Available online at www.sciencedirect.com
ScienceDirect
C. R. Acad. Sci. Paris, Ser. I 346 (2008) 229–232
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Analyse numérique

Accélération de schémas d'intégration temporelle pour les équations différentielles algébriques linéaires

Mouhamad Al Sayed Ali

Université de Brest, département de mathématiques, 6, avenue Le-Gorgeu, CS 93837, 29238 Brest cedex 3, France

Reçu le 7 février 2007 ; accepté après révision le 11 décembre 2007

Disponible sur Internet le 14 janvier 2008

Présenté par Olivier Pironneau

Résumé

On propose une approche permettant d'accélérer les schémas d'intégration temporelle implicites pour la résolution des équations différentielles algébriques linéaires à coefficients constants. ***Pour citer cet article : M. Al Sayed Ali, C. R. Acad. Sci. Paris, Ser. I 346 (2008).***

Abstract

Acceleration of implicit schemes for linear differential-algebraic equations. We propose an approach to accelerate implicit schemes associated to linear differential-algebraic equations with constant coefficients. ***To cite this article: M. Al Sayed Ali, C. R. Acad. Sci. Paris, Ser. I 346 (2008).***

1. Introduction

De nombreuses discrétisations de problèmes physiques aboutissent à des équations différentielles algébriques linéaires à coefficients constants évoluant dans un espace vectoriel de grande dimension. Les schémas d'intégration temporelle (Crank–Nicolson, Runge–Kutta implicite, etc.), permettant d'en calculer une solution approchée, nécessitent la résolution d'un nombre si important de systèmes linéaires de grande taille que leurs coûts de calcul représentent généralement une grande partie du coût global du schéma.

Sachant que généralement ces systèmes linéaires sont résolus via une méthode itérative, se pose le problème du choix de la solution initiale, autrement dit d'une approximation de la solution. Intuitivement il est clair que plus la solution initiale est proche de la solution exacte, plus vite la méthode itérative devrait converger.

Dans cette Note, nous proposons et justifions l'utilisation d'une approche permettant de calculer une « bonne » solution initiale. L'idée est de combiner deux outils classiques de l'analyse numérique que sont les schémas multipas [3] et les techniques d'approximation de Petrov–Galerkin [5].

Adresse e-mail : alsayed@univ-brest.fr.

1631-073X/$ – see front matter
doi:10.1016/j.crma.2007.12.011

Le principe de cette approche est d'obtenir une solution approchée à un instant donné à partir des r éléments calculés lors des instants précédents comme le font les schémas multi-pas. Cependant contrairement à ces derniers, cette solution approchée n'est pas calculée en se basant sur un développement de type Taylor mais via une approximation de Petrov–Galerkin qui, dans ce cas, se trouve être optimale.

Afin de décrire plus précisément cette approche, supposons que l'équation différentielle qu'on souhaite résoudre s'écrit :

$$B\dot{y}(t) = Ay(t) + f(t) \quad \forall t \in [t_0, T], \qquad y(t_0) = y^{(0)}, \tag{1}$$

où $y^{(0)} \in \mathbb{R}^n$, A et B sont deux matrices carrées réelles d'ordre n et $f : [t_0, T] \to \mathbb{R}^n$ est de classe $C^{r+\nu-1}$, où $r \ll n$ et ν est l'indice du faisceau $\lambda B - A$ supposé régulier, c'est-à-dire la taille du plus grand bloc nilpotent dans la forme canonique de Weierstrass de $\lambda B - A$ (voir [4, p. 17]). Nous supposons également que $(y^{(0)}, f(t_0))$ vérifie la condition de consistance (voir [4, p. 17]) permettant d'assurer l'existence d'une unique solution y de (1) de classe C^r.

