HABILITATION À DIRIGER DES RECHERCHES

présentée devant

L'Université de Rennes 1
Institut Mathématique de Rennes

par

Philippe CHARTIER [1]

Méthodes numériques pour les équations différentielles ordinaires et algébriques avec application aux systèmes hamiltoniens

soutenue le 21 Janvier 2000 devant le jury composé de

MM	J. Della Dora, Pr. Institut National Polytechnique de Grenoble	Rapporteur
	E. Hairer, Pr. Université de Genève	Rapporteur
	Ch. Lubich, Pr. University of Tübingen	Rapporteur
	F. Bonnans, DR INRIA et MdC. Ecole Polytechnique	Examinateur
	G. Caloz, Pr. Université de Rennes I	Examinateur
	M. Crouzeix, Pr. Université de Rennes I	Examinateur
	M. Pierre, Pr. Ecole Normale Supérieure de Cachan	Examinateur

1. INRIA-Rennes, Campus de Beaulieu, 35042 Rennes cedex, Email : Philippe.Chartier@inria.fr

À Léo, Owen, et Ghislaine,

Remerciements

Je tiens à remercier vivement Michel Crouzeix dont l'aide et les encouragements m'ont été précieux dans l'élaboration de ce document.

Je tiens à exprimer ma sincère gratitude à Jean Della Dora, Ernst Hairer et Christian Lubich pour avoir accepté la charge de rapporteur. Je souhaite en particulier les remercier de l'intérêt qu'ils ont bien voulu porter aux travaux présentés dans cette thèse.

Je voudrais également exprimer ma reconnaissance à Frédéric Bonnans et Michel Pierre pour leur participation au jury et à Gabriel Caloz pour la présidence qu'il a bien voulu exercer le temps de ce jury.

Enfin, je désire remercier très chaleureusement Jocelyne Erhel et Bernard Philippe pour leur sympathie, leur patience et leur aide.

Table des matières

Chapitre 1

Cadre général

Les équations différentielles ordinaires et les équations différentielles algébriques apparaissent dans un nombre important d'applications liées à des disciplines variées de la physique ou de la chimie (par exemple). Elles forment un cadre naturel au sein duquel nombre de systèmes complexes peuvent être modélisés.

En complément des équations purement différentielles, les modèles nécessitent fréquemment l'adjonction d'équations algébriques, destinées à rendre compte des lois de conservation, des contraintes cinétiques ou géométriques, des lois de Kirchhoff, etc.

Les équations différentielles ordinaires et algébriques interviennent typiquement dans les situations suivantes :

- modélisation de systèmes mécaniques à plusieurs corps (robotique),
- simulation de réseaux électriques et/ou électroniques,
- modélisation de processus chimiques (cinétique),
- discrétisation (en espace) d'équations aux dérivées partielles,
- contrôle optimal,
- mécanique du point (calcul de trajectoire), ...

La nécessité de concevoir, analyser et développer de nouveaux schémas numériques naît des constatations suivantes :

- les modèles sont de taille croissante et l'obtention d'une solution numérique de précision acceptable requiert donc une grande puissance de calcul. Face à ces exigences, et suivant une tendance observée bien plus précocement dans d'autres domaines, le parallélisme s'impose comme un élément de réponse.

Par ailleurs, des gains de vitesse d'exécution peuvent résulter d'une accélération de la méthode de Newton, qui est au coeur de tous les solveurs implicites ou plus simplement d'un schéma plus efficace.

- la présence de contraintes introduit des difficultés numériques nouvelles, non observables dans le cas des équations différentielles ordinaires. L'étude de ces phénomènes a été initiée au début des années 1980 et se poursuit aujourd'hui avec le développement de méthodes numériques adaptées.
- il semble parfois désastreux, aux yeux du physicien par exemple, qu'une méthode numérique, même précise, fournisse une solution d'énergie non constante, alors que la théorie atteste du contraire. Que le comportement d'une solution numérique puisse différer qualitativement à ce point de la solution exacte a motivé la construction de schémas préservant certains invariants du problème. C'est le cas des méthodes symplectiques ou symétriques, qui conservent la symplecticité ou la symétrie du flot, et dans une moindre mesure des méthodes pseudo-symplectiques ou pseudo-symétriques, qui conservent "approximativement" la symplecticité ou la symétrie.

Ces points sont abordés à des degrés divers dans cette thèse.

Chapitre 2

Quelques propriétés géométriques des systèmes différentiels

Ce chapitre est destiné à introduire les systèmes différentiels avec invariant et/ou dont le flot est symétrique ou symplectique. Les premiers paragraphes sont consacrés aux invariants linéaires et quadratiques. Les systèmes conservant le volume ou les équations sur les groupes de Lie sont ensuite évoqués brièvement. Tous ces systèmes ont fait et font encore l'objet de recherches qui sont sommairement décrites et référencées. Mais on s'intéresse surtout dans cette partie aux systèmes hamiltoniens ou réversibles en temps. C'est à ces systèmes que les deux chapitres suivants sont consacrés.

2.1 Introduction et notations

On considère le système différentiel[1]

$$y'(x) = f(y(x)), \tag{2.1}$$

où y est un vecteur de $\mathbb{R}^m$ et f une fonction de $\mathbb{R}^m$ à valeurs dans $\mathbb{R}^m$.

Pour une *valeur initiale* donnée y_0, il peut exister zéro, une ou plusieurs solutions $y(x)$ vérifiant (2.1) et la *condition initiale* $y(x_0) = y_0$. Ces solutions peuvent être locales ou s'étendre à tout un intervalle de la forme $[x_0,X]$ ou $[x_0, +\infty[$.

1. Dans ce chapitre et les deux suivants, on désigne les composantes des vecteurs y et f de (2.1) par un indice en lettre minuscule ou chiffre, ou par un exposant en lettre capitale lorsqu'il y a risque de confusion.

Diverses conditions d'existence et d'unicité existent dans la littérature des équations différentielles ordinaires (Hartman [Har64], Arnold [Arn74]) ou dans celle de l'analyse numérique de ces équations (Crouzeix-Raviart [CR80], Crouzeix-Mignot [CM89], Butcher [But87], Hairer-Nørsett-Wanner [HNW93]). Cependant, dans le cadre de l'intégration numérique, l'unicité est essentielle et on se contente donc ici de rappeler le théorème suivant :

Théorème 1 *On considère le système (2.1) sur un intervalle I_0 de la forme $[x_0,X]$ ou $[x_0,X[$ et on suppose que f est continue sur $\mathbb{R}^m$ et qu'il existe une constante réelle l telle que*

$$\forall\, y,z \in \mathbb{R}^m, \; < f(y) - f(z), y - z > \leq l |y-z|^2,$$

où $< .,. >$ est un produit scalaire sur $\mathbb{R}^m$ et $|\,.\,|$ la norme associée. Alors le système admet une solution unique sur I_0.

On suppose par la suite que f est de classe C^k, pour $k \geq 1$ suffisamment grand (de sorte que (2.1) possède une solution unique).

2.2 Définition d'un invariant

Définition 1 *Un invariant I (encore appelé* intégrale première*) du système (2.1) est une fonction définie sur $\mathbb{R}^m$ à valeurs dans $\mathbb{R}$ de classe C^1 telle que*

$$\forall\, y \in \mathbb{R}^m, (\nabla_y I)^T f(y) = 0.$$

Il résulte de la définition que pour toute solution $y(x)$ de (2.1) on a :

$$I(y(x)) = I(y(x_0)) = \text{const.}$$

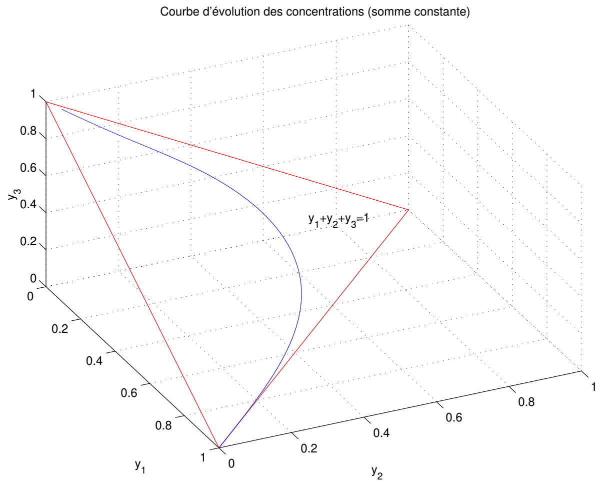
Étant donnée une méthode d'intégration quelconque, il est possible de définir une étape de projection sur l'invariant $I(y(x)) = \text{const}$ sans modifier l'ordre de convergence de la méthode. Cette technique est par exemple décrite dans [HW96], Section VII. 2.

2.2.1 Invariants linéaires et quadratiques

La définition d'un invariant linéaire ou quadratique semblant évidente, on se contente ici d'en donner deux exemples :

Exemple 1 (Cinétique chimique) *On considère les réactions suivantes :*

$$\begin{array}{rcl} A & \xrightarrow{k_1} & B \\ B+B & \xrightarrow{k_2} & B+C \\ B+C & \xrightarrow{k_3} & A+C \end{array}$$



Si on désigne respectivement par y_1, y_2 et y_3 les concentrations respectivement en A, B et C, l'évolution du composé est régie par les équations :

$$\left\{\begin{array}{lcl} y_1' & = & -k_1y_1 + k_3y_2y_3 \\ y_2' & = & k_1y_1 - k_3y_2y_3 - k_2y_2^2 \\ y_3' & = & k_2y_2^2 \end{array}\right. \tag{2.2}$$

Dans l'exemple donné par Robertson (voir [Rob66]), les conditions initiales sont prises égales à

$$y_1(0) = 1, y_2(0) = 0, y_3(0) = 0,$$

et $k_1 = 0.04$, $k_2 = 3.10^7$, $k_3 = 10^4$. Pour ces valeurs des constantes, le système est raide. Ici, afin de visualiser l'invariant, on prend $k_1 = k_2 = k_3 = 1$. Il est alors aisé de vérifier que $I(y) = y_1 + y_2 + y_3$ (traduction de la conservation de la masse) est un invariant pour (2.2). Le long de toute solution exacte $y(x)$ de (2.2), on a $I(y(x)) = 1$ (voir figure 1).

Exemple 2 (Mécanique du solide) *On considère le mouvement d'un solide autour de son centre de gravité. Si $\mathcal{I}_1$, $\mathcal{I}_2$ et $\mathcal{I}_3$ désignent ses moments d'inertie par rapport aux trois axes principaux et y_1, y_2, y_3 les coordonnées du moment angulaire, les équations d'Euler-Lagrange s'écrivent :*

$$\begin{bmatrix} y_1' \\ y_2' \\ y_3' \end{bmatrix} = \begin{bmatrix} 0 & \frac{y_3}{\mathcal{I}_3} & -\frac{y_2}{\mathcal{I}_2} \\ -\frac{y_3}{\mathcal{I}_3} & 0 & \frac{y_1}{\mathcal{I}_1} \\ \frac{y_2}{\mathcal{I}_2} & -\frac{y_1}{\mathcal{I}_1} & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}. \tag{2.3}$$

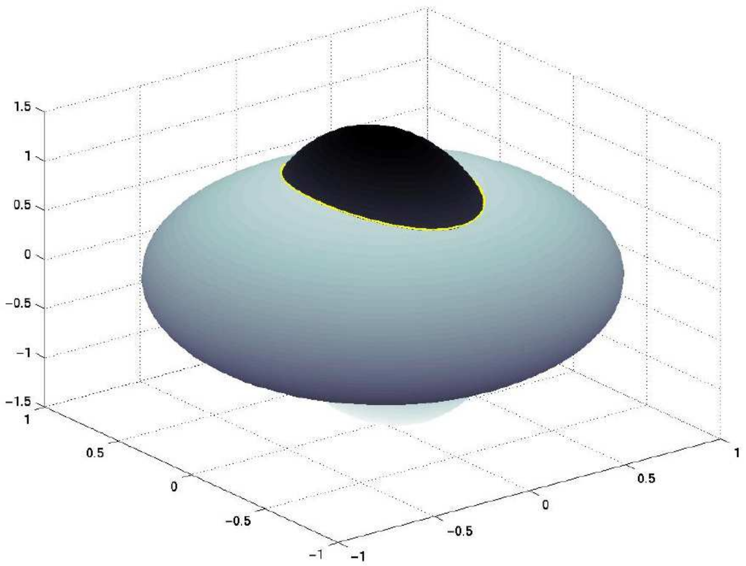


FIG. 2.1 – *Courbe solution (en jaune), sphère et ellipsoïde des invariants*

On peut alors vérifier (voir figure 2) que $I_1(y) = y_1^2 + y_2^2 + y_3^2$ *(énergie cinétique), d'une part, et* $I_2(y) = (\frac{y_1}{\mathcal{I}_1})^2 + (\frac{y_2}{\mathcal{I}_2})^2 + (\frac{y_3}{\mathcal{I}_3})^2$*, d'autre part, sont des invariants du système (2.3).*

Toutes les méthodes d'intégration usuelles conservent les invariants linéaires. C'est en particulier le cas des méthodes de Runge-Kutta (voir Hairer [Hai99]). Les méthodes de Runge-Kutta symplectiques (voir chapitre suivant) conservent en outre les invariants quadratiques (sans qu'il soit nécessaire de recourir à une étape de projection). Pour une démonstration de ce résultat, on peut consulter par exemple Cooper [Coo87].

2.2.2 Préservation du volume

Certains systèmes ont la propriété de conserver tout volume U le long des solutions exactes. Pour donner une définition précise de la conservation du volume, on introduit la fonction *flot*, ψ_x, associée au système (2.1) :

$$\begin{array}{cccc} \psi_x & : & I\!R^m & \rightarrow & I\!R^m \\ & & y_0 & \mapsto & y_x \end{array}$$

telle que y_x soit la solution exacte $y(x)$ de (2.1) au point x avec comme condition initiale $y(0) = y_0$.

Théorème 2 *Supposons que (2.1) vérifie la condition suivante :*

$$Tr(\frac{\partial f}{\partial y}) = \sum_{i=1}^{m} \frac{\partial f_i}{\partial y_i} = 0. \tag{2.4}$$

Alors le flot associé à (2.1) conserve les volumes.

Exemple 3 *Soit le système*

$$\left\{ \begin{array}{l} y_1' = y_2, \\ y_2' = -y_1 \end{array} \right. .$$

La fonction flot s'écrit ici

$$\psi_x(y^0) = \left(\begin{array}{c} y_1^0 \cos(x) + y_2^0 \sin(x) \\ -y_1^0 \sin(x) + y_2^0 \cos(x) \end{array} \right).$$

On peut alors vérifier que

$$\begin{aligned} \forall U \subset I\!R^m, \textit{Vol}(\psi_x(U)) &= \int_{\psi_x(U)} dy = \int_U \left| \det \frac{\partial \psi_x}{\partial y}(y^0) \right| dy^0 \\ &= \int_U \left| \begin{array}{cc} \cos(x) & \sin(x) \\ -\sin(x) & \cos(x) \end{array} \right| dy^0 = \int_U dy^0 = \textit{Vol}(U). \end{aligned}$$

On n'abordera pas dans cette thèse le problème de la conservation des volumes par des méthodes numériques. On peut cependant noter qu'en dimension 2, le flot associé à un système différentiel préserve les volumes si et seulement si le système est *hamiltonien* (voir paragraphes suivants). Les méthodes *symplectiques* (voir chapitre 2.1) conservent alors le volume. En dimension supérieure, les méthodes dites de "splitting" et de "correction" permettent également de préserver les volumes. Elles sont exposées dans Feng Kang & Shang Zai-jiu [KZj95], Quispel [Qui95] et Quispel & Dyt [QD98].

2.2.3 Équations différentielles sur des groupes de Lie

Définition 2 *Un groupe de Lie est un groupe G muni d'une structure de variété différentielle et tel que le produit sur G soit différentiable. De plus, soit I l'élément neutre de G : l'espace tangent à G en I, noté $T_I G$ est appelé* algèbre de Lie.

Remarque 1 *On peut montrer que $T_I G$ est naturellement muni d'une structure d'algèbre, d'où son nom.*

Exemple 4 *Le groupe des matrices orthogonales*

$$G = \{M \in \mathbb{R}^{m\times m}, M^T M = I\}$$

est un groupe de Lie de dimension $m(m-1)/2$ (c'est une sous-variété de $\mathbb{R}^{m\times m}$). Son algèbre de Lie associée est l'espace vectoriel des matrices antisymétriques

$$T_I G = \{M \in \mathbb{R}^{m\times m}, M + M^T = 0\}.$$

Soit alors le système différentiel

$$Y' = B(Y)Y, \tag{2.5}$$

où l'on suppose que pour tout $Y \in G$, $B(Y) \in T_I G$. Si $Y(0) \in G$ alors $Y(x) \in G$ pour tout x. La solution $Y(x)$ évolue sur le groupe de Lie G et satisfait donc les contraintes définissant la variété attachée à G; dans l'exemple ci-dessus, on a ainsi $\sum_k Y_{i,k} Y_{k,j} = \delta_{i,j}$ pour tout i et tout j.

Récemment, de nombreuses recherches ont été engagées sur les méthodes de Lie et les méthodes de Runge-Kutta-Lie en particulier (voir Iserles [Ise84], Crouch & Grossman [CR93], Iserles & Nørsett [IN97], Zanna [Zan97] ou Munthe-Kass [MK]). L'idée inhérente à ces méthodes consiste à récrire l'équation (2.5) dans l'algèbre de Lie au moyen de l'exponentielle (exp est en effet un difféomorphisme de $T_I G$ sur G). L'intérêt majeur de ce changement de variable est que les invariants attachés à cette nouvelle formulation sont linéaires et donc naturellement conservés par toute méthode d'intégration. Son inconvénient majeur tient aux calculs liés à exp.

2.3 Propriétés fondamentales des systèmes hamiltoniens ou réversibles en temps

2.3.1 Systèmes hamiltoniens

On s'intéresse dans ce paragraphe aux systèmes de la forme

$$\left\{ \begin{array}{rcrl} p' & = & - & \nabla_q H, \\ q' & = & & \nabla_p H \end{array} \right. \tag{2.6}$$

où

$$\nabla_p H = \left[\frac{\partial H}{\partial p^1}, \dots, \frac{\partial H}{\partial p^d}\right]^T \text{ et } \nabla_q H = \left[\frac{\partial H}{\partial q^1}, \dots, \frac{\partial H}{\partial q^d}\right]^T,$$

H étant une fonction de $\mathbb{R}^{2d}$ dans $\mathbb{R}$ suffisamment régulière. Les systèmes de la forme (2.6) sont évidemment de la forme (2.1) avec $m = 2d$ mais possèdent en outre certaines propriétés caractéristiques qu'il est parfois essentiel de conserver dans les applications à la physique par exemple. La propriété la plus élémentaire est la conservation de l'hamiltonien le long des solutions exactes de (2.6) :

Théorème 3 *Soit* $(p(x),q(x))$ *une solution exacte du système hamiltonien (2.6) avec comme condition initiale* $(p(x_0),q(x_0)) = (p_0,q_0)$. *Alors on a*

$$\forall\, x \geq x_0,\ H(p(x),q(x)) = H(p_0,q_0).$$

Remarque 2 *Autrement dit,* H *est un invariant de (2.6).*

Le seconde propriété est une propriété de conservation des aires. Soit donc ω^2 la 2-forme différentielle définie par

$$\forall\, v_1, v_2,\ \in\ \mathbb{R}^{2d},\ \omega^2(v_1,v_2) = \sum_{I=1}^{d} (dp^I \wedge dq^I)(v_1,v_2)$$

où les $2d$ termes dp^I et dq^I désignent les formes différentielles suivantes [2]

$$dp^I : \begin{array}{ccc} \mathbb{R}^{2d} & \rightarrow & \mathbb{R} \\ v & \mapsto & v^I \end{array} \qquad dq^I : \begin{array}{ccc} \mathbb{R}^{2d} & \rightarrow & \mathbb{R} \\ v & \mapsto & v^{I+d} \end{array}$$

et où l'expression $(dp^I \wedge dq^I)(v_1,v_2)$ désigne le produit extérieur

$$(dp^I \wedge dq^I)(v_1,v_2) = dp^I(v_1)dq^I(v_2) - dp^I(v_2)dq^I(v_1) = v_1^I v_2^{I+d} - v_1^{I+d} v_2^I.$$

La figure 2.2 permet d'interpréter ω^2 : pour une paire de vecteurs v_1 et v_2 de $\mathbb{R}^{2d}$, $(dp^I \wedge dq^I)(v_1,v_2)$ représente *l'aire orientée* de la projection sur le plan (p^I,q^I) du parallélogramme engendré par v_1 et v_2 ; $\omega^2(v_1,v_2)$ est alors la somme de ces contributions sur chacun des plans (p^I,q^I).

Soient maintenant V une variété de $\mathbb{R}^{2d}$ de dimension 2, (p_0,q_0) un point sur V et $v_{0,1}$, $v_{0,2}$ deux vecteurs tangents en ce point à V. On considère $v_{x,1}$, $v_{x,2}$ les images de $v_{0,1}$, $v_{0,2}$ par ψ_x. Les vecteurs $v_{x,1}$ et $v_{x,2}$ sont tangents à $\psi_x(V)$ en $\psi_x(p_0,q_0)$. On a alors :

Théorème 4 *Le flot* ψ_x *associé à un système hamiltonien de la forme (2.6) est* symplectique, *i.e. est tel que*

$$\forall\, x,\ \omega^2(v_{0,1},v_{0,2}) = \omega^2(v_{x,1},v_{x,2}).$$

2. v^K désigne la Kième composante de v.

