RAPPORT DE STAGE DE DEA/ENSIMAG

Estimation de l'erreur globale lors de l'intégration numérique des systèmes algébro-différentiels d'indice un et deux

Claude-Pierre Jeannerod
ENSIMAG - 3ème année

1996 - 1997

LMC - IMAG / DER - EDF
Responsables : Gilles Villard
René Aïd

Remerciements

Je tiens à remercier Gilles Villard, responsable de l'équipe Calcul Formel (LMC-IMAG) d'avoir bien voulu encadrer ce stage de DEA ainsi que Laurent Levacher de la Direction des Etudes et Recherches d'EDF sans qui ce travail n'aurait pas été possible et qui m'a permis de le faire dans les meilleures conditions. Je remercie également et tout particulièrement René Aïd et Josselin Visconti, thésards de l'équipe Calcul Formel, pour le temps qu'ils ont bien voulu m'accorder, le suivi qu'ils ont assuré pendant toute cette période, leurs précieux conseils et leur bonne humeur.

Introduction

Les équation algébro-différentielles (EAD) apparaissent naturellement dans de nombreuses applications (modélisation de la dynamique des réseaux électriques, thermodynamique, contrôle optimal, ...) mais présentent, tant du point de vue numérique que du point de vue analytique, des difficultés qu'on ne rencontre pas avec les équations différentielles ordinaires (EDO). Cependant, qu'il s'agisse d'EAD ou d'EDO, la résolution numérique de tels problèmes pose la question de la mesure de l'erreur par rapport à la solution exacte. Comme dans la plupart des cas réels on ne dispose pas de cette solution, l'erreur commise ne peut être calculée exactement. Certaines situations exigent pourtant une bonne connaissance de cette grandeur et il faut donc pouvoir l'*estimer* précisément.

On se limite ici à l'étude des EAD semi-explicites d'indice 1 et 2 pour lesquelles la question de l'estimation de l'erreur globale est posée à partir de résultats obtenus pour les EDO. Le premier chapitre présente l'application qui est à l'origine de ce travail. Il s'agit d'un problème de simulation de réseaux électriques traité par la société EDF. Cette formulation du problème permet aussi de préciser le cadre dans lequel l'estimation de l'erreur globale doit se faire (méthodes numériques, estimateurs,...). On rappelle ensuite au chapitre 2 quelques éléments de la théorie de l'estimation de l'erreur globale pour les EDO dont on se servira par la suite. Le chapitre 3 est consacré aux EAD semi-explicites d'indice 1. On y propose, pour une classe de méthodes d'intégration, une extension possible des principaux résultats vus pour les EDO. Le cas de l'indice 2 est envisagé au chapitre 4 : au vu des tests réalisés, quels résultats théoriques peut-on espérer dans ce cas ? On s'intéresse enfin à la possibilité de recourir à l'estimation de l'erreur globale pour les EAD d'indice 1 lors du traitement des problèmes à perturbation singulière.

Table des matières

Chapitre 1

Position du problème

1.1 Formulation

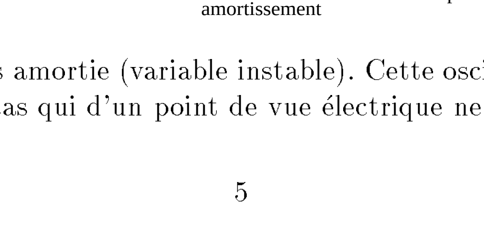
Actuellement la modélisation de la dynamique des réseaux électriques constitue un thème de recherche important, notamment pour la société Electricité de France (EDF). Dans ce contexte, EDF a développé des outils de simulation dont le logiciel EUROSTAG. Celui-ci réalise en particulier l'intégration numérique des équations différentielles qui modélisent de tels réseaux.

1.1.1 Stabilité d'un réseau

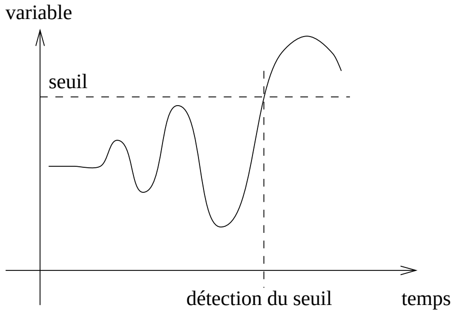
Suite à un incident sur le réseau électrique (chute d'un arbre sur une ligne, augmentation brutale de la demande en électricité de la part des usagers, ...), certaines variables du système simulé par EUROSTAG deviennent *instables*. Des actions doivent alors être menées pour ramener le réseau dans une situation normale. Dans certains cas le réseau parvient à se stabiliser, dans d'autres il faut délester certaines zones géographiques afin de ne pas perturber l'ensemble du réseau électrique. Ces actions de stabilisation sont automatiques et sont menées suite à la détection de seuils ou à la comptabilisation du nombre de battements des oscillations.

Le problème peut être formulé de la façon suivante. Considérons une variable du système qui, suite à un évènement, est perturbée.

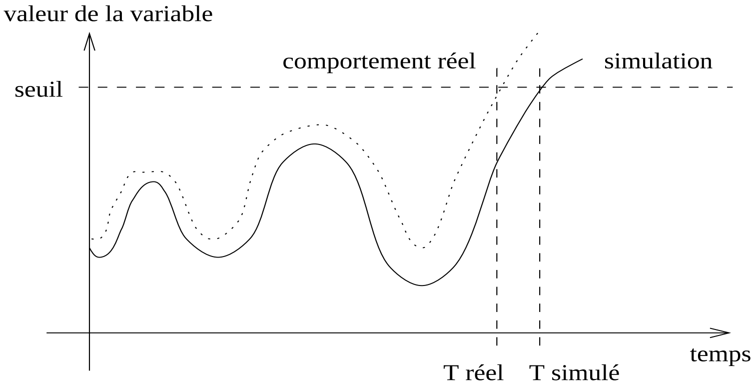
Soit l'oscillation est amortie et la variable retourne à un point d'équilibre (variable stable).



Soit l'oscillation n'est pas amortie (variable instable). Cette oscillation peut être entrenue et bornée ou divergente, deux cas qui d'un point de vue électrique ne sont pas acceptables.



Des actions doivent donc être menées afin de ramener le système dans un nouvel état d'équilibre. Pour cela, on cherche à détecter l'instant de dépassement du seuil par la variable suffisamment tôt. Il s'agit donc de détecter ce seuil de façon sûre.



Il s'avère qu'une erreur sur la valeur de la variable est commise lors de l'intégration numérique réalisée par le simulateur. La validité de la détection du seuil à ne pas dépasser étant fondamentale, il faut donc pouvoir estimer cette erreur précisément.

1.1.2 Le modèle mathématique

Au niveau de modélisation choisi par EUROSTAG, la description mathématique d'un réseau électrique conduit à un système d'EAD de la forme :

$$\begin{array}{rcl} y' & = & f(y,z) \\ 0 & = & g(y,z) \end{array} \tag{1.1}$$

Un tel système est dit *semi-explicite* au sens où la partie différentielle (première équation) et la partie algébrique (deuxième équation) sont séparées. Dans le cas contraire, on parle d'EAD *implicite*

$$F(Y,Y') = 0 \tag{1.2}$$

où $\frac{\partial F}{\partial Y'}$ est singulière.

Il s'agit alors de déterminer le comportement de la solution de ce système à partir d'une condition initiale sur les fonctions y et z : $y(0) = y_0, z(0) = z_0$. On choisit ces valeurs de sorte que : $g(y_0, z_0) = 0$ (conditions initiales *consistantes*).

1.1.3 Nature du système traité par EUROSTAG

Le système différentiel intégré par le simulateur EUROSTAG est comme on l'a vu un système algébro-différentiel. Aux difficultés propres à cette classe d'équations s'ajoutent les caractéristiques suivantes :

- ce système est non linéaire.
- il présente des constantes de temps d'ordres de grandeur très différentes (phénomènes très rapides de l'ordre de la milliseconde à des phénomènes lents de l'odre de la minute), ce qui en fait un système *raide* (cf. [14]).
- les fonctions f et g ainsi que leurs dérivées premières peuvent présenter des discontinuités par rapport à la variable différentielle y. Ceci est dû à l'existence de mécanismes de régulation des tensions discontinues.

EUROSTAG calcule la solution de ce système par des méthodes numériques de type BDF et Adams, à pas et ordre variables.

1.2 Estimation de l'erreur globale

1.2.1 Les systèmes étudiés

Cette étude porte sur une version simplifiée du cas traité par le simulateur EUROSTAG. On s'intéresse ici à des systèmes d'équations différentielles algébriques semi-explicites d'indice un, c'est-à-dire de la forme (1.1) et tels que :

- f et g sont suffisamment différentiables.
- $(g_z)^{-1}$ existe et est borné dans un voisinage de la solution.

Cette dernière hypothèse est celle qui fait entrer les équations étudiées dans la classe des EAD d'indice un. Quant à la régularité de f et g, elle permet de disposer d'un bon cadre théorique.

Le calcul d'une solution approchée du système (1.1) se fera avec une méthode non nécessairement identique à celle utilisée dans EUROSTAG. En effet, les deux solveurs dont on dispose, RADAU5 et DASSL, nous conduiront à étudier en priorité les méthodes sur lesquelles ils reposent.

1.2.2 Quelle erreur estimer ?

Lors de l'intégration d'un système différentiel, on peut considérer 4 types d'erreurs : l'*erreur de modélisation* (liée au choix du modèle), l'*erreur numérique* (qui dépend du calculateur utilisé), l'*erreur de troncature* et l'*erreur globale*, qui est celle que nous allons étudier. Rappelons que l'on cherche à déterminer si la simulation numérique fournit dans tous les cas des résultats fiables, ce qui peut exiger une évaluation précise de l'erreur commise lors de l'intégration numérique.

Les erreurs qui nous intéressent sont donc l'erreur de troncature et l'erreur globale. Toutes les deux dépendent de la méthode d'intégration employée.

La première est l'erreur commise à chaque pas d'intégration, en supposant que celle-ci s'effectue à partir des valeurs exactes de la solution au temps précédent. Lorsqu'elle se propage au cours du temps, elle conduit à l'erreur globale, qui mesure l'erreur entre la solution exacte et son approximation numérique. En général l'estimation de l'erreur globale coûte cher et on utilise plutôt l'erreur de troncature, d'autant plus qu'il existe une relation liant ces deux types d'erreur. D'après cette relation, si l'erreur de troncature reste en deçà d'un seuil alors l'erreur globale restera aussi faible. Cependant, pour certains systèmes dits *instables*, cette relation n'est plus vérifiée et l'erreur de troncature peut rester faible alors que l'erreur globale croît très rapidement. Ce phénomène, lié à l'instabilité du système intégré (cas du problème d'EDF) et au choix de la méthode numérique utilisée, motive donc la recherche d'estimateurs fiables de l'*erreur globale*.

1.2.3 Quels estimateurs pour quelles méthodes?

Il s'agit au cours de cette étude de considérer trois types d'estimateurs figurant parmi les techniques d'estimation de l'erreur globale les plus efficaces [2]:

- extrapolation de Richardson (RS),
- algorithme de Zadunaisky (ZD),
- calcul d'une Correction Globale (SC, pour *Solving for the Correction*).

Ces méthodes d'estimation ont déjà été largement étudiées dans le cadre des équations différentielles ordinaires et pour des méthodes à pas et ordre variables dans [2, 3, 15, 16].

Précisons enfin qu'on ne cherche pas à construire des estimateurs propres aux systèmes algébro-différentiels semi-explicites mais plus simplement à utiliser, et éventuellement adapter, les trois méthodes classiques au cas d'indice 1. Cette extension de RS, ZD et SC doit se fera essentiellement autour des méthodes de Runge-Kutta implicites (RADAU5), qui sont particulièrement adaptées aux systèmes algébro-différentiels. Les méthodes BDF fortement stables (DASSL) seront étudiées pour RS uniquement.

Chapitre 2

Estimation de l'erreur globale pour les EDO

Soit l'équation différentielle ordinaire :

$$y' = f(t, y), \quad y(0) = y_0 \tag{2.1}$$

où $t \in [0, T], T > 0$ et f est une fonction de $\mathbb{R} \times \mathbb{R}^n$ dans $\mathbb{R}^n$ localement lipschitzienne en y. Cela assure l'existence et l'unicité de la solution passant par tout point d'un voisinage de $(0, y_0)$. On prend désormais $n = 1$ sans perte de généralité.

2.1 Méthodes à un pas

2.1.1 Méthodes de Runge-Kutta pour les EDO

Les méthodes de Runge-Kutta ont été initialement conçues pour résoudre des EDO du type (2.1). Ces méthodes sont dites *à un pas* au sens où, à partir d'une valeur approchée y_n de la solution exacte à l'instant t_n, elles permettent de construire une approximation y_{n+1} à l'instant $t_{n+1} = t_n + h$ grâce aux formules

$$y_{n+1} = y_n + h \sum_{i=1}^{s} \underline{\mathrm{b}}_i Y'_{ni} \tag{2.2a}$$

où Y'_{ni} est défini par [1]

$$Y'_{ni} = f(Y_{ni})$$

et où les pas intermédiaires (appelés aussi *points*) sont donnés par

$$Y_{ni} = y_n + h \sum_{j=1}^{s} \underline{\mathrm{a}}_{ij} Y'_{nj} \qquad i = 1, \ldots, s. \tag{2.2b}$$

La méthode est déterminée par les coefficients $\underline{\mathrm{a}}_{ij}$, $\underline{\mathrm{b}}_i$ et s est le nombre de points. Si $\underline{\mathrm{a}}_{ij} = 0$ pour $i \leq j$, on peut calculer les valeurs numériques aux pas intermédiaires $Y_{n1}, \ldots, Y_{ns}$ les unes

1. Y'_{ni} est une notation qui ne représente pas une dérivée mais permet de faire apparaître dans le schéma numérique la forme de l'EDO initiale (2.1) : dans ce sens, $Y'_{ni} = f(Y_{ni})$ 'ressemble' à $y' = f(y)$.

après les autres grâce à (2.2b) en évaluant explicitement la fonction f. Ces méthodes sont dites *explicites* (ERK pour Explicit Runge-Kutta). Sinon, (2.2b) forme un système non linéaire en les Y_{ni} et la méthode est dite *implicite* (IRK pour Implicit Runge-Kutta).
On décrit habituellement les méthodes de Runge-Kutta de façon condensée avec les *tableaux de Butcher* en posant $\underline{c}_i = \sum_{j=1}^{s} \underline{a}_{ij}\,, \quad i = 1, \ldots, s$:

$$\begin{array}{c|c} \underline{c}_i & \underline{a}_{ij} \\ \hline & \underline{b}_j \end{array}$$

Une méthode explicite s'écrit donc

$$\begin{array}{c|ccc} & 0 & \cdots & 0 \\ \underline{c}_i & & \ddots & \vdots \\ & \underline{a}_{ij} & & 0 \\ \hline & \underline{b}_j & & \end{array}$$

On rappelle maintenant quelques définitions concernant la méthode (2.2) et la solution numérique obtenue avec cette méthode.

Définition 1 *L'*erreur globale *de la solution numérique est donnée à chaque instant t_n par*

$$E_n = y_n - y(t_n)$$

*On définit l'*erreur locale *en t_{n+1} par*

$$e_{n+1} = y_{n+1} - u(t_{n+1})$$

où u est la solution exacte de l'équation différentielle :

$$\begin{array}{rcl} u' & = & f(t, u) \\ u(t_n) & = & y_n \end{array}$$

*Enfin l'*erreur de troncature *est définie par*

$$\varepsilon_{n+1} = y(t_{n+1}) - y(t_n) - h \sum_{i=1}^{s} \underline{b}_i \tilde{Y}'_{ni}$$

où

$$\tilde{Y}'_{ni} = f\left(y(t_n) + h \sum_{j=1}^{s} \underline{a}_{ij} \tilde{Y}'_{nj} \right)$$

Cette équation est obtenue en remplaçant dans (2.2) la solution numérique y_n par la valeur exacte $y(t_n)$. La relation (2.2) n'est alors plus vérifiée et ε_{n+1} définit la quantité dont la solution exacte y ne vérifie pas le schéma numérique.

Pour les rapports entre ces trois types d'erreurs, on peut consulter [6].

Définition 2 *La méthode (2.2) est dite* d'ordre p *quand*

$$\varepsilon_n = \mathcal{O}(h^{p+1})$$

Ceci implique en général $E_n = \mathcal{O}(h^p)$.

Définition 3 *La méthode (2.2) est* stable *s'il existe une constante M indépendante de h telle que les valeurs obtenues par*

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i Y'_{ni}$$

$$z_{n+1} = z_n + h \sum_{i=1}^{s} b_i Z'_{ni} + \epsilon_n$$

avec Y'_{ni} et Z'_{ni} définis comme dans (2.2) vérifient

$$\|z_n - y_n\|_\infty \leq M \|\epsilon_n\|_\infty$$

Définition 4 *La méthode (2.2) est* convergente *si* $\lim_{h\emptyset_y 0} \|E_n\|_\infty = 0$.

2.1.2 Développement asymptotique de l'erreur globale

On suppose qu'on intègre (2.1) avec la méthode (2.2). L'erreur globale commise en y permet de définir la notion d'estimateur.

Définition 5 *On appelle* estimateur de E_n *une suite (E_n^*) telle que $E_n^* = E_n(1 + \mathcal{O}(h^k))$ uniformément en n, avec $k \geq 1$.*

Lorsqu'un tel estimateur existe, il est dit *d'ordre relatif k*. Comme on l'a vu lors de la formulation du problème, on s'intéresse aux trois estimateurs RS, ZD et SC. Le premier est directement construit à partir du développement asymptotique de l'erreur globale et les deux autres, bien que fondés sur une fonction d'interpolation des y_n, doivent leur caractère asymptotiquement correct à ce même développement. Le théorème suivant [13] en précise l'allure dans le cas des EDO et pour une intégration à pas constant avec une méthode à un pas.

On suppose que la méthode numérique utilisée est d'ordre p et que l'erreur de troncature en t_n admet le développement asymptotique suivant

$$\varepsilon_{n+1} = d_{p+1}(t_n)h^{p+1} + \ldots + d_{N+1}(t_n)h^{N+1} + \mathcal{O}(h^{N+2}) \tag{2.3}$$

Théorème 1 *Si l'on intègre (2.1) avec une méthode numérique à un pas, stable, d'ordre $p \geq 1$, on a alors*

$$y_n - y(t_n) = h^p a_p(t_n) + \ldots + h^N a_N(t_n) + \mathcal{O}(h^{N+1}) \tag{2.4}$$

où a_k, $k = p, \ldots, N$ est solution de l'equation variationnelle

$$\begin{array}{rcl} a'_k & = & f_y(t,y)a_k + \Psi_k(t) \\ a_k(0) & = & 0 \end{array} \tag{2.5}$$

où Ψ_k est un terme inhomogène. Dans le cas $k = p$, on a $\Psi_p \equiv d_{p+1}$ et l'équation vérifiée par a_p est appelée équation variationnelle de l'erreur.

2.1.3 Trois techniques d'estimation de l'erreur globale

Les trois techniques d'estimation de l'erreur globale étudiées sont à rapprocher de deux principes courants en intégration numérique : les méthodes d'extrapolation qui consistent à considérer une autre grille pour l'intégration et les méthodes de correction de défaut itérée (*iterated defect correction* en anglais) pour lesquelles on perturbe légèrement le problème de départ. Dans les deux cas, l'objectif initial est de construire une méthode numérique d'ordre supérieur [13], [18]. La figure 2.1 montre comment se répartissent RS, ZD et SC.

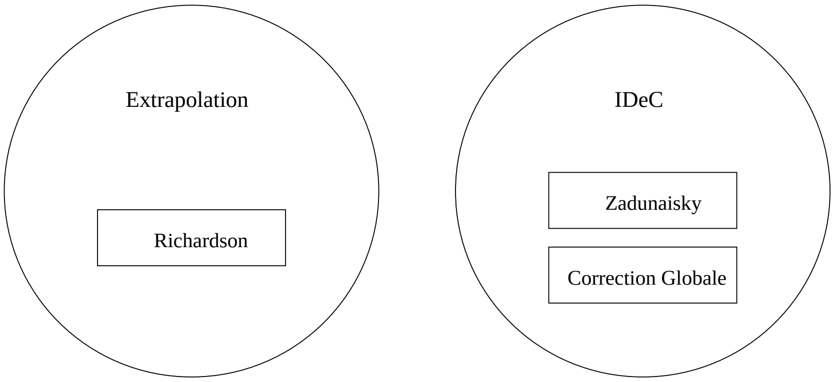


FIG. 2.1 –

• Extrapolation de Richardson (RS)

Le principe de cette méthode est le suivant. On réalise simultanément deux intégrations numériques, l'une avec le pas h, l'autre avec le pas $h/2$. Les valeurs obtenues à l'instant t_n sont notées y_n et y^*_{2n} respectivement. On obtient alors en tronquant le développement de l'erreur globale $E_n = y_n - y(t_n)$ à l'ordre $p+1$

$$\begin{aligned} y_n &= y(t_n) + h^p a_p(t_n) + \mathcal{O}(h^{p+1}) \\ y^*_{2n} &= y(t_n) + (h/2)^p a_p(t_n) + \mathcal{O}(h^{p+1}) \end{aligned}$$

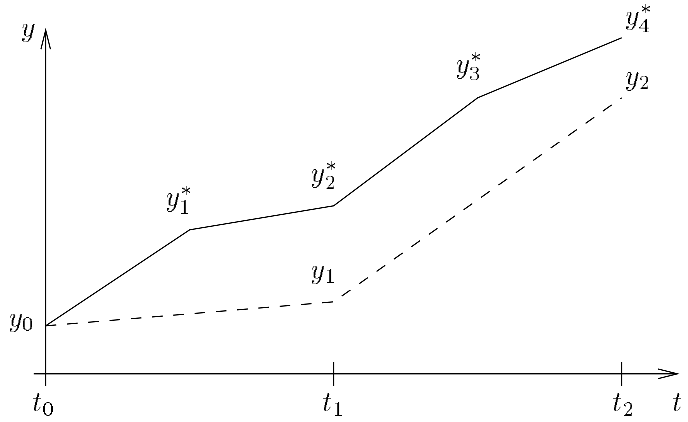
Et donc, en éliminant $y(t_n)$

$$\frac{y_n - y^*_{2n}}{1 - 2^{-p}} = h^p a_p(t_n) + \mathcal{O}(h^{p+1}) \tag{2.6}$$

On en déduit directement l'estimateur RS défini par

$$E^*_{RS,n} = \frac{y_n - y^*_{2n}}{1 - 2^{-p}}$$

D'après (2.6), cet estimateur est asymptotiquement correct d'ordre relatif 1. La figure ci-dessous montre comment les deux intégrations se font de façon totalement indépendante : y^*_4 est obtenu à partir de y_3^* et non de y_1.



- **Algorithme de Zadunaisky (ZD)**

Il s'agit de construire un problème perturbé mais proche de (2.1) et dont on connaisse la solution exacte. On le définit de la façon suivante

$$\hat{y}' = f(t, \hat{y}) + d_h(t) \tag{2.7}$$

où d_h est le défaut introduit.

Généralement on prend $d_h(t) = P'_h(t) - f(t, P_h(t))$ avec P_h un polynôme d'interpolation des y_n qui dépend du pas h. En intégrant (2.7) avec la même méthode que celle utilisée pour (1.1), on obtient des valeurs $\hat{y}_n$. Une estimation de l'erreur globale E_n à l'instant t_n est alors donnée par

$$E^*_{ZD,n} = \hat{y}_n - y_n$$

Pratiquement, on utilise plusieurs polynômes $P_{h,k}$ de degré m qui interpolent les y_{km+i}, $i = 0, \ldots, m$, $k \geq 0$. Les m premiers pas de l'intégration de (1.1) permettent de construire $P_{h,0}$ et le problème perturbé correspondant

$$\begin{aligned} \hat{y}' &= f(t, \hat{y}) + P'_{h,0}(t) - f(t, P_{h,0}(t)) \\ \hat{y}(0) &= y_0 \end{aligned}$$

En intégrant ce problène perturbé avec la même méthode que celle utilisée pour (1.1), on obtient les $\hat{y}_i$, $i = 0, \ldots, m$ et une estimation de l'erreur globale E_i :

$$E^*_{ZD,i} = \hat{y}_i - y_i \,, \quad i = 0, \ldots, m$$

Ce processus, répété pour chaque séquence de m pas, fournit donc bien une estimation $E^*_{ZD,n}$ de l'erreur globale à chaque instant t_n.

En ce qui concerne l'ordre de cet estimateur, on a le résultat suivant, dû à Frank et Uebe-rhuber [7, 8].

Théorème 2 *Pour une méthode d'intégration à pas constant d'ordre p, si l'erreur globale du problème initial (2.1) et celle du problème perturbé (2.7) admettent chacune un développement asymptotique de la forme (2.4) et si P_k est le polynôme de degré m qui interpole les $y_{km+i}, i = 0, \ldots, m$, alors l'ordre relatif de l'estimateur de Zadunaisky est :*

$$\begin{array}{ccc} 0 & si & m \leq p \\ m-p & si & p \leq m \leq 2p \\ p & si & 2p \leq m \end{array}$$

Cette méthode permet donc d'obtenir au mieux un estimateur d'ordre relatif p (en prenant $m = 2p$), ce qui représente cependant une amélioration considérable de l'ordre fourni par l'extrapolation de Richardson. De plus, cette précision est obtenue sans surcoût significatif. La preuve du théorème 2 est détaillée dans [8]. On se contente ici d'en donner le principe en vue d'une extension aux EAD d'indice 1.

Schéma de la preuve du théorème 2 :
Comme on intègre le problème de départ (1.1) et le problème perturbé (2.8) avec la même méthode d'ordre p, on a, concernant l'erreur globale pour cette méthode, les développements asymptotiques suivants :

$$\begin{aligned} y_n &= y(t_n) + h^p a_p(t_n) + \ldots \\ \hat{y}_n &= \hat{y}(t_n) + h^p \hat{a}_p(t_n) + \ldots \end{aligned}$$

que l'on note $DA_y(t_n)$ et $DA_{\hat{y}}(t_n)$ respectivement.
L'estimation de l'erreur globale étant donnée par $\hat{y}_n - \hat{y}(t_n) = \hat{y}_n - y_n$, on cherche l'ordre jusqu'auquel les deux développements coïncident.

On étudie l'ordre des différences $a_p(t) - \hat{a}_p(t)$. Cela se fait en deux temps. Les équations variationnelles associées aux deux développements asymptotiques sont, pour $k \geq p$

$$\begin{aligned} {a_k}' &= f_y(y) a_k + \Psi_k(t) \\ \hat{a}_k' &= f_y(\hat{y}) \hat{a}_k + \hat{\Psi}_{h,k}(t) \end{aligned}$$

On obtient alors par différence l'expression suivante

$$\hat{a}_k' - {a_k}' = f_y(y)[\hat{a}_k(t) - a_k(t)] + [f_y(\hat{y}) - f_y(y)]\hat{a}_k(t) + \hat{\Psi}_{h,k}(t) - \Psi_k(t) \qquad (2.8)$$

Comme f_y est lipschitzienne, on en déduit :

$$[f_y(\hat{y}) - f_y(y)]\hat{a}_k(t) = \mathcal{O}(h^p)$$

Il faut enfin évaluer $\hat{\Psi}_{h,k}(t) - \Psi_k(t)$, ce qui constitue le coeur du résultat. Pour cela, on a besoin du lemme suivant :

Lemme 1 *Soit $\mu = min(m, 2p)$. Le défaut $d_h(t)$ du problème perturbé vérifie pour tout k*

$$d_h^{(k)}(t) = \mathcal{O}(h^{max(0, min(p, \mu - k))})$$

Le formalisme des arbres de Butcher permettant décrire pour $p \leq k \leq 2p-1$ les fonctions Ψ_k et $\hat{\Psi}_{h,k}$ comme des combinaisons linéaires des différentielles élémentaires de y et $\hat{y}$:

$$\Psi_k(t) = \sum_{u \in T,\, \rho(u)=k+1} a_k(u) F(u)(y)(t)$$
$$\hat{\Psi}_{h,k}(t) = \sum_{u \in T,\, \rho(u)=k+1} a_k(u) \hat{F}_h(u)(\hat{y})(t)\,,$$

on se ramène à l'étude $\hat{F}_h(u)(\hat{y})(t) - F(u)(y)(t)$. Or cette différence est une somme de termes dont le terme dominant en h est $d_h^{(k)}(t)$. En appliquant le lemme précédent, on a alors l'ordre de $\hat{F}_h(u)(\hat{y})(t) - F(u)(y)(t)$ et donc celui de $\hat{\Psi}_{h,k}(t) - \Psi_k(t)$. Il apparaît que ce terme domine toujours sur $[f_y(\hat{y}) - f_y(y)]\hat{a}_k(t)$. Le lemme de Gronwall appliqué à (2.8) permet de conclure. ■

- **Calcul d'une correction globale (SC)**

l'idée est assez proche de la technique précédente. On définit, toujours à l'aide d'un polynôme P_h interpolant les y_n, la fonction

$$\mathcal{E}_h = P_h - y$$

Elle vérifie l'équation différentielle ordinaire

$$\mathcal{E}_h' = P_h'(t) - f(t, P_h(t) - \mathcal{E}_h) \tag{2.9}$$

L'intégration de cette EDO avec une méthode non nécessairement identique à celle utilisée pour (2.1) fournit alors directement une estimation $\varepsilon_{h,n}$ de l'erreur globale $\varepsilon_h(t_n)$ à l'instant t_n :

$$E^*_{SC,n} = \varepsilon_{h,n}$$

Lors de l'implémentation de cet algorithme, on considère en fait les mêmes polynômes d'interpolation $P_{h,k}$ que pour ZD et on définit les fonctions

$$\mathcal{E}_{h,k} = P_{h,k} - y$$

vérifiant chacune l'equation différentielle ordinaire

$$\mathcal{E}_{h,k}' = P_{h,k}' - f(t, P_{h,k} - \mathcal{E}_{h,k}) \quad t \in [\,t_{km}, t_{(k+1)m}\,] \tag{2.10}$$

On choisit $\mathcal{E}_{h,0}(0) = 0$ et, pour assurer la continuité de la solution, $\mathcal{E}_{h,k}(t_{km})$ prendra la dernière valeur numérique calculée sur $[\,t_{(k-1)m}, t_{km}\,]$. De même que pour ZD, répéter ce procédé conduit à une estimation de l'erreur globale à chaque instant t_n. Par ailleurs, le théorème 2 est encore valable ici. Dans [1], on démontre en effet le théorème suivant.

