MÉTHODES DE RÉSOLUTION D'ÉQUATIONS DIFFÉRENTIELLES POUR DES PROBLÈMES RAIDES

DIDIER AUROUX

Groupe de travail théorique réalisé sous la direction de M. Schatzman dans le cadre du DEA d'Analyse numérique, Equations aux dérivées partielles et Calcul scientifique Ecole Doctorale MATHIF de Lyon

RÉSUMÉ. Un nouveau type de méthodes de résolution d'équations différentielles raides est construit, combinant la A-stabilité (ou L-stabilité) avec une certaine facilité de mise en œuvre. Ces méthodes sont construites de telle sorte que leur région de stabilité soit la même que pour des méthodes de Runge-Kutta.

1. Introduction

La plupart des méthodes de résolution de systèmes d'équations différentielles se classent dans les deux groupes suivants : les méthodes multipas et les méthodes de type Runge-Kutta. Lorsqu'on s'intéresse à des problèmes raides, ces types de méthodes présentent tous deux quelques difficultés : les méthodes multipas linéaires ne sont pas A-stables si l'ordre est supérieur à 2 (théorème de Dahlquist page 265), et les méthodes Runge-Kutta (bien que A-stables même si l'ordre est arbitrairement grand) ont des coûts d'implémentation énormes.

Il devient donc nécesaire d'émettre des hypothèses sur les coefficients des méthodes RK afin de les rendre implémentables. Un premier critère consiste à ne garder que les méthodes dont la matrice a un spectre réduit à un point ($\sigma(A) = \{\lambda\}$). Ces méthodes *uni-implicites* (SIRK) ont un coût comparable aux méthodes pour lesquelles la matrice est triangulaire inférieure (DIRK). Néanmoins ces méthodes conduisent à de grandes constantes d'erreur, et nous avons encore une trop grande classe de méthodes pour en trouver une sous-classe acceptable pour les problèmes raides.

En émettant des hypothèses supplémentaires et en cherchant de nouveaux critères de choix, un nouveau type de méthodes précises et efficaces (DIMSIM) a été identifié. Nous nous restreindrons aux méthodes qui possèdent une *stabilité inhérente Runge-Kutta* : nous choisirons les coefficients de sorte que la zone de stabilité soit la même que pour une méthode Runge-Kutta, et la forme *inhérente* de cette propriété repose sur des hypothèses émises sur les matrices qui garantissent la stabilité RK avant même d'avoir cherché à imposer d'autres conditions (telles que des conditions d'ordre).

Nous étudierons ensuite divers cas en fonction du nombre d'étapes et de l'ordre de la méthode, puis une implémentation de quelques-unes de ces méthodes sera présentée.

2. Notations et généralités sur les méthodes DIMSIM

Nous nous intéressons ici au problème $y'(x) = f(x, y(x))$, que l'on peut approcher de diverses manières. Une discrétisation simple de ce problème est la suivante : $y_n = y_{n-1} + hf(x_n, y_n)$ où h est le pas de calcul $x_n - x_{n-1}$. Ainsi, de nombreuses méthodes reposent sur la résolution de façon successive de systèmes $Y = Z + h\lambda f(X, Y)$ où Y est l'inconnue, X une quantité connue du temps, et Z la solution du système précédent. Le fait qu'on soit amené à résoudre plusieurs fois le même type de systèmes nous incite à considérer des méthodes uni-implicites.

De façon générale, nous supposerons que chaque étape de la méthode utilise r variables de l'étape précédente et en produit autant pour l'étape suivante. On suppose également qu'il y a s pas élémentaires, que l'ordre de la méthode globale est p et que l'ordre des méthodes

Date: 2 mars 2000 *Email :* Didier.Auroux@ens-lyon.fr .

élémentaires est q. Une méthode de ce type peut alors se représenter sous la forme assez générale :

$$\left[\begin{array}{cc} A & U \\ B & V \end{array}\right]$$

où les matrices A, B, U et V sont de dimensions respectives $s \times s$, $r \times s$, $s \times r$ et $r \times r$. À la $n^{\text{ième}}$ étape, on utilise r quantités calculées à l'étape $n-1$: $y_1^{[n-1]} \dots y_r^{[n-1]}$ et on produit r nouvelles valeurs en résolvant le système

$$\boxed{\begin{array}{rcl} Y_i & = & \displaystyle\sum_{j=1}^{s} a_{ij} h F_j \quad + \quad \sum_{j=1}^{r} u_{ij} y_j^{[n-1]}, \quad i = 1 \dots s \\ y_i^{[n]} & = & \displaystyle\sum_{j=1}^{s} b_{ij} h F_j \quad + \quad \sum_{j=1}^{r} v_{ij} y_j^{[n-1]}, \quad i = 1 \dots r \end{array}} \tag{1}$$

où $Y_1 \dots Y_s$ sont les états intermédiaires et $F_1 \dots F_s$ les dérivées correspondantes.

Pour réduire la catégorie de méthodes ainsi décrites, le type de méthodes DIMSIM suppose que A soit triangulaire inférieure. On suppose également que l'ordre de la méthode et celui des méthodes élémentaires sont égaux ($p = q$), et proches de r et s. Le choix de la structure de A est crucial puisqu'il déterminera une grande partie des coûts de calcul (méthode implicite). Pour une implémentation parallèle, les méthodes de type 4 sont obtenues à partir d'une matrice A multiple de l'identité ($A = \lambda I$), et celles de type 3 découlent de $A = 0$ pour des problèmes raides, A diagonale sinon. Pour une implémentation classique (séquentielle), nous avons le choix entre les méthodes de type 2 (pour des problèmes raides) où A est triangulaire inférieure avec tous les éléments diagonaux égaux à une constante positive λ, et celles de type 1 (autres types de problèmes) où A est triangulaire inférieure à diagonale nulle. Dans la suite, nous nous restreindrons aux méthodes de type 2.

Notons J la matrice suivante :

$$J = \left[\begin{array}{ccccccc} 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 1 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 & 0 \end{array}\right],$$

soit $c_1 \dots c_s \in \mathbb{R}^s$, soient $t_1 \dots t_p$ p réels, notons ψ la fonction $(z \mapsto 1 + t_1 z + \dots + t_p z^p)$. Soit enfin le vecteur $Z = [1\ z\ z^2\ \dots\ z^p]^T$. Alors nous avons

$$\boxed{\psi(z) \exp(cz) = \Phi Z + \mathcal{O}(z^{p+1}),} \tag{2}$$

avec $\Phi = [\Phi_0\ \Phi_1\ \dots\ \Phi_p]$ et $(\Phi_k)_i = \sum_{j=0}^{k} t_j \frac{c_i^{k-j}}{(k-j)!}$.

3. Stabilité inhérente de Runge-Kutta

3.1. **Matrice de stabilité.** Considérons l'équation différentielle $x' = \lambda x$. Un schéma de Runge-Kutta pour résoudre cette équation s'écrit :

$$\begin{array}{rcl} X_{n,i} & = & x_n \quad + \quad \displaystyle\sum_{j=1}^{r} a_{ij} \lambda h X_{n,j}, \quad i = 1 \dots r \\ x_{n+1} & = & x_n \quad + \quad \displaystyle\sum_{j=1}^{r} b_j \lambda h X_{n,j}, \end{array}$$

On a donc $x_{n+1} = M(\lambda h) x_n$ où $M(z)$ est une fraction rationnelle en z ne dépendant que des coefficients de la méthode. $M(z)$ est alors appelée la *matrice de stabilité* du schéma numérique.

Remarque : Dans ce cas, M étant à coefficients matriciels de dimension 1, M sera notée R et appelée *fonction de stabilité*.

Dans le cas étudié ici, en supposant que $f(x,y) = \lambda y$, en notant $Y = (Y_1 \dots Y_s)^T$, $y^{[n]} = (y_1^{[n]} \dots y_r^{[n]})^T$ et $y^{[n-1]} = (y_1^{[n-1]} \dots y_r^{[n-1]})^T$, on a $Y = (1 - h\lambda A)^{-1} U y^{[n-1]}$ et donc $y^{[n]} = M(h\lambda) y^{[n-1]}$ avec $M(z) = V + zB(I - zA)^{-1}U$. M est donc la matrice de stabilité.

