = 2\,Har \longrightarrow \text{unité de l'énergie}$$

$$\hbar = m_e = |e| = 1 \Rightarrow \frac{|e|^2}{4\pi\epsilon_0} = 1 \Rightarrow \frac{\hbar^2}{2\,m_e} = 1/2$$

*B. **Méthode variationnelle linéaire***

Cette méthode est plus générale, elle porte également le nom de "méthode de Ritz". Elle est basée sur un choix particulier de fonctions d'essai $\tilde{\psi}$, construites à partir d'un développement linéaire des fonctions propres $\{\varphi_j\}_{j=1,n}$ soit :

$$\tilde{\psi} = \sum_j c_j\,\varphi_j \tag{24}$$

Les coefficients de la combinaison linéaire c_j étant inconnus. Les coefficients variationnels $\{c_1, c_2, \cdots, c_n\}$ sont ceux minimisant la valeur moyenne de l'Hamiltonien.

*b) **Première approche**:* Afin d'illustrer cette procédure, nous considérons le cas où :

$$\tilde{\psi} = c_1\,\varphi_1 + c_2\,\varphi_2 \tag{25}$$

Ensuite, nous effectuerons une généralisation pour une combinaison à n fonctions propres. Nous avons :

$$\tilde{E} = \frac{\left\langle\tilde{\psi}|\hat{\mathcal{H}}|\tilde{\psi}\right\rangle}{\left\langle\tilde{\psi}|\tilde{\psi}\right\rangle}$$

$$\tilde{E} = \frac{\left\langle c_1\,\varphi_1 + c_2\,\varphi_2|\hat{\mathcal{H}}|c_1\,\varphi_1 + c_2\,\varphi_2\right\rangle}{\langle c_1\,\varphi_1 + c_2\,\varphi_2|c_1\,\varphi_1 + c_2\,\varphi_2\rangle} = \frac{A(c_1,c_2)}{B(c_1,c_2)}$$

$$A(c_1,c_2) = \left\langle c_1\,\varphi_1|\hat{\mathcal{H}}|c_1\,\varphi_1\right\rangle + \left\langle c_1\,\varphi_1|\hat{\mathcal{H}}|c_2\,\varphi_2\right\rangle + \left\langle c_2\,\varphi_2|\hat{\mathcal{H}}|c_1\,\varphi_1\right\rangle + \left\langle c_2\,\varphi_2|\hat{\mathcal{H}}|c_2\,\varphi_2\right\rangle$$

$$= c_1^2\left\langle\varphi_1|\hat{\mathcal{H}}|\varphi_1\right\rangle + c_1\,c_2\left\langle\varphi_1|\hat{\mathcal{H}}|\varphi_2\right\rangle + c_2\,c_1\left\langle\varphi_2|\hat{\mathcal{H}}|\varphi_1\right\rangle + c_2^2\left\langle\varphi_2|\hat{\mathcal{H}}|\varphi_2\right\rangle$$

$$\Rightarrow A(c_1,c_2) = c_1^2\,\hat{\mathcal{H}}_{11} + 2\,c_1\,c_2\,\hat{\mathcal{H}}_{12} + c_2^2\,\hat{\mathcal{H}}_{22}$$

L'opérateur $\hat{\mathcal{H}}$ est hermétique alors $\hat{\mathcal{H}}_{12} = \hat{\mathcal{H}}_{21}$. Calculons désormais le dénominateur $B(c_1,c_2)$, de façon analogue nous obtenons :

$$B(c_1,c_2) = c_1^2\,\langle\varphi_1|\varphi_1\rangle + c_1\,c_2\,\langle\varphi_1|\varphi_2\rangle + c_2\,c_1\,\langle\varphi_2|\varphi_1\rangle + c_2^2\,\langle\varphi_2|\varphi_2\rangle$$

$$B(c_1,c_2) = c_1^2\,\hat{\mathcal{O}}_{11} + 2\,c_1\,c_2\,\hat{\mathcal{O}}_{12} + c_2^2\,\hat{\mathcal{O}}_{22}$$

Il en découle,

$$\tilde{E} = \frac{c_1^2\,\hat{\mathcal{H}}_{11} + 2\,c_1\,c_2\,\hat{\mathcal{H}}_{12} + c_2^2\,\hat{\mathcal{H}}_{22}}{c_1^2\,\hat{\mathcal{O}}_{11} + 2\,c_1\,c_2\,\hat{\mathcal{O}}_{12} + c_2^2\,\hat{\mathcal{O}}_{22}} \tag{26}$$

Qu'on écrira sous la forme :

$$\tilde{E}\left[c_1^2\,\hat{\mathcal{O}}_{11} + 2\,c_1\,c_2\,\hat{\mathcal{O}}_{12} + c_2^2\,\hat{\mathcal{O}}_{22}\right] = \left[c_1^2\,\hat{\mathcal{H}}_{11} + 2\,c_1\,c_2\,\hat{\mathcal{H}}_{12} + c_2^2\,\hat{\mathcal{H}}_{22}\right] \tag{27}$$

Minimisons cette dernière expression, soit $\dfrac{\partial \tilde{E}}{\partial c_j}(c_j = c_j^{(0)}) = 0$. Nous commencerons par dériver (27) par rapport au premier coefficient c_1 :

$$\underbrace{\frac{\partial \tilde{E}}{\partial c_1}}_{=0}\left[c_1^2\,\hat{\mathcal{O}}_{11} + 2\,c_1\,c_2\,\hat{\mathcal{O}}_{12} + c_2^2\,\hat{\mathcal{O}}_{22}\right] + \tilde{E}\,\left[2\,c_1\,\hat{\mathcal{O}}_{11} + 2\,c_2\,\hat{\mathcal{O}}_{12}\right] = \left[2\,c_1\,\hat{\mathcal{H}}_{11} + 2\,c_2\,\hat{\mathcal{H}}_{12}\right]$$

$$\Rightarrow \tilde{E}\,\left[2\,c_1\,\hat{\mathcal{O}}_{11} + 2\,c_2\,\hat{\mathcal{O}}_{12}\right] = \left[2\,c_1\,\hat{\mathcal{H}}_{11} + 2\,c_2\,\hat{\mathcal{H}}_{12}\right]$$

$$\Rightarrow 2\,c_1\,\hat{\mathcal{H}}_{11} + 2\,c_2\,\hat{\mathcal{H}}_{12} - \tilde{E}\,\left[2\,c_1\,\hat{\mathcal{O}}_{11} + 2\,c_2\,\hat{\mathcal{O}}_{12}\right] = 0$$

$$\Rightarrow c_1\,\left[\hat{\mathcal{H}}_{11} - \tilde{E}\,\hat{\mathcal{O}}_{11}\right] + c_2\,\left[\hat{\mathcal{H}}_{12} - \tilde{E}\,\hat{\mathcal{O}}_{12}\right] = 0 \tag{28}$$

De façon analogue avec le deuxième coefficient c_2, on obtient :

$$c_1\,\left[\hat{\mathcal{H}}_{21} - \tilde{E}\,\hat{\mathcal{O}}_{21}\right] + c_2\,\left[\hat{\mathcal{H}}_{22} - \tilde{E}\,\hat{\mathcal{O}}_{22}\right] = 0 \tag{29}$$

A partir des équations algébriques (28) et (29), nous obtenons le système matriciel suivant :

$$\begin{pmatrix} \hat{\mathcal{H}}_{11} - \tilde{E}\,\hat{\mathcal{O}}_{11} & \hat{\mathcal{H}}_{12} - \tilde{E}\,\hat{\mathcal{O}}_{12} \\ \hat{\mathcal{H}}_{21} - \tilde{E}\,\hat{\mathcal{O}}_{21} & \hat{\mathcal{H}}_{22} - \tilde{E}\,\hat{\mathcal{O}}_{22} \end{pmatrix} \times \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{30}$$

La solution triviale est donnée par $c_1 = 0$ et $c_2 = 0$. Cette solution n'est pas intéressante physiquement. Ce système d'équations admet une solution, autre que la solution triviale, si et seulement si le déterminant :

$$\begin{vmatrix} \hat{\mathcal{H}}_{11} - \tilde{E}\,\hat{\mathcal{O}}_{11} & \hat{\mathcal{H}}_{12} - \tilde{E}\,\hat{\mathcal{O}}_{12} \\ \hat{\mathcal{H}}_{21} - \tilde{E}\,\hat{\mathcal{O}}_{21} & \hat{\mathcal{H}}_{22} - \tilde{E}\,\hat{\mathcal{O}}_{22} \end{vmatrix} = 0 \tag{31}$$

Le calcul de ce déterminant donne :

$$\left(\hat{\mathcal{H}}_{11} - \tilde{E}\,\hat{\mathcal{O}}_{11}\right) \times \left(\hat{\mathcal{H}}_{22} - \tilde{E}\,\hat{\mathcal{O}}_{22}\right) - \left(\hat{\mathcal{H}}_{12} - \tilde{E}\,\hat{\mathcal{O}}_{12}\right) \times \left(\hat{\mathcal{H}}_{21} - \tilde{E}\,\hat{\mathcal{O}}_{21}\right) = 0 \tag{32}$$

La base $\{\varphi_1, \varphi_2\}$ étant orthonormée, ainsi $\hat{\mathcal{O}}_{12} = \hat{\mathcal{O}}_{21} = 0$ et $\hat{\mathcal{O}}_{11} = \hat{\mathcal{O}}_{22} = 1$. De plus l'opérateur Hamiltonien est hermétique $\Rightarrow \hat{\mathcal{H}}_{12} = \hat{\mathcal{H}}_{21}$ et après réarrangement nous obtenons :

$$\tilde{E}^2 + (\hat{\mathcal{H}}_{11} - \hat{\mathcal{H}}_{22})\tilde{E} + \hat{\mathcal{H}}_{11}\,\hat{\mathcal{H}}_{22} - \hat{\mathcal{H}}_{12}^2 = 0 \tag{33}$$

Il en résulte un polynôme de second ordre en $\tilde{E}$. La résolution de cette équation donnera deux racines $\tilde{E}_1$ et $\tilde{E}_2$ conduisant à la détermination des coefficients de la combinaison c_1 et c_2 permettant la minimisation de l'énergie du système quantique. La généralisation du déterminant ci-dessus est immédiate :

$$|\hat{\mathcal{H}}_{ij} - \tilde{E}\,\hat{\mathcal{O}}_{ij}| = 0 \tag{34}$$

C'est le déterminant *séculaire*. C'est un système d'équations de degré n en $\tilde{E}$ menant à n valeurs propres de l'énergie. Soulignons finalement que l'accroissement du nombre de paramètres ajustables améliore le résultat, mais risque d'accroitre également la complexité du problème.

c) ***Deuxième approche****:* dans cette approche, la fonction d'essai est toujours une combinaison linéaire de fonctions connues (par exemple les orbitales atomiques). Ces fonctions sont normalisées.

$$\tilde{\psi} = \sum_{j=1}^{n} c_j \, \varphi_j$$

et,

$$\delta_{ij} = \langle \varphi_i | \varphi_j \rangle = \begin{cases} 1 & \text{si} \quad i = j \\ \\ 0 & \text{si} \quad i \neq j \end{cases} \tag{35}$$

La condition de normalisation $\langle \varphi_j | \varphi_j \rangle = 1$ impose que la modification d'un coefficient engendre simultanément une modification des $n-1$ autres coefficients de la combinaison afin de maintenir la norme constante. En effet, pour $n = 2$ il vient :

$$|\tilde{\psi}|^2 = \left\langle \tilde{\psi} | \tilde{\psi} \right\rangle = \text{cst} \tag{36}$$

$$\Rightarrow |\tilde{\psi}|^2 = \langle c_1 \, \varphi_1 + c_2 \, \varphi_2 | c_1 \, \varphi_1 + c_2 \, \varphi_2 \rangle = c_1^2 \, \langle \varphi_1 | \varphi_1 \rangle + c_1 \, c_2 \, \langle \varphi_1 | \varphi_2 \rangle + c_2 \, c_1 \, \langle \varphi_2 | \varphi_1 \rangle + c_2^2 \, \langle \varphi_1 | \varphi_1 \rangle$$

$$\Rightarrow |\tilde{\psi}|^2 = c_1^2 + c_2^2 + 2 \, c_1 \, c_2 \, \delta_{12} = \text{cst}$$

Toute variation de c_1 entraine mécaniquement celle de c_2 et inversement. Le Lagrangien du système, sous la contrainte de normalisation s'écrit :

$$\mathcal{L} = \left\langle \tilde{\psi} | \hat{\mathcal{H}} | \tilde{\psi} \right\rangle - \lambda \underbrace{\left[\left\langle \tilde{\psi} | \tilde{\psi} \right\rangle - 1 \right]}_{\text{contrainte}} \tag{37}$$

Avec λ est le multiplicateur de Lagrange. Ce coefficient mesure le poids de la contrainte imposée.

$$\mathcal{L} = \underbrace{\left\langle \tilde{\psi} | \hat{\mathcal{H}} | \tilde{\psi} \right\rangle}_{\text{énergie}} - \lambda \left\langle \tilde{\psi} | \tilde{\psi} \right\rangle + \lambda$$

Ainsi, le multiplicateur λ doit avoir nécessairement les dimensions d'une énergie. Notons :

$$\lambda = \tilde{\varepsilon}$$

$$\mathcal{L} = \left\langle \tilde{\psi} | \hat{\mathcal{H}} | \tilde{\psi} \right\rangle - \tilde{\varepsilon} \left\langle \tilde{\psi} | \tilde{\psi} \right\rangle + \tilde{\varepsilon}$$

$$\Rightarrow \mathcal{L} = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i \, c_j \, \hat{\mathcal{H}}_{ij} - \tilde{\varepsilon} \sum_{i=1}^{n} \sum_{j=1}^{n} c_i \, c_j \, \hat{\mathcal{O}}_{ij} + \tilde{\varepsilon}$$

La minimisation du Lagrangien donne :

$$\Rightarrow \frac{d\mathcal{L}}{dc_i} = \sum_{j=1}^{n} c_j \, \hat{\mathcal{H}}_{ij} - \tilde{\varepsilon} \sum_{j=1}^{n} c_j \, \hat{\mathcal{O}}_{ij} = 0$$

$$\Rightarrow \sum_{j=1}^{n} c_j \left[\hat{\mathcal{H}}_{ij} - \tilde{\varepsilon} \, \hat{\mathcal{O}}_{ij} \right] = 0$$

Nous retrouvons ainsi le déterminant séculaire :

$$\left[\hat{\mathcal{H}}_{ij} - \tilde{\varepsilon} \, \hat{\mathcal{O}}_{ij} \right] = 0$$

Exercice ❷ ☞ Ⓡ

Considérons un électron astreint à se mouvoir le long d'un segment (ou dans une boîte unidimensionnelle) de longueur a. Le comportement de cet électron est décrit par la fonction d'essai :

$$\varphi_n(x) = x^n\,(a-x)^n \tag{38}$$

Afin de simplifier le problème, on prendra a égale à l'unité de longueur et $\{\varphi_n(x)\}_{n=1,2}$. Ainsi, la fonction d'essai "globale" de ce système s'écrit :

$$\tilde{\psi}(x) = \sum_n c_n\,\varphi_n(x) = c_1\,\underbrace{x\,(a-x)}_{\varphi_1} + c_2\,\underbrace{x^2\,(a-x)^2}_{\varphi_2} \tag{39}$$

Nous souhaitons estimer l'énergie totale de l'électron par la méthode des variations linéaires (Méthode de Ritz).