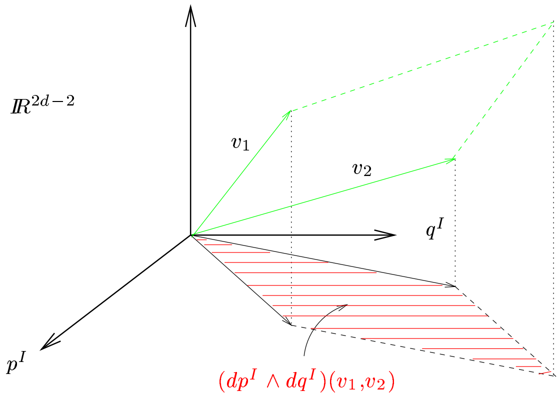


FIG. 2.2 – *Produit extérieur de deux formes linéaires*

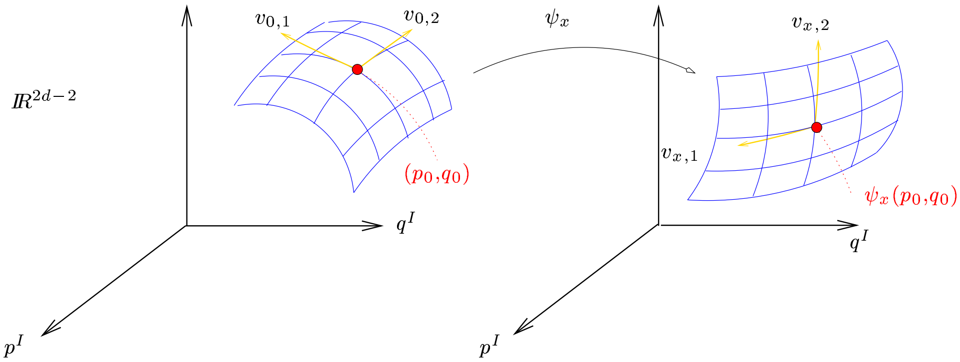


FIG. 2.3 – *Images de* V*,* (p_0,q_0) *et* $v_{0,1}$ *et* $v_{0,2}$ *par* ψ_x

De manière équivalente, ψ_x *vérifie l'égalité suivante :*

$$\left(\frac{\partial \psi_x(p,q)}{\partial(p,q)}\right)^T J \left(\frac{\partial \psi_x(p,q)}{\partial(p,q)}\right) = J,$$

où J *est la matrice*

$$\left[\begin{array}{cc} 0 & I_{I\!R^d} \\ -I_{I\!R^d} & 0 \end{array}\right].$$

Remarque 3 *Tout système hamiltonien vérifie la condition (2.4).*

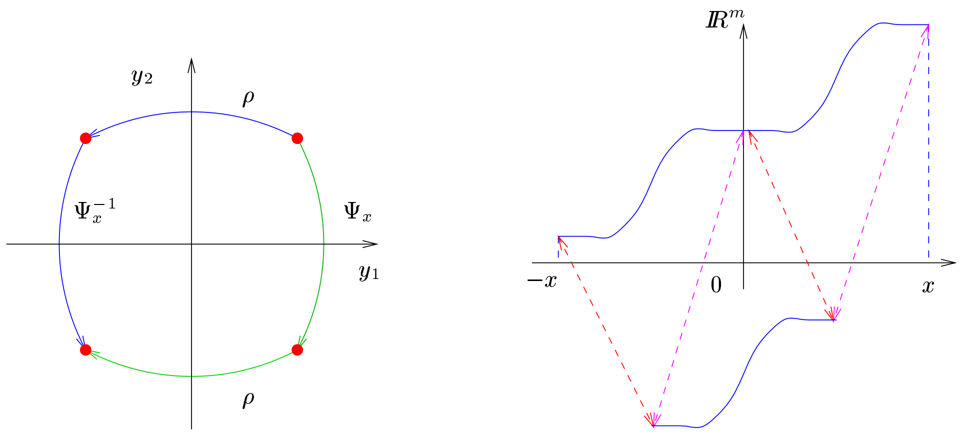


FIG. 2.4 – *Réversibilité du flot en dimension* 2 *(gauche);* $\Psi_x^{-1} = \Psi_{-x}$ *(droite)*

2.3.2 Systèmes réversibles en temps

Définition 3 *Le système (2.1) est dit* réversible *si pour tout isomorphisme* ρ *de* $I\!R^n$ *on a (voir figure 2.4) :*

$$f\,o\,\rho = -\rho\,o\,f.$$

Si (2.1) est réversible, alors le flot ψ_x associé vérifie

$$\rho\,o\,\psi_x = \psi_x^{-1}\,o\,\rho.$$

Cette égalité s'obtient aisément en constatant que (2.1) étant autonome, on a $\psi_x^{-1} = \psi_{-x}$ (voir figure 2.4).

Les systèmes réversibles sont importants à maints égards. En premier lieu, de nombreux systèmes hamiltoniens sont réversibles en temps. C'est le cas en particulier si $H(p,q) = H(-p,q)$ dans (2.6). On verra dans les paragraphes suivants, qu'il existe des méthodes symétriques (adaptées aux systèmes réversibles) à pas variables, ce qui n'est pas le cas des méthodes symplectiques (adaptées elles aux systèmes hamiltoniens). En second lieu, certains systèmes issus de la physique sont réversibles sans être hamiltoniens et la préservation des symétries est alors le pendant de celle de l'hamiltonien ou de la symplecticité du flot. L'étude des systèmes réversibles est donc également intéressante en soi.

Chapitre 3

Intégration numérique géométrique

Les systèmes hamiltoniens ou réversibles en temps sont des exemples particuliers d'équations différentielles ordinaires. A ce titre, les méthodes numériques habituelles sont susceptibles de leur être appliquées. Cependant, qu'il s'agisse de méthodes de Runge-Kutta ou de méthodes multipas, les schémas couramment utilisés ne possèdent généralement aucune des propriétés spécifiques aux systèmes hamiltoniens ou réversibles en temps. Ainsi, l'hamiltonien n'est pas conservé le long des solutions numériques obtenues, pas plus que le flot numérique n'est en général symplectique ou réversible en temps.

Existe-t-il, parmi les méthodes de Runge-Kutta, des méthodes possédant les propriétés évoquées ci-avant? Bien que la réponse soit affirmative dans les deux cas, il convient d'étudier séparément les méthodes symplectiques et les méthodes symétriques.

3.1 Méthodes de Runge-Kutta symplectiques

Il n'existe aucune méthode conservant à la fois l'hamiltonien et le caractère symplectique du flot [Sco91] (en dehors de cas particuliers: par exemple lorsque la fonction hamiltonienne est un invariant quadratique). Il existe cependant des méthodes conservant soit l'un, soit l'autre.

3.1.1 Méthodes de Runge-Kutta partitionnées

On rappelle qu'une méthode de Runge-Kutta (partitionnée) $(\mathcal{R},\hat{\mathcal{R}})$ associe deux méthodes de Runge-Kutta $\mathcal{R} = (A,b)$ et $\hat{\mathcal{R}} = (\hat{A},\hat{b})$ et s'écrit pour un système de la forme (2.6) :

$$\begin{aligned} P_i &= p_0 - h\sum_{j=1}^{s} a_{i,j}\nabla_q H(P_j,Q_j),\ i=1,\ldots,s, && (3.1)\\ Q_i &= q_0 + h\sum_{j=1}^{s} \hat{a}_{i,j}\nabla_p H(P_j,Q_j),\ i=1,\ldots,s, && (3.2)\\ p_1 &= p_0 - h\sum_{j=1}^{s} b_j\nabla_q H(P_j,Q_j), && (3.3)\\ q_1 &= q_0 + h\sum_{j=1}^{s} \hat{b}_j\nabla_p H(P_j,Q_j), && (3.4)\end{aligned}$$

où les (P_i,Q_i) sont les approximations internes et (p_1,q_1) l'approximation fournie par $(\mathcal{R},\hat{\mathcal{R}})$ de la solution exacte au point $x_1 = x_0 + h$. Les matrices A et $\hat{A}$ de $\mathbb{R}^{s\times s}$, les vecteurs b et $\hat{b}$ de $\mathbb{R}^s$ sont les *coefficients* de la méthode. Lorsque $A = \hat{A}$ et $b = \hat{b}$, on retrouve alors la formulation classique (voir (3.7,3.8)) d'une méthode de Runge-Kutta.

Pour un système hamiltonien déterminé et une méthode fixée, on considère le flot *numérique* ψ_h, indicé par le pas d'intégration h, comme étant la fonction :

$$\begin{array}{rccc} \psi_h : & \mathbb{R}^{2d} & \to & \mathbb{R}^{2d} \\ & (p_0,q_0) & \mapsto & (p_1,q_1) \end{array}$$

où (p_1,q_1) est obtenu par application de (3.1,3.2,3.3,3.4).

3.1.2 Caractérisation des méthodes symplectiques

Définition 4 *Une méthode de Runge-Kutta est dite* symplectique *si pour toute fonction hamiltonienne suffisamment régulière et pour tout pas d'intégration h, le flot numérique ψ_h associé est une fonction symplectique, c'est-à-dire si*

$$\left(\frac{\partial\psi_h}{\partial(p,q)}\right)^T J \left(\frac{\partial\psi_h}{\partial(p,q)}\right) = J. \tag{3.5}$$

On peut désormais caractériser simplement les méthodes de Runge-Kutta symplectiques $(\mathcal{R},\hat{\mathcal{R}})$ à l'aide de la matrice M, dite matrice de *stabilité algébrique* (voir Burrage & Butcher [BB79] ou Crouzeix [Cro79]), définie par

$$M = B\hat{A} + A^T\hat{B} - b\hat{b}^T \in \mathbb{R}^{s\times s}$$

où $B = \mathrm{diag}(b_1,\ldots,b_s)$ et $\hat{B} = \mathrm{diag}(\hat{b}_1,\ldots,\hat{b}_s)$.

Théorème 5 (Lasagni [Las88], Sanz-Serna [SS88], Suris [Sur89]) *Une méthode de Runge-Kutta* $(\mathcal{R},\hat{\mathcal{R}}) = (A,b,\hat{A},\hat{b})$ *est symplectique si*

$$\left\{ \begin{array}{ccc} M & = & 0 \\ b & = & \hat{b} \end{array} \right. .$$

démonstration : on pose $u_j = -\nabla_q H(P_j,Q_j)$ et $v_j = \nabla_p H(P_j,Q_j)$. Par linéarité, on a

$$dp_1^L = dp_0^L + h\sum_j b_j du_j^L,$$

$$dq_1^L = dq_0^L + h\sum_j \hat{b}_j dv_j^L,$$

de sorte que

$$\begin{aligned} dp_1^L \wedge dq_1^L &= dp_0^L \wedge dq_0^L + h\sum_j (b_j du_j^L \wedge dq_0^L + \hat{b}_j dp_0^L \wedge dv_j^L) \\ &+ h^2 \sum_{j,k} b_j \hat{b}_k du_j^L \wedge dv_k^L. \end{aligned} \tag{3.6}$$

De (3.1,3.2) on tire

$$dp_0^L = dP_j^L - h\sum_k b_k du_k^L \text{ et } dq_0^L = dQ_j^L - h\sum_k \hat{b}_k dv_k^L,$$

ce qui donne

$$\begin{aligned} h\sum_j b_j du_j^L \wedge dq_0^L &= h\sum_j b_j du_j^L \wedge dQ_j^L - h^2 \sum_{j,k} b_j \hat{a}_{j,k} du_j^L \wedge dv_k^L, \\ h\sum_j \hat{b}_j dp_0^L \wedge dv_j^L &= h\sum_j \hat{b}_j dP_j^L \wedge dv_j^L - h^2 \sum_{j,k} \hat{b}_j a_{j,k} du_k^L \wedge dv_j^L. \end{aligned}$$

En regroupant dans (3.6) les termes en h^2, on obtient

$$\begin{aligned} dp_1^L \wedge dq_1^L &= dp_0^L \wedge dq_0^L + h\sum_j b_j (du_j^L \wedge dQ_j^L + dP_j^L \wedge dv_j^L) \\ &- h^2 \sum_{j,k} m_{jk} du_j^L \wedge dv_k^L. \end{aligned}$$

Il reste à prouver que le terme $\sum_j b_j (du_j^L \wedge dQ_j^L + dP_j^L \wedge dv_j^L)$ disparaît dans $\sum_L dp_1^L \wedge dq_1^L$. Pour cela, on considère les expressions

$$\begin{aligned} du_j^L &= -\sum_K (\frac{\partial^2 H}{\partial q^L p^K} dP_j^K + \frac{\partial^2 H}{\partial q^L q^K} dQ_j^K), \\ dv_j^L &= \sum_K (\frac{\partial^2 H}{\partial p^L p^K} dP_j^K + \frac{\partial^2 H}{\partial p^L q^K} dQ_j^K). \end{aligned}$$

Il vient alors

$$\begin{aligned}\sum_L (du_j^L \wedge dQ_j^L + dP_j^L \wedge dv_j^L) = \sum_{K,L} (\quad & - \frac{\partial^2 H}{\partial q^L p^K} dP_j^K \wedge dQ_j^L \\ & - \frac{\partial^2 H}{\partial q^L q^K} dQ_j^K \wedge dQ_j^L \\ & + \frac{\partial^2 H}{\partial p^L p^K} dP_j^L \wedge dP_j^K \\ & + \frac{\partial^2 H}{\partial p^L q^K} dP_j^L \wedge dQ_j^K).\end{aligned}$$

Le premier et le dernier termes s'annulent par symétrie des indices. Le fait que le deuxième terme (de même que le troisième) s'annule se déduit aisément des propriétés du produit extérieur et des dérivées partielles, à savoir

$$\begin{aligned} dQ_j^L \wedge dQ_j^K &= -dQ_j^K \wedge dQ_j^L, \\ dQ_j^L \wedge dQ_j^L &= 0, \\ \frac{\partial^2 H}{\partial q^L q^K} &= \frac{\partial^2 H}{\partial q^K q^L}. \end{aligned}$$

□

Remarque 4 *Dans le cas où H est séparable, i.e. $H(p,q) = T(p) + U(q)$, la condition $M = 0$ est suffisante.*

Remarque 5 *Une autre caractérisation est possible en termes d'arbres racinés (Calvo & Sanz-Serna [CSS94], Hairer [Hai94]) et sera explicitée au chapitre 3.*

Les méthodes basées sur les formules de quadrature de Gauss sont symplectiques. Ce résultat est une conséquence immédiate du fait que pour $A = \hat{A}$, $b = \hat{b}$, la matrice M correspondant à la méthode de Gauss est nulle (voir Butcher [But87], théorème 356F). Une démonstration directe de la symplecticité des méthodes de Gauss est également possible (voir Hairer & Wanner [HW96], théorème 16.5).

Les méthodes de Radau IA, Radau IIA, Lobatto IIIA, IIIB et IIIC ne sont pas symplectiques (voir Hairer & Wanner [HW96] pour une définition de ces méthodes). Par contre, la méthode partitionnée Lobatto IIIA-Lobatto IIIB l'est (voir Jay [Jay94]) et possède un certain nombre de propriétés remarquables pour les systèmes hamiltoniens avec contrainte holonome.

Une méthode explicite ne peut être symplectique. Cette règle souffre (ou plutôt bénéficie) d'une notable exception. Pour les systèmes à hamiltonien séparable, il est en effet possible de construire des méthodes partitionnées implicites en théorie mais possédant en pratique une formulation explicite (voir Sanz-Serna & Calvo [SSC94], paragraphe 8.4). On peut illustrer la situation générale au moyen de la paire Euler implicite-Euler explicite, pour laquelle les conditions du théorème 5 sont trivialement vérifiées. Les formules (3.1,3.2,3.3,3.4) s'écrivent ici

$$\begin{aligned} p_1 &= p_0 - h\nabla_q H(p_1,q_1), \\ q_1 &= q_0 + h\nabla_p H(p_0,q_0). \end{aligned}$$

Mais l'hamiltonien étant séparable, on a en fait

$$\begin{aligned} q_1 &= q_0 + h\nabla_p T(p_0), \\ p_1 &= p_0 - h\nabla_q U(q_1). \end{aligned}$$

Remarque 6 *Dans le cas où $H(p,q) = \frac{1}{2}p^T M p + U(q)$, le système (2.6) peut aussi s'écrire*

$$q'' = -M\nabla_q U(q).$$

Il est alors possible de recourir aux méthodes de Runge-Kutta-Nyström, plus adaptée à l'ordre 2. Divers auteurs ont en effet construit des méthodes explicites symplectiques de ce type (voir Suris [Sur89], Calvo [Cal92], Calvo & Sanz-Serna [CSS92], Okunbor & Skeel [OS94]).

3.2 Méthodes symétriques

3.2.1 Définitions

Pour un système de la forme (2.1), une méthode de Runge-Kutta $\mathcal{R} = (A,b)$ s'écrit :

$$Y_i = y_0 + h\sum_{j=1}^{s} a_{i,j} f(Y_j),\ i = 1,\ldots,s, \tag{3.7}$$

$$y_1 = y_0 + h\sum_{j=1}^{s} b_j f(Y_j). \tag{3.8}$$

Les Y_i sont des approximations internes. y_1 est une approximation de la solution exacte au point $x_1 = x_0 + h$. La matrice A de $\mathbb{R}^{s\times s}$ et le vecteur b de $\mathbb{R}^s$ sont les *coefficients* de la méthode. Sous forme plus compacte, un pas de $\mathcal{R}$ s'écrit encore :

$$\begin{aligned} Y &= (e\otimes I_{\mathbb{R}^m})y_0 + h(A\otimes I_{\mathbb{R}^m})f(Y), \\ y_1 &= y_0 + h(b^T\otimes I_{\mathbb{R}^m})f(Y), \end{aligned}$$

avec les notations [1]

$$\begin{aligned} Y &= [Y_1^T,\ldots,Y_s^T]^T, \\ f(Y) &= [f(Y_1)^T,\ldots,f(Y_s)^T]^T, \end{aligned}$$

et $e = [1,\ldots,1]^T \in I\!R^s$.

De même que pour les systèmes hamiltoniens, on peut définir pour toute fonction f (2.1) et tout pas h la fonction flot ψ_h :

$$\begin{array}{rccc} \psi_h \; : & I\!R^m & \to & I\!R^m \\ & y_0 & \mapsto & y_1 \end{array}$$

où y_1 est l'approximation obtenue par application des formules (3.7, 3.8).

Une fonction flot ψ_h est dite réversible si pour tout isomorphisme ρ de $I\!R^m$ et toute fonction f de (2.1) tels que

$$f \, o \, \rho = -\rho \, o \, f \tag{3.9}$$

on a :

$$\psi_h \, o \, \rho = \rho \, o \, \psi_h^{-1}. \tag{3.10}$$

Stoffer [Sto88] a montré que les méthodes de Runge-Kutta réversibles sont en fait les méthodes *symétriques*, c'est-à-dire les méthodes dont le flot associé vérifie

$$\psi_h = \psi_{-h}^{-1}.$$

Définition 5 *Une méthode de Runge-Kutta $\mathcal{R}$ est dite symétrique si le flot numérique associé vérifie l'identité (voir figure 3.2.1) :*

$$\psi_h \, o \, \psi_{-h} = Id_{I\!R^m}. \tag{3.11}$$

3.2.2 Caractérisation des méthodes symétriques

On a alors la caractérisation suivante des méthodes symétriques :

Théorème 6 (Stetter [Ste73], Wanner [HNW93]) *Une méthode de Runge-Kutta $\mathcal{R} = (A,b)$ est symétrique si ses coefficients vérifient les égalités :*

$$\begin{aligned} PA + AP &= eb^T, && (3.12) \\ b &= Pb, && (3.13) \end{aligned}$$

1. L'emploi de f ici est abusif mais habituel.

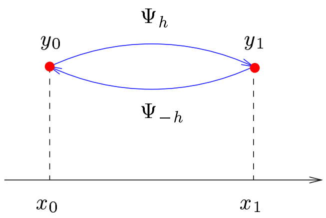


FIG. 3.1 – *Symétrie du flot numérique*

où P est la matrice

$$P = \begin{bmatrix} 0 & \dots & 0 & 1 \\ \vdots & \iddots & \iddots & 0 \\ 0 & \iddots & \iddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix}.$$

Soit $c = Ae$: si les c_i sont distincts et ordonnés par ordre croissant et si les b_i sont non nuls alors cette condition est en outre nécessaire.

démonstration : Pour $y \in I\!R^m$, $\tilde{y} = \Psi_h(\Psi_{-h}(y))$ est donné par les équations :

$$\begin{aligned} Y &= (e \otimes I_{I\!R^m})y + h(A \otimes I_{I\!R^m})f(Y), \\ y_1 &= y + h(b^T \otimes I_{I\!R^m})f(Y), \\ \tilde{Y} &= (e \otimes I_{I\!R^m})y_1 - h(A \otimes I_{I\!R^m})f(\tilde{Y}), \\ \tilde{y} &= y - h(b^T \otimes I_{I\!R^m})f(\tilde{Y}). \end{aligned}$$

Soit alors

$$\begin{aligned} \Delta Y &= (P \otimes I_{I\!R^m})\tilde{Y} - Y, \\ \Delta F &= (P \otimes I_{I\!R^m})f(\tilde{Y}) - f(Y), \end{aligned}$$

Un simple calcul en tenant compte de (3.12) et de (3.13) donne :

$$\Delta Y = h(PA \otimes I_{I\!R^m})\Delta F.$$

Si f est supposée Lipschitziennne, alors pour toute norme $|.|$ sur $I\!R^{m\times s}$, il existe une constante $C > 0$ telle que

$$\begin{aligned} |\Delta Y| &\leq h|PA \otimes I_{I\!R^m}||\Delta F|, \\ &\leq hC|\Delta Y|. \end{aligned}$$

Pour h suffisamment petit, cela entraîne $\Delta Y = 0$ et par suite $\tilde{y} = y$.
La réciproque s'obtient en considérant des fonctions f de la forme :

$$f(y_1,y_2) = \begin{pmatrix} 1 \\ g(y_1) \end{pmatrix}$$

(voir Hairer, Nørsett & Wanner [HNW93], pp. 222). □

Parmi les méthodes classiques, les méthodes de Gauss, de Lobatto IIIA, IIIB et IIIS (voir Chan [Cha90]) sont symétriques. Les méthodes de Radau IA et IIA ne le sont pas. Notons enfin que comme dans le cas des méthodes symplectiques, une méthode de Runge-Kutta symétrique est nécessairement implicite.

3.3 Exemple de l'oscillateur harmonique

On considère l'exemple très simple de l'oscillateur harmonique, dont l'hamiltonien s'écrit :

$$H(p,q) = \frac{1}{2}mp^2 + \frac{1}{2}kq^2, m = k = 1.$$

La figure 3.3 montre l'évolution d'un ensemble de valeurs initiales (tête de chat [Arn67]), après $800 + i/8$ périodes, $i = 1,\ldots,8$. L'expérience est répétée pour trois méthodes : BUTCHER6 (voir [HNW87], pp. 189), méthode non-symplectique d'ordre 6 et GAUSS4 et GAUSS6 respectivement, méthodes de Gauss symplectiques d'ordres respectifs 4 et 6. La propriété de conservation des aires apparaît ici clairement pour la solution exacte et les méthodes symplectiques GAUSS4 et GAUSS6 (sur la figure, GAUSS6 se confond pratiquement avec la solution exacte). La "croissance" du chat est par contre très nette pour BUTCHER6.

La solution exacte du système est une rotation de la forme :

$$\begin{pmatrix} p(x) \\ q(x) \end{pmatrix} = \begin{pmatrix} \cos(x) & \sin(x) \\ -\sin(x) & \cos(x) \end{pmatrix} \begin{pmatrix} p_0 \\ q_0 \end{pmatrix}.$$

Le flot numérique ψ_h associé à une méthode de Runge-Kutta $\mathcal{R} = (A,b)$ est de la même forme

$$\begin{pmatrix} p_1 \\ q_1 \end{pmatrix} = \psi_h \begin{pmatrix} p_0 \\ q_0 \end{pmatrix} = \begin{pmatrix} \cos(w) & \sin(w) \\ -\sin(w) & \cos(w) \end{pmatrix} \begin{pmatrix} p_0 \\ q_0 \end{pmatrix}.$$

avec $w = \arg(1 + ihb^T(I_{\mathbb{R}^2} - ihA)e)$. On peut en outre montrer (voir Chan & Murua [CM]) que si $\mathcal{R}$ est d'ordre p, alors $w = h + \mathcal{O}(h^{p+1})$ et que si $\mathcal{R}$ est

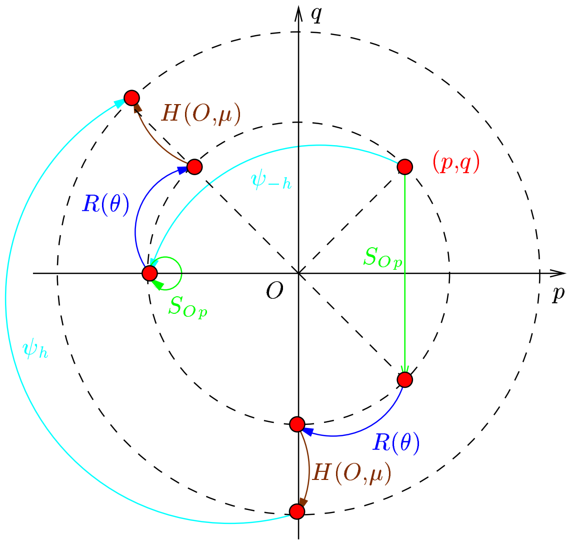


FIG. 3.2 – $\rho\, o\, \psi_h = \psi_{-h}\, o\, \rho$

symétrique, w est une fonction *impaire* de h, de sorte que $\psi_h^{-1} = \psi_{-h}$. Il est par ailleurs aisé de vérifier que les isomorphismes ρ de $\mathbb{R}^2$ satisfaisant (3.9) avec f donnée par (2.6) sont de la forme :

$$\rho(\theta,\mu) = \underbrace{\begin{bmatrix} \mu & 0 \\ 0 & \mu \end{bmatrix}}_{H(O,\mu)} \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}}_{S_{Op}} \underbrace{\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}}_{R(\theta)}, \ \mu \in \mathbb{R},\ \theta \in [0,2\pi[.$$

Il est donc aisé de vérifier (3.10) (voir figure 3.2).