Théorème 3 *Sous les hypothèses du théorème 2 et si on intègre (2.9) avec une méthode à un pas d'ordre q ($q \neq p$), on a*

$$\varepsilon_{h,n} = \mathcal{O}(h^{max(0, min(p, m-k))})$$

Remarque : en pratique, on intègre les deux problèmes avec la même méthode d'ordre p, ce qui permet d'obtenir une estimation de l'erreur globale d'ordre relatif au plus égal à p.

La preuve du théorème 3 consiste à se ramener à celle du théorème 2 et l'idée qu'on en donne donnés ci-dessous servira lors de l'étude des EAD d'indice 1.

Schéma de la preuve du théorème 3 :
Comme P_h est le même polynôme que pour ZD, il est en particulier solution du problème perturbé (2.7) que l'on note

$$\hat{y}' = \hat{f}(\hat{y}, t)$$

En notant de plus

$$\varepsilon_h' = f_{\varepsilon_h}(\epsilon_h, t)$$

pour le problème (2.9), on a

$$\varepsilon_h = P_h - y \quad \Rightarrow \quad \varepsilon_h' = P_h' - f(y)$$

Comme $P_h(t) = \hat{y}(t)$, cela s'écrit

$$f_{\varepsilon_h}(\epsilon_h, t) = \hat{f}(\hat{y}, t) - f(y)$$

Les différentielles élémentaires vérifient alors

$$F_{\varepsilon_h}(u)(\varepsilon_h(t)) = \hat{F}_h(u)(\hat{y}(t)) - F(u)(y(t))$$

et on peut ainsi utiliser la preuve du théorème 2. ■

On résume ici les résultats obtenus avec une méthode d'ordre p, en précisant de plus le coût de chaque algorithme, l'unité choisie étant le nombre d'évaluations de la fonction f pour calculer un pas (une intégration nécessite un seul appel à f à chaque pas).

estimateur	coût	ordre relatif
RS	3	1
ZD	2	p
SC	2	p

2.2 Méthodes multipas

2.2.1 Méthodes BDF

Le principe des méthodes *multipas* repose sur l'idée suivante. Pour calculer y_{n+1} à l'instant t_{n+1}, on n'utilise plus seulement la valeur précédente y_n (méthodes à un pas), mais *plusieurs* valeurs parmi celles calculées auparavant. Si on choisit d'exploiter les k dernières valeurs obtenues, cela s'écrit de façon générale

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(y_{n+j}) \tag{2.11}$$

Comme on veut déterminer y_{n+k} à partir de $y_{n+k-1}, \ldots, y_n$ donnés, on suppose $\underline{\alpha}_k \neq 0$. Si $\underline{\beta}_k \neq 0$, la méthode est *implicite*. Dans le cas contraire elle est *explicite*.

A chaque méthode on associe les deux polynômes suivants, appelés respectivement *premier* et *deuxième polynômes caractéristiques*

$$
\begin{aligned}
\rho(z) &= \underline{\alpha}_k z^k + \underline{\alpha}_{k-1} z^{k-1} + \cdots + \underline{\alpha}_0 \\
\sigma(z) &= \underline{\beta}_k z^k + \underline{\beta}_{k-1} z^{k-1} + \cdots + \underline{\beta}_0
\end{aligned}
$$

Les méthodes qui nous intéressent vérifient $\underline{\beta}_k = 1$ et $\underline{\beta}_j = 0,\ j = 0, \ldots, k-1$. Elles sont donc de la forme

$$
\underline{\alpha}_k y_{n+k} + \underline{\alpha}_{k-1} y_{n+k-1} + \cdots + \underline{\alpha}_0 y_n = h f(y_{n+k}) \tag{2.12}
$$

Ces méthodes, dites BDF-k (pour *Backward Differentiation Formulas* à k pas), sont toujours implicites.

Comme pour les méthodes de Runge-Kutta, on rappelle quelques notions classiques qui seront utiles pour la suite. L'erreur globale et l'erreur locale répondent aux mêmes définitions. L'erreur de troncature aussi mais sa forme est liée à la méthode utilisée.

Définition 6 *L'*erreur de troncature *pour la méthode multipas (2.11) est*

$$
\varepsilon_{n+k} = \sum_{j=0}^{k} \underline{\alpha}_j y(t_{n+j}) - h \sum_{j=0}^{k} \underline{\beta}_j f(y(t_{n+j}))
$$

Cette équation est obtenue en remplaçant dans (2.11) la solution numérique y_n par la valeur exacte $y(t_n)$ et on a la même interprétation que dans le cas des méthodes à un pas.

Définition 7 *La méthode (2.11) est dite* d'ordre p *quand $\varepsilon_n = \mathcal{O}(h^{p+1})$.*

Proposition 1 *Les méthodes BDF-k sont d'ordre k.*

Définition 8 *La méthode (2.11) est dite* convergente *quand pour tout problème $y' = f(y)$ suffisamment régulier, on a :*

$$
\lim_{h \to 0} \|y_{n+k} - y(t_{n+k})\|_\infty = 0\,,
$$

dès que les valeurs initiales vérifient pour tout $i = 0, \ldots, k$:

$$
\lim_{h \to 0} \|y_i - y(t_i)\|_\infty = 0.
$$

Définition 9 *La méthode (2.11) est dite* stable *si son premier polynôme caratéristique ρ est tel que toutes ses racines sont de module ≤ 1 et les racines de module 1 sont simples. Elle est* fortement stable *si toutes les racines de ρ, sauf 1, sont de module < 1. Les racines de module 1 sont dites* essentielles. *1 est la* racine principale.

Proposition 2 *Les méthodes BDF-k sont convergentes d'ordre k mais ne sont fortement stables que pour $k < 7$.*

Il existe une autre façon de présenter la notion de stabilité [13, 11]: c'est une approche matricielle qui va permettre dans la partie suivante d'obtenir le développement asymptotique de l'erreur globale pour les méthodes multipas.

Soit S la matrice donnée par

$$S = \begin{pmatrix} -\underline{\alpha}_{k-1} & -\underline{\alpha}_{k-2} & \cdots & \cdots & -\underline{\alpha}_0 \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}$$

où les $\underline{\alpha}_i$ sont les coefficients de la méthode (2.11). S est donc une matrice compagnon de taille k dont les valeurs propres sont les racines du premier polynôme caractéristique de (2.11). Cette méthode est alors dite *stable* si S^n est bornée uniformément pour tout $n \geq 0$ et *fortement stable* si

$$E := \lim_{n \to \infty} S^n$$

existe. Dans ce cas, 1 est l'unique valeur propre de S de module 1 et le rayon spectral de $S - E$ vérifie $\rho(S - E) < 1$.

2.2.2 Développement asymptotique de l'erreur globale

L'idée développée par Hairer et al. [13, 11] consiste à étendre aux méthodes multipas les résultats obtenus pour les méthodes à un pas. Pour cela il faut voir une méthode multipas comme une méthode à un pas dans un espace de dimension supérieure.

Pour la méthode (2.11), cela donne

$$u_{n+1} = S u_n + h\Phi(t_n, u_n, h)$$

avec $u_n = (y_{n+k-1}, \ldots, y_n)^T \in \mathbb{R}^k$ et

$$\Phi(t_n, u_n, h) = e_1.\underline{\beta}_k f(y_{n+k}) \,, \quad e_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^k$$

Le problème est maintenant de savoir de quelle EDO les u_n sont la solution numérique. On introduit donc une fonction u considérée comme *exacte* au sens où u_n est une approximation de $u(t_n, h)$ pour tout $n \geq 0$ obtenue par résolution numérique de $u' = F(u)$. A priori F n'est pas facile à expliciter mais pour u, un choix assez naturel serait

$$u(t, h) = (y(t + (k-1)h), \ldots, y(t + h), y(t))^T$$

En étendant la définition de Φ à $\mathbb{R} \times \mathbb{R}^k \times \mathbb{R}$ et comme f a été supposée suffisamment différentiable, on peut définir la matrice G suivante

$$G(t) = \frac{\partial \Phi}{\partial u}\,(t, u(t,0), 0))$$

Si on applique cette définition à la méthode BDF-k (2.12), cela donne

$$\begin{aligned} G(t) &= e_1 . \underline{\beta}_k \left(-\underline{\alpha}_{k-1}, \ldots, -\underline{\alpha}_0\right) . f_y(t,y) \\ &= \begin{pmatrix} -\underline{\beta}_k \underline{\alpha}_{k-1} & \cdots & -\underline{\beta}_k \underline{\alpha}_0 \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} . f_y(t,y) \end{aligned}$$

On peut maintenant énoncer le théorème donnant le développement asymptotique de l'erreur globale dans le cas des méthodes multipas *fortement stables* [10, 11]. Ce résultat s'appliquera donc aussi aux BDF-k.

Théorème 4 *Si l'on intègre (2.1) avec une méthode multipas fortement stable, d'ordre $p \geq 1$ et si l'erreur de troncature admet le développement asymptotique (dans $\mathbb{R}^k$)*

$$\begin{aligned} \varepsilon_0 &= \gamma_0 + \gamma_1 h + \ldots + \gamma_N h^N + \mathcal{O}(h^{N+1}) \\ \varepsilon_{n+1} &= \left(d_0(t_n) + \delta_n^0\right) + \ldots + \left(d_{N+1}(t_n) + \delta_n^{N+1}\right) h^{N+1} + \mathcal{O}(h^{N+2}) \end{aligned}$$

pour $nh \geq const$ et avec $d_j(t)$ régulière et $\delta_n^j = \mathcal{O}(\rho^n)$ pour tout $\rho \in]\,\rho(S-E), 1\,[$, alors l'erreur globale possède un développement asymptotique de la forme

$$u_n - u(t_n) = \left(a_p(t_n) + \alpha_n^p\right) h^p + \ldots + \left(a_N(t_n) + \alpha_n^N\right) h^N + \mathcal{O}(h^{N+1}) \tag{2.13}$$

où $\alpha_n^j = \mathcal{O}(\rho^n)$ pour tout $\rho \in]\,\rho(S-E), 1\,[$ et où les a_j sont donnés par

$$a_j'(t) = E\,G(t)\,a_j(t) + \Psi_j(t) \tag{2.14}$$

où Ψ_j est un terme inhomogène fonction du temps et où $E\,G(t)\,\mathbb{1} = \mathbb{1}\,f_y(t,y)$, $\mathbb{1} = (1, \ldots, 1)^T$. Pour $j = p$, on a

$$\begin{aligned} a_p'(t) &= E\,G(t)\,a_p(t) - E\,\varepsilon_{n+1}(t) - (I - S + E)^{-1}\,\varepsilon_p(t) \\ a_p(0) &= -E\,\gamma_p - E \sum_{j=0}^{\infty} \delta_j^p + (I - S + E)^{-1}\,\varepsilon_p(0) \end{aligned}$$

Le cas plus général des méthodes stables qui ne sont pas fortement stables est traité dans un théorème, dû à Gragg [10, 11], qu'on ne donne pas ici puiqu'on se restreint à l'étude des BDF-k.

2.2.3 Possibilités d'estimation de l'erreur globale

On considère une méthode de type (2.12) qu'on suppose être fortement stable. On peut alors appliquer le théorème 4.

• Estimateur RS

On a $\alpha_n^p = \mathcal{O}(\alpha^n)$ pour tout $\alpha \in]\,\rho(S-E), 1\,[$ donc α_n^p tend vers zéro quand $n \to \infty$ plus vite que toute puissance de α. Il existe ainsi $n_0 \geq 0$ tel qu'à tout instant $t_n \geq t_{n_0}$ cette perturbation soit $\ll a_p(t_n)$. Le développement (2.13) s'écrit alors

$$u_n - u(t_n) = a_p(t_n) + \mathcal{O}(h^{p+1})$$

On a la même relation pour $y_n - y(t_n)$ par définition de u_n et $u(t_n)$. Par conséquent, peu après l'instant initial, RS est d'ordre relatif 1.

• Estimateurs ZD et SC

Il n'existe pas à notre connaissance de résultat donnant, comme pour les méthodes à un pas, l'ordre relatif de ces estimateurs. L'équation variationnelle (2.14) est cette fois un système d'EDO de dimension k. En ce qui concerne l'algorithme ZD, une remarque de Frank [7] indique qu'il n'est valable que pour les méthodes fortement stables.

Chapitre 3

Estimation de l'erreur globale pour les EAD semi-explicites d'indice 1

On considère le système (1.1) en prenant $y(0) = y_0$ et $z(0) = z_0$ comme conditions initiales consistantes et f et g suffisamment différentiables. On rappelle que l'indice 1 est obtenu en supposant que $(g_z)^{-1}$ existe et est borné dans un voisinage de la solution. On a alors $z = G(y)$ d'après le théorème des fonctions implicites et le problème (1.1) équivaut dans ce cas à l'EDO

$$y' = f(y, G(y)) \tag{3.1}$$

3.1 Extension des méthodes IRK et BDF aux EAD

3.1.1 Le choix des méthodes implicites

On remarque que la notation $Y'_{ni} = f(Y_{ni})$ de la méthode (2.2) ressemble à l'équation de départ $y' = f(y)$ alors que les relations (2.2a) et (2.2b) dépendent des coefficients de la méthode et du pas d'intégration h, mais ne dépendent pas de la forme de l'équation différentielle à intégrer. L'idée proposée dans [12] pour étendre ces méthodes à l'EAD implicite (1.2) est basée sur cette observation. On définit par conséquent y_{n+1} comme la solution du système (2.2a), (2.2b) et on adapte $Y'_{ni} = f(Y_{ni})$ à la forme du problème considéré

$$F(Y_{ni}, Y'_{ni}) = 0 \tag{3.2}$$

Cependant les méthodes explicites ne permettent pas toujours de traiter les EAD implicites. Considérons par exemple l'équation différentielle algébrique

$$\begin{aligned} y' &= z \\ 0 &= y \end{aligned} \tag{3.3}$$

La méthode (2.2a-2.2b-3.2) donne alors

$$\begin{aligned} Y'_{ni} &= Z_{ni} \\ 0 &= Y_{ni} \end{aligned}$$

et les Z_{ni} sont déterminés par (2.2b), ce qui n'est possible que si la matrice $\underset{\sim}{A} = (\underset{\sim}{a}_{ij})$ est inversible. Or cette condition n'est jamais vérifiée par les méthodes de Runge-Kutta explicites, alors qu'elle peut l'être si la méthode est implicite.

Considérons par exemple la méthode d'Euler. Ses deux versions, explicite et implicite, permettent de comparer dans le cas du problème (3.3) les deux types de méthodes.

• Euler explicite : $y_{n+1} = y_n + hf(y_n)$

$$\begin{aligned} y_{n+1} &= y_n + hz_n \\ 0 &= y_n \end{aligned}$$

ici $s = 1$, $\underline{b}_1 = 1$, $\underline{A} = (\underline{a}_{11}) = 0$ et il est alors impossible de calculer z_{n+1}.

• Euler implicite : $y_{n+1} = y_n + hf(y_{n+1})$

$$\begin{aligned} y_{n+1} &= y_n + hz_{n+1} \\ 0 &= y_{n+1} \end{aligned}$$

On a toujours $s = 1$ et $\underline{b}_1 = 1$ mais la matrice $\underline{A}$ est cette fois inversible ($\underline{A} = I_1$) et le schéma numérique donne $y_{n+1} = z_{n+1} = 0$.

Cela illustre l'avantage des IRK sur les ERK lorsqu'on cherche à résoudre numériquement des systèmes algébro-différentiels. Les méthodes à un pas que l'on considérera à partir de la section suivante seront donc toujours des IRK.

Il est cependant possible d'appliquer les ERK aux EAD semi-explicites d'indice 1 semi-explicites en utilisant une approche différente, dite *semi-explicite* [12]. En effet, la partie différentielle étant distincte de la partie algébrique, on applique à (1.1) une ERK de type (2.2). Les valeurs des variables algébriques sont ensuite obtenues en résolvant la contrainte algébrique avec la méthode de Newton. D'où le schéma

$$y_{n+1} = y_n + h \sum_{i=1}^{s} \underline{b}_i Y'_{ni} \tag{3.4a}$$

$$Y_{ni} = y_n + h \sum_{j=1}^{i-1} \underline{a}_{ij} Y'_{nj}, \quad i = 1, \dots, s \tag{3.4b}$$

$$Y'_{ni} = f(Y_{ni}, Z_{ni}), \quad i = 1, \dots, s \tag{3.4c}$$

$$0 = g(Y_{ni}, Z_{ni}) \tag{3.4d}$$

et z_{n+1} est donné par

$$0 = g(y_{n+1}, z_{n+1}) \tag{3.4e}$$

En partant de $Y_{n1} = y_n$, la résolution de (3.4d) permet d'en déduire Z_{n1}. L'équation (3.4c) donne Y'_{n1} et en reportant dans (3.4b), on obtient Y_{n2} *explicitement.* On peut alors en déduire Z_{n2} et ainsi de suite. Ce schéma s'applique aussi à notre exemple bien que dans ce cas la contrainte ne dépende pas de z : on part toujours de $Y_{n1} = y_n$. En insérant (3.4a) dans (3.4d) avec $i = 2$, on obtient Z_{n1}. L'équation (3.4c) donne alors Y_{n2} etc. On peut continuer ainsi jusqu'à Z_{ns} et y_{n+1} mais z_{n+1} ne peut pas être calculé. Hairer et al. suggèrent de prendre dans ce cas $c_s = 1$ de façon à avoir $z_{n+1} = Z_{ns}$ comme approximation de $z(t_{n+1})$.

3.1.2 Deux approches pour les IRK

La résolution de (1.1) peut se faire de deux façons dans le cas des méthodes IRK. La première consiste à *résoudre l'EDO équivalente* (3.1) avec la méthode (2.2) pour obtenir y_n. La variable algébrique est déduite de y_n par $z_n = G(y_n)$. Cela équivaut à

$$y_{n+1} = y_n + h \sum_{i=1}^{s} \underline{\mathrm{b}}_i Y'_{ni} \tag{3.5a}$$

$$Y_{ni} = y_n + h \sum_{j=1}^{s} \underline{\mathrm{a}}_{ij} Y'_{nj}, \quad i = 1, \dots, s \tag{3.5b}$$

$$Y'_{ni} = f(Y_{ni}, Z_{ni}), \quad i = 1, \dots, s \tag{3.5c}$$

$$0 = g(Y_{ni}, Z_{ni}) \tag{3.5d}$$

et z_{n+1} est donné par

$$0 = g(y_{n+1}, z_{n+1}) \tag{3.5e}$$

Cette approche est indifférente à la nature de la méthode de Runge-Kutta utilisée et l'emploi d'une méthode explicite est tout à fait possible, notamment si le système à résoudre est non raide. On retrouve alors (3.4). De plus elle a l'avantage d'être équivalente à la résolution de (3.1) par la même méthode. En revanche, cette approche ne permet pas de traiter les EAD implicites, pour lesquelles les variables différentielles et algébriques ne sont pas nécessairement dissociées.

Avec l'*approche directe*, cela devient possible. On considère en effet les deux variables de la même façon, ce qui ne suppose pas une forme séparée de l'équation de départ. Le schéma obtenu ne correspond plus à la méthode (2.2) : on conserve (3.5a-d) mais l'équation (3.5e) est remplacée par deux relations donnant z_{n+1}. On obtient le schéma suivant, identique en y et en z

$$y_{n+1} = y_n + h \sum_{i=1}^{s} \underline{\mathrm{b}}_i Y'_{ni} \tag{3.6a}$$

$$Y_{ni} = y_n + h \sum_{j=1}^{s} \underline{\mathrm{a}}_{ij} Y'_{nj}, \quad i = 1, \dots, s \tag{3.6b}$$

$$Y'_{ni} = f(Y_{ni}, Z_{ni}), \quad i = 1, \dots, s \tag{3.6c}$$

$$0 = g(Y_{ni}, Z_{ni}) \tag{3.6d}$$

$$z_{n+1} = z_n + h \sum_{i=1}^{s} \underline{\mathrm{b}}_i Z'_{ni} \tag{3.6e}$$

$$Z_{ni} = z_n + h \sum_{j=1}^{s} \underline{\mathrm{a}}_{ij} Z'_{nj}, \quad i = 1, \dots, s \tag{3.6f}$$

La relation (3.6f) définit Z'_{ni} de façon unique si la matrice $\underset{\sim}{A} = (\underset{\sim}{a}_{ij})$ des coefficients de la méthode numérique est non singulière. On considère donc désormais de telles méthodes.

Cette approche, en plus de la possibilité de traiter les problèmes implicites, présente l'intérêt suivant. Si l'on applique une méthode de Runge-Kutta au problème à perturbation singulière

$$\begin{aligned} y' &= f(y,z) \\ \epsilon z' &= g(y,z), \end{aligned} \tag{3.7}$$

la solution numérique obtenue (qui dépend de ϵ) tend, lorsque $\epsilon \to 0$, vers la solution numérique résultant de l'approche directe avec la même méthode. On reviendra sur ce point au dernier chapitre.

Cette approche présente pourtant un inconvénient majeur. En effet, en remplaçant (3.5e) par (3.6e) et (3.6f), on n'a plus $z_n = G(y_n)$ et appliquer le schéma (3.6) ne revient plus à résoudre l'EDO équivalente avec la même méthode.

Cependant pour certaines méthodes de Runge-Kutta dites *stiffly accurate* (SAIRK pour *Stiffly Accurate Implicit Runge-Kutta*) les deux approches sont équivalentes. On a en effet la définition suivante.

Définition 10 *Une méthode de Runge-Kutta définie par* $\underset{\sim}{b}^T = (\underset{\sim}{b}_1, \ldots, \underset{\sim}{b}_s)$ *et* $\underset{\sim}{A} = (\underset{\sim}{a}_{ij})$ *est une* méthode SAIRK *si ces coefficients vérifient*

$$\forall\, i = 1, \ldots, s \quad \underset{\sim}{b}_i = \underset{\sim}{a}_{si}$$

Ceci équivaut à $y_{n+1} = Y_{ns}$ et $z_{n+1} = Z_{ns}$. Puisque $Z_{ns} = G(Y_{ns})$, on a $z_{n+1} = G(y_{n+1})$ et la contrainte (3.5e) est vérifiée. Une propriété résume la situation.

Propriété 1 *Pour les méthodes SAIRK, les schémas (3.5) et (3.6) sont équivalents (i. e. les solutions numériques obtenues sont les mêmes dans les deux cas).*

L'intérêt de ces méthodes vient du fait qu'elles permettent d'une part de traiter des systèmes algébro-différentiels implicites tout en conservant les propriétés éventuelles relatives à la résolution numérique de l'EDO équivalente. D'autre part, on garantit que la solution numérique calculée reste sur la variété des contraintes.

3.1.3 Méthodes BDF

L'approche la plus simple pour les EAD d'indice 1 sous forme semi-explicite est présentée par Brenan et al. dans [5]. Elle consiste à appliquer une méthode multipas à la partie différentielle de (1.1) et à résoudre l'équation algébrique avec la méthode de Newton. Le schéma est alors

$$\sum_{j=0}^{k} \underset{\sim}{\alpha}_j y_{n+j} = h \sum_{j=0}^{k} \underset{\sim}{\beta}_j f(y_{n+k}, z_{n+k}) \tag{3.8a}$$

$$0 = g(y_{n+k}, z_{n+k}) \tag{3.8b}$$

Cela revient encore à résoudre l'EDO équivalente (3.1). Cette façon de faire garantit en particulier que les méthodes appliquées ont les mêmes propriétés de stabilité et de convergence que si elles l'étaient à des EDO non raides.

En pratique, on travaille plutôt sur des systèmes sous forme implicite de type (1.2). C'est notamment le cas pour DASSL. Le schéma numérique est obtenu en discrétisant la dérivée de Y

$$F\left(Y_{n+k}, \frac{\rho Y_{n+k}}{h}\right) = 0 \tag{3.9}$$

où ρY_{n+k} est la dérivée du polynôme interpolant $Y_n, \ldots, Y_{n+k}$. On a alors

$$\rho Y_{n+k} = \sum_{j=0}^{k} \underline{\alpha}_j Y_{n+j}$$

Pour une fonction F quelconque, un tel schéma n'équivaut pas à résoudre l'EDO sous-jacente. Cependant lorsqu'il s'agit de systèmes semi-explicites, F peut sécrire

$$F\left(\begin{pmatrix} y' \\ z' \end{pmatrix}, \begin{pmatrix} y \\ z \end{pmatrix}\right) = \begin{pmatrix} y' - f(y,z) \\ -g(y,z) \end{pmatrix}$$

avec $Y = (y,z)^T$. Dans ce cas, l'équation (3.9) devient

$$\begin{aligned} \frac{\rho y_{n+k}}{h} - f(y_{n+k}, z_{n+k}) &= 0 \\ g(y_{n+k}, z_{n+k}) &= 0 \end{aligned}$$

où $Y_{n+k} = (yn+k, z_{n+k})^T$. Si la méthode est une BDF-k, on retrouve bien (3.8) moyennant une normalisation par $\underline{\beta}_k \neq 0$.

Une autre façon de faire existe, due à Gear [9]. Cette approche, appelée *approche directe* dans [14], consiste à appliquer une BDF-k au problème à pertubation singulière (3.7). Cela donne

$$\begin{aligned} \sum_{j=0}^{k} \underline{\alpha}_j y_{n+j} &= h \sum_{j=0}^{k} \underline{\beta}_j f(y_{n+k}, z_{n+k}) \\ \epsilon \sum_{j=0}^{k} \underline{\alpha}_j y_{n+j} &= h \sum_{j=0}^{k} \underline{\beta}_j g(y_{n+k}, z_{n+k}) \end{aligned}$$

En posant $\epsilon = 0$, on obtient

$$\begin{aligned} \sum_{j=0}^{k} \underline{\alpha}_j y_{n+j} &= h \sum_{j=0}^{k} \underline{\beta}_j f(y_{n+k}, z_{n+k}) \\ 0 &= h \sum_{j=0}^{k} \underline{\beta}_j g(y_{n+k}, z_{n+k}) \end{aligned}$$

ce qui permet aussi d'appliquer une méthode multipas linéaire aux EAD semi-explicites d'indice 1 et si la méthode est une BDF-k, on retrouve également (3.8).

Les deux approches sont bien sûr équivalentes pour les BDF-k.

3.2 Estimation de l'erreur globale

En vue de l'extension des estimateurs au problème (1.1), on utilise l'approche directe présentée au chapitre 2 pour les IRK (3.6). Pour les BDF fortement stables, on utilise l'approche indirecte (3.8).

Remarque : lorsque l'indice est 1, on a $z = G(y)$ et aussi $Z_{ni} = G(Y_{ni})\ \forall\, n,\ \forall\, i$. D'après le schéma (3.6), on voit donc que le calcul des $\{y_i\}$ est indépendant de celui des $\{z_i\}$ et par conséquent y_n est en fait la solution numérique d'une EDO en y. L'étude qui suit portera donc essentiellement sur la variable algébrique z.

3.2.1 Développement asymptotique de l'erreur globale pour les IRK

Parmi les différentes classes de méthodes IRK, celle qui nous intéresse plus particulièrement regroupe les schémas numériques vérifiant la condition $C(q)$ suivante

$$C(q): \qquad \sum_{j=1}^{s} \underline{a}_{ij}\underline{c}_j^{k-1} = \frac{\underline{c}_i^k}{k} \qquad \text{pour } k = 1, \ldots, q \text{ et tout } i$$

On aura également besoin du concept de *A-stabilité*. Si on applique une méthode de Runge-Kutta à l'EDO scalaire $y' = \lambda y$, on obtient $y_{n+1} = R(h\lambda)\, y_n$ où la *fonction de stabilité* R est donnée par

$$R(w) = 1 + w\underline{b}^T\,(I - w\underline{A})^{-1}\,\mathbb{1}$$

Ici $\underline{A} = (\underline{a}_{ij})$ et $\underline{b}^T = (\underline{b}_1, \ldots, \underline{b}_s)$ définissent les coefficients de la méthode de Runge-Kutta et $\mathbb{1} = (1, \ldots, 1)^T$. La méthode de Runge-Kutta est dite *A-stable* si

$$|R(w)| \leq 1 \quad \text{pour} \quad Re\, w \leq 0$$

Le comportement à h fixé de la solution numérique $y_{n,h}$ dépend de R : elle reproduit le comportement de la solution exacte y pour $\lambda < 0$ si et seulement si $|R(h\lambda)| < 1$. Dans le cas des équations algébro-différentielles, la limite à l'infini de la fonction de stabilité joue un rôle important : d'elle dépend notamment la forme du développement asymptotique de l'erreur globale. Si on suppose que la matrice $\underline{A}^{-1}$ existe, cette limite est donnée par

$$R(\infty) = 1 - \underline{b}^T\underline{A}^{-1}\,\mathbb{1}$$

En particulier, la stabilité à l'infini d'une SAIRK est nulle. En effet, d'après [12], on a le résultat

Propriété 2 *Si $\underline{A}^{-1}$ existe alors*

$$(\underline{b}_i = \underline{a}_{si}, \quad \forall\, i = 1, \ldots, s) \Rightarrow (R(\infty) = 0)$$

On peut maintenant s'intéresser aux développements asymptotiques de l'erreur globale commise en y et en z. La composante en y du problème (1.1) étant la solution numérique de l'EDO équivalente (3.1), on a pour y le développement (2.4)

$$y_n - y(t_n) = h^p a_p(t_n) + \ldots + h^N a_N(t_n) + \mathcal{O}(h^{N+1})$$

Pour la variable algébrique, Hairer et al. donnent dans [12] le résultat suivant.