1) Déterminer les coefficients de la combinaison linéaire c_1 et c_2. Écrire l'expression de $\tilde{\psi}(x)$ conduisant à la meilleure combinaison linéaire possible (au sens de l'énergie minimale).
2) Estimer l'énergie variationnelle $\tilde{E}$. Comparer le résultat obtenu à l'énergie exacte $E_0 = \dfrac{h^2}{8\,m\,a^2}$. Commenter
3) Comparer le graphe de $\tilde{\psi}(x)$ à celui de la fonction d'onde exacte de la particule sur un segment dans son état fondamental, soit :

$$\psi_0(x, a=1) = \sqrt{2}\,\sin(\pi\,x) \tag{40}$$

4) L'analogie avec la particule dans une boîte unidimensionnelle est-elle justifiée ?

On donne :

$$\int_0^1 x^m\,(1-x)^n\,dx = \frac{m!\,n!}{(m+n+1)!} \tag{41}$$

et,

$$V = \varepsilon_0\,x \tag{42}$$

C. Méthode des perturbations

1) ***Système quantique non-dégénéré:*** cette méthode est largement utilisée pour le traitement quantique de système complexes de type polyélectroniques (atome, molécule, solide ... etc) afin de résoudre l'équation de Schrödinger. Le principe de base de cette méthode d'approximation repose sur la décomposition de l'Hamiltonien du système en deux contributions distinctes :

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}^{(0)} + \gamma\,\hat{\mathcal{H}}^{(1)} \tag{43}$$

Dans cette configuration les fonctions propres $\psi_k^{(0)}$ et les valeurs propres $E_k^{(0)}$ sont supposées connues avec exactitude, soit :

$$\hat{\mathcal{H}}^{(0)}\,\psi_k^{(0)} = E_k^{(0)}\,\psi_k^{(0)} \tag{44}$$

La deuxième contribution $\gamma\,\hat{\mathcal{H}}^{(1)}$ constitue la perturbation du système. Une correction de l'Hamiltonien sans perturbation $\hat{\mathcal{H}}^{(0)}$. Le paramètre $\gamma << 1$ est un facteur d'échelle. Lorsque $\gamma = 0$ cela signifie que le système n'est soumis à aucune perturbation. En revanche, lorsque $\gamma = 1$ cela signifie

que le système est totalement perturbé. Toute la problématique consiste à trouver les fonctions et les valeurs propres de l'Hamiltonien global $\hat{\mathcal{H}}$:

$$\hat{\mathcal{H}}\,\psi_k = E_k\,\psi_k \tag{45}$$

Un développement en série de ψ_k et E_k donne respectivement :

$$\psi_k = \psi_k^{(0)} + \gamma\,\psi_k^{(1)} + \gamma^2\,\psi_k^{(2)} + O(\gamma^3) \tag{46}$$

et,

$$E_k = E_k^{(0)} + \gamma\,E_k^{(1)} + \gamma^2\,E_k^{(2)} + O(\gamma^3) \tag{47}$$

La quantité $O(\gamma^3)$ exprime les termes d'ordre supérieur ou égale à trois. L'influence de ces termes n'est pas significative. Tenant compte des équations (43), (45), (46) et (47) il vient :

$$\left[\hat{\mathcal{H}}^{(0)} + \gamma\,\hat{\mathcal{H}}^{(1)}\right]\left[\psi_k^{(0)} + \gamma\,\psi_k^{(1)} + \gamma^2\,\psi_k^{(2)}\right] = \left[E_k^{(0)} + \gamma\,E_k^{(1)} + \gamma^2\,E_k^{(2)}\right]\left[\psi_k^{(0)} + \gamma\,\psi_k^{(1)} + \gamma^2\,\psi_k^{(2)}\right]$$

$$\Rightarrow \left[\hat{\mathcal{H}}^{(0)} + \gamma\,\hat{\mathcal{H}}^{(1)}\right]\left[\psi_k^{(0)} + \gamma\,\psi_k^{(1)} + \gamma^2\,\psi_k^{(2)}\right] - \left[E_k^{(0)} + \gamma\,E_k^{(1)} + \gamma^2\,E_k^{(2)}\right]\left[\psi_k^{(0)} + \gamma\,\psi_k^{(1)} + \gamma^2\,\psi_k^{(2)}\right] = 0$$

La distribution terme par terme et la factorisation de γ conduit à :

$$\gamma\left[\hat{\mathcal{H}}^{(0)}\,\psi_k^{(1)} + \hat{\mathcal{H}}^{(1)}\,\psi_k^{(0)} - E_k^{(0)}\,\psi_k^{(1)} - E_k^{(1)}\,\psi_k^{(0)}\right] + \gamma^2\left[\hat{\mathcal{H}}^{(0)}\,\psi_k^{(2)} + \hat{\mathcal{H}}^{(1)}\,\psi_k^{(1)} - E_k^{(0)}\,\psi_k^{(2)} - E_k^{(1)}\,\psi_k^{(1)} - E_k^{(2)}\,\psi_k^{(0)}\right] = 0$$

$$\gamma\left[\psi_k^{(0)}(\hat{\mathcal{H}}^1 - E_k^{(1)}) + \psi_k^{(2)}(\hat{\mathcal{H}}^0 - E_k^{(0)})\right] + \gamma^2\left[\psi_k^{(0)}\,E_k^{(2)} + \psi_k^{(1)}(\hat{\mathcal{H}}^{(1)} - E_k^{(1)}) + \psi_k^{(2)}(\hat{\mathcal{H}}^{(0)} - E_k^{(0)})\right] = 0$$

Les termes en $O(\gamma^3)$ n'interviennent pas dans les calculs. Commençons par le terme de premier ordre en γ :

$$\gamma\left[\psi_k^{(0)}(\hat{\mathcal{H}}^{(1)} - E_k^{(1)}) + \psi_k^{(2)}(\hat{\mathcal{H}}^0 - E_k^{(0)})\right] = 0$$

$$\Rightarrow \left[\psi_k^{(0)}(\hat{\mathcal{H}}^{(1)} - E_k^{(1)}) + \psi_k^{(1)}(\hat{\mathcal{H}}^{(0)} - E_k^{(0)})\right] = 0$$

Multiplions à gauche par le conjugué de $\psi_k^{(0)}$ soit $\psi_k^{(0)*}$ et intégrons :

$$\left\langle\psi_k^{(0)}|\hat{\mathcal{H}}^{(1)}|\psi_k^{(0)}\right\rangle - \left\langle\psi_k^{(0)}|E_k^{(1)}|\psi_k^{(0)}\right\rangle + \left\langle\psi_k^{(0)}|\hat{\mathcal{H}}^{(0)}|\psi_k^{(1)}\right\rangle - \left\langle\psi_k^{(0)}|\hat{\mathcal{H}}^{(0)}|\psi_k^{(1)}\right\rangle - E_k^{(1)}\left\langle\psi_k^{(0)}|\psi_k^{(1)}\right\rangle = 0$$

$$\Rightarrow E_k^{(1)} = \left\langle\psi_k^{(0)}|\hat{\mathcal{H}}^{(1)}|\psi_k^{(0)}\right\rangle + \left\langle\psi_k^{(0)}|\hat{\mathcal{H}}^{(0)}|\psi_k^{(1)}\right\rangle - E_k^{(0)}\left\langle\psi_k^{(0)}|\psi_k^{(1)}\right\rangle \tag{48}$$

Avec $\hat{\mathcal{H}}^{(0)}$ étant un opérateur hermétique, nous pouvons écrire :

$$\left\langle\psi_k^{(0)}|\hat{\mathcal{H}}^{(0)}|\psi_k^{(1)}\right\rangle = \left\langle\hat{\mathcal{H}}^{(0)}\,\psi_k^{(0)}|\psi_k^{(1)}\right\rangle = E_k^{(0)}\left\langle\psi_k^{(0)}|\psi_k^{(1)}\right\rangle \tag{49}$$

Injectons (49) dans (48) nous obtenons l'expression finale :

$$E_k^{(1)} = \left\langle\psi_k^{(0)}|\hat{\mathcal{H}}^{(1)}|\psi_k^{(0)}\right\rangle \tag{50}$$

C'est l'énergie de la perturbation au premier ordre en γ. Calculer $E_k^{(1)}$ revient donc à déterminer la moyenne de l'opérateur perturbation dans l'état non perturbé $\psi_k^{(0)}$. Il en résulte que la méthode

des perturbations est conditionnée par la possibilité :

1) D'une décomposition de l'Hamiltonien global en deux contributions. Une première contribution prédominante $\hat{\mathcal{H}}^{(0)}$ décrivant le système en absence de toutes perturbations. Une deuxième contribution $\gamma\,\hat{\mathcal{H}}^{(1)}$ tenant compte uniquement de l'effet de la perturbation du système en question.

2) De connaitre exactement les fonctions propres $\psi_k^{(0)}$ du terme non pertubatif.

Cherchons désormais la fonction d'onde "perturbée" $\psi_k^{(1)}$ au premier ordre. Cette fonction se développe linéairement sur la base orthogonale des fonctions propres de $\hat{\mathcal{H}}^{(0)}$, soit :

$$\psi_k^{(1)} = \sum_l c_{kl}\,\psi_l^{(0)} \tag{51}$$

D'un autre coté nous avons déjà écrit pour la perturbation de premier ordre :

$$\Rightarrow \left[\psi_k^{(0)}(\hat{\mathcal{H}}^{(1)} - E_k^{(1)}) + \psi_k^{(1)}(\hat{\mathcal{H}}^{(0)} - E_k^{(0)})\right] = 0 \tag{52}$$

En substituant (51) dans (52) il vient :

$$(\hat{\mathcal{H}}^{(1)} - E_k^{(1)})\,\psi_k^{(0)} + \sum_l c_{kl}\,\psi_l^{(0)}\,(\hat{\mathcal{H}}^{(0)} - E_k^{(0)}) = 0 \tag{53}$$

$$\Rightarrow c_{kl}\,(\hat{\mathcal{H}}^{(0)} - E_k^{(0)})\,\psi_l^{(0)} = -(\hat{\mathcal{H}}^{(1)} - E_k^{(1)})\,\psi_k^{(0)} \tag{54}$$

Multiplions à gauche par $\psi_l^{(0)*}$ et intégrons :

$$c_{kl}\left\langle\psi_l^{(0)}|\hat{\mathcal{H}}^{(0)}|\psi_l^{(0)}\right\rangle - E_k^{(0)}\underbrace{\left\langle\psi_l^{(0)}|\psi_l^{(0)}\right\rangle}_{=1} = -\left\langle\psi_l^{(0)}|\hat{\mathcal{H}}^{(1)}|\psi_k^{(0)}\right\rangle - E_k^{(1)}\underbrace{\left\langle\psi_l^{(0)}|\psi_k^{(0)}\right\rangle}_{=0}$$

$$c_{kl}\left\langle\psi_l^{(0)}|\hat{\mathcal{H}}^{(0)}|\psi_l^{(0)}\right\rangle - E_k^{(0)} = -\left\langle\psi_l^{(0)}|\hat{\mathcal{H}}^{(1)}|\psi_k^{(0)}\right\rangle$$

$$c_{kl}\left[E_l^{(0)} - E_k^{(0)}\right] = -\left\langle\psi_l^{(0)}|\hat{\mathcal{H}}^{(1)}|\psi_k^{(0)}\right\rangle$$

$$c_{kl} = -\frac{\left\langle\psi_l^{(0)}|\hat{\mathcal{H}}^{(1)}|\psi_k^{(0)}\right\rangle}{E_l^{(0)} - E_k^{(0)}} \tag{55}$$

Il en résulte que l'expression finale de la fonction d'onde globale du système s'écrit :

$$\psi_k = \psi_k^{(0)} + \gamma\,\psi_k^{(1)}$$

$$\Rightarrow \psi_k = \psi_k^{(0)} - \gamma\sum_{l\neq k}\frac{\left\langle\psi_l^{(0)}|\hat{\mathcal{H}}^{(1)}|\psi_k^{(0)}\right\rangle}{E_l^{(0)} - E_k^{(0)}}\,\psi_k^{(0)} \tag{56}$$

Avec un raisonnement similaire, il est possible d'atteindre la fonction d'onde perturbée d'ordre n.

2) ***Système quantique dégénéré****:* Dans ce cas de figure, qui très fréquent, nous tacherons de répondre à la question : parmi les fonctions propres de $\hat{\mathcal{H}}^{(0)}$, quelle est la fonction qui servira au calcul de la fonction d'onde perturbée ? Mathématiquement, la fonction d'onde d'un état n fois dégénérés est :

$$\psi_k^{(0)} = \sum_n c_n \, \psi_{k,n}^{(0)} \tag{57}$$

Pour la perturbation de premier ordre, nous avons déjà écrit :

$$\psi_k^{(0)}(\hat{\mathcal{H}}^{(1)} - E_k^{(1)}) + \psi_k^{(1)}(\hat{\mathcal{H}}^{(0)} - E_k^{(0)}) = 0 \tag{58}$$

Par substitution de (57) dans (58) nous obtenons :

$$\psi_k^{(1)}(\hat{\mathcal{H}}^{(0)} - E_k^{(0)}) = -\sum_n c_n \, (\hat{\mathcal{H}}^{(1)} - E_k^{(1)}) \, \psi_{k,n}^{(0)} \tag{59}$$

Multiplions à gauche par $\psi_{k,m}^{(0)*}$ (le conjugué complexe de l'une des fonctions de la base $\psi_{k,n}^{(0)}$) et intégrons :

$$\left\langle \psi_{k,m}^{(0)} | \hat{\mathcal{H}}^{(0)} | \psi_k^{(1)} \right\rangle - E_k^{(0)} = -\sum_n c_n \left[\left\langle \psi_{k,m}^{(0)} | \hat{\mathcal{H}}^{(1)} | \psi_{k,n}^{(0)} \right\rangle - E_k^{(1)} \left\langle \psi_{k,m}^{(0)} | \psi_{k,n}^{(0)} \right\rangle \right] \tag{60}$$

Avec $\hat{\mathcal{H}}^{(0)}$ étant un opérateur hermétique, nous pouvons écrire :

$$\left\langle \psi_{k,m}^{(0)} | \hat{\mathcal{H}}^{(0)} | \psi_{k,n}^{(1)} \right\rangle = \left\langle \psi_{k,m}^{(0)} | \hat{\mathcal{H}}^{(0)} \, \psi_{k,n}^{(1)} \right\rangle = \left\langle \hat{\mathcal{H}}^{(0)} \, \psi_{k,m}^{(0)} | \psi_{k,n}^{(1)} \right\rangle = E_k^{(0)}$$

Avec,

$$\hat{\mathcal{H}}^{(0)} \, \psi_{k,m}^{(0)} = E_k^{(0)} \, \psi_{k,m}^{(0)}$$

De (60) nous en déduisons :

$$\sum_n c_n \left[\hat{\mathcal{H}}_{m,n}^{(1)} - \delta_{m,n} \, E_k^{(1)} \right] = 0 \tag{61}$$

$$\Rightarrow \left[\hat{\mathcal{H}}_{m,n}^{(1)} - \delta_{m,n} \, E_k^{(1)} \right] = 0 \tag{62}$$

Écrivons ce déterminant séculaire pour $n = 2 \, (m = n)$:

$$\begin{pmatrix} \hat{\mathcal{H}}_{11}^{(1)} & \hat{\mathcal{H}}_{12}^{(1)} \\ \hat{\mathcal{H}}_{21}^{(1)} & \hat{\mathcal{H}}_{22}^{(1)} \end{pmatrix} - E_k^{(1)} \times \begin{pmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{pmatrix} = 0$$

La base $\psi_{k,n}^{(0)}$ étant orthogonale : $\delta_{21} = \delta_{12} = 0$ alors :

$$\begin{pmatrix} \hat{\mathcal{H}}_{11}^{(1)} - E_k^{(1)} \, \delta_{21} & \hat{\mathcal{H}}_{12}^{(1)} \\ \hat{\mathcal{H}}_{21}^{(1)} & \hat{\mathcal{H}}_{22}^{(1)} - E_k^{(1)} \, \delta_{22} \end{pmatrix} = 0$$

La résolution de ce déterminant donne un polynôme caractéristique d'ordre deux en $E_k^{(1)}$. Les racines de ce polynôme sont $E_{k,1}^{(1)}$ et $E_{k,2}^{(1)}$. Nous terminons cette section en disant que la théorie des perturbations consiste en une succession de corrections d'un problème non perturbé. Les méthodes de variation et de perturbation sont susceptibles d'obtenir de très bons résultats si elles sont appliquées de façon rigoureuse.