3.2. **Fonction de stabilité.** Nous supposerons désormais que $Ve_1 = e_1$ (e_1 étant le premier vecteur de la base), et que seule la première ligne des matrices $BA - JB$ et $BU - JV + VJ$ est non nulle. On a alors le théorème suivant :

Théorème 1 ([But]). *Notons $(X-1)P(X)$ le polynôme caractéristique de V, alors le polynôme caractéristique de la matrice de stabilité M est de la forme $(X - R(z))P(X)$.*

Démonstration [But] : Par hypothèse, on a

$$\begin{aligned} BA &= JB + e_1(e_1^T BA), \\ BU &= JV - VJ + e_1(e_1^T BU + e_1^T VJ). \end{aligned} \tag{3}$$

En notant $a = e_1^T BA$ et $b = e_1^T BU + e_1^T VJ$, (3) devient

$$BA = JB + e_1 a \tag{4}$$

$$BU = JV - VJ + e_1 b \tag{5}$$

et (4) devient

$$\begin{aligned} B(I - zA) &= (I - zJ)B - ze_1 a \\ (I - zJ)B &= B(I - zA) + ze_1 a \\ (I - zJ)B(I - zA)^{-1} &= B + ze_1 a(I - zA)^{-1}. \end{aligned} \tag{6}$$

En utilisant (5) et (6), et en travaillant sur une matrice semblable à M (ce qui ne change pas le polynôme caractéristique), on obtient :

$$\begin{aligned} M &\sim (I - zJ)M(I - zJ)^{-1} \\ &= (I - zJ)V(I - zJ)^{-1} + z(I - zJ)B(I - zA)^{-1}U(I - zJ)^{-1} \\ &= (I - zJ)V(I - zJ)^{-1} + z(Bze_1 a(I - zA)^{-1})U(I - zJ)^{-1} \\ &= \left((I - zJ)V + z(BU + ze_1 a(I - zA)^{-1}U)\right)(I - zJ)^{-1} \\ &= V + \left(ze_1 b + z^2 e_1 a(I - zA)^{-1}U\right)(I - zJ)^{-1}. \end{aligned} \tag{7}$$

La matrice ainsi obtenue en (7) est identique à V sauf la première ligne et donc son polynôme caractéristique est de la forme $(X - R(z))P(X)$, avec

$$\begin{aligned} R(z) &= e_1^T \left((I - zJ)M(I - zJ)^{-1}\right) e_1 \\ &= e_1^T MZ \\ &= e_1^T \left(V + zB(I - zA)^{-1}U\right) Z. \end{aligned} \tag{8}$$

□

R est alors appelée *fonction de stabilité.*

$R(z)$ peut être interprété comme étant la solution après un pas de temps du problème

$$\boxed{y' = \lambda y, \quad y_0 = 1, \quad z = h\lambda} \tag{9}$$

(problème connu sous le nom de l'*équation test de Dahlquist*). La méthode est stable si toutes les valeurs propres de M sont incluses dans le disque unité. Nous supposerons donc par la suite que les zéros de P sont situés dans le disque unité. $S = \{z \in \mathbb{C}; |R(z)| \leqslant 1\}$ est donc la *zone de stabilité* de la méthode. La solution exacte de (9) est e^z et donc si la méthode est d'ordre p, $R(z)$ doit satisfaire

$$e^z - R(z) = \mathcal{O}(h^{p+1}) = \mathcal{O}(z^{p+1}) = Cz^{p+1} + \mathcal{O}(z^{p+2}) \quad \text{quand} \quad z \to 0, \tag{10}$$

C étant la constante d'erreur de la fonction de stabilité. En utilisant la méthode définie par (1) pour résoudre le problème test de Dahlquist (9), nous obtenons

$$\begin{pmatrix} I - zA & 0 \\ -zB & 1 \end{pmatrix} \begin{pmatrix} Y \\ y^{[n]} \end{pmatrix} = \begin{pmatrix} U \\ V \end{pmatrix} y^{[n-1]},$$

et donc le dénominateur de $R(z)$ est $\det(I - zA)$ et le numérateur est

$$\det \begin{pmatrix} I - zA & U \\ -zB & V \end{pmatrix}. \tag{11}$$

Nous avons donc $R(z) = \frac{P(z)}{Q(z)}$ où P et Q sont deux polynômes de degré s, et $Q(z) = \det(I - zA) = (1 - \lambda z)^s = \sum_{i=0}^{s} q_i z^i$. Nous savons aussi que $R(z) = e^z + \mathcal{O}(z^{p+1})$, donc $P(z) = e^z Q(z) + \mathcal{O}(z^{p+1}) = (\sum_{i=0}^{p} \frac{z^i}{i!})(1 - \lambda z)^s + \mathcal{O}(z^{p+1}) = \sum_{i=0}^{s} p_i z^i$, avec $p_i = \sum_{k=0}^{i} \frac{q_k}{(s-k)!}$. Nous obtenons $P(z) = (-1)^s \sum_{j=0}^{s} L_s^{(s-j)}(\frac{1}{\lambda})(\lambda z)^j$, avec $L_s(X) = \sum_{j=0}^{s} (-1)^j C_s^j \frac{X^j}{j!}$ le polynôme de Laguerre de degré s.

Dans le cas $s = p + 1$, nous aurons donc

$$\boxed{R(z) = \frac{(-1)^{p+1} \sum_{i=0}^{p} L_{p+1}^{(p+1-i)}\left(\frac{1}{\lambda}\right) \lambda^i z^i}{(1 - \lambda z)^{p+1}},} \tag{12}$$

avec λ choisi pour la A-stabilité (voir les définitions 1 à 3 ci-dessous). La L-stabilité en découlera automatiquement puisque $R(\infty) = 0$ (le coefficient en $z^s = z^{p+1}$ ayant été choisi nul). Dans le cas $s = p + 2$, nous avons un degré de liberté supplémentaire α :

$$\boxed{R(z) = \frac{(-1)^{p+2} \sum_{i=0}^{p} L_{p+2}^{(p+2-i)}\left(\frac{1}{\lambda}\right) \lambda^i z^i + \alpha z^{p+1}}{(1 - \lambda z)^{p+2}},} \tag{13}$$

le terme en z^s ayant encore été choisi nul de façon à assurer la L-stabilité.

3.3. **A-stabilité, L-stabilité.** La solution exacte de (9) est stable dans tout le demi-plan gauche $\mathbb{C}^-$, et une propriété intéressante des méthodes cherchées est que ce soit aussi leur domaine de stabilité :

Définition 1 (A-stabilité). *La méthode est dite* A-stable *si* $\{\mathcal{R}ez \leqslant 0\} \subset \{|R(z)| \leqslant 1\}$.

Nous avons alors le critère suivant de A-stabilité :

Proposition 1. *La méthode est A-stable si et seulement si*

$$\begin{array}{ll} |R(iy)| \leqslant 1 \ \forall y \in \mathbb{R} & (i) \\ R(z) \text{ est analytique sur } \mathcal{R}ez < 0 & (ii) \end{array} \tag{14}$$

La démonstration découle de façon immédiate du principe du maximum appliqué à R sur $\mathbb{C}^-$.

La A-stabilité peut être dans certains cas un critère trop fort de stabilité, et nous devons donc introduire une notion plus générale :

Définition 2 (A(α)-stabilité). *Une méthode est dite* A(α)-stable *si le secteur angulaire* $\{z \in \mathbb{C};\ |\arg(-z)| \leqslant \alpha,\ z \neq 0\}$ *est inclus dans son domaine de stabilité.*

Il est clair que la A-stabilité n'est autre que la A($\frac{\pi}{2}$)-stabilité. Celà étant, ce genre de critère de stabilité est trop faible en un certain sens. En effet, supposons désormais que $\mathbb{C}^-$ soit *exactement* la zone de stabilité de la méthode. Nous en déduisons que $R(iy) = 1,\ \forall y \in \mathbb{R}$ et donc $\lim_{|z| \to \infty} |R(z)| = 1$ puisque R est une fraction rationnelle. Par conséquent, si $\mathcal{R}ez \ll 0$, $|R(z)|$ sera très proche de 1 et donc la dissipation des modes instables aura lieu très lentement, ce qui n'est pas forcément une bonne chose. Pour y remédier, nous définissons un deuxième type de stabilité :

Définition 3 (L-stabilité). *La méthode est dite* L-stable *si elle est A-stable et si*

$$\lim_{z \to \infty} R(z) = 0.$$

Puisque nous cherchons des méthodes L-stables, nous avons besoin d'évaluer la limite de $R(z)$ en l'infini :

$$R(z) = e_1^T V Z + z e_1^T B \left(-z^{-1} A^{-1} \sum_{i=0}^{\infty} z^{-i} A^{-i} \right) UZ.$$

z n'apparaît qu'avec une puissance négative ou nulle dans cette formule, et donc il suffit de chercher le coefficient de z^0 :

$$R(\infty) = 1 - \sum_{i=1}^{p} e_1^T B A^{-i} U e_i = 0. \tag{15}$$

Ceci nous fournit une condition supplémentaire sur les coefficients de la méthode. Dans le cas $s = p+2$, puisque nous avons un degré de liberté supplémentaire, il faut évaluer le coefficient de z^{-1} et on trouve

$$\lim_{z\to\infty} zR(z) = -\sum_{i=1}^{p} e_1^T B A^{-i-1} U e_i = (-1)^p \alpha \lambda^{-p-2}. \tag{16}$$

4. Conditions d'ordre et calcul des coefficients de la méthode

4.1. Conditions d'ordre. Nous cherchons des méthodes pour lesquelles $q = p$. Sauf indication contraire, nous supposerons désormais que $s = p+1$. En notant $\phi(z)$ le vecteur composé des r approximations importées de l'étape précédente, on suppose que chaque composante de $\phi(z)$ est un polynôme de degré p en z, l'ordre de la méthode élémentaire étant égal à p, nous avons donc

$$\begin{aligned} \exp(cz) &= zA\exp(cz) + U\phi(z) + \mathcal{O}(z^{p+1}), &(17)\\ \exp(z)\phi(z) &= zB\exp(cz) + V\phi(z) + \mathcal{O}(z^{p+1}). &(18)\end{aligned}$$

En effet, si $\phi(z)$ est la solution de $y' = \lambda y$ à l'instant t, la solution à l'instant $t+h$ est $\phi(z)\exp(h\lambda) = \phi(z)\exp(z)$. La dérivée à l'instant t est $\lambda\exp(cz)$ ce qui explique (17).