Théorème 5 *Soit un système de la forme (1.1) qu'on intègre avec un ensemble de conditions initiales consistantes. On suppose de plus que la méthode IRK utilisée est d'ordre p, vérifie C(q) ($p \geq q+1$) et que $\underset{\sim}{A}^{-1}$ existe. On a alors*

$$\begin{array}{lll} a) & SAIRK: & z_n - z(t_n) = h^p b_p(t_n) + \cdots + h^{N-1} b_{N-1}(t_n) + \mathcal{O}(h^N) \\ b) & -1 \leq R(\infty) < 1: & z_n - z(t_n) = h^{q+1}(b_{q+1}(t_n) + \beta_n^{q+1}) + \cdots \\ & & \cdots + h^{N-1}(b_{N-1}(t_n) + \beta_n^{N-1}) + \mathcal{O}(h^N) \end{array}$$

où les perturbations β_n^ksont indépendantes du pas et telles que :

$$\beta_n^k = \begin{cases} 0 \quad \forall n \geq 0 & si \quad R(\infty) = 0 \\ \mathcal{O}(\alpha^n) \quad \forall n, \forall \alpha \in \,]\,|R(\infty)|, 1\,[& si \quad |R(\infty)| < 1 \\ (-1)^{n+1} b_k(0) \quad \forall n \geq 0 & si \quad R(\infty) = -1 \end{cases}$$

$$\begin{array}{lll} c) & R(\infty) = 1: & z_n - z(t_n) = h^q b_q(t_n) + \cdots + h^{N-1} b_{N-1}(t_n) + \mathcal{O}(h^N) \end{array}$$

On voit donc que certaines méthodes (SAIRK notamment) conduisent à des développements asymptotiques de même nature pour y et z, alors que pour d'autres des *perturbations* apparaissent dans le développement relatif à la variable algébrique. Ces perturbations ne dépendent pas du pas d'intégration au sens où la numérotation de la grille d'intégration suffit à les identifer. Si l'on réalise par exemple deux intégrations avec les pas h et $h/2$ et qu'on note $\beta_{n,h}^k$ et $\beta_{n,h/2}^k$ ($k = p, \ldots, N$) les perturbations correspondantes, on a ainsi pour tout k fixé $\beta_{n,h}^k = \beta_{2n,h/2}^k$ et les deux développements contiennent les mêmes termes perturbatifs.

3.2.2 Extension des estimateurs pour les IRK

A partir de ce résultat on étudie les propriétés des trois estimateurs pour la variable algébrique seulement, la variable différentielle étant, on l'a vu, la variable d'une EDO.

• Estimateur de Richardson

Cet estimateur est défini de façon naturelle par

$$E_{RS,n}^{*,y} = \frac{y_n - y_{2n}^*}{1 - 2^{-p}}$$
$$E_{RS,n}^{*,z} = \frac{z_n - z_{2n}^*}{1 - 2^{-p}}$$

Théorème 6 *On suppose qu'on intègre l'EAD (1.1) semi-explicite d'indice 1 avec un ensemble de conditions initiales consistantes et une méthode IRK d'ordre p, vérifiant C(q) avec $p \geq q+1$ et de matrice $\underset{\sim}{A}$ inversible. L'extrapolation de Richardson conduit alors à un estimateur de l'erreur globale sur la variable algébrique d'ordre*

$$\begin{array}{lll} a) & p+1 & si \text{ la méthode est une } SAIRK \\ b) & q+2 & si \quad -1 < R(\infty) < 1 \\ c) & q+1 & si \quad R(\infty) = 1 \\ d) & q+2 & si \quad R(\infty) = -1 \quad et \quad b_{q+1}(0) \ll b_{q+1}(t_n)\ \forall n > 0 \end{array}$$

Dans tous les cas, l'ordre relatif de cet estimateur est égal à 1.

Preuve du théorème 6 :
- Pour a) et c), le développement asymptotique en z n'est pas perturbé (théorème 5) donc c'est immédiat.
- Pour b), la perturbation du premier terme du développement asymptotique vérifie

$$\beta_n^{q+1} = \mathcal{O}(\alpha^n) \quad \forall n, \forall \alpha \in \,]\, |R(\infty)|, 1\, [$$

En reprenant l'argument utilisé pour démontrer l'ordre de RS dans le cas des méthodes BDf pour les EDO, on en déduit que la perturbation du premier terme passe donc dans le deuxième terme et on est ramené aux cas a) et c).
- Pour d), $\beta_n^{q+1} = (-1)^{n+1} b_{q+1}(0)$ donc il suffit que $b_{q+1}(0) \ll b_{q+1}(t_n)$ $\forall n > 0$ pour retrouver les cas précédents.
- Enfin, d'après le théorème 5 et par définition de l'ordre relatif d'un estimateur, celui-ci vaut bien 1 dans tous les cas. ∎

- **Algorithme de Zadunaisky**

On sait que pour la variable différentielle y, appliquer une IRK quelconque à (1.1) équivaut à appliquer cette méthode à l'EDO (3.1). Selon le même principe, on cherche à appliquer ZD à (1.1) de sorte que ce soit équivalent à l'appliquer à l'EDO (3.1). On considère donc le problème perturbé

$$\begin{array}{rcl} \hat{y}' & = & f(\hat{y}, \hat{z}) + d_h(t) \\ 0 & = & g(\hat{y}, \hat{z}) \end{array} \tag{3.10}$$

avec $d_h(t) = P_h'(t) - f\left(P_h(t), G(P_h(t))\right)$ où G est la fonction habituelle issue du théorème des fonctions implicites (g_z inversible $\Rightarrow z = G(y)$) et où le polynôme P_h est de degré m et interpole les $\{y_n\}$, comme pour les EDO.
L'estimateur de Zadunaisky répond alors à la définition suivante :

$$E^{*,y}_{ZD,n} = \hat{y}_n - y_n$$
$$E^{*,z}_{ZD,n} = \hat{z}_n - G(y_n)$$

En pratique, l'alorithme nécessite le calcul de P_h et donc une résolution à chaque instant t_n de $g(y_n, \cdot) = 0$ par la méthode de Newton (pour évaluer $G(P_h(t_n))$). Ce surcoût disparaît évidemment pour toutes les méthodes qui garantissent $g(y_n, z_n) = 0$, notamment les SAIRK.
Comme le problème perturbé (3.10) est équivalent au problème perturbé obtenu en appliquant ZD directement à l'EDO (3.1), on en déduit d'après la théorie de l'estimation de l'erreur globale pour les EDO (lemme 1), que d'une part, $d_h(t)$ a la propriété

Propriété 3

$$d_h^{(k)}(t) = \mathcal{O}(h^{max(0, min(p, m-k))})$$

et que d'autre part (théorème 2), ZD fournit une estimation de l'erreur globale en y d'ordre m (et d'ordre relatif $m - p$). On rappelle qu'en effet les développements asymptotiques coïncident jusqu'à l'ordre m :

$$y_n = y(t_n) + DA_y(t_n)$$
$$\hat{y}_n = \hat{y}(t_n) + DA_{\hat{y}}(t_n)$$

et

$$DA_{\hat{y}}(t_n) = DA_y(t_n) + \mathcal{O}(h^m)$$

On cherche maintenant à étendre ce résultat à la variable algébrique z. La question est donc de savoir jusqu'à quel ordre les développements asymptotiques donnés par

$$\begin{aligned} z_n &= z(t_n) + DA_z(t_n) \\ \hat{z}_n &= \hat{z}(t_n) + DA_{\hat{z}}(t_n) \end{aligned}$$

coïncident, c'est-à-dire à quel ordre $\hat{z}_n - \hat{z}(t_n)$ estime l'erreur globale $z_n - z(t_n)$.

Remarque : dans tous les cas du théorème 5, DA_z se déduit de DA_y (cf. théorème 3.1 dans [12]) et le fait que DA_y et $DA_{\hat{y}}$ coïncident jusqu'à l'ordre m va conduire à une propriété similaire pour DA_z et $DA_{\hat{z}}$.

On peut maintenant énoncer un théorème donnant l'ordre de ZD pour la variable algébrique.

Théorème 7 *Pour le problème (1.1) semi-explicite d'indice 1 qu'on résoud avec un ensemble de conditions initiales consistantes et une IRK d'ordre p satisfaisant $C(q)$ $(p \geq q+1)$ et de matrice de coefficients A inversible, l'algorithme ZD appliqué avec (3.10) a les propriétés suivantes.*

$\hat{z}_n - G(y_n)$ est un estimateur de l'erreur globale sur la variable algébrique d'ordre

$$\begin{array}{lll} a) & min(m, 2p) & \textit{si la méthode est une SAIRK} \\ b) & min(m, p+q+1) & \textit{si} \quad -1 \leq R(\infty) < 1 \\ c) & min(m-1, p+q) & \textit{si} \quad R(\infty) = 1 \end{array}$$

Remarque :
En pratique, on prendra comme pour les EDO $m = 2p$ pour que l'estimateur soit d'ordre maximal. Dans ce cas, puisque $p \geq q+1$ et compte-tenu de l'ordre de l'erreur globale (théorème 5), l'ordre relatif de l'estimateur est le même dans les trois cas a), b) et c).

cas	ordre	ordre relatif
a)	$2p$	p
b)	$p+q+1$	p
c)	$p+q$	p

La preuve du théorème 7 est basée sur la preuve du théorème 3.1 de [12] où l'ordre de convergence de la variable algébrique est déduit de celui de la variable différentielle. On a aussi besoin du lemme suivant (qui utilise le formalisme des arbres de Butcher pour les EAD d'indice 1 [14]).

Lemme 2 *Soit (3.10) le problème perturbé de (1.1) par $d_h(t)$ vérifiant la propriété 3. On a alors, pour deux B-séries associées à un ensemble quelconque a de coefficients*

$$a) \sum_{u \in DAT_y,\, \rho(u) \geq k} h^{\rho(u)} a(u) \hat{F}_h(u)\,(\hat{y}(t), \hat{z}(t)) = \sum_{u \in DAT_y,\, \rho(u) \geq k} h^{\rho(u)} a(u) F(u)\,(y(t), z(t)) + \mathcal{O}(h^{min(m+1, p+k)})$$

$$b) \sum_{v \in DAT_z,\, \rho(v) \geq k} h^{\rho(v)} a(v) \hat{F}_h(v)\,(\hat{y}(t), \hat{z}(t)) = \sum_{v \in DAT_z,\, \rho(v) \geq k} h^{\rho(v)} a(v) F(v)\,(y(t), z(t)) + \mathcal{O}(h^{min(m+1, p+k)})$$

Preuve du lemme 2 :
On note $D_y(t)$ et $D_z(t)$ les différences de B-séries correspondant aux cas a) et b) respectivement. Il faut montrer que

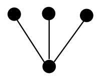
$$D_y(t) = \mathcal{O}(h^{min(m+1,p+k)})$$
$$D_z(t) = \mathcal{O}(h^{min(m+1,p+k)})$$

Considérons la différence de deux différentielles élémentaitres

$$\hat{F}_h(u)\,(\hat{y}(t),\hat{z}(t)) - F(u)\,(y(t),z(t))$$

Le problème (3.10) n'est pas autonome donc pour calculer $\hat{F}_h(u)(\hat{y}(t),\hat{z}(t))$, il faut d'abord se ramener à un problème autonome en ajoutant une variable. On peut alors calculer facilement cette différence avec les arbres de Butcher de la façon suivante.

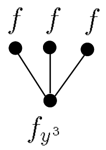
Soit par exemple l'arbre

$u =$ 

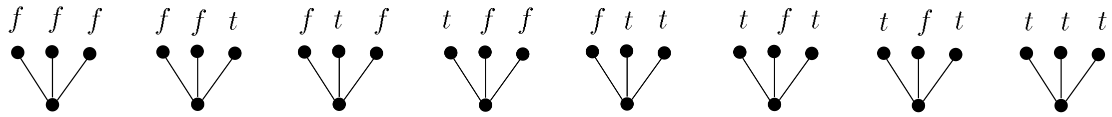
On a

$$F(u)\,(y(t),z(t)) = f_{y^3}fff\,(y(t),z(t))$$

car u s'écrit en fait



Pour $\hat{F}_h(u)\,(\hat{y}(t),\hat{z}(t))$, il faut tenir compte du temps t, c'est-à-dire considérer toutes les façons d'instancier les feuilles par f ou t :



En regroupant les arbres topologiquement équivalents, il vient finalement

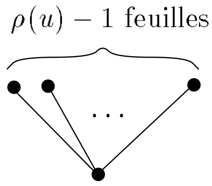
$$\hat{F}_h(u)\,(\hat{y}(t),\hat{z}(t)) = \left[f_{y^3}fff + 3f_{y^2t}ff + 3f_{yt^2}f + f_{t^3}\right](\hat{y}(t),\hat{z}(t))$$

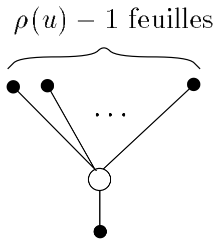
Ainsi,

$$\hat{F}_h(u)\,(\hat{y}(t),\hat{z}(t)) - F(u)\,(y(t),z(t)) = -3f_{y^2t}ff - 3f_{yt^2}f - f_{t^3}$$

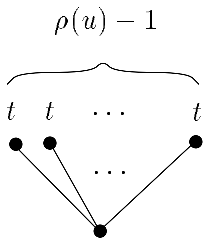
L'ordre en h de cette expression est imposé par le terme contenant la plus haute dérivée du défaut $d_h(t)$ d'après la propriété 3. Ici, c'est $f_{t^3} = d_h^{(3)}(t)$.

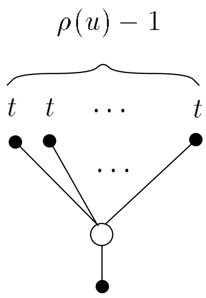
Considérons maintenant la somme $D_y(t)$. Son ordre en h est imposé, pour $\rho(u)$ donné, par les différences $\hat{F}_h(u)\,(\hat{y}(t),\hat{z}(t)) - F(u)\,(y(t),z(t))$ contenant la plus haute dérivée de $d_h(t)$, c'est-à-dire les dérivées en temps d'ordre le plus élevé. Ce sont donc les termes associés aux arbres du type

$u =$ 

et $u =$ 

En effet, pour ces arbres, $\hat{F}_h(u)\,(\hat{y}(t),\hat{z}(t)) - F(u)\,(y(t),z(t))$ fait intervenir



et


c'est-à-dire

$$f_{t^{\rho(u)-1}} \quad \text{et} \quad f_z(-g_z^{-1})g_{t^{\rho(u)-1}}$$

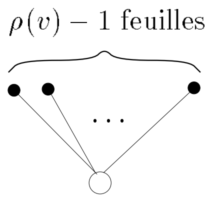
Finalement, si $m+1 \geq \rho(u)$, on a pour $u \in DAT_y$,

$$\hat{F}_h(u)\,(\hat{y}(t),\hat{z}(t)) - F(u)\,(y(t),z(t)) \sim d^{(\rho(u)-1)}(t) = \mathcal{O}(h^{min(p,m-\rho(u)+1)})$$

d'après la propriété 3. Et comme $\rho(u) \geq k$, on a enfin

$$D_y(t) = \mathcal{O}(h^{min(p+k,m+1)})$$

On raisonne de la même façon pour $D_z(t)$, et en tenant compte du fait que la partie algébrique du problème (3.10) n'est pas perturbée, on peut montrer que pour $\rho(v)$ donné l'ordre en h est imposé par les arbres du type

$v =$ 

Ainsi, si $m+1 \geq \rho(v)$, on a pour $v \in DAT_z$

$$\hat{F}_h(v)\,(\hat{y}(t), \hat{z}(t)) - F(v)\,(y(t), z(t)) \sim d^{(\rho(v)-1)}(t) = \mathcal{O}(h^{min(p, m-\rho(u)+1)})$$

et

$$D_z(t) = \mathcal{O}(h^{min(p+k, m+1)})$$

$\blacksquare$

Preuve du théorème 7:
$*$ Pour le cas a), c'est immédiat. En effet, comme $g(y_n, z_n) = 0$ pour tout n, on a

$$\begin{aligned} z_n &= G(y_n) \\ &= G(y(t_n) + DA_y(t_n)) \end{aligned}$$

Par un développement de Taylor à l'ordre 1 et comme $z(t) = G(y(t))$ pour tout t, on obtient

$$z_n = z(t_n) + DA_z(t_n)$$

avec

$$DA_z(t_n) = h^p G_y(y(t_n)) \cdot a_p(t_n) + \ldots + h^{2p-1} \cdot G_y(y(t_n)) a_{2p-1}(t_n) + \mathcal{O}(h^{2p})$$

Les coefficients du développement asymptotique du problème perturbé sont obtenus de la même façon à partir de $DA_{\hat{y}}(t_n)$. Pour k donné, la différence des coefficients des développements en z vaut

$$G_y(y(t_n)) \cdot a_k(t_n) - G_y(y(t_n)) \cdot \hat{a}_k(t_n) = G_y(y(t_n)) \cdot (a_k(t_n) - \hat{a}_k(t_n))$$

et dépend donc linéairement de la différence (pour le même k) des coefficients des développements en y.
Comme y est la variable d'une EDO, la preuve du théorème 2 permet de conclure. $\blacksquare$

$*$ Pour les cas b) et c), on suit la preuve du théorème 3.1 de [12]. On a

$$z_{n+1} = R(\infty) z_n + \underline{b}^T \underline{A}^{-1} Z_n$$

avec $Z_n = (Z_{n1}, \ldots, Z_{ns})^T$. Comme la méthode est d'ordre p et vérifie $C(q)$,

$$z(t_{n+1}) = z(t_n) + h \sum_{i=1}^{s} \underline{b}_i z'(t_n + \underline{c}_i h) + taylor(t_n)$$

$$z(t_n + \underline{c}_i h) = z(t_n) + h \sum_{j=1}^{s} \underline{a}_{ij} z'(t_n + \underline{c}_j h) + TAYLOR_i(t_n)$$

où $taylor(t_n)$ et $TAYLOR_i(t_n)$ sont des développements de Taylor de valuation $p+1$ et $q+1$ respectivement. La solution exacte z vérifie alors

$$z(t_{n+1}) = R(\infty) z(t_n) + \underline{b}^T \underline{A}^{-1} \hat{Z}_n + taylor(t_n) - \underline{b}^T \underline{A}^{-1} TAYLOR(t_n)$$

avec $\hat{Z}_n = (z(t_n + \underset{\sim}{c}_1 h), \ldots, z(t_n + \underset{\sim}{c}_s h))^T$ et $TAYLOR(t_n) = (TAYLOR_1(t_n), \ldots, TAYLOR_s(t_n))^T$.
On vérifie facilement que la série

$$B_z(y(t_n), z(t_n)) = -taylor(t_n) + \underset{\sim}{b}^T \underset{\sim}{A}^{-1} TAYLOR(t_n)$$

est une B-série d'arbres de DAT_z, de valuation $q+1$ (car $p \geq q+1$). On a donc, en posant $\Delta z_n = z_n - z(t_n)$ et $\Delta Z_n = Z_n - \hat{Z}_n$,

$$\Delta z_{n+1} = R(\infty)\Delta(z_n) + \underset{\sim}{b}^T \underset{\sim}{A}^{-1} \Delta Z_n + B_z(y(t_n), z(t_n))$$

ce qui s'écrit aussi

$$\boxed{\Delta z_{n+1} = \rho \Delta z_n + \delta_n}$$

avec $\rho = R(\infty)$ et $\delta_n = \underset{\sim}{b}^T \underset{\sim}{A}^{-1} \Delta Z_n + B_z(y(t_n), z(t_n))$.

Comme on cherche l'ordre en h de $\delta_n - \hat{\delta}_n$, il faut étudier séparément les différences $\Delta Z_n - \Delta\hat{Z}_n$ et $B_z(y(t_n), z(t_n)) - B_z(\hat{y}(t_n), \hat{z}(t_n))$. D'après le lemme 2, cette dernière expression vaut

$$B_z(y(t_n), z(t_n)) = B_z(\hat{y}(t_n), \hat{z}(t_n) + \mathcal{O}(h^{min(m+1, p+q+1)}) \tag{3.11}$$

Pour $\Delta Z_n - \Delta\hat{Z}_n$, on procède de la façon suivante. La i-ème composante de ΔZ_n vérifie

$$\Delta Z_{n|_i} = Z_{ni} - z(t_n + \underset{\sim}{c}_i h)$$

Afin de dérminer l'ordre en h de $\Delta Z_{n|_i} - \Delta\hat{Z}_{n|_i}$, on part de la variable différentielle y pour laquelle on étudie $\Delta Y_{n|_i} - \Delta\hat{Y}_{n|_i}$. On a

$$Y_{ni} - y(t_n + \underset{\sim}{c}_i h) = (Y_{ni} - y_n) - (y(t_n + \underset{\sim}{c}_i h) - y(t_n)) + (y_n - y(t_n))$$

D'après la théorie des EDO, on sait que :
∗ $y_n - y(t_n)$ admet un développement asymptotique de valuation p.
∗ $(Y_{ni} - y_n) - (y(t_n + \underset{\sim}{c}_i h) - y(t_n))$ est une B-série par rapport à $y(t_n)$ de valuation $q+1$ (car la méthode vérifie $C(q)$).
Ainsi, toujours d'après la théorie de l'erreur globale pour les EDO,

$$y_n - y(t_n) = \hat{y}_n - \hat{y}(t_n) + \mathcal{O}(h^{min(m, 2p)})$$

et d'après le lemme 2,

$$[(Y_{ni} - y_n) - (y(t_n + \underset{\sim}{c}_i h) - y(t_n))] - \left[\left(\hat{Y}_{ni} - \hat{y}_n\right) - (\hat{y}(t_n + \underset{\sim}{c}_i h) - \hat{y}(t_n))\right] = \mathcal{O}(h^{min(m+1, p+q+1)})$$

Finalement,

$$[Y_{ni} - y(t_n + \underset{\sim}{c}_i h)] - \left[\hat{Y}_{ni} - \hat{y}(t_n + \underset{\sim}{c}_i h)\right] = \mathcal{O}(h^{min(m, p+q+1)})$$

On en déduit, en utilisant $Z_{ni} = G(Y_{ni})$ et $Z(t) = G(y(t))$,

$$[Z_{ni} - z(t_n + \underset{\sim}{c}_i h)] - \left[\hat{Z}_{ni} - \hat{z}(t_n + \underset{\sim}{c}_i h)\right] = \mathcal{O}(h^{min(m, p+q+1)})$$

et donc

$$\Delta Z_n = \Delta \hat{Z}_n + \mathcal{O}(h^{min(m,p+q+1)}) \tag{3.12}$$

De (3.11) et (3.12), on déduit donc

$$\boxed{\delta_n = \hat{\delta}_n + \mathcal{O}(h^{min(m,p+q+1)})}$$

On a enfin, par récurrence et comme les conditions initiales ont été prises consistantes,

$$\Delta z_n = \rho^n \Delta z_0 + \sum_{i=1}^{n} \rho^{n-i} \delta_i \,, \quad \Delta z_0 = 0$$

Pour conclure, les cas b) et c) sont traités séparément.

b) $-1 \leq \rho < 1$
$*$ Si $|\rho| < 1$, pour tout i il existe $C_i \in \mathbb{R}$ tel que $|\delta_i - \hat{\delta}_i| \leq C_i \, h^{min(m,p+q+1)}$ et on a

$$\begin{aligned} |\Delta z_n - \Delta \hat{z}_n| &\leq \sum_{i=1}^{n} |\rho|^{n-i} \left| \delta_i - \hat{\delta}_i \right| \\ &\leq (max_i C_i) \, \frac{1}{1-|\rho|} h^{min(m,p+q+1)} \end{aligned}$$

donc

$$\Delta z_n - \Delta \hat{z}_n = \mathcal{O}(h^{min(m,p+q+1)})$$

$*$ Si $\rho = -1$, on regroupe les termes selon la parité de $n-i$, $i = 1, \ldots, n$ et le même raisonnement que ci-dessus sur chacune des deux sommes obtenues donne le résultat.

c) $\rho = 1$

$$|\Delta z_n - \Delta \hat{z}_n| \leq n \, (max_i C_i) \, h^{min(m,p+q+1)}$$

Mais comme $n = t_n/h$, on perd un ordre et finalement

$$\Delta z_n - \Delta \hat{z}_n = \mathcal{O}(h^{min(m-1,p+q)})$$

Ces deux dernières expressions donnent les résultats annoncés pour les cas b) et c). ■

Remarque : en pratique, RADAU5 résoud des systèmes de la forme $MY' = F(Y)$ et non (1.1). Les variables différentielle et algébrique ne sont donc pas séparées et on ne peut plus définir ZD par (3.10). Une façon naturelle de construire ZD dans le cas d'EAD implicites consiste alors à perturber le problème (1.1) pour *toutes* les variables :

$$M\hat{Y}' = F(\hat{Y}) + d_h(t)$$

avec $d_h(t) = MP_h'(t) - F(P_h(t))$ et où P_h est un polynôme interpolant les $\{y_n\}$. Cette formulation a été choisie lors de l'implémentation de ZD dans RADAU5. Sur (1.1), cela s'écrit

$$\begin{aligned} \hat{y}' &= f(\hat{y}, \hat{z})) + d_{1,h}(t) \\ 0 &= g(\hat{y}, \hat{z}) + d_{2,h}(t) \end{aligned}$$

avec

$$\begin{aligned} d_{1,h} &= P_{1,h}'(t) - f(P_{1,h}(t), P_{2,h}(t)) \\ d_{2,h}(t) &= -g(P_{1,h}(t), P_{2,h}(t)) \end{aligned}$$

où $P_{1,h}$ et $P_{2,h}$ interpolent les $\{y_n\}$ et les $\{z_n\}$ respectivement.

$*$ Pour une SAIRK, cette formulation de ZD est équivalente à celle utilisée pour établir le théorème 7.
$*$ Si la méthode est une IRK quelconque, cela induit une réduction d'ordre sur la variable algébrique, le défaut $d_{2,h}$ n'étant plus de valuation p mais $q+1 \leq p$. Dans le cas semi-explicite, cette approche est évidemment sans intérêt les variables différentielle et algébrique étant clairement séparées.

- **Technique de correction de défaut**

Pour SC, on a un résultat analogue au théorème 7 en utilisant les mêmes polynômes que pour ZD, à savoir $P_{1,h}$ interpolant $\{y_n\}$ et $P_{2,h} = G(P_{1,h})$. On a donc

$$\begin{array}{rcl} \varepsilon_1' &=& P_{1,h}' - f\,(P_{1,h} - \varepsilon_1,\, G(P_{1,h} - \varepsilon_2) \\ 0 &=& g(P_{1,h} - \varepsilon_1, P_{2,h} - \varepsilon_2) \end{array} \tag{3.13}$$

où ε_1 et ε_2 dépendent bien sûr du pas h (on supprime cet indice pour plus de clarté). L'estimateur SC est alors défini par

$$\begin{aligned} E_{SC,n}^{*,y} &= \varepsilon_{1,n} \\ E_{SC,n}^{*,z} &= \varepsilon_{2,n} \end{aligned}$$

Cela revient à l'appliquer à l'EDO èquivalente donc $\varepsilon_{1,n}$ est une estimation d'ordre relatif p (théorie des EDO).

Théorème 8 *Pour le problème (1.1) semi-explicite d'indice 1 qu'on résoud avec un ensemble de conditions initiales consistantes et une IRK d'ordre p satisfaisant $C(q)$ $(p \geq q+1)$ et de matrice de coefficients $\underline{A}$ inversible, l'algorithme SC appliqué avec (3.13) a les propriétés suivantes.*

$\varepsilon_{2,n}$ est un estimateur de l'erreur globale sur la variable algébrique d'ordre

a)	$min(m, 2p)$	*si la méthode est une SAIRK*
b)	$min(m, p+q+1)$	*si* $-1 \leq R(\infty) < 1$
c)	$min(m-1, p+q)$	*si* $R(\infty) = 1$

preuve du théorème 8 :
L'idée consiste alors, comme pour les EDO, à se ramener à la preuve de ZD. Avec SC construit comme ci-dessus, la variable différentielle ε_1 est la variable d'une EDO et on a le résultat pour ε_1. Pour la variable algébrique ε_2, on utilise la linéarité de la différentielle élémentaire et le fait que pour tout $k \geq 0$, $\varepsilon_2^{(k)}(t) = P_{2,h}^{(k)}(t) - z^{(k)}(t)$ pour se ramener à la preuve de ZD (théorème 7). ∎

Remarque : de même que pour ZD, le traitement d'EAD implicites impose d'écrire SC comme pour une EDO de dimension 2 :

$$\varepsilon_h = P_h - Y$$

$$M\,\varepsilon_h' = MP_h' - F(P_h' - \varepsilon_h)$$

ce qui pour le problème (1.1) s'écrit

$$\begin{aligned} \varepsilon_{1,h}' &= P_{1,h}' - f(P_{1,h} - \varepsilon_{1,h}, G(P_{1,h}) - \varepsilon_{2,h}) \\ 0 &= g(P_{1,h} - \varepsilon_{1,h}, G(P_{1,h}) - \varepsilon_{2,h}) \end{aligned}$$

car $P_{2,h} = G(P_{1,h})$.

∗ Si la méthode est une SAIRK, cette approche est encore équivalente à celle utilisée dans le théorème 8 .
∗ Sinon, la réduction d'ordre est la même qu'avec ZD : pour la variable algébrique, l'ordre relatif est $q+1$ au lieu de p.