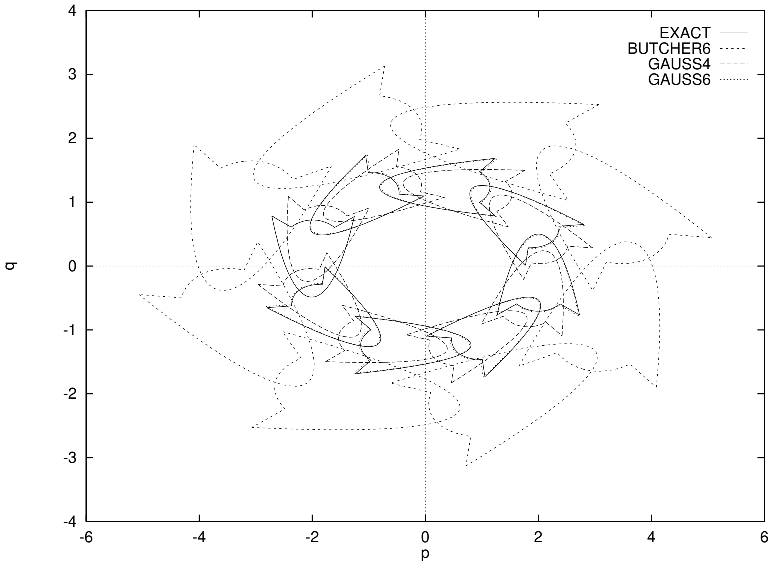


FIG. 3.3 – *Tête de chat après passage à la machine*

Chapitre 4

Méthodes de Runge-Kutta à croissance linéaire de l'erreur

4.1 Éléments d'analyse rétrograde

Soit la suite

$$y_0, y_1, \dots, y_n, \dots$$

des approximations numériques aux points

$$x_0, x_1 = x_0 + h, \dots, x_n = x_0 + nh, \dots,$$

de la solution exacte de (2.1), obtenue par une méthode de Runge-Kutta $\mathcal{R}$ à pas constant h. L'idée de l'analyse rétrograde consiste à interpréter cette suite comme la solution *exacte* $\tilde{y}(x)$ aux points

$$x_0, x_1, \dots, x_n, \dots,$$

de l'équation modifiée

$$\tilde{y}'(x) = f(\tilde{y}(x)) + hf_2(\tilde{y}(x)) + \cdots + h^{I-1}f_I(\tilde{y}(x)) + \dots, \tag{4.1}$$

avec comme condition initiale $\tilde{y}(x_0) = y_0$ (les fonctions f_i peuvent être calculées récursivement en fonction de f et de ses dérivées, voir Hairer [Hai94]). Le second membre de (4.1) est une série généralement non convergente, qu'il convient de tronquer après le terme en h^{I-1}. Sous certaines hypothèses (dont l'analycité de f), Hairer et Lubich ont montré le résultat suivant (voir Hairer & Lubich [HL97]) :

$$\exists\, C > 0, \exists\, h_0 > 0, \forall h \leq \frac{h_0}{2}, \exists\, I(h), \|y_1 - \tilde{y}_I(x_0 + h)\| \leq Ce^{-\frac{h_0}{h}},$$

où $\tilde{y}_I$ est solution de (4.1) tronquée après le terme en h^{I-1} et où h_0 et C sont des constantes dépendant des coefficients de la méthode utilisée et de majorations de f et de ses dérivées sur le domaine où elle est holomorphe.

Il est essentiel de noter que :
- (4.1) est *hamiltonien* si (2.1) l'est et si la méthode numérique employée est *symplectique*;
- (4.1) est *réversible en temps* si (2.1) l'est et si la méthode numérique employée est *symétrique*.

Ces propriétés, démontrées dans Hairer [Hai94], ont d'importantes conséquences (voir Calvo & Hairer [CH95]) :
- pour un système hamiltonien, l'analyse rétrograde permet d'expliquer la quasi conservation de l'hamiltonien par les méthodes symplectiques sur des intervalles de temps très longs;
- pour un système hamiltonien ou réversible en temps, supposé *intégrable*, l'analyse rétrograde permet d'expliquer la croissance linéaire de l'erreur globale.

Pour une méthode non symplectique d'ordre p et un pas h suffisamment petit, l'erreur globale

$$E = \|y(X) - y_N\|$$

sur un intervalle d'intégration de longueur $X - x_0 = Nh$ vérifie

$$E \leq L(X - x_0)^2 h^p,$$

où L est une constante dépendant du problème et de la méthode. Le comportement de l'erreur globale pour une méthode symplectique est beaucoup plus favorable puisque on a alors

$$E \leq \hat{L}(X - x_0) h^p,$$

où $\hat{L}$ est une constante dépendant du problème et de la méthode. Il est nécessaire de rappeler que ces observations ne sont valables et démontrées que dans les cas ci-dessus décrits.

4.2 Méthodes pseudo-symplectiques

Comme il a été dit au chapitre précédent, une méthode symplectique ou symétrique est nécessairement implicite (en dehors de certaines exceptions importantes en pratique et déjà signalées). Le gain potentiel qu'induit la croissance linéaire de l'erreur globale est susceptible d'être oblitéré par le coût d'une méthode implicite.

L'idée est ici d'affaiblir la définition de la symplecticité (resp. de la symétrie) afin de rendre possible la construction de méthodes de Runge-Kutta *explicites* dites *pseudo-symplectiques* (resp. *pseudo-symétriques*) se comportant essentiellement comme des méthodes symplectiques (resp. symétriques).

4.2.1 Première caractérisation : définition

Définition 6 (Aubry & Chartier [AC98c]) *Une méthode à un pas, utilisée avec un pas h constant, appliquée à un système hamiltonien (2.6), est dite pseudo-symplectique d'ordre de pseudo-symplecticité q (d'o.p.s. q), si son flot numérique Ψ_h vérifie*

$$\left(\frac{\partial\Psi_h}{\partial(p,q)}\right)^T J \left(\frac{\partial\Psi_h}{\partial(p,q)}\right) = J + \mathcal{O}\left(h^{q+1}\right). \tag{4.2}$$

On dit qu'une méthode est d'ordre (p,q) si elle est d'ordre classique p et d'ordre de pseudo-symplecticité q (voir Moan [Moa96] pour la notation [1]).

Remarque 7 *La formule (4.2) est bien sûr à rapprocher de (3.5) d'où le terme $\mathcal{O}(h^{q+1})$ est absent.*

4.2.2 Deuxième caractérisation : B-séries

On rappelle quelques définitions relatives aux arbres (racinés). Le lecteur désireux de se familiariser avec ce formalisme peut consulter les monographies de Butcher [But87] ou de Hairer, Nørsett & Wanner [HNW93].

Définition 7 *L'ensemble des arbres, noté T, est défini récursivement par :*

1. $\tau \in T$,
2. $[t_1,t_2,\ldots,t_n] \in T$ *si et seulement si pour tout i de $\{1,\ldots,n\}$, $t_i \in T$.*

L'ordre d'un arbre t, noté $\rho(t)$, est son nombre de nœuds. De plus, si $t = [t_1,\ldots,t_n]$, $t \circ t'$ désigne l'arbre $[t_1,\ldots,t_n,t']$.

1. Pour des raisons qui tiennent à l'habitude, p et q désignent à la fois les variables du système hamiltonien (2.6) et les ordres de convergence et de pseudo-symplecticité. Leur signification est cependant claire d'après le contexte

Soit maintenant $\mathcal{R} = (A,b)$ une méthode de Runge-Kutta à s étapes internes.

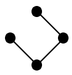
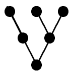
Définition 8 *La fonction $\Phi(t)$ de T à valeurs dans $I\!R^s$ est définie récursivement par :*

1. $\Phi(\tau) = e$*,*

2. $\Phi([t_1,t_2,\ldots,t_n]) = \prod_{i=1}^{n} A\Phi(t_i)$*,*

où le produit de vecteur est à entendre composante par composante.

Des exemples d'arbres et les valeurs des fonctions associées sont donnés dans le tableau 4.1. Dans Calvo & Sanz-Serna [CSS94], les auteurs établissent une condi-

t	• τ	 $[\tau,[\tau]]$	 $[[\tau],[\tau,\tau]]$
$\rho(t)$	1	4	6
$\Phi(t)$	e	$Ae.A^2e$	$A^2e.A(Ae.Ae)$

TAB. 4.1 – *Exemples d'arbres et valeurs des fonctions associées*

tion nécessaire et suffisante pour qu'une B-série définisse une méthode symplectique[2]. Afin de ne pas accroître l'importance du formalisme, on particularise ici ce résultat au cas des méthodes de Runge-Kutta.

Théorème 7 (Calvo & Sanz-Serna, [CSS94]) *Une méthode de Runge-Kutta $\mathcal{R} = (A,b)$ est symplectique si et seulement si $b^T\Phi(\tau) = 1$ et pour toute paire d'arbres $u,v \in T$,*

$$b^T\Phi(u \circ v) + b^T\Phi(v \circ u) = \left(b^T\Phi(u)\right)\left(b^T\Phi(v)\right). \tag{4.3}$$

Pour une méthode pseudo-symplectique d'o.p.s. q, il vient alors :

Corollaire 1 *Une méthode de Runge-Kutta $\mathcal{R} = (A,b)$ est pseudo-symplectique d'o.p.s. q si et seulement si $b^T\Phi(\tau) = 1$ et pour toute paire d'arbres $u,v \in T$ telle que $\rho(u) + \rho(v) \leq q$, la relation (4.3) est satisfaite.*

Ces conditions ont été aussi obtenues indépendamment par Moan [Moa96].

Remarque 8 *Toute méthode à un pas d'ordre classique p est d'ordre de pseudo-symplecticité p.*

Les conditions de pseudo-symplecticité d'ordre 4, obtenues à partir de (4.3), sont regroupées dans le tableau 4.2.

2. Hairer a généralisé ce résultat aux P-séries [Hai94].

ordre	conditions de pseudo-symplecticité	
1	$b^T e = 1$	(1)
2	$b^T c = \frac{1}{2}$	(2)
3	$b^T Ac + b^T c^2 = b^T c$	(3)
4	$b^T A^2 c + b^T (c.Ac) = b^T Ac$	(4)
	$b^T (c.Ac) = \frac{1}{2}(b^T c)^2$	(5)
	$b^T Ac^2 + b^T c^3 = b^T c^2$	(6)

TAB. 4.2 – *Conditions de pseudo-symplecticité d'ordre* 4

4.2.3 Construction de méthodes pseudo-symplectiques explicites

Ce paragraphe regroupe différents résultats concernant l'existence de méthodes pseudo-symplectiques explicites. Il est généralement beaucoup plus difficile de construire des méthodes explicites d'ordre élevé que des méthodes implicites : afin de guider la recherche de telles méthodes, on donne ici des bornes concernant le nombre d'étapes s (i.e. la dimension des matrice et vecteur de coefficients) et le nombre d'étapes séquentielles σ, c'est-à-dire le nombre d'étapes internes devant être impérativement calculées de manière consécutive.

Théorème 8 (Aubry & Chartier [AC98c]) *Soit $\mathcal{R} = (A,b)$ une méthode de Runge-Kutta explicite d'ordre de pseudo-symplecticité $2p$ et d'ordre p, alors le nombre d'étapes séquentielles, noté σ, satisfait les inégalités suivantes :*

- *si $p = 2r$, $\sigma \geq 3r$ et*
- *si $p = 2r + 1$, $\sigma \geq 3r + 2$ où $r \geq 1$.*

Ce théorème fournit donc une borne inférieure σ_{min} sur σ. On peut obtenir très facilement une méthode d'ordre $(p,2p)$ avec $\sigma = 2p$ (il suffit d'itérer par point fixe une méthode de Gauss), ce qui donne une borne supérieure σ_{max} sur σ. Il s'avère que la borne inférieure peut être atteinte à l'ordre $(2,4)$:

Théorème 9 (Aubry & Chartier [AC98c]) *Toute méthode de Runge-Kutta d'ordre $(2,4)$ avec un nombre minimal $s = 3$ d'étapes internes, est de la forme :*

$$\begin{array}{c|ccc}
0 & 0 & 0 & 0 \\
\frac{1}{4}\frac{8\,b_3 - 3}{2\,b_3 - 1} & \frac{1}{4}\frac{8\,b_3 - 3}{2\,b_3 - 1} & 0 & 0 \\
1 & \frac{1}{2}\frac{16\,{b_3}^2 - 8\,b_3 + 1}{b_3\,(8\,b_3 - 3)} & \frac{1}{2}\frac{2\,b_3 - 1}{b_3\,(8\,b_3 - 3)} & 0 \\
\hline
 & \frac{3\,b_3 - 1}{8\,b_3 - 3} & -2\frac{(2\,b_3 - 1)^2}{8\,b_3 - 3} & b_3.
\end{array} \tag{4.4}$$

A l'ordre $(3{,}6)$ existent plusieurs familles de méthodes avec $s = \sigma_{min} = 5$. L'optimisation des coefficients d'erreur a conduit à la méthode PS63 (Aubry & Chartier [AC98c]). On a par contre le résultat négatif suivant :

Théorème 10 (Aubry & Chartier [AC98c]) *Il n'existe pas de méthodes de Runge-Kutta explicites d'ordre* $(4{,}8)$ *avec* $\sigma = 6$ *étapes séquentielles et* $s = \sigma = 6$ *étapes internes.*

Bien que non constructif, le théorème suivant assure l'existence de méthodes $(4{,}8)$ et $(5{,}10)$ avec un nombre minimum d'étapes séquentielles (respectivement 6 et 8). A titre de comparaison, la méthode DOP853 de Dormand & Prince (voir Hairer, Nørsett & Wanner [HNW93] pp. 181) est d'ordre 8 avec $s = \sigma = 12$.

Théorème 11 (Aubry & Chartier [AC98a]) *Il existe des méthodes de Runge-Kutta pseudo-symplectiques d'ordre* $(p,\sigma+2)$ *avec* σ *étapes séquentielles où* σ *est donné par*

$$\sigma = \begin{cases} 3r & \textit{si } p = 2r, \\ 3r+2 & \textit{si } p = 2r+1 \textit{ avec } r \geq 1. \end{cases}$$

4.2.4 Propriétés des méthodes pseudo-symplectiques

Par construction, l'erreur globale des méthodes pseudo-symplectiques croît linéairement en fonction de la longueur de l'intervalle d'intégration (les démonstrations dans le cas des méthodes symplectiques s'adaptent aisément au cas pseudo-symplectique : voir Aubry & Chartier [AC98c]), et ce dans toutes les situations où l'on peut observer un tel comportement pour les méthodes symplectiques.

Par contre, l'hamiltonien n'est pas conservé par une méthode pseudo-symplectique. Bien plus, on observe une dérive séculaire de l'erreur (voir les expériences numériques en fin de chapitre). Cette erreur est néanmoins comparable à celle observée pour une méthode non symplectique d'ordre q.

On rappelle enfin ce qui semble une évidence, à savoir que le flot associé à une méthode pseudo-symplectique n'est en aucun cas symplectique et que les invariants quadratiques ne sont conservés qu'à un terme en $\mathcal{O}(h^q)$ près.

4.3 Méthodes pseudo-symétriques

4.3.1 Première caractérisation : définition

De manière comparable aux méthodes pseudo-symplectiques, on peut définir une méthode pseudo-symétrique de la façon suivante :

Définition 9 *Une méthode de Runge-Kutta $\mathcal{R}$ est dite pseudo-symétrique d'ordre de pseudo-symétrie q (en abrégé d'o.p.st. q) si le flot numérique associé Ψ_h vérifie l'identité suivante :*

$$\Psi_h \, o \, \Psi_{-h} = Id_{\mathbb{R}^m} + \mathcal{O}(h^{q+1}). \tag{4.5}$$

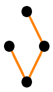
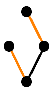
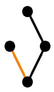
4.3.2 Deuxième caractérisation : B-séries

Théorème 12 (Chartier & Lapôtre [CL98]) *Une méthode de Runge-Kutta $\mathcal{R} = (A,b)$ est symétrique si et seulement si ses coefficients satisfont les relations*

$$\forall t \in T, \, b^T\Phi(t) = \sum_{p \subseteq t} \mu(p) \frac{\sigma(t)}{\sigma(p)} \prod_{s \in d_p(t)} b^T\Phi(s), \tag{4.6}$$

où p décrit l'ensemble des partitions de t, où $\sigma(t)$ et $\sigma(p)$ sont les "symétries" de t et p et $\mu(p)$ un coefficient rationnel attaché à p, et où $d_p(t)$ est l'ensemble différence.

$\sigma(t)$ et $\sigma(p)$ peuvent être définis et calculés récursivement. Il en va de même pour $\mu(p)$. Leur définition précise est donnée dans Chartier & Lapôtre [CL98]. Afin d'illustrer la formule (4.6), quelques partitions et les fonctions associées sont représentées dans le tableau 4.3.

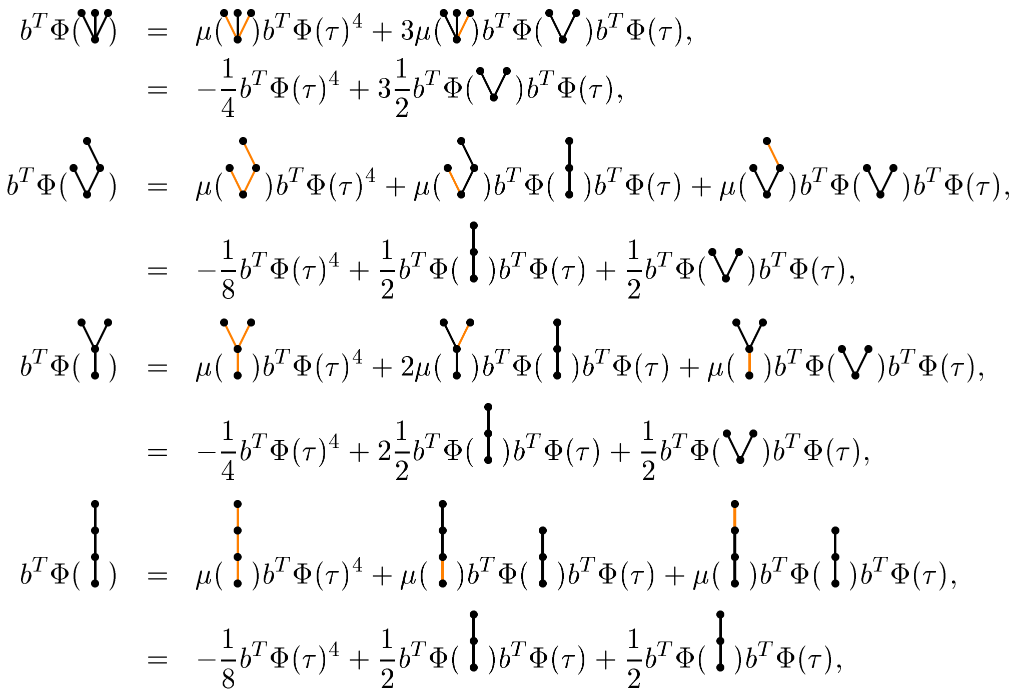
p	$\emptyset$										
$\sigma(p)$	1	1	1	2	1	1	1	1	1	1	1
$\mu(p)$	1	$\frac{-1}{2}$	0	0	$\frac{-1}{8}$	0	0	0	$\frac{-1}{2}$	0	$\frac{-1}{2}$

TAB. 4.3 – *Quelques partitions et fonctions associées.*

Remarque 9 *Pour un arbre d'ordre impair, la formule (4.6) est trivialement vérifiée. En outre, $\mu(p) = 0$ pour toute partition p d'un arbre d'ordre pair. Les conditions du théorème portent donc uniquement sur des arbres d'ordre pair et s'expriment uniquement en fonction des coefficients associés aux arbres d'ordre impair.*

Corollaire 2 (Chartier & Lapôtre) *Une méthode de Runge-Kutta $\mathcal{R} = (A,b)$ est pseudo-symétrique d'o.p.st. q si et seulement si ses coefficients satisfont les relations (4.6) pour tout arbre t tel que $\rho(t) \leq q$.*

Ce dernier corollaire a l'avantage de fournir explicitement les conditions algébriques en les coefficients de la méthode pour qu'elle soit pseudo-symétrique. A titre d'exemple, les conditions jusqu'à l'ordre 4 s'écrivent :



4.4 Construction de méthodes pseudo-symétriques explicites

Comme dans le cas de la pseudo-symplecticité, il est possible de déterminer les nombres σ_{min} et σ_{max} et d'établir certains résultats d'existence ou de non-existence. Néanmoins, les résultats obtenus à ce jour sont encore très partiels :

Théorème 13 (Chartier & Lapôtre [CL98]) *Toute méthode de Runge-Kutta pseudo-symétrique d'ordre $(2,4)$ avec un nombre minimal $s = 3$ d'étapes internes, est de la forme :*

$$
\begin{array}{c|ccc}
0 & 0 & 0 & 0 \\
-\frac{b_3-1}{(2b_3+1)} & -\frac{b_3-1}{2b_3+1} & 0 & 0 \\
\frac{3}{2}\frac{1}{(2b_3+1)} & \frac{1}{8}\frac{(4b_3-1)^2}{b_3(2b_3+1)(b_3-1)} & -\frac{1}{8}\frac{(2b_3+1)}{b_3(b_3-1)} & 0 \\
\hline
 & -b_3+\frac{1}{2} & \frac{1}{2} & b_3
\end{array}
. \qquad (4.7)
$$

Pour $b_3 = \frac{1}{4}$ la méthode

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \hline & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{array}, \tag{4.8}$$

est également pseudo-symplectique d'ordre $(2{,}4)$ *(voir Aubry & Chartier [AC98c] et aussi Hairer, Nørsett & Wanner [HNW93] pp. 223 où cette méthode est également citée).*

Comme dans le cas des méthodes pseudo-symplectiques, la construction de méthodes pseudo-symétriques explicites se heurte aux difficultés pratiques liées à la résolution du système polynomial constitué des conditions d'ordre et de pseudo-symétrie. Afin de réduire le nombre de conditions de pseudo-symétrie, il est possible d'élargir la classe des méthodes pseudo-symétriques aux méthodes «conjuguées» à une méthode pseudo-symétrique. On dit qu'une méthode de flot numérique ψ_h est conjuguée à une méthode de flot $\hat{\psi}_h$ si il existe une méthode de flot θ_h telle que

$$\psi_h = \theta_h^{-1}\, o\, \hat{\psi}_h\, o\, \theta_h.$$

Le comportement à long terme de ψ est essentiellement le même que celui de $\hat{\psi}_h$ puisqu'on a

$$\psi_h^n = (\theta_h^{-1}\, o\, \hat{\psi}_h\, o\, \theta_h)^n = \theta_h^{-1}\, o\, \hat{\psi}_h^n\, o\, \theta_h.$$

En particulier, si $\hat{\psi}_h$ est symétrique ou pseudo-symétrique (en fait ce type de construction est également envisageable dans le contexte des systèmes hamiltoniens), alors ψ_h possède d'excellentes caractéristiques pour l'intégration des systèmes réversibles en temps. On peut noter qu'il n'est pas nécessaire de connaître θ_h ni $\hat{\psi}_h$; leur existence suffit à assurer de bonnes propriétés à ψ_h.