Nous allons nous restreindre au cas où $\phi(z)$ est à $\mathcal{O}(z^{p+1})$ près $\phi_1(z)Z$. Pour des raisons de simplicité des calculs, nous noterons dans la suite $\phi_1(z) = \psi(z)^{-1} + \mathcal{O}(z^{p+1})$ avec $\psi(z) = 1 + \sum_{i=1}^{p} t_i z^i$, $t \in \mathbb{R}^p$. (17) et (18) peuvent alors se réécrire

$$\begin{aligned} \psi(z)(I - zA)\exp(cz) &= UZ + \mathcal{O}(z^{p+1}), &(19)\\ \exp(z)Z &= z\psi(z)B\exp(cz) + VZ + \mathcal{O}(z^{p+1}). &(20)\end{aligned}$$

Supposons que (19) soit vrai, ainsi que la première ligne de (20). En notant $u_1 \dots u_{p+1}$ les vecteurs colonnes de U et respectivement $v_1^T \dots v_{p+1}^T$ et $b_1^T \dots b_{p+1}^T$ les vecteurs lignes de V et B, on peut réécrire les hypothèses précédant le théorème 1 (4) et (5) sous la forme suivante :

$$\begin{aligned} b_i^T A &= b_{i-1}^T, \quad 2 \leqslant i \leqslant p+1, &(21)\\ b_i^T U &= v_{i-1}^T - v_i^T J, \quad 2 \leqslant i \leqslant p+1. &(22)\end{aligned}$$

La $i-1^{\text{ème}}$ ligne de (20) s'écrit

$$\exp(z)z^{i-2} = z\psi(z)b_{i-1}^T\exp(cz) + v_{i-1}^T Z + \mathcal{O}(z^{p+1}),$$

en éliminant b_{i-1} et v_{i-1} à l'aide de (21) et (22), et en multipliant par z, on obtient

$$\begin{aligned} \exp(z)z^{i-1} &= z\left(z\psi(z)\exp(cz)b - i^T A + bi^T UZ + v_i^T JZ\right) + \mathcal{O}(z^{p+1})\\ &= z\psi(z)b_i^T\exp(cz) + v_i^T zJZ + \mathcal{O}(z^{p+1})\\ &= z\psi(z)b_i^T\exp(cz) + v_i^T Z + \mathcal{O}(z^{p+1}),\end{aligned}$$

qui est exactement la $i^{\text{ème}}$ ligne de (20).

Il suffit donc de s'assurer que (19) et la première ligne de (20) ont lieu pour satisfaire les conditions d'ordre requises.

4.2. Construction de U. (19) peut être réécrite sous la forme

$$\psi(z)\exp(cz) = (I - zA)^{-1}UZ + \mathcal{O}(z^{p+1}) = \Phi Z$$

d'après la définition de Φ (2). Les colonnes de U peuvent donc se calculer de la façon suivante :

$$\boxed{u_i = \Phi_{i-1} - A\Phi_{i-2}, \qquad i = 1\dots r,} \tag{23}$$

en utilisant la convention $\Phi_{-1} = 0$.

4.3. **Construction de B.** Nous noterons dans la suite $\beta^T = b_2^T$ la deuxième ligne de B. D'après (21), la première ligne de B est $\beta^T A$. Nous supposerons dorénavant que la deuxième ligne de V est nulle. Alors les deux premiers coefficients du vecteur produit lors de la première étape sont d'après (17) et (18)

$$(24) \quad e^z\Phi(z) = z\begin{pmatrix} b_1^T e^{cz} \\ b_2^T e^{cz} \\ \vdots \end{pmatrix} + \begin{pmatrix} v_1^T\Phi(z) \\ v_2^T\Phi(z) \\ \vdots \end{pmatrix} = \begin{pmatrix} z\beta^T A e^{cz} + \beta^T U\Phi(z) \\ z\beta^T e^{cz} \\ \vdots \end{pmatrix} = \begin{pmatrix} \beta^T e^{cz} \\ z\beta^T e^{cz} \\ \vdots \end{pmatrix}.$$

Si l'on impose que la somme des éléments de β vaut 1, alors les deux premières valeurs construites peuvent être vues respectivement comme une combinaison linéaire des états et comme la même combinaison linéaire des états dérivés (ce qui peut être vu comme une condition de consistance du schéma).

Il peut être pratique d'imposer que le rayon spectral de la matrice obtenue à partir de V en supprimant les deux premières lignes et colonnes soit nul, et la façon la plus simple de le faire est de supposer que, mis à part la première ligne, tous les coefficients sur et au-dessus de la diagonale de V sont nuls. Ainsi, d'après le théorème 1, le spectre de la matrice de stabilité est $\{0, R(z)\}$ et donc la méthode est effectivement stable exactement sur $\{z \in \mathbb{C}; |R(z)| \leqslant 1\}$.

Dans ces conditions, la deuxième ligne de (20) est alors

$$\exp(z)z = z\psi(z)\beta^T \exp(cz) + \mathcal{O}(z^{p+1})$$

et donc

$$(25) \qquad \psi(z)\sum_{i=1}^{s}\beta_i \exp(zc_i) = \exp(z) + \mathcal{O}(z^{p+1}).$$

En comparant la définition de Φ (2) et (25), on trouve

$$\boxed{\beta^T\Phi_i = \tfrac{1}{i!}, \qquad i = 0\dots p.} \tag{26}$$

Les équations (21) permettent alors de déduire B de β.

4.4. **Construction de V.** Nous cherchons maintenant à calculer les éléments de la matrice V. D'après (22), $v_1 = \beta^T U$ et en utilisant (23) et (26), on obtient

$$\boxed{v_{1i} = \tfrac{1}{i!} - \beta^T A\Phi_{i-2}, \qquad i = 1\dots p+1.} \tag{27}$$

Il nous reste à trouver les éléments de V en dessous de la diagonale. En utilisant (5) (découlant des hypothèses émises pour le théorème 1) sous la forme $\beta^T A^{2-i}U = v_{i-1} - v_i J$, et en substituant les colonnes de U calculées à l'aide de (23), on trouve

$$\begin{aligned} (28)\quad v_{ij} + \beta^T A^{2-i}\Phi_{j-2} &= v_{i-1,j-1} + \beta^T A^{3-i}\Phi_{j-3}, \qquad 3 \leqslant i \leqslant p+1, \quad 2 \leqslant j \leqslant p+1, \\ (29)\qquad\qquad v_{i,p+1} &= \beta^T A^{1-i}\Phi_p - \beta^T A^{2-i}\Phi_{p-1}, \qquad 2 \leqslant i \leqslant p+1. \end{aligned}$$

On constate donc que $v_{ij} + \beta^T A^{2-i}\Phi_{j-2}$ est constant sur chacune des sous-diagonales de V, et pour $j = 1$ cette quantité est nulle puisque $v_{ij} = 0$ et $\Phi_{-1} = 0$. Nous avons donc

$$\boxed{v_{ij} = -\beta^T A^{2-i}\Phi_{j-2}, \qquad i > j,} \tag{30}$$

ce qui nous permet de finir la construction de la matrice V.

4.5. **Construction de A.** D'après (26), (28) et (29), nous avons

$$(31) \qquad \beta^T A^{-i}\Phi_j = \frac{1}{(j-i)!}, \qquad 0 \leqslant i \leqslant p-1, \quad i \leqslant j \leqslant p.$$

4.5.1. $s = p+1$. Nous pouvons déduire de (31) le système d'équations suivant :

$$(32) \qquad \beta^T(I - \lambda A^{-1})^i\Phi_j = \sum_{k=0}^{i} C_i^k \frac{(-\lambda)^k}{(j-k)!}, \qquad 0 \leqslant i \leqslant j \leqslant p.$$

Comme le vecteur ligne $\beta^T(I - \lambda A^{-1})^i$ a ses i dernières composantes nulles, il y a juste assez d'équations dans (32) pour le calculer, sous réserve que ce système d'équations soit libre. Ceci sera le cas si l'on choisit les c_i distincts deux à deux.

Considérons alors la matrice W formée des lignes $\beta^T(I-\lambda A^{-1})^i$ pour $i = p\dots 0$. C'est une matrice $(p+1)\times(p+1)$ triangulaire inférieure. Nous avons alors

$$\boxed{A = \lambda W^{-1}(I-J)^{-1}W,} \tag{33}$$

ce qui nous permet d'implémenter A.

4.5.2. $s = p+2$. Dans ce cas, pour $i \in [0,p]$, il nous manque une équation pour déterminer les éléments non-nuls de $\beta^T(I-\lambda A^{-1})^i$. Pour $i = p$, nous pouvons utiliser la condition supplémentaire que nous avions obtenu en (16). En effet, un calcul analogue à celui effectué en (15) nous donne :

$$\lim_{z\to\infty} zR(z) = -\sum_{i=1}^{p} e_1^T B A^{-i-1} U e_i = (-1)^p \alpha \lambda^{-p-2}, \tag{34}$$

et avec la définition de U donnée par (23), on obtient

$$\lim_{z\to\infty} zR(z) = -\beta^T A^{-p}\Phi_{p-1}. \tag{35}$$

En comparant ces deux dernières équations, on trouve

$$\beta^T A^{-p}\Phi_{p-1} = (-1)^{p+1}\alpha\lambda^{-p-2}, \tag{36}$$

ce qui nous fournit l'équation supplémentaire recherchée.