3.2.3 Développement asymptotique de l'erreur globale pour les BDF

On utilise l'approche *indirecte*, qui revient, comme on l'a vu dans la partie consacrée à l'intégration numérique des EAD, à résoudre l'EDO équivalente $y' = f(y, G(y))$ car on exige à chaque instant t_n que la contrainte algébrique $g(y_n, z_n) = 0$ soit satisfaite. On suppose également que la méthode utilisée est fortement stable afin de pouvoir appliquer le théorème 4. Avec l'approche choisie ici la variable différentielle apparaît comme la variable d'une EDO. On sait donc déjà que le développement asymptotique de l'erreur globale commise sur cette variable a la forme (2.13). Il reste à étudier ce développement pour la variable algébrique.

Avec l'approche *indirecte* et d'après le théorème 4, on a

$$\begin{aligned} z_n &= G(y_n) \\ &= G\left(y(t_n) + (a_p(t_n) + \alpha_n^p)\, h^p + \ldots\right) \end{aligned}$$

Le développement de Taylor au premier ordre de cette expression donne

$$z_n = G(y(t_n)) + G_y\left(y(t_n)\right).\left((a_p(t_n) + \alpha_n^p)\, h^p\right) + \ldots + \mathcal{O}(h^{2p})$$

Si $N+1 \geq 2p$, on obtient finalement

$$z_n - z(t_n) = (b_p(t_n) + \beta_n^p) + \ldots + \left(b_N(t_n) + \beta_n^N\right) + \mathcal{O}(h^{N+1}) \tag{3.14}$$

avec pour tout $j = p, \ldots, N$

$$b_j(t_n) = G_y\left(y(t_n)\right).a_j(t_n) \quad \text{et} \quad \beta_n^j = G_y\left(y(t_n)\right).\alpha_n^j$$

Pour Richardson, l'estimation en z est du même ordre relatif que celle en y, car le développement (3.14) a la même forme que (2.13).

Pour les deux autres techniques, si la preuve de leurs propriétés repose sur les mêmes arguments que dans le cas des méthodes à un pas, on peut s'attendre à retrouver pour z les résultats obtenus pour y. En définitive, la difficulté ne réside pas dans le passage des EDO aux EAD d'indice 1 puisque l'approche *indirecte* permet d'une part de considérer y comme la variable d'une EDO et d'autre part de déduire directement l'erreur globale commise en z de celle commise en y. Mais on a vu que l'ordre relatif de ZD et SC n'est apparemment pas établi dans le cas des EDO, ce qui constitue un obstacle majeur à l'extension de ces algorithmes aux systèmes algébro-différentiels.

3.3 Résultats numériques

3.3.1 Tests avec RADAU5

Cette section est consacrée aux tests concernant l'estimation de l'erreur globale pour les méthodes de Runge-Kutta implicites, de type SAIRK exclusivement. On cherche moins ici à confirmer les résultats annoncés au début de cette partie quà comparer avec le cas des EDO et mettre à jour certains phénomènes numériques, souvent liés aux problèmes traités et se situant à la limite du cadre théorique utilisé.

Tous les tests ont été faits avec une version modifiée du solveur RADAU5 développé par E. Hairer et G. Wanner et dont le code est disponible à l'adresse

`http://www.unige.ch/math/folks/hairer/software.html`

Il permet de résoudre numériquement les EDO raides ainsi que les EAD implicites d'indice 1, 2 et 3. La méthode d'intégration implémentée est une SAIRK d'ordre 5 [1]. La solution est obtenue après quelques itérations (≤ 6) de l'algorithme de Newton et, dans sa version standard, RADAU5 choisit automatiquement le pas en fonction des tolérances absolue et relative données en entrée. Afin de disposer d'une intégration à pas constant et retrouver ainsi le cadre théorique, le code a été légèrement modifié. A l'origine, RADAU5 calcule $z_n = G(y_n)$ par Newton en jouant à la fois sur le nombre d'itérations de cet algorithme et le pas d'intégration h : la tolérance fixée est obtenue soit par une itération supplémentaire, soit par une réduction de h. A pas constant, on a donc compensé l'impossiblité pour h d'être diminué par un nombre d'itérations maximum ($= 6$). Cette modification ralentit certes un peu l'exécution du code d'intégration mais permet d'être toujours en-dessous de la tolérance fixée.

L'ordre p utilisé pour définir RS a été pris égal à 5 d'après ce qui précède. Pour ZD et SC, on utilise des polynômes de Lagrange par morceaux, calculés par différences divisées. Leur degré m est toujours 10 de façon à atteindre l'ordre maximum de ces deux estimateurs à moindre coût : $m = 10 = 2p$. Une fois m choisi, on effectue m intégrations du problème de départ puis on intègre le problème perturbé sur les m derniers pas avec la même méthode numérique.

Pour les cas simples 1, 2, 3 présentés plus loin et l'EDO $y' = cos(t).y$, la jacobienne est calculée analytiquement. Dans les autres cas, RADAU5 l'évalue numériquement.

Les premiers tests portent sur l'EDO scalaire non autonome

$$y' = cos(t).y\,, \qquad y(0) = 1 \quad \text{sur } [\,0, 20\,]$$

La solution est $y(t) = e^{sin(t)}$. Cette équation permet de valider les modifications apportées à RADAU5 en confrontant les résultats obtenus à ce qui a été fait précédemment pour les EDO [3].

1. Cet ordre de convergence est valable pour les EDO et les EAD d'indice 1. Lorsque l'indice est ≥ 2, il ne vaut plus que 3 pour la variable algébrique [4, 12].

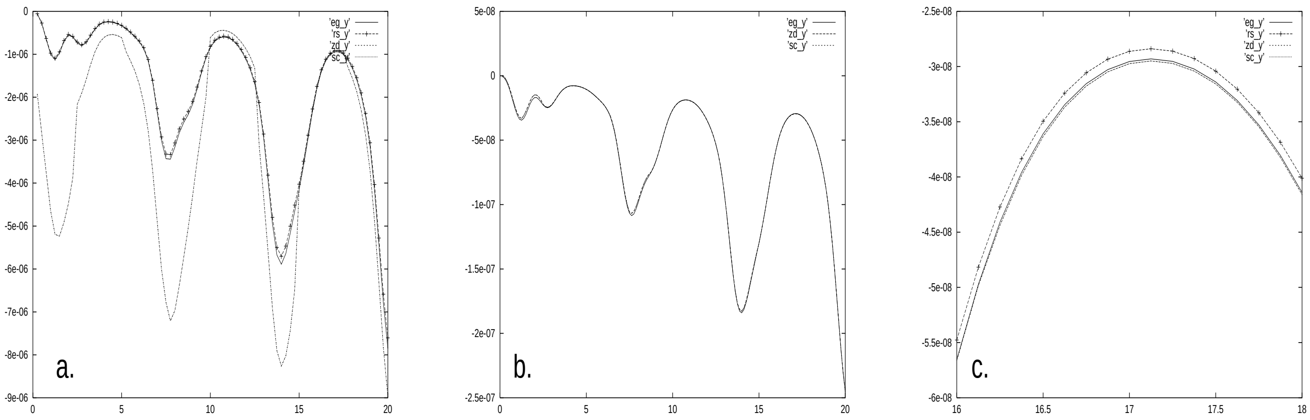


FIG. 3.1 – $y' = cos(t).y$ *à pas constant sur* $[0, 20]$

Dans le cas à pas constant, RS donne de très bons résultats avec le pas $h = 0.25$, SC est assez précis mais moins bon que RS et ZD a plus de mal à converger (fig. 3.1-a). Avec un pas plus petit, $h = 0.125$, SC et ZD donnent une excellente estimation (fig. 3.1-b), meilleure que celle de RS (fig. 3.1-c). Ceci confirme que la convergence de ZD et SC est plus longue à s'installer mais, quand c'est le cas, la précision obtenue est plus grande.
A pas variable, RS donne une estimation très fiable pour $rtol = atol = 10^{-5}$ (fig. 3.2-a). Avec $rtol = atol = 10^{-8}$, SC finit par converger et ZD s'est un peu amélioré (fig. 3.2-b). Pour ZD, la convergence est plus lente. Avec $atol = rtol = 10^{-9}$, l'estimation de ZD est beaucoup moins bonne qu'avec DOPRI [3]. Ceci est dû vraisemblablement au fait que RADAU5 utilise une méthode de Runge-Kutta *implicite* alors que DOPRI repose sur une ERK. Ce premier constat est important : même si intégrer une EAD semi-explicite d'indice revient à résoudre une EDO , l'implicitation faite par le solveur doit être prise en compte. (fig. 3.2-c).

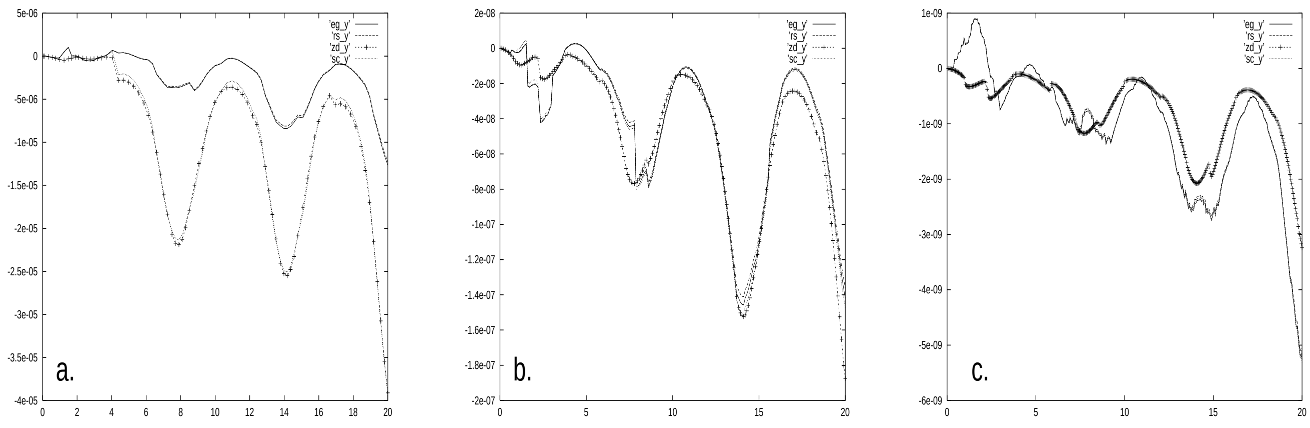


FIG. 3.2 – $y' = cos(t).y$ *à pas variable sur* $[0, 20]$

On considère maintenant 3 systèmes algébro-différentiels très simples afin de comparer les 3 estimateurs sur les points suivants : installation de la convergence, capacité à prendre en compte l'erreur due à l'algorithme de Newton et cas des problèmes très stables. Ces tests ne concernent que la variable algébrique.

Les systèmes étudiés ici sont tous des EAD semi-explicites d'indice 1, non linéaires et autonomes dont on connaît la solution exacte.

système 1. $\begin{cases} y' = y + z^2\,, & y(0) = 1 \\ 0 = y - z^2\,, & z(0) = 1 \end{cases}$ sur $[\,0:10\,]$

système 2. $\begin{cases} y' = y\,, & y(0) = 1 \\ 0 = 2 + y^2 z\,, & z(0) = -2 \end{cases}$ sur $[\,0:10\,]$

système 3. $\begin{cases} y' = y^2 + z\,, & y(0) = 1 \\ 0 = 2y^2 + z\,, & z(0) = -2 \end{cases}$ sur $[\,0:1\,]$

La solution exacte est respectivement pour chacun de ces trois systèmes $(e^{2t},\, e^t)$, $(e^t, -2e^{-2t})$ et $(\frac{1}{1+t}, -\frac{2}{(1+t)^2})$.

On visualise, pour la variable algébrique et en fonction du temps, l'erreur globale (eg_z) et les 3 estimations obtenues (rs_z), (zd_z) et (sc_z).

$*$ Installation de la convergence

On commence par intégrer le système 1 à pas constant sur $[\,0, 10\,]$ en faisant varier le nombre de pas. On voit (fig. 3.3-a) que l'estimation de Richardson est correcte même pour un nombre de pas faible (20 pas sur $[\,0, 10\,]$) alors que ZD et SC n'arrivent pas à suivre l'erreur globale. Lorsque le nombre de pas augmente, ZD et SC gagnent en précision mais, pour 30 et 40 pas, restent encore moins bons que RS (fig. 3.3-b,c). Le test réalisé avec 50 pas (fig. 3.3-d) montre que les 3 estimateurs estiment l'erreur globale de façon satisfaisante sur tout l'intervalle $[\,0, 10\,]$. Si on revient à $[\,0, 4\,]$, il est encore difficile de comparer les 3 estimations qui semblent aussi bonnes l'une que l'autre (fig. 3.3-e). Un zoom sur $[\,3, 3.5\,]$ permet pourtant de constater que ZD et SC sont désormais meilleurs que RS (fig. 3.3-f). Le phénomène observé est caratéristique des méthodes d'ordre élevé : la convergence est généralement plus longue à s'installer mais, une fois installée, est plus rapide. Quand on divise le pas par 10, on gagne un facteur 10 pour RS contre 10^5 pour ZD et SC.

A pas variable, ce système montre que l'installation de la convergence peut être plus lente pour ZD que pour SC . En effet, si $rtol = atol = 10^{-6}$, SC estime convenablement alors que pour une même convergence relative de ZD il faut attendre 10^{-8}. Enfin, à 10^{-12}, ZD et SC sont meilleurs que RS.

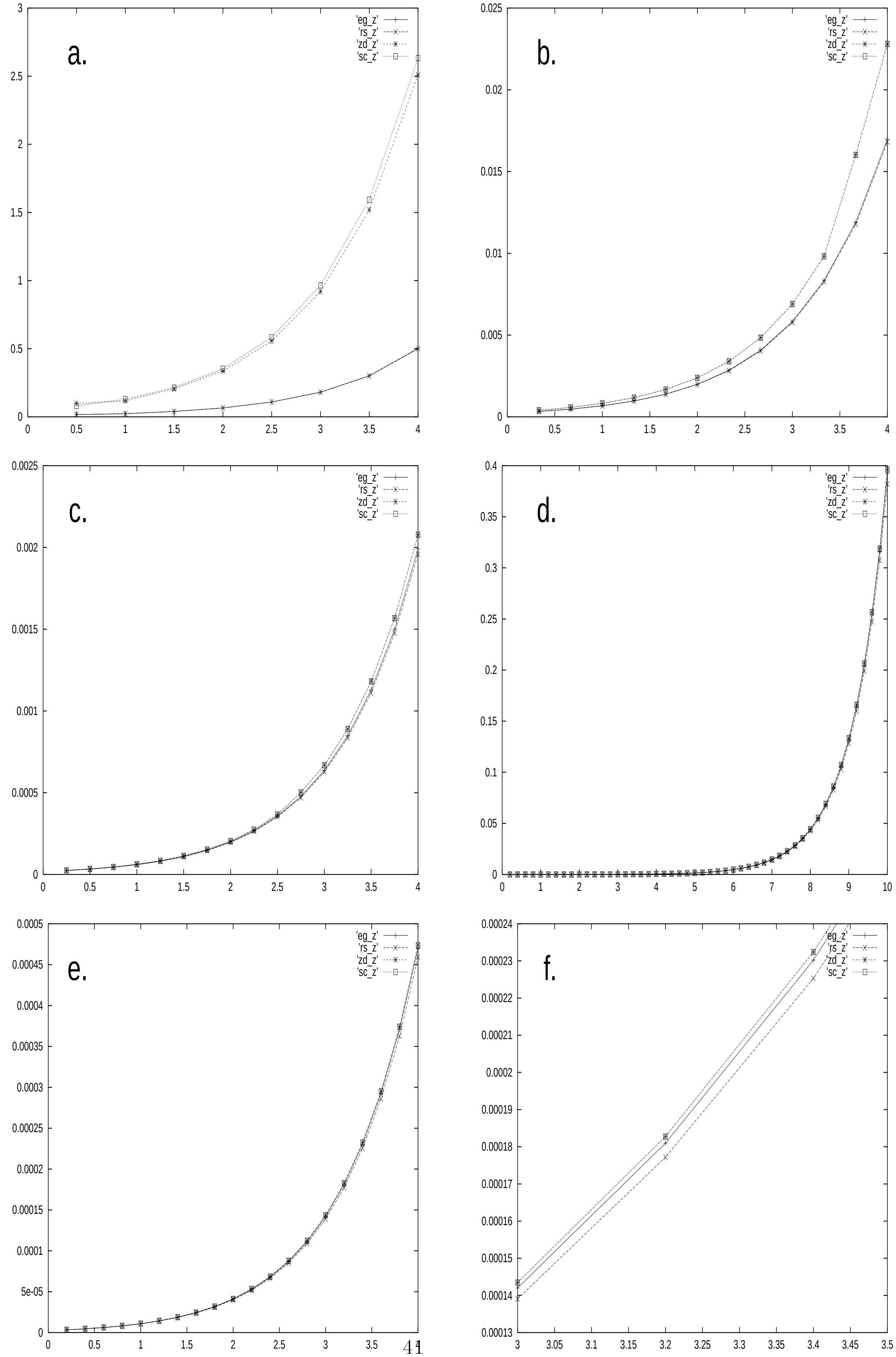


FIG. 3.3 – *Vitesse de convergence à pas constant : système 1. avec 20, 30, 40 pas sur [0,4] puis 50 pas sur [0,10], [0,4] et [3,3.5]*

$*$ **Capacité à prendre en compte l'erreur due à la méthode de Newton**
La résolution de $g(y_n, z_n) = 0$ fournit des valeurs $\tilde{z}_n$ telles que $\tilde{z}_n - z_n = \mathcal{O}(h^\alpha)$ est l'erreur commise par l'algorithme de Newton. Pour RS, cette errreur est incorporée dans $\mathcal{O}(h^{p+1})$. Elle est également prise en compte par SC dans le calcul de l'erreur globale mais pas par ZD. En effet, $\varepsilon_{n,2}$ est obtenu directement par Newton et comprend cette composante de l'erreur globale. Pour ZD, cette composante n'est a priori pas la même quand on intègre le problème initial et le problème perturbé. On illustre cette remarque avec le système 2 pour lequel l'algorithme de Newton n'a pas convergé lors de l'intégration à pas variable avec $rtol = atol = 10^{-6}$. Ceci introduit une erreur non négligeable qui expliquerait les fortes oscillations de l'erreur globale (fig. 3.4-a). Si RS et SC fournissent une très bonne estimation, ZD 'décroche' dès l'apparition de ces oscillations. En revanche, à pas constant le nombre d 'itérations de Newton est fixé à 6 et l'erreur introduite est alors toujours négligeable. C'est bien ce qu'on observe : avec 50 pas, l'erreur globale est très régulière (forme d'exponentielle inverse) et son estimation ne pose plus de problème, quelle que soit la technique utilisée (fig. 3.4-b).

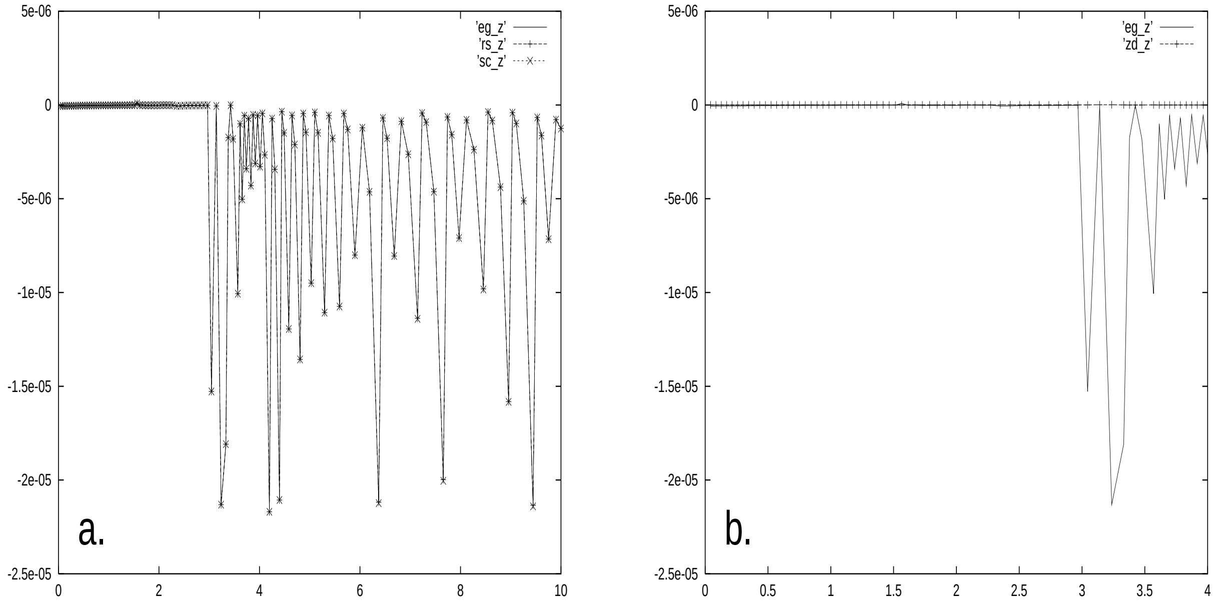


FIG. 3.4 – *Sensibilité à la régularité de l'erreur globale : système 2 avec $rtol = atol = 10^{-6}$ sur [0,10] et sur [0,4]*

$*$ **Cas des problèmes très stables**
Par exemple, le système 3 intégré sur $[\,0, 1\,]$ à pas constant a la particularité d'être très stable. Cet exemple nous assure donc que l'erreur à estimer sera toujours très faible et que l'estimation ne pourra être fiable, quelque soit l'estimateur utilisé , la précision machine étant atteinte. On est donc en présence d'un problème très stable qu'une méthode implicite est capable d'intégrer avec un pas h grand (faisceau d'exponentielles croissantes par exemple). Les estimateurs sont donc forcément imprécis puisque leur précision dépend de h. Ce n'est pas grave dans la mesure où les problèmes qu'on peut résoudre très précisément avec de grands pas ne sont pas ceux pour

lesquels l'estimation de l'erreur globale présente un intérêt : on sait que l'erreur diminue très vite avec le pas.

En définitive, ces tests ont permis d'observer pour la variable algébrique des phénomènes déjà mentionnés dans le cadre EDO [3]. En particulier, RS et SC semblent moins sensibles à la forme de l'erreur globale que ZD, ce qui pose la question d'une éventuelle adaptation de cet algorithme. Si l'installation de la convergence est plus longue pour SC que pour RS, l'estimation obtenue est finalement plus précise, ce que prévoyaient la théorie et les tests réalisés sur les EDO [2]. Dans la plupart des cas, on a observé comme pour les EDO que RS donnait en moyenne 1 chiffre significatif de l'erreur globale exacte et ZD et SC 4, 5 chiffres. On n'a donc pas observé de difficulté due à la variable algébrique elle-même, conformément à ce qui a été dit au début de cette partie dans le cas des méthodes SAIRK, et les limitations constatées (notamment lorsque le pas varie) viennent plutôt de l'allure de l'erreur globale (trop irrégulière / trop faible) d'une part et du caractère implicite de la méthode d'intégration d'autre part. En effet, SC donne de meilleurs résultats que ZD car il intègre notamment la composante de l'erreur due à la méthode de Newton utilisée pour résoudre les itérations implicites.

On teste enfin les estimateurs sur 3 exemples d'indice 1 plus réalistes présentés dans [12] : une EAD semi-explicite de dimension 3 issue de la thermodynamique, une EAD implicite de dimension 8 modélisant un amplificateur et la formulation d'indice 1 du pendule non linéaire (EAD semi-explicite de dimension 5). Une présentation détaillée se trouve en annexe et on se contente ici de donner l'allure générale de ces systèmes. Dans tous les cas la solution exacte n'est pas connue et on ne dipose pas de l'erreur globale exacte. La référence utilisée pour tester les estimateurs est alors une approximation *assez précise* de l'erreur globale, obtenue en résolvant une première fois le problème avec un nombre de pas très grand : la solution numérique calculée est substituée dans la définition de l'erreur globale à la solution exacte qu'on ne connaît pas.

$$\text{système 4a} \qquad \begin{cases} y_1' = -y_2^2\,, & y_1(0) = P_0 \\ 0 = f_a(y_1, y_2, y_3)\,, & y_2(0) = u_0 \\ 0 = g_a(y_2, y_3)\,, & y_3(0) = y_0 \end{cases} \qquad \text{sur } [\,0, 8.047\cdot 10^6\,]$$

où f_a, g_a, P_0, u_0 et y_0 sont donnés en annexe. Les solutions de ce problème sont très régulières. Une situation un peu différente est donnée par

$$\text{système 4b} \qquad \begin{cases} y_1' = -y_2^2\,, & y_1(0) = P_0 \\ 0 = f_b(y_1, y_2, y_3)\,, & y_2(0) = u_0 \\ 0 = g_b(y_2, y_3)\,, & y_3(0) = y_0 \end{cases} \qquad \text{sur } [\,0, 3.2188\cdot 10^7\,]$$

où f_b, g_b diffèrent de f_a et g_a par le choix des constantes. Cette deuxième version présente une singularité vers $1.3.10^7$ [12].

$$\text{système 5} \qquad MY' = f(Y)\,, \quad Y(0) \in \mathbb{R}^8 \qquad \text{sur } [\,0, 0.2\,]$$

où M est une matrice constante singulière de dimension 8, de la forme

$$M = \begin{pmatrix} C_1 & -C_1 & & & & & & \\ -C_1 & C_1 & & & & & & \\ & & C_2 & & & & & \\ & & & C_3 & -C_3 & & & \\ & & & -C_3 & C_3 & & & \\ & & & & & C_4 & & \\ & & & & & & C_5 & -C_5 \\ & & & & & & -C_5 & C_5 \end{pmatrix}$$

Ce système, bien qu'étant d'indice 1, n'est donc pas semi-explicite.

$$\text{système 6} \qquad \begin{cases} z_1' = z_3\,, & z_1(0) = 1 \\ z_2' = z_4\,, & z_2(0) = 0 \\ z_3' = -z_1 z_5\,, & z_3(0) = 0 \\ z_4' = -z_2 z_5 - g\,, & z_4(0) = 0 \\ 0 = z_3^2 + z_4^2 - z_5 L^2 - z_2 g\,, & z_5(0) = 1 \end{cases} \quad \text{sur } [\,0, 30\,]$$

où g et L sont deux constantes.

A partir du bilan dressé pour la première série de tests, on s'intéresse ici autant aux variables différentielles qu'aux variables algébriques. On visualise donc les approximations des erreurs globales, eg_i ainsi que les estimations calculées rs_i, zd_i et sc_i, où i désigne le numéro ou le nom de la variable étudiée. Les systèmes 4a, 4b, 5 et 6 permettent d'observer les trois phénomènes suivants.

$*$ **Influence de la nature de la solution numérique sur la qualité de l'estimation**
Dans la version 4a, les 3 solutions numériques obtenues sont des droites. L'interpolation polynomiale de la solution numérique redonne donc les mêmes valeurs. Pour ZD, cela se traduit par un défaut nul : le problème de départ et le problème perturbé sont identiques et par différence, on obtient 0 comme estimation de l'erreur globale. Le même raisonnement semble valable pour SC, pourtant l'expérience montre que cet estimateur se comporte davantage comme RS et prend des valeurs de l'ordre de 10^{-16}. Cet exemple constitue donc un cas pour lequel le type de défaut choisi ne convient pas et pose le problème du choix du défaut lorsque le système admet des solutions polynomiales. Avec la version 4b, la situation est radicalement différente puisque la solution numérique n'approche plus un polynôme et présente même une singularité (fig. 3.5). Ceci est d'autant plus intéressant que celle-ci apparaît relativement tard : on a le temps d'observer le comportement des estimateurs dans la zone 'calme' avant de considérer la zone transitoire.
Il s'avère qu'avec 800 pas, l'erreur globale est nulle sur $[\,0, 1.28752 \cdot 10^8\,]$, après quoi elle prend très rapidement des valeurs élevées, ce qui correspond à l'apparition de la singularité pour la solution numérique (fig. 3.5). Sur $[\,0, 1.2 \cdot 10^7\,]$, il semble que ZD et SC suivent l'erreur globale plus longtemps que RS, avant d'être définitivement perturbés à l'approche de la singularité (fig. 3.6-a).