Exercice ❸ ☞ Ⓢ

Identifier les termes $\hat{\mathcal{H}}^{(0)}$, $\hat{\mathcal{H}}^{(1)}$, $\psi^{(0)}$ et $E^{(0)}$ pour les problèmes suivants :

1) Un oscillateur gouverné par le potentiel :

$$V(x) = \frac{k\,x^2}{2} + \frac{\gamma\,x^3}{6} + \frac{\gamma\,x^4}{24} \tag{63}$$

2) Particule dans une boîte à une dimension.

$$V(x) = \begin{cases} 0 & 0 < x < a/2 \\ b & a/2 < x < a \end{cases} \tag{64}$$

3) Un atome d'hydrogène soumis à champ électrique d'intensité $|\vec{\epsilon}|$. L'Hamiltonien de système s'écrit :

$$\hat{\mathcal{H}} = -\frac{\hbar^2}{2\,m}\,\nabla^2 - \frac{e^2}{4\,\pi\,\epsilon_0\,r} + e\,\epsilon\,\cos\theta \tag{65}$$

Exercice ❹ ☞ Ⓢ

Un électron astreint à se mouvoir le long d'un segment de longueur a $(0 \leq x \leq a)$. Sur la portion $x = a/4$ à $x = 3\,a/4$ la particule est régit par le potentiel $V(x) = \varepsilon$. En dehors de cette portion le potentiel est nul $V(x) = 0$.

1) Calculer l'énergie de la première correction $E^{(1)}$ pour :

   a- L'état fondamental $E^{(0)}_{n=1}$.
   b- Le premier état excité $E^{(0)}_{n=2}$.

2) Déterminer la fonction d'onde normalisée de la correction de premier ordre en utilisant jusqu'à $n = 4$ des fonctions d'onde non perturbées.

Désormais la particule est régit par le potentiel $V(x) = \varepsilon_0 \sin\left(\frac{3\,\pi\,x}{a}\right)$ sur tout le segment.

1) Calculer la correction de second ordre de l'énergie de l'état fondamental du système perturbé en utilisant jusqu'à $n = 4$ fonctions d'onde non perturbées.

Exercice ❺ ☞ Ⓢ

Le potentiel de Morse décrit l'énergie potentielle d'interaction d'une molécule diatomique. L'expression de ce potentiel est :

$$V(x) = D\,(1 - e^{-\beta\,x})^2 \tag{66}$$

Avec D et β sont des paramètres d'ajustement pour un système donné. Par exemple pour H_2 : $D = 7.61 \times 10^{-19}$ et $\beta = 0.019\,pm^{-1}$.

1) Montrer que le potentiel de Morse peut s'écrire :

$$\hat{\mathcal{H}} = -\frac{\hbar^2}{2\,m}\,\frac{d^2}{dx^2} + a\,x^2 + b\,x^3 + c\,x^4 + \cdots + O(x^n) \tag{67}$$

2) Exprimer les paramètres D et β en fonction des coefficients du développement en série ci-dessus.
3) Identifier les termes $\hat{\mathcal{H}}^{(0)}$, $\psi^{(0)}$ et $E^{(0)}$.

4) En utilisant la méthode des perturbations, calculer l'énergie de la première perturbation. La fonction d'onde est :

$$\psi_k^{(0)}(x) = \left[\frac{a}{\pi}\right]^{(1/4)} e^{-\alpha x^2} \tag{68}$$

III. Méthode de Hartree-Fock

Dans le modèle des électrons indépendants, la fonction d'onde totale s'écrit comme un produit des spin-orbitales, c'est le produit de *Hartree* :

$$\Psi(r) = \prod_{i=1}^{N} \phi_i(r) \tag{69}$$

Toutefois, cette approximation s'est révélée rapidement incohérente ou incompatible avec le principe de *Pauli*, selon lequel une fonction d'onde doit être antisymétrique par rapport à l'échange de spin entre deux électrons. Cette incompatibilité est levée par l'écriture de la fonction d'onde totale du système étudié comme un déterminant de *Slater*. Selon ce dernier, la fonction d'onde [1] d'un système à deux électrons (1) et (2) $\uparrow\downarrow$ (ou $\downarrow\uparrow$) s'écrit selon le déterminant suivant :

$$\Psi(1,2) = \frac{1}{\sqrt{2}} \begin{vmatrix} \phi_1(1) & \phi_1(2) \\ \phi_2(1) & \phi_2(2) \end{vmatrix} \quad \Rightarrow \Psi = \frac{1}{\sqrt{2}} \left[\phi_1(1)\,\phi_2(2) - \phi_2(1)\,\phi_1(2)\right] \tag{70}$$

Avec ϕ_i est une spin-orbitale [2], c'est le produit d'une fonction spatiale (orbitale) par une fonction de spin :

$$\phi_i = \underbrace{\varphi_i}_{\text{fonction spatiale}} \times \underbrace{\sigma_i}_{\text{fonction de spin}} \qquad \text{avec} \quad \sigma_1 = \alpha\,(\uparrow) \quad \sigma_2 = \beta\,(\downarrow) \tag{71}$$

Ce qui donne pour notre système à deux électrons, les deux spin-orbitales $\phi_1 = \varphi_1 \times \alpha$ et $\phi_2 = \varphi_1 \times \beta$. L'orbitale spatiale φ_1 est une orbitale atomique de type $|1s\rangle$, ou $|2s\rangle$, ou $|2p\rangle$ ... etc. Le déterminant de *Slater* est réécrit alors selon :

$$\Psi(1,2) = \frac{1}{\sqrt{2}} \begin{vmatrix} \varphi_1(1)\,\alpha(1) & \varphi_1(2)\,\alpha(2) \\ \varphi_1(1)\,\beta(1) & \varphi_1(2)\,\beta(2) \end{vmatrix} \tag{72}$$

$$\Rightarrow \Psi(1,2) = \frac{1}{\sqrt{2}} \left[\varphi_1(1)\,\alpha(1)\,\varphi_1(2)\,\beta(2) - \varphi_1(1)\,\beta(1)\,\varphi_1(2)\,\alpha(2)\right] \tag{73}$$

Réécrivons le même déterminant en permutant les deux électrons, nous obtenons :

$$\Psi(2,1) = \frac{1}{\sqrt{2}} \begin{vmatrix} \varphi_1(2)\,\alpha(2) & \varphi_1(1)\,\alpha(1) \\ \varphi_1(2)\,\beta(2) & \varphi_1(1)\,\beta(1) \end{vmatrix} \tag{74}$$

$$\Rightarrow \Psi(2,1) = \frac{1}{\sqrt{2}} \left[\varphi_1(2)\,\alpha(2)\,\varphi_1(1)\,\beta(1) - \varphi_1(2)\,\beta(2)\,\varphi_1(1)\,\alpha(1)\right] \tag{75}$$

1. Selon la théorie quantique, on ne peut associer une trajectoire à un électron dans un atome. On définit plutôt une fonction d'onde qui une fonction des coordonnées de la particule en question. Le carré de cette fonction d'onde représente la probabilité de présence de la particule en tout points de l'espace.

2. Une orbitale est une fonction d'onde mono-électronique, sa connaissance permet de déterminer l'énergie et la probabilité de présence de l'électron qui l'occupe. Une orbitale moléculaire est délocalisée spatialement sur l'ensemble de la molécule.

En comparant les équations (73) et (75) il vient :

$$\Psi(1,2) = -\Psi(2,1) \tag{76}$$

Nous remarquons qu'en permettant les deux électrons, la fonction d'onde totale change de signe, par conséquent la fonction d'onde Ψ est antisymétrique et donc le principe de *Pauli* est respecté. Regardons maintenant que se passera t-il si deux spin-orbitales sont occupées avec deux électrons de même spin soit avec la configuration électronique $\uparrow\uparrow$ ou bien $\downarrow\downarrow$. La fonction d'onde du système s'écrit alors selon :

$$\Psi(1,2) = \frac{1}{\sqrt{2}} \begin{vmatrix} \varphi_1(1)\,\alpha(1) & \varphi_1(2)\,\alpha(2) \\ \varphi_1(1)\,\alpha(1) & \varphi_1(2)\,\alpha(2) \end{vmatrix} \tag{77}$$

$$\Rightarrow \Psi(1,2) = \frac{1}{\sqrt{2}} \left[\varphi_1(1)\,\alpha(1)\,\varphi_1(2)\,\alpha(2) - \varphi_1(1)\,\alpha(1)\,\varphi_1(2)\,\alpha(2)\right] = 0 \tag{78}$$

$$\Rightarrow \text{ la densité de probabilité de présence } \int_v |\Psi(1,2)|^2\, dv = 0 \tag{79}$$

Cela signifie que les configurations électroniques $\uparrow\uparrow$ et $\downarrow\downarrow$ sont interdites. Autrement dit, les deux électrons (fermions) ne peuvent avoir le même état quantique, c'est-à-dire avoir les mêmes valeurs des quarte nombres quantiques. Ce principe impose par exemple que deux électrons de valence de deux atomes peuvent former une liaison chimique et limite aussi le nombre d'électron par couche électronique.

Par ailleurs, le système des unités atomiques simplifie grandement l'écriture mathématique des Hamiltoniens. Dans ce système, plusieurs grandeurs sont ramenées à l'unité. L'Hamiltonien électronique exacte d'une molécule s'écrit :

$$\hat{\mathcal{H}}_e = \sum_i^N \hat{h}^c(i) + \sum_{i\neq j}^N \frac{1}{r_{ij}} = \sum_i^N \hat{h}^c(i) + \sum_{i\neq j}^N \hat{h}(i,j) \quad \text{avec} \quad \hat{h}_i^c = \sum_i^N \frac{-\nabla^2(i)}{2} - \sum_i^N \sum_k^M \frac{-Z_k}{r_{ik}} \tag{80}$$

Où le premier terme $\hat{h}_i^c$ est appelé *l'Hamiltonien de cœur*. C'est l'expression mathématique de l'énergie d'un électron se baignant dans le champ électrostatique des noyaux en absence des $(N-1)$ autres électrons. Le deuxième terme de l'hamiltonien électronique exacte exprime la répulsion électrostatique entre deux électrons i et j. Afin de simplifier la description, considérons un système à deux électrons notés (1) et (2) :

$$\hat{\mathcal{H}}_e = \hat{h}^c(1) + \hat{h}^c(2) + \frac{1}{r_{12}} \tag{81}$$

La fonction d'onde de ce système à deux électrons s'écrit selon le déterminant de *Slater* :

$$\Psi = \frac{1}{\sqrt{2}} \begin{vmatrix} \phi_1(1) & \phi_1(2) \\ \phi_2(1) & \phi_2(2) \end{vmatrix} \quad \Rightarrow \Psi = \frac{1}{\sqrt{2}} \left[\phi_1(1)\,\phi_2(2) - \phi_2(1)\,\phi_1(2)\right] \tag{82}$$

L'énergie moyenne associée à cette fonction d'onde s'écrit :

$$E = \left\langle \Psi | \hat{\mathcal{H}}_e | \Psi \right\rangle = \left\langle \Psi | \hat{h}^c(1) + \hat{h}^c(2) | \Psi \right\rangle + \left\langle \Psi | \hat{h}(1,2) | \Psi \right\rangle \tag{83}$$

$$\Rightarrow E = \underbrace{\left\langle \Psi|\hat{h}^c(1)|\Psi \right\rangle}_{\text{T1}} + \underbrace{\left\langle \Psi|\hat{h}^c(2)|\Psi \right\rangle}_{\text{T2}} + \underbrace{\left\langle \Psi|\hat{h}(1,2)|\Psi \right\rangle}_{\text{T3}} \tag{84}$$

Commençons par développer le premier terme T1 de l'énergie électronique totale,

$$\left\langle \Psi|\hat{h}^c(1)|\Psi \right\rangle = \left\langle \frac{1}{\sqrt{2}}\left[\phi_1(1)\,\phi_2(2) - \phi_2(1)\,\phi_1(2)\right]|\hat{h}^c(1)|\frac{1}{\sqrt{2}}\left[\phi_1(1)\,\phi_2(2) - \phi_2(1)\,\phi_1(2)\right] \right\rangle \tag{85}$$

$$= \frac{1}{2}\left\langle \phi_1(1)\,\phi_2(1)|\hat{h}^c(1)|\phi_1(1)\,\phi_2(1) \right\rangle + \frac{1}{2}\left\langle \phi_2(1)\,\phi_1(2)|\hat{h}^c(1)|\phi_2(1)\,\phi_1(2) \right\rangle$$

$$-\frac{1}{2}\left\langle \phi_1(1)\,\phi_2(1)|\hat{h}^c(1)|\phi_2(1)\,\phi_1(2) \right\rangle - \frac{1}{2}\left\langle \phi_2(1)\,\phi_1(2)|\hat{h}^c(1)|\phi_1(1)\,\phi_2(2) \right\rangle$$

$$\Rightarrow \left\langle \Psi|\hat{h}^c(1)|\Psi \right\rangle = \frac{1}{2}\left\langle \phi_1(1)|\hat{h}^c(1)|\phi_1(1) \right\rangle \langle\phi_2(1)|\phi_2(1)\rangle + \frac{1}{2}\left\langle \phi_2(1)|\hat{h}^c(1)|\phi_2(1) \right\rangle \langle\phi_1(2)|\phi_1(2)\rangle$$

$$-\frac{1}{2}\left\langle \phi_1(1)|\hat{h}^c(1)|\phi_2(1) \right\rangle \langle\phi_2(1)|\phi_1(2)\rangle - \frac{1}{2}\left\langle \phi_2(1)|\hat{h}^c(1)|\phi_1(1) \right\rangle \langle\phi_1(2)|\phi_2(2)\rangle$$

En raison de l'orthonormalisation des spin-orbitales $\langle\phi_1|\phi_1\rangle = 1$ ou $(\langle\phi_2|\phi_2\rangle = 1)$ et $\langle\phi_1|\phi_2\rangle = 0$ ou $(\langle\phi_2|\phi_1\rangle = 0)$ il en découle :

$$\Rightarrow \underbrace{\left\langle \Psi|\hat{h}^c(1)|\Psi \right\rangle}_{\text{T1}} = \frac{1}{2}\left\langle \phi_1(1)|\hat{h}^c(1)|\phi_1(1) \right\rangle + \frac{1}{2}\left\langle \phi_2(1)|\hat{h}^c(1)|\phi_2(1) \right\rangle \tag{86}$$

De manière analogue, le deuxième terme T2 s'obtient en remplaçant simplement l'électron (1) par l'électron (2) soit :

$$\Rightarrow \underbrace{\left\langle \Psi|\hat{h}^c(2)|\Psi \right\rangle}_{\text{T2}} = \frac{1}{2}\left\langle \phi_1(2)|\hat{h}^c(2)|\phi_1(2) \right\rangle + \frac{1}{2}\left\langle \phi_2(2)|\hat{h}^c(2)|\phi_2(2) \right\rangle \tag{87}$$