4.6. **Récapitulatif de la construction de la méthode.**

4.6.1. *Hypothèses émises sur les matrices.* Les hypothèse émises sur la méthode peuvent partiellement se résumer à :

$$A = \begin{bmatrix} \lambda & 0 & \dots & 0 \\ * & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ * & \dots & * & \lambda \end{bmatrix} \qquad V = \begin{bmatrix} 1 & * & \dots & \dots & * \\ 0 & 0 & \dots & \dots & 0 \\ \vdots & * & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & * & \dots & * & 0 \end{bmatrix}.$$

D'autres hypothèses ont été émises, telles que les hypothèses préliminaires au théorème 1.

4.6.2. *Construction de la méthode.* Sur le diagramme suivant, nous pouvons voir l'ordre de construction des divers éléments définissant la méthode.

$$\left\{\begin{array}{lllllll} c & & & & & & \xrightarrow{(23)} U \\ t & \xrightarrow{(2)} \Phi & \xrightarrow{(26)} \beta & \xrightarrow{(32)} W & \xrightarrow{(33)} A & & \xrightarrow{(21)} B \\ \lambda & & & & & & \xrightarrow{(22)} V \end{array}\right.$$

Il suffit donc de définir c, t et λ (qui sera choisi pour la A et L-stabilité), pour en déduire tout le reste.

5. Exemples de méthodes numériques

5.1. **Méthodes choisies.** Deux méthodes ont été particulièrement étudiées. La première, pour laquelle $p = 2$ et $s = 3$, est représentée par la matrice suivante :

$$\left[\begin{array}{ccc|cc} \frac{4}{9} & 0 & 0 & 1 & \frac{5}{9} \\ \frac{29}{72} & \frac{4}{9} & 0 & 1 & \frac{47}{72} \\ \frac{1295}{522} & -\frac{248}{261} & \frac{4}{9} & 1 & \frac{13}{522} \\ \hline \frac{4301}{4698} & -\frac{320}{783} & \frac{8}{81} & 1 & \frac{1853}{4698} \\ \frac{11}{9} & -\frac{4}{9} & \frac{2}{9} & 0 & 0 \end{array}\right]. \tag{37}$$

Cette méthode a été construite en fixant $\lambda = \frac{4}{9}$ pour la L-stabilité, $c = [0\ \frac{1}{2}\ 1]^T$, $r = 2$, et $t_2 = \lambda t_1$, en choisissant de façon arbitraire $t_1 = 1$.

La deuxième méthode mise en œuvre est d'ordre 4 ($p = 4$) et donc $s = 5$. La valeur de λ est choisie pour la stabilité ($\lambda = \frac{1}{4}$). Pour la consistance de la méthode, en supposant $r = p$, on doit avoir $t_4 = \lambda t_3$. La méthode est construite pour $t_1 = \frac{1}{2}$, $t_2 = \frac{1}{16}$, $t_3 = -\frac{1}{96}$ et donc $t_4 = -\frac{1}{384}$, ce qui correspond à $\beta^T = [\frac{1}{4}\ 0\ \frac{1}{2}\ 0\ \frac{1}{4}]$. Les matrices A, B, U et V peuvent alors être construites, et l'on obtient :

$$\left[\begin{array}{ccccc|cccc}
\frac{1}{4} & 0 & 0 & 0 & 0 & 1 & \frac{1}{4} & -\frac{1}{16} & -\frac{5}{192} \\
\frac{47}{64} & \frac{1}{4} & 0 & 0 & 0 & 1 & -\frac{15}{64} & -\frac{43}{128} & -\frac{79}{1024} \\
\frac{24197}{14476} & \frac{678}{3619} & \frac{1}{4} & 0 & 0 & 1 & -\frac{4013}{3619} & -\frac{45673}{57904} & -\frac{52343}{347424} \\
\frac{7102302807}{1544183872} & \frac{987465}{24127873} & \frac{10395}{26668} & \frac{1}{4} & 0 & 1 & -\frac{6223228775}{1544183872} & -\frac{7146274211}{3088367744} & -\frac{261876469}{655936512} \\
-\frac{117251104}{55207845} & -\frac{27818059}{55207845} & \frac{7255}{6102} & -\frac{59}{135} & \frac{1}{4} & 1 & \frac{690268919}{220831380} & \frac{9714337}{6543152} & \frac{55597}{23161} \\
\hline
\frac{4213687}{11471760} & -\frac{92627}{2876940} & \frac{5153}{12204} & -\frac{59}{540} & \frac{1}{16} & 1 & \frac{830293}{2876940} & -\frac{3301}{84976} & -\frac{197}{9024} \\
\frac{1}{4} & 0 & \frac{1}{2} & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 \\
-\frac{427}{135} & \frac{788}{135} & -\frac{148}{27} & \frac{236}{135} & 1 & 0 & \frac{8}{135} & 0 & 0 \\
-\frac{611188}{15255} & \frac{1161632}{15255} & -\frac{191464}{3051} & \frac{1888}{135} & 4 & 0 & \frac{132512}{15255} & \frac{248}{113} & 0
\end{array}\right]. \tag{38}$$

5.2. **Implémentation.** Ces deux méthodes ont été implémentées en Matlab. La routine principale se trouve en annexe 1. Les deux méthodes précédentes ont été testées avec plusieurs systèmes d'équations différentielles, afin d'évaluer leurs performances.

5.2.1. *Un problème simple.* La première équation différentielle testée est la suivante :

$$y' = -y, \tag{39}$$

avec comme condition initiale $y(0) = 1$, système que l'on cherchera à intégrer sur l'intervalle de temps $[0, 1]$.

La solution *exacte* de (39) ($t \mapsto \mathrm{e}^{-t}$) est représentée sur la figure 1 (en annexe 2). Les figures 2-a,b,c montrent l'erreur entre la solution exacte et celle calculée à l'aide de la première méthode de Butcher implémentée, en fonction du pas de temps utilisé lors de l'intégration, cette erreur étant calculée respectivement en norme 1, norme 2 et norme infinie (ou norme sup). Enfin, la figure 3 montre l'évolution de la solution calculée lorsque l'on diminue le pas de temps h. La convergence vers la solution exacte lorsque $h \to 0$ apparaît clairement.

Les figures 4-a,b,c, et 5 sont les figures correspondantes dans le cas où l'on utilise la 2$^{\text{ème}}$ méthode implémentée. Là encore, la convergence est assez claire, et est similaire à celle de la première méthode (bien que l'ordre ne soit pas le même).

Dans la suite, seule la figure montrant l'erreur en fonction du pas de temps, calculée à l'aide de la norme infinie, sera montrée, les deux autres figures correspondant aux deux autres normes choisies pour mesurer l'erreur étant très largement similaires.

5.2.2. *Equation de Van der Pol.* L'équation de Van der Pol est

$$y'' + \varepsilon(y^2 - 1)y' + y = 0, \tag{40}$$

ce qui se ramène au système suivant :

$$\begin{array}{rcl} y_1' & = & y_2 \\ y_2' & = & \varepsilon(1 - y_1^2)y_2 - y_1, \quad \varepsilon > 0, \end{array} \tag{41}$$

avec les conditions initiales $y_1(0) = -0.5$ et $y_2(0) = 0$.

La solution *exacte* du système précédent est représentée sur la figure 6, pour le temps t variant de 0 à 20, et la solution devient rapidement cyclique.

La figure 7-a montre l'évolution de la solution calculée à l'aide de la méthode définie par (37) du système (41) lorsque h diminue, et la figure 7-b montre l'évolution de l'erreur entre ces solutions calculées et la solution exacte en fonction de h.

Les figures 8-a et b correspondent aux mêmes calculs pour la deuxième méthode (définie par (38)).

On constate une fois de plus qu'il y a peu d'écart entre les résultats obtenus pour l'une ou l'autre des deux méthodes testées, et que la convergence a bien lieu vers la solution exacte dans les deux cas.

5.2.3. *Attracteur étrange, modèle de Lorenz.* Le système des équations de Lorenz est le suivant :

$$\begin{aligned} y_1' &= -\sigma y_1 + \sigma y_2, \\ y_2' &= -y_1 y_3 + r y_1 - y_2, \\ y_3' &= y_1 y_2 - b y_3. \end{aligned} \tag{42}$$

Les valeurs des paramètres sont $b = \frac{8}{3}$, $\sigma = 10$ et $r = 28$. La condition initiale est $y_1(0) = -8$, $y_2(0) = 8$ et $y_3(0) = 27$.

La figure 9 montre l'allure de la solution *supposée exacte* (construite comm les autres solutions *exactes* à l'aide du programme `ode45` de Matlab, avec une tolérance relative de 10^{-12} et une tolérance absolue de 10^{-18}) pour t variant de 0 à 15. Néanmoins, nous nous limiterons à l'intervalle de temps $[0, 5]$ dans la suite pour tester les méthodes, le système de Lorenz étant trop chaotique pour espérer obtenir la solution exacte sans choisir le pas de temps h de manière exagérément petit.

Les figures 10-a et b montrent respectivement l'erreur en fonction du pas de temps pour la méthode n^{o}1, et l'évolution de la solution calculée lorsque h diminue (toujours pour la première méthode).

Les figures 11-a,b sont les figures analogues dans le cas de la méthode n2.

