# Contribution à l'estimation asymptotique de l'erreur globale des méthodes d'intégration numériques à un pas. Application à la simulation des réseaux électriques

René Aïd

**THÈSE**

présentée par

**René AÏD**

pour obtenir le grade de DOCTEUR

de l'INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

(Arrêté ministériel du 30 mars 1992)

(spécialité : **Mathématiques Appliquées**)

# Contribution à l'estimation asymptotique de l'erreur globale des methodes d'intégration numérique à un pas Application à la simulation des réseaux électriques

**Date de la soutenance** : 07 Janvier 1998
**Composition du Jury**:

| | |
|---|---|
| P. Witomski | président |
| E. Hairer | rapporteur |
| A. Ronveaux | rapporteur |
| J. Della Dora | directeur de thèse |
| G. Villard | examinateur |
| P. Panciatici | examinateur |
| L. Levacher | examinateur |

Thèse financée par la Direction des Études et Recherches d'Électricité de France et préparée au Laboratoire LMC-IMAG.

*Sommaire:*

*Démonstration expérimentale d'une organisation tomatotopique chez la Cantatrice.*

L'auteur étude les fois que le lancement de la tomate il provoquit la *réaction yellante* chez la Chantatrice et demonstre que divers plusieures aires de la cervelle elles était implicatées dans le response, en particular, le trajet légumier, les nucléi thalameux et le fiçure musicien de l'hémisphère nord.

Georges Perec, *Cantatrix Sopranica L.*, Librairie du $XX^e$ siècle, Seuil

**Remerciements**

Naïf, je croyais une thèse le résultat d'un travail personnel et solitaire. Mais, à l'heure de la conclure, je réalise que s'il s'agit bien d'un ouvrage personnel, il n'est en rien solitaire. Par leur présence ou leur intervention, de nombreuses personnes m'ont permis d'arriver à ce résultat.

Au cours de ces trois années, trois d'entre elles ont occupé une place fondamentale pour ma formation. Il s'agit de Jean Della Dora, de Gilles Villard et de Laurent Levacher.

Jean Della Dora, professeur à l'Inpg, m'a séduit quand j'étais encore un tout jeune étudiant de l'Ensimag. A l'époque je n'étais encore que de la pâte à pizza informe dont il se servait pour nous représenter des surfaces de Riemann. Je ne saurais dire ce qui de ses cours ou de sa personne a fait de moi son fan au point de ne pouvoir imaginer faire ma thèse sans lui. Aussi, je tiens à le remercier d'avoir accepté de m'avoir pris sous son aile. Pendant trois ans, j'ai grandi non seulement dans la lumière qu'il projetait sur les équations différentielles, mais surtout, dans le plaisir de nos conversations, de sa gentillesse et de sa joie de vivre. Quand je me sentais déprimé par la difficulté des choses, c'est vers lui que je me tournais pour trouver l'énergie de poursuivre.

Gilles Villard, chargé de recherches au Cnrs, a accepté d'assurer ma formation au travail de chercheur. Il a eu beaucoup à faire. Et, c'est à lui que s'adressent mes remerciements les plus vifs. Maintenant que le temps a passé, je mesure de façon plus juste ce qu'il a souffert pour insuffler en moi un peu de la rigueur nécessaire à la conduite d'une thèse, des normes de la rédaction d'un document scientifique mais aussi et surtout simplement, pour supporter un caractère si différent du sien.

Laurent Levacher, ingénieur d'études à la Der d'Edf à l'origine de l'élaboration de ce sujet, a beaucoup œuvré pour le bon déroulement de ma thèse. Il a veillé à une bonne et rentable utilisation de ces trois années. Ce faisant, il a permis que quelque chose se construise. Et, c'est un honneur qu'il ait accepté de faire partie de mon jury. Je reste en admiration devant son sens de l'action et pour toutes ces raisons, je voudrais ici lui témoigner ma reconnaissance.