Déterminer les conditions sur ψ_h et les coefficients de la méthode de Runge-Kutta sous-jacente pour qu'un tel θ_h et un tel $\hat{\psi}_h$ existent fait l'objet de premiers résultats cités dans Chartier & Lapôtre [CL98].

4.5 Expériences numériques

Soit un point matériel attiré vers l'origine dans un plan, avec une force inversement proportionnelle au carré de la distance (problème à deux corps ou problème

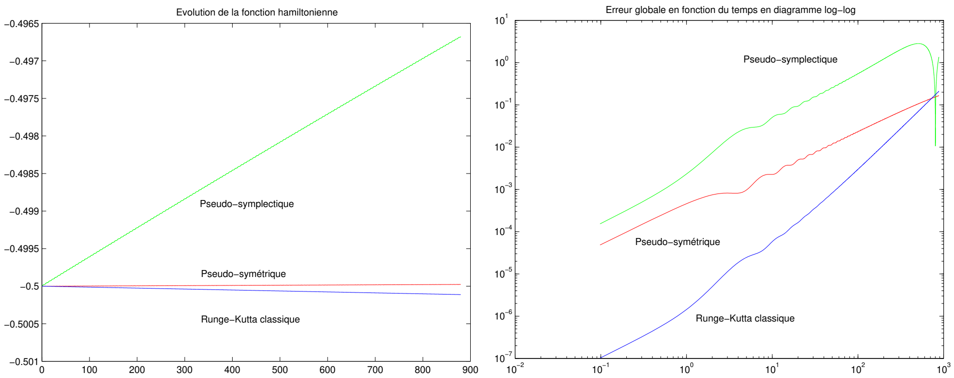


FIG. 4.1 – *Comportement des méthodes (4.4, $b_3 = 4/5$), (4.7, $b_3 = 9/10$) et RK4 pour le problème de Kepler*

de Kepler). Dans sa formulation a-dimensionnelle, la fonction hamiltonienne correspondante s'écrit :

$$H(p,q) = \frac{1}{2}\left(p_1^2 + p_2^2\right) - \frac{1}{\sqrt{q_1^2 + q_2^2}}, \tag{4.9}$$

et par suite, les équations du mouvement sont de la forme :

$$p_i' = -\frac{q_i}{\left(q_1^2 + q_2^2\right)^{\frac{3}{2}}},\, q_i' = p_i,\, i = 1{,}2. \tag{4.10}$$

On considère les conditions initiales $(p_1(0), p_2(0), q_1(0), q_2(0)) = (1,0,0,1)$, de sorte que la solution exacte est $(\cos t, -\sin t, \sin t, \cos t)$. La figure 4.1 représente l'évolution de l'hamiltonien et de l'erreur globale (en fonction du temps) avec un pas constant $h = \frac{2\pi}{256}$ durant 1000 périodes. Elle permet d'observer la pente 2 de l'erreur de la méthode de Runge-Kutta classique RK4 (évolution quadratique) et les pentes 1 (croissance linéaire) des méthodes pseudo-symplectique et pseudo-symétrique. On rappelle en effet que le problème de Kepler est réversible en temps.

Remarque 10 *Le fait que (4.7) soit considérablement plus précise que (4.4) tient à l'impossibilité d'optimiser les coefficients d'erreur de (4.4). Les coefficients d'erreur à l'ordre 3 sont en effet indépendants de b_3.*

Chapitre 5

Systèmes algébro-différentiels

5.1 Introduction et notations

Dans ce chapitre, on s'intéresse à une sous-classe des systèmes de la forme

$$F(y,y') = 0, \tag{5.1}$$

où F est une fonction régulière de $\mathbb{R}^{2m}$ vers $\mathbb{R}^m$. On suppose en outre ici que $F_{y'}$ est singulière en tout point de $\mathbb{R}^{2m}$, de sorte qu'il n'est pas possible de recourir au théorème des fonctions implicites et ce faisant de reformuler (5.1) comme une équation différentielle ordinaire.

Lorsque $F_{y'}$ est supposée singulière comme ici, l'existence d'une solution ne peut être envisagée que pour des valeurs initiales (y_0,y_0') *consistantes*, i.e. telles que $F(y_0,y_0') = 0$. D'une façon générale, les conditions pour qu'une telle solution existe sont nettement plus complexes que dans le cas des EDO. Brenan, Campbell & Petzold ont donc introduit le concept de résolubilité [BCP89] (voir aussi Campbell & Gear [CG95]). La résolubilité de (5.1) assure que les solutions forment localement une sous-variété de dimension r.

Définition 10 (Brenan, Campbell & Petzold [BCP89]) *Soit Ω un ouvert connexe non vide de $\mathbb{R}^{2m}$. L'équation (5.1) est localement résoluble sur Ω si il existe un ouvert Λ non vide de $\mathbb{R}^r$, un intervalle I de $\mathbb{R}$ et une fonction Φ, tels que :*

1. *$(x,\lambda) \mapsto (x,\Phi(x,\lambda))$ est un difféomorphisme de $I \times \Lambda$ dans $I \times \mathbb{R}^m$.*
2. *Pour tout $\lambda \in \Lambda$, $\Phi(.,\lambda)$ est une solution de (5.1) sur I.*
3. *Pour tout x de I et tout λ de Λ, $(\Phi(x,\lambda),\Phi_x(x,\lambda)) \in \Omega$.*
4. *Toute solution $y(x)$ de (5.1), telle que $(y(x_0),y'(x_0)) \in \Omega$ pour un $x_0 \in I$, est de la forme $y(x) = \Phi(x,\lambda)$ avec $\lambda \in \Lambda$.*

Un couple (x_0,y_0) est dit consistant si il existe un $\lambda \in \Lambda$ tel que $y_0 = \Phi(x_0,\lambda)$.

Ici, on supposera toujours que (5.1) est résoluble, ce qui deviendra évident pour les systèmes sous forme de Hessenberg qu'on considère par la suite. En particulier, pour toutes valeurs initiales consistantes (au sens de la définition précédente), on aura existence et unicité de la solution.

L'étude des équations algébro-différentielles (EDA) de la forme (5.1) est détaillée dans Brenan, Campbell & Petzold [BCP89]. Pour un brève introduction sur le sujet, on peut aussi consulter Jay [Jay94] ou März [Mär92].

5.2 Indice d'une EDA

5.2.1 Indice de différentiation

Soit $F^{(k)}$ le vecteur de $\mathbb{R}^{m(k+1)}$ constitué de F et de ses k premières dérivées par rapport à la variable d'intégration x :

$$F^{(k)}(y,\ldots,y^{(k+1)}) := \begin{bmatrix} F(y,y') \\ F_y(y,y')y' + F_{y'}(y,y')y'' \\ \vdots \\ \frac{d^k}{dx^k}F(y,y') \end{bmatrix}. \tag{5.2}$$

L'indice de différentiation est alors défini de la manière suivante :

Définition 11 (Gear [GP84]) *L'indice de différentiation ν_i de la variable y_i est le plus petit entier k tel que y_i' soit déterminée de façon unique et continue en fonction des y_j par l'équation $F^{(k)}(y,y',\ldots,y^{(k+1)}) = 0$. L'indice de différentiation ν de l'équation (5.1) est alors le plus grand des indices de différentiation des variables y_i.*

Exemple 5 *On considère l'EDA suivante (voir Hairer & Wanner [HW96]) :*

$$\begin{aligned} y_2' + y_1 &= \delta_1, \\ y_3' + y_2 &= \delta_2, \\ y_3 &= \delta_3. \end{aligned}$$

Les indices de différentiation des diverses composantes sont $\nu_1 = 3$, $\nu_2 = 2$ et $\nu_3 = 1$. Le problème est donc d'indice de différentiation 3.

5.2.2 Indice de perturbation

L'indice de perturbation est une mesure de la sensibilité aux perturbations de l'équation (5.1).

Définition 12 (Hairer, Lubich & Roche [HLR89]) *Soit y une solution de (5.1) sur un intervalle $[x_0,X]$. L'indice de perturbation de (5.1) est le plus petit entier k*

tel que pour toute fonction $\hat{y}$ de classe C^1 vérifiant

$$F(\hat{y}(x),\hat{y}'(x)) = \delta(x),$$

il existe une majoration de la forme

$$\|\hat{y}(x) - y(x)\| \leq C\left(\|\hat{y}(x_0) - y(x_0)\| + \sup_{\xi\in[x_0,X]} \|\int_{x_0}^{\xi} \delta(\tau)d\tau\| + \sum_{j=0}^{k-1} \sup_{\xi\in[x_0,X]} \|\delta^{(j)}(\xi)\|\right)$$

valable pour toute fonction δ de dérivées suffisamment petites et où C est une constante ne dépendant que de F et de $(X - x_0)$.

Les relations entre ces deux définitions sont abordées dans Hairer, Lubich & Roche [HLR89], pp. 13, ainsi que dans Gear [Gea90]. On peut cependant montrer que pour les systèmes semi-explicites sous forme de Hessenberg, les deux notions coïncident.

5.3 Structure des EDA

Un système est dit semi-explicite s'il est de la forme

$$\begin{cases} y'(x) &= f(y(x),z(x)), \\ 0 &= g(y(x),z(x)). \end{cases}$$

La variable y est alors désignée comme la variable différentielle et z comme la variable algébrique. Il convient de noter que l'on peut toujours se ramener à un système de cette forme en posant dans (5.1) :

$$\begin{cases} y' &= z, \\ 0 &= F(y(x),z(x)). \end{cases}$$

Définition 13 (Brenan, Campbell & Petzold [BCP89]) *Une équation différentielle algébrique semi-explicite est dite sous forme de Hessenberg de taille $r \geq 2$ si elle est de la forme*

$$\begin{cases} {y^1}' &= F^1(y^1,y^2), \\ {y^2}' &= F^2(y^1,y^2,y^3), \\ &\vdots \\ {y^{r-1}}' &= F^{r-1}(y^1,y^2,\ldots,y^{r-1},y^r), \\ 0 &= F^r(y^1), \end{cases}$$

avec $F^r_{y^1} \times F^1_{y^2} \times \cdots \times F^{r-1}_{y^r}$ supposée inversible au voisinage de la solution. Un système de taille 1 est par définition de la forme :

$$\left\{ \begin{array}{lcl} y'(x) & = & f(y(x),z(x)), \\ 0 & = & g(y(x),z(x)), \end{array} \right. .$$

avec g_z inversible.

Dans la suite, on s'intéresse uniquement aux systèmes semi-explicites sous forme de Hessenberg de taille 1, 2 ou 3, avec les notations suivantes, empruntées à Hairer, Lubich & Roche [HLR89] :

Indice 1 :

$$\left\{ \begin{array}{lcl} y'(x) & = & f(y(x),z(x)), \\ 0 & = & g(y(x),z(x)), \end{array} \right. \tag{5.3}$$

où g_z est supposée inversible dans un voisinage de la solution exacte (hypothèse (H_1)).

Indice 2 :

$$\left\{ \begin{array}{lcl} y'(x) & = & f(y(x),z(x)), \\ 0 & = & g(y(x)), \end{array} \right. \tag{5.4}$$

où $g_y f_z$ est supposée inversible dans un voisinage de la solution exacte (hypothèse (H_2)).

Indice 3 :

$$\left\{ \begin{array}{lcl} y'(x) & = & f(y(x),z(x)), \\ z'(x) & = & k(y(x),z(x),u(x)), \\ 0 & = & g(y(x)), \end{array} \right. \tag{5.5}$$

où $g_y f_z k_u$ est supposée inversible dans un voisinage de la solution exacte (hypothèse (H_3)).

La valeur initiale (y_0,z_0) ou (y_0,z_0,u_0) suivant les cas est dite consistante si elle vérifie la contrainte et éventuellement la (ou les) contrainte(s) cachée(s), i.e.

- $g(y_0,z_0) = 0$ pour l'indice 1.
- $g(y_0) = 0$ et $g_y(y_0)f(y_0,z_0) = 0$ pour l'indice 2.
- $g(y_0) = 0$, $g_y(y_0)f(y_0,z_0) = 0$ et $(g_{yy}(f,f) + g_y f_y f + g_y f_z k)|_{(y_0,z_0,u_0)}$ pour l'indice 3.

Alors, pour toute valeur initiale consistante, on montre l'existence locale et l'unicité d'une solution régulière (voir Hairer, Lubich & Roche [HLR89]).

5.4 Exemple : mécanique lagrangienne

Soit un système mécanique de coordonnées $(q_1, \dots, q_n)$ et $(u_1, \dots, u_n)$. Le vecteur q représente la position du système, le vecteur $u = q'$ sa vitesse. On suppose que ce système est soumis aux lois de la mécanique Newtonienne et qu'en outre, sa position est *contrainte* par les équations

$$g_1(q) = 0, \ \dots, g_m(q) = 0.$$

Le lagrangien de ce système s'écrit :

$$L = T - U - \lambda^T g,$$

où T est l'énergie cinétique du système, U son énergie potentielle et $\lambda = [\lambda_1, \dots, \lambda_m]^T$ le vecteur des *multiplicateurs de Lagrange*. Les équations du mouvement sont alors de la forme :

$$\frac{d}{dt}\left[\begin{array}{c} \nabla_u L \\ \nabla_{\lambda'} L \end{array}\right] = \left[\begin{array}{c} \nabla_q L \\ \nabla_\lambda L \end{array}\right].$$

Pour les systèmes mécaniques contraints, $T = T(u)$ avec T_{uu} symétrique définie

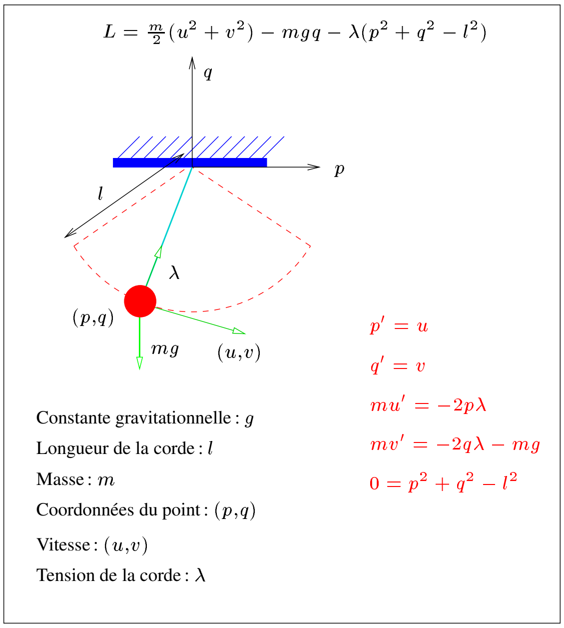


FIG. 5.1 – *Illustration : pendule simple*

positive et $U = U(q)$, de sorte que les équations deviennent :

$$\left\{ \begin{array}{lcl} q' & = & u, \\ T_{uu}u' & = & \nabla_q U - \left(\frac{\partial g}{\partial q}\right)^T \lambda, \\ 0 & = & g(q). \end{array} \right. \tag{5.6}$$

Sous cette forme, (5.6) est d'indice 3 dès lors que $g_q(T_{uu})^{-1}(g_q)^T$ est inversible, i.e. g_q de rang maximum. Si l'on remplace la dernière équation de (5.6) par sa dérivée par rapport à x, respectivement sa dérivée seconde,

$$0 = g_q u,$$

respectivement

$$0 = g_{qq}(u,u) + g_q(T_{uu})^{-1}\left(\nabla_q U - g_q)^T\lambda\right),$$

le système est d'indice 2 (hypothèse (H_2)), respectivement d'indice 1 (hypothèse (H_1)). Les contraintes dérivées de $g(q) = 0$ sont toutes satisfaites par la solution exacte du problème, pour peu que les conditions initiales soient consistantes. Ce n'est évidemment pas le cas de la solution numérique, qui ne satisfait que l'une d'entre elles, suivant la formulation choisie. Afin de réduire l'indice (et donc la difficulté du problème), tout en permettant que les contraintes d'origine $0 = g(q)$ soient satisfaites, Gear, Gupta & Leimkuhler ont introduit [GGL85] d'autres multiplicateurs dans la première équation de (5.6) :

$$\left\{ \begin{array}{lcl} q' & = & u - \left(\frac{\partial g}{\partial q}\right)^T \mu, \\ T_{uu}u' & = & \nabla_q U - \left(\frac{\partial g}{\partial q}\right)^T \lambda, \\ 0 & = & g(q), \\ 0 & = & \left(\frac{\partial g}{\partial q}\right) u. \end{array} \right.$$

Le système est alors d'indice 2 dès lors que $(\frac{\partial g}{\partial q})(T_{uu})^{-1}(\frac{\partial g}{\partial q})^T$ est inversible, i.e. $\frac{\partial g}{\partial q}$ de rang maximum. Cette formulation est particulièrement intéressante et justifie en partie que l'on s'intéresse dans cette thèse plus particulièrement aux systèmes d'indice 2 (les autres motivations sont liées au comportement des méthodes numériques pour les différents indices : voir chapitre suivant).

Chapitre 6

Méthodes de Runge-Kutta pour des équations algébro-différentielles

Dans ce chapitre, on privilégie le cas de l'indice 2, qui fait l'objet du chapitre suivant, i.e. des systèmes de la forme (5.4)

$$\begin{cases} y'(x) &= f(y(x),z(x)) \in \mathbb{R}^m, \\ 0 &= g(y(x)) \in \mathbb{R}^n, \end{cases}$$

où l'on suppose pour préciser les notations que y est un vecteur de $\mathbb{R}^m$, z un vecteur de $\mathbb{R}^n$, f une fonction C^k de $\mathbb{R}^{m+n}$ dans $\mathbb{R}^m$ et g une fonction C^{k+1} de $\mathbb{R}^m$ dans $\mathbb{R}^n$ avec $k \geq 1$. Cependant, on donnera chaque fois que possible les résultats correspondants pour les indices 1 et 3.

6.1 Définition d'une méthode de Runge-Kutta

Afin de déterminer le forme que revêt une méthode de Runge-Kutta appliquée à un système d'indice 2, on considère (5.4) comme limite du problème de perturbation singulière

$$P_\varepsilon : \begin{cases} y'(x) &= f(y(x),z(x)) \\ \varepsilon z'(x) &= g(y(x)) \end{cases} .$$

Pour une méthode de Runge-Kutta $\mathcal{R} = (A,b)$ appliquée à (P_ε), on a :

$$\begin{aligned}
Y_i &= y_0 + h\sum_{j=1}^{s} a_{i,j} f(Y_j,Z_j), i = 1,\dots,s,\\
Z_i &= z_0 + \frac{h}{\varepsilon}\sum_{j=1}^{s} a_{i,j} g(Y_j), i = 1,\dots,s,\\
y_1 &= y_0 + h\sum_{j=1}^{s} b_j f(Y_j,Z_j),\\
z_1 &= z_0 + \frac{h}{\varepsilon}\sum_{j=1}^{s} b_j g(Y_j).
\end{aligned}$$

Soient $\omega_{i,j}$ les coefficients de A^{-1} (on suppose A inversible). Il vient :

$$hg(Y_i) = \varepsilon\sum_{j=1}^{s}\omega_{i,j}(Z_j - z_0), i = 1,\dots,s,$$

et par suite la définition de z_1 devient indépendante de ε :

$$z_1 = (1 - \sum_{i,j=1}^{s} b_i\omega_{i,j})z_0 + \sum_{i,j=1}^{s} b_i\omega_{i,j}Z_j.$$

En faisant tendre ε vers 0, on obtient donc :

Définition 14 *Soit $\mathcal{R} = (A,b)$ une méthode de Runge-Kutta avec A inversible. Un pas de $\mathcal{R}$ pour le système (5.4) s'écrit :*

$$\begin{aligned}
Y_i &= y_0 + h\sum_{j=1}^{s} a_{i,j} f(Y_j,Z_j), i = 1,\dots,s, && (6.1)\\
0 &= g(Y_i), i = 1,\dots,s, && (6.2)\\
y_1 &= y_0 + h\sum_{j=1}^{s} b_j f(Y_j,Z_j), && (6.3)\\
z_1 &= (1 - \sum_{i,j=1}^{s} b_i\omega_{i,j})z_0 + \sum_{i,j=1}^{s} b_i\omega_{i,j}Z_j. && (6.4)
\end{aligned}$$

Remarque 11 *En procédant de façon similaire, on obtient la formulation d'une méthode de Runge-Kutta pour l'indice* 1 *(méthode directe) et pour l'indice* 3*.*

Remarque 12 *On peut écrire (6.1,6.2,6.3,6.4) sous la forme plus compacte :*

$$\begin{aligned} Y &= e \otimes y_0 + h(A \otimes I_{\mathbb{R}^m}) f(Y,Z), && (6.5)\\ 0 &= g(Y), && (6.6)\\ y_1 &= y_0 + h(b^T \otimes I_{\mathbb{R}^m}) f(Y,Z), && (6.7)\\ z_1 &= (1 - b^T A^{-1} e) z_0 + (b^T A^{-1} \otimes I_{\mathbb{R}^n}) Z, && (6.8) \end{aligned}$$

où $e = [1, \ldots, 1]^T \in \mathbb{R}^s$ *et*

$$\begin{aligned} Y &= [Y_1^T, \ldots, Y_s^T]^T,\\ Z &= [Z_1^T, \ldots, Z_s^T]^T,\\ f(Y,Z) &= [f(Y_1,Z_1)^T, \ldots, f(Y_s,Z_s)^T]^T,\\ g(Y) &= [g(Y_1)^T, \ldots, g(Y_s)^T]^T. \end{aligned}$$

Chaque pas de $\mathcal{R}$ implique la résolution du système non-linéaire constitué des équations (6.1,6.2). L'existence et l'unicité d'une solution sont assurées sous les hypothèses (H_2), $g(y_0) = \mathcal{O}(h^2)$ et $g_y(y_0) f(y_0,z_0) = \mathcal{O}(h)$ (voir théorème 4.1. de Hairer, Lubich & Roche [HLR89]).

Remarque 13 *Pour l'indice* 1, *(H_1) suffit. Pour l'indice* 3, *les hypothèses sont au contraire plus fortes (voir théorème 6.1. de [HLR89]).*

Le jacobien de ce système est de la forme

$$J = \begin{bmatrix} I_{\mathbb{R}^s} \otimes I_{\mathbb{R}^m} - hA \otimes f_y & -hA \otimes f_z \\ I_{\mathbb{R}^s} \otimes g_y & 0 \end{bmatrix}$$

et son inverse J^{-1} est de norme non bornée pour h tendant vers 0. Le théorème de Newton-Kantorovich ne peut donc pas être appliqué. On peut cependant montrer que la fonction Θ de l'itération du point fixe

$$\Theta(Y,Z) = \begin{bmatrix} Y \\ Z \end{bmatrix} - J^{-1} \begin{bmatrix} Y - e \otimes y_0 - h(A \otimes I_{\mathbb{R}^m}) f(Y,Z) \\ Z \end{bmatrix}$$

est bornée pour h tendant vers 0.