Pour la première méthode, on constate qu'il y a convergence de la solution calculée vers la solution (considérée comme) exacte, mais il faut prendre un pas de temps h de l'ordre de 10^{-2} au minimum si l'on veut une solution acceptable sur l'intervalle de temps $[0, 5]$. Pour des pas de temps plus grands, la solution ne peut être considérée comme juste (cas où le système oscille autour du bon état et non pas autour de l'autre) que pour des temps plus petits.

Avec la deuxième méthode, on prend conscience de la nature fortement non linéaire du système (42), puisque pour des pas de temps supérieurs à environ 4.10^{-2}, la solution produite est totalement fausse, et part à l'infini presque immédiatement (et n'est donc pas représentée sur les différentes courbes puisque sans aucun intérêt). Pour des pas de temps plus petits, la solution calculée ressemble beaucoup plus à celles obtenues par la première méthode pour des pas de temps comparables. La convergence vers la solution exacte a donc bien lieu, mais est moins apparente lorsque le pas de temps initial est choisi trop grand.

5.2.4. *Angles d'Euler.* Le système d'équations monodimensionnelles d'Euler est

$$\begin{aligned} I_1 y_1' &= (I_2 - I_3) y_2 y_3, \\ I_2 y_2' &= (I_3 - I_1) y_3 y_1, \\ I_3 y_3' &= (I_1 - I_2) y_1 y_2 + f(x), \end{aligned} \tag{43}$$

où l'on prend $I_1 = 0,5$, $I_2 = 2$, $I_3 = 3$, $y_1(0) = 1$, $y_2(0) = 0$, $y_3(0) = 0,9$ et

$$f(x) = \begin{cases} 0,25 \sin^2(x) & \text{si } 3\pi \leqslant x \leqslant 4\pi \\ 0 & \text{sinon} \end{cases} \tag{44}$$

La solution du système (43) est représentée sur la figure 12 où l'on voit l'évolution des 3 angles d'Euler pour t variant entre 0 et 20. On peut remarquer le changement dans les oscillations du troisième angle vers $t = 10$, ce qui correspond à l'influence du terme $f(x)$ dans (43).

La figure 13 montre l'évolution de l'erreur en fonction du pas de temps (toujours en norme sup) pour des solutions calculées à l'aide de la méthode (37). Il y a donc convergence vers la solution exacte lorsque h tend vers 0.

Les figures 14-a,b,c montrent l'évolution de la solution calculée respectivement pour $h = 1$, $h = 0.5$ et $h = 0.125$. Cette dernière est comparable avec la figure 12, et donc on peut pratiquement considérer qu'on a atteint la solution exacte dès $h = 0.125$.

Sur la figure 15, on peut voir l'évolution de l'erreur en fonction du pas de temps pour des solutions calculées avec la méthode (38). On constate qu'il y a convergence vers la solution exacte, mais comme dans le cas du système de Lorenz, uniquement à partir d'un certain rang, les solutions calculées pour des pas de temps de l'ordre de 1 étant totalement fausses. La figure 16-a montre la solution calculée pour $h = 0.5$ et on constate effectivement une divergence assez nette lorsque t croît. Pour h plus petit, comme le montre la figure 16-b ($h = 0.25$), la solution trouvée ressemble beaucoup plus à la solution exacte du problème, et on a des résultats semblables à ceux obtenus avec la première méthode.

5.2.5. *Un problème de mécanique céleste : les Pléïades.* Nous avons le système suivant, qui modélise les mouvements de 7 astres *livrés à eux-mêmes* :

$$\begin{array}{rcl} x_i'' & = & \sum_{j\neq i} m_j \frac{(x_j - x_i)}{r_{ij}}, \\ y_i'' & = & \sum_{j\neq i} m_j \frac{(y_j - y_i)}{r_{ij}}, \end{array} \tag{45}$$

pour i variant de 1 à 7, et avec $m_i = i$ et $r_{ij} = \left((x_i - x_j)^2 + (y_i - y_j)^2\right)^{\frac{3}{2}}$.
Les conditions initiales sont $x(0) = (3;3;-1;-3;2;-2;2)$, $y(0) = (3;-3;2;0;0;-4;4)$, $x'(0) = (0;0;0;0;0;1,75;-1,5)$ et $y'(0) = (0;0;0;-1,25;1;0;0)$.

La solution que l'on considèrera comme exacte (calculée en précisions relative 10^{-8} et absolue 10^{-12}, toujours à l'aide du solveur `ode45` de Matlab) est représentée sur la figure 17-a. On constate qu'il y a un certain nombre de *collisions* entre les différentes *planètes*. La figure 17-b montre la solution calculée à l'aide de la première méthode implémentée, mais pour un temps t variant dans un intervalle de temps beaucoup plus restreint, $t \in [0, 1.2]$ au lieu de $t \in [0, 3]$. En effet, il n'a pas été possible d'intégrer de manière satisafaisante le système (45) à partir de l'instant $t = 1.2$, temps auquel survient la première *collision* entre deux planètes, ce qui crée une très nette complication du problème (certains termes r_{ij} tendant vers 0), empêchant même la résolution du système par le programme `ode45` avec des précisions aussi fines que pour les cas précédents.

5.2.6. *Hanging rope.* Un dernier problème à grand nombre de degrés de liberté est le problème de la corde pendante (hanging rope). Il s'agit de simuler le mouvement d'une corde, initialement au repos, soumise uniquement à son poids. Le système d'équations s'écrit sous la forme :

$$\begin{array}{rcl} \displaystyle\sum_{k=1}^{n} a_{lk}\ddot{\theta}_k & = & \displaystyle -\sum_{k=1}^{n} b_{lk}\dot{\theta}_k^2 - n\left(n + \frac{1}{2} - l\right)\sin\theta_l \\ & & -n^2 \sin\theta_l F_x(t) + \left\{ \begin{array}{ll} n^2\cos\theta_l F_y(t) & \text{si } l \leqslant \frac{3n}{4} \\ 0 & \text{si } l > \frac{3n}{4}, \end{array} \right. \qquad l = 1\ldots n, \end{array} \tag{46}$$

avec $a_{lk} = g_{lk}\cos(\theta_l - \theta_k)$, $b_{lk} = g_{lk}\sin(\theta_l - \theta_k)$ et $g_{lk} = n + \frac{1}{2} - \max(l, k)$. La condition initiale est $\theta_l(0) = \dot{\theta}_l(0) = 0$, et n est le nombre de degrés de liberté souhaité.

La figure 18-a montre la solution du système précédent, pour $n = 10$ paramètres, obtenue avec le même solveur que pour le problème des Pleiades, pour un temps t variant entre 0 et 1. L'intégration de ce système par cette méthode s'est avérée pratiquement impossible pour des temps plus grands, ou pour un nombre plus élevé de degrés de liberté.

La solution obtenue à l'aide de la première méthode de Butcher implémentée, toujours pour $n = 10$ degrés de liberté, est représentée sur la figure 18-b, pour un pas de temps $h = 0.1$. Les deux méthodes étudiées ici n'ont pas permis l'intégration du système (46) pour un pas de temps vraiment plus petit, ni sur une période temporelle plus grande.

6. Conclusion

On constate que les deux méthodes considérées sont sensiblement équivalentes, mis à part de très mauvaises solutions calculées lorsque le pas de temps est trop grossier.

Au niveau des différents problèmes testés, on note (de façon assez rassurante) une très bonne convergence des solutions calculées vers la solution exacte pour les cas linéaires, mais aussi pour le problème de Van der Pol, l'attracteur de Lorenz, et les angles d'Euler. En revanche, les méthodes étudiées atteignent leurs limites lorsqu'on les teste sur des problèmes à nombreux degrés de liberté (comme celui de la corde pendante), mais aussi sur des problèmes très raides, comme celui des Pleiades.

Il faut aussi noter que la convergence de la $2^{\text{ème}}$ méthode n'est souvent pas plus rapide que celle de la $1^{\text{ère}}$ méthode, alors que l'ordre est 4 au lieu de 2. Comme les matrices considérées lors des calculs sont (deux fois) plus grandes avec la deuxième méthode, les calculs sont relativement plus longs, et, pour les problèmes considérés, on peut se demander si celà vaut le coup d'utiliser cette méthode (n'amenant pas forcément une meilleure convergence vers la solution exacte).

En théorie, plus s est grand, plus les approximations construites sont précises, puisqu'on dispose des $\frac{h^k}{k!}y^{(k)}(x_i)$ pour un plus grand nombre d'entiers k, et donc, comme pour des formules de Taylor, $y(x_{i+1})$ devrait pouvoir être connu avec plus de précision, et donc l'ordre est plus élevé.

En pratique, les approximations contiennent des erreurs de calcul, et disposer de plus de termes (contenant donc globalement plus d'erreurs) n'est peut-être pas forcément une bonne chose.

Il semble donc judicieux de tester plusieurs méthodes sur le problème que l'on considère, afin de choisir celle qui se révèle en pratique la plus efficace.

Références

[But] J.C. Butcher, *General linear methods for stiff differential equations*, October 14, 1999.

[CM84] M. Crouzeix et A.L. Mignot, *Analyse numérique des équations différentielles*, Masson, 1984.

[HW91] E. Hairer et G. Wanner, *Solving ordinary differential equations ii*, Springer-Verlag, 1991.