Je souhaite aussi remercier plusieurs autres personnes de la Der d'Edf. Tout d'abord, je tiens à remercier Patrick Panciatici et Ollivier Fillatre, ingénieurs d'études, d'avoir accepté de faire partie du comité technique de suivi de ma thèse. Par leurs réflexions et leurs conseils, ils m'ont permis de mieux cerner leur problème. Et, c'est pour moi une source de satisfaction que Patrick Panciatici ait aussi accepté de faire partie de mon jury. Enfin, je tiens à remercier M. Bruno Meyer, chef du Département CER, qui à l'époque où je cherchais un sujet et un financement de thèse était chef du groupe OSR et m'a fait confiance pour ce travail.

Je voudrais maintenant remercier Ernst Hairer, professeur à l'Université de Genève, d'avoir accepté de jouer le rôle de rapporteur et de m'avoir aidé à améliorer de façon substantielle ce manuscrit. J'ai d'abord connu Ernst Hairer

par ses livres et ses articles puis, plusieurs fois, j'ai eu l'occasion de l'écouter exposer ses recherches et de discuter avec lui. J'ai découvert en lui non seulement le chercheur et l'enseignant mais aussi une personne d'une grande gentillesse. C'est déjà pour moi un grand honneur qu'il ait accepté de lire ce travail et de le juger.

Bien que je ne puisse pas faire état d'une semblable connaissance pour André Ronveaux, professeur à la Faculté Notre-Dame de la Paix de Namur, je tiens à le remercier d'avoir accepté de rapporter sur ce manuscrit et de montrer ainsi qu'il reste accessible à des mathématiciens non-spécialistes de la question.

Je voudrais aussi remercier Patrick Witomski, professeur à l'Université Joseph Fourier, d'avoir accepté de présider mon jury. C'est toujours un honneur d'être jugé par des personnes que l'on estime. Et, c'est dès ma première année de l'Ensimag, que j'ai découvert et apprécié cet homme grand par la taille et par l'esprit.

Je tiens aussi à remercier René Alt, professeur à l'Université de Paris VI, de m'avoir accordé du temps pour discuter de certains points techniques de mes recherches, Philippe Chartier, chargé de recherches à l'Irisa, pour m'avoir permis de m'exprimer devant un public intéressé, et Michel Crouzeix, professeur à l'Université de Rennes, pour ses remarques et ses suggestions.

Je voudrais aussi adresser mes plus sincères remerciements à Claudine Chaffy, Françoise Jung et Jean-Baptiste Nugeyre, tous trois membres de l'équipe de Calcul Formel du Lmc, qui se sont penchés sur moi avec une grande gentillesse et une grande patience pour m'expliquer de nombreux points d'analyse complexe et du calcul avec les flottants. Cela a toujours été pour moi un grand plaisir de m'entretenir avec eux. Je n'oublie pas non plus François Robert dont le soutien moral et les conseils de routard de la recherche m'ont été précieux.

J'aimerais aussi dire combien j'ai trouvé agréable le quotidien avec mes collègues thésards. En particulier, je voudrais remercier celui qui a partagé le même bureau que moi depuis toujours, Laurent Testard. Je le remercie d'avoir supporté mes traits de caractère les plus irritants. Je voudrais dire combien cela m'a fait plaisir d'avoir pu trouver avec lui un problème à la jonction de nos deux sujets et d'avoir travaillé avec lui. De même, je voudrais dire à Evelyne Hubert combien elle a été et reste pour moi une source constante d'étonnements et de réjouissances. Gabriel Thomas a aussi beaucoup compté depuis que l'on fréquente les mêmes cours et les mêmes cafés, non seulement par les explications qu'il m'a prodiguées mais aussi par sa bonhomie. Et, les derniers mois de ma thèse ont été littéralement enchantés par l'arrivée du jeune Claude-Pierre Jeannerod. Cela a été pour moi une réelle source de satisfaction d'encadrer son DEA en collaboration avec Josselin Visconti que je tiens aussi à remercier pour nos nombreuses discussions.

Je voudrais aussi remercier les deux étudiants, Rodolphe Chopinet et Philippe Poncet, qui ont permis que certains des résultats expérimentaux présentés ici existent.