Remarque 14 *Pour l'indice* 1, *les hypothèses du théorème de Newton-Kantorovich sont vérifiées : appliquer* $\mathcal{R}$ *à un système algébro-différentiel d'indice* 1 *revient à appliquer cette même méthode au système différentiel ordinaire obtenu par le théorème des fonctions implicites. Pour l'indice* 3, *la fonction* Θ *correspondante n'est plus bornée et l'itération du point fixe devient très sensible aux perturbations (voir Hairer, Lubich & Roche [HLR89], pp. 92-95).*

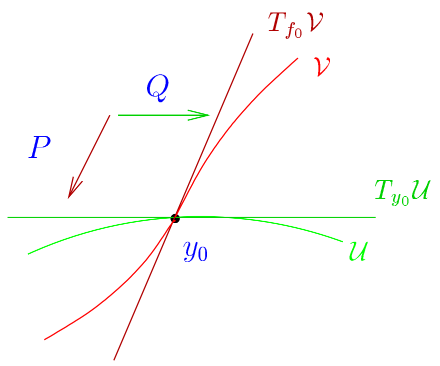


FIG. 6.1 – *Projections P et Q*

6.2 Principaux résultats de convergence

6.2.1 Erreur locale

Soit $(y(x),z(x))$ une solution de (5.4) telle que $(y(x_0),z(x_0)) = (y_0,z_0)$. Les erreurs locales $\delta e_y(x)$ et $\delta e_z(x)$ sont définies par :

$$\begin{array}{rcl} \delta e_y & := & y_1 - y(x+h), \\ \delta e_z & := & z_1 - z(x+h). \end{array}$$

On définit les projections (voir figure 6.1) :

$$Q(y,z) := \left(f_z(g_y f_z)^{-1} g_y\right)(y,z) \text{ et } P := I_{\mathbb{R}^m} - Q. \tag{6.9}$$

On est alors amené à considérer les arbres et fonctions associées définies ci-après.

Définition 15 *Les ensembles d'arbres $T2_y$ et $T2_z$ sont définis récursivement par :*

1. $\tau \in T2_y$,
2. $[t_1,\ldots,t_m,u_1,\ldots,u_n]_y \in T2_y$ *si et seulement si* $t_1,\ldots,t_m \in T2_y$ *et* $u_1, \ldots,$ $u_n \in T2_z$,
3. $[t_1,\ldots,t_m]_z \in T2_z$ *si et seulement si* $t_1,\ldots,t_m \in T2_y$ *et si* $m = 1$ *alors* $t_1 \neq [u]_y$ *où* $u \in T2_z$.

Par convention, un arbre de $T2_y$ sera désigné par la lettre t et sa racine représentée par un noeud "maigre" de couleur noire, un arbre de $T2_z$ par la lettre u et sa racine par un noeud "gras" de couleur orange. Par ailleurs, on appelle ordre et on note $\rho(t)$ la différence entre le nombre de noeuds maigres et le nombre de noeuds gras.

Définition 16 *La densité d'un arbre t, noté $\gamma(t)$, est définie récursivement par :*

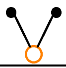
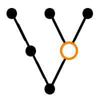
1. $\gamma(\tau) = 1$,

2. *Si* $t = [t_1, \ldots, t_m, u_1, \ldots, u_n]_y$ *alors* $\gamma(t) = \rho(t) \prod_{i=1}^{m} \gamma(t_i) \prod_{j=1}^{n} \gamma(u_j)$,
3. *Si* $u = [t_1, \ldots, t_m]_z$ *alors* $\gamma(u) = \frac{1}{\rho(u)+1} \prod_{i=1}^{m} \gamma(t_i)$.

Définition 17 *La fonction* Φ*, que nous appellerons fonction d'ordre, associée à la méthode* $\mathcal{R}$ *est définie récursivement sur les ensembles d'arbres par :*

1. $\Phi(\tau) = e$,
2. *Si* $t = [t_1, \ldots, t_m, u_1, \ldots, u_n]_y$ *alors* $\Phi(t) = \prod_{i=1}^{m} A\Phi(t_i) \prod_{j=1}^{n} \Phi(u_j)$,
3. *Si* $u = [t_1, \ldots, t_m]_z$ *alors* $\Phi(u) = A^{-1} \prod_{i=1}^{m} A\Phi(t_i)$.

Ci-dessous sont représentés quelques arbres et leurs fonctions associées (voir 6.1). On a alors le théorème fondamental suivant :

w	τ	 $[\tau,\tau]_z$	 $[[\tau]_y]_y$	 $[[\tau]_y,[\tau,\tau]_z]_y$
$\rho(t)$	1	1	3	4
$\gamma(t)$	1	$\frac{1}{2}$	6	4
$\Phi(t)$	e	$A^{-1}(Ae.Ae)$	A^2e	$(A^2e).\big(A^{-1}(Ae.Ae)\big)$

TAB. 6.1 – *Quelques arbres et fonctions associées pour les EDA d'indice 2*

Théorème 14 (Hairer, Lubich & Roche [HLR89]) *Soient* $\delta e_y(x)$ *et* $\delta e_z(x)$ *les erreurs locales respectivement pour la composante différentielle* y *et la composante algébrique* z *d'une méthode de Runge-Kutta* $\mathcal{R} = (A,b)$*. On a alors*

$$\begin{array}{lll} \delta e_y(x) = \mathcal{O}(h^{p+1}) & ssi & \forall\, t \in T2_y, \rho(t) \le p, \gamma(t) b^T \Phi(t) = 1, \\ \delta e_z(x) = \mathcal{O}(h^{q+1}) & ssi & \forall\, u \in T2_z, \rho(u) \le q, \gamma(u) b^T A^{-1} \Phi(u) = 1. \end{array}$$

6.2.2 Erreur globale

Théorème 15 (Hairer, Lubich & Roche [HLR89]) *On suppose* $g_y f_z$ *inversible au voisinage de la solution exacte* $(y(x),z(x))$ *du problème (5.4) et les valeurs initiales* $(y_0,z_0) = (y(x_0),z(x_0))$ *consistantes. Si la méthode de Runge-Kutta* $\mathcal{R} = (A,b)$ *est telle que*

$$\left\{ \begin{array}{ccc} \det A & \neq & 0, \\ |1 - b^T A^{-1} e| & < & 1 \end{array} \right.$$

et si l'erreur locale vérifie

$$\left\{ \begin{array}{lcl} \delta e_y(x) & = & \mathcal{O}(h^r), \\ P(x)\delta e_y(x) & = & \mathcal{O}(h^{r+1}), \end{array} \right. \tag{6.10}$$

où $P(x)$ est la projection (6.9) évaluée en $(y(x),z(x))$ alors la méthode $\mathcal{R}$ est convergente d'ordre r pour la composante différentielle, i.e.

$$\Delta E_y = y_N - y(x_N) = \mathcal{O}(h^r)$$

pour $x_N := x_0 + Nh \leq$ Const. Si de plus $\delta e_y(x) = \mathcal{O}(h^{r+1})$ alors $g(y_N) = \mathcal{O}(h^{r+1})$.

Remarque 15 *Lorsque $|1 - b^T A^{-1} e| = 1$, comme c'est le cas pour les méthodes de Gauss, l'ordre de convergence pour la composante différentielle et algébrique est donné par le théorème 4.9 de Hairer, Lubich & Roche [HLR89].*

Pour ce qui concerne la composante algébrique, le résultat suivant exprime le caractère essentiellement local de l'erreur sur la composante algébrique. Dans le chapitre suivant, cette caractéristique interviendra de différentes façons.

Théorème 16 (Hairer, Lubich & Roche [HLR89]) *On suppose $g_y f_z$ inversible au voisinage de la solution exacte $(y(x),z(x))$ du problème (5.4), les valeurs initiales $(y_0,z_0) = (y(x_0),z(x_0))$ consistantes. Si la méthode de Runge-Kutta $\mathcal{R} = (A,b)$ est telle que*

$$\left\{ \begin{array}{ccc} \det A & \neq & 0, \\ |1 - b^T A^{-1} e| & < & 1, \end{array} \right.$$

si y_N satisfait

$$\left\{ \begin{array}{lcl} y_N - y(x_N) & = & \mathcal{O}(h^r), \\ g(y_N) & = & \mathcal{O}(h^{r+1}), \end{array} \right.$$

et si

$$\delta e_z(x) = \mathcal{O}(h^r),$$

alors la méthode $\mathcal{R}$ est convergente d'ordre r pour la composante algébrique, i.e.

$$\Delta E_z = z_N - z(x_N) = \mathcal{O}(h^r)$$

pour $x_N := x_0 + Nh \leq$ Const.

Chapitre 7

Quelques résultats liés à l'indice 2

7.1 Une loi de composition

Soient $\mathcal{R}_1 = (A_1,b_1)$ et $\mathcal{R}_2 = (A_2,b_2)$ deux méthodes de Runge-Kutta. On a coutume de les représenter par leur tableau de coefficients, dit "tableau de Butcher" :

$$\mathcal{R}_1 = \begin{array}{c|c} c_1 & A_1 \\ \hline & b_1^T \end{array} \qquad \text{et} \qquad \mathcal{R}_2 = \begin{array}{c|c} c_2 & A_2 \\ \hline & b_2^T \end{array},$$

où $c_1 = A_1 e_1$, $e_1 = [1,\dots,1]^T \in I\!R^{s_1}$ et $c_2 = A_2 e_2$, $e_2 = [1,\dots,1]^T \in I\!R^{s_2}$. Si l'on considère deux pas d'intégration consécutifs, l'un rh par $\mathcal{R}_1$, l'autre $(1-r)h$ par $\mathcal{R}_2$ (dans cet ordre), on obtient un nouveau schéma numérique qui fournit une approximation de la solution exacte au point $x_0 + h$. Il s'avère, comme cela peut être facilement vérifié à partir des formules (6.5,6.6,6.7,6.8), que ce schéma est encore une méthode de Runge-Kutta $\mathcal{R}$ de tableau :

$$\begin{array}{c|cc} rc_1 & rA_1 & 0 \\ re_2 + (1-r)c_2 & re_2 b_1^T & (1-r)A_2 \\ \hline & rb_1^T & (1-r)b_2^T \end{array}.$$

Le théorème 14 fournit des conditions algébriques portant sur les matrices et vecteurs de coefficients pour que $\mathcal{R}_1$ ou $\mathcal{R}_2$ atteignent un ordre de convergence local donné. Ces conditions s'expriment en fonction des fonctions Φ_1 et Φ_2 définies sur l'ensemble des arbres de $T2_y \cup T2_z$. Qu'en est-il de la méthode composée $\mathcal{R}$?

Dans le cas des équations différentielles ordinaires, Butcher a été à l'origine d'une loi de composition opérant sur les arbres et permettant d'exprimer Φ en fonction de Φ_1 et de Φ_2 (voir Butcher [But87]) :

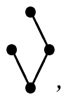
Théorème 17 (Butcher [But87]) *Soient $\mathcal{R}_1 = (A_1,b_1)$ et $\mathcal{R}_2 = (A_2,b_2)$ deux méthodes de Runge-Kutta, et Φ_1 et Φ_2 leurs fonctions d'ordre associées. Soit par ailleurs T l'ensemble des arbres de $T2_y$ ne possédant pas de noeud gras (seuls les arbres de ce type interviennent dans le cas des EDO). Alors, la fonction Φ correspondant à la méthode composée $\mathcal{R} = (1-r)\mathcal{R}_2 \, o \, r\mathcal{R}_1$ s'écrit :*

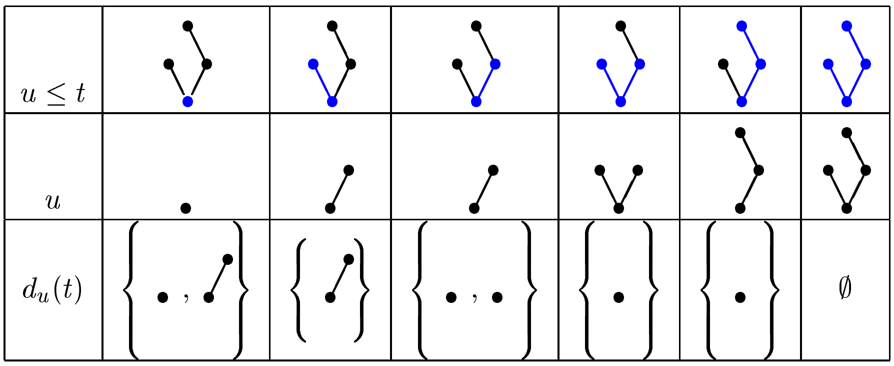
$$\forall t \in T, \; \Phi(t) = r^{\rho(t)-1} \begin{bmatrix} \Phi_1(t) \\ \Phi^{(2)}(t) \end{bmatrix} \tag{7.1}$$

où

$$\Phi^{(2)}(t) = \sum_{u \leq t} \left(\frac{1-r}{r} \right)^{\rho(u)} \left(\prod_{s \in d_u(t)} b_1^T \Phi_1(s) \right) \Phi_2(u). \tag{7.2}$$

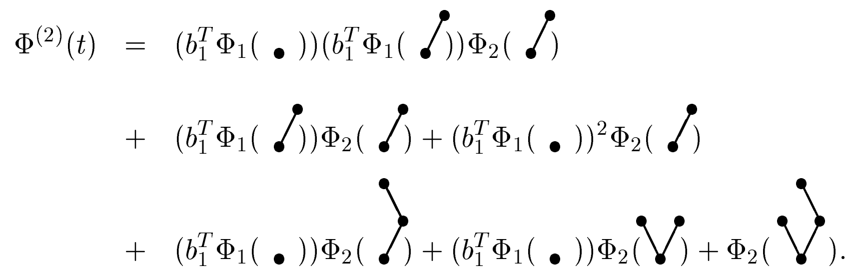
Ici, u désigne un sous-arbre de t et $d_u(t)$ l'ensemble des arbres obtenus lorsque u est extrait de t.

Exemple 6 *Pour l'arbre $t =$*  *, les différents sous-arbres u et les ensembles $d_u(t)$ correspondants sont regroupés dans le tableau 7.1. L'expression de $\Phi^{(2)}(t)$*



TAB. 7.1 – *Sous-arbres et ensembles différences de $[\tau,[\tau]]$*

est alors, par exemple pour $r = \frac{1}{2}$ *:*



Cette loi a par la suite donné naissance au concept de B-série, déjà évoqué, et a été réinterprétée en terme de composition de B-séries. Jay a étendu le concept de B-série aux EDA d'indice 3 (voir Jay [Jay94]). L'objet du présent chapitre est d'étudier le cas des EDA d'indice 2.

7.1.1 Résultat principal

Le théorème 17 s'étend de façon presque identique au cas des EDA d'indice 2 si l'on prend soin d'introduire de nouveaux arbres et de définir les notions de sous-arbre et d'ensemble différence adaptées à la nouvelle situation.

Théorème 18 (Chan & Chartier [CC96]) *Soient* $\mathcal{R}_1 = (A_1, b_1)$ *et* $\mathcal{R}_2 = (A_2, b_2)$ *deux méthodes de Runge-Kutta telles que* A_1 *et* A_2 *soient inversibles, et* Φ_1 *et* Φ_2 *leurs fonctions d'ordre associées. Soit par ailleurs* $ET2_y$ *et* $ET2_z$ *les sur-ensembles de* $T2_y$ *et* $T2_z$ *définis dans [CC96]. Alors, la fonction* Φ *correspondant à la méthode composée* $\mathcal{R} = (1-r)\mathcal{R}_2 \, o \, r\mathcal{R}_1$ *s'écrit :*

$$\forall\, t \in T2_y \cup T2_z,\ \Phi(t) = r^{\rho(t)-1} \begin{bmatrix} \Phi_1(t) \\ \Phi^{(2)}(t) \end{bmatrix} \tag{7.3}$$

où

$$\Phi^{(2)}(t) = \sum_{u \leq \tilde{t}} (-1)^{\sigma(u)} \left(\frac{1-r}{r}\right)^{\rho(u)} \left(\prod_{s \in d_u(\tilde{t})} (b_1^T \Phi_1(s)) \right) \Phi_2(u). \tag{7.4}$$

Ici, $\tilde{t}$ *désigne le "splitting" de* t*,* u *un sous-arbre de* $\tilde{t}$ *dans* $ET2_y \cup ET2_z$*,* $d_u(t)$ *l'ensemble des arbres obtenus lorsque* u *est extrait de* t *et* $\sigma(u)$ *le nombre de noeuds gras de* t*.*

Le lecteur intéressé trouvera les définitions précises de $\rho(t)$, $\tilde{t}$, u, $d_u(\tilde{t})$ et σ dans Chan & Chartier [CC96], ainsi qu'une formulation différente de ce résultat.

7.1.2 Applications

Il est aisé de vérifier à l'aide de cette loi de composition (théorème 18) qu'une méthode de Runge-Kutta composée de deux méthodes d'ordre p_1 et p_2 est d'ordre p supérieur ou égal à $\min(p_1,p_2)$. Cependant, il est des cas où p est strictement supérieur à $\max(p_1,p_2)$. Par exemple, si l'on prend pour $\mathcal{R}_1 = \mathcal{R}_2$ la méthode du point milieu

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

alors, il est aisé de vérifier que $p_1 = p_2 = 1$ tandis que $p = 2$. D'une façon générale, on a :

Théorème 19 (Chan & Chartier [CC96]) *Soit $\mathcal{R} = (A,b)$, $A \in I\!R^{s\times s}$, une méthode de Runge-Kutta symétrique, telle que A soit inversible. Alors on a :*

$$s \text{ impair} \implies \exists\, q \in \mathbb{N}, \delta e_y = \mathcal{O}(h^{2q+1}).$$

En particulier, les méthodes de Gauss avec un nombre impair d'étapes internes sont d'ordre de convergence (global) pair lorsqu'elles sont utilisées à pas constant. Ce résultat est également démontré dans Hairer, Lubich & Roche [HNW87] par une technique de développement asymptotique de l'erreur globale.

Comme il a été mentionné plus avant, il est possible de restaurer l'ordre de convergence habituel (celui des EDO) d'une méthode de Runge-Kutta appliquée à (5.4) en projetant la solution numérique sur la contrainte, et ce après chaque pas. La preuve de ce résultat s'appuie sur le théorème suivant :

Théorème 20 (Chan & Chartier [CC96]) *Soit $\mathcal{R} = (A,b)$ une méthode de Runge-Kutta avec A inversible. Si $\mathcal{R}$ satisfait les conditions simplificatrices $B(p)$, $C(\eta)$ et $D(\xi)$ (voir Butcher [But87]) avec $p \leq 2\eta$ et $p \leq \eta + \xi + 1$ (c'est le cas des méthodes de Radau IA et de Gauss en particulier), alors :*

$$\delta e_y = -(f_z(g_y f_z)^{-1})(y(x_0+h),z(x_0+h))\theta(x_0+h) + \mathcal{O}(h^{p+1}),$$

où θ est une fonction régulière telle que $\theta(x_0+h) = \mathcal{O}(h^{\eta+1})$.

Plusieurs démonstrations ont été données antérieurement (voir Ascher & Petzold [AP91] ou Hairer & Wanner [HW96]). Celle de Chan & Chartier [CC96] repose sur le théorème 18.

	indice 0 (EDO)	indice 1	indice 2	indice 3
y	h^{2s-1}	h^{2s-1}	h^{2s-1}	h^{2s-1}
z	-	h^{2s-1}	h^s	h^s
u	-	-	-	h^{s-1}

TAB. 7.2 – *Ordres de convergence pour les méthodes de Radau IIA à s étapes internes*

7.2 Une technique de «smoothing»

L'application des théorèmes 14, 15 et 16 au cas des méthodes de Radau IIA (voir Hairer, Wanner & Nørsett [HNW93]) conduit aux ordres de convergence donnés dans la quatrième colonne du tableau 7.2. En regard des autres grandes classes de méthodes de Runge-Kutta (Gauss, RadauIA, ...), les méthodes de Radau IIA semblent particulièrement bien adaptées à la résolution des EDA. Pour un nombre d'étapes s fixé, nombre qui constitue une mesure du coût, elles possèdent en effet les ordres de convergence les plus élevées. Elles sont en outre "stiffly accurate", c'est-à-dire telles que la dernière ligne de A se confond avec b^T. Cette propriété implique en particulier que l'on a $y_1 = Y_s$ dans (6.7) et $z_1 = Z_s$ dans (6.8). La solution numérique produite par la méthode satisfait ainsi la contrainte $g(y_1) = 0$.

Il reste que l'ordre de convergence de la composante algébrique z est considérablement plus faible que celui de la composante différentielle. Ce phénomène, dit de "réduction d'ordre", est pénalisant dans certaines situations où une valeur précise de z est nécessaire. C'est le cas par exemple dans certaines applications physiques. Différentes techniques permettent de restaurer l'ordre de convergence habituel (i.e. $2s-1$) :

1. on peut projeter la solution sur la contrainte cachée, i.e. résoudre l'équation $g_y(y_1)f(y_1,\tilde{z}_1) = 0$ et remplacer z_1 par $\tilde{z}_1$. Cette solution exige de connaître une approximation à $\mathcal{O}(h^p)$ près de $g_y(y_1)$.
2. on peut considérer la composition sur plusieurs pas de la méthode de Radau IIA et construire ainsi une approximation de z_1 d'ordre plus élevé. C'est la technique qui est développée dans [AC98b] et exposée ci-après.

7.2.1 Modification du schéma numérique

Lorsqu'on applique une méthode de Radau IIA, $\mathcal{R} = (A,b)$, à un système différentiel algébrique d'indice deux, le schéma numérique est donné par les équations (6.5,6.6,6.7,6.8). Cependant, comme mentionné ci-avant, $\mathcal{R}$ est telle que $1 - b^T A^{-1} e = 0$ de sorte que (6.8) devient ici :

$$z_1 = (b^T A^{-1} \otimes I_{I\!R^n})Z.$$

Rien n'interdit de remplacer le vecteur b^TA^{-1} par un vecteur w^T ad hoc. Le schéma numérique obtenu devient alors :

$$\left\{\begin{array}{lcl} Y & = & e\otimes y_0 + h(A\otimes I_{\mathbb{R}^m})f(Y,Z), \\ 0 & = & g(Y), \\ y_1 & = & y_0 + h(b^T\otimes I_{\mathbb{R}^m})f(Y,Z), \\ z_1 & = & (w^T\otimes I_{\mathbb{R}^n})Z, \end{array}\right. \tag{7.5}$$

et l'ordre de convergence de la composante différentielle est inchangé. On note $\mathcal{R}_w = (A,b,w)$ la méthode ainsi obtenue et

$$\begin{array}{c|c} c & A \\ \hline & b^T \\ & w^T \end{array}$$

son tableau. Si on choisit les composantes de w de façon à maximiser l'ordre de l'erreur δe_z, il est clair que $w^T = b^TA^{-1}$ et $\mathcal{R}_w$ se confond avec $\mathcal{R}$.

Pour obtenir davantage de degrés de liberté, il est nécessaire de considérer la méthode $\mathcal{R}^\sigma$ obtenue par composition σ fois de la méthode $\mathcal{R}$. Le pas d'intégration h_i étant variable, on introduit les ratios $r_i = h_i/H$, $i = 1,\ldots,\sigma$ avec $H = \sum_{i=1}^{\sigma} h_i$. La méthode composée

$$\mathcal{R}^\sigma = (r_1H\mathcal{R})\, o\, (r_2H\mathcal{R})\, o\, \ldots\, o\, (r_\sigma H\mathcal{R})$$

est alors caractérisée par le tableau

$$\begin{array}{c|c} \mathcal{C} & \mathcal{A} \\ \hline & \mathcal{B}^T \end{array}$$

ou sous forme développée

$$\begin{array}{c|cccc} r_1c & r_1A & 0 & \ldots & 0 \\ r_2c + r_1e & r_1eb^T & r_2A & \ddots & \vdots \\ \vdots & \vdots & & & 0 \\ r_\sigma c + (\sum_{k=1}^{\sigma-1} r_k)e & r_1eb^T & \ldots & r_{\sigma-1}eb^T & r_\sigma A \\ \hline & r_1b^T & r_2b^T & \ldots & r_\sigma b^T \end{array}$$

Comme précédemment, on adjoint au tableau le vecteur $\mathcal{W} \in \mathbb{R}^{\sigma\times s}$. Pour $\sigma > 1$ suffisamment grand, le nombre de degrés de liberté autorise un ordre supérieur dans $\delta e_z = \mathcal{O}(h^{q+1})$. L'erreur en z étant de nature purement locale, cela suffit à accroître l'ordre de *convergence* de la méthode. La technique est schématisée sur la figure 7.1.