Ainsi pour l'opérateur mono-électronique nous écrirons :

$$\left\langle \Psi|\hat{h}^c(1) + \hat{h}^c(2)|\Psi \right\rangle = \frac{1}{2}\left[\left\langle \phi_1(1)|\hat{h}^c(1)|\phi_1(1) \right\rangle + \left\langle \phi_2(1)|\hat{h}^c(1)|\phi_2(1) \right\rangle + \left\langle \phi_1(2)|\hat{h}^c(2)|\phi_1(2) \right\rangle + \left\langle \phi_2(2)|\hat{h}^c(2)|\phi_2(2) \right\rangle\right] \tag{88}$$

Les quatre intégrales de l'équation (88) dépendent de la nature mathématique des fonctions ϕ_1 et ϕ_2 indépendamment des électrons (1) et (2). Par conséquent :

$$\left\langle \phi_1(1)|\hat{h}^c(1)|\phi_1(1) \right\rangle = \left\langle \phi_1(2)|\hat{h}^c(2)|\phi_1(2) \right\rangle = \hat{\mathcal{H}}_1$$

$$\left\langle \phi_2(1)|\hat{h}^c(1)|\phi_2(1) \right\rangle = \left\langle \phi_2(2)|\hat{h}^c(2)|\phi_2(2) \right\rangle = \hat{\mathcal{H}}_2$$

Nous remarquons que l'indice du nombre $\hat{\mathcal{H}}_i$ se réfère exclusivement à la spin-orbitale ϕ_i. Il en ressort :

$$\Rightarrow \left\langle \Psi|\hat{h}^c(1) + \hat{h}^c(2)|\Psi \right\rangle = \hat{\mathcal{H}}_1 + \hat{\mathcal{H}}_2 \tag{89}$$

Développons désormais le terme T3 de l'opérateur bi-électronique $\hat{h}(1,2)$:

$$\left\langle \Psi|\hat{h}(1,2)|\Psi\right\rangle = \frac{1}{2}\left\langle [\phi_1(1)\,\phi_2(2)-\phi_2(1)\,\phi_1(2)]\,|\hat{h}(1,2)|\,[\phi_1(1)\,\phi_2(2)-\phi_2(1)\,\phi_1(2)]\right\rangle \quad \text{avec} \quad \hat{h}(1,2)=\frac{1}{r_{12}}$$

$$= \frac{1}{2}\underbrace{\left\langle \phi_1(1)\,\phi_2(2)|\hat{h}(1,2)|\phi_1(1)\,\phi_2(2)\right\rangle}_{I1} + \frac{1}{2}\underbrace{\left\langle \phi_2(1)\,\phi_1(2)|\hat{h}(1,2)|\phi_2(1)\,\phi_1(2)\right\rangle}_{I2}$$

$$-\frac{1}{2}\underbrace{\left\langle \phi_1(1)\,\phi_2(2)|\hat{h}(1,2)|\phi_2(1)\,\phi_1(2)\right\rangle}_{I3} - \frac{1}{2}\underbrace{\left\langle \phi_2(1)\,\phi_1(2)|\hat{h}(1,2)|\phi_1(1)\,\phi_2(2)\right\rangle}_{I4}$$

Nous remarquons que les couples d'intégrales I1 et I2 puis I3 et I4 sont identiques, nous avons juste permuté les électrons (1) et (2).

$$\Rightarrow \left\langle \Psi|\hat{h}(1,2)|\Psi\right\rangle = \underbrace{\left\langle \phi_1(1)\,\phi_2(2)|\hat{h}(1,2)|\phi_1(1)\,\phi_2(2)\right\rangle}_{I5} - \underbrace{\left\langle \phi_1(1)\,\phi_2(2)|\hat{h}(1,2)|\phi_2(1)\,\phi_1(2)\right\rangle}_{I6}$$

Qui s'écrit facilement sous la forme simplifiée :

$$\Rightarrow \left\langle \Psi|\hat{h}(1,2)|\Psi\right\rangle = \langle \phi_1\,\phi_2|\phi_1\,\phi_2\rangle - \langle \phi_1\,\phi_2|\phi_2\,\phi_1\rangle \tag{90}$$

Nous rappelons qu'une spin-orbitale ϕ_i s'écrit systématiquement sous forme d'un produit d'une fonction spatiale $\varphi(i)$ multipliée par une fonction de spin σ_i.

$$\Rightarrow I5 = \left\langle \varphi_1(1)\,\sigma_1(1)\,\varphi_2(2)\,\sigma_2(2)|\hat{h}(1,2)|\varphi_1(1)\,\sigma_1(1)\,\varphi_2(2)\,\sigma_2(2)\right\rangle$$

$$= \left\langle \varphi_1(1)\,\varphi_2(2)|\hat{h}(1,2)|\varphi_1(1)\,\varphi_2(2)\right\rangle \langle \sigma_1(1)\,\sigma_2(2)|\sigma_1(1)\,\sigma_2(2)\rangle$$

$$= \underbrace{\left\langle \varphi_1(1)\,\varphi_2(2)|\hat{h}(1,2)|\varphi_1(1)\,\varphi_2(2)\right\rangle}_{\mathcal{J}_{12}} \underbrace{\langle \sigma_1(1)|\sigma_1(1)\rangle}_{=1} \underbrace{\langle \sigma_2(2)|\sigma_2(2)\rangle}_{=1}$$

L'intégrale $\mathcal{J}_{12}$ est appelée *intégrale Coulombienne*. Avec un raisonnement similaire, nous obtenons :

$$\Rightarrow I6 = \left\langle \varphi_1(1)\,\sigma_1(1)\,\varphi_2(2)\,\sigma_2(2)|\hat{h}(1,2)|\varphi_2(1)\,\sigma_2(1)\,\varphi_1(2)\,\sigma_1(2)\right\rangle$$

$$= \left\langle \varphi_1(1)\,\varphi_2(2)|\hat{h}(1,2)|\varphi_2(1)\,\varphi_1(2)\right\rangle \langle \sigma_1(1)\,\sigma_2(2)|\sigma_2(1)\,\sigma_1(2)\rangle$$

$$= \underbrace{\left\langle \varphi_1(1)\,\varphi_2(2)|\hat{h}(1,2)|\varphi_2(1)\,\varphi_1(2)\right\rangle}_{\mathcal{K}_{12}} \underbrace{\langle \sigma_1(1)|\sigma_2(1)\rangle}_{=1} \underbrace{\langle \sigma_2(2)|\sigma_1(2)\rangle}_{=1}$$

L'intégrale $\mathcal{K}_{12}$ est appelée *intégrale d'échange*. Cette intégrale est une entité mathématique purement quantique découlant de la propriété d'antisymétrie que doivent vérifier les fonctions d'onde afin de ne pas violer le principe d'exclusion de *Pauli*. En termes plus précis, cette intégrale permet de tenir compte de l'état de spin (multiplicité) de l'atome ou de la molécule. A partir de l'équation (III), si les deux électrons se trouvant dans les orbitales φ_i et φ_j ont des spins différents ($\langle\sigma_1(1)|\sigma_2(1)\rangle = 0$, ie : si $\sigma_2 \neq \sigma_1$) alors le terme d'échange $\mathcal{K}_{ij}$ est annihilé. Dans ce cas, la répulsion électrostatique moyenne prend en compte uniquement le terme *Coulombien*. Il en découle de cette analyse que la répulsion électrostatique moyenne entre deux électrons dans des orbitales φ_i et φ_j vaut :

$$\hat{h}(1,2) = \frac{1}{r_{12}} = \mathcal{J}_{ij} \quad \text{si les électrons sont de spins différents}$$

$$\hat{h}(1,2) = \frac{1}{r_{12}} = \mathcal{J}_{ij} - \mathcal{K}_{ij} \quad \text{si les électrons sont de mêmes spins}$$

Ainsi, l'énergie électronique totale du système à deux électrons s'écrit en définitive suivant :

$$E_e^T = \hat{h}^c(1) + \hat{h}^c(2) + \mathcal{J}_{12} - \mathcal{K}_{12} \tag{91}$$

Ce résultat se généralise immédiatement pour un système à N électrons selon la relation :

$$E_e^T = \sum_i^N \hat{h}^c(i) + \sum_{i \neq j}^N [\mathcal{J}_{ij} - \mathcal{K}_{ij}] \tag{92}$$

D'après l'équation (92), l'énergie totale d'un électron (i) est égale à l'énergie de son interaction avec le champ électrostatique des noyaux (premier terme) et la somme des répulsions électrostatiques entre l'électron en question occupant l'orbitale φ_i et les autres électrons (j) occupant les orbitales φ_j. Dans le cas d'un système à couche fermée (closed shell), les orbitales spatiales φ_i sont doublement occupées par deux électrons de spins différents α (spin up) et β (spin down) :

$$\underbrace{\{\phi_1 = \varphi_1 \times \alpha\,,\, \phi_2 = \varphi_1 \times \beta\}}_{\text{1ier doublet}} \quad \underbrace{\{\phi_3 = \varphi_2 \times \alpha\,,\, \phi_4 = \varphi_2 \times \beta\}}_{\text{2ème doublet}} \cdots \underbrace{\{\phi_{n-1} = \varphi_{n/2} \times \alpha\,,\, \phi_n = \varphi_{n/2} \times \beta\}}_{\text{nième doublet}}$$

Tenant compte de cette propriété, l'opérateur mono-électronique $\hat{h}^c(i)$ se répète deux fois selon :

$$\left\langle \varphi_1(1)\,\sigma_1(1) | \hat{h}^c(1) | \varphi_1(1)\,\sigma_1(1) \right\rangle = \left\langle \varphi_2(2)\,\sigma_2(2) | \hat{h}^c(1) | \varphi_2(2)\,\sigma_2(2) \right\rangle$$

$$\Rightarrow \sum_{i=1}^N \hat{h}^c(i) = 2 \sum_{i=1}^{N/2} \hat{h}^c(i) \tag{93}$$

Avec un raisonnement similaire, l'intégrale *Coulombienne* $\mathcal{J}_{ij}$ se répète quatre fois, ce qui donne :

$$\Rightarrow \sum_{i \neq j}^N \mathcal{J}_{ij} = 4 \sum_{i \neq j}^{N/2} \mathcal{J}_{ij} \tag{94}$$

D'un autre côté, l'intégrale *d'échange* $\mathcal{K}_{ij}$ se répète deux fois, ce qui donne :

$$\Rightarrow \sum_{i \neq j}^N \mathcal{K}_{ij} = 2 \sum_{i \neq j}^{N/2} \mathcal{J}_{ij} \tag{95}$$

En substituant (93), (94) et (95) dans (92), nous obtenons finalement :

$$E_e^T = 2 \sum_{i=1}^{N/2} \hat{h}^c(i) + \sum_{i \neq j}^{N/2} [2\,\mathcal{J}_{ij} - \mathcal{K}_{ij}] \tag{96}$$

L'équation (96) est l'énergie électronique totale dans le cadre de la théorie de *Hartree-Fock Restreinte* (Restricted Hartree-Fock). Cette même équation peut s'écrire aussi en fonction des énergies des orbitales spatiales (φ_i) doublement occupées. Il en ressort que l'énergie électronique totale (E_e^T) est égale à la somme des énergies des orbitales occupées corrigée de la somme des énergies de répulsion électrostatique entre électrons :

$$E_e^T = \sum_i^{occ} E_i - \sum_{i \neq j}^{occ} [2\,\hat{\mathcal{J}}_{ij} - \hat{\mathcal{K}}_{ij}] \tag{97}$$

Le premier terme exprime l'énergie cinétique de la paire d'électrons se trouvant dans l'orbitale spatiale φ_i et de son énergie d'interaction électrostatique avec les noyaux. Le deuxième terme exprime l'énergie de répulsion de la paire d'électrons dans l'orbitale spatiale φ_i avec toutes les autres paires d'électrons se trouvant dans les orbitales moléculaires φ_j. Par ailleurs, l'énergie totale de la molécule (E_m^T) est égale à l'énergie électronique totale à laquelle il faudra ajouter la répulsion électrostatique entre une paire de noyaux :

$$E_m^T = \sum_{i}^{occ} E_i - \sum_{i\neq j}^{occ} [2\,\hat{\mathcal{J}}_{ij} - \hat{\mathcal{K}}_{ij}] + \sum_{k\neq l}^{M} \frac{Z_k\, Z_l}{r_{kl}} \tag{98}$$

C'est l'énergie totale de la molécule qui est minimisée lors d'une opération d'optimisation de la géométrie moléculaire. Nous cherchons sa conformation la plus stable et les valeurs optimales des ses paramètres géométriques. En outre, l'optimisation de la géométrie est la procédure qui consiste à trouver la configuration de l'énergie minimale de la molécule. Cette opération calcule la fonction d'onde et l'énergie de la géométrie initiale et procède ensuite à la recherche d'une nouvelle géométrie d'énergie plus faible. Cette opération est répétée jusqu'à ce que la géométrie d'énergie la plus faible soit trouvée. La force sur chaque atome est calculée en évaluant le gradient de l'énergie par rapport aux positions atomiques. Des algorithmes d'optimisation très sophistiqués sont ensuite utilisés à chaque étape pour sélectionner une nouvelle géométrie, visant une convergence rapide vers la géométrie de plus basse d'énergie. Dans la géométrie finale d'énergie minimale, la force sur chaque atome est nulle. Il est important de noter que cette opération d'optimisation ne convergera pas nécessairement vers un minimum global ayant la plus basse énergie moléculaire. L'optimisation s'arrête lorsqu'elle l'algorithme bute sur un point stationnaire (ou un point critique) pour lequel le gradient de la fonction énergie est nul. Ce point stationnaire, peut être un minimum global, un minimum local ou carrément un point selle (géométrie ou molécule de transition). Cela se produira en particulier si nous limitons la symétrie de la molécule de sorte que l'algorithme d'optimisation sera incapable d'explorer tout l'espace des configurations moléculaires. Il est donc fortement recommandé de commencer une opération d'optimisation de la géométrie moléculaire avec une petite base de fonctions avant de passer à une base plus étendue. Il est possible ensuite de lancer l'optimisation finale de la géométrie à partir de la géométrie obtenue avec la base réduite.