Enfin, je ne pense pas que j'aurais été capable de venir à bout de ce travail sans l'aide précieuse de Charles Maffré de Lastens, l'affection d'Ibtissam

El Boukhari, l'amitié d'Hervé Raynaud et l'amour de Brigitte *Bib* Piallat. Qu'ils soient ici remerciés.

# Table des matières

# Introduction

Le 19 décembre 1978, un accroissement rapide de la demande en électricité a conduit à un écroulement de l'ensemble du réseau électrique français. L'interruption de la distribution dura plusieurs heures, de huit heures et demi du matin jusqu'au milieu de l'après-midi. Le coût d'une telle défaillance du réseau n'est pas chiffré. Mais, si on l'évalue en nombres d'heures de travail perdues, il est de l'ordre de la centaine de millions d'heures.

Bien que ce type de panne générale soit rare, un réseau électrique n'en est jamais à l'abri. En France, un autre événement de cette ampleur est survenue le 12 janvier 1987 et, il est possible de recenser par an, dans le monde, de très nombreux incidents de cette espèce.

L'origine de ces pannes tient à la nature même de la gestion de l'énergie électrique. Comme elle ne peut être stockée avec un faible coût, l'exploitant d'un réseau doit en permanence maintenir un équilibre entre l'offre disponible et la demande potentielle. Quand le réseau s'éloigne trop de cette zone d'équilibre, ce genre de panne peut alors survenir. C'est notamment le cas en période de grand froid - tous les radiateurs sont branchés simultanément -, ou de grande chaleur - *idem* avec la climatisation.

Pour s'assurer de la fiabilité d'un réseau électrique, la Direction des Études et Recherches d'Électricité de France (DER d'EDF) a développé en collaboration avec la société belge Tractebel, le logiciel EUROSTAG. Le principe de ce simulateur est de permettre à l'utilisateur d'une part, de concevoir et de dessiner facilement un réseau comptant quelques nœuds à plusieurs centaines, d'autre part, d'établir des scénarii d'incidents (coupure d'une ligne, arrêt d'une tranche, augmentation brutale de la demande...) et de simuler alors l'évolution du réseau au cours du temps. À partir de ces résultats, il est possible de dire si les systèmes automatiques de régulation du réseau (automates) sont suffisants à rétablir un état d'équilibre, même dégradé, après une avarie.

Le modèle mathématique choisi dans EUROSTAG pour effectuer ses simulations est un *système d'équations algébro-différentielles d'indice un*, muni d'une condition initiale qui est un point d'équilibre du système. Ce modèle résulte d'une première approximation. Un modèle plus précis devrait tenir compte du temps de propagation du champ électrique et conduirait à un système d'équations aux dérivées partielles. La dimension du système peut varier entre une centaine de variables dans le cas d'une étude d'un réseau régional, à plusieurs milliers de variables pour un réseau national.

La nature même des scénarii étudiés conduit à des systèmes différentiels dont les temps caractéristiques diffèrent de plusieurs ordres de grandeurs (de la milliseconde pour un court-circuit, à la seconde au cours d'un état d'équilibre), ce qui en fait des systèmes *raides*. Enfin, toutes les variables sont astreintes à demeurer dans des intervalles de valeurs à l'intérieur d'une zone de fonctionnement du réseau. Cela a comme conséquence d'imposer des *discontinuités* par rapport aux variables d'état du système à la fonction le décrivant. Typiquement, si une variable atteint un certain seuil, une action est déclenchée qui correspond à un changement du mode de calcul des fonctions donnant le système.

Le franchissement de ces seuils est un moment crucial. Un des problèmes pour EUROSTAG est de les détecter. L'instant pour lequel ce seuil est atteint est essentiel dans les études de re-stabilisation du réseau.

Comme le modèle mathématique est le résultat d'une approximation physique de la réalité, que de nombreuses discontinuités sont mises en jeu, que l'intégration numérique est effectuée sur de longs intervalles de temps, on est en droit de se demander si la simulation est fiable, c'est-à-dire si qualitativement, le système se re-stabilise vraiment après l'intervention des automates et si, quantitativement, les valeurs après le passage d'un seuil ont un rapport avec la réalité.