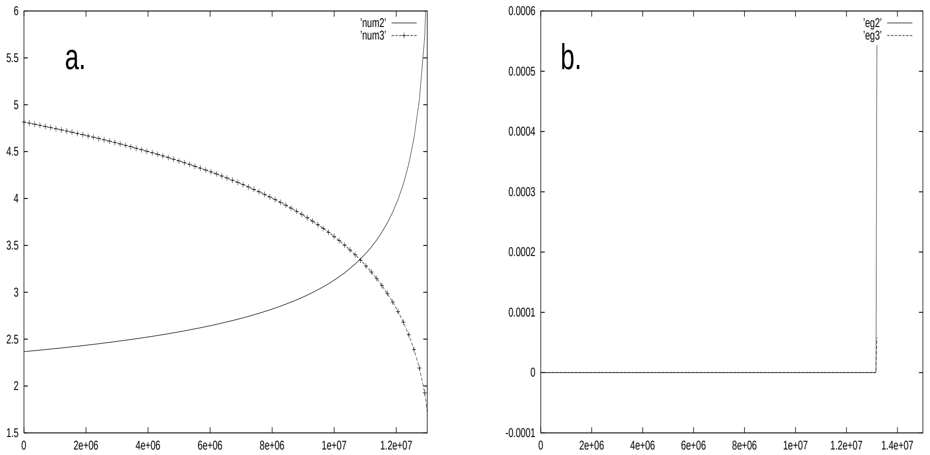


Fig. 3.5 – *système 4b : 800 pas sur* $[\,0, 1.5\cdot 10^7\,]$*, variables* y_2 *et* y_3

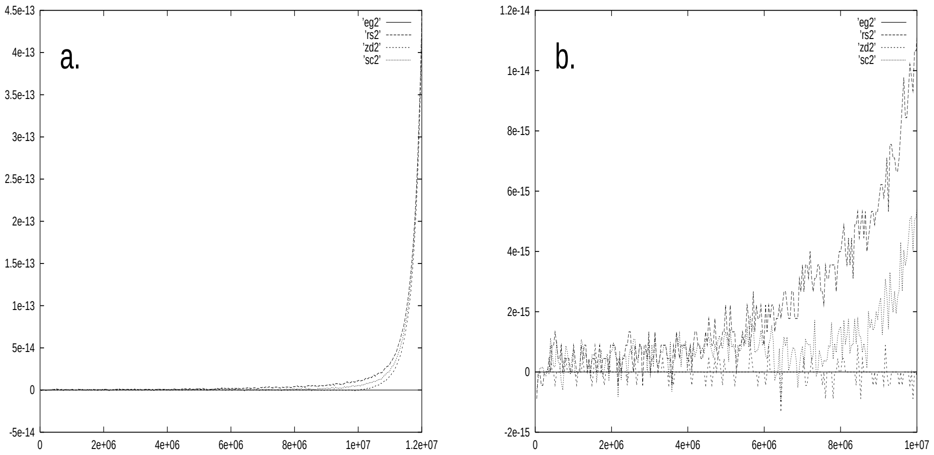


Fig. 3.6 – *système 4b : loin de la singularité*

∗ **Limitations de RS**
Tous les tests réalisés jusqu'à présent montrent que RS, malgré un ordre relatif minimum, est le plus fiable des 3 estimateurs, notamment lorsque l'erreur globale varie beaucoup : RS parvient en effet à suivre au moins l'allure de celle-ci, ce qui n'est pas le cas pour ZD et SC. Le système 5 constitue un cas légèrement nouveau au sens où RS n'arrive pas à suivre les oscillations de l'erreur globale et prend des valeurs beaucoup plus importantes (fig. 3.7). La situation est la même à pas variable.

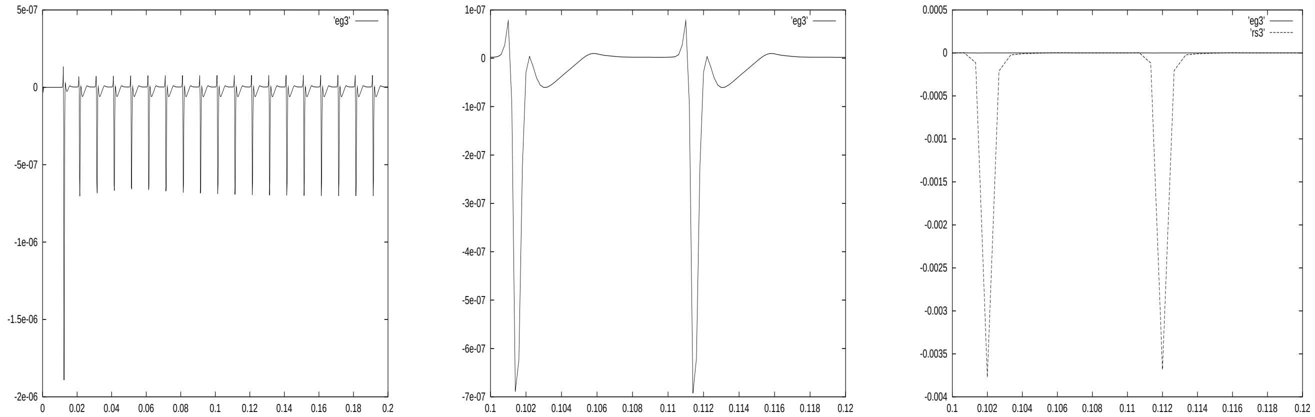


FIG. 3.7 – *système 5 :* 500 *pas sur* $[0, 0.2]$ - *mauvaise estimation pour RS*

∗ **Prévision d'un mauvais comportement de la solution numérique**
On intègre le système 6. sur $[0, 30]$. La solution numérique calculée par RADAU5 diverge (pour toutes les variables) vers $t = 19$ (fig. 3.8-a).

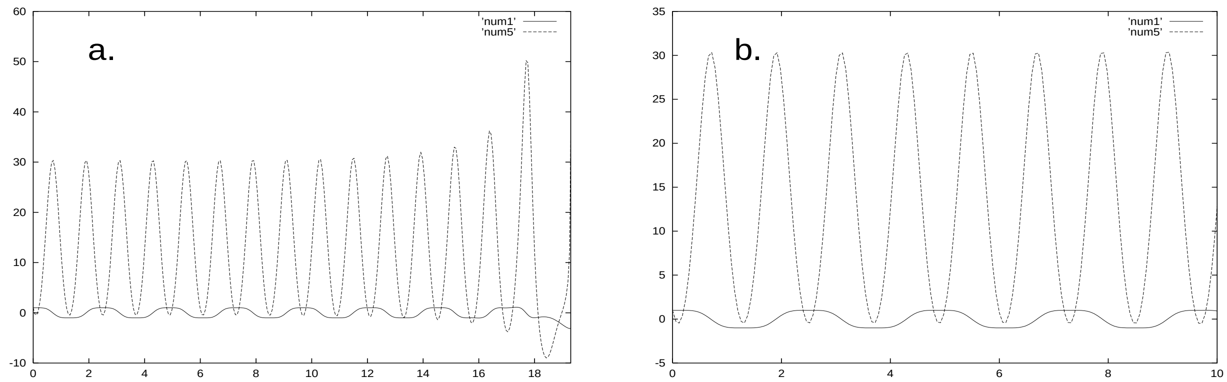


FIG. 3.8 – *système 6 :* 500 *pas sur* $[0, 30]$

Cette divergence est assez soudaine et n'est pas détectable à l'oeil nu si on se restreint à $[0, 10]$ par exemple (fig. 3.8-b). Pourtant l'erreur globale présente une instabilité déjà clairement perceptible sur cet intervalle (fig. 3.9, premier graphique). En revanche si l'intégration se fait sans problème sur un intervalle plus grand, $[0, 40]$ par exemple, l'erreur globale ne diverge pas. Elle constitue donc un moyen d'anticiper une éventuelle divergence de la solution numérique. Pouvoir l'estimer correctement est donc particulièrement souhaitable dans ce cas. Les deux autres graphiques montrent que c'est le cas, même avec l'estimateur de Zadunaisky qui est pourtant le plus difficile à faire converger.

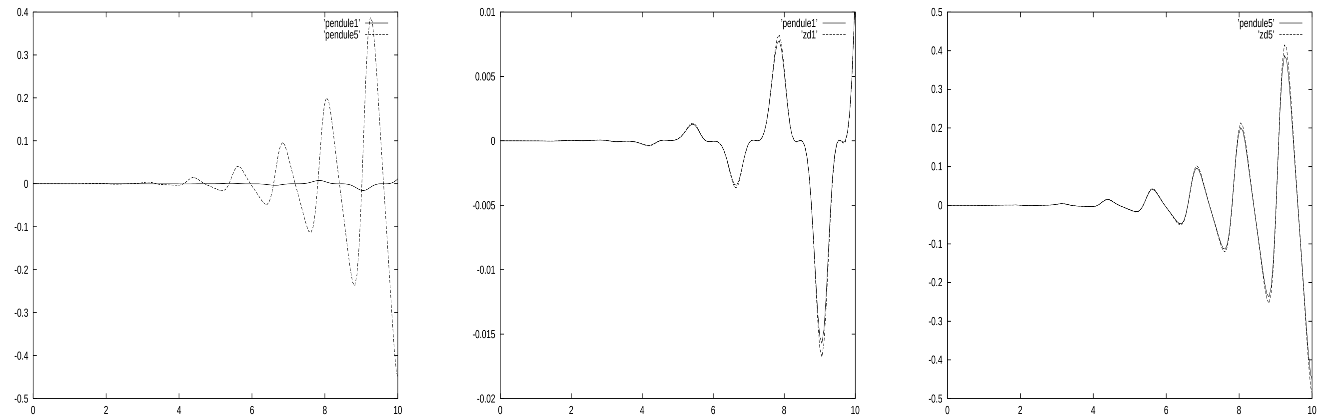


FIG. 3.9 – *système 6 : anticipation de la divergence de la solution numérique par l'erreur globale et son estimation*

Finalement, les systèmes 4, 5 et 6 confirment les expérimentations précédentes en soulignant le manque de souplesse de ZD, aussi bien lorsque l'erreur est extrêmement faible que lorsqu'elle oscille rapidement. SC et RS sont eux aussi sensibles à la forme de l'erreur globale mais seulement lorsque celle-ci varie de façon très brutale. Enfin, l'exemple du pendule illustre bien l'intérêt d'estimer l'erreur globale précisément.

Conclusion

Ces observations mettent en lumière les difficultés que pose la résolution des EAD semi-explicites d'indice 1 par des méthodes implicites de type SAIRK. La théorie prévoyait une situation en tout point analogue au cas des EDO mais on a vu que le fait que la méthode soit implicite permet de départager ZD et SC, ce dernier estimateur étant le plus fiable des deux. D'autre part il semblerait qu'une solution numérique très instable rende vaine toute tentative d'estimation, même avec RS. Ce cas où aucun des trois estimateurs ne fonctionne est toujours lié à la nature de la solution numérique, ce qui pose deux questions. Une solution numérique très irrégulière conduit-elle systématiquement, comme le laissent penser les tests réalisés, à une erreur globale elle-aussi très irrégulière? La difficulté pour les EAD d'indice 1 ne porte-t-elle pas sur l'intégration numérique du système plutôt que sur l'estimation de l'erreur globale?

3.3.2 Tests avec DASSL

On présente ici quelques tests réalisés avec une version modifiée du solveur DASSL et permettant d'étudier le comportement de l'estimateur RS pour les méthodes BDF. Le code original est accessible depuis le serveur FTP

`ftp://ftp.cwi.nl/pub/IVPtestset`

Pour une présentation détaillée, on peut se reporter à [5]. Petzold y indique notamment que DASSL est avant tout conçu pour résoudre numériquement des EDO et EAD d'indice 1 et que l'intégration repose sur une méthode BDF-k (avec k = 5) utilisant l'approche *indirecte* vue au chapitre 3. Cette méthode est fortement stable et entre donc dans le cadre théorique utilisé pour les méthodes multipas. De plus, l'intégration se fait à pas *et* ordre *variables*, l'ordre variant de 1 à 5. Les systèmes traités sont toujours sous forme implicite et l'ordre est le même pour toutes les variables. La principale modification apportée au code initial a été de proposer une intégration *à pas constant* et de permettre de *borner l'ordre de la méthode*. Cette dernière option est importante pour définir l'estimateur de Richardson. En effet, le premier pas est effectué avec une méthode d'ordre 1 (Euler implicite) et l'ordre est ensuite augmenté autant que possible. Il peut cependant être diminué si le respect des tolérances l'exige. Une façon de définir RS serait donc d'appliquer la formule vue pour les EDO avec la bonne valeur de l'ordre p à chaque pas. Cependant, la pratique montre que le deuxième pas est toujours effectué avec une méthode d'ordre 2. Si les pas suivants imposent une réduction ou une stagnation de l'ordre, il faudra finalement prendre $p = 2$ pour construire RS. Si $p \in \{3, 4, 5\}$, comme d'une part l'erreur globale est obtenue par accumulation de l'erreur locale à chaque pas, et comme d'autre part, $t_n \sim n\, h$, l'erreur globale en fin d'intégration correspond finalement à celle qu'on aurait obtenue avec une méthode d'ordre 2. Cela explique le choix $p = 2$ dans la définition de RS.

Le premier problème testé est encore l'EDO $y' = cos(t)\, y$ qui permet de retrouver les observations faites avec DOPRI. On voit cependant (fig. 3.10) que RS met plus de temps à converger que dans le cas des méthodes à un pas où $h = 0.25$ donnait déjà d'excellents résultats. Cependant on dispose avec $h = 0.01$ d'une estimation visiblement fiable.

On a aussi testé les EAD 1, 2 et 3 à pas constant sur $[\,0, 0.5\,]$. Avec 100 pas, on arrive à obtenir des estimations assez satisfaisantes lorsque l'erreur globale est lisse au départ (fig. 3.11-a). Mais la figure 3.11-b montre que le fait que le premier pas soit calculé avec une méthode d'ordre 1 peut pénaliser RS en particulier lorsque l'erreur globale présente des irrégularités en début d'intégration.La figure 3.11-c montre une erreur globale moins violente et RS marche mieux.

On illustre enfin sur le pendule non linéaire d'indice 1 (système 6) la possibilité pour l'estimation de l'erreur globale de prédire, comme pour les méthodes à un pas, une éventuelle divergence numérique.

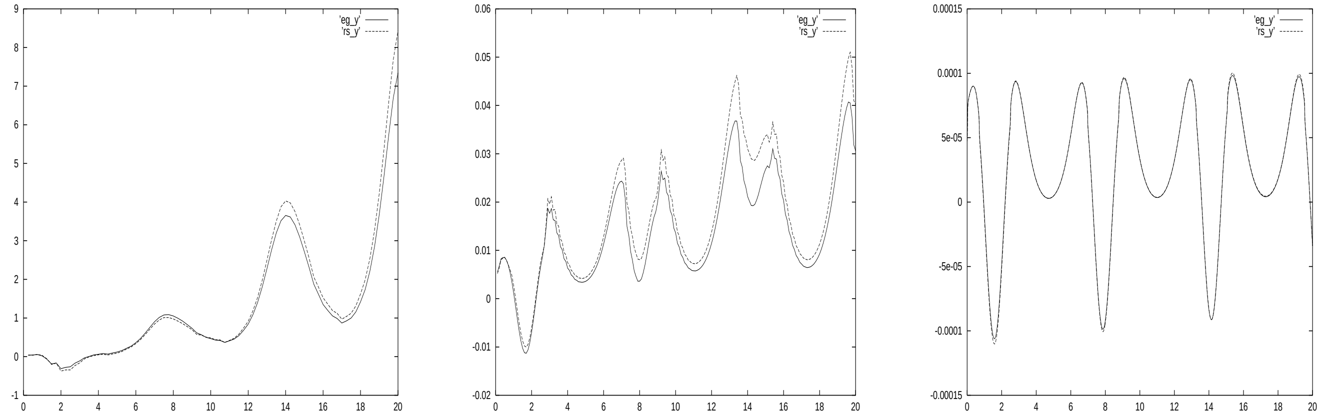


Fig. 3.10 – $y' = cos(t) \cdot y$ *pour* $h \in \{0.25,\, 0.1\; 0.01\}$ $[\,0, 20\,]$

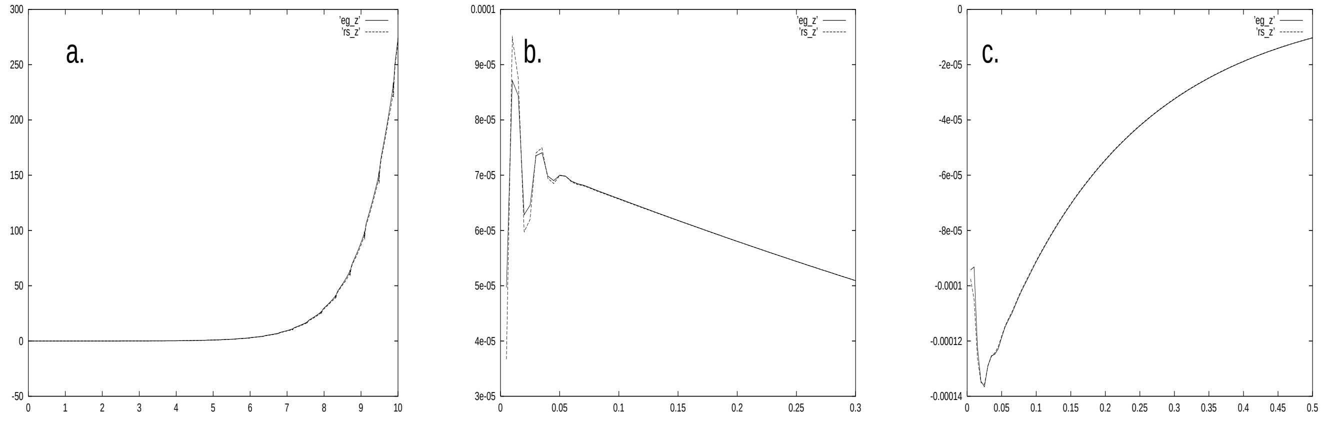


Fig. 3.11 – *Systèmes 1, 2 et 3 avec* $h = 5 \cdot 10^{-3}$ *sur* $[\,0, 0.5\,]$

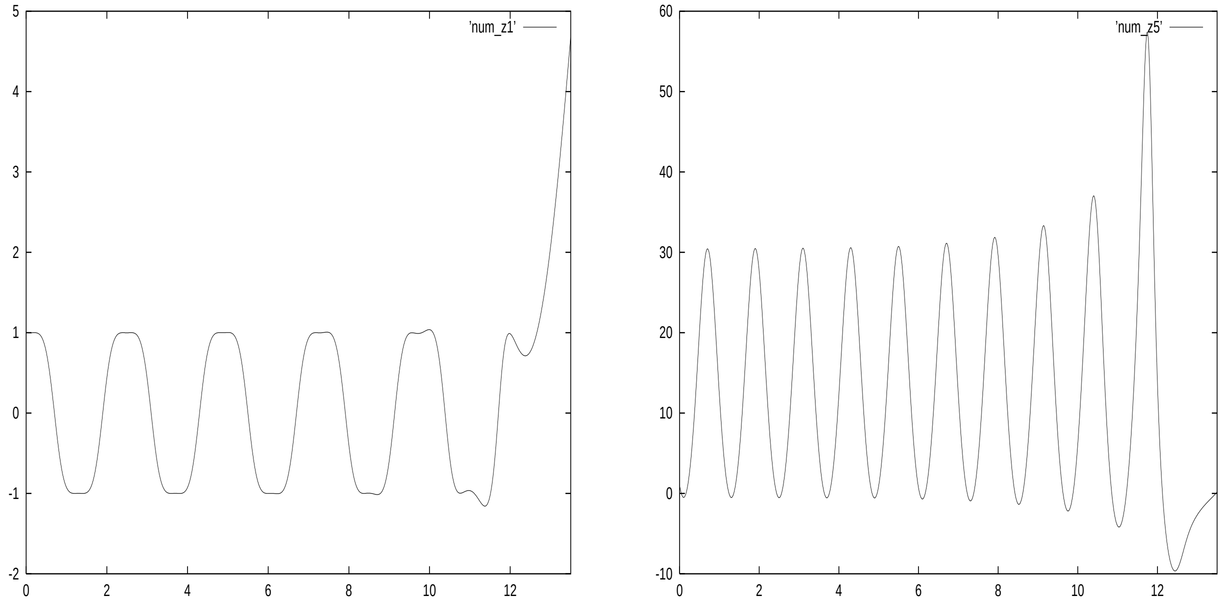


FIG. 3.12 – *Pendule non linéaire d'indice 1 avec* $h = 5 \cdot 10^{-3}$ *sur* $[\,0:40\,]$ *: solutions numériques*

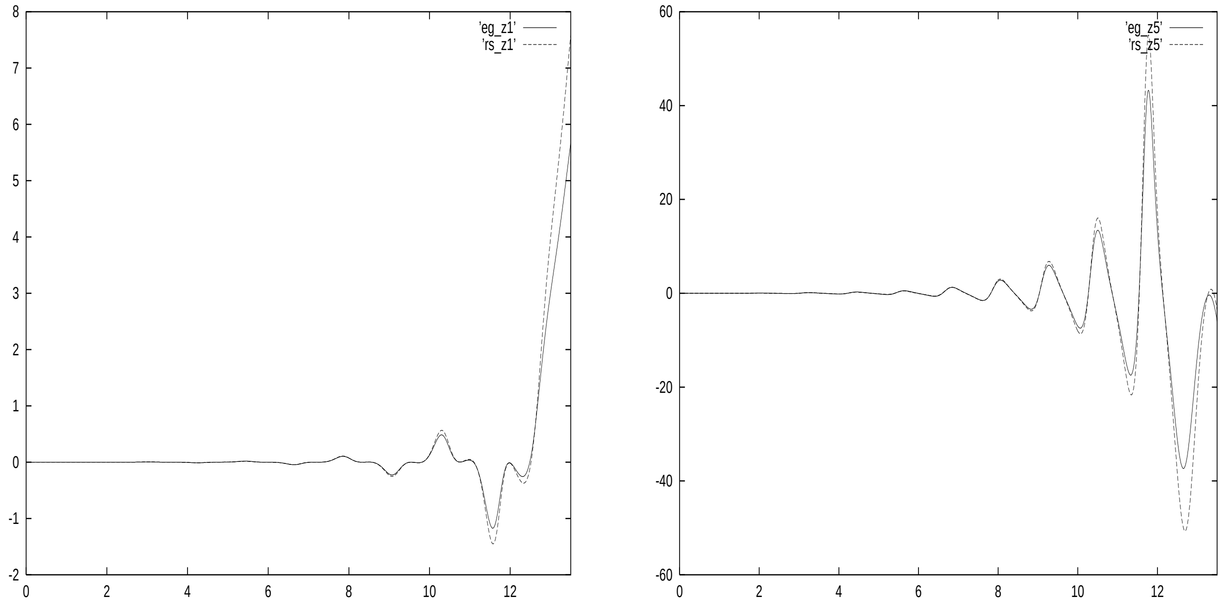


FIG. 3.13 – *Pendule non linéaire d'indice 1 avec* $h = 5 \cdot 10^{-3}$ *sur* $[\,0:40\,]$ *: estimation RS*

La figure 3.12 montre que pour les variable z_1 (différentielle) et z_5 (algébrique), la divergemce numérique a lieu autour de $t = 13$ alors que l'erreur globale (et son estimation par RS) diverge déjà vers $t = 12$ 9fig. 3.13).

Chapitre 4

Estimation de l'erreur globale pour les EAD semi-explicites d'indice 2

4.1 Motivation

Le pendule non linéaire peut être formulé de plusieurs façons, et notamment conduire à une EAD d'indice 1, 2 ou 3. Après les tests réalisés sur la formulation d'indice 1, il était tentant de s'intéresser à l'estimation de l'erreur globale sur la version d'indice 2. Cette formulation (détaillée dans la partie consacrée aux tests) donne des résultats encourageants pour les 3 estimateurs avec RADAU5. En revanche, DASSL ne traite pas les problèmes d'indice 2 [5] et nous n'avons fait aucun test sur des méthodes multipas.

La figure 4.1 montre pour RS les résultats obtenus à pas constant sur un grand intervalle de temps. Les variables étudiées sont z_1, z_3 et z_5. On obtient une très bonne estimation, les deux courbes étant confondues.

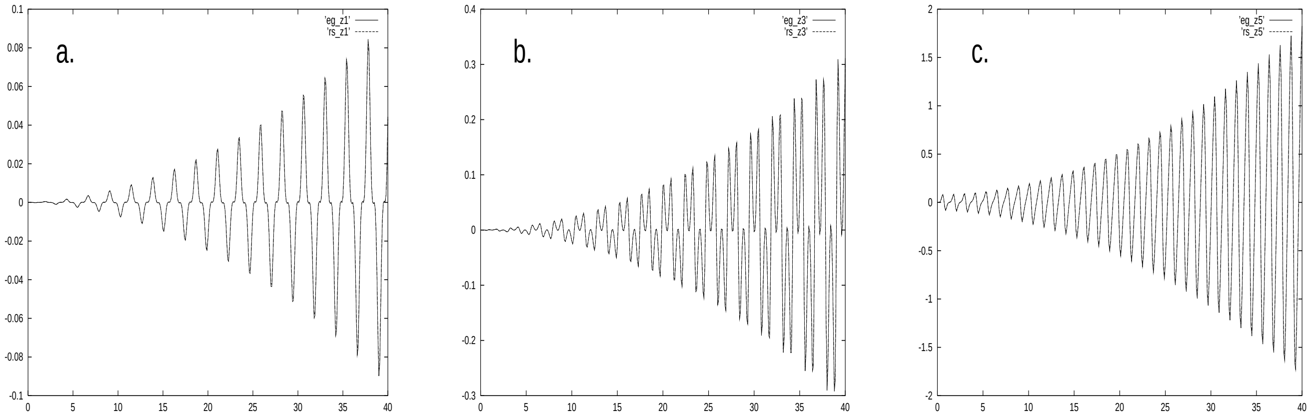


FIG. 4.1 – *Pendule non linéaire d'indice 2 : à pas constant $h = 0.1$ sur* $[0:40]$

On compare ensuite les 3 estimateurs en se restreignant à $[21, 23]$ pour pouvoir distinguer les trois courbes.

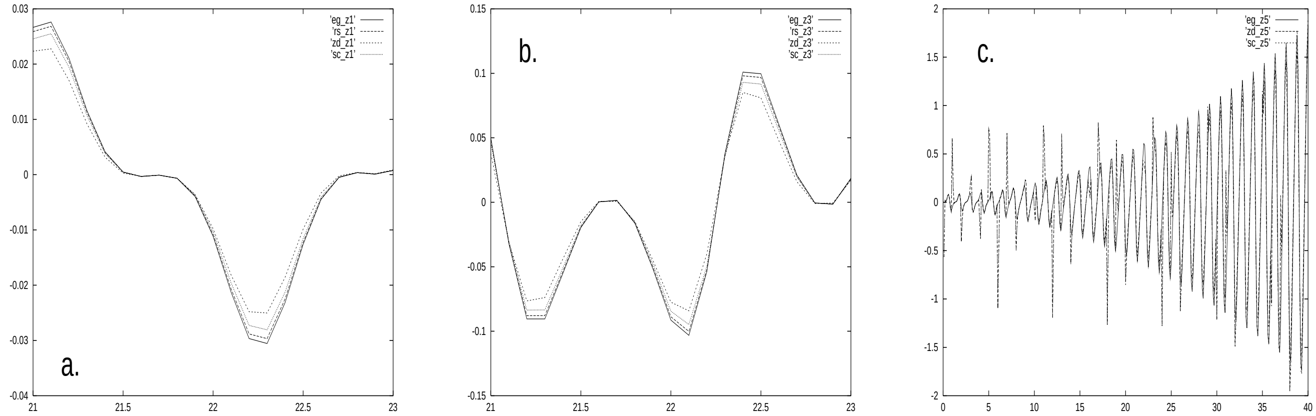


FIG. 4.2 – *Pendule non linéaire d'indice 2 : à pas constant* $h = 0.1$ *sur* $[0:40]$

Il apparaît qu'avec le pas choisi, RS est meilleur que SC qui est lui-même plus précis que ZD (fig. 4.2-a,b), observation qui avait déjà été faite dans le cas de l'indice 1. Mais sur la figure 4.2-c, ZD présente de fortes oscillations, ce qui laisse penser que le pas a été choisi trop grand.

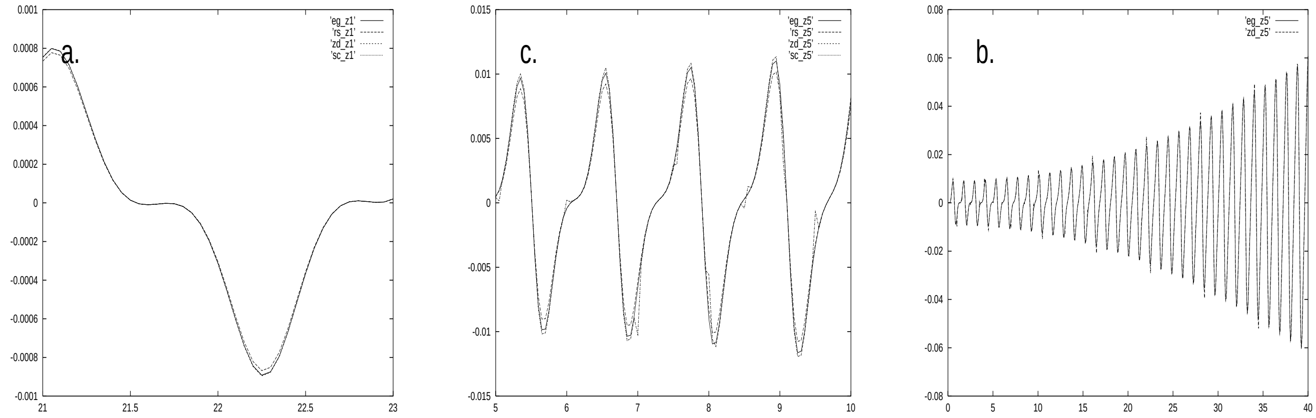


FIG. 4.3 – *Pendule non linéaire d'indice 2 : à pas constant* $h = 0.05$ *sur* $[0:40]$

Diviser le pas h par 2 suffit à améliorer considérablement la situation (fig. 4.3). Les figures 4.3-a,b montrent que ZD $\equiv$ SC sont désormais meilleurs que RS. Cela tient au fait que les oscillations pour ZD s'écartent moins brutalement de l'erreur globale qu'avant (fig. 4.3-c).

4.2 Extension des méthodes IRK et BDF aux EAD

On considère le système (1.1) en supposant cette fois que g_z est singulière et que deux dérivations de la contrainte algébrique suffisent à faire apparaître z'. Un tel système est dit *semi-explicite d'indice 2*. Dans [14], Hairer montre qu'il est possible de restreindre l'étude de tels problèmes aux EAD de la forme

$$\begin{aligned} y' &= f(y, z) \\ 0 &= g(y) \end{aligned} \tag{4.1}$$

avec

$$g_y(y) f_z(y, z) \text{ inversible borné dans un voisinage de la solution}$$

à condition que la transformation qui conduit à (4.1) ne change pas l'indice (i.e. ne nécessite aucune différenciation) et préserve les méthodes numériques. Comme c'est souvent le cas, la classe de problèmes (4.1) est assez représentative des EAD semi-explicites d'indice 2.