IV. Équation de Hartree-Fock

Dans ce calcul variationnel, nous utilisons la méthode de *Lagrange* à travers la fonctionnelle $\mathcal{L}$ définie ci-dessous. Nous exigeons, par le biais des multiplicateurs de Lagrange, que l'ensemble des orbitales ϕ_i demeure orthogonal tout au long du processus de minimisation. La condition à remplir est alors :

$$\mathcal{L} = E - \sum_{i,j}^{N} \lambda_{ij} \left[\langle\phi_i|\phi_j\rangle - \delta_{ij}\right] \quad \Rightarrow \mathcal{L} = E - \sum_{i,j}^{N} \lambda_{ij} \langle\phi_i|\phi_j\rangle = \underbrace{-\sum_{i,j}^{N} \lambda_{ij}\,\delta_{ij}}_{\text{cst}}$$

$$\delta\mathcal{L} = \delta E - \sum_{i,j}^{N} \lambda_{ij} \left[\langle\delta\phi_i|\phi_j\rangle + \langle\phi_i|\delta\phi_j\rangle\right] = 0 \tag{99}$$

D'un autre côté, nous avons l'énergie de *Hartree-Fock* :

$$E = \sum_{i}^{N} \left\langle\phi_i|\hat{h}_i|\phi_i\right\rangle + \frac{1}{2}\sum_{i,j}^{N} \left[\left\langle\phi_j|\hat{\mathcal{J}}_i|\phi_j\right\rangle - \left\langle\phi_j|\hat{\mathcal{K}}_i|\phi_j\right\rangle\right]$$

Nous devons ainsi minimiser l'expression de l'énergie de *Hartree-Fock* par rapport aux changements des orbitales $\phi_i \longrightarrow \phi_i + \delta\phi_i$

$$\Rightarrow \delta E = \sum_i^N \left\langle \delta\phi_i|\hat{h}_i|\phi_i\right\rangle + \left\langle \phi_i|\hat{h}_i|\delta\phi_i\right\rangle + \frac{1}{2}\sum_{i,j}^N \left[\left\langle \delta\phi_j|\hat{\mathcal{J}}_i|\phi_j\right\rangle + \left\langle \phi_j|\hat{\mathcal{J}}_i|\delta\phi_j\right\rangle - \left\langle \delta\phi_j|\hat{\mathcal{K}}_i|\phi_j\right\rangle - \left\langle \phi_j|\hat{\mathcal{K}}_i|\delta\phi_j\right\rangle\right]$$

$$\Rightarrow \delta E = \sum_i^N \left\langle \delta\phi_i|\hat{h}_i|\phi_i\right\rangle + \left\langle \phi_i|\hat{h}_i|\delta\phi_i\right\rangle + \frac{1}{2}\sum_{i,j}^N \left[\left\langle \delta\phi_j|\hat{\mathcal{J}}_i - \hat{\mathcal{K}}_i|\phi_j\right\rangle + \left\langle \phi_j|\hat{\mathcal{J}}_i - \hat{\mathcal{K}}_i|\delta\phi_j\right\rangle\right]$$

L'opérateur de Fock s'écrit

$$\hat{\mathcal{F}}_i = \hat{h}_i + \sum_j^N \underbrace{\left[\hat{\mathcal{J}}_j - \hat{\mathcal{K}}_j\right]}_{V_{\rm HF}(i)} \quad \text{ou avec la notation} \quad \hat{\mathcal{F}}(i) = \hat{h}(i) + \sum_j^N \underbrace{\left[\hat{\mathcal{J}}_j(i) - \hat{\mathcal{K}}_j(i)\right]}_{V_{\rm HF}(i)} \tag{100}$$

Avec $V_{\rm HF}(i)$ est le potentiel de Hartree-Fock. C'est le champ électrostatique moyen, créé par les électrons j ($N-1$ électrons restants), ressenti par l'électron i. Par conséquent,

$$\Rightarrow \delta E = \sum_i^N \left\langle \delta\phi_i|\hat{\mathcal{F}}_i|\phi_i\right\rangle + \left\langle \phi_i|\hat{\mathcal{F}}_i|\delta\phi_i\right\rangle \tag{101}$$

En substituant (99) dans (101) nous obtenons :

$$\Rightarrow \delta\mathcal{L} = \sum_i^N \left\langle \delta\phi_i|\hat{\mathcal{F}}_i|\phi_i\right\rangle + \left\langle \phi_i|\hat{\mathcal{F}}_i|\delta\phi_i\right\rangle + \sum_{i,j}^N \lambda_{ij}\ [\langle\delta\phi_i|\phi_j\rangle + \langle\phi_i|\delta\phi_j\rangle] \tag{102}$$

$$\Rightarrow \delta\mathcal{L} = \sum_i^N \left[2\left\langle \delta\phi_i|\hat{\mathcal{F}}_i|\phi_i\right\rangle + \sum_j^N \lambda_{ij}\ \langle\delta\phi_i|\phi_j\rangle\right] = 0 \tag{103}$$

L'équation (103) est vérifiée si :

$$2\left\langle \delta\phi_i|\hat{\mathcal{F}}_i|\phi_i\right\rangle + \sum_j^N \lambda_{ij}\ \langle\delta\phi_i|\phi_j\rangle = 0 \quad \Rightarrow \left\langle \delta\phi_i|\hat{\mathcal{F}}_i|\phi_i\right\rangle = -\frac{1}{2}\sum_j^N \lambda_{ij}\ \langle\delta\phi_i|\phi_j\rangle \tag{104}$$

L'équation (104) est équivalente à l'équation aux valeurs propres ci-dessous :

$$\hat{\mathcal{F}}_i\phi_i = -\frac{1}{2}\sum_j^N \lambda_{ij}\,\phi_j = \epsilon_{ij}\,\phi_j \quad \text{avec} \quad \epsilon_{ij} = -\frac{1}{2}\sum_j^N \lambda_{ij} \tag{105}$$

L'Hamiltonien mono-électronique $\hat{\mathcal{F}}_i$ n'est pas exacte dans le sens où la *corrélation électronique* n'est pas prise en compte dans l'interaction électron-électron. Autrement dit le terme,

$$\underbrace{\sum_{i\neq j}^N V_{ij}}_{\text{terme exacte}} \simeq \underbrace{\sum_j^N \left[\hat{\mathcal{J}}_j(i) - \hat{\mathcal{K}}_j(i)\right]}_{\text{terme approximatif}} \tag{106}$$

Dans le terme approximatif chaque électron i subit le champ moyen des autres électrons j. Ces derniers sont considérés comme une sorte de nuage électronique où chaque électron occupe une position r_j fixe. Or en réalité la dynamique de chaque électron j ($N-1$ électrons restants) influence celle de son voisin. C'est la raison pour laquelle l'énergie électronique totale calculée par l'équation de *Hartree-Fock* est approximative (mais très proche de la valeur exacte). Cet écart énergétique entre les valeurs exacte et calculée est dû à la non prise en compte de la *corrélation électronique* dans l'opérateur de *Fock*. Notons toutefois que la prise en compte de cette relation corrélative fait perdre

à l'opérateur de *Fock* son caractère mono-électronique.

$$\begin{aligned}
\hat{\mathcal{F}}_1\phi_1(1) &= \epsilon_{11}\,\phi_1(1) + \epsilon_{12}\,\phi_2(1) + \cdots + \epsilon_{1n}\,\phi_n(1) \quad i=1\\
\hat{\mathcal{F}}_2\phi_2(1) &= \epsilon_{12}\,\phi_1(1) + \epsilon_{22}\,\phi_2(1) + \cdots + \epsilon_{2n}\,\phi_n(1) \quad i=2\\
&\vdots\\
\hat{\mathcal{F}}_n\phi_n(1) &= \epsilon_{n1}\,\phi_1(1) + \epsilon_{n2}\,\phi_2(1) + \cdots + \epsilon_{nn}\,\phi_n(1) \quad i=n
\end{aligned}$$

Sous forme matricielle,

$$\hat{\mathcal{F}}\begin{pmatrix}\phi_1(1)\\ \phi_2(1)\\ \vdots\\ \\ \phi_n(1)\end{pmatrix} = \begin{pmatrix}\epsilon_{11} & \epsilon_{12} & \cdots & \epsilon_{1n}\\ \epsilon_{12} & \epsilon_{22} & \cdots & \epsilon_{2n}\\ \vdots & \vdots & \vdots & \vdots\\ \\ \epsilon_{n1} & \epsilon_{n2} & \cdots & \epsilon_{nn}\end{pmatrix} \times \begin{pmatrix}\phi_1(1)\\ \phi_2(1)\\ \vdots\\ \\ \phi_n(1)\end{pmatrix} \Leftrightarrow \hat{\mathcal{F}}\,\Phi = \varepsilon\,\Phi \tag{107}$$

Il nous reste maintenant l'opération de *diagonalisation* de la matrice des énergies propres des orbitales. Formellement la matrice ε est diagonalisable, s'il existe une matrice diagonale ε_i telle que :

$$\varepsilon = Q^{-1}\,\varepsilon_i\,Q \tag{108}$$

Ainsi,

$$\hat{\mathcal{F}}\,\Phi = Q^{-1}\,\varepsilon_i\,Q\,\Phi \tag{109}$$

$$\hat{\mathcal{F}}\,\underbrace{Q^{-1}\,\Phi\,Q}_{\psi} = \underbrace{Q^{-1}\,Q}_{\mathbf{1}}\,\varepsilon_i\,\underbrace{Q^{-1}\,\Phi\,Q}_{\psi} \Rightarrow \hat{\mathcal{F}}\,\psi = \varepsilon_i\,\psi \tag{110}$$

D'un point de vue pratique, nous choisissons un jeu de M orbitales $\{\chi_k\}_{k=1,M}$ qui constitueront la base dans laquelle les solutions ψ des équations seront exprimées, soit :

$$\psi_i = \sum_{k=1}^{M} c_{ik}\,\chi_k \tag{111}$$

$$\Rightarrow \sum_{k=1}^{M} c_{ik}\,\hat{\mathcal{F}}\,|\chi_k\rangle = \varepsilon_i \sum_{k=1}^{M} c_{ik}\,|\chi_k\rangle \tag{112}$$

En multipliant par le bras $\langle\chi_l|$ il vient,

$$\Rightarrow \sum_{k=1}^{M} c_{ik}\,\langle\chi_l|\hat{\mathcal{F}}\,|\chi_k\rangle = \varepsilon_i \sum_{k=1}^{M} c_{ik}\,\langle\chi_l|\chi_k\rangle \quad \forall\, l = 1,2,3,\cdots M \tag{113}$$

$$\Rightarrow \sum_{k=1}^{M} c_{ik}\left[\underbrace{\langle\chi_l|\hat{\mathcal{F}}\,|\chi_k\rangle}_{\mathbf{F}} - \varepsilon_i\,\underbrace{\langle\chi_l|\chi_k\rangle}_{\mathbf{S}}\right] = 0 \tag{114}$$

Ce système de M équations à M inconnues n'a de solutions non nulles que si le déterminant :

$$\det |\mathbf{F} - \varepsilon\, \mathbf{S}| = 0 \tag{115}$$

Nous avons M valeurs possibles pour ε qui sont les valeurs propres correspondant à l'énergie des orbitales ψ que nous cherchons. Les matrices $\mathbf{F}$, $\mathbf{S}$ sont de taille $M \times M$ et ε est une matrice diagonale $M \times M$ dont les éléments de la diagonale principale $\epsilon_1, \epsilon_2, \cdots, \epsilon_M$. Le calcul des intégrales de recouvrement $S_{lk} = \langle \chi_l | \chi_k \rangle$ est trivial alors que celui de $F_{lk} = \langle \chi_l | \hat{\mathcal{F}} | \chi_k \rangle$ suppose que l'on connaisse un jeu d'orbitales ψ_j pouvant former un déterminant de *Slater* qui soit fonction propre du système. Or c'est précisément c'est l'ensemble des orbitales ψ_j que nous cherchons en diagonalisant l'opérateur de Fock. Autrement dit, nous avons besoin de connaitre à l'avance les orbitales ψ_j pour construire les équations permettant de calculer ψ_j !. Pour lever ce paradoxe, l'algorithme est initialisé en choisissant un jeu d'orbitales $\psi_j^{(0)}$ combinaisons linéaire des fonctions de base χ_k (orbitales atomiques) qui formeront un déterminant appelé "orbitales d'essai" ou "fonctions d'essai". On les utilise pour calculer les éléments de la matrice $F^{(0)}$ et résoudre les équations séculaires donnant $\psi_j^{(1)}$ et ainsi de suite de façon itérative jusqu'à ce que $\varepsilon^{(n+1)} = \varepsilon^{(n)}$ (il n'y a plus de différences entre deux jeux d'orbitales successives $\psi_j^{(n)} = \psi_j^{(n+1)}$). On dit que la cohérence entre les orbitales de départ $\psi_j^{(0)}$ et les orbitales permettant la convergence $\psi_j^{(n+1)}$. D'où le nom du champ auto-cohérent ou Self Consistent Field (SCF) en anglais.

A. *Signification physique de* ε_i

Le théorème de *Koopmans* fournit une interprétation physique à l'énergie des orbitales moléculaires ε_i. Elle est définie comme l'opposé de l'énergie d'ionisation associée à l'expulsion d'un électron de l'orbitale ϕ_p, soit :

$$\varepsilon_p = E_N - E_{N-1}^p \quad \text{ou} \quad E_{N-1}^p - E_N = -\varepsilon_p \tag{116}$$

Avec E_N et E_{N-1} sont respectivement les énergies des formes neutre et cationique de l'atome ou de la molécule en question. Ce théorème se démontre de la façon suivante : commençons par écrire l'énergie électronique totale calculée par la méthode de *Hartree-Fock* pour la forme neutre (atome ou molécule),

$$E_N = \sum_i^N \hat{h}_i + \sum_{i=1}^N \sum_{i \neq j}^N \left[\hat{\mathcal{J}}_{ij} - \hat{\mathcal{K}}_{ij} \right] \tag{117}$$

Isolons désormais un électron, noté p :

$$E_N = \sum_i^{N-1} \hat{h}_i + \sum_{i=1}^{N-1} \sum_{i \neq j}^{N-1} \left[\hat{\mathcal{J}}_{ij} - \hat{\mathcal{K}}_{ij} \right] + \hat{h}_p + \sum_i^{N-1} \left[\hat{\mathcal{J}}_{ip} - \hat{\mathcal{K}}_{ip} \right] \tag{118}$$

Écrivons l'énergie électronique totale calculée par la méthode de *Hartree-Fock* pour la forme cationique :

$$E_{N-1}^p = \sum_i^{N-1} \hat{h}_i + \sum_{i=1}^{N-1} \sum_{i \neq j}^{N-1} \left[\hat{\mathcal{J}}_{ij} - \hat{\mathcal{K}}_{ij} \right] \tag{119}$$

Par soustraction des équations (119) et (117), il vient :

$$E_{N-1}^p - E_N = -\left[\underbrace{\hat{h}_p + \sum_i^{N-1} \left[\hat{\mathcal{J}}_{ip} - \hat{\mathcal{K}}_{ip} \right]}_{\varepsilon_p} \right] \tag{120}$$

$$\Rightarrow \quad E_{N-1}^{p} - E_N = -\varepsilon_p \tag{121}$$

Lors de l'expulsion d'un électron d'un atome ou d'une molécule pour former un cation, on augmente le potentiel attractif ressenti par les autres électrons. A l'inverse lors de l'addition d'un électron à une molécule pour former un anion, on augmente le potentiel répulsif ressenti par les autres électrons. Les niveaux d'énergie du cation seront plus bas que ceux de la molécule neutre correspondante. D'un autre coté, les niveaux d'énergie de l'anion seront plus élevés que ceux de la molécule neutre correspondante. En outre, la taille des orbitales diminuera dans l'ordre suivant : anion, neutre, cation.