Annexe 1 - Programme principal

```
function [t,y] = main(pb,niter,deltat,schema)

%
% function [t,y] = main(pb,niter,deltat,schema)
%
% Intègre le système d'edp contenu dans 'pb',
% en effectuant 'niter' pas de temps de taille 'deltat',
% à l'aide du schéma 'schema' de Butcher.
% Output : 't' = temps, 'y' = solution du système.
%

%
% Paramètres de la méthode (matrices A,B,U,V, scalaires r,s)
%

if schema == 1     % Schéma défini par 37
 AA = ⟨ ... ⟩ ; BB = ⟨ ... ⟩ ; UU = ⟨ ... ⟩ ; VV = ⟨ ... ⟩ ;
else               % Schéma défini par 38
 AA = ⟨ ... ⟩ ; BB = ⟨ ... ⟩ ; UU = ⟨ ... ⟩ ; VV = ⟨ ... ⟩ ;
end

[r,s] = size(BB) ;

%
% Définition du problème (dimension, conditions initiales ... )
%

n = niter ;
h = deltat ;

if pb == 1         % — essai — (problème test linéaire)
 f = 'essai' ;  dim = 1 ;  y0 = [1] ;
elseif pb == 2     % — vdp — (Van der Pol)
 f = 'vdp' ;  dim = 2 ;  y0 = [-0.5 ;0] ;
elseif pb == 3     % — lorenz —
 f = 'lorenz' ;  dim = 3 ;  y0 = [-8 ;8 ;27] ;
elseif pb == 4     % — euler —
 f = 'euler' ;  dim = 3 ;  y0 = [1 ;0 ;0.9] ;
elseif pb == 5     % — pleiades —
 f = 'pleiades' ;  dim = 28 ;
 y0=[3 ;0 ;3 ;0 ;3 ;0 ;-3 ;0 ;-1 ;0 ;2 ;0 ;-3 ;0 ;0 ;-1.25 ;
     2 ;0 ;0 ;1 ;-2 ;1.75 ;-4 ;0 ;2 ;-1.5 ;4 ;0] ;
elseif pb == 6     % — hanging rope —
 f = 'rope' ;  dim = 2*10 ;  y0 = zeros(dim,1) ;
end

%
% Construction de y^[0]
%
% A partir de la CI y0, on construit le 1er vecteur de la méthode y00
% en rajoutant des estimations de h y'(x), (h^2/2) y''(x) ...
%
```

```
y00 = zeros(r*dim,1);              % Initialisation au vecteur nul
for j = 1 :dim
 y00((j-1)*r+1) = y0(j);           % On part de y0
end
ref = 1;
for i = 1 :r-1                     % Derivée i^ème
 ystart = zeros(s*dim,1);
 ref = ref*h/i;
 for j = 1 :dim
  ystart((j-1)*s+1) = y00((j-1)*r+i);
 end
 ystart = feval(f,ystart);
 for j = 1 :dim
  y00((j-1)*r+i+1,1) = ystart((j-1)*s+1)*ref;
 end
end

%
% Matrices tensorisées
%

A = kron(eye(dim),AA); B = kron(eye(dim),BB);
U = kron(eye(dim),UU); V = kron(eye(dim),VV);

%
% Initialisation à la condition initiale
%

t = zeros(n,1);
y = zeros(r*dim,n+1);
y( :,1) = y00;

%
% Point de départ du schéma élémentaire
%

Y( :,1) = U*y00;

%
% Boucle des calculs
%

for k = 1 :n
 ylast = y( :,k);

% Résolution du schéma élémentaire, 'stage' étant la fonction
% de Y à annuler pour résoudre le schéma élémentaire.

 [Y( :,k+1),zero,exitflag] = fsolve('stage',Y( :,k));

% Calcul de la solution au temps t + δ t

 F = feval(f,Y( :,k+1));
 y( :,k+1) = h*(B*F)+V*y( :,k);
 t(k) = (k-1)*h;
```

```
end

⟨Sauvegarde de la solution⟩

%
% Tracé de la solution
%

if pb == 1
 % — essai —
 plot(t,y(1,1 :n))
 ⟨Réglage du titre, sous-titre, légende ... ⟩
elseif pb == 2
 % — vdp —
 plot(y(1, :),y(1+r, :))
 ⟨Réglage du titre, sous-titre, légende ... ⟩
elseif pb == 3
 % — lorenz —
 plot(t,y(1,1 :n),t,y(1+r,1 :n),t,y(1+2*r,1 :n))
 ⟨Réglage du titre, sous-titre, légende ... ⟩
elseif pb == 4
 % — euler —
 plot(t,y(1,1 :n),t,y(1+r,1 :n),t,y(1+2*r,1 :n))
 ⟨Réglage du titre, sous-titre, légende ... ⟩
elseif pb == 5
 % — pleiades —
 plot(y(1, :),y(1+2*r, :),y(1+4*r, :),y(1+6*r, :),y(1+8*r, :),y(1+10*r, :),
      y(1+12*r, :),y(1+14*r, :),y(1+16*r, :),y(1+18*r, :),y(1+20*r, :),
      y(1+22*r, :),y(1+24*r, :),y(1+26*r, :))
 ⟨Réglage du titre, sous-titre, légende ... ⟩
elseif pb == 6
 % — hanging rope —
 for k = 1 :n+1
  y1(1,k) = 0 ;
  y2(1,k) = 0 ;
 % passage en coordonnées cartésiennes à partir des angles
  for j = 1 :dim/2
   y1(j+1,k) = y1(j,k)+cos(y((j-1)*2*r+1,k)) ;
   y2(j+1,k) = y2(j,k)+sin(y((j-1)*2*r+1,k)) ;
  end
 end
 tt = kron(t,ones(dim/2+1,1)) ;
 plot3(tt,y1,y2)
 ⟨Réglage du titre, sous-titre, légende ... ⟩
end
```

Annexe 2 - Figures

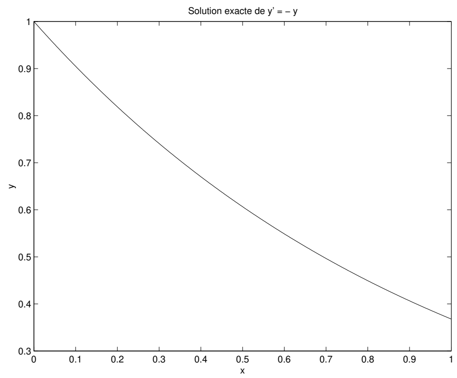
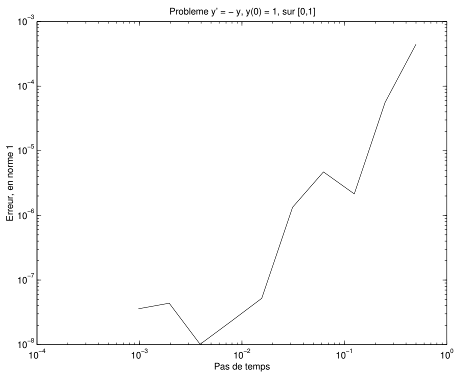
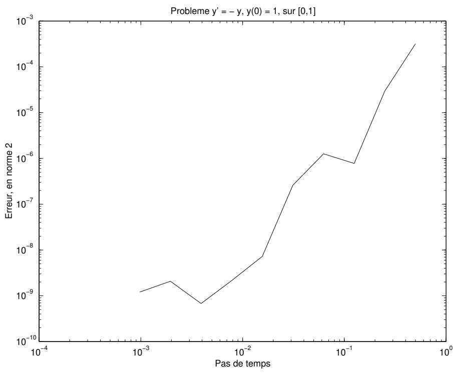


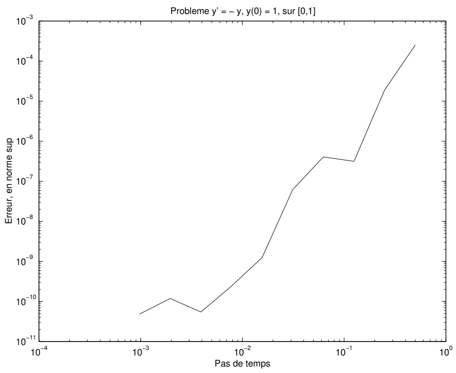
Fig. 1. Solution exacte du problème-test linéaire (39).



(a)



(b)



(c)

Fig. 2. Différence entre la solution exacte et la solution approchée (méthode 1) en fonction du pas de temps, mesurée en norme 1 (a), 2 (b) et infinie (c).

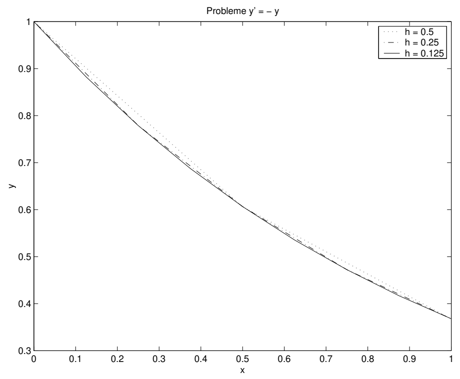


FIG. 3. Evolution de la solution approchée (méthode 1) lorsque h tend vers 0.