Une manière de répondre à cette question est de parvenir à encadrer la solution exacte du système différentiel dans un tuyau centré sur la solution numérique de façon à pouvoir anticiper la détection du seuil (FIG. 0.1). Il s'agirait donc de borner finement l'erreur commise entre la solution numérique et la solution exacte.

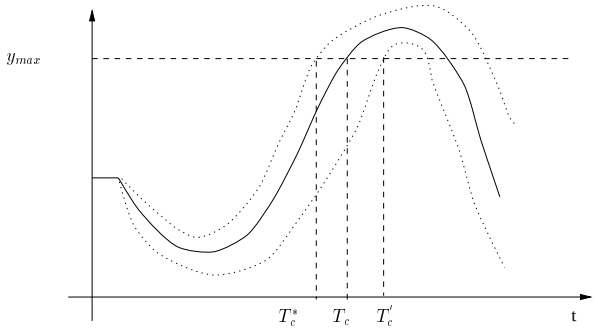



FIG. 0.1 – *Détection d'un seuil*

Devant le nombre des difficultés mathématiques apparaissant dans les systèmes différentiels traités par EUROSTAG, la question qui nous était posée était de savoir dans quelle mesure il était possible de réaliser un tel encadrement dans des cas plus simples. En somme, il nous était demandé de ne retenir de toutes les

difficultés énumérées ci-dessus, que la plus classique d'entre elles, celle de *la validation des solutions numériques d'équations différentielles ordinaires* (EDO):

$$\begin{array}{lcl} \dot{y} & = & f(y), \\ y(0) & = & y_0, \end{array} \tag{0.1}$$

intégrées sur un intervalle de temps $[0, T]$.

Comme les systèmes traités par EUROSTAG sont des systèmes algébriques et différentiels semi-explicites et d'indice un, ils ne diffèrent pas essentiellement des équations différentielles ordinaires [BCP89]. En revanche, s'il est clair qu'à ce niveau, nous ne considérons plus l'erreur de modélisation, il reste que la différence entre la solution numérique et la solution exacte comprend deux types d'erreur différents, une erreur due à la méthode d'intégration utilisée et une erreur arithmétique, c'est-à-dire de l'erreur due à la représentation des nombres sur un calculateur. De ces deux types d'erreurs, nous avons fait l'hypothèse que l'erreur arithmétique était négligeable par rapport à celle de la méthode d'intégration.

L'obtention de bornes de l'erreur commise par une solution numérique d'équations différentielles ordinaires n'est pas un problème récent. La majoration classique de cette erreur donnée par la relation [Hen62]

$$\|y_n - y(t_n)\| \leq C \, \frac{e^{L(t_n - t_0)} - 1}{L} \max_{0 \leq i \leq n-1} h_i^p,$$

où $y_n$ est la solution numérique de (0.1) approchant la solution exacte $y(t_n)$, $L$ et $C$ sont des constantes dépendant respectivement du problème et de la méthode d'intégration, $p$ l'ordre de la méthode d'intégration et $h_i$, le pas d'intégration tel que $t_{i+1} = t_i + h_i$, $t_0 = 0$, est beaucoup trop pessimiste pour pouvoir être utilisée.

Un domaine de l'analyse numérique, *l'analyse d'intervalle*, s'est constitué autour de la seule question de l'obtention d'encadrements sûrs. L'analyse d'intervalle tient compte non seulement de l'erreur de la méthode de résolution, mais aussi de l'erreur arithmétique. Cela impose de remplacer tous les calculs sur les nombres par des calculs sur des intervalles contenant ces nombres [Moo76]. Depuis, des travaux ont été effectués pour porter les méthodes de l'analyse d'intervalle à l'intégration d'équations différentielles ordinaires [DS76, Rih94]. Dans le cas de l'intégration numérique des EDO, cette approche conduit à une explosion des encadrements successifs [Rhi97]. Après quelques itérations, le seul intervalle sûr est $\mathbb{R}$ lui-même.