V. Formalisme de la DFT

Le terme exacte $\sum_{i\neq j} \dfrac{e^2}{r_{ij}}$ traduit mathématiquement la répulsion électrostatique entre deux électrons (i) et (j) séparés par une distance r_{ij}. Il faut noter que la distance inter-électronique r_{ij} varie à chaque instant t car les électrons tendent à s'éviter au cours de leur processus dynamique, c'est la *corrélation électronique*. La répulsion exacte n'est pas calculable car nous n'avons pas accès à la valeur de r_{ij} à chaque instant. Selon la théorie de *Hartree-Fock*, la répulsion électrostatique exacte entre deux électrons est approchée par un terme moyen :

$$\sum_i \sum_{i\neq j} [2\,\mathcal{J}_{ij} - \mathcal{K}_{ij}] \tag{122}$$

Ce dernier terme est approximatif car l'énergie dûe à la corrélation électronique (E_c) n'est pas prise en compte. C'est pour cette raison que l'énergie totale d'une molécule (E_m^T) calculée par l'équation de Hartree-Fock est systématiquement supérieure à son énergie exacte (E_0) soit :

$$E_c = E_0 - E_m^T < 0 \tag{123}$$

Ainsi pour compléter la théorie de *Hartree-Fock*, plusieurs modèles ont été proposés afin de tenir compte de la *corrélation électronique*. Dans cette section, nous aborderons l'une de ces méthodes emblématique qui est la théorie de la densité de la fonctionnelle ou DFT (Density Functional theory, en anglais).

a) ***Théorème de Hohenberg-Kohn****: Ce théorème se démontre comme suit, considérons deux potentiels* $\hat{\mathcal{U}}_{eN}^{A} \neq \hat{\mathcal{U}}_{eN}^{B}$*. Les deux Hamiltoniens des deux systèmes s'écrivent :*

$$\hat{\mathcal{H}}_1 = \hat{\mathcal{T}}_e + \hat{\mathcal{U}}_{ee} + \hat{\mathcal{U}}_{eN}^{A} \quad \textit{tel que} \quad \hat{\mathcal{H}}_1\,[\psi_1] = E_1\,\psi_1 \tag{124}$$

$$\hat{\mathcal{H}}_2 = \hat{\mathcal{T}}_e + \hat{\mathcal{U}}_{ee} + \hat{\mathcal{U}}_{eN}^{B} \quad \textit{tel que} \quad \hat{\mathcal{H}}_2\,[\psi_2] = E_2\,\psi_2 \tag{125}$$

Où ψ_1 *et* ψ_2 *sont respectivement les fonctions d'onde de l'état fondamental des systèmes (1) et (2). Soit* $\rho(r)$ *la densité électronique correspondant aux deux systèmes. Autrement dit,* $|\psi_1|^2 = \rho(r)$ *et* $|\psi_2|^2 = \rho(r)$*. Nous en déduisons ce qui suit :*

$$\underbrace{\langle \psi_1|\hat{\mathcal{H}}_1|\psi_1\rangle}_{E_1} < \langle \psi_2|\hat{\mathcal{H}}_1|\psi_2\rangle$$

$$E_1 < \langle \psi_2|\hat{\mathcal{H}}_2 + \hat{\mathcal{U}}_{eN}^{A} - \hat{\mathcal{U}}_{eN}^{B}|\psi_2\rangle$$

$$E_1 < \underbrace{\langle \psi_2|\hat{\mathcal{H}}_2|\psi_2\rangle}_{E_2} + \underbrace{\langle \psi_2|\hat{\mathcal{U}}_{eN}^{A} - \hat{\mathcal{U}}_{eN}^{B}|\psi_2\rangle}_{\int [\hat{\mathcal{U}}_{eN}^{A} - \hat{\mathcal{U}}_{eN}^{B}]\,\rho(r)\,dr}$$

$$\Rightarrow \quad E_1 < E_2 + \int [\hat{\mathcal{U}}_{eN}^{A} - \hat{\mathcal{U}}_{eN}^{B}]\, \rho(r)\, dr \tag{126}$$

D'un autre côté nous pouvons écrire :

$$\underbrace{\langle \psi_2 | \hat{\mathcal{H}}_2 | \psi_2 \rangle}_{E_2} < \langle \psi_1 | \hat{\mathcal{H}}_2 | \psi_1 \rangle$$

$$E_2 < \langle \psi_1 | \hat{\mathcal{H}}_1 + \hat{\mathcal{U}}_{eN}^{B} - \hat{\mathcal{U}}_{eN}^{A} | \psi_1 \rangle$$

$$E_2 < \underbrace{\langle \psi_1 | \hat{\mathcal{H}}_1 | \psi_1 \rangle}_{E_1} + \underbrace{\langle \psi_1 | \hat{\mathcal{U}}_{eN}^{B} - \hat{\mathcal{U}}_{eN}^{A} | \psi_1 \rangle}_{\int [\hat{\mathcal{U}}_{eN}^{B} - \hat{\mathcal{U}}_{eN}^{A}]\, \rho(r)\, dr}$$

$$\Rightarrow \quad E_2 < E_1 + \int [\hat{\mathcal{U}}_{eN}^{B} - \hat{\mathcal{U}}_{eN}^{A}]\, \rho(r)\, dr \tag{127}$$

Ou encore sous la forme équivalente :

$$\Rightarrow \quad E_2 < E_1 - \int [\hat{\mathcal{U}}_{eN}^{A} - \hat{\mathcal{U}}_{eN}^{B}]\, \rho(r)\, dr \tag{128}$$

En additionnant les inéquations (126) *et* (128)*, nous obtenons,*

$$\Rightarrow \quad E_1 + E_2 < E_2 + E_1 \quad ! \tag{129}$$

La contradiction exprimée par l'inéquation (129) *est levée en considérant que la densité électronique* $\rho(r)$ *est unique pour un potentiel externe donné.*

A. *Premier niveau d'approximation*

L'Hamiltonien (énergie totale) global exacte d'un atome ou d'une molécule est exprimé sous la forme :

$$\hat{\mathcal{H}} = \underbrace{-\sum_{i=1}^{N} \frac{\hbar^2\, \nabla_i^2}{2\, m_e}}_{\hat{T}_e} \underbrace{-\sum_{h=1}^{M} \frac{\hbar^2\, \nabla_h^2}{2\, m_N}}_{\hat{T}_N} \underbrace{-\sum_{h=1}^{M} \sum_{i=1}^{N} \frac{Z_h\, e^2\, \nabla_h^2}{4\, \pi\, \epsilon_0\, r_{ih}}}_{\hat{U}_{eN}} + \underbrace{\frac{1}{2} \sum_{i=j}^{N} \frac{e^2}{4\, \pi\, \epsilon_0\, r_{ij}}}_{\hat{U}_{ee}} + \underbrace{\frac{1}{2} \sum_{h=k}^{M} \frac{Z_h\, Z_k\, e^2}{4\, \pi\, \epsilon_0\, r_{hk}}}_{\hat{U}_{NN}} \tag{130}$$

Où les deux premiers termes ($\hat{T}_e$ et $\hat{T}_N$) désignent respectivement les opérateurs associés aux énergies cinétiques des électrons et des noyaux. Les termes $\hat{U}_{eN}$, $\hat{U}_{ee}$ et $\hat{U}_{NN}$ désignent respectivement les opérateurs associés aux énergies potentielles d'interaction électron-noyau, électron-électron et noyau-noyau. Un tel système est impossible à résoudre compte tenu du nombre élevé de variables ($3\,N + 3\,M$) et d'interactions mises en jeu dans ce type de problèmes. Un premier niveau de simplification consiste à appliquer l'approximation de *Born-Oppenheimer*. Cette approximation est basée sur l'idée que la masse des noyaux est environ 1840 fois plus grande que celle des électrons, donc nous pouvons considérer que le mouvement relatif des noyaux est suffisamment long par rapport à celui des électrons ce qui justifie de figer la position des noyaux. Cette première approximation permet de ramener un système à plusieurs électrons et noyaux de départ à un système poly-électroniques

seulement. Comme conséquence de cette approximation le terme $\hat{T}_N$ s'annule et le terme $\hat{U}_{NN}$ est réduit à une constante. L'Hamiltonien (130) est ainsi réduit à la forme :

$$\hat{\mathcal{H}} = \underbrace{-\sum_{i=1}^{N} \frac{\hbar^2 \nabla_i^2}{2\, m_e}}_{\hat{T}_e} \underbrace{-\sum_{h=1}^{M}\sum_{i=1}^{N} \frac{Z_h\, e^2\, \nabla_h^2}{4\,\pi\,\epsilon_0\, r_{ih}}}_{\hat{U}_{eN}\equiv\hat{U}_{ext}} + \underbrace{\frac{1}{2}\sum_{i=j}^{N} \frac{e^2}{4\,\pi\,\epsilon_0\, r_{ij}}}_{\hat{U}_{ee}} \tag{131}$$

L'Hamiltonien (131) décrit des systèmes à électrons seuls en interaction mutuelle et en déplacement dans le potentiel externe des noyaux $\hat{U}_{ext}$. En adoptant le système des unités atomiques [3], l'expression de l'Hamiltonien (131) devient :

$$\hat{\mathcal{H}} = \underbrace{-\sum_{i=1}^{N} \frac{\nabla_i^2}{2}}_{\hat{T}_e} \underbrace{-\sum_{h=1}^{M}\sum_{i=1}^{N} \frac{Z_h\, \nabla_h^2}{r_{ih}}}_{\hat{U}_{eN}\equiv\hat{U}_{ext}} + \underbrace{\frac{1}{2}\sum_{i=j}^{N} \frac{1}{r_{ij}}}_{\hat{U}_{ee}} \tag{132}$$

Le terme $\hat{U}_{ext}$ traduit le potentiel crée par l'ensemble des noyaux et ressenti par chaque électron du système quantique étudié. Afin de révéler cet aspect, réécrivons (132) sous la forme développée :

$$\hat{\mathcal{H}} = \sum_{i=1}^{N} \underbrace{\left\{ -\frac{\nabla_i^2}{2} + \hat{u}_{ext}(r_i) \right\}}_{\hat{h}_i} + \hat{U}_{ee} \quad \text{avec} \quad \hat{U}_{ext} = \sum_{i=1}^{N} \hat{u}_{ext}(r_i) \tag{133}$$

$$\hat{\mathcal{H}} = \sum_{i=1}^{N} \hat{h}_i + \hat{U}_{ee} \tag{134}$$

Le théorème de *Hohenberg-Kohn* démontré au début de la section, stipule que :

$$E[\hat{u}_{ext}(\rho)] \equiv E[\rho] \tag{135}$$

Selon le théorème de *Hohenberg-Kohn*, puisque la densité électronique $\rho(r)$, détermine le nombre d'électron total N et le potentiel externe $\hat{U}_{ext}$, elle devrait également déterminer toutes les propriétés de l'état fondamental y compris l'énergie cinétique des électrons $\hat{T}_e$ et l'énergie de l'interaction entre les électrons $\hat{U}_{ee}$. Ainsi, l'énergie totale de l'état fondamental est une fonction de la densité avec les composantes suivantes :

$$E[\rho] = \left\langle \psi | \hat{\mathcal{H}} | \psi \right\rangle = \underbrace{\left\langle \psi | \hat{T}_e + \hat{U}_{ee} | \psi \right\rangle}_{F_{HK}[\rho]} + \left\langle \psi | \hat{U}_{ext} | \psi \right\rangle \tag{136}$$

$$= F_{HK}[\rho] + \underbrace{\int \hat{u}_{ext}[\rho]\ \rho(r)\ dr}_{\hat{U}_{ext}[\rho]} \tag{137}$$

$$= F_{HK}[\rho] + \hat{U}_{ext}[\rho] \tag{138}$$

L'expression résultante est celle d'un Hamiltonien décrivant un système polyélectroniques en interaction mutuelle et en déplacement dans la potentiel externe $\hat{U}_{ext}$ généré par les noyaux. La fonctionnelle $F_{HK}[\rho]$ est *universelle* car dépendante exclusivement des électrons et totalement indépendante de la nature du système quantique (atomes, molécules, solide ... etc.) étudié. Cette information est

3. Le rayon de Bohr $a_0 = 5.2911\, nm$ est pris comme unité de base des longueurs et le *Rydberg* ou le *Hartree* comme celles des énergies sachant que $1\, Ryd = 13.60\, eV$ et $1\, Ha = 2\, Ryd = 27.21\, eV$.

contenue dans l'expression du potentiel externe.

Hohenberg-Kohn contournent le problème de la résolution de l'équation de *Schrödinger* à plusieurs électrons en formulant une fonctionnelle $F_{HK}[\rho]$ sous l'influence d'un potentiel externe donné. Désormais toute la problématique consiste à déterminer la formule de $F_{HK}[\rho]$. Il n'existe pas de formules analytiques pour les fonctionnelles de la densité relatives à l'énergie cinétique $T_e[\rho]$ et à l'interaction électron-électron $U_{ee}[\rho]$. Cette problématique a été résolue par *Kohn* et *Sham* dont le principe est donné dans la prochaine section.

B. *Méthode de Kohn-Sham*

Afin de contourner cette difficulté, *Kohn-Sham* ont imaginé un système fictif sans interaction ($T_f[\rho] \neq 0$ et $U_f[\rho] = 0$) ayant la même densité électronique que le système quantique étudié (ou système réel). A partir de cette idée, la fonctionnelle de *Hohenberg-Kohn* est réécrite sous la forme :

$$
\begin{aligned}
F_{HK}[\rho] &= T_e[\rho] + U_{ee}[\rho] && (139)\\
&= T_e[\rho] + U_{ee}[\rho] + \{T_f[\rho] - T_f[\rho]\} && (140)\\
&= T_f[\rho] + U_{ee}[\rho] + \underbrace{T_e[\rho] - T_f[\rho]}_{E_c[\rho]} && (141)\\
&= T_f[\rho] + U_{ee}[\rho] + E_c[\rho] + \{U_H[\rho] - U_H[\rho]\} && (142)\\
&= T_f[\rho] + U_H[\rho] + E_c[\rho] + \underbrace{\{U_{ee}[\rho] - U_H[\rho]\}}_{E_x[\rho]} && (143)\\
&= T_f[\rho] + U_H[\rho] + \underbrace{E_c[\rho] + E_x[\rho]}_{E_{xc}[\rho]} && (144)\\
&= T_f[\rho] + U_H[\rho] + E_{xc}[\rho] && (145)\\
& && (146)
\end{aligned}
$$

Où le terme $E_{xc}[\rho]$ est la fonctionnelle *d'échange-corrélation*. Cette fonctionnelle contient toutes les interactions électron-électron non classiques. Elle s'écrit comme la somme d'une fonctionnelle d'échange $E_x[\rho]$ (interactions de même spin) et d'une fonctionnelle de corrélation $E_c[\rho]$ (interactions entre spins différents). Le lien avec le système étudié (avec interaction) se fait en définissant une énergie d'échange-corrélation par :

$$
E_{xc}[\rho] = \underbrace{\{T_e[\rho] - T_f[\rho]\}}_{E_c[\rho]} + \underbrace{\{U_{ee}[\rho] - U_H[\rho]\}}_{E_x[\rho]} \qquad (147)
$$

Le terme U_H est le potentiel électrostatique du système fictif qui considère le nuage électronique figé sur des positions fixes (ce qui néglige la corrélation électronique). L'étude de la fonctionnelle $E_{xc}[\rho]$ sera détaillée dans les prochaines sections. La fonctionnelle de l'énergie totale correspondant à ce système fictif (sans interactions mutuelle) et en déplacement dans le potentiel externe des noyaux est égale au système réel (avec interaction) selon :

$$
\begin{aligned}
E[\rho] &= F_{HK}[\rho] + \underbrace{\int \hat{u}_{ext}(r)\,\rho(r)\,dr}_{\hat{U}_{ext}[\rho]} && (148)\\
&= T_f[\rho] + U_H[\rho] + E_{xc}[\rho] + \hat{U}_{ext}[\rho] && (149)
\end{aligned}
$$

Parmi les nombreuses densités électroniques possibles, celle correspondant à l'état fondamental $\rho_0(r)$ est obtenue en appliquant le principe variationnel minimisant l'énergie totale $E[\rho]$ pour un potentiel externe bien défini :

$$E_0 = E_v[\rho_0] = \min_{\rho} E_v[\rho] \tag{150}$$

L'indice v dans l'équation (150) souligne que le principe variationnel s'applique uniquement à des densités électroniques *v-représentables*. Cela signifie l'existence d'une correspondance entre densité électronique et le potentiel externe au travers du premier théorème de *Hohenberg-Kohn*. Néanmoins, les conditions pour qu'une densité électronique soit *v-représentable* sont inconnues. Par voie de conséquence, l'utilisation du principe variationnel (150) est impossible, puisqu'il peut conduire sans la contrainte de *v-représentable* à des densités dénouées de sens physique. Afin de surmonter cette impossibilité, nous imposons à la densité électronique d'être *n-représentable* seulement. Cela signifie qu'on impose à la densité électronique d'être positive ou nulle en tout point de l'espace.