Les schémas d'intégration temporelle usuels permettant de résoudre (1) sont basés sur une discrétisation temporelle, que nous supposerons uniforme pour simplifier, définie par

$$t_i = t_0 + ih \quad \text{avec } h = \frac{T - t_0}{N},\ 0 \leqslant i \leqslant N \text{ et } N \in \mathbb{N}.$$

Ces schémas calculent les vecteurs $y_0, y_1, \ldots, y_N$ tels que y_i soit une approximation de $y(t_i)$ pour $i = 0, \ldots, N$. Dans un souci de clarté, nous supposons que la suite $(y_i)_{q+1 \leqslant i \leqslant N}$ est définie de la manière suivante

$$y_0 = y^{(0)}, \qquad y_{i+1} = y_i + hz_i \quad \text{pour } i = q, q+1, \ldots, N-1 \tag{2}$$

où, pour $i = 1, \ldots, q$, y_i est calculé par un schéma à i pas stable et de même ordre que (2), et z_i est la solution du système linéaire

$$Cz_i = b_i \tag{3}$$

où C est une matrice inversible de la forme $C = B - h\gamma A$, $\gamma \in \mathbb{R}$, et b_i est un vecteur de la forme

$$b_i = \sum_{l=0}^{q} (\xi_l A + \psi_l B) y_{i-l} + \sum_{l=-1}^{q} \phi_l f(t_{i-l}),$$

où ξ_l, ψ_l, ϕ_l sont des réels.

Parmi les schémas vérifiant ce formalisme, on trouve de nombreux schémas tels que ceux de Euler implicite, Crank–Nicolson ou Adams–Moulton (voir [3]).

L'étape cruciale, en ce qui concerne cette note, se situe lors de la résolution du système (3). En pratique, il n'est jamais résolu exactement. On calcule seulement une approximation $\tilde{z}_i$, généralement par une méthode itérative, du système

$$C z = \tilde{b}_i, \tag{4}$$

vérifiant

$$\|\tilde{b}_i - C\tilde{z}_i\| \leqslant \varepsilon \tag{5}$$

avec

$$\tilde{b}_i = \sum_{l=0}^{q} (\xi_l A + \psi_l B) \tilde{y}_{i-l} + \sum_{l=-1}^{q} \phi_l f(t_{i-l}), \tag{6}$$

et la suite $(\tilde{y}_i)_{q+1 \leqslant i \leqslant N}$ est définie de manière analogue à (2) :

$$\tilde{y}_0 = y^{(0)}, \qquad \tilde{y}_{i+1} = \tilde{y}_i + h\tilde{z}_i \quad \text{pour } i = q, q+1, \ldots, N-1, \tag{7}$$

où, pour $i = 1, \ldots, q$, $\tilde{y}_i$ est calculé par un schéma à i pas. Dans (5), ε est un paramètre de précision et le symbole $\| \ \|$ désigne la norme euclidienne.

Rappelons que le but de notre approche est précisément de calculer une approximation de la solution du système (4) qui sera utilisée comme solution initiale dans une méthode itérative afin d'aboutir plus rapidement au calcul de $\tilde{z}_i$.

Le principe de l'approche proposée nécessite un sous-espace vectoriel $\mathcal{V}_i$ de $\mathbb{R}^n$ à partir duquel on détermine la solution initiale, notée $z_i^{(0)}$, dite de Petrov–Galerkin, qui est définie comme l'unique solution du problème

$$\text{trouver } z_i^{(0)} \in \mathcal{V}_i \quad \text{tel que} \quad \langle \tilde{b}_i - C z_i^{(0)}, C v \rangle = 0 \text{ pour tout } v \in \mathcal{V}_i, \tag{8}$$

où $\langle , \rangle$ désigne le produit scalaire euclidien sur $\mathbb{R}^n$. La résolution de ce problème implique un coût de calcul très modeste dès lors que la dimension de $\mathcal{V}_i$ est petite. L'intérêt de $z_i^{(0)}$ est de vérifier la condition de minimization

$$\|\tilde{b}_i - C z_i^{(0)}\| = \min_{z \in \mathcal{V}_i} \|\tilde{b}_i - Cz\|.$$

La question qui se pose alors est le choix de ce sous-espace $\mathcal{V}_i$. En se basant sur la théorie des schémas multi-pas, il apparaît que prendre pour $\mathcal{V}_i$ le sous-espace engendré par les vecteurs $\tilde{z}_{i-r}, \ldots, \tilde{z}_{i-1}$ constitue un choix performant.