La validation de la solution numérique de (0.1) est moins exigeante que le strict encadrement décrit plus haut. Dans la mesure où de nombreuses approximations ont été faites pour rendre le système différentiel intégrable dans des temps raisonnables, on peut se contenter d'une estimation de l'erreur commise.

On prendra alors comme tuyau encadrant de la solution de (0.1), la solution numérique $\pm$ l'estimation.

C'est l'approche que nous avons adoptée dans ce travail. Nous nous sommes intéressés à des *méthodes d'estimation asymptotique*. Le principe de ces méthodes est d'améliorer le résultat numérique obtenu. Effectivement, il suffit de disposer d'une valeur de la solution exacte meilleure $\tilde{y}_n$ que $y_n$ pour utiliser $y_n - \tilde{y}_n$ comme d'une estimation de l'erreur commise. Quand on dispose d'un estimateur asymptotique, il est aussi facile d'en dériver des bornes asymptotiques de l'erreur globale [BS66, Sha84].

L'intérêt de ces méthodes est de rester dans la droite file des méthodes d'intégration numérique elles-mêmes. Elles sont donc implantables facilement dans un logiciel industriel. Leur inconvénient, comme les méthodes d'intégration numérique, est de ne fournir un résultat qui n'est garanti qu'à la limite, quand le pas tend vers zéro. La question qui se pose alors à nous est de savoir dans quelle mesure pour une intégration donnée, ces estimateurs sont fiables, c'est-à-dire s'il n'est pas nécessaire d'attendre la convergence de l'estimateur pour obtenir au moins l'ordre de grandeur de l'erreur et son premier chiffre significatif.

De plus, on ne cherche pas à répondre à cette question dans le cas de la méthode particulière implantée dans EUROSTAG. Le problème est plutôt la recherche de procédés d'estimation d'erreur que l'on puisse greffer à n'importe quelle méthode. Et, bien que la méthode d'intégration numérique implantée dans EUROSTAG soit une méthode de prédiction-correction à pas et à ordre variables utilisant une mélange de méthodes Adams et BDF [ABJ93], comme nous nous intéressons avant tout aux procédés généraux d'estimation de l'erreur globale, il était beaucoup plus simple de nous focaliser sur les méthodes à un pas variable.

Ce mémoire est construit de la façon suivante.

La Partie I présente et analyse les estimateurs asymptotiques que nous avons retenus pour notre étude. L'analyse se restreint aux seules méthodes d'intégrations à un pas. Cette partie s'attache à étudier l'influence des variations du pas sur ces techniques d'estimation.

Les résultats théoriques concernant ces estimateurs n'étant qu'asymptotiques, il est nécessaire de les doubler de tests numériques. La Partie II est consacrée à ce besoin. On reproduit les tests publiés dans [AL97]. À ceux-ci nous avons ajouté des tests effectués sur EUROSTAG.

La Partie III ne concerne pas directement le travail effectué autour d'EUROSTAG. Elle présente d'une part une application du contrôle de l'erreur locale au contournement automatique des singularités isolées des EDO, d'autre part les effets d'une singularité sur l'erreur globale.

# Première partie

# Analyse de quelques estimateurs asymptotiques

# Chapitre 1

# Introduction

Dans [Ske86], on trouve recensés un grand nombre d'estimateurs possibles. Ce qui a guidé notre choix est la généricité des techniques d'estimation. Nous souhaitions pouvoir les utiliser sans avoir à entrer dans les paramètres de la méthode à un pas. Ainsi, la technique d'Epstein et Hicks [EH79] décrite dans le cas particulier de la méthode d'Euler ne pouvait pas être retenue, de même que celle faisant intervenir deux méthodes d'intégration différentes ou celle applicable à des méthodes de Runge-Kutta dont l'équation variationnelle vérifie une propriétée particulière (*Method with an Exact Principal Error Equation*). En revanche, il aurait été possible de retenir la technique qui consiste à effectuer deux intégrations avec deux tolérances différentes [Ste80b]. Nous avons toutefois préféré la technique comparable de Richardson.