Tenant compte de cette contrainte, parmi l'infinité de fonction d'onde ψ qui s'intègrent en $\rho_0(r)$, la fonction d'onde de l'état fondamental ψ_0 est celle minimisant la fonctionnelle de *Hohenberg-Kohn*. Il en résulte que le principe variationnel (150) peut être réécrit en substituant la contrainte de *v-représentabilité* par celle de la *n-représentabilité* selon :

$$\text{contrainte} \quad \longrightarrow \int \rho(r)\, dr - N = 0 \tag{151}$$

Il s'agit d'un problème d'optimisation (minimisation de l'énergie totale) avec contrainte. Le *Lagrangien* correspondant à ce système s'écrit :

$$E[\rho] - \lambda \left[\int \rho(r)\, dr - N \right] \tag{152}$$

Avec λ étant le *multiplicateur de Lagrange*. Plus ce coefficient est grand plus le poids de la contrainte en question est important. La minimisation de (152) donne :

$$\delta \left\{ E[\rho] - \lambda \left[\int \rho(r)\, dr - N \right] \right\} = 0 \tag{153}$$

Où N étant le nombre total d'électrons, c'est un paramètre constant. Il en résulte :

$$\delta E[\rho] - \lambda\, \delta \left\{ \int \rho(r)\, dr \right\} = 0 \tag{154}$$

Rappelons que la différentielle d'une fonctionnelle s'écrit :

$$F[f + \delta f] - F[f] = \delta F = \int \frac{\delta F}{\delta f(x)}\, \delta f(x)\, dx \tag{155}$$

Tenant compte de (155), l'équation (154) s'écrira :

$$\int \frac{\delta E[\rho]}{\delta \rho}\, \delta\rho\, dr - \lambda \int \delta\rho(r)\, dr = 0 \tag{156}$$

Les deux intégrales de (156) dépendent de la même variable et elles ont les mêmes bornes d'intégration. Nous pouvons les écrire sous forme d'une seule intégrale :

$$\int \left[\frac{\delta E[\rho]}{\delta \rho}\, \delta\rho - \lambda\, \delta\rho(r) \right] dr = 0 \tag{157}$$

L'égalité exprimée par l'équation (157) est vérifiée si :

$$\lambda = \frac{\delta E[\rho]}{\delta \rho} \tag{158}$$

Tenant compte de l'équation (138), il vient :

$$\lambda = \frac{\delta E[\rho]}{\delta \rho} = \hat{U}_{ext}(r) + \frac{\partial F_{HK}[\rho(r)]}{\partial \rho(r)} \tag{159}$$

Dans le cadre de la *DFT*, le *multiplicateur de Lagrange* est définie comme un *potentiel chimique*. Ce descripteur est très important car il est relié, entre autre, à la réactivité chimique. Ce descripteur est largement étudié dans le cadre de la *DFT* conceptuelle. Tenant compte de (146), il vient :

$$\Rightarrow \quad \frac{\delta E[\rho]}{\delta \rho} = \frac{\partial T_f[\rho(r)]}{\partial \rho(r)} + \hat{U}_{ext}(r) + \hat{U}_H(r) + \underbrace{\frac{\partial E_{xc}[\rho(r)]}{\partial \rho(r)}}_{\hat{U}_{xc}(r)} \tag{160}$$

Ou de façon équivalente :

$$\Rightarrow \quad \frac{\delta E[\rho]}{\delta \rho} = \frac{\partial T_f[\rho(r)]}{\partial \rho(r)} + \hat{U}_{eff}(r) \tag{161}$$

En combinant les équations (160) et (160) il vient :

$$\hat{U}_{eff}(r) = \hat{U}_{ext}(r) + \hat{U}_H(r) + \hat{U}_{xc}(r) \tag{162}$$

Chaque électron du système fictif [4] (sans interaction mutuelle entre électrons) ressent individuellement un potentiel effectif de *Kohn-Sham* $\hat{u}_{eff}(r_i)$ qui est la somme des termes électrostatique (U_H), potentiel externe $\hat{U}_{ext}$ et du potentiel d'échange-corrélation U_{xc} selon l'Hamiltonien mono-électronique suivant :

$$\hat{h}^{KS} = \frac{-\nabla_i^2}{2} + \hat{u}_{eff}(r_i) \quad \text{avec} \quad \hat{U}_{eff} \equiv \sum_i^N \hat{u}_{eff}(r_i) \tag{163}$$

C'est l'opérateur mono-électronique de *Kohn-Sham*. Nous obtenons les fameuses équations de *Kohn-Sham* :

$$\Rightarrow \quad \left[\frac{-\nabla_i^2}{2} + \hat{u}_{eff}(r_i)\right] \theta_i^{KS} = \varepsilon_i^{KS}\, \theta_i^{KS} \quad \text{avec} \quad \rho(r) = \sum_i^N |\theta_i^{KS}|^2 \tag{164}$$

Où θ_i^{KS} sont appelées les orbitales de *Kohn-Sham*. Cette équation mono-électronique s'écrit aussi sous la forme :

$$\hat{h}^{KS}(1)\theta^{KS}(1) = \varepsilon_i^{KS}\, \theta_i^{KS}(1) \tag{165}$$

Selon les calculs des propriétés électroniques reposant sur la DFT, la minimisation de l'énergie totale du système se fait donc en résolvant de manière itérative ou auto-cohérente (algorithme SCF) les équations de *Kohn-Sham* de la même manière que dans le cas de la méthode *Hartree-Fock* que nous avons déjà établi dans le section précédente Eq.(115). Ce sont des équations de type Schrödinger, dont les solutions sont des orbitales mono-électroniques. Dans cette configuration, $\hat{u}_{eff}(r_i)$ est ajusté afin que θ_i^{KS} minimise l'Hamiltonien du système réel. Les orbitales de *Kohn-Sham* doivent reproduire la densité électronique exacte de l'état fondamental du système étudié.

4. Le système fictif d'électrons sans interaction qui a la même densité électronique que le système réel avec interaction

C. *Fonctionnelle échange-corrélation*

Du fait du principe de *Pauli*, il faut tenir compte de l'anti-symmétrisation de la fonction d'onde et donc du fait que les électrons de même spin se repoussent fortement. Le terme E_x doit contenir ce facteur, mais il est difficile d'expliciter une fonctionnelle en $\rho(r)$ ayant cette fonction. De même, il manque la corrélation électronique qui doit être contenue dans E_c qui va plutôt reproduire la répulsion entre électrons de spins différents. Ces deux derniers termes sont difficiles à obtenir et jusqu'à présent seules des expressions approchées existent. La première approximation est nommée LDA (Local Density Approximation). Cette approximation [5] considère que pour les systèmes inhomogènes dont la densité varie lentement, le système semble localement avoir une densité constante. Par conséquent, le potentiel externe sera également constant et le système est similaire au gaz d'électrons homogène. Utilisons ce principe pour construire une approximation locale de $E_{xc}[\rho]$

$$E_{xc}[\rho] = \int \rho(r)\,\epsilon_{xc}[\rho(r)]\,dr = \int \rho(r)\,\{\epsilon_x[\rho(r)] + \epsilon_c[\rho(r)]\}\,dr \tag{166}$$

Où $\epsilon_{xc}[\rho(r)]$ désigne la densité d'énergie (c'est-à-dire une énergie par électron) au point r dans l'espace, qui ne dépend que de la densité en ce point. Cette énergie par particule est pondérée avec la probabilité $\rho(r)$ qu'il y ait un électron à cette position. En développant la dernière équation il vient :

$$\Rightarrow \quad E_{xc}[\rho] = \underbrace{\int \rho(r)\,\epsilon_x[\rho(r)]\,dr}_{E_x} + \underbrace{\int \rho(r)\,\epsilon_c[\rho(r)]\,dr}_{E_c} \tag{167}$$

La contribution de l'échange pour gaz d'électrons homogène est connue de manière analytique :

$$E_x[\rho] = \int \rho(r)\,\frac{3}{4}\left[\frac{3}{\pi}\,\rho(r)\right]^{1/3} dr \tag{168}$$

$$\Rightarrow E_x[\rho] = \frac{3}{4}\left[\frac{3}{\pi}\right]^{1/3} \int \rho(r)^{4/3}\,dr \tag{169}$$

Cette dernière équation stipule qu'il suffit de connaître la densité électronique en un point donné de l'espace et l'énergie d'échange est alors simplement l'intégrale sur la densité à la puissance 4/3. La contribution d'échange au potentiel $\hat{U}_{xc}(r)$ peut être calculée directement suivant :

$$U_x(r) = \frac{\delta E_x[\rho]}{\delta \rho} = \epsilon_x[\rho(r)] + \rho(r)\,\frac{\partial \epsilon_x[\rho(r)]}{\partial \rho(r)} \tag{170}$$

$$\Rightarrow \quad U_x(r) = \frac{3}{4}\left[\frac{3}{\pi}\right]^{1/3} \rho(r)^{1/3} + \rho(r)\,\frac{3}{4}\left[\frac{3}{\pi}\right]^{1/3} \frac{1}{\rho^{2/3}} \tag{171}$$

$$\Rightarrow \quad U_x(r) = \left[\frac{3}{\pi}\right]^{1/3} \rho(r)^{1/3} \tag{172}$$

C'est la forme la plus simple de la contribution de l'échange. Il existe une multitude d'expressions analytiques, nous en donnerons uniquement deux expressions. Nous commençons par la fonctionnelle d'échange de *Dirac-Slater* qui est donnée par :

$$E_x^{LSD}[\rho_\alpha, \rho_\beta] = \int \rho(r)\,\epsilon_x[\rho(r), \xi]\,dr \tag{173}$$

5. L'approximation LDA a été développée originalement pour les métaux en supposant que la densité est constante dans le solide. LDA tend à sous-estimer les énergies d'échange par près de 10%, et à sur-estimer la corrélation par plus de $\times 2$.

$$\epsilon_x[\rho(r),\xi] = \epsilon_x^0[\rho(r)] + \left\{\epsilon_x^1[\rho(r)] - \epsilon_x^0[\rho(r)]\right\}\, f(\xi) \tag{174}$$

$$\epsilon_x^0[\rho(r)] = \epsilon_x[\rho(r),0] = \frac{3}{4}\left[\frac{3}{\pi}\right]^{1/3}\rho^{1/3} \quad \text{et} \quad \epsilon_x^1[\rho(r)] = \epsilon_x[\rho(r),1] = 2^{1/3}\,\frac{3}{4}\left[\frac{3}{\pi}\right]^{1/3}\rho^{1/3} \tag{175}$$

$$f(\xi) = \frac{(1+\xi)^{4/3} + (1-\xi)^{4/3} - 2}{2\,(2^{1/3}-1)} \quad \text{et} \quad \xi = \frac{\rho_\alpha - \rho_\beta}{\rho_\alpha + \rho_\beta} \tag{176}$$

La deuxième fonctionnelle d'échange que nous donnons est celle de *Becke*, définie par :

$$E_x^{BEC}[\rho_\alpha,\rho_\beta] = E_x^{LSD}[\rho_\alpha,\rho_\beta] - \sum_\sigma^{\alpha,\beta}\int \rho_\sigma(r)\,\epsilon_x[\rho_\sigma(r),\xi_\sigma]\,dr \tag{177}$$

$$\epsilon_x[\rho_\sigma(r),\xi_\sigma] = \rho_\sigma(r)^{1/3} \times \frac{0.0042\,\xi_\sigma^2}{1 + 0.0252\,\sinh^{-1}(\xi_\sigma)} \quad \text{avec} \quad \xi_\sigma = \frac{|\nabla\rho_\sigma(r)|}{\rho_\sigma^{4/3}} \tag{178}$$

Cette fonctionnelle constitue une correction de celle de *Dirac-Slater* par l'introduction du gradient de la densité électronique $\nabla\rho_\sigma(r)$. Ce gradient corrige les insuffisances de l'approximation de la densité locale (LDA). En effet, le gradient $\nabla\rho_\sigma(r)$ prend en compte les inhomogénéités locales de la densité électronique. Dans la littérature, cette correction porte le nom de *l'approximation du gradient généralisé ou GGA*. Dans le cadre de cette approximation, la fonctionnelle d'échange-corrélation est donnée par la forme générale :

$$E_x^{GGA}[\rho(r),\nabla\rho_\sigma(r)] = \int f(\rho(r),\nabla\rho_\sigma(r))\,dr \tag{179}$$

En fonction de la formule de $f(\rho(r),\nabla\rho_\sigma(r))$, différentes expressions analytiques ont été développées. Nous donnons ci-dessous, une formule analytique décrivant la fonctionnelle de la corrélation E_c dans le cadre de l'approximation *GGA*. Nous donnerons la plus utilisée pour les molécules organiques qui celle de *Lee, Yang et Parr* :

$$E_c^{LYP}[\rho_\alpha,\rho_\beta] = -a\int \frac{\gamma(r)}{1 + d\,\rho(r)^{-1/3}} \times$$

$$\left\{\rho(r) + 2\,b\,\rho(r)^{-5/3}\left[c_x\,\rho_\beta(r)^{8/3} - t_w(r) + \frac{1}{9}\left(\rho_\alpha(r)t_w^a + \rho_\beta(r)t_w^b\right) + \frac{1}{18}\left(\rho_\alpha(r)\nabla^2\rho_\alpha(r)\right)\right]e^{-c\,\rho(r)^{-1/3}}\right\} \tag{180}$$

Où

$$\gamma(r) = 2\left[1 - \frac{\rho_\alpha^2(r) + \rho_\beta^2(r)}{\rho(r)^2}\right] \quad \text{et} \quad t_w(r) = \frac{1}{8}\,\frac{|\nabla\rho(r)|^2}{\rho(r)} - \frac{1}{8}\,\nabla^2\rho(r)$$

Les constantes valent : $c_x = 2^{2/3}\,\dfrac{3}{10}\,(3\pi^2)^{2/3}$, $a = 0.049$, $b = 0.132$, $c = 0.253$ et $d = 0.349$. Avec un niveau de précision similaire, les exigences de calcul avec les méthodes *DFT* sont bien moindres qu'avec les méthodes ab initio. C'est pourquoi les méthodes *DFT* sont largement utilisées dans le calcul des propriétés électroniques des molécules organiques. Par ailleurs, comme nous l'avons mentionné précédemment, il existe de nombreuses approximations de la fonctionnelle d'échange-corrélation. Ces dernières sont désignées dans le logiciel *Gaussian* par les initiales des auteurs dont la première partie désigne la partie échange et la deuxième la celle de la corrélation.