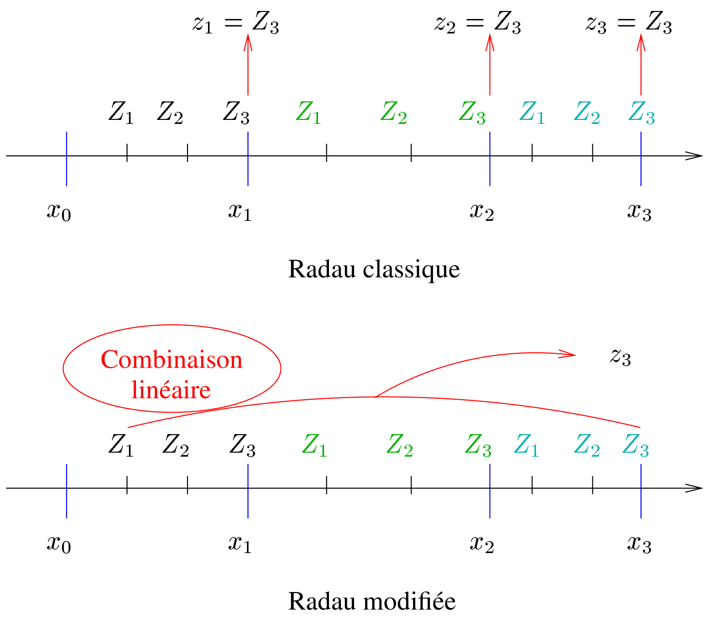


FIG. 7.1 – *Modification de la méthode de Radau IIA*

7.2.2 Construction de $\mathcal{W}$

Corollaire 3 (Aubry & Chartier [AC98b]) *Pour une méthode de Runge-Kutta $\mathcal{R}_w = (A,b,w)$, on considère δe_y et δe_z les erreurs locales sur la composante différentielle y et sur la composante algébrique z obtenue par (7.5). Alors, on a*

$$\delta e_y = O(h^{p+1}) \quad ssi \quad \forall t \in T2_y, \rho(t) \leq p, \gamma(t) b^T \Phi(t) = 1, \qquad (7.6)$$
$$\delta e_z = O(h^{q+1}) \quad ssi \quad \forall u \in T2_z, \rho(u) \leq q, \gamma(u) w^T \Phi(u) = 1. \qquad (7.7)$$

Exemple 7 *La méthode de Radau IIA à deux étapes s'écrit :*

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & \frac{-1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$

Elle est d'ordre (global) 3 en y et 2 en z. Pour obtenir une méthode d'ordre 3 en z (comme en y) $\mathcal{W}$ doit satisfaire les équations suivantes (données par le théorème 18) :

$$\begin{array}{lcll} \sum_{i=1}^{\sigma s} \mathcal{W}_i & = & 1 & (0), \\ \sum_{i=1}^{\sigma s} \mathcal{W}_i \mathcal{C}_i & = & 1 & (1), \end{array} \qquad \begin{array}{lcll} \sum_{i=1}^{\sigma s} \mathcal{W}_i \mathcal{C}_i^2 & = & 1 & (2), \\ \sum_{i,j=1}^{\sigma s} \mathcal{W}_i \mathcal{A}_{i,j}^{-1} \mathcal{C}_j^3 & = & 3 & (3). \end{array}$$

La valeur $\sigma = 2$ conduit alors à un système de 4 équations à 4 inconnues, dont on peut exprimer l'unique solution (pour $r \in]0,1]$) en fonction du ratio $r = \frac{h_n}{h_{n-1}}$:

$$\begin{array}{rclcrcl} \mathcal{W}_1 &=& \frac{3}{2r}\frac{r^5-4r^4+7r^3-7r^2+4r-1}{4r^3-8r^2+5r-2}, & & \mathcal{W}_3 &=& \frac{1}{2r}\frac{r^5-6r^4+11r^3-5r^2-4r+3}{4r^3-8r^2+5r-2}, \\ \mathcal{W}_2 &=& -\frac{3}{2}\frac{r^4-6r^3+12r^2-10r+3}{4r^3-8r^2+5r-2}, & & \mathcal{W}_4 &=& -\frac{1}{2}\frac{r^4-8r^3+12r^2-6r+3}{4r^3-8r^2+5r-2}. \end{array}$$

7.2.3 Application à la méthode à 3 étapes internes

La méthode de Radau IIA à 3 étapes internes s'écrit :

$$\mathcal{R} = \begin{array}{c|ccc} \frac{4-\sqrt{6}}{10} & \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\ \frac{4+\sqrt{6}}{10} & \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\ 1 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \\ \hline & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \end{array}.$$

Afin d'obtenir une méthode d'ordre de convergence 5 en z, le vecteur $\mathcal{W}$ doit vérifier un système linéaire (S) composé de 10 équations. Pour $\sigma = 3$, $\mathcal{W}$ ne comporte que neuf degrés de liberté. On montre cependant :

Théorème 21 (Aubry & Chartier [AC98b]) *Pour toutes valeurs non nulles de r_1 et r_2, le système (S) possède une solution.*

La méthode de Radau IIA est au coeur du code RADAU5 développé par Hairer & Wanner (adresse URL : "http://www.unige.ch/math/folks/hairer/softwares.html") et considéré aujourd'hui comme l'un des meilleurs codes de résolution des EDA d'indice un et deux (en termes de rapidité et de fiabilité). Ce code a été modifié dans Aubry & Chartier [AC98b] afin d'inclure l'amélioration proposée ici. Le code correspondant est disponible à l'adresse "http://www.irisa.fr/aladin/perso/chartier.html" (code RADAU5M).

7.2.4 Expérience numérique : problème des sept corps

Le problème des sept corps est décrit dans Hairer & Wanner [HW96]. On applique les codes RADAU5 et RADAU5M à la formulation d'indice deux suivante :

$$\left\{ \begin{array}{lcl} q'(x) &=& v(x), \\ v'(x) &=& M(q(x))^{-1}\left(f(q(x),v(x)) - G^T(q(x))\lambda(x)\right), \\ 0 &=& G(q(x))v(x), \end{array} \right.$$

pour $x \in [0,3.10^{-2}]$ et des conditions initiales consistantes.

La figure 7.2 représente le nombre de chiffres significatifs de la composante algébrique en fonction du temps CPU. L'erreur prise en considération est l'erreur maximum sur les sorties “continues” aux instants $x_i = x_{i-1} + 0.0003$.

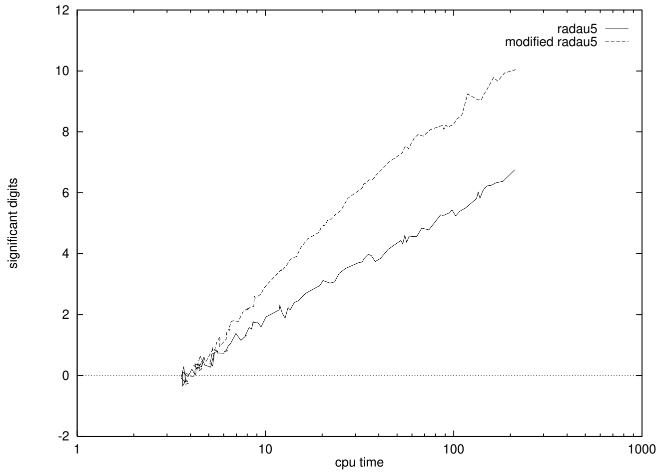


FIG. 7.2 – *Précision en fonction du temps* CPU *pour le problème des sept corps*

7.3 Super-convergence des méthodes violant la contrainte

Comme cela a été mentionné en divers endroits de cette thèse, les méthodes de Runge-Kutta projetées [1] permettent de remédier à la réduction d'ordre sur la composante différentielle. L'idée est de projeter la solution numérique après chaque pas d'intégration sur la variété $\mathcal{U} := \{y \,|\, g(y) = 0\}$. L'objet de ce chapitre est de montrer que pour les méthodes de Radau IA et les méthodes de Gauss, il n'est pas nécessaire d'effectuer la projection après chaque intégration mais uniquement après la dernière. Comme suggéré dans Burrage & Chan [BC93] ou Chan [Cha89], cet ultime pas d'intégration peut aussi être remplacé par un pas d'une méthode de Radau IIA de même ordre que la méthode de Radau IA. On obtient alors un ordre de convergence de $2s - 2$.

7.3.1 Arbres spéciaux

Définition 18 *Les ensembles d'arbres spéciaux $ST2_y$ et $ST2_z$ sont définis récursivement par :*

1. $\emptyset_y \in ST2_y$,
2. $[\tau_x, \ldots, \tau_x, \omega]_y \in ST2_y$ *ssi* $\omega \in ST2_y \cup ST2_z$,
3. $[\underbrace{\tau_x, \ldots, \tau_x}_{m}, t]_z \in ST2_z$ *ssi* $t \in ST2_y$ *et soit* $m > 0$ *ou* $m = 0$ *et* $t \neq [u]_y$ *avec* $u \in ST2_z$.

1. Les méthodes projetées ont été introduites par Ascher & Petzold [AP91]

Par convention, on prend $\tau_y = \bullet = [\emptyset_y]_y$ *et* $\tau_z = \circ = [\emptyset_y]_z$.

Le nouveau type de nœuds τ_x sera représenté par une flèche bleue. Des exemples d'arbres spéciaux et de leurs fonctions associées sont donnés dans le tableau 7.3.1.

Définition 19 *Soit ρ_0 un entier non nul. La fonction ρ est définie récursivement sur $ST2_y \cup ST2_z$ par :*

1. $\rho(\emptyset_y) = \rho_0$,
2. $\rho([\underbrace{\tau_x, \ldots, \tau_x}_{m}, \omega]_y) = m + 1 + \rho(\omega)$,
3. $\rho([\underbrace{\tau_x, \ldots, \tau_x}_{m}, t]_z) = m - 1 + \rho(t)$.

$\rho(\omega)$ *est l'*ordre *de* ω.

$\rho(\emptyset_y)$ est pris égal à ρ_0, où ρ_0 sera ici l'ordre de la perturbation Δy_0 sur la valeur initiale y_0.

Définition 20 *La fonction α est définie récursivement sur $ST2_y \cup ST2_z$ par :*

1. $\alpha(\emptyset_y) = 1$,
2. $\alpha([\underbrace{\tau_x, \ldots, \tau_x}_{m}, \omega]_y) = \begin{pmatrix} \rho(\omega) + m \\ \rho(\omega) \end{pmatrix} \alpha(\omega)$,
3. $\alpha([\underbrace{\tau_x, \ldots, \tau_x}_{m}, t]_z) = \begin{pmatrix} \rho(t) + m \\ \rho(t) \end{pmatrix} \alpha(t)$.

$\alpha(\omega)$ *est le nombre de façon d'étiqueter de manière monotone un arbre* ω.

Définition 21 *La fonction γ est définie récursivement sur $ST2_y \cup ST2_z$ par :*

1. $\gamma(\emptyset_y) = \rho_0!$,
2. $\gamma([\underbrace{\tau_x, \ldots, \tau_x}_{m}, \omega]_y) = (\rho(\omega) + m + 1)\, \gamma(\omega)$,
3. $\gamma([\underbrace{\tau_x, \ldots, \tau_x}_{m}, t]_z) = \frac{1}{(\rho(t)+m)}\, \gamma(t)$.

$\gamma(\omega)$ *est appelée* densité *de* ω.

Définition 22 *La fonction Φ est définie récursivement sur $ST2_y \cup ST2_z$ par :*

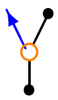
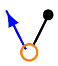
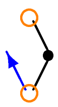
1. $\Phi(\emptyset_y) = A^{-1}$,
2. $\Phi([\underbrace{\tau_x, \ldots, \tau_x}_{m}, t]_y) = C^m A \Phi(t)$,
3. $\Phi([\underbrace{\tau_x, \ldots, \tau_x}_{m}, u]_y) = C^m \Phi(u)$,
4. $\Phi([\underbrace{\tau_x, \ldots, \tau_x}_{m}, t]_z) = A^{-1} C^m A \Phi(t)$,

avec $C = \mathrm{diag}(c_1,\ldots,c_s)$. Φ *est dite* fonction d'ordre.

Définition 23 *La fonction F est définie récursivement sur $ST2_y \cup ST2_z$ par :*

1. $F(\emptyset_y) = I_{\mathbb{R}^m}$,
2. $F([\underbrace{\tau_x,\ldots,\tau_x}_{m},t]_y) = \left.\frac{\partial^m f_y(y(x),z(x))}{\partial x^m}\right|_{x_0} F(t)$,
3. $F([\underbrace{\tau_x,\ldots,\tau_x}_{m},u]_y) = \left.\frac{\partial^m f_z(y(x),z(x))}{\partial x^m}\right|_{x_0} F(u)$,
4. $F([\underbrace{\tau_x,\ldots,\tau_x}_{m},t]_z) = (-g_y f_z)^{-1}\,(y(x_0),z(x_0))\,\left.\frac{\partial^m g_y(y(x),z(x))}{\partial x^m}\right|_{x_0} F(t)$.

$F(\omega)$ *est l'*élément différentiel *associé à* ω.

ω	$\rho(\omega)$	$\alpha(\omega)$	$\gamma(\omega)$	$\Phi(\omega)$	$F(\omega)$
$\emptyset_y$	ρ_0	1	$\rho_0!$	A^{-1}	$I_{\mathbb{R}^m}$
	ρ_0+1	1	$(\rho_0+1)!$	I_s	f_y
	ρ_0	1	$\rho_0!$	A^{-1}	$f_z(-g_y f_z)^{-1}g_y = Q$
	ρ_0+2	1	$(\rho_0+2)!$	A	$f_y f_y$
	ρ_0+1	1	$(\rho_0+1)!$	I_s	$f_z(-g_y f_z)^{-1}g_y f_y$
	ρ_0+2	ρ_0+2	$(\rho_0+1)!$	$A^{-1}CA$	$f_z(-g_y f_z)^{-1}g_y^{(1)} f_y$
	ρ_0+1	ρ_0+2	$\frac{(\rho_0+1)!}{(\rho_0+2)}$	$A^{-1}CA$	$(-g_y f_z)^{-1}g_y^{(1)} f_y$
	ρ_0	ρ_0+1	$\frac{\rho_0!}{(\rho_0+1)}$	$A^{-1}C$	$(-g_y f_z)^{-1}g_y^{(1)} f_z(-g_y f_z)^{-1}g_y$

TAB. 7.3 – *Exemples d'arbres spéciaux et de leurs fonctions associées*

7.3.2 Influence des perturbations

Le résultat présenté ci-dessous est une version plus précise du théorème 4.2 de [HLR89].

Lemme 1 (Aubry & Chartier [AC96]) *Soit $\mathcal{R} = (A,b)$ une méthode de Runge-Kutta à s étapes internes avec une matrice A inversible. En complément du système (6.5,6.6,6.7,6.8), on considère le système perturbé*

$$\begin{aligned} \hat{Y} &= e \otimes \hat{y}_0 + h(A \otimes I_{I\!R^m}) f(\hat{Y},\hat{Z}), && (7.8)\\ 0 &= g(\hat{Y}), && (7.9)\\ \hat{y}_1 &= \hat{y}_0 + h(b^T \otimes I_{I\!R^m}) f(\hat{Y},\hat{Z}), && (7.10)\\ \hat{z}_1 &= (1 - b^T A^{-1} e)\hat{z}_0 + (b^T A^{-1} \otimes I_{I\!R^n})\hat{Z}, && (7.11) \end{aligned}$$

Si la méthode de Runge-Kutta $\mathcal{R}$ vérifie $C(q)$ avec $q \geq 1$, et si

$$\begin{aligned} y_0 &= y(x_0) + \mathcal{O}(h^\eta), && (7.12)\\ \hat{y}_0 &= y(x_0) + \mathcal{O}(h^\eta), \eta \geq 2, && (7.13) \end{aligned}$$

où $(y(x),z(x))$ est la solution exacte à l'instant x, alors on a

$$\begin{aligned} h\Delta F &= \sum_{t \in ST2_y,\, \rho(t) \leq \mu+\eta,\, t \neq \emptyset_y} \frac{h^{\rho(t)-\eta}}{\rho(t)!} \alpha(t)\gamma(t)(\Phi(t)e \otimes F(t)\Delta y_0)\\ &+ \mathcal{O}(h^{\mu+1}\|\Delta y_0\|),\\ \Delta Z &= \sum_{u \in ST2_z,\, \rho(u) \leq \mu-1+\eta} \frac{h^{\rho(u)-\eta}}{\rho(u)!} \alpha(u)\gamma(u)(\Phi(u)e \otimes F(u)\Delta y_0)\\ &+ \mathcal{O}(h^{\mu}\|\Delta y_0\|),\\ \Delta y_1 &= \Delta y_0 + \sum_{t \in ST2_y,\, \rho(t) \leq \mu+\eta,\, t \neq \emptyset_y} \frac{h^{\rho(t)-\eta}}{\rho(t)!} \alpha(t)\gamma(t) \left(b^T \Phi(t) e\right) F(t)\Delta y_0\\ &+ \mathcal{O}(h^{\mu+1}\|\Delta y_0\|), \end{aligned}$$

où $\rho_0 = \eta$,

$$\begin{aligned} \Delta y_0 &= \hat{y}_0 - y_0,\\ \Delta y_1 &= \hat{y}_1 - y_1,\\ \Delta F &= \left((f(\hat{Y}_1,\hat{Z}_1) - f(Y_1,Z_1))^T, \ldots, (f(\hat{Y}_s,\hat{Z}_s) - f(Y_s,Z_s))^T\right)^T,\\ \Delta Z &= \left((\hat{Z}_1 - Z_1)^T, \ldots, (\hat{Z}_s - Z_s)^T\right)^T \end{aligned}$$

et $\mu = \min(\eta - 2, q - 1)$.

La technique de démonstration est similaire à celle utilisée dans Jay [Jay93, Jay94].

démonstration : Soit $\Delta Y = \Big((\hat{Y}_1 - Y_1)^T, \ldots, (\hat{Y}_s - Y_s)^T\Big)^T$. On considère les fonctions $f_y(y(x_0+c_ih),z(x_0+c_ih))$, $f_z(y(x_0+c_ih),z(x_0+c_ih))$ et $g_y(y(x_0+c_ih)$ comme des fonctions de $x = x_0 + c_ih$. Leurs développements de Taylor en x_0 donne :

$$\begin{aligned}
\Delta Y &= (e \otimes \Delta y_0) + (A \otimes I_{\mathbb{R}^m})(h\Delta F), && (7.14)\\
(h\Delta F) &= \sum_{k=0}^{\mu} \frac{h^{k+1}}{k!}(C^k \otimes f_y^{(k)})\Delta Y + \sum_{k=0}^{\mu} \frac{h^{k+1}}{k!}(C^k \otimes f_z^{(k)})(\Delta Z)\\
&+ \mathcal{O}(h^{\mu+1}\|\Delta y_0\|), && (7.15)\\
0 &= \sum_{k=0}^{\mu} \frac{h^{k}}{k!}(C^k \otimes g_y^{(k)})\Delta Y + \mathcal{O}(h^{\mu+1}\|\Delta y_0\|). && (7.16)
\end{aligned}$$

On cherche à exprimer $h\Delta F$ et ΔZ sous la forme :

$$\begin{aligned}
h\Delta F &= \sum_{k=0}^{\mu} \frac{h^k}{(k+\eta)!}\phi^k(e \otimes \Delta y_0) + \mathcal{O}(h^{\mu+1}\|\Delta y_0\|), && (7.17)\\
\Delta Z &= \sum_{k=-1}^{\mu-1} \frac{h^k}{(k+\eta)!}\psi^k(e \otimes \Delta y_0) + \mathcal{O}(h^{\mu}\|\Delta y_0\|). && (7.18)
\end{aligned}$$

En insérant (7.14) dans (7.15), il vient :

$$\begin{aligned}
(h\Delta F) &= \sum_{k=0}^{\mu} \frac{h^{k+1}}{k!}(C^k \otimes f_y^{(k)})(e \otimes \Delta y_0) + \sum_{k=0}^{\mu} \frac{h^{k+1}}{k!}(C^k A \otimes f_y^{(k)})(h\Delta F)\\
&\quad + \sum_{k=0}^{\mu} \frac{h^{k+1}}{k!}(C^k \otimes f_z^{(k)})\Delta Z + \mathcal{O}(h^{\mu+1}\|\Delta y_0\|).
\end{aligned}$$

On obtient ainsi :

$$\begin{aligned}
\phi^0 &= \eta(I_s \otimes f_z)\psi^{-1}, && (7.19)\\
\phi^j &= \sum_{k=0}^{j}(j+\eta)\begin{pmatrix} j-1+\eta \\ k \end{pmatrix}(C^k \otimes f_z^{(k)})\psi^{j-k-1} +\\
&\quad \frac{(j+\eta)!}{(j-1)!}(C^{j-1} \otimes f_y^{(j-1)}) && (7.20)\\
&\quad + \sum_{k=0}^{j-1}(j+\eta)\begin{pmatrix} j-1+\eta \\ k \end{pmatrix}(C^k A \otimes f_y^{(k)})\phi^{j-k-1},\ j = 1, \ldots, \mu.
\end{aligned}$$

De même, en insérant (7.14) dans (7.16) :

$$0 = \sum_{k=0}^{\mu} \frac{h^k}{k!} (C^k \otimes g_y^{(k)})(e \otimes \Delta y_0) + \sum_{k=0}^{\mu} \frac{h^k}{k!} (C^k A \otimes g_y^{(k)})(h\Delta F) + \mathcal{O}(h^{\mu+1} \|\Delta y_0\|),$$

Par comparaison des termes de même puissance en h, on tire :

$$0 = (C^j \otimes g_y^{(j)}) + \sum_{k=0}^{j} \frac{j!}{k!(j-k+\eta)!} (C^k A \otimes g_y^{(k)}) \phi^{j-k},\ j = 0, \ldots, \mu.$$

En tenant compte de (7.20) dans l'expression précédente, on peut alors obtenir les relations de récurrence suivantes :

$$\psi^{-1} = (\eta - 1)!(A^{-1} \otimes (-g_y f_z)^{-1} g_y), \tag{7.21}$$

$$\begin{aligned} \psi^{j-1} = {} & \frac{(j+\eta-1)!}{j!} (A^{-1} C^j \otimes (-g_y f_z)^{-1} g_y^{(j)}) \\ & + \frac{(j+\eta-1)!}{(j-1)!} (C^{j-1} \otimes (-g_y f_z)^{-1} g_y f_y^{(j-1)}) \\ & + \sum_{k=1}^{j} \frac{1}{(j+\eta)} \begin{pmatrix} j+\eta \\ k \end{pmatrix} (A^{-1} C^k A \otimes (-g_y f_z)^{-1} g_y^{(k)}) \phi^{j-k} \\ & + \sum_{k=1}^{j} \begin{pmatrix} j-1+\eta \\ k \end{pmatrix} (C^k \otimes (-g_y f_z)^{-1} g_y f_z^{(k)}) \psi^{j-k-1} \\ & + \sum_{k=0}^{j-1} \begin{pmatrix} j-1+\eta \\ k \end{pmatrix} (C^k A \otimes (-g_y f_z)^{-1} g_y f_y^{(k)}) \phi^{j-k-1},\ j = 1, \ldots, \mu. \end{aligned} \tag{7.22}$$

Une récurrence sur j conduit finalement aux développements annoncés. $\square$

7.3.3 Structure de l'erreur pour les méthodes de Radau IA

Théorème 22 (Aubry & Chartier [AC96]) *Sous les hypothèses du lemme 1, en supposant que $\mathcal{R} = (A,b)$ est la méthode de Radau IA à s étapes internes et pour $\eta = s$, la perturbation Δy_1 du lemme 1 prend la forme*

$$\Delta y_1 = P_0 \Delta y_0 + \mathcal{O}(h \|P_0 \Delta y_0\|) + \mathcal{O}(h^{s-1} \|Q_0 \Delta y_0\|) \tag{7.23}$$

où Q_0 et P_0 sont les projections définies en (6.9) et évaluées en $(y(x_0),z(x_0))$.