Une méthode IRK appliquée à (4.1) s'écrit

$$\begin{aligned} y_{n+1} &= y_n + h \sum_{i=1}^{s} \underline{b}_i Y'_{ni} \\ Y_{ni} &= y_n + h \sum_{j=1}^{s} \underline{a}_{ij} Y'_{nj} \\ Y'_{ni} &= f(Y_{ni}, Z_{ni}) \\ 0 &= g(Y_{ni}) \\ z_{n+1} &= z_n + h \sum_{i=1}^{s} \underline{b}_i Z'_{ni} \\ Z_{ni} &= z_n + h \sum_{j=1}^{s} \underline{a}_{ij} Z'_{nj} \end{aligned}$$

Ce schéma correspond à l'approche *directe* vue pour l'indice 1. L'approche qui consistait à résoudre l'EDO équivalente n'a bien sûr plus de sens ici puisque la contrainte algébrique est indépendante de z.

Pour les méthodes multipas linéaires, on a comme pour l'indice 1 les deux approches.
L'approche *indirecte* donne

$$\begin{aligned} \sum_{i=0}^{k} \alpha_i y_{n+i} &= h \sum_{i=0}^{k} \beta_k f(y_{n+k}, z_{n+k}) \\ 0 &= g(y_{n+k}) \end{aligned}$$

et l'approche *directe* consiste là encore à poser $\epsilon := 0$ dans le problème à pertubation singulière associé qui a cette fois la forme

$$\begin{aligned} y' &= f(y,z) \\ \epsilon z' &= g(y) \end{aligned}$$

On obtient alors

$$\sum_{i=0}^{k} \alpha_i y_{n+i} = h \sum_{i=0}^{k} \beta_k f(y_{n+k}, z_{n+k})$$

$$0 = \sum_{i=0}^{k} \beta_i g(y_{n+k})$$

Pour les BDF-k, ces deux définitions sont équivalentes.

4.3 Peut-on encore appliquer les 3 estimateurs?

Pour les méthodes multipas, il n'existe pas à notre connaissance d'étude donnant le développement asymptotique de l'erreur globale pour les EAD d'indice 2. On envisage donc seulement le cas des méthodes à un pas, IRK notamment [12].

Par ailleurs, il est désormais impossible de se ramener à une EDO en y à partir de (4.1) et la théorie de l'estimation de l'erreur globale pour les EDO n'est plus applicable ici. Il faut donc étudier le développement asymptotique de l'erreur globale commise en chaque variable, différentielle *et* algébrique.

4.3.1 Développement asymptotique de l'erreur globale pour les IRK

Comme pour l'indice 1, on s'intéresse aux IRK vérifiant la condition $C(q)$. Afin de pouvoir préciser l'allure du développement asymptotique de l'erreur globale, une hypothès supplémentaire sur la méthode est nécessaire, dite *condition* $B(p)$ et définie par

$$B(p): \qquad \sum_{i=1}^{s} \underline{b}_i \underline{c}_i^{k-1} = \frac{1}{k} \qquad \text{pour } k = 1, \ldots, p$$

Le théorème suivant [12] précise l'allure du développement asymptotique en y et en z pour les EAD semi-explicites d'indice 2 de la forme (4.1).

Théorème 9 *On suppose qu'on intègre le système (4.1) avec un ensemble de conditions intiales consistantes et avec une méthode IRK d'ordre p, de matrice de coefficients $\underline{A}$ inversible, vérifiant $|R(\infty)| < 1$ et satisfaisant $B(p)$, $C(q)$, ($p \geq q+1$, $q \geq 1$). On a alors*

$$\begin{aligned} y_n - y(t_n) &= h^{q+1}\left(a_{q+1}(t_n) + \alpha_n^p\right) + \ldots + h^N\left(a_N(t_n) + \alpha_n^N\right) + \mathcal{O}(h^{N+1}) \\ z_n - z(t_n) &= h^{q+1}\left(b_{q+1}(t_n) + \beta_n^r\right) + \ldots + h^N\left(b_N(t_n) + \beta_n^N\right) + \mathcal{O}(h^{N+1}) \end{aligned}$$

où les perturbations α_n^k et β_n^k sont indépendantes du pas et telles

$$\alpha_n^k, \beta_n^k = \mathcal{O}(\alpha^n) \quad \forall \alpha \quad \in \quad]\,|R(\infty)|, 1\,[$$

Si $R(\infty) = 0$, ces perturbations sont nulles pour tout $n \geq 1$ et tout k.

D'après ce théorème et d'après la propriété 2, le cas particulier des méthodes SAIRK conduit donc à des développements asymptotiques *non perturbés*, comme pour l'indice 1 mais, même pour une SAIRK, l'équation variationnelle donnant les fonctions $a_k(t)$ est plus compliquée que celle obtenue pour les EDO (et donc l'indice 1). Hairer et al. [12] montrent qu'on obtient non plus une équation différentielle en $a_k(t)$ mais en $(Pa_k)(t)$ où P est la projection définie par

$$P(t) = I - Q(t)\,, \quad Q(t) = (f_z(g_y f_z)^{-1} g_y)(y(t), z(t))$$

En effet, les fonctions $P\,a_k$ sont solution de l'EDO

$$(P\,a_k)' = P\,a_k + P'\,(P\,a_k) + P'\,(I - P\,a_k)$$

Les b_k sont définies de la même façon. Si on suppose les conditions $B(p)$ et $C(q)$ vérifiées, on a donc pour l'indice 2 un seul type de développement asymptotique, de même valuation pour y et z. Pourtant, les méthodes utilisées en pratique donnent un ordre plus faible pour la variable algébrique. Pour les méthodes RADAU IIA(s) ($s = 5$ dans RADAU5), par exemple, l'ordre de y est $2s - 1$ alors qu'il n'est que s pour z.

4.3.2 Extension des techniques d'estimation pour les IRK

• Estimateur RS

On le définit comme pour l'indice 1 et son ordre relatif est 1 quelle que soit l'IRK utilisée. En effet, si la méthode est une SAIRK, les perturbations sont nulles (car $R(\infty) = 0$). Sinon, on reprend la preuve du théorème 6.

• Estimateurs ZD et SC

Comme la contrainte algébrique ne dépend pas de z, il faut définir ZD et SC comme si l'EAD était implicite. On construit donc le problème perturbé

$$\begin{aligned} \hat{y}' &= f(\hat{y}, \hat{z}) + d_{1,h}(t)\,, \quad d_{1,h}(t) = P'_{1,h}(t) - f(P_{1,h}(t), P_{2,h}(t)) \\ 0 &= g(\hat{y}) + d_{2,h}(t)\,, \quad d_{2,h}(t) = -g(P_{1,h}(t)) \end{aligned}$$

Pour SC, on a également

$$\begin{aligned} \varepsilon_{1,h} &= P_{1,h} - y \\ \varepsilon_{2,h} &= P_{2,h} - z \end{aligned}$$

$$\begin{aligned} \varepsilon'_{1,h} &= P_{1,h}(t) - f(P_{1,h}(t) - \varepsilon_{1,h}, P_{2,h}(t) - \varepsilon_{2,h}) \\ 0 &= g(P_{1,h}(t) - \varepsilon_{1,h}, P_{2,h}(t) - \varepsilon_{2,h}) \end{aligned}$$

On peut donc s'attendre, d'après les remarques faites lors de l'extension des estimateurs aux EAD d'indice 1, à observer une réduction d'ordre de ZD et SC pour la variable algébrique z et ce même pour les méthodes de type SAIRK. La question est alors de savoir si la réduction d'ordre sur l'estimation pour z est la même que celle sur l'ordre de l'erreur globale en z.

4.4 Résultats numériques

On présente ici 3 exemples d'EAD semi-explicites d'indice 2. La formulation en indice 2 du pendule non linéaire (dont les tests ont été présentés au début de cette partie) conduit à un système de dimension 5. Le deuxième exemple est une EAD de dimension 7 permettant de simuler la dynamique d'un fluide utilisé en pétrochimie. Enfin, on étudie l'EAD réduite d'une EDO raide de dimension 15 modélisant un petit circuit électrique. Le but de ces tests est avant tout de montrer que le passage de l'indice 1 à l'indice 2 ne pose pas de problème en pratique et qu'on peut encore espérer obtenir, même pour des EAD ayant un intérêt réel, des estimations fiables avec les trois algorithmes étudiés.

$$\text{système 7.} \qquad \begin{cases} z_1' = z_3\,, & z_1(0) = 1 \\ z_2' = z_4\,, & z_2(0) = 0 \\ z_3' = -z_1 z_5\,, & z_3(0) = 0 \\ z_4' = -z_2 z_5 - g\,, & z_4(0) = 0 \\ 0 = z_1 z_3 + z_2 z_4\,, & z_5(0) = 1 \end{cases} \qquad \text{sur } [\,0:30\,]$$

$$\text{système 8.} \qquad MY' = f(Y)\,, \quad Y(0) \in \mathbb{R}^7 \qquad \text{sur } [\,0:40\,]$$

où M est une matrice constante singulière de dimension 7, de la forme

$$M = \begin{pmatrix} 1 & & & & & & \\ & 1 & \frac{1}{15} & & & & \\ & & 0 & 1 & & & \\ & & & 0 & & & \\ & & & & 0 & & \\ & & & & & 0 & \\ & & & & & & 0 \end{pmatrix}$$

Comme pour le système 5, cette EAD n'est pas semi-explicite mais implicite. On considère enfin l'EDO raide définie comme suit.

$$\text{système 9.} \qquad MY' = f(Y)\,, \quad Y(0) \in \mathbb{R}^{15} \qquad \text{sur } [\,0:2.10^{-4}\,]$$

où M est la matrice diagonale de dimension 15 suivante

$$M = Diag\,(C, C, C_S, C_S, C_S, C_S, C_P, L_h, L_h, L_{S2}, L_{S3}, L_{S2}, L_{S3}, L_{S1}, L_{S1})$$

En posant $C_S := 0$, on obtient une EAD semi-explicite d'indice 2.

Là encore, la solution exacte de ces problèmes n'est pas connue et on en calcule une approximation de la même façon que pour les exemples précédents.

On examine ici les systèmes 8 et 9, le pendule d'indice 2 (système 7) ayant déj'a été testé. De plus, les tests sont menés à pas constant (pour disposer facilement d'une approximation de l'erreur globale). A pas variable, il est possible de supposer que RS est suffisamment fiable et ainsi de vérifier que ZD et SC donnent des résultats cohérents (i.e. ne s'écartent pas trop de l'estimation RS).

La figure 4.4 montre qu'on retrouve pour le système 8 des observations faites dans le cas de l'indice 1 : le pas étant assez grand (h = 0.2), RS est meilleur que ZD et SC qui ont plus de mal à suivre le pic de l'erreur globale en t = 10.

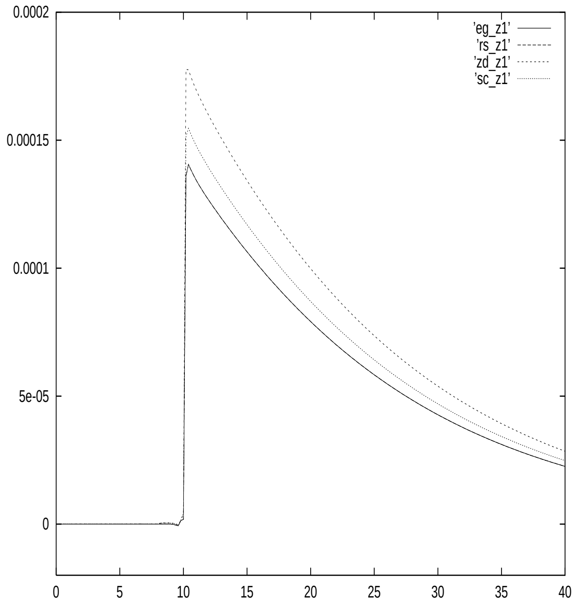


FIG. 4.4 – *Système 8 : à pas constant* $h = 0.2$ *sur* $[\,0 : 40\,]$

On considère ensuite la variable y_3 pour laquelle l'erreur globale présente un fort transitoire toujours vers t = 10 : apparemment l'estimation est satisfaisante (fig. 4.5).

L'étude séparée des trois parties de la courbe (avant/ pendant/après le transitoire) montre que lorsque l'erreur globale est régulière, ZD ≡ SC sont meilleurs que RS (fig. 4.6-a) alors ZD et SC perdent un peu en précision dans la zone 'agitée' (fig. 4.6-b). Mais après le transitoire, ZD et SC ne parviennent pas à donner une meilleure estimation que RS, bien que l'erreur globale soit aussi régulière qu'au début (fig. 4.6-c).

En passant de h=0.2 à h=0.1, on espère améliorer la situation. En fait, l'erreur globale est beaucoup plus irrégulière pour z_1 (fig. 4.7-a) et ZD et SD sont incapables de l'estimer pour $t \geq 10$, ce qui en regard des tests faits en indice 1, n'est guère surprenant.

Pour z_3, les choses vont effectivement mieux et en se restreignant à $[\,8, 12\,]$, on voit que SC estime très précisément, que ZD est à peine moins bon et que tous deux sont bien meilleurs que RS (fig. 4.7-b).

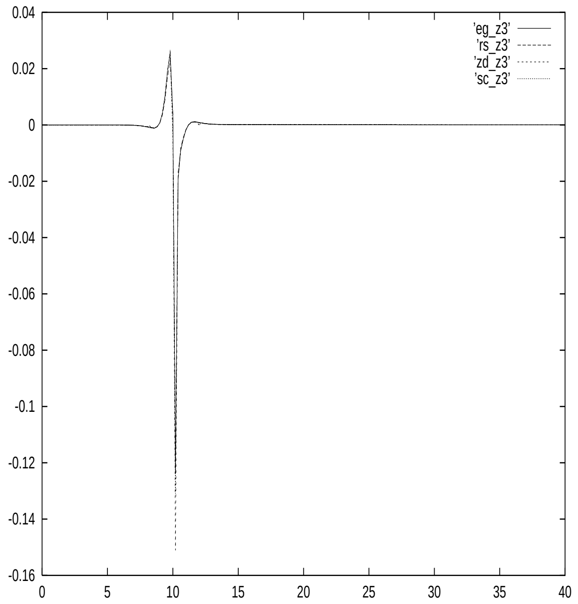


FIG. 4.5 – *Système 8 : à pas constant* $h = 0.2$ *sur* $[\,0 : 40\,]$

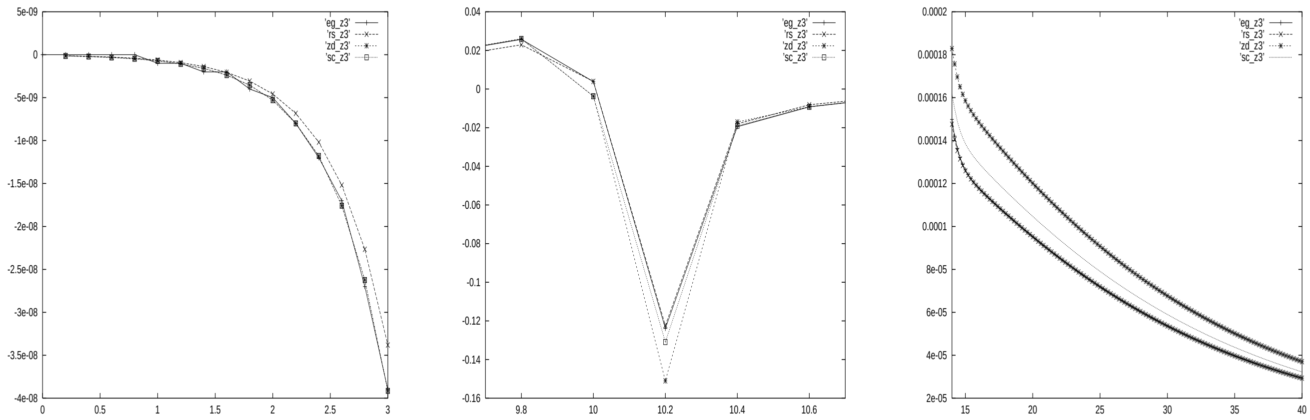


FIG. 4.6 – *Système 8 : à pas constant* $h = 0.2$ *sur* $[\,0, 40\,]$

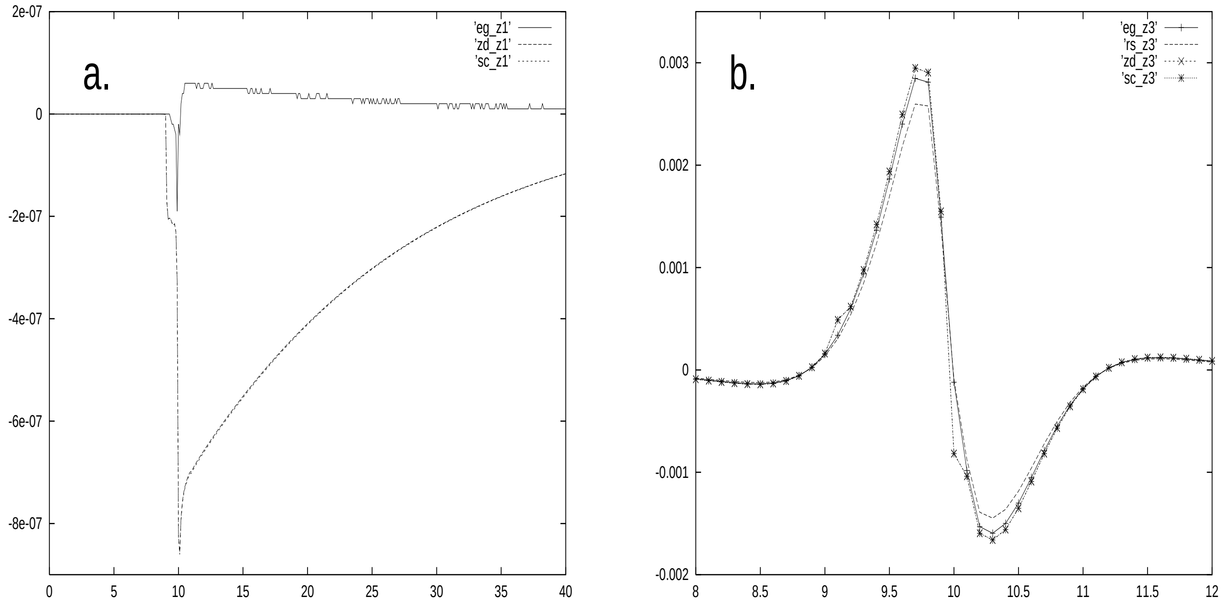


FIG. 4.7 – *Système 8 : à pas constant* $h = 0.1$ *sur* $[\,0, 40\,]$

Enfin le système 9 permet de retrouver l'intérêt d'une bonne estimation de l'erreur globale déjà mis en évidence sur le pendule d'indice 1.

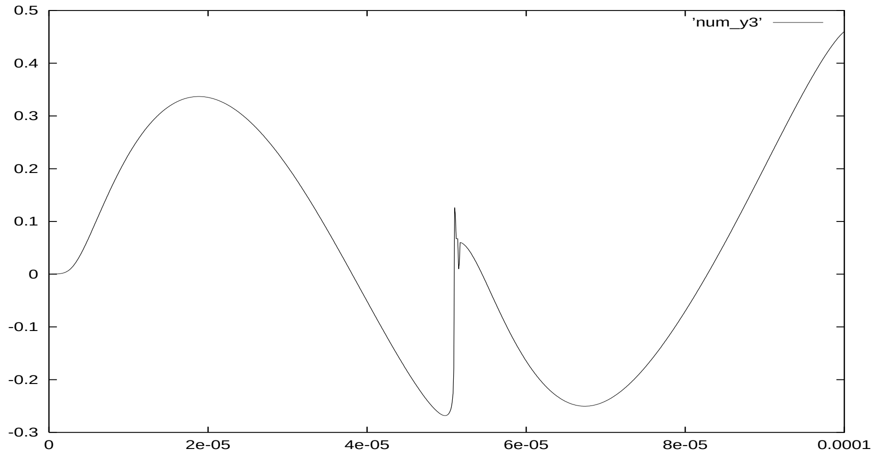


FIG. 4.8 – *Système 9 : à pas constant* $h = 2 \cdot 10^{-7}$ *sur* $[\,0, 2 \cdot 10^{-4}\,]$

En effet, l'intégration numérique à pas constant h $= 2{\cdot}10^{-7}$ sur $[\,0, 2{\cdot}10^{-4}\,]$ diverge vers $t = 10^{-4}$. Par exemple, la solution numérique pour la variable y_3 (fig. 4.8) présente une premier pic vers $t = 5 \cdot 10^{-5}$ avant de diverger complètement à partir de $t = 10^{-4}$.

Comme pour le pendule d'indice 1, l'erreur globale prévoit en quelque sorte l'instabilité numérique en prenant des valeurs relativement grandes dès $t = 5.09 \cdot 10^{-5}$. En effet, si on ajuste les paramètres (réduction du pas notamment) de façon à ce que l'intégration se passe bien sur $[\,0, 2 \cdot 10^{-4}\,]$, l'erreur globale ne diverge plus. Il s'avère que pour les 3 algorithmes, l'estimation calculée, même si elle ne suit pas parfaitement la courbe de l'erreur globale, reflète bien le comportement de celle-ci (fig. 4.9).

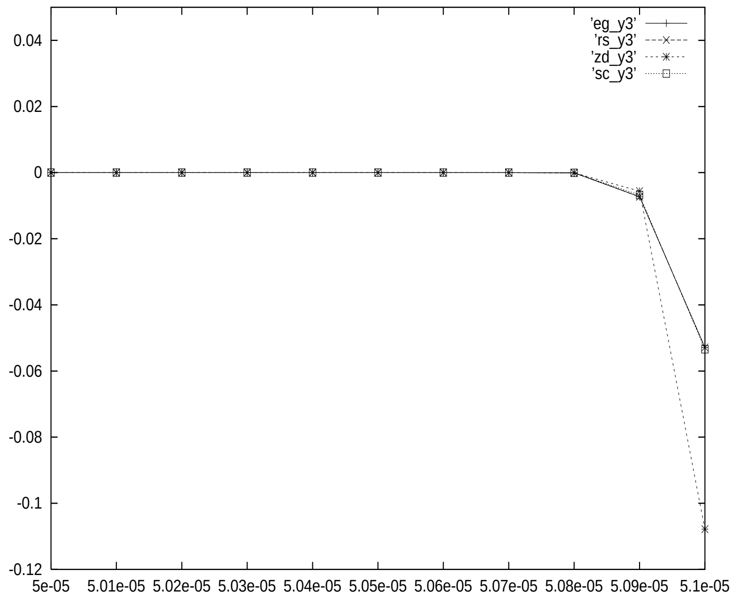


FIG. 4.9 – *Système 9 : à pas constant* $h = 2 \cdot 10^{-7}$ *sur* $[\,0, 2 \cdot 10^{-4}\,]$

Conclusion

On retrouve donc les principaux points vus pour l'indice 1 concernant le comportement des trois estimateurs lorsque l'erreur est très régulière ou présente de grosses irrégularités. La possibilité de *prédire* une éventuelle instablilité numérique apparaît également ici. Notons enfin qu'à pas variable, l'erreur globale peut présenter des oscillations très rapides et brutales (comme celles observées en indice avec le système 5) qui mettent en défaut les trois algorithmes.

Chapitre 5

Application aux problèmes à perturbation singulière

On étudie dans cette partie le lien entre les EAD semi-explicites d'indice 1 et certaines EDO raides, les problèmes à perturbation singulière. Souvent, dans les applications, les EAD viennent en fait d'EDO raides dont la raideur a été supposée infinie comme hypothèse simplificatrice du problème physique. Réciproquement, on peut chercher à construire, pour une EAD donnée, sa *régularisation*, c'est-à-dire une EDO raide dont la solution analytique tend vers celle de l'EAD lorsque la raideur μ tend vers l'infini (en pratique, on considère plutôt l'inverse de la raideur $\epsilon = \frac{1}{\mu}$ qu'on fait tendre vers zéro). D'un point de vue purement numérique, la question qui se pose alors est la suivante [12]: quel lien existe-t-il entre l'ordre d'une méthode numérique appliquée à une EAD et celui de la même méthode appliquée à sa régularisation, i.e. dans quelle mesure la solution numérique de l'EAD tend-elle vers celle de l'EDO associée? A partir des réponses apportées par Hairer et al. dans [12], il est alors légitime de chercher à étendre la question à l'estimation de l'erreur globale.

5.1 Un exemple : l'équation de Van der Pol

Les problèmes à pertubation singulière sont en général caractérisés par la présence d'un paramètre ϵ très petit (ou μ très grand) et tels que la dérivée d'ordre maximum disparaît lorsque $\epsilon \to 0$. De tels systèmes présentent des instabilités observables, par exemple, sous forme d'oscillations. La zone où se trouvent les instabilités est appelée *transitoire*. D'un point de vue purement numérique, il est d'usage de considérer qu'une équation est raide lorsque certaines méthodes implicites (BDF notamment) marchent beaucoup mieux que les méthodes explicites [17]. On va voir que les EAD semi-explicites d'indice un sont en fait un cas limite de ce type de problèmes. Pour faire apparaître un tel problème comme la régularisation possible d'une EAD semi-explicite, il faut réécrire l'EDO de départ de façon à la rendre elle aussi semi-explicite. Considérons par exemple l'équation de Van der Pol [14]

$$\epsilon z'' + (z^2 - 1)z' + z = 0$$

Grâce à l'identité

$$\epsilon z'' + (z^2 - 1)z' = \frac{d}{dx}\left(\epsilon + \left(\frac{z^3}{3} - z\right)\right)$$

et au changement de variable $y = \epsilon + \left(\frac{z^3}{3} - z\right)$, l'équation (5.1) se réécrit de de la façon suivante

$$\begin{array}{rcl} y' & = & -z \\ \epsilon z' & = & y - \left(\frac{z^3}{3} - z\right) \end{array}$$

Le problème de départ a donc été mis sous la forme semi-explicite (3.7). Cependant les fonctions f et g obtenues ne sont pas uniques. On aurait pu faire un autre changement de variable, $y' = z$ par exemple, et obtenir (3.7) avec d'autres fonctions f et g.

5.2 Développement asymptotique de l'erreur globale pour les problèmes à perturbation singulière

On s'intéresse désormais au comportement de la *solution numérique* lorsque $\epsilon \to 0$. Soit un problème à perturbation singulière sous forme semi-explicite du type (3.7). On suppose de plus que f, g et les conditions initiales $y(0)$ et $z(0)$ ne dépendent pas de ϵ. On appelle *équation réduite* associée à (3.7), l'EAD semi-explicite de type (1.1) obtenue en posant $\epsilon = 0$. On se place toujours dans le cas d'indice 1 : g_z est inversible dans un voisinage de la solution de (1.1).

Développement des solutions analytiques régulières
Avant de traiter le cas de la solution numérique, on étudie la structure de la solution analytique de (3.7), notamment lorsque celle-ci est *régulière*, c'est-à-dire de la forme

$$\begin{array}{l} y(t) = y_0(t) + \epsilon y_1(t) + \epsilon^2 y_2(t) + \dots \\ z(t) = z_0(t) + \epsilon z_1(t) + \epsilon^2 z_2(t) + \dots \end{array} \tag{5.1}$$

où y_k, z_k sont des coefficients *réguliers* indépendants de ϵ. De telles solutions existent, pourvu que l'on se place sur un intervalle borné ne contenant pas le transitoire.
En insérant (5.1) dans (3.7) et en comparant les puissances de ϵ, on trouve que les coefficients des développements sont solution des EAD suivantes

$$\epsilon^0 : \quad \begin{array}{rcl} {y_0}' & = & f(y_0, z_0) \\ 0 & = & g(y_0, z_0) \end{array} \tag{5.2}$$

$$\epsilon^1 : \quad \begin{array}{rcl} {y_1}' & = & f_y(y_0, z_0) y_1 + f_z(y_0, z_0) z_1 \\ {z_0}' & = & g_y(y_0, z_0) y_1 + g_z(y_0, z_0) z_1 \end{array} \tag{5.3}$$

$$\dots$$

$$\epsilon^\nu : \quad \begin{array}{rcl} {y_\nu}' & = & f_y(y_0, z_0) y_\nu + f_z(y_0, z_0) z_\nu + \varphi_\nu(y_0, z_0, \dots, y_{\nu-1}, z_{\nu-1}) \\ {z_{\nu-1}}' & = & g_y(y_0, z_0) y_\nu + g_z(y_0, z_0) z_\nu + \psi_\nu(y_0, z_0, \dots, y_{\nu-1}, z_{\nu-1}) \end{array} \tag{5.4}$$

En particulier, on voit d'après (1.1) que $y_0(x)$, $z_0(x)$ est solution de l'EAD réduite (1.1). Puisque g_z est inversible, la deuxième équation de (5.3) donne z_1 en fonction de ${z_0}'$ et y_1. En reportant cette expression dans la première équation de (5.3), on obtient une équation différentielle en $y_1(t)$. On peut donc déterminer $y_1(t)$ et $z_1(t)$. De la même façon, on obtient $y_2(t)$ et $z_2(t)$ à partir de l'EAD associée à ϵ^2, etc. Ce procédé de construction des coefficients de (5.1) montre qu'on peut choisir les valeurs initiales $y_j(0)$ arbitrairement mais pas les $z_j(0)$. Par conséquent, toute solution de (3.7) ne peut pas s'écrire sous la forme (5.1), ce qui justifie l'hypothèse de départ.