VI. Annexe : Rappels mathématiques

La définition formelle d'un espace vectoriel sur un corps $\mathbb{K}$ (ou un $\mathbb{K}$-espace vectoriel) est un ensemble non vide $\mathcal{E}$ muni d'une loi de composition interne, notée $(+)$ qui est l'addition vectorielle :

$$\begin{aligned} \mathcal{E} \times \mathcal{E} &\longmapsto \mathcal{E} \\ (v_1, v_2) &\longmapsto v_1 + v_2 \end{aligned}$$

La somme de deux éléments $(v_1, v_2 \in \mathcal{E}^2)$ de l'espace vectoriel $\mathcal{E}$ est aussi un élément de l'espace vectoriel $(v_1 + v_2 \in \mathcal{E})$. Le mot *interne* signifie que l'addition vectorielle est réalisée uniquement sur les éléments appartenant à l'espace vectoriel lui-même. L'espace vectoriel $\mathcal{E}$ est également muni d'une loi de composition externe, noté $(\cdot)$ soit $\forall \lambda \in \mathbb{K}, \forall v \in \mathcal{E}$:

$$\begin{aligned} \mathbb{K} \times \mathcal{E} &\longmapsto \mathcal{E} \\ (\lambda, v) &\longmapsto \lambda \cdot v \end{aligned}$$

On appelle les éléments de $\mathcal{E}$ des *vecteurs* et les éléments de $\mathbb{K}$ des *scalaires*. La loi de composition *externe* sur l'espace vectoriel $\mathcal{E}$ est la multiplication d'un vecteur par un scalaire λ.

i. Axiomes à la loi interne

- Commutativité : $\forall v_1, v_2 \in \mathcal{E}, v_1 + v_2 = v_2 + v_1$.
- Associativité : $\forall v_1, v_2 \in \mathcal{E}, v_1 + (v_2 + v_3) = (v_1 + v_2) + v_3$.
- Élément neutre : $\exists! 0_{\mathcal{E}} \in \mathcal{E}, \forall v \in \mathcal{E}, v + 0_{\mathcal{E}} = v$. Avec $0_{\mathcal{E}}$ est le vecteur nul.
- Symétrique : $\exists! v^{'} \in \mathcal{E}, \forall v \in \mathcal{E}, v^{'} + v = 0_{\mathcal{E}} \Rightarrow v^{'} = -v$.

ii. Axiomes à la loi externe

- Élément neutre : $\exists! \lambda \in \mathbb{K}, \forall v \in \mathcal{E}, \lambda \cdot v = v \Rightarrow \lambda = 1$.
- Distributivité par rapport à l'addition vectorielle : $\forall \lambda \in \mathbb{K}, \forall v_1, v_2 \in \mathcal{E}, \lambda\,(v_1 + v_2) = \lambda\, v_1 + \lambda\, v_2$.
- Distributivité par rapport à l'addition des scalaires : $\forall \lambda_1, \lambda_2 \in \mathbb{K}, \forall v \in \mathcal{E}, (\lambda_1 + \lambda_2)\, v = \lambda_1\, v + \lambda_2\, v$.

Le scalaire λ engendre un accroissement ou un rétrécissement d'un vecteur. La loi interne $(+)$ et la loi externe $(\cdot)$ doivent satisfaire ces axiomes pour que $(\mathcal{E}, +, \cdot)$ soit un espace vectoriel sur le corps $\mathbb{K}$.

iii. Combinaison linéaire

$\forall n \in \mathbb{N}^*$, soit $\{v_n\}_{n\in\mathbb{N}^*}$ des vecteurs d'un espace vectoriel $\mathcal{E}$. Le vecteur :

$$v = \sum_{i=1}^{n} \lambda_i\, v_i \tag{181}$$

est appelé combinaison linéaire des vecteurs $\{v_n\}_{n\in\mathbb{N}^*}$. Les scalaires $\{\lambda_n\}_{n\in\mathbb{N}^*}$ sont les coefficients de la combinaison linéaire.

Exemple 1 : dans l'espace vectoriel sur le corps $\mathbb{R}^3$ $(\mathbb{K} = \mathbb{R})$, le vecteur $(3, 3, 1)$ est combinaison linéaire des vecteurs $(1, 1, 0)$ et $(1, 1, 1)$:

$$(3, 3, 1) = 2\,(1, 1, 0) + (1, 1, 1) \quad \text{avec} \quad \lambda_1 = 2, \lambda_2 = 1$$

Exemple 2 : considérons le $\mathbb{R}^3$-espace vectoriel, $v_1, v_2 \in \mathcal{E}$ telle que $v_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$ et $v_2 = \begin{pmatrix} 6 \\ 4 \\ 2 \end{pmatrix}$

– Montrer que $v = \begin{pmatrix} 9 \\ 2 \\ 7 \end{pmatrix}$ est une combinaison linéaire de v_1 et v_2.

Cherchons $\lambda_1, \lambda_2 \in \mathbb{R}$ tel que :

$$\begin{pmatrix} 9 \\ 2 \\ 7 \end{pmatrix} = \lambda_1 \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} + \lambda_2 \begin{pmatrix} 6 \\ 4 \\ 2 \end{pmatrix} \tag{182}$$

$$\Rightarrow \begin{cases} 9 = \lambda_1 + 6\,\lambda_2 \\ 2 = 2\,\lambda_1 + 4\,\lambda_2 \\ 7 = -\lambda_1 + 2\,\lambda_2 \end{cases} \Rightarrow \begin{cases} \lambda_1 = -3 \\ \lambda_2 = 2 \end{cases}$$

iv. Espace vectoriel Euclidien

Un espace vectoriel est *Euclidien* si en plus des axiomes définissant un espace vectoriel $(\mathcal{E}, +, \cdot)$ on lui associe un produit scalaire, c'est-à-dire une forme, bilinéaire, symétrique et définie positive. Un espace vectoriel seul n'a pas la notion de *distance* (ou de norme). Afin de donner une structure à cet espace, une sorte de maillage, on doit lui associer un produit scalaire pour calculer des angles et des longueurs. Parce que le produit scalaire est définie positif qu'on peut définir une norme. La norme Euclidienne est définie avec les propriétés suivantes :

– $\forall v \in \mathcal{E}, ||v||_2 = \sqrt{v \cdot v}$ si $||v||_2 = 0 \Leftrightarrow v = 0$.
– $\forall v_1, v_2 \in \mathcal{E}, ||v_1 + v_2||_2 \leq ||v_1||_2 + ||v_2||_2$ (inégalité triangulaire).
– Les inégalités de Cauchy-Schwartz montre que la norme Euclidienne est une vraie norme :

$$\forall v_1, v_2 \in \mathcal{E}, ||v_1 \cdot v_2||_2 \leq ||v_1||_2 \cdot ||v_2||_2$$

Exemple 1 :

$$\sum_i (x_i\, y_i)^2 \leq (\sum_i x_i^2) \cdot (\sum_i y_i^2)$$

Exemple 2 :

$$[f(x) \cdot g(x)]^2 \leq [f(x)]^2 \cdot [g(x)]^2$$

Preuve : Soit la distance entre deux vecteurs :

$$|v_1 - \lambda\, v_2||_2^2 = ||v_1||^2 - 2\lambda\, v_1\, v_2 + \lambda^2\, ||v_2||_2 = f(\lambda) \geq 0 \tag{183}$$

Nous remarquons que $f(\lambda)$ a une forme parabolique donc elle admet un minimum pour λ^* :

$$f^{'}(\lambda = \lambda^*) = 2\, v_1\, v_2 + 2\,\lambda^*\, ||v_2||_2 = 0 \Rightarrow \lambda^* = \frac{v_1 \cdot v_2}{||v_2||_2^2} \tag{184}$$

En substituant (184) dans (183), il vient :

$$||v_1||_2^2 - \frac{(v_1 \cdot v_2)^2}{||v_2||_2^2} \geq \Rightarrow ||v_1||_2^2 \geq \frac{(v_1 \cdot v_2)^2}{||v_2||_2^2}$$

$$\Rightarrow (v_1 \cdot v_2)^2 \leq ||v_1||_2^2 \cdot ||v_2||_2^2 \Leftrightarrow ||v_1 \cdot v_2||^2 \leq ||v_1||_2^2 \cdot ||v_2||_2^2$$

Exemples de normes :

- Norme 1 sur $\mathbb{R}^n$ ou $\mathbb{C}^n$, $||v|| = \sum_{i=1}^{n} x_i$. Avec x_i sont les coordonnées du vecteur v.
- Norme 2 sur $\mathbb{R}^n$ ou $\mathbb{C}^n$, $||v||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$.
- Norme ∞ sur $\mathbb{R}^n$ ou $\mathbb{C}^n$, $||v||_\infty = \max_{i=1,n} |x_i|$.

Théorème : $\forall v_1, v_2 \in \mathcal{E}$, la distance $d(v_1, v_2) = ||v_1 - v_2||$ est alors une *métrique* sur l'espace vectoriel $\mathcal{E}$

v. Produit scalaire sur un $\mathbb{R}$-espace vectoriel

Le but est de donner une notion de continuité dans l'espace des états de façon à traduire la continuité des évolutions des systèmes dans l'espace physique. Afin de passer d'un état physique à un autre on doit définir la notion de distance dans l'espace mathématique (espace vectoriel). Autrement dit, la distance ou la norme dans cet espace abstrait traduira la continuité de l'évolution des états dans l'espace physique (ou espace des mesures). Comme nous l'avons mentionné précédemment, formellement le produit scalaire est définie comme une *forme*, *bilinéaire*, *symétrique* et *définie positive*. Nous allons expliquer chaque terme de cette définition.

une *forme* est une application :

$$f : \mathcal{E} \times \mathcal{E} \longmapsto \mathcal{R}$$
$$\forall v_1, v_2 \in \mathcal{E} \longmapsto f(v_1, v_2)$$

Nous formons, c'est le cas de le dire, un scalaire $f(v_1, v_2)$ à partir de deux vecteurs. Le terme *bilinéaire* signifie que cette application est linéaire à gauche (par rapport à v_1) et à droite (par rapport à v_2).

$$\forall\, v_1, v_2, v_3 \in \mathcal{E}, \forall \lambda \in \mathcal{R} :$$
$$f(v_1, v_2 + \lambda\, v_3) = f(v_1, v_2) + \lambda\, f(v_1, v_3)$$
$$f(v_1 + \lambda\, v_2, v_3) = f(v_1, v_3) + \lambda\, f(v_2, v_3)$$

Le terme *symétrique* signifie : $f(v_1, v_2) = f(v_2, v_1)$. Le terme *positive* signifie $\forall\, v \in \mathcal{E}, f(v, v) \geq 0$ un produit scalaire sur lui-même donne un scalaire positif ou nul. Finalement le terme *définie positive* signifie :

$$\forall\, v \in \mathcal{E}, f(v, v) = 0 \Rightarrow v = 0_\mathcal{E} \quad \text{si} \quad v \neq 0_\mathcal{E} \Rightarrow f(v, v) > 0$$

Ce sont les propriétés standards d'un produit scalaire typiquement Euclidien dans un $\mathbb{R}$-espace vectoriel. Par ailleurs, nous pouvons associer à n'importe quelle forme linéaire de ce type une norme $\forall\, v \in \mathcal{E}$, $||v||^2 = f(v, v)$. Les propriétés de cette norme ont été déjà définies dans la section précédente. L'*Orthogonalité* s'exprime par :

$$\forall\, v_1, v_2 \in \mathcal{E}, v_1 \perp v_2 \Rightarrow f(v_1, v_2) = 0$$

Une base est orthonormée si $\forall\, i \neq j \in \mathbb{N}^* \Rightarrow f(v_i, v_j) = 0$ et $f(v_i, v_i) = 1$. La norme des vecteurs constituant la base est égale à 1. On peut passer d'une base quelconque à une base orthonormée au moyen de l'algorithme de *Gram-Schmidt* :

$$u_k = v_k - \sum_{i=1}^{k-1} \frac{v_k\, u_i}{u_i\, u_i}\, u_i \tag{185}$$

Avec $\{v_k\}_{k\in\mathbb{N}^*}$ sont les vecteurs de la base quelconque et $\{u_i\}_{i\in\mathbb{N}^*}$ sont les vecteurs de la base orthogonale. La base orthonormée est obtenue en divisant chaque vecteur de la base orthogonale par sa norme.

$$\text{Base orthonormée} \quad \Rightarrow \frac{u_k}{||u_k||}$$

vi. Espace hermitien : Pour les espaces vectoriels Euclidiens, nous avons

$$\forall v \in \mathcal{E} \Rightarrow v \cdot v = \sum_{i=1}^{n} x_i^2 > 0 \quad \text{seulement si} \quad x_i \in \mathbb{R}$$

Dans un espace vectoriel hermitien ($\mathbb{K} = \mathbb{C}$), afin de disposer d'une norme $||\cdot|| \in \mathbb{R}$ il faudra :

$$\forall z \in \mathcal{C} \Rightarrow z \cdot z = \sum_{i=1}^{n} |z_i|^2 = \sum_{i=1}^{n} z_i \cdot z_i^* > 0 \quad \text{produit scalaire hermitien}$$

Ainsi,

$$\forall\, x\, z \in \mathcal{C}^2 \quad : w \cdot z = \sum_{i=1}^{n} w_i\, z_i^* \neq z \cdot w \quad \text{avec} \quad z \cdot w = \sum_{i=1}^{n} z_i\, w_i^*$$

En effet,

$$z \cdot w = \sum_{i=1}^{n} z_i\, w_i^* = \sum_{i=1}^{n} [z_i^*]^*\; w_i^* = \left[\sum_{i=1}^{n} z_i^*\, w_i^*\right]^* = [w \cdot z]^*$$

Définition : Un espace hermitien est un espace vectoriel sur le corps $\mathbb{C}$ muni d'un produit scalaire hermitien : $\mathcal{E} \times \mathcal{E} \longmapsto \mathbb{C}$ ayant les propriétés suivantes, $\forall \lambda_1, \lambda_2 \in \mathbb{C}^2, \forall z_1, z_2 \in \mathcal{E}^2$:

- Linéarité à gauche : $(\lambda_1\, z_1 + \lambda_2\, z_2)\, z_2 = \lambda_1\, (z_1 \cdot z_2) + \lambda_2\, (z_2 \cdot z_3)$.
- Sesquilinéaire à droite : $z_1 \cdot (\lambda_1\, z_2 + \lambda_2\, z_3)\, z_2 = \lambda_1^*\, z_1 \cdot z_2 + \lambda_2^*\, z_1 \cdot z_3$. En notation de Dirac $|\lambda\, \psi\rangle = \lambda\, |\psi\rangle$ (linéarité à droite) et $\langle \lambda\, \varphi| = \lambda^* \langle \varphi|$ (sesquilinéaire à gauche).
- $z_1 \cdot z_2 = (z_2 \cdot z_1)^*$.
- Défini positif : $z \cdot c > 0$ si $z \neq 0$.
- On définit la norme hermitienne : $||z|| = \sqrt{z \cdot z}$.

Les autres propriétés d'un espace vectoriel Euclidien (inégalité triangulaire, inégalité de Cauchy-Schwartz, ...) restent valables pour le produit scalaire hermitien, même l'algorithme de Gram-Schmidth.

Références

F. Filbet, Analyse numérique : algorithme et étude mathématique, Édition Dunod, ***2013****.*

J. Ouin, Algorithmique : calcul numérique, Édition Ellipses, ***2013****.*

C. Brezinski, M. Redivo-Zaglia, Méthodes numériques itératives, Édition Ellipses, ***2006****.*

J.P. Grivet, Méthodes numériques appliquées, Édition EDP Sciences, ***2013****.*

J. Chaskalovic, Méthodes mathématiques et numériques pour les équations aux dérivées partielles, Édition Lavoisier, ***2013****.*

M. Bernadou, Le calcul scientifique, Édition PUF, ***2001****.*

G. Marchouk, Méthodes de calcul numérique, Édition MIR, Moscou, ***1980****.*

R. Richtmyer, K. Morton, Difference methods for initial value problem, Wiley, New York, ***1967****.*

D. J. Griffiths, Introduction to Quantum Mechanics, Prentice Hall, Englewood Cliffs, New Jersey, ***1995****.*

C. Cohen-Tannoudji, B. Diu, and F. Laloë, Quantum Mechanics, John Wiley & Sons, New York, ***1977****.*

R.G. Parr and W. Yang, Density-Functional Theory of Atoms and Molecules, Oxford University Press, New York, ***1989****.*

W. Koch and M.C. Holthausen, A Chemist's Guide to Density Functional Theory, WILEYVCH, ***2001****.*

W. Kohn and L.J. Sham, Self Consistent Equations Including Exchange and Correlation Effects, Phys. Rev. 140, A1133, ***1965****.*

A. J. Austin, Studies in Computational Quantum Chemistry, MedCrave Group LLc, ***2016****.*

K. I. Ramachamdran, G. Deepa, K. Namboori, Computational Chemistry and Molecular Modeling, Springer ***2008****.*

P. Hohenberg and W. Kohn, Inhomogeneous Electron Gas, Phys. Rev. 136, B864, ***1964****.*