Afin d'obtenir une confirmation numérique du résultat montré précédemment et illustré figure 7.3, on considère le système suivant extrait de Jay [Jay93] :

$$\left\{ \begin{array}{lcl} y_1'(x) & = & y_1(x) y_2^2(x) z^2(x), \\ y_2'(x) & = & y_1^2(x) y_2^2(x) - 3 y_2^2(x) z(x), \\ 0 & = & y_1^2(x) y_2(x) - 1, \end{array} \right. \tag{7.24}$$

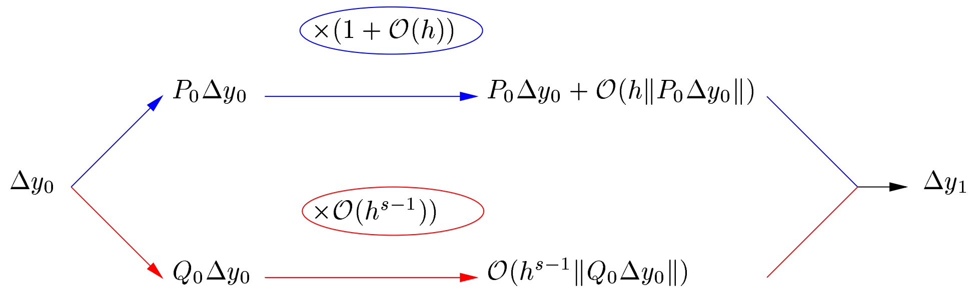


FIG. 7.3 – *Action de la méthode Radau IA sur une perturbation* Δy_0

dont la solution exacte est $(y_1(x),y_2(x),z(x)) = (e^x,e^{-2x},e^{2x})$. On prend en outre

$$y_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + h^s \begin{pmatrix} 1 \\ -1000 \end{pmatrix} \text{ and } \hat{y}_0 = y_0 + 10^3 h^s \begin{pmatrix} 2 \\ -3 \end{pmatrix} + h^{2s} \begin{pmatrix} 1000 \\ 1 \end{pmatrix}.$$

On peut vérifier que $Q(1,1,1)(2,-3)^T = (2,-3)^T$. La figure 7.4 représente Δy_1 en fonction de h pour les méthodes de Radau IA avec $s = 3$ et $s = 4$. Les ordres observés sont respectivement 5 et 7, pour respectivement la première et la seconde méthode, et ce en accord avec les résultats du théorème 22. Le lemme

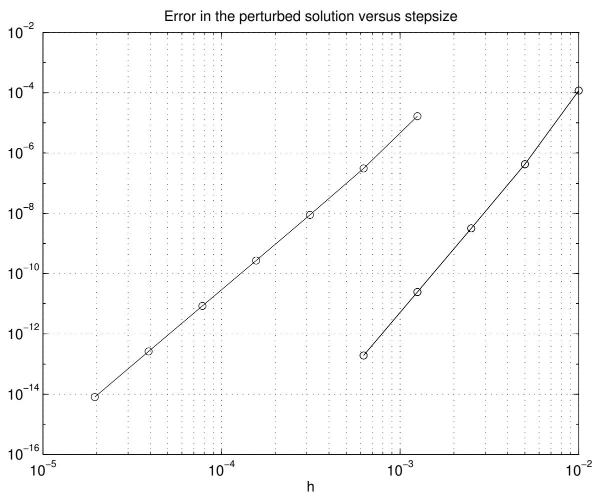


FIG. 7.4 – Δy_1 *en fonction de* h *pour les méthodes de Radau IA* $s = 3$ *et* $s = 4$.

suivant concerne l'accumulation des erreurs et consiste en une légère modification du lemme 4.5 de Hairer, Lubich & Roche [HLR89].

Lemme 2 (Aubry & Chartier [AC96]) *On suppose que la suite $\{\Delta y_n\}$ vérifie*

$$\Delta y_{n+1} = P_n \Delta y_n + \mathcal{O}(h\|P_n \Delta y_n\|) + \mathcal{O}(h^r \|Q_n \Delta y_n\|),$$

où P_n et Q_n sont des projections telles que $P_n + Q_n = I$ et $P_{n+1} = P_n + \mathcal{O}(h)$. Alors pour h suffisamment petit et $nh \leq Const$, on a

$$\|\Delta y_n\| \leq C\left(\|P_0 \Delta y_0\| + h^r \|Q_0 \Delta y_0\|\right). \tag{7.25}$$

Il vient finalement pour les méthodes de Radau IA :

Théorème 23 (Aubry & Chartier [AC96]) *Pour la méthode de Radau IA à s étapes internes appliquée à un système différentiel algébrique d'indice deux de la forme (5.4), l'erreur globale vérifie*

$$\begin{aligned} y_n - y(x_n) &= \mathcal{O}(h^s), \\ P(x_n)(y_n - y(x_n)) &= \mathcal{O}(h^{2s-2}), \textit{for } s \geq 2, \end{aligned}$$

où $P(x_n)$ est la projection définie en (6.9) et évaluée en $(y(x_n), z(x_n))$.

On considère à nouveau l'exemple (7.24) sur l'intervalle $[0,0.5]$. La figure 7.3.3 représente la solution numérique obtenue avec les méthodes de Radau IA à 3 et 4 étapes internes, avec et sans la projection après l'intégration finale. Le pas est choisi de manière aléatoire autour d'une valeur de référence h. Les pentes observées sont en accord avec le théorème 23.

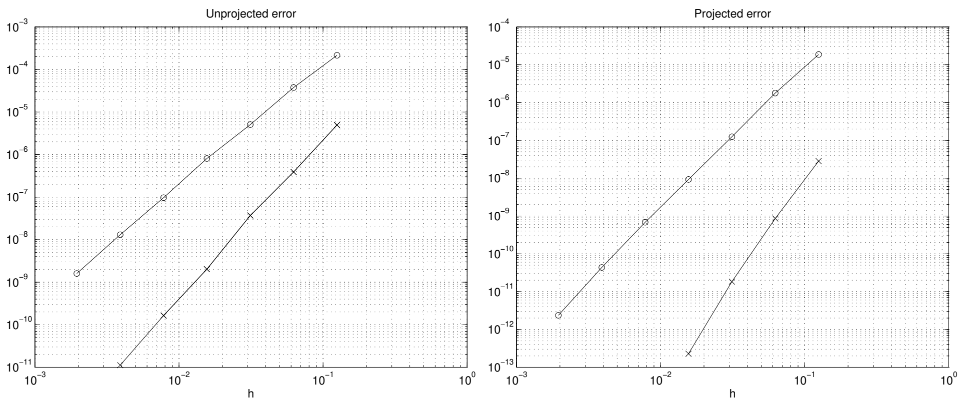


FIG. 7.5 – $y_n - y(x_n)$ *et* $P(x_n)\,(y_n - y(x_n))$ *pour des méthodes de Radau IA* $(s = 3 : o; s = 4 : \times)$

Chapitre 8

Une généralisation des méthodes de Runge-Kutta : les méthodes générales linéaires

Les *méthodes générales linéaires* sont apparues à l'origine comme un moyen d'unifier la théorie des méthodes à un pas et des méthodes multipas. Il est de fait possible, dans une large mesure, d'élaborer des théorèmes de convergence et de stabilité conformes à ceux existants pour les méthodes de Runge-Kutta et multipas. Cependant, les méthodes générales linéaires ont depuis été étudiées pour elles-mêmes, car elles constituent une classe strictement plus grande que la simple réunion des méthodes connues jusqu'alors. En particulier, elles ont un fort potentiel pour le parallélisme, au sens ou l'on peut construire des méthodes autorisant *intrinsèquement* certains calculs *concurrents* (typiquement plusieurs évaluations simultanées du second membre de (2.1)).

8.1 Formalisme

Pour le problème (2.1), une méthode générale linéaire se définit par :

– une *fonction valeur exacte* $z(x,h)$ prenant ses valeurs dans $\mathbb{R}^{r\times m}$ et fournissant l'*interprétation* de la méthode, -chacune des r composantes de $z(x,h)$ représente une fonction à valeur dans $\mathbb{R}^m$ liée à la solution exacte $y(x)$ de (2.1)-,

– une *procédure de démarrage* $\mathcal{S}$ fournissant les approximations initiales $y^{[0]}$ au point x_0 :

$$\begin{array}{rcl} \mathcal{S} : \mathbb{R}^m & \to & \mathbb{R}^{r\times m}, \\ y_0 & \mapsto & y^{[0]} = \mathcal{S}(y_0) = z(x_0,h_0) + \mathcal{O}(h_0^{p+1}), \end{array}$$

– un *système d'équations* permettant de calculer une nouvelle approximation $y^{[n+1]}$ de $z(x_{n+1},h_{n+1})$ à partir d'une approximation $y^{[n]}$ de $z(x_n,h_n)$. Dans leur version à pas constant ($h_n = h$), ces relations sont de la forme :

$$\begin{array}{rcl} Y_i^{[n+1]} & = & h\sum_{j=1}^s a_{ij} f(Y_j^{[n+1]}) \quad + \quad \sum_{j=1}^r u_{ij} y_j^{[n]}, \quad i = 1,\dots,s, \\ y_i^{[n+1]} & = & h\sum_{j=1}^s b_{ij} f(Y_j^{[n+1]}) \quad + \quad \sum_{j=1}^r v_{ij} y_j^{[n]}, \quad i = 1,\dots,r, \end{array}$$

ou de manière équivalente mais plus compacte :

$$\begin{array}{rcl} Y^{[n+1]} & = & h(A \otimes I_{\mathbb{R}^m}) F(Y^{[n+1]}) \quad + \quad (U \otimes I_{\mathbb{R}^m}) y^{[n]}, \\ y^{[n+1]} & = & h(B \otimes I_{\mathbb{R}^m}) F(Y^{[n+1]}) \quad + \quad (V \otimes I_{\mathbb{R}^m}) y^{[n]}, \end{array}$$

où $A = (a_{ij})$, $U = (u_{ij})$, $B = (b_{ij})$, $V = (v_{ij})$ et où

$$Y^{[n+1]} = \begin{bmatrix} Y_1^{[n+1]} \\ Y_2^{[n+1]} \\ \vdots \\ Y_s^{[n+1]} \end{bmatrix} \text{ et } F(Y^{[n+1]}) = \begin{bmatrix} f(Y_1^{[n+1]}) \\ f(Y_2^{[n+1]}) \\ \vdots \\ f(Y_s^{[n+1]}) \end{bmatrix}. \tag{8.1}$$

Les quatre matrices A, U, B et V caractérisent la méthode et sont pour cette raison généralement regroupées dans un tableau de la forme :

$$\mathcal{M} = \left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right]. \tag{8.2}$$

Exemple 8 *La fonction $z(x,h)$ possède une forme simple pour les méthodes connues telles les méthodes de Runge-Kutta ou les méthodes multipas. Pour une méthode de Runge-Kutta, on a en effet*

$$z(x,h) = y(x),$$

alors que pour une méthode à r pas, il vient

$$z(x,h) = \begin{bmatrix} y(x-(r-1)h) \\ y(x-(r-2)h) \\ \vdots \\ y(x) \end{bmatrix}.$$

8.2 DIMSIMs

Les méthodes DIMSIM ("Diagonally Implicit Multi-Stage Method") ont été introduites par Butcher en 1993 [But93]. L'idée clé des méthodes DIMSIM consiste à considérer des fonctions valeur exacte de la forme suivante : si p désigne l'ordre de la méthode, on suppose que $z(x,h)$ s'obtient comme combinaison linéaire des fonctions $y^{(k)}(x)$:

$$z(x,h) = \sum_{k=0}^{p} h^k \alpha_k \otimes y^{(k)}(x),\ \alpha_k \in I\!R^r, k = 0, \dots, p.$$

On suppose en outre que les approximations internes $Y_j^{[0]}$ sont telles que :

$$Y_j^{[0]} = y(x_0 + c_j h) + \mathcal{O}(h^{q+1}),$$

avec $q = p$ ou $q = p - 1$.

8.2.1 Différents groupes de méthodes DIMSIMS

Les méthodes DIMSIM ont été divisées en 4 classes suivant la structure de la matrice A (voir tableau 8.1), A étant néanmoins toujours de la forme :

$$\begin{bmatrix} \lambda & 0 & \dots & 0 \\ \times & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \times & \dots & \times & \lambda \end{bmatrix}.$$

Il est aisé de constater que les méthodes des classes 3 et 4 sont des méthodes

Type	λ	$a_{i,j}, j < i$	Parallélisme	Classe de problèmes
1	0	Arbitraire	Non	Non-raide
2	$\neq 0$	Arbitraire	Non	Raide
3	0	0	Oui	Non-raide
4	$\neq 0$	0	Oui	Raide

TAB. 8.1 – *Les* 4 *classes de méthodes DIMSIMs.*

susceptibles d'être parallélisées.

8.2.2 Stabilité linéaire

Le domaine de stabilité d'une méthode numérique est une caractéristique fondamentale. Pour le problème test

$$\begin{cases} y'(x) &= \mu y(x), \\ y(x_0) &= y_0, \end{cases}$$

une méthode DIMSIM devient

$$\begin{aligned} y^{[n+1]} &= (h\mu)BY^{[n+1]} + Vy^{[n]}, \\ Y^{[n+1]} &= (h\mu)AY^{[n+1]} + Uy^{[n]}, \end{aligned}$$

i.e.

$$y^{[n+1]} = M(h\mu)y^{[n]},$$

où $M(z)$ est la *matrice de stabilité*

$$M(z) := V + zB(I_s - zA)^{-1}U.$$

De même que pour une méthode de Runge-Kutta ou une méthode multipas, une méthode DIMSIM est dite A-stable si et seulement si son *domaine de stabilité*

$$S := \{z \in \mathbb{C}, \exists\, C \in \mathbb{R}_+^*, \forall\, n \in \mathbb{N}, \|M(z)\|^n \leq C\}$$

contient $\mathbb{C}^-$.

Théorème 24 (Chartier [Cha94]) *Soit $\pi(t,z) := \det(M(z) - tI_{\mathbb{R}^r})$ le polynôme caractéristique de $M(z)$ et soient ϵ_j les coefficients de ω^j, $j = 0, \ldots, 2r$, dans*

$$W(\omega,y) = (\omega - 1)^{2r}\pi\left(\frac{\omega+1}{\omega-1}, iy\right)\pi\left(\frac{\omega+1}{\omega-1}, -iy\right).$$

Soit en outre $\mathcal{H}$ la matrice $2r \times 2r$

$$\mathcal{H} = \begin{bmatrix} \epsilon_{2r-1} & \epsilon_{2r-3} & \epsilon_{2r-5} & \cdots & \epsilon_{1-2r} \\ \epsilon_{2r} & \epsilon_{2r-2} & \epsilon_{2r-4} & \cdots & \epsilon_{2-2r} \\ 0 & \epsilon_{2r-1} & \epsilon_{2r-3} & \cdots & \epsilon_{3-2r} \\ 0 & \epsilon_{2r} & \epsilon_{2r-2} & \cdots & \epsilon_{4-2r} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \epsilon_{2r} & \cdots & \epsilon_0 \end{bmatrix},$$

où $\epsilon_j = 0$ pour $j < 0$, et où $\Delta(y)$ désigne son mineur principal d'ordre $(2r - 1)$. Alors si l'ensemble des conditions suivantes sont satisfaites la méthode est A-stable :

$$\left\{ \begin{array}{l} \exists y^* \in I\!R_+^*, \forall \omega \in \mathbb{C}, W(\omega, y^*) = 0 \Rightarrow Re(\omega) < 0, \\ \forall y \in I\!R_+^*, \epsilon_0(y) \neq 0, \\ \forall y \in I\!R_+^*, \epsilon_{2r}(y) \neq 0, \\ \forall y \in I\!R_+^*, \Delta(y) \neq 0. \end{array} \right. \tag{8.3}$$

8.3 Construction de méthodes de type 4

Théorème 25 (Butcher & Chartier [BC94, BC95]) *On considère une méthode DIMSIM à r pas et r étapes internes de la forme :*

$$\left[\begin{array}{c|c} A & I_{I\!R^r} \\ \hline B & V \end{array} \right],$$

avec $A = diag(\lambda_1, \ldots, \lambda_r)$ et $M(\infty) = V - BA^{-1} = 0$. Alors elle est d'ordre r si et seulement si

$$\begin{array}{rcl} \lambda_i & = & \dfrac{l_i(1+c_i)}{l_i(1+c_i) + l_i'(1+c_i)}, i = 1, \ldots, r, \\ v_{i,j} & = & \dfrac{l_j^2(1+c_i)}{(l_i(1+c_i) + l_i'(1+c_i))l_j(c_j)}, i = 1, \ldots, r, j = 1, \ldots, r, \end{array}$$

où les l_i sont définis par

$$l_i(x) = \frac{\prod_{j=1,\, j\neq i}^{r}(x - c_j)}{\prod_{j=1,\, j\neq i}^{r}(c_i - c_j)}.$$

Si $c_i = -r+1+i, i = 1, \ldots, r$, alors la méthode obtenue n'est autre que la méthode de différentiation rétrograde à r pas (voir Hairer, Nørsett & Wanner [HNW93]). Cependant, cette méthode n'est pas A-stable et cela semble être plus largement une caractéristique commune aux méthodes de cette classe (voir Butcher & Chartier [BC94]). Afin d'obtenir des méthodes A-stables, on relaxe donc certaines hypothèses :

Théorème 26 (Butcher & Chartier [BC94, BC95]) *On considère une méthode DIMSIM à r pas et r étapes internes de la forme :*

$$\left[\begin{array}{c|c} A & I_{I\!R^r} \\ \hline B & V \end{array} \right],$$

où $A = diag(\lambda_1, \ldots, \lambda_r)$ et $M(\infty) = V - BA^{-1} = 0$. Alors elle est d'ordre $r-1$ si et seulement si

$$V = L - AL',$$

où les éléments (i,j) de L et L' sont respectivement

$$l_j(1+c_i) \text{ et } l'_j(1+c_i).$$

Des méthodes A-stables d'ordres 1 à 12 ont ainsi été construites (voir Chartier [Cha94]).

8.4 Expériences numériques

On présente maintenant la méthode développée dans Chartier [Cha98b] ainsi que quelques résultats numériques. Elle a été obtenue par le théorème 26 avec $r = 6$ et

$$\lambda_i = \mu + \nu c_i.$$

Pour

$$z(x,h) = \begin{bmatrix} y(x) \\ hy'(x) \\ \vdots \\ \frac{h^5}{5!} y^{(5)} \end{bmatrix},$$

son tableau $\mathcal{M}$ est de la forme :

$$\left[\begin{array}{cccccc|cccccc}
2/5 & 0 & 0 & 0 & 0 & 0 & 1 & -2/5 & 0 & 0 & 0 & 0 \\
0 & \frac{7}{15} & 0 & 0 & 0 & 0 & 1 & -\frac{4}{15} & -\frac{11}{75} & -\frac{6}{125} & -\frac{1}{75} & -\frac{32}{9375} \\
0 & 0 & \frac{8}{15} & 0 & 0 & 0 & 1 & -2/15 & -\frac{4}{15} & -\frac{24}{125} & -\frac{208}{1875} & -\frac{544}{9375} \\
0 & 0 & 0 & 3/5 & 0 & 0 & 1 & 0 & -\frac{9}{25} & -\frac{54}{125} & -\frac{243}{625} & -\frac{972}{3125} \\
0 & 0 & 0 & 0 & 2/3 & 0 & 1 & 2/15 & -\frac{32}{75} & -\frac{96}{125} & -\frac{1792}{1875} & -\frac{9728}{9375} \\
0 & 0 & 0 & 0 & 0 & \frac{11}{15} & 1 & \frac{4}{15} & -\frac{7}{15} & -6/5 & -\frac{29}{15} & -8/3 \\
\hline
0 & 0 & 0 & 0 & 0 & \frac{11}{15} & 1 & \frac{4}{15} & -\frac{7}{15} & -6/5 & -\frac{29}{15} & -8/3 \\
-2/5 & \frac{35}{12} & -\frac{80}{9} & 15 & -\frac{50}{3} & \frac{1507}{180} & 0 & 2/3 & -2/15 & -\frac{12}{5} & -\frac{92}{15} & -\frac{34}{3} \\
-\frac{25}{6} & \frac{2135}{72} & -\frac{260}{3} & \frac{535}{4} & -\frac{1925}{18} & \frac{275}{8} & 0 & 0 & 1/3 & -6/5 & -\frac{34}{5} & -\frac{56}{3} \\
-\frac{175}{12} & \frac{7175}{72} & -\frac{2450}{9} & \frac{1475}{4} & -\frac{8875}{36} & \frac{4675}{72} & 0 & 0 & 0 & 0 & -\frac{44}{15} & -\frac{44}{3} \\
-\frac{125}{6} & \frac{9625}{72} & -\frac{1000}{3} & \frac{1625}{4} & -\frac{4375}{18} & \frac{1375}{24} & 0 & 0 & 0 & 0 & -1/3 & -16/3 \\
-\frac{125}{12} & \frac{4375}{72} & -\frac{1250}{9} & \frac{625}{4} & -\frac{3125}{36} & \frac{1375}{72} & 0 & 0 & 0 & 0 & 0 & -2/3
\end{array}\right],$$

avec $c = [0,\frac{1}{5},\frac{2}{5},\frac{3}{5},\frac{4}{5},1]^T$. En utilisant les résultats de Butcher, Chartier & Jackiewicz [BCJ97], on peut estimer l'erreur locale

$$err = h_n^6 (v^T \varphi_5) y^{(6)}(x_{n-1}),$$

où v^T est le vecteur propre à gauche de V associé à la valeur propre 1 et tel que $v_1 = 1$ et où

$$\varphi_p = [\frac{1}{6!\,0!}, \frac{1}{5!\,1!}, \ldots, \frac{1}{1!\,5!}]^T - \frac{1}{5!} B c^5.$$

Cette méthode a été codée en incluant toutes les techniques habituelles de contrôle de l'erreur et de changement de pas (son comportement est décrit par quelques indicateurs statistiques en figure 8.4). Les tests présentés ci-dessous (figure 8.2) démontrent son efficacité (comparée au code VODE qui comme RADAU5 fait référence). Le problème choisi comme test est celui du Brusselator (voir Hairer & Wanner [HW96], pp. 5-7) :

$$\begin{array}{rcl} \frac{\partial u}{\partial t} & = & A + u^2 v - (B+1)u + \beta \frac{\partial^2 u}{\partial x^2} \\ \frac{\partial v}{\partial t} & = & Bu - u^2 v + \beta \frac{\partial^2 v}{\partial x^2} \end{array} \tag{8.4}$$

avec $0 \leq x \leq 1$, $A = 1$, $B = 3$, $\beta = 1/50$ et les conditions aux bords

$$\begin{array}{ll} u(0,t) = u(1,t) = 1, & v(0,t) = v(1,t) = 3, \\ u(x,0) = 1 + sin(2\pi x), & v(x,0) = 3. \end{array} \tag{8.5}$$

Le système est discrétisé par différence finie sur une grille de N points donnant lieu à un système d'EDO de dimension $2N$. On prend ici $N = 20$.