Développement de la solution numérique

Après avoir considéré la solution exacte de (3.7), on étudie maintenant sa solution numérique et on présente un résultat de convergence dans le cas des méthodes IRK.
Considérons la méthode

$$\begin{pmatrix} y_{n+1} \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} y_n \\ z_n \end{pmatrix} + h \sum_{i=1}^{s} \underline{\mathrm{b}}_i \begin{pmatrix} k_{ni} \\ l_{ni} \end{pmatrix} \tag{5.5}$$

où

$$\begin{pmatrix} k_{ni} \\ \epsilon l_{ni} \end{pmatrix} = \begin{pmatrix} f(Y_{ni}, Z_{ni}) \\ g(Y_{ni}, Z_{ni}) \end{pmatrix} \tag{5.6}$$

et où les points intermédiaires sont donnés par

$$\begin{pmatrix} Y_{ni} \\ Z_{ni} \end{pmatrix} = \begin{pmatrix} y_n \\ z_n \end{pmatrix} + h \sum_{j=1}^{s} \underline{\mathrm{a}}_{ij} \begin{pmatrix} k_{nj} \\ l_{nj} \end{pmatrix} \tag{5.7}$$

On se place encore dans une zone ne contenant pas le transitoire et on suppose qu'avec la tolérance fixée la méthode a permis d'approcher la solution régulière. On peut donc considérer que les valeurs initiales se trouvent sur la solution régulière (i.e. de la forme (5.1)). Par ailleurs, dans la phase transitoire, la méthode doit prendre des pas de taille $\mathcal{O}(\epsilon)$, alors qu'en dehors de cette zone, les pas peuvent être choisis plus grands. On suppose donc que le pas h est grand par rapport à ϵ. On cherche alors un développement en ϵ de la solution numérique. On développe pour cela formellement les quantités apparaissant dans le schéma numérique en puissances de ϵ avec des coefficients indépendants de ϵ.

$$y_n = y_n^0 + \epsilon y_n^1 + \epsilon^2 y_n^2 + \dots \tag{5.8}$$
$$Y_{ni} = Y_{ni}^0 + \epsilon Y_{ni}^1 + \epsilon^2 Y_{ni}^2 + \dots \tag{5.9}$$
$$k_{ni} = k_{ni}^0 + \epsilon k_{ni}^1 + \epsilon^2 k_{ni}^2 + \dots \tag{5.10}$$

Les mêmes développements sont également valables pour z_n, Z_{ni} et l_{ni}. Par linéarité de (5.5) et (5.7), on a

$$\begin{pmatrix} y_{n+1}^\nu \\ z_{n+1}^\nu \end{pmatrix} = \begin{pmatrix} y_n^\nu \\ z_n^\nu \end{pmatrix} + h \sum_{i=1}^{s} b_i \begin{pmatrix} k_{ni}^\nu \\ l_{ni}^\nu \end{pmatrix}$$

et

$$\begin{pmatrix} Y_{ni}^\nu \\ Y_{ni}^\nu \end{pmatrix} = \begin{pmatrix} y_n^\nu \\ z_n^\nu \end{pmatrix} + h \sum_{j=1}^{s} a_{ij} \begin{pmatrix} k_{nj}^\nu \\ l_{nj}^\nu \end{pmatrix}$$

En insérant (5.9,5.10) dans (5.6) et par comparaison des puissances de ϵ, on obtient

$$\epsilon^0 : \quad \begin{array}{rcl} {k^0_{ni}}' &=& f(Y^0_{ni}, Z^0_{ni}) \\ 0 &=& g(Y^0_{ni}, Z^0_{ni}) \end{array} \tag{5.11}$$

$$\epsilon^1 : \quad \begin{array}{rcl} {k^1_{ni}}' &=& f_y(Y^0_{ni}, Z^0_{ni})Y^1_{ni} + f_z(Y^0_{ni}, Z^0_{ni})Z^1_{ni} \\ {l^0_{ni}}' &=& g_y(Y^0_{ni}, Z^0_{ni})Y^1_{ni} + g_z(Y^0_{ni}, Z^0_{ni})Z^1_{ni} \end{array} \tag{5.12}$$

$\ldots$

$$\epsilon^\nu : \quad \begin{array}{rcl} {k^\nu_{ni}}' &=& f_y(Y^0_{ni}, Z^0_{ni})Y^\nu_{ni} + f_z(Y^0_{ni}, Z^0_{ni})Z^\nu_{ni} + \varphi_\nu(Y^0_{ni}, Z^0_{ni}, \ldots, Y^{\nu-1}_{ni}, Z^{\nu-1}_{ni}) \\ {l^{\nu-1}_{ni}}' &=& g_y(Y^0_{ni}, Z^0_{ni})Y^\nu_{ni} + g_z(Y^0_{ni}, Z^0_{ni})Z^\nu_{ni} + \psi_\nu(Y^0_{ni}, Z^0_{ni}, \ldots, Y^{\nu-1}_{ni}, Z_{\nu-1}) \end{array} \tag{5.13}$$

Comme (5.6) a la même forme que (3.7), les formules de (5.11)-(5.13) sont les mêmes que celles de (5.2)-(5.4). Cela peut s'interpréter de la façon suivante : les coefficients y^0_n, z^0_n, y^1_n, z^1_n, $\ldots$ représentent la solution numérique de la méthode de Runge-Kutta (5.5)-(5.7) appliquée aux EAD (5.2)-(5.4). En particulier, y^0_n, z^0_n est la solution numérique de l'EAD reduite (1.1).

Développement asymptotique de l'erreur globale
Formellement, les équations (5.8) et (5.1) donnent par soustraction

$$y_n - y(t_n) = \sum_{\nu \geq 0} \epsilon^\nu \left(y^\nu_n - y_\nu(t_n)\right) \tag{5.14}$$

$$z_n - z(t_n) = \sum_{\nu \geq 0} \epsilon^\nu \left(z^\nu_n - z_\nu(t_n)\right) \tag{5.15}$$

A partir de l'étude des différences $\Delta y^\nu_n = y^\nu_n - y_\nu(t_n)$ et $\Delta z^\nu_n = z^\nu_n - z_\nu(t_n)$, Hairer et al. [14] donnent une estimation rigoureuse des restes de (5.14,5.15).

Théorème 10 *On considère le problème raide (3.7) ayant $y(0)$, $z(0)$ pour conditions initiales consistantes. On suppose que ce problème admet une solution régulière. Si on lui applique une méthode IRK (5.5)-(5.7) d'ordre p à q points $(1 \leq q < p)$ et si l'on suppose que cette méthode est A-stable, que la fonction de stabilité vérifie $|R(\infty)| < 1$ et que les valeurs propres de la matrice $\underline{A}$ des coefficients de la méthode sont à partie réelle positive, alors, pour toute constante $c > 0$ fixée, l'erreur globale vérifie pour $\epsilon \leq ch$ et $\nu \leq q + 1$*

$$\begin{array}{l} y_n - y(t_n) = \Delta y^0_n + \epsilon \Delta y^1_n + \ldots + \epsilon^\nu \Delta y^\nu_n + \mathcal{O}(\epsilon^{\nu+1}) \\ z_n - z(t_n) = \Delta z^0_n + \epsilon \Delta z^1_n + \ldots + \epsilon^\nu \Delta z^\nu_n + \mathcal{O}(\epsilon^{\nu+1}) \end{array} \tag{5.16}$$

L'estimation (5.16) est valable uniformément pour $h \leq h_0$ et $nh \leq Const.$

On voit donc qu'à un instant t_n donné et pour un pas h fixé, l'erreur globale d'un problème à perturbation singulière converge vers l'erreur globale de son EAD réduite lorsque $\epsilon \to 0$.

5.3 Continuité de l'estimation de l'erreur globale pour les IRK

- **Estimateur de Richardson**

D'après le théorème 10, on a

$$\frac{y_n - y^*_{2n}}{1 - 2^{-p}} \rightarrow \frac{y^0_n - y^{0,*}_{2n}}{1 - 2^{-p}} \quad \text{quand} \quad \epsilon \to 0$$

On a la même relation pour la variable algébrique z. Ainsi, pour une méthode IRK et un pas h donnés, l'estimation de RS pour l'EDO (3.7) tend vers l'estimation de RS pour l'EAD réduite lorsque $\epsilon \to 0$.

On illustre cette propriété avec le système 1 dont la régularisation est

$$\begin{aligned} y' &= y + z^2 \\ \epsilon z' &= y - z^2 \end{aligned}$$

On intègre cette EDO raide à pas constant pour différentes valeurs de ϵ et on observe la convergence de l'estimation de l'erreur globale vers celle obtenue sur le système 1 lorsque $\epsilon \to 0$. La figure 5.1-a montre les cas où $\epsilon \in \{10^{-1}, 10^{-2}, 10^{-3}\}$ avec RS. Pour $\epsilon = 10^{-5}$, la limite est quasiment atteinte (fig. 5.1-b). Avec $\epsilon = 10^{-6}$, on améliore définitivement la convergence (fig. 5.1-c).

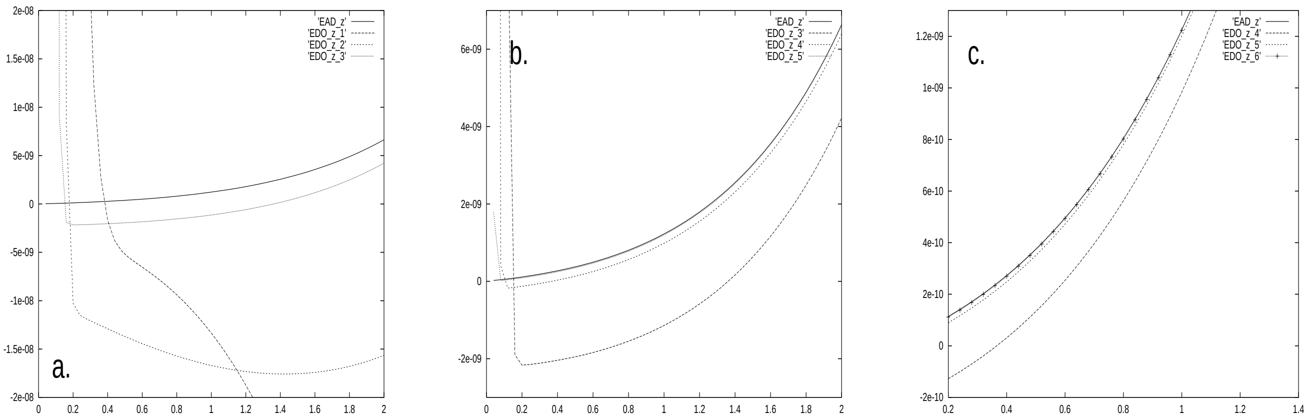


FIG. 5.1 – *RS sur le système 1 et son EDO sous-jacente : pas constant* $h = 0.04$ *sur* $[0, 2]$

Le système 9 constitue un exemple plus intéressant (bien que d'indice 2 et pas 1) puisqu'il a un sens même lorsque C_S (qui joue ici le rôle de ϵ) est très petit [12].
On considère la partie $[0, 5 \cdot 10^{-5}]$ qui précède le transitoire et présente la régularité requise pour appliquer le théorème 10. L'EDO a été intégrée avec $C_S = 10^{-9}$ et $C_S = 10^{-12}$ et on compare pour ZD et SC les estimations de l'erreur globale obtenues pour la variable $U_1(t)$ (désignée par y_1 sur les graphiques) avec ces deux valeurs de ϵ et celles obtenues avec l'EAD réduite ($C_S := 0$). Avec $C_S = 10^{-9}$, l'EAD ne donne pas une bonne approximation l'estimation de l'erreur globale obtenue pour sa régularisation. Mais pour $C_S = 10^{-12}$ la convergence est quasiment atteinte (fig. 5.2). On confirme donc que pour RS, poser $C_S := 0$ est une simplification qui préserve l'estimation de l'erreur globale dès que l'EDO de départ a un paramètre ϵ suffisamment petit.

- **Estimateurs ZD et SC**

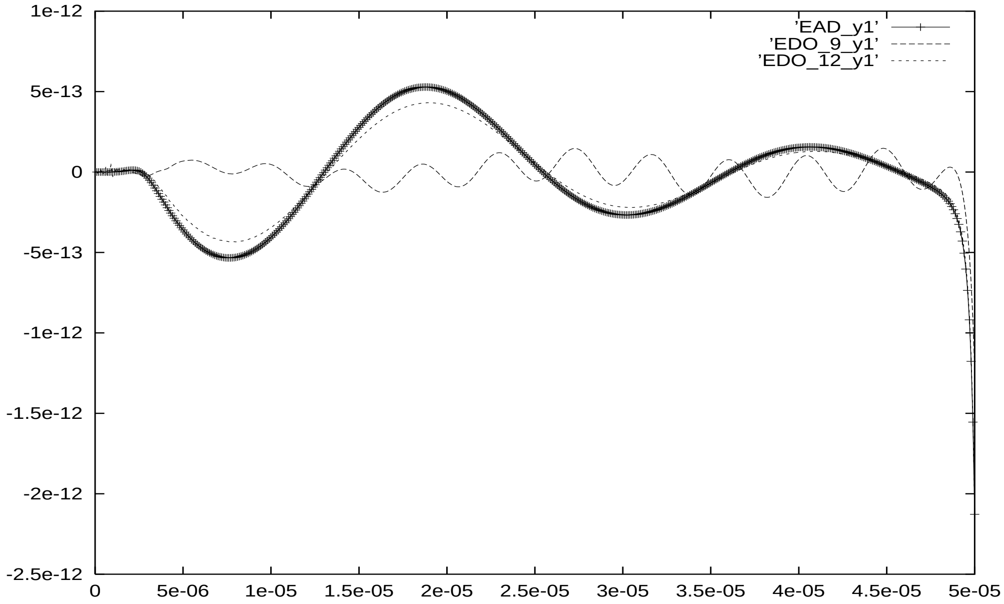


FIG. 5.2 – *RS sur le système 9 : pas constant* $h = 10^{-7}$ *sur* $[\,0, 2 \cdot 10^{-4}\,]$

Pour ces deux estimateurs, la continuité de l'estimation en ϵ n'est plus triviale. Dans le cas de ZD, une façon de faire pourrait être de construire deux problèmes perturbés, l'un associé à l'EDO raide et dépendant de ϵ, l'autre associée à l'EAD réduite et défini comme d'habitude pour les EAD. Une autre possibilité, qui a l'avantage de convenir pour tous les estimateurs, serait d'étudier les développements asymptotiques des deux problèmes : peut-on conclure à une convergence des termes a_k^ϵ associés à l'EDO raide vers les a_k associés à l'EAD réduite quand $\epsilon \to 0$?
Les tests qui suivent laissent supposer que c'est le cas puisqu'on observe les mêmes phénomènes qu'avec RS.

Sur l'exemple 1, le passage à la limite pour ZD et SC a lieu également pour $\epsilon = 10^{-6}$. De plus, le système 9, testé dans les mêmes conditions qu'avec RS, donne également des résultats très satisfaisants (fig. 5.3).

On voit que ZD ayant convergé, l'estimation de l'erreur globale est la même que pour RS et que le passage à la limite se fait de la même façon. L'algorithme SC donne aussi les mêmes résultats.

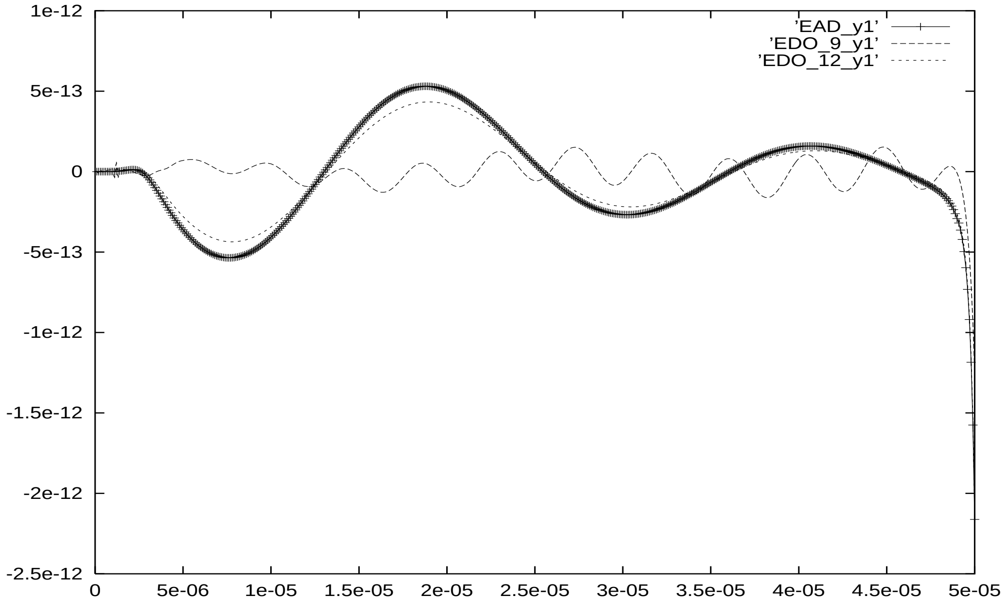


FIG. 5.3 – *ZD sur le système 9 : pas constant* $h = 10^{-7}$ *sur* $[\,0, 2 \cdot 10^{-4}\,]$

Chapitre 6

Conclusion

Cette étude, initialement motivée par le besoin d'estimer précisément l'erreur globale lors de l'étude de la dynamique des réseaux électriques par EDF, a permis d'envisager trois techniques d'estimation pour les EAD semi-explicites d'indice 1 et 2. On a vu que le choix de la méthode d'intégration est capital car il conditionne d'une part les propriétés théoriques des estimateurs et d'autre part leur implémentation.

Les propriétés de RS, ZD et SC rappelées pour les EDO ont été généralisées au cas de l'indice 1 pour des IRK quelconques. Puisque lorsque l'indice est 1, résoudre une EAD semi-explicite revient à résoudre une EDO en y et à en déduire z par la méthode de Newton, le travail a porté uniquement sur la variable algébrique z. Les SAIRK permettent pour chaque estimateur de retrouver exactement le même ordre relatif que dans le cas des EDO alors que les autres IRK entraînent une réduction d'ordre, sauf pour RS. Ceci est vrai lorsqu'on applique les estimateurs aux EAD de la même façon qu'aux EDO. On a cependant mis en évidence une adaptation possible de ZD et SC qui évite cette réduction d'ordre et s'applique aux systèmes semi-explicites. Les test réalisés avec la SAIRK de RADAU5 montrent toutefois que l'estimation n'est pas toujours aussi précise qu'en théorie. En effet, le choix d'une méthode d'intégration implicite rend ZD moins fiable que SC et RS. (l'erreur due à la méthode de Newton n'est pas prise en compte par ZD). De plus, certaines solutions numériques très instables empêchent toute estimation de l'erreur globale qui présente elle-même dans ce cas de fortes oscillations. Ces deux points consituent la principale originalité par rapport aux EDO. Dans le cas des méthodes multipas, on a choisi de s'intéresser aux BDF-k pour lesquelles seul l'ordre de RS est connu pour les EDO. Son extension à l'indice 1 a été précisée et implémentée dans DASSL. Les tests réalisés ont alors permis de retrouver les observations faites sur les EDO. ZD et SC n'ont quant à eux fait encore l'objet d'aucun développement pour ce type de méthodes, qu'il s'agisse d'EDO ou d'EAD et une étude théorique approfondie semble indispensable avant toute mise en oeuvre.

Pour l'indice 2, l'expérience montre que l'estimation de l'erreur globale ne pose pas plus de problèmes que pour l'indice 1. Pourtant, la théorie est ici plus difficile à établir. Bien que RS soit encore d'ordre relatif 1, il semblerait que pour une IRK quelconque, ZD et SC perdent en précision par rapport à l'indice 1. Seul le cas de RS a été démontré, les deux autres techniques exigeant une étude précise du développement asymptotique de chacune des variables, différentielle et algébrique. Il est en effet désormais impossible de se ramener à une EDO en la variable différentielle.

Enfin, l'étude des problèmes à perturbation singulière dans le cas des IRK met en évidence un lien étroit entre l'estimation de l'erreur globale lors de l'intégration de tels problèmes et celle de l'erreur globale commise lors de l'intégration des EAD d'indice 1. En effet, l'estimation de RS obtenue pour un problème à perturbation singulière de paramètre ϵ tend vers celle obtenue pour l'EAD réduite lorsque $\epsilon \to 0$. Comme on préfère souvent réduire la dimension d'un système différentiel quitte à faire apparaître une ou plusieurs contraintes algébriques, il était important de savoir si ce genre de simplification dans la formulation du problème étudié ne nuit pas à la qualité de l'estimation. Si le problème a été résolu pour RS, il reste à établir une éventuelle continuité de l'estimation de l'erreur globale pour ZD et SC.

Ce travail a permis certes d'apporter quelques réponses au problème de l'estimation de l'erreur globale pour les EAD d'indice 1 et 2 mais a surtout soulevé plusieurs questions. En particulier, qu'en est-il des estimateurs ZD et SC pour les méthodes BDF-k? Il serait raisonnable de commencer par considérer le cas des EDO avant d'envisager une extension aux EAD semi-explicites. Pour les méthodes à un pas, le cas de l'indice 2 ouvre des perspectives intéressantes et nécessitera sans doute des idées plus novatrices dans la mesure où la théorie de l'estimation de l'erreur globale pour les EDO ne s'applique plus. La réduction d'ordre supposée pour ZD et SC est-elle la même que celle concernant l'ordre de la méthode? Par ailleurs, la question de la continuité des estimateurs lorsqu'on remplace un problème à perturbation singulière par son EAD réduite reste posée pour ZD et SC.

Enfin, d'une point de vue pratique, un test 'en vraie grandeur' est à envisager. Nous disposons pour cela d'une maquette de réseau électrique proposée par EDF. Ce problème peut être formulé de différentes façons et conduire notamment à une EAD d'indice 1 ou 2. Il serait alors intéressant de comparer, au moins pour RS, les estimations obtenues dans les deux cas.

Bibliographie

[1] R. Aïd. Asymptotic global error estimation for variable step-size and one-step methodes. Rapport Technique, LMC-IMAG, Grenoble, 1997.

[2] R. Aïd. Estimation de l'erreur globale pour l'intégration numérique d'équations différentielles ordinaires. Rapport Technique, LMC-IMAG, Grenoble, 1996.

[3] R. Aïd and L. Levacher. Numerical investigations on global error estimation for ordinary differential equations. *J. on Comput. and Appl. Math.*, 1996. submitted to.

[4] A. Aubry and P. Chartier. Augmenter l'ordre de convergence des composantes algébriques pour des méthodes RADAU IIA appliquées à des EDA d'indice 2. *Rapport de recherche de l'INRIA*, Rennes, 1995.

[5] K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical Solution of Initial-Value Problems in Differential- Algebraic Equations.* North-Holland, 1989.

[6] J.R. Dormand. *Numerical Methods for Differential Equations.* CRC, 1996.

[7] R. Franck and C. W. Ueberhuber. Iterated defect correction for differential equations, Part I : theoretical results. *Computing*, 20:207–228, 1978.

[8] R. Frank and C. W. Ueberhuber. Iterated defect correction for Runge-Kutta metho ds. Technical Report 14/75, Institut für Numerische Mathematik, T. U. Wien, 1975.

[9] C.W. Gear. *Numerical Initial Value Problems in Ordinary Differential Equations.* Prentice-Hall, 1971.

[10] W. B. Gragg. Repeated extrapolation to the limit in the numerical solution of ordinary differential equations. *Thèse, University of California*, Los Angeles 1964.

[11] E. Hairer and C. Lubich. Asymptotic expansions of the global error of fixed-stepsize methods. *Numer. Math.*, 45:345–360, 1984.

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

[13] E. Hairer, S.P. Norsett, and G. Wanner. *Solving Ordinary Differential Equations I : Non Stiff Problems.* Springer-Verlag, 1980.

[14] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II : Stiff and Differential-Algebraic Problems.* Springer-Verlag, 1991.

[15] A. Iserles. *Numerical Analysis of Differential Equations.* Cambridge University Press, 1996.

[16] L. Levacher and R. Aïd. Etude d'erreurs et de procédés de stabilisation pour la simulation numérique de phénomènes instables (rapport 4). Rapport technique HR-41/96-052, Electricité de France - Direction des Etudes et Recherches, Département MOS, Clamart, France, 1996.

[17] R. O'Malley. *Singular Perturbation Methods for Ordinary Differential Equations.* Springer-Verlag, 1991.

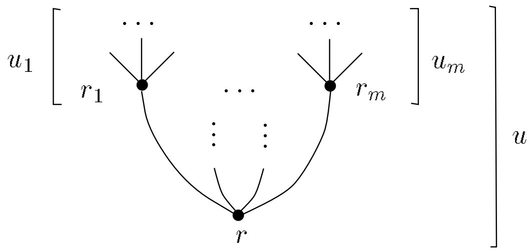
[18] S. Somali and B. Karasözen. Iterated defect correction methods for differential-algebraic equations. *Tr. J. of Mathematics*, 16:230–249, 1992.

Annexe A

Arbres, B-séries et différentielles élémentaires

A.1 Arbres et différentielles élémentaires pous les EDO

On note $\tau = \bullet$ l'arbre réduit à un noeud. Soient $u_1, \ldots, u_m$ des arbres de racines $r_1, \ldots, r_m$ respectivement. On note $u = [\, u_1, \ldots, u_m \,]$ le nouvel arbre obtenu en reliant toutes les racines à un même noeud appelé r. Ce noeud est alors la racine de u.



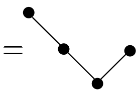
On note $\rho(u)$ le nombre de noeuds de l'arbre u et T l'ensemble des arbres.

Définition 11 *Soit $f : I\!R^N \to I\!R^N$ une application infiniment différentiable. La* différentielle élémentaire de f associée à l'arbre u *est l'application $F(u) : I\!R^N \to I\!R^N$ définie par*

$$F(u)(y) = \frac{d^{(m)}}{dy} [\, F(u_1)(y), \ldots, F(u_m)(y) \,]$$

avec $u = [\, u_1, \ldots, u_m \,]$ et $F(\bullet)(y) = f(y)$.