Or il s'avère qu'en pratique, certaines approximations de Petrov–Galerkin $z_i^{(0)}$ vérifient directement (5). Il suffit alors de poser $\tilde{z}_i = z_i^{(0)}$. Il n'est donc plus nécessaire, durant cette itération, ni d'utiliser la méthode itérative, ni de modifier le sous-espace $\mathcal{V}_i$. Il est à noter que ces situations sont très intéressantes d'un point de vue pratique car elles sont synonymes d'une forte réduction du coût de calcul. Ceci nous amène à considérer le sous-espace $\mathcal{V}_i$ comme étant le sous-espace vectoriel engendré par les r derniers vecteurs, $\{\tilde{z}_{i-l_j}\}_{1 \leqslant j \leqslant r}, l_j < l_{j+1}$, dont le calcul a nécessité l'utilisation de la méthode itérative. Dans le théorème suivant nous montrons que $z_i^{(0)}$ est une bonne solution initiale pour le système (4) :

Théorème 1.1. *Si f est de classe C^{r+v+l_1-2} et le schéma* (2) *et celui du démarrage sont stables et d'ordre p, alors l'approximation de Petrov–Galerkin vérifie*

$$\|\tilde{b}_i - C z_i^{(0)}\| = \mathrm{O}(h^p) + \mathrm{O}(h^{r+l_1-1}) + \mathrm{O}(\varepsilon) \quad \textit{pour } i = q, \ldots, N-1.$$

Preuve. Il suffit de montrer l'existence d'un vecteur $z \in \mathcal{V}_i$ tel que

$$\|\tilde{b}_i - Cz\| = \mathrm{O}(h^p) + \mathrm{O}(h^{r+l_1-1}) + \mathrm{O}(\varepsilon).$$

Puisque $\tilde{z}_{i-k} = z_{i-k}^{(0)} \in \mathcal{V}_i$ pour $k = 1, \ldots, l_1 - 1$, le sous-espace $\mathcal{V}_i$ est, en fait, le sous-espace vectoriel engendré par $\tilde{z}_{i-1}, \ldots, \tilde{z}_{i-l_1+1}$ et $\tilde{z}_{i-l_1}, \ldots, \tilde{z}_{i-l_r}$.

Posons $m = r + l_1 - 1$ et $l_j = j - r$ pour $j = r+1, \ldots, m$. En utilisant, par exemple, les polynômes d'interpolation de Lagrange, on montre qu'il existe des constantes $\{\alpha_{j,m}\}_{1 \leqslant j \leqslant m}$ telles que

$$\left\| y(t_{i-l}) - \sum_{j=1}^{m} \alpha_{j,m}\, y(t_{i-l-l_j}) \right\| = \mathrm{O}(h^m), \quad 0 \leqslant l \leqslant q \tag{9}$$

et

$$\left\| f(t_{i-l}) - \sum_{j=1}^{m} \alpha_{j,m}\, f(t_{i-l-l_j}) \right\| = \mathrm{O}(h^m), \quad -1 \leqslant l \leqslant q. \tag{10}$$

Soit $z = \sum_{j=1}^{m} \alpha_{j,m} \tilde{z}_{i-l_j} \in \mathcal{V}_i$. En utilisant (5), (6) et (10) on obtient

$$\begin{aligned} \|\tilde{b}_i - Cz\| &\leqslant \left\| \tilde{b}_i - \sum_{j=1}^{m} \alpha_{j,m} \tilde{b}_{i-l_j} \right\| + \left\| \sum_{j=1}^{m} \alpha_{j,m} (\tilde{b}_{i-l_j} - C\tilde{z}_{i-l_j}) \right\| \\ &= \sum_{l=0}^{q} \|\xi_l A + \psi_l B\| \left\| \tilde{y}_{i-l} - \sum_{j=1}^{m} \alpha_{j,m} \tilde{y}_{i-l-l_j} \right\| + \mathrm{O}(h^m) + \mathrm{O}(\varepsilon). \end{aligned}$$

Montrons que $\|\tilde{y}_{i-l} - \sum_{j=1}^{m} \alpha_{j,m} \tilde{y}_{i-l-l_j}\| = \mathrm{O}(h^p) + \mathrm{O}(h^m) + \mathrm{O}(\varepsilon)$, ce qui achèvera la démonstration.