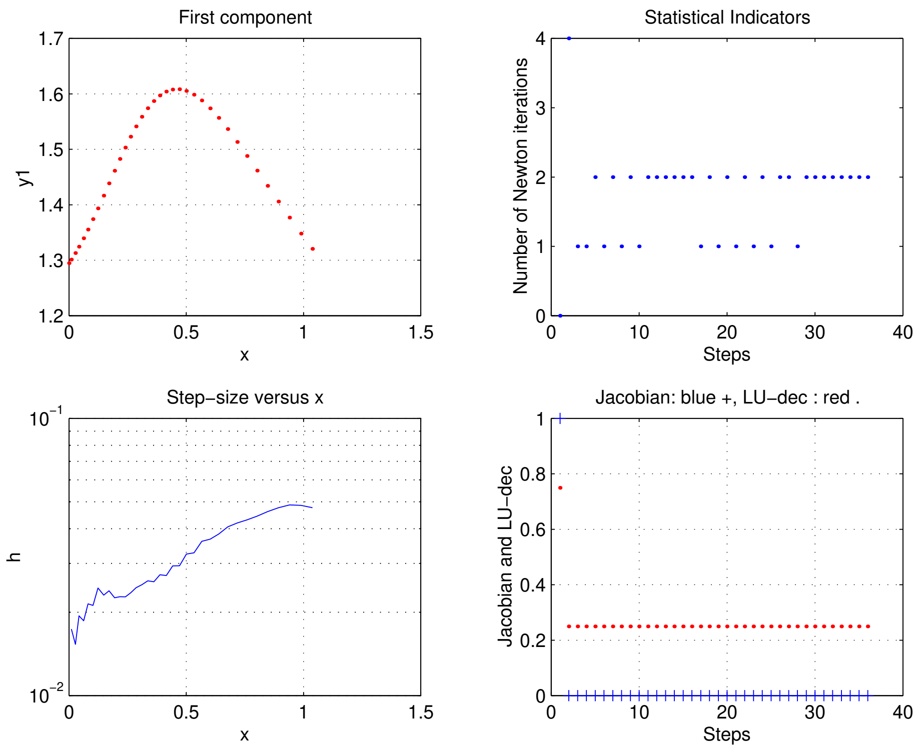


FIG. 8.1 – *Quelques indicateurs statistiques*

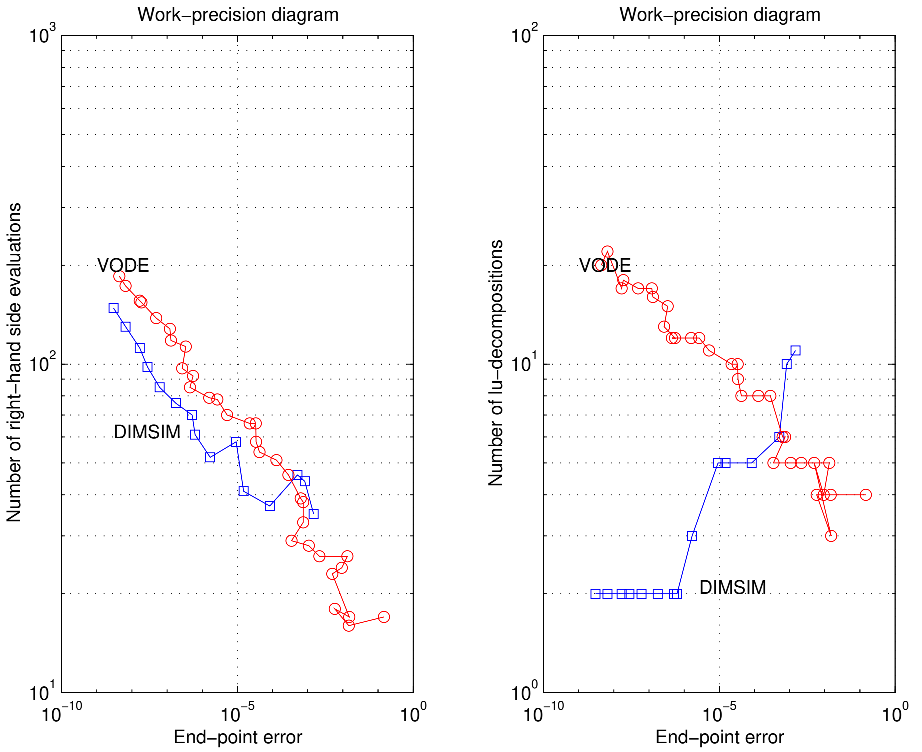


FIG. 8.2 – *Courbe coût/précision pour les codes DIMSIM et VODE*

Chapitre 9

Synthèse

9.1 Une technique de “shooting”

La première partie de [Cha93b] est consacrée à l'étude d'une méthode dérivée du processus de Picard, donc itérative. L'intervalle d'intégration est partitionné en sous-intervalles, sur lesquels une approximation de la solution peut être calculée indépendamment des autres sous-intervalles. Un processus de correction portant sur les valeurs aux bornes est alors appliqué. Il a été montré en particulier que ce processus est convergent pour les systèmes dissipatifs. Implémentée sur hypercube, cette méthode offre un parallélisme assez massif et permet des accélérations importantes, mais se révèle peu efficace pour les systèmes non-dissipatifs. Ces limitations sont dues à la non-convergence de la méthode de Newton en dehors d'un voisinage de la solution exacte [CP93].

9.2 DIMSIM : Diagonally Implicit Multi-Stage Integration Methods

S'apparentant aux méthodes générales linéaires, elles en partagent le formalisme assez lourd et la difficulté d'analyse. Néanmoins, elles réunissent certains des avantages des méthodes de Runge-Kutta (stabilité pour les problèmes raides) et des méthodes multipas (coût comparable). Ainsi, dans [Cha94] est construite une classe de schémas L-stables d'ordres 1 à 12. Cette classe de méthodes constitue la sous-classe des méthodes DIMSIM implicites et parallèles mais a été introduite indépendamment des méthodes DIMSIM (dues à J.C. Butcher). Le parallélisme est ici utilisé non pas pour accroître directement la vitesse d'exécution, mais pour améliorer la stabilité et ainsi lever certaines restrictions portant sur la nature des problèmes traités. En d'autres termes, ces schémas ouvrent de nouveaux champs d'applications. Des formules explicites de calcul de leurs coefficients sont détaillées dans [BC95], alors que leur application aux équations différentielles algébriques est envisagée dans [BC94] et [Cha93a]. Les problèmes liés à l'implémentation ont été pour l'essentiel résolus, grâce à une réécriture adéquate des formules d'intégration, permettant une estimation rigoureuse (asymptotiquement correcte) des erreurs locales et autorisant des changements de pas d'un coût négligeable. L'étude de la stabilité à pas variable s'en trouve en outre simplifiée [BCJ97].

9.3 Conditions d'ordre et composition de méthodes

Les conditions d'ordre (conditions algébriques portant sur les coefficients d'une méthode de Runge-Kutta pour qu'elle soit d'un ordre donné) peuvent être obtenues à l'aide du formalisme des arbres de Butcher, comme dans le cas des équations différentielles ordinaires. Dans [CC96], la loi de composition des arbres, correspondant à la composition de deux méthodes de Runge-Kutta, est généralisée au cas des équations différentielles algébriques. Cette extension permet de démontrer certains résultats théoriques dans un cadre plus général.

9.4 Construction d'une méthode de Radau IIA modifiée

Il est désormais bien établi et analysé qu'une méthode de Runge-Kutta converge avec un ordre réduit lorsqu'on l'applique à un système algébro-différentiel. Ainsi, la méthode de Radau IIA d'ordre 5, squelette du code RADAU5 développé par E. Hairer et G. Wanner de l'université de Genève et considéré comme l'un des meilleurs codes du domaine public, ne converge qu'avec l'ordre 3 pour ce qui est de la composante algébrique. Le travail effectué ici a consisté en la définition d'un nouveau schéma d'ordre 5 aussi bien pour la composante différentielle que pour la composante algébrique. Ce nouveau schéma a été incorporé dans le code RADAU5 et apporte une amélioration notable de la précision sans augmentation des temps de calcul [AC98b].

9.5 Analyse de la propagation des erreurs

Une technique classique permettant de remédier au phénomène de réduction d'ordre consiste à projeter la solution numérique sur les contraintes algébriques et ce après chaque pas d'intégration. L'analyse de la propagation des erreurs montre néanmoins que l'accumulation des erreurs locales s'effectue sur un sous-espace particulier de l'espace dans lequel vivent les solutions et qu'il est possible d'accroître l'ordre d'approximation de la solution numérique en projetant uniquement à la fin de l'intervalle. Ce résultat n'a pour le moment été prouvé en toute rigueur que dans le cas des méthodes Radau IA [AC96]. Le cas des méthodes de Gauss est en cours d'examen.

9.6 Méthodes pseudo-symplectiques

Il s'agit de construire des méthodes de Runge-Kutta *explicites* d'ordre p, dites *pseudo-symplectiques* d'ordre $2p$, c'est à dire "approchant" une méthode symplectique avec un ordre $2p$. L'intérêt de cette définition réside dans le fait qu'une méthode symplectique a une erreur globale croissant de manière linéaire par rapport à la longueur de l'intervalle d'intégration, lorsque la solution exacte est périodique ou lorsque le système est intégrable. Ainsi, une méthode pseudo-symplectique est-elle susceptible de reproduire ce comportement, ce qui en fait un schéma performant pour la résolution de systèmes hamiltoniens sur de longs intervalles de temps. Des méthodes d'ordres $2-4$ et $3-6$ ont ainsi été construites et testées avec succès. La construction de méthodes d'ordre plus élevé reste cependant délicate étant donnée la complexité des équations d'ordre [AC98c, AC98b].

9.6.1 Méthodes SIRK (Singly-Implicit Runge-Kutta)

La propriété fondamentale des méthodes SIRK est la suivante : la matrice de coefficients A de ces méthodes possède une seule valeur propre λ de multiplicité s. Ainsi, si J est la jacobienne $m \times m$ du système différentiel, la décomposition LU (lower-upper) de la matrice $(I_s \otimes I_m - hA \otimes J)$, dont le coût est prédominant dans les formules de passage d'un pas au suivant, peut être évitée et remplacée par la décomposition LU de la matrice $I_m - h\lambda J$. Modulo quelques transformations linéaires, le coût de ces méthodes est alors ramené à un niveau comparable à celui des méthodes multipas. Les méthodes SIRK *classiques* souffrent néanmoins d'un handicap important qui les a jusqu'alors confinées à des applications particulières. Il est lié à la localisation des approximations intermédiaires. L'extension proposée permet de lever cette contrainte tout en conservant la propriété fondamentale des méthodes classiques, et ce en ayant recours à la notion d'"ordre effectif" [BC97, BC99].

9.7 Relaxation non-linéaire

La méthode de Newton modifiée reste au coeur de l'intégration numérique aussi bien pour les méthodes SIRK que pour les autres. La possibilité de découpler le système non-linéaire à résoudre en sous-systèmes est conditionnée à l'application de certaines transformations linéaires, de coût négligeable pour les systèmes de grande taille, mais important pour ceux de faible dimension. Même dans les cas favorables, la résolution de ces différents sous-systèmes reste séquentielle. L'introduction d'un niveau supplémentaire d'itérations (relaxation non-linéaire), permet au contraire un découplage complet et l'utilisation du parallélisme. Le choix de la matrice d'itération sous-jacente est crucial et a fait l'objet d'une étude de convergence dans le cas des méthodes de Runge-Kutta "classiques" [Cha98a].

9.8 Méthodes de Runge-Kutta d'ordre élevé

Il existe différentes classes de méthodes de Runge-Kutta d'ordre élevé. Certaines possèdent une ou plusieurs des propriétés de stabilité algébrique, de symplecticité ou de symétrie. Les relations entre ces différentes classes et leurs propriétés ont fait l'objet d'une étude systématique dans [CC00].

Bibliographie

[AC96] A. Aubry and P. Chartier. On the structure of errors for Radau IA methods applied to index-2 DAEs. *Journal of Applied Numerical Mathematics*, 22(1-3):23–34, Novembre 1996. Special Issue: Celebrating The Centenary of Runge-Kutta Methods.

[AC98a] A. Aubry and P. Chartier. A note on pseudo-symplectic Runge-Kutta methods. *BIT*, 38(4):802–806, 1998.

[AC98b] A. Aubry and P. Chartier. On Improving the Convergence of Radau IIA methods when Applied to Index-2 DAEs. *SIAM Journal on Numerical Analysis*, 35(4):1347–1367, 1998.

[AC98c] A. Aubry and P. Chartier. Pseudo-symplectic Runge-Kutta methods. *BIT*, 38(3):439–461, 1998.

[AP91] U.M. Ascher and L.R. Petzold. Projected implicit Runge-Kutta methods for differential-algebraic equations. *SIAM Journal of Numerical Analysis*, 4:1097–1120, 1991.

[Arn67] V.I. Arnold. *Méthodes mathématiques pour la mécanique classique*. Éditions MIR, 1967.

[Arn74] Arnold. *Equations Différentielles Ordinaires*. Editions de Moscou, 1974.

[BB79] K. Burrage and J.C. Butcher. Stability criteria for implicit Runge-Kutta methods. *SIAM Journal on Numerical Analysis*, 16:46–57, 1979.

[BC93] K. Burrage and R.P.K. Chan. On smoothing and order reduction effects for implicit Runge-Kutta formulae. *Journal of Computational and Applied Mathematics*, 45:17–27, 1993. North-Holland, CAM 1275.

[BC94] J. C. Butcher and P. Chartier. The construction of DIMSIMs for stiff ODEs and DAEs. Technical Report 308, University of Auckland, Auckland, New Zealand, July 1994.

[BC95] J.C. Butcher and P. Chartier. Parallel general linear methods for stiff ordinary differential and differential algebraic equations. *Applied Numerical Mathematics*, 17:213–222, 1995.

[BC97] J. C. Butcher and P. Chartier. A generalization of singly-implicit Runge-Kutta methods. *Applied Numerical Mathematics*, 24:343–350, 1997.

[BC99] J. C. Butcher and P. Chartier. The effective order of Singly-Implicit Runge-Kutta methods. *Numerical Algorithms*, 20(4):269–284, 1999.

[BCJ97] J. C. Butcher, P. Chartier, and Z. Jackiewicz. Nordsieck representation of DIMSIMs. *Numerical Algorithms*, 16(2):209–230, 1997.

[BCP89] K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical solution of initial value problems in differential-algebraic equations*. North Holland, 1989. New-York.

[But87] J.C. Butcher. *The Numerical Analysis of Ordinary Differential Equations*. John Wiley and Sons, 1987.

[But93] J.C. Butcher. Diagonally-Implicit Multi-Stage Integration Methods. *Applied Numerical Mathematics*, 11(380):347–363, 1993. North-Holland.

[Cal92] M.P. Calvo. *Canonical Runge-Kutta-Nyström Methods*. PhD thesis, Universidad de valladolid, 1992.

[CC00] R.P.K. Chan and P. Chartier. Classification of High-Order Implicit Runge-Kutta Methods. *New-Zealand Journal of Mathematics*, 2000. To appear.

[CC96] R.P.K Chan and P. Chartier. A Composition Law for Runge-Kutta Methods Applied to Index-2 Differential-Algebraic Equations. *BIT*, 36(2):229–246, 96.

[CG95] S.L. Campbell and C.W. Gear. The index of general nonlinear DAEs. *Numerisch Mathematik*, 72:173–196, 1995.

[CH95] M.P. Calvo and E. Hairer. Accurate long-term integration of dynamical systems. *Journal of Applied Numerical Mathematics*, 18:95–105, 1995.

[Cha89] R.P.K. Chan. *Extrapolation of Runge-Kutta Methods for Stiff Initial-Value Problems*. PhD thesis, University of Auckland, 1989.

[Cha90] R.P.K. Chan. On Symmetric Runge-Kutta Methods of High Order. *Computing*, 45:301–309, 1990.

[Cha93a] P. Chartier. General Linear Methods for differential Algebraic Equations. Technical Report 711, IRISA, mar 1993.

[Cha93b] P. Chartier. *Parallélisme dans la résolution numérique des problèmes de valeur initiale pour les équations différentielles ordinaires et algébriques*. PhD thesis, Université de Rennes 1, Rennes, France, June 1993.

[Cha94] P. Chartier. L-stable parallel one-block methods for ordinary differential equations. *SIAM Journal of Numerical Analysis*, 31(2), April 1994.

[Cha98a] P. Chartier. On diagonally-iterated Runge-Kutta methods. *Journal of Comp. and Appl. Math.*, 81(1), 1998.

[Cha98b] P. Chartier. The potential of parallel multi-value methods for the simulation of large real-life problems. *CWI Quaterly*, 11(1):7–32, 1998.

[CL98] P. Chartier and E. Lapôtre. Reversible B-series. Technical Report 1221, IRISA, 1998.

[CM] R. Chan and A. Murua. Extrapolation of symplectic methods for Hamiltonian problems. In preparation.

[CM89] M. Crouzeix and A.L. Mignot. *Analyse numérique des équations différentielles*. Masson, 1989. Collection maîtrise de mathématiques pures.

[Coo87] G.J. Cooper. Stability of Runge-Kutta methods for trajectory problems. *IMA Journal of Numerical Analysis*, 7:1–13, 1987.

[CP93] P. Chartier and B. Philippe. A Parallel Shooting Technique for Solving Dissipative ODE's. *Computing*, 51(3-4):209–213, 1993.

[CR80] M. Crouzeix and P.A. Raviart. Approximation des problèmes d'évolution. Lecture Notes, 1980.

[CR93] P.E. Crouch and Grossman R. Numerical integration of ordinary differential equations on manifolds. *J. Nonlinear Sci.*, 3:1–33, 1993.

[Cro79] M. Crouzeix. Sur la B-stabilité des méthodes de Runge-Kutta. *Numerische Mathematik*, 32:75–82, 1979.

[CSS92] M.P. Calvo and J.M. Sanz-Serna. Order conditions for canonical Runge-Kutta-Nyström methods. *BIT*, (32):131–142, 1992.

[CSS94] M.P. Calvo and J.M. Sanz-Serna. Canonical B-series. *Numerische Mathematik*, 67:161–175, 1994. Springer-Verlag.

[Gea90] C.W. Gear. Differential-Algebraic equations, indices and integral algebraic equations. *SIAM Journal of Numerical Analysis*, 27:1527–1534, 1990.

[GGL85] C.W. Gear, G.K. Gupta, and B. Leimkuhler. Automatic integration of Euler-lagrange equations with constraints. *Journal of Computing and Applied Mathematics*, 12, 13:77–90, 1985.

[GP84] C.W. Gear and L.R. Petzold. ODE methods for the solution of differential/algebraic systems. *SIAM Journal of Numerical Analysis*, 21:716–728, 1984.

[Hai94] E. Hairer. Backward analysis of numerical integrators and symplectic methods. *Annals of Numerical Mathematics*, 1:107–132, 1994.

[Hai99] E. Hairer. Numerical geometric integration. Lecture Notes, http://www.unige.ch/math/folks/hairer/polycop.html, 1998-1999.

[Har64] P. Hartman. *Ordinary Differential Equations*. John Wiley and sons, London, 1964.

[HL97] E. Hairer and C. Lubich. The life-span of backward error analysis for numerical integrators. *Numerische Mathematik*, 76:441–462, 1997.

[HLR89] E. Hairer, C. Lubich, and M. Roche. *The Numerical Solution of Differential Algebraic Systems by Runge-Kutta Methods*. Springer-Verlag, 1989. Lecture Notes in Mathematics 1409.

[HNW87] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations, Nonstiff Problems*. Springer-Verlag, first edition, 1987. Volume 1.

[HNW93] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations, Nonstiff Problems*. Springer-Verlag, second revised edition, 1993. Volume 1.

[HW96] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations, Stiff Problems and Differential Algebraic Problems*. Springer-Verlag, second revised edition, 1996. Volume 2.

[IN97] A. Iserles and S.P. Nørsett. On the solution of linear differential equations in lie groups. Technical Report NA3, DAMTP, University of Cambridge, 1997.

[Ise84] A. Iserles. Solving linear ordinary differential equations by exponentials of iterated commutators. *Numerische Mathematik*, 45:183–199, 1984.

[Jay93] L. Jay. Convergence of a class of Runge-Kutta methods for differential-algebraic systems of index 2. *BIT*, 33:137–170, 1993.

[Jay94] L.O. Jay. *Runge-Kutta methods for differential-algebraic equations of index 3 with applications to Hamiltonian systems*. PhD thesis, Université de Genève, 1994.

[KZj95] Feng Kang and Shang Zai-jiu. Volume-preserving algorithms for source-free dynamical systems. *Numerische Mathematik*, 71:451–463, 1995.

[Las88] F.M. Lasagni. Canonical Runge-Kutta methods. *ZAMP*, 39:952–953, 1988.

[MK] H. Munthe-Kass. High order Runge-Kutta methods on manifolds. *J. Appl. Numer. Maths.* To appear.

[Moa96] C. Moan. The numerical solution of ordinary differential equations with conservation laws. Master's thesis, Department of Mathematical Sciences, Norwegian University of Science and Technology, March 1996.

[Mär92] R. März. Numerical methods for differential algebraic equations. *Acta Numerica*, 1:141–198, 1992.

[OS94] D. Okunbor and R.D. Skeel. Canonical Runge-Kutta-Nyström methods of order 5 and 6. *Journal of Computational and Applied Mathematics*, 51(3):375–382, 1994.

[QD98] G.R.W. Quispel and C.P. Dyt. Volume-preserving integrators have linear error growth. *Phys. Lett.*, 242A:25–30, 1998.

[Qui95] G.R.W. Quispel. Volume-preserving integrators. *Phys. Lett.*, 206A:26–30, 1995.

[Rob66] H. H. Robertson. The solution of a set of reaction rate equations. In *Numerical Analysis: An introduction*, pages 178–182. Academic Press, London, 1966.

[Sco91] C. Scovel. Symplectic numerical integration of hamiltonian systems. In T. Ratiu, editor, *The Geometry of Hamiltonian Systems*, pages 463–496. Springer Verlag, New York, 1991.

[SS88] J.M. Sanz-Serna. Runge-Kutta schemes for Hamiltonian systems. *BIT*, 28:877–883, 1988.

[SSC94] J.M. Sanz-Serna and M.P. Calvo. *Numerical Hamiltonian Problems*. Number 7 in Applied Mathematics and Mathematical Computation. Chapman and Hall, 1994.

[Ste73] H. J. Stetter. *Analysis of discretization methods for ordinary differential equations*. Springer, 1973.

[Sto88] D. Stoffer. On reversible and canonical integration methods. Technical Report 88-05, ETH-Zürich, 1988.

[Sur89] Y.B. Suris. The canonicity of mappings generated by Runge-Kutta type methods when integrating the systems $\ddot{x} = -\partial u/\partial x$. *Comput. Maths. Phys.*, 29:138–144, 1989.

[Zan97] A. Zanna. Collocation and relaxed collocation for the fer and the magnus expansions. Technical Report NA17, DAMTP, University of Cambridge, 1997.

Table des figures

Liste des tableaux