Exemple : Soit u tel que

$u =$ 

On a $u = [\,u_1, u_2\,]$ avec $u_1 = [\,\bullet\,]$ et $u_2 = \bullet$. Donc

$$\begin{aligned} F(u)(y) &= f^{(2)}(y)\,[\,F(u_1)(y), F(u_2)(y)\,]\,, \\ F(u_1)(y) &= f^{(1)}F(\bullet)(y)\,, \\ F(u_2)(y) &= f(y) \end{aligned}$$

D'où finalement

$$F(u)(y) = f''\,f'\,ff = f_{y^2}\,f_y\,ff$$

• Cas des EDO autonomes

On considère l'EDO autonome

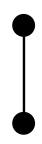
$$y' = f(y) \tag{A.1}$$

Puisque la différentielle élémentaire associée à l'arbre τ est f, on note

$$y' = \bullet$$

En dérivant (A.1) une fois par rapport au temps, il vient

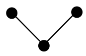
$$\begin{aligned} y'' &= f_y y' \\ &= F([\,\bullet\,])(y) \end{aligned}$$

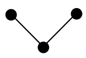
et comme $[\bullet] =$ , on note finalement

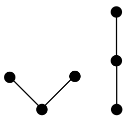
$y'' =$  ou encore 

Une dérivation supplémentaire donne

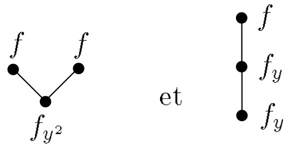
$$\begin{aligned} y''' &= f_{y^2}{y'}^2 + f_y y'' \\ &= f_{y^2} f^2 + f_y f_y f \end{aligned}$$

Soit $u =$ . On a $u = [[\bullet]]$ donc $F(u)(y) = f^{(1)}(y)\,[\,F([\bullet])(y)\,] = f_y f_y f$.
Soit $u =$ . On a $u = [\,\bullet, \bullet\,]$ donc $F(u)(y) = f^{(2)}(y)\,[\,F(\bullet)(y), F(\bullet)(y)\,] = f_{y^2} ff$.

y''' est donc la somme de deux différentielles élémentaires, l'une associée à , l'autre à .
On note alors

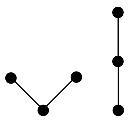
$y''' =$ 

et ces deux arbres s'écrivent aussi :

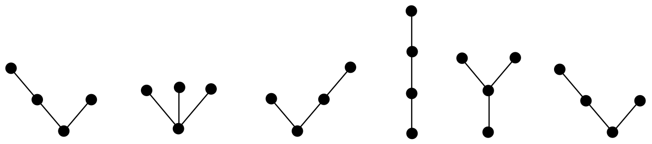


On arrive ainsi à exprimer $y^{(k)}$ comme une somme de différentielles élémentaires associées à des arbres ayant chacun k noeuds. On dispose de plus d'une règle de construction portant sur les arbres et permettant de passer de $y^{(k)}$ à $y^{(k+1)}$ sans aucune différenciation.
Règle de construction : *pour passer des arbres associés à $y^{(k)}$ à ceux associés à $y^{(k+1)}$, il suffit d'ajouter*  *à tous les noeuds.*

Exemple : on a vu que

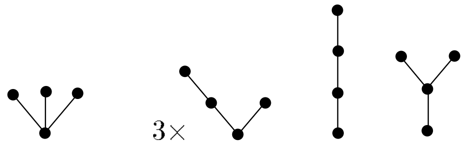
$y^{(3)} =$ 

donc

$y^{(4)} =$ 

(1) (2) (3) (4) (5) (6)

or les arbres (1), (3) et (6) sont topologiquement équivalents donc

$y^{(4)} =$ 

Par le calcul, on doit obtenir les différentielles correspondantes :

$$y^{(4)} = f_{y^3}\, fff + 3\, f_{y^2}\, f\, f_y\, f + f_y\, f_{y^2}\, f^2 + f_y\, f_y\, f_y\, f$$

En dérivant $y^{(3)}$, on obtient effectivement ce qu'on cherche :

$$\begin{aligned} y^{(4)} &= f_{y^3}\, y'^3 + f_{y^2} \cdot 2y'y'' + f_{y^2}\, y'\, y'' + f_y\, y''' \\ &= f_{y^3}\, f^3 + 2f_{y^2}\, f\, f_y\, f + f_{y^2}\, f\, f_y\, f + f_y\, \left(f_{y^2}\, f^2 + f_y\, f_y\, f\right) \\ &= f_{y^3}\, fff + 3f_{y^2}\, f\, f_y\, f + f_y\, f_{y^2}\, f^2 + f_y\, f_y\, f_y\, f \end{aligned}$$

• Cas des EDO non autonomes

On considère cette fois l'EDO non autonome (2.1)

$$y' = f(t, y)$$

On se ramène à un système autonome équivalent

$$Y = \begin{pmatrix} t \\ y \end{pmatrix}, \quad Y' = F(Y)$$

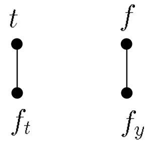
dont on ne regarde que la deuxième composante (car $Y'_{|1} = 1$, $Y''_{|1} = Y'''_{|1} = \ldots = 0$). On a alors $Y'_{|2} = y'$ et on note $y' = \bullet$, par analogie avec le cas autonome. Pour la dérivée seconde, on a

$$\begin{aligned} Y'' &= F_Y \cdot Y' \\ &= \begin{pmatrix} 0 & 0 \\ f_t & f_y \end{pmatrix} \cdot \begin{pmatrix} 1 \\ f \end{pmatrix} \end{aligned}$$

d'où

$$Y''_{|2} = y'' = f_t + f_y\, f$$

ce qu'on note



On dérive encore une fois :

$$Y''' = F_{Y^2} \cdot F^2 + F_Y F_Y F$$

et on explicite chacun des termes de cette somme :

$$\begin{aligned} F_{Y^2} \cdot F^2 &= \left(\frac{\partial}{\partial t} \begin{pmatrix} 0 & 0 \\ f_t & f_y \end{pmatrix}, \quad \frac{\partial}{\partial y} \begin{pmatrix} 0 & 0 \\ f_t & f_y \end{pmatrix} \right) \cdot \begin{pmatrix} 1 \\ f \end{pmatrix}^2 \\ &= \begin{pmatrix} 0 & 0 \\ f_{t^2} + f_{ty}\, f & f_{yt} + f_{y^2}\, f \end{pmatrix} \cdot \begin{pmatrix} 1 \\ f \end{pmatrix} \end{aligned}$$

donc

$$F_{Y^2}\, F^2{}_{|2} = f_{t^2} + f_{ty}\, f + f_{yt}\, f + f_{y^2}\, f f$$

$$F_Y F_y F = \begin{pmatrix} 0 & 0 \\ f_t & f_y \end{pmatrix} \begin{pmatrix} 0 & 0 \\ f_t & f_y \end{pmatrix} \begin{pmatrix} 1 \\ f \end{pmatrix}$$

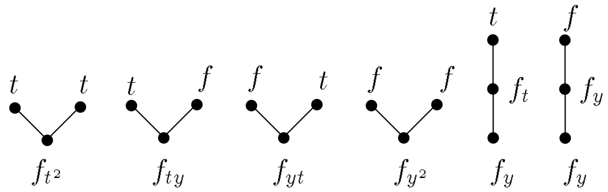
donc

$$F_Y F_Y F_{|2} = f_y\, f_t + f_y\, f_y\, f$$

Finalement, $y^{(3)}$ sécrit

$$y^{(3)} = f_{t^2} + f_{ty}\, f + f_{yt}\, f + f_{y^2}\, f\, f + f_y\, f_t + f_y\, f_y\, f$$

ce qui correspond aux arbres

$y^{(3)} =$ 

Règle de construction : *pour passer des arbres associés à $y^{(k)}$ à ceux associés à $y^{(k+1)}$, il suffit d'ajouter* 

et 

à tous les noeuds quand c'est possible.

- le seul noeud auquel on ne peut rien ajouter est $\overset{t}{\bullet}$.

- ici encore, $y^{(k)}$ est une combinaison linéaire d'arbres ayant tous exactement k noeuds.

A.2 Arbres et différentielles élémentaires pour les EAD semi-explicites d'indice 1

Soit $DAt = DAT_y \cup DAT_z$ l'ensemble des arbres. DAT_y contient tous les arbres dont la racine est $\bullet$, et DAT_z tous ceux dont la racine est $\bigcirc$. Soient $u_1, \ldots, u_m \in DAT_y$ et $v_1, \ldots, v_n \in DAT_z$. Alors si on relie toutes les racines de $u_1, \ldots, u_m, v_1, \ldots, v_n$ par $\bullet$ (qui devient la nouvelle racine), on obtient l'arbre $[\,u_1, \ldots, u_m, v_1, \ldots, v_n\,]_y \in DAT_y$. Si on les relie par $\bigcirc$, le nouvel arbre ainsi construit est $[\,u_1, \ldots, u_m, v_1, \ldots, v_n\,]_z \in DAT_z$.

Définition 12 *Soit $f : I\!R^N \to I\!R^N$ une application infiniment différentiable. La* différentielle élémentaire de f associé à l'arbre u $\in DAT_y$ *est l'application $F(u) : I\!R^N \times I\!R^N \to I\!R^N$ définie par*

$$F(u)(y,z) = \frac{\partial^{m+n} f}{\partial\, y^m\, z^m}\,(F(u_1)(y,z), \ldots, F(u_m)(y,z), F(v_1)(y,z), \ldots, F(v_n)(y,z))$$

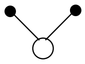
avec $u = [\,u_1, \ldots, u_m\, v_1, \ldots, v_n\,]_y \in DAt_y$ et $F(\tau_y) = f$.

On définit de la même façon la différentielle élémentaire de f associé à l'arbre v $\in DAT_z$. *C'est l'application $F(v) : I\!R^N \times I\!R^N \to I\!R^N$ donnée par*

$$F(v)(y,z) = (-g_z^{-1})\,\frac{\partial^{m+n} g}{\partial\, y^m\, z^m}\,(F(u_1)(y,z), \ldots, F(u_m)(y,z), F(v_1)(y,z), \ldots, F(v_n)(y,z))$$

avec $v = [\,u_1, \ldots, u_m\, v_1, \ldots, v_n\,]_z \in DAT_z$ et $F(\tau_z) = \left(-g_z^{-1}\right)\, g_y\, f$.

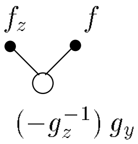
Exemples:
1) Soit $u = [\,\tau_z\,]_y = (-g_z^{-1})\, g_y\, f \in DAT_y$. On a donc

$u =$ 

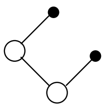
D'après la définition

$$F(u)(y,z) = \frac{\partial f}{\partial z}\,(F(\tau_z)(y,z))$$
$$= f_z\,\left(-g_z^{-1}\right)\,g_y\,f$$

et l'arbre u sécrit finalement

$u =$ 

2) Soit $v = [\,\tau_z, \tau_y\,]_z \in DAT_z$. On a

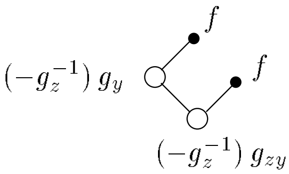
$v =$ 

$$F(v)(y,z) = (-g_z^{-1})\,\frac{\partial^2 g}{\partial y \partial z}\,(F(\tau_z)(y,z),\, F(\tau_y)(y,z))$$

Comme $F(\tau_z)(y,z) = (-g_z^{-1})\, g_y\, g$ et $F(\tau_y)(y,z) = f$,

$$F(v)(y,z) = (-g_z^{-1})\, g_{zy}\, \left((-g_z^{-1})\, g_y\, f\,,\, f\right)$$

et l'arbre v s'écrit



• Cas des EAD autonomes

Dérivons la contrainte alébrique $0 = g(y,z)$ de (1.1):

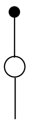
$$0 = g_y\, y' + g_z\, z'$$

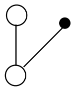
L'indice valant 1, on a $z' = (-g_z{}^{-1})\, g_y\, f$. On représente y' par l'arbre réduit à un seul noeud $\tau_y = \bullet$ et on note $\tau_z = \bigcirc$ pour désigner l'opérateur $(-g_z{}^{-1})g$. On a donc

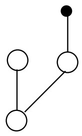
$y' = \bullet$ et $z' =$ 

Règles de construction de $y^{(k+1)}$ et $z^{(k+1)}$ à partir de $y^{(k)}$ et $z^{(k)}$:

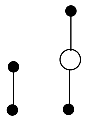
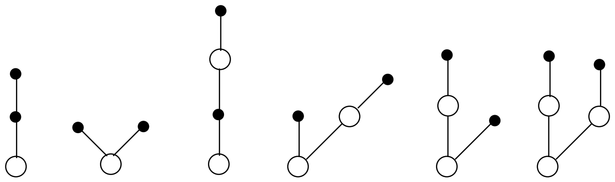
(a) ajouter sur n'importe quel noeud 

(b) ajouter sur n'importe quel noeud 

(c) remplacer toutes les ocurrences de $\bigcirc$ par 

(d) remplacer toutes les ocurrences de $\bigcirc$ par 

On en déduit

$y'' =$  et $z'' =$ 

$f_y\, f \quad f_z\,(-g_z^{-1})\, g_y\, f \quad$ etc.

- On remarque qu'ici les arbres associés aux différentielles élémentaires de $y^{(k)}$, $z^{(k)}$ n'ont en général pas k noeuds mais souvent beaucoup plus (cf. z'' par exemple).

• **Cas des EAD non autonomes**

On considère l'EAD semi-explicite non autonome

$$\begin{aligned} y' &= f(t, y, z) \\ 0 &= g(t, y, z) \end{aligned}$$

qu'on suppose d'indice 1. Comme pour les EDO, on prend comme système autonome équivalent

$$Y = \begin{pmatrix} t \\ y \end{pmatrix}, \qquad \begin{array}{ccc} Y' & = & F(Y, z) \\ 0 & = & G(Y, z) \end{array}$$

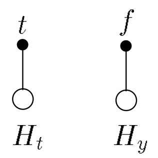
On note cette fois H la fonction obtenue par le théorème des fonctions implicites : $z = H(Y)$. En dérivant la contrainte du système autonome et comme l'indice est 1, il apparaît que H est défini par

$$H = (-g_z)^{-1}\, g$$

On note encore $Y'_{|2} = y' = \bullet$ et pour z', on a

$$\begin{aligned} z' = H_Y \cdot F = [\,H_t,\, H_y\,] \cdot \begin{bmatrix} 1 \\ f \end{bmatrix} \\ = H_t + H_y\, f \\ = (-g_z^{-1})\, g_t + (-g_z^{-1})\, g_y\, f \end{aligned}$$

et z' est donc défini par *deux* arbres.

$z' =$ 

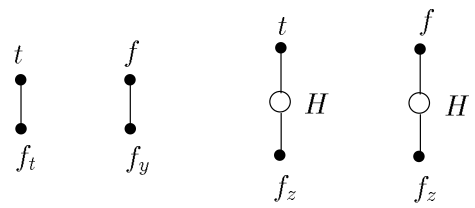
$$Y'' = F_Y\, F + F_z\, H_Y\, F$$

donc la deuxième composante contient les deux termes suivants

$$F_Y\, F_{|2} = [\,f_t,\, f_y\,] \cdot \begin{bmatrix} 1 \\ f \end{bmatrix} = f_t + f_y\, f$$

$$F_z\, H_Y\, F_{|2} = f_z\, [\,H_t,\, H_y\,] \cdot \begin{bmatrix} 1 \\ f \end{bmatrix} = f_z\, H_t + f_z\, H_y\, f$$

qu'on représente par les arbres



A.3 Séries de Butcher (ou B-séries)

Définition 13 *Soient T l'ensemble de tous les arbres u, $h \in I\!R$ fixé et a une fonction $T \to I\!R$. On appelle* B-série de valuation k associée à $y' = f(y)$ et à l'ensemble $a(T)$ *la série de différentielles élémentaires*

$$\sum_{u \in T,\, \rho(u) \geq k} h^{\rho(u)}\, a(u)\, F(u)(y)$$

Par exemple, une développement de Taylor de valuation k est une B-série de même valuation car, d'après la première partie, pour tout j, $y^{(j)}$ s'exprime comme une combinaison linéaire des différentielles élémentaires associées à tous les arbres ayant exactement j noeuds.

Définition 14 *Soient $DAT = DAT_y \cup DAT_z$ l'ensemble de tous les arbres u, v, $h \in I\!R$ fixé et a une fonction $DAT \to I\!R$. On appelle* B-série en y de valuation k associée à $y' = f(y, z)$ $0 = g(y, z)$ et à l'ensemble $a(DAT_y)$ *la série de différentielles élémentaires*

$$\sum_{u \in DAT_y,\, \rho(u) \geq k} h^{\rho(u)}\, a(u)\, F(u)(y, z)$$

De la même façon, on définit la B-série en z de valuation k associée à $y' = f(y,z)$ $0 = g(y,z)$ et à l'ensemble $a(DAT_z)$ *par*

$$\sum_{v \in DAT_z,\, \rho(v) \geq k} h^{\rho(v)}\, a(v)\, F(v)(y,z)$$

Annexe B

Description des EAD testées

On détaille ici les systèmes 4, 5, 8, 9 cités par Hairer dans [12].

• Système 4

Ce problème modélise l'écoulement d'un liquide dans un pipe-line qui peut présenter un étranglement. Les équations sont

$$\pi\sqrt{\frac{R}{2\rho}}(R-y)^2\sqrt{-P'}\left(2.5\ln\left[\sqrt{\frac{\rho R}{2}}.\frac{y}{\mu}.\sqrt{-P'}-5\right]+10.5\right)$$
$$-bQ_{co}-\frac{P_0}{P}Q_{co}(1-b)=0$$

$$2\pi\sqrt{\frac{R}{2\rho}}\sqrt{-P'}\left((2.5Ry-1.25y^2)\ln\left[\sqrt{\frac{\rho R}{2}}.\frac{y}{\mu}.\sqrt{-P'}-5\right]\right.$$
$$\left.+3Ry-2.125y^2-13.6R\mu\sqrt{\frac{2}{\rho R}}\frac{1}{\sqrt{-P'}}\right)-Q_a=0$$

où R, ρ, μ, b, Q_{co}, Q_a, P_0 et L sont des paramètres. Ces équations forment un système implicite de la forme

$$\begin{aligned} f\left(\sqrt{-P'},y,P\right) &= 0 \\ g\left(\sqrt{-P'},y\right) &= 0 \end{aligned}$$

qu'on intègre pour $t\in[0,L]$. Les valeurs à calculer sont la pression $P(t)$ du liquide et l'épaisseur $y(t)$ du pipeline. Contrairement à DASSL, RADAU5 ne peut résoudre directement un tel système. Pour obtenir une formulation du type $M.Y'=F(Y)$, on introduit la variable $u=\sqrt{-P'}$ et le problème initial s'écrit désormais

$$\begin{aligned} P' &= -u^2 \\ 0 &= f(u,y,P) \\ 0 &= g(u,y) \end{aligned}$$

La valeur initiale de P est $P(0) = P_0$. On déduit ensuite des deux contraintes algébriques les valeurs de $u(0)$ et $y(0)$. Deux jeux de donnés sont proposés dans [12]. Le premier correspond à un écoulement standard :

$$\begin{array}{cccc} R = 45.72 & \rho = 0.814 & \mu = 0.098 & b = 0.606 \\ Q_{co} = 1.1531 \cdot 10^6 & Q_a = 2.035 \cdot 10^5 & P_0 = 1.457 \cdot 10^8 & L = 8.047 \cdot 10^6 \end{array}$$

Les valeurs initiales consistantes pour u et y sont alors

$$u(0) = 1.6514698693706613577094 07$$
$$y(0) = 4.8774007077380742483128 39$$

Pour simuler un étranglement, on peut prendre les valeurs suivantes.

$$\begin{array}{cccc} R = 45.72 & \rho = 0.814 & \mu = 0.098 & b = 0.345 \\ Q_{co} = 1.7153 \cdot 10^6 & Q_a = 3.027 \cdot 10^5 & P_0 = 1.378 \cdot 10^8 & L = 3.2188 \cdot 10^7 \end{array}$$

et les conditions initiales consistantes pour u et y sont cette fois

$$u(0) = 2.3665271547403188758740152$$
$$y(0) = 4.8146652453167371525906527$$

Ce cas fait apparaître une singularité vers $1.3 \cdot 10^7$, alors que la solution dans la première version est très lisse (droite pour P et y). D'après [12], la matrice

$$\begin{pmatrix} \partial f/\partial u & \partial f/\partial y \\ \partial g/\partial u & \partial g/\partial y \end{pmatrix}$$

est inversible, d'inverse borné continu au voisinage de la solution. Ce problème est donc bien d'indice 1.

• Système 5

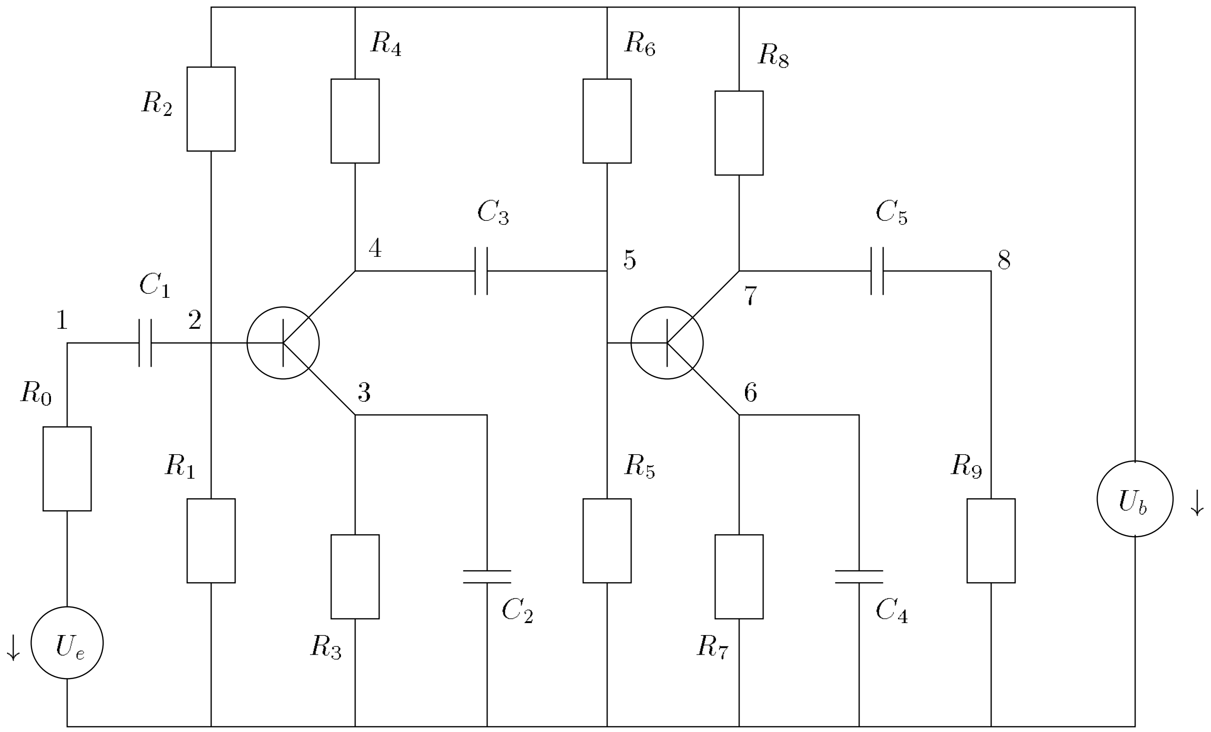
La modélisation des cicrcuits électriques conduit souvent à des EAD. Le système étudié ici correspond en particulier à un amplificateur à deux transistors. Les équations qui décrivent le circuit proviennent directement de l'application des lois de Kirchhoff et forment un système implicite $M.Y' = F(Y)$ de dimension 8 d'indice 1 dont les variables sont les tensions U_i, $i = 1, \ldots, 8$.

$$\begin{array}{rcl} -C_1(U_1' - U_2') & = & U_1/R_0 - U_e(t)/R_0 \\ C_1(U_1' - U_2') & = & (1-\alpha)f(U_2 - U_3) - U_b/R_2 + U_2(1/R_1 + 1/R_2) \\ -C2U_3' & = & U_3/R_3 - f(U_2 - U3) \\ -C_3(U_4' - U_5') & = & U_4/R_4 - U_b/R_4 + \alpha f(U_2 - U_3) \\ C_3(U_4' - U_5') & = & U_5(1/R_5 + 1/R_6) - U_b/R_6 + (1-\alpha)f(U_5 - U_6) \\ -C_4U_6' & = & U_6/U_7 - f(U_5 - U_6) \\ -C_5(U_7' - U_8') & = & U_7/U_8 - U_b/R_8 + \alpha f(U_5 - U_6) \\ C_5(U_7' - U_8') & = & U_8/R_9 \end{array}$$

où

$$f(U_i - U_j) = \beta\, e^{(\frac{U_i - U_j}{U_F} - 1)}$$

est la fonction caractéristique des transistors.



Les paramètres sont

$$U_b = 6, \quad U_F = 0.026, \quad \alpha = 0.99, \quad \beta = 10^{-6}$$

$$R_0 = 1000\,, \quad R_k = 9000\,, k = 1, \dots, 9$$

$$C_k = k.10^{-6}\,, k = 1, \dots, 5$$

et la tension en entrée est donnée par

$$U_e(t) = 0.1 \cdot sin(200\pi t)$$

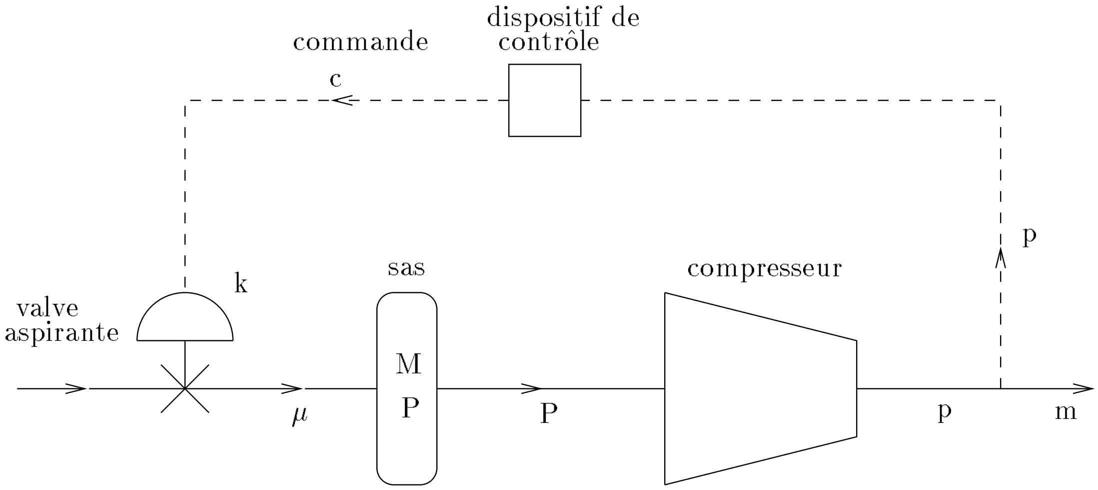
La variable intéressante est $U_8(t)$ et correspond à la tension amplifiée obtenue pour un signal donné en entrée $U_e(t)$. Lors des tests, nous avons étudié cette variable avec $t \in$

$[0, 0.2]$. Les conditions initiales proposées dans [12] sont

$$\begin{array}{l} U_1(0) = 0\,, \quad U_2(0) = U_3(0) = U_b/(R_2/R_1+1)\,, \quad U_4(0) = U_b \\ U_5(0) = U_6(0) = U_b/(R_6/R_5+1)\,, \quad U_7(0) = U_b\,, \quad U_6(0) = 0 \end{array}$$

• Système 8

Il s'agit d'un procédé de contrôle utilisé en pétrochimie. Une valve aspirante permet de réguler la pression en sortie d'un compresseur (cf. schéma ci-dessous).



La modélisation de ce phénomène conduit à

$$\begin{aligned} k' &= \frac{1}{20}\,(c-k) \\ c' + \frac{1}{15}\,p' &= -\frac{1}{75}\,(p-99.1) \\ M' &= \mu - m \\ \frac{p}{P} &= 3.35 - 0.075\,m + 0.001\,m^2 \\ P^2 &= 49.58^2 - \left(\frac{\mu}{1.2\cdot k}\right)^2 \\ M &= 20\,P \\ \mu &= 15 + 5\,tanh(t-10) \end{aligned}$$

où k, c, p, P, m, M et μ sont les variables. k représente la position de la valve, c la commande, p la pression en sortie du compresseur, P celle en entrée, m la masse du flux sortant, M la masse dans le sas et μ celle du flux entrant.

Ce système est initialement d'indice 2 mais si l'on remplace la dernière équation algébrique par une relation de la forme

$$m = f(t)$$

où f est une fonction du temps donnée, l'indice vaut alors 1 [12]. Seule la version d'indice 2 a été testée lors de l'estimation de l'erreur globale.
Des conditions initiales consistantes peuvent être obtenues comme suit. On choisit

$$k(0) = 0.25\,, \quad p(0) = 99.$$

Hairer [12] montre que les autres c. i. sont alors déterminées de façon unique et valent

$$\begin{array}{rcl} c(0) & = & 0.2500145529515559 \\ P(0) & = & 36.70238799190793 \\ m(0) & = & 9.998240239254898 \\ M(0) & = & 734.0477598381585 \\ \mu(0) & = & 15 + 5\,tanh(-10) \end{array}$$

• **Système 9**

On a vu dans la partie consacrée aux tests réalisés sur des EDA d'indice 2 que ce système est au départ une EDO raide. Il modélise un circuit électrique devant fournir à un amplificateur un signal obtenu par combinaison d'un signal $e_1(t)$ de fréquence faible et un signal $e_2(t)$ de fréquence élevée. Cela se traduit par les 15 équations différentielles

$$
\begin{aligned}
CU_1' &= I_1 - \frac{I_3}{2} + \frac{I_4}{2} + I_7 - \frac{U_1}{R} \\
CU_2' &= I_2 - \frac{I_5}{2} + \frac{I_6}{2} + I_8 - \frac{U_2}{R} \\
C_S U_3' &= I_3 - G(UD_1) + G(UD_4) \\
C_S U_4' &= -I_4 + G(UD_2) - G(UD_3) \\
C_S U_5' &= I_5 + G(UD1) - G(UD_3) \\
C_S U_6' &= -I_6 - G(UD_2) + G(UD_4) \\
C_P U_7' &= -\frac{U_7}{R_i} + G(UD_1) + G(UD_2) - G(UD3) - G(UD_4) \\
L_h I_1' &= -U_1 \\
L_h I_2' &= -U_2 \\
L_{S2} I_3' &= \frac{U_1}{2} - U_3 - R_{g2} \cdot I_3 \\
L_{S3} I_4' &= -\frac{U_1}{2} + U_4 - R_{g3} \cdot I_4 \\
L_{S2} I_5' &= \frac{U_2}{2} - U_5 - R_{g2} \cdot I_5 \\
L_{S3} I_6' &= -\frac{U_2}{2} + U_6 - R_{g3} \cdot I_6 \\
L_{S1} I_7' &= -U_1 + e_1(t) - (R_0 + R_{g1}) \cdot I_7 \\
L_{S1} I_8' &= -U_2 - (R_a + R_{g1}) \cdot I_8
\end{aligned}
$$

où

$$G(UD) = 40.67286402 \cdot 10^{-9} \cdot [\, exp(17.7493332 \cdot UD) - 1\,]$$

est la caractéristique des diodes et où

$$
\begin{aligned}
UD_1 &= U_3 - U_5 - U_7 - e_2(t) \\
UD_2 &= -U_4 + U_6 - U_7 - e_2(t) \\
UD_3 &= U_4 + U_5 + U_7 + e_2(t) \\
UD_4 &= -U_3 - U_6 + U_7 + e_2(t)
\end{aligned}
$$

sont les tensions aux diodes.
Les paramètres sont définis par

$$
\begin{array}{ccc}
R_{g1} = 36.3\,, & R_{g2} = R_{g3} = 17.3\,, & R_0 = R_i = 50\,, \\
R_a = 600\,, & R = 2500\,, & \\
C = 16 \cdot 10^{-9}\,, & C_P = 10 \cdot 10^{-9}\,, & \\
L_h = 4.45\,, & L_{S1} = 0.002\,, & L_{S2} = L_{S3} = 0.0005
\end{array}
$$

et les signaux d'entrée sont

$$e_1(t) = 0.5\, sin(2\pi 10^3 t)\,, \quad e_2(t) = 2\, sin(2\pi 10^4 t).$$

On prend comme conditions initiales

$$\begin{aligned} U_i(0) &= 0\,, \quad i = 1, \ldots, 7\,, \\ I_i(0) &= 0\,, \quad i = 1, \ldots, 8. \end{aligned}$$

Pour C_S, trois cas de figure sont envisagés dans [12] :

$$C_S = 10^{-9}\,, \quad C_S = 2 \cdot 10^{-12}\,, \quad C_S = 0$$

Dans les deux premiers cas, on a une EDO raide. Le dernier cas conduit à une EAD d'indice 2. La variable $I_8(t)$ est proportionnelle au signal de sortie et est donc intéressante d'un point de vue physique. Sur le plan numérique, la variable qui reflète le mieux la complexité de ce problème est $U_3(t)$. En effet lors de l'intégration sur $[\,0, 2 \cdot 10^{-4}\,]$ de ces 3 versions, cette variable présente les oscillations les plus fortes.