Comme le schéma est stable et d'ordre p, on a (voir [2, p. 72]) :

$$\max_{q+1 \leqslant i \leqslant N} \|\tilde{y}_i - y_i\| = \mathrm{O}(\varepsilon) \quad \text{et} \quad \|y_{i-l} - y(t_{i-l})\| = \mathrm{O}(h^p) \quad \text{pour } l = 0, \ldots, q.$$

D'autre part, puisque les vecteurs $y_1, \ldots, y_q$ sont calculés par un schéma à i pas, stable et d'ordre p, on montre comme précédemment que $\max_{1\leqslant i\leqslant q} \|\tilde{y}_i - y_i\| = \mathrm{O}(\varepsilon)$ et $\max_{1\leqslant i\leqslant q} \|y(t_i) - y_i\| = \mathrm{O}(h^p)$. On en déduit que $\|y(t_{i-k}) - \tilde{y}_{i-k}\| = \mathrm{O}(h^p) + \mathrm{O}(\varepsilon)$, pour $k = 0, \ldots, i$.

En utilisant (9), on obtient pour $l = 0, \ldots, q$:

$$\left\| \tilde{y}_{i-l} - \sum_{j=1}^{m} \alpha_{j,m} \tilde{y}_{i-l-l_j} \right\| = \left\| \left(\tilde{y}_{i-l} - y(t_{i-l})\right) + \sum_{j=1}^{m} \alpha_{j,m} \left(y(t_{i-l-l_j}) - \tilde{y}_{i-l-l_j}\right) \right\| + \mathrm{O}\left(h^m\right)$$
$$= \mathrm{O}\left(h^p\right) + \mathrm{O}\left(h^m\right) + \mathrm{O}(\varepsilon). \qquad \square$$

Dans le cas où B est la matrice d'identité ou une matrice facilement inversible, on peut obtenir un résultat plus précis en spécifiant le schéma implicite utilisé. Par exemple, lorsque B est la matrice d'identité, on montre le résultat suivant (voir [1]) :

Théorème 1.2. *Si f est de classe C^r, $r \leqslant 5$, et la suite $(y_i)_{0\leqslant i\leqslant N}$ définie dans* (2) *est obtenue par le schéma d'Euler implicite ou celui de Crank–Nicolson, alors l'approximation de Petrov–Galerkin, obtenue avec $\mathcal{V}_i = \mathrm{vect}(\tilde{z}_{i-l}, 1 \leqslant l \leqslant r)$ vérifie*

$$\left\| \tilde{b}_i - C z_i^{(0)} \right\| = \mathrm{O}\left(h^5\right) + \mathrm{O}(\varepsilon).$$

Nous n'avons pas effectué la démonstration pour $r > 5$ car le calcul devient encombrant. Les aspects théorique et algorithmique sont présentés en détail dans [1].

Remerciements

Je remercie Miloud Sadkane pour les discussions mathématiques et pour ses conseils.

Références

[1] M. Al Sayed Ali, Accélération de schémas d'intégration temporelle pour la résolution d'équations différentielles, Thèse de l'Université de Bretagne Occidentale, Brest, 2007.

[2] K.E. Brenan, S.L. Campbell, L.R. Petzold, Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations, second ed., SIAM Publications, Philadelphia, PA, 1996.

[3] M. Crouzeix, A.L. Mignot, Analyse numérique des équations différentielles, 2e édition, Masson, Paris, 1991.

[4] P. Kunkel, V. Mehrmann, Differential-Algebraic Equations Analysis and Numerical Solution, EMS Publishing House, Zürich, Switzerland, 2006.

[5] Y. Saad, Iterative Methods for Sparse Linear Systems, second ed., SIAM, Philadelphia, PA, 2003.