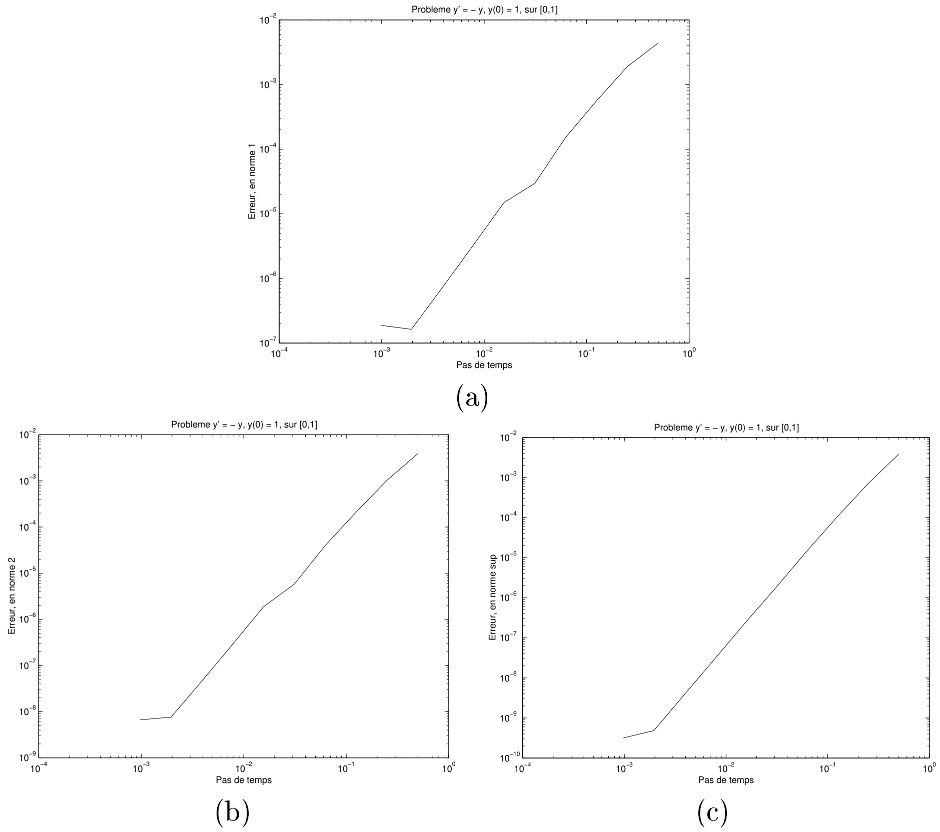


FIG. 4. Différence entre la solution exacte et la solution approchée (méthode 2) en fonction du pas de temps, mesurée en norme 1 (a), 2 (b) et infinie (c).

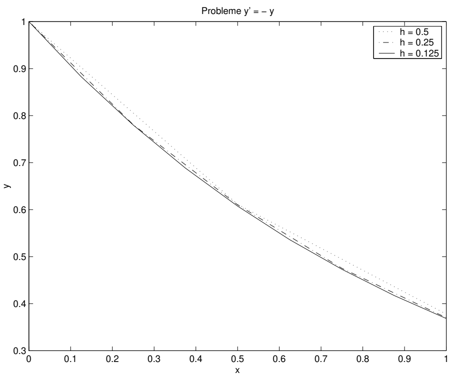


FIG. 5. Evolution de la solution approchée (méthode 2) lorsque h tend vers 0.

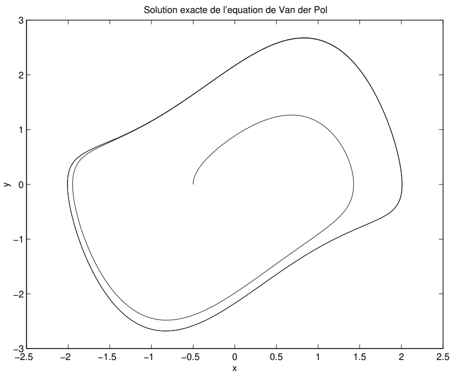
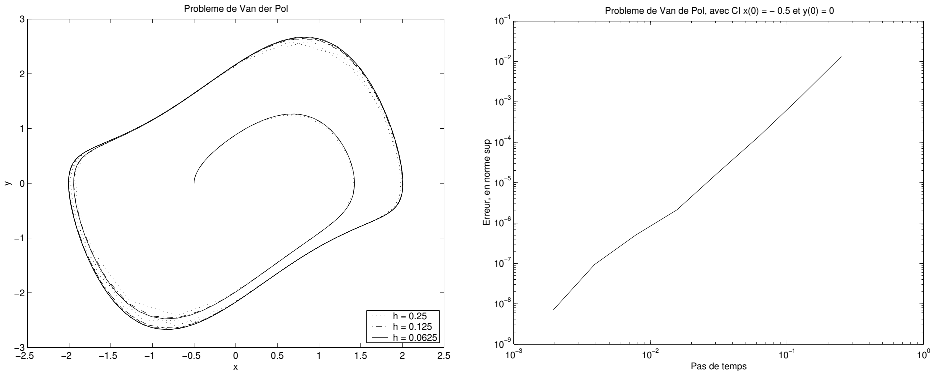
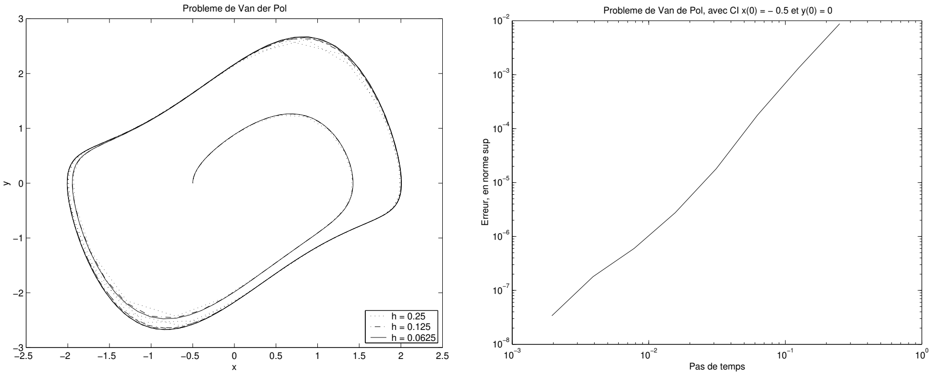


FIG. 6. Solution exacte de l'équation de Van der Pol (41).



(a) (b)

FIG. 7. Evolution de la solution approchée (méthode 1) (a) et de son écart (en norme infinie) avec la solution exacte (b) en fonction du pas de temps.



(a) (b)

Fig. 8. Evolution de la solution approchée (méthode 2) (a) et de son écart (en norme infinie) avec la solution exacte (b) en fonction du pas de temps.

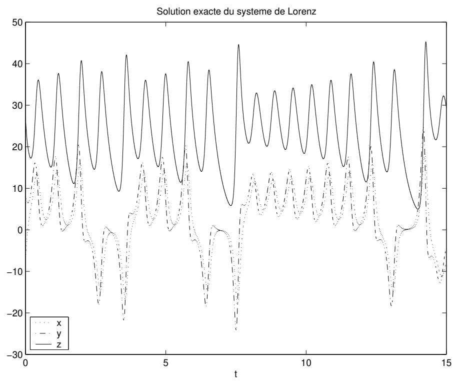
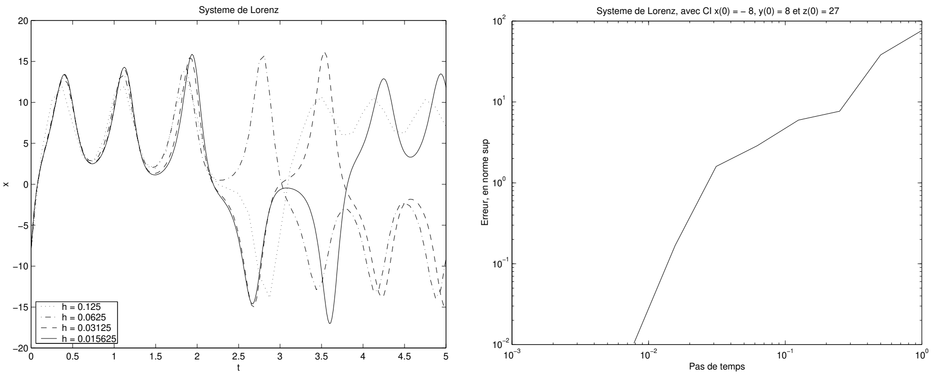
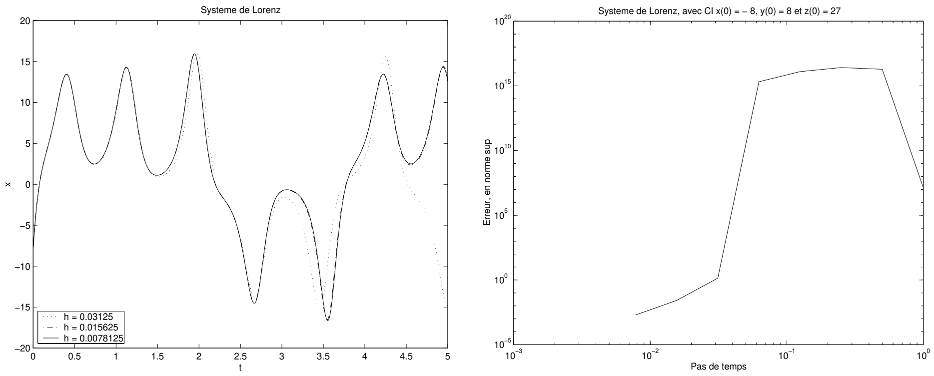


Fig. 9. Solution exacte du système de Lorenz (42).



(a) (b)

Fig. 10. Evolution de la solution approchée (méthode 1) (a) et de son écart (en norme infinie) avec la solution exacte (b) en fonction du pas de temps.



(a) (b)

FIG. 11. Evolution de la solution approchée (méthode 2) (a) et de son écart (en norme infinie) avec la solution exacte (b) en fonction du pas de temps.

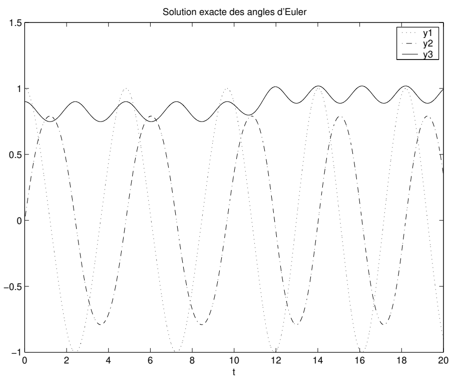


FIG. 12. Solution exacte du problème des angles d'Euler (43).

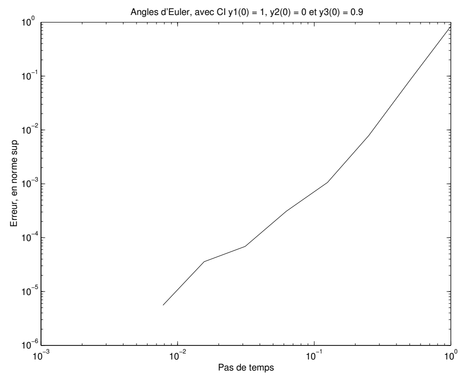


FIG. 13. Différence entre la solution exacte et la solution approchée (méthode 1) en fonction du pas de temps.

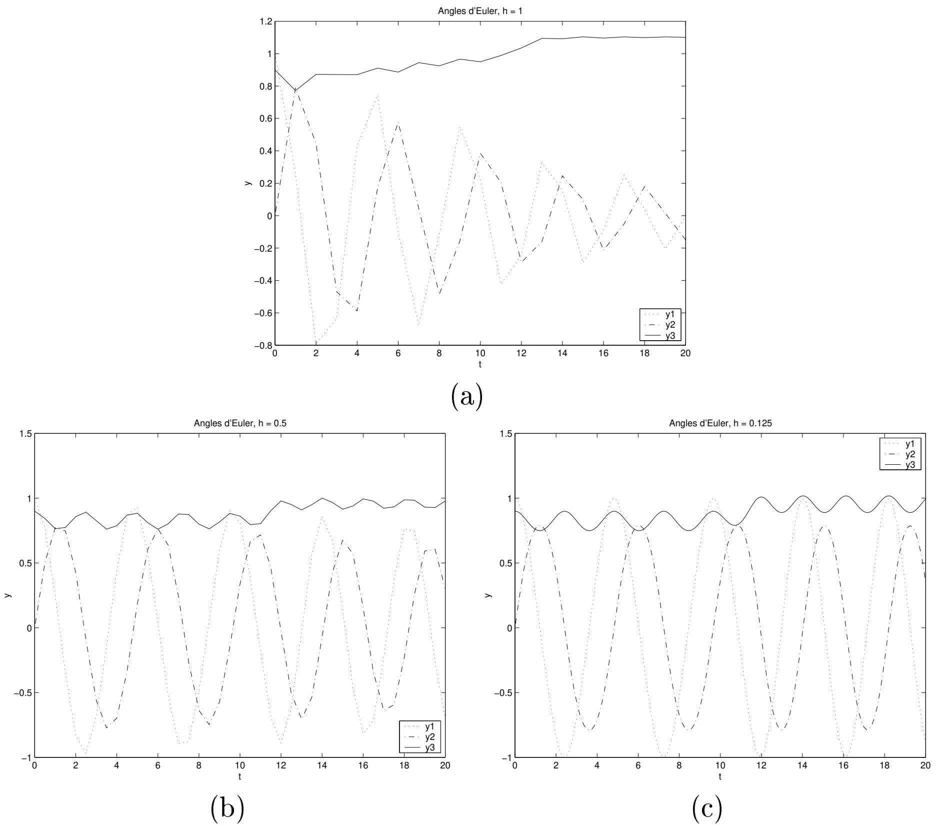


FIG. 14. Solution calculée à l'aide de la méthode 1 pour $h = 1$ (a), $h = 0.5$ (b) et $h = 0.125$ (c).

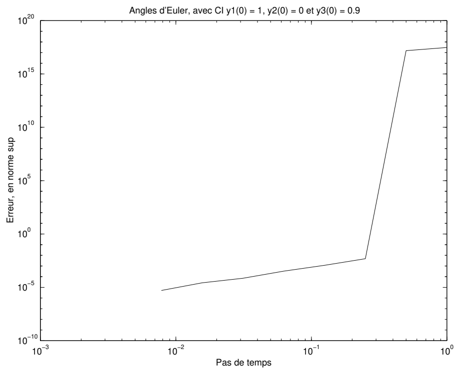
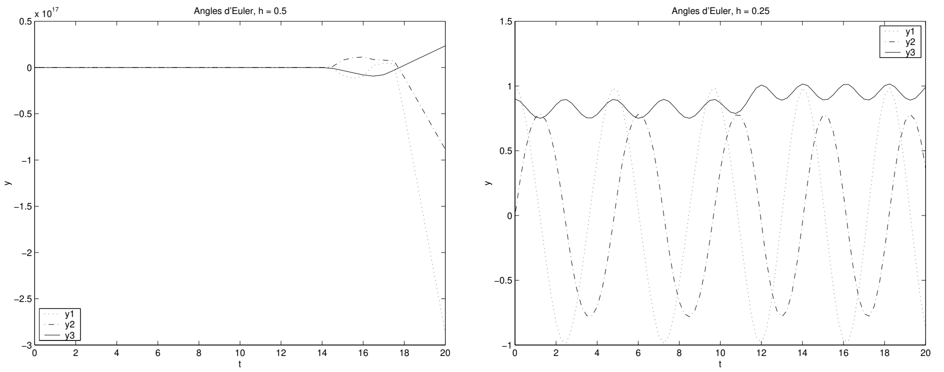
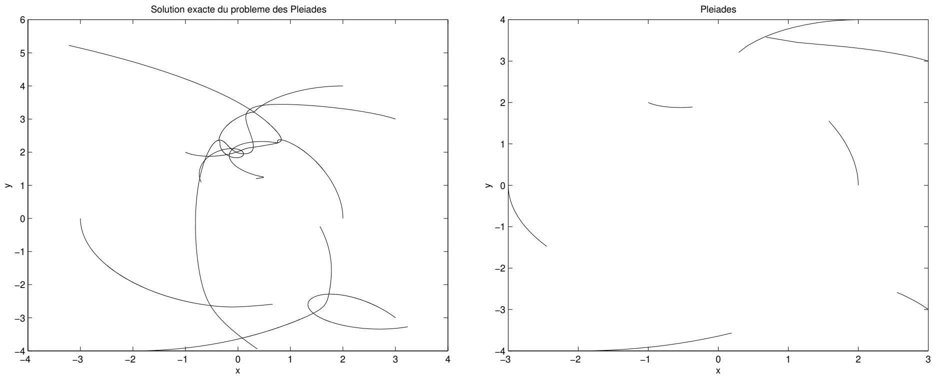


FIG. 15. Différence entre la solution exacte et la solution approchée (méthode 2) en fonction du pas de temps.



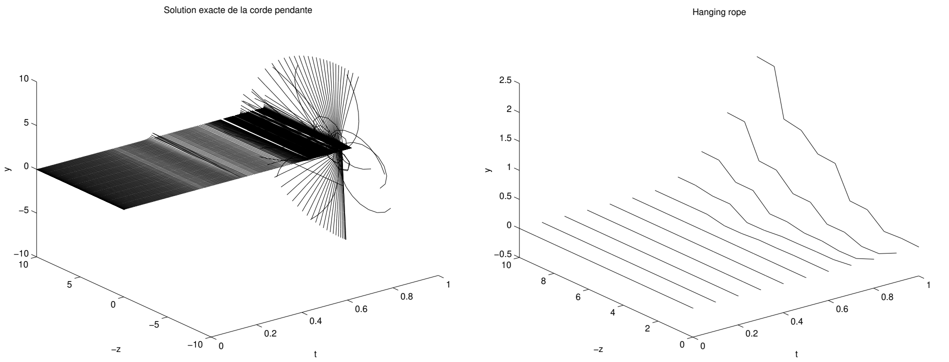
(a) (b)

FIG. 16. Solution calculée à l'aide de la méthode 2 pour $h = 0.5$ (a) et $h = 0.25$ (b).



(a) (b)

FIG. 17. Solutions *exacte* (a) et approchée (b) (méthode 1, sur un intervalle de temps plus court) du problème des Pleiades (45).



(a) (b)

FIG. 18. Solutions *exacte* (a) et approchée (b) (méthode 1, pour un pas de temps grossier) du problème de la corde pendante (46).