Nous avons donc retenu quatre estimateurs :

- l'estimateur de Richardson (RS) [SW76],
- les estimateurs de Zadunaisky (ZD) [Zad76],
- l'intégration de l'équation variationnelle (EV) [Ste74, Pro80],
- le calcul d'une correction globale[1] (SC) [Ske86].

Ces estimateurs fonctionnent tous selon un principe commun de calcul de variation. Ils effectuent tous une seconde intégration numérique en parallèle de la première. Cette seconde intégration est obtenue en faisant varier un des paramètres de la solution numérique du problème (0.1), ou bien en effectuant une mise à l'échelle de l'erreur globale. Ainsi, l'estimateur de Richardson se calcule en utilisant deux solutions numériques de (0.1) obtenue sur deux grilles différentes. Les estimateurs de Zadunaisky utilisent la solution numérique d'un problème voisin de (0.1) calculé sur la même grille. L'intégration de l'équation variationnelle utilise la connaissance de l'équation vérifiée par le terme dominant de l'erreur globale. Le calcul d'une correction globale généralise cette approche.

Le Chapitre 2 est consacré aux rappels des notions dont nous avons besoin pour conduire notre analyse. On rappelle en particulier, la forme du développe-

1. Solving for the Correction, d'où SC.

ment asymptotique de l'erreur globale des méthodes à un pas (§2.1), les notions de base des méthodes de Runge-Kutta (§2.2). On rappelle aussi au paragraphe §2.3, les principes du contrôle de l'erreur locale dans un code d'intégration numérique. On définit une restriction que l'on appelle *contrôle théorique* et qui correspond au but que cherche à réaliser un contrôle effectif.

Le Chapitre 3 est consacré aux estimateurs asymptotiques. Plus qu'à leur coût, on s'intéresse à leur *ordre relatif de convergence*. Un estimateur asymptotique $\hat{E}_n$ de $E_n$ est dit *valide d'ordre relatif* $r > 0$ quand il vérifie $\hat{E}_n = E_n\,(1 + \mathbf{O}(h^r))$. Pour une intégration à pas variable, on remplace dans la relation précédente, $h$ par $H = \max_i h_i$ ou par $\tau$, la tolérance utilisateur. Dans le cas d'une intégration à pas constant, l'ordre de convergence relatif des trois premiers estimateurs est déjà bien connu. Pour les deux premiers estimateurs, il est de 1 [Hen62, Pro80]. L'ordre relatif maximum des estimateurs de Zadunaisky est égal à l'ordre de la méthode d'intégration numérique [FU75, Hai78].

Contrairement à l'estimateur de Richardson, il est possible de construire une pléthore de variantes autour de l'estimateur classique de Zadunaisky [Zad66]. Celles que l'on rencontre dans la littérature sont construites en utilisant des procédés d'intégration numérique de Lagrange de la solution numérique [Zad76, Hai78, DDP84, DP85] ou des tangentes de la solution numérique [DDP84, DP85], ou d'une combinaison des deux par de l'interpolation d'Hermite [DDP84, DP85]. La technique de Zadunaisky a aussi donné lieu à des estimations de l'erreur locale dans [CK63, Alt82].

Contrairement à l'estimateur de Richardson dont la preuve de l'ordre de convergence tient en une demi-ligne (§3.1), celle des estimateurs de Zadunaisky demande un peu d'effort. Pour obtenir une présentation suffisamment générale et explicative du fonctionnement de ces estimateurs, on les considère du point de vue des perturbations régulières du problème (0.1). De cette manière, on essaie de faire une synthèse des différents résultats connus sur ces estimateurs que l'on trouve répartis dans [FU75, Hai78, DDP84, DP85]. Il est alors possible de montrer qu'une partie de ce qui a été fait dans le cas non-autonome (§3.2.1) admet un pendant autonome (§3.2.2). En effectuant un parallèle entre la technique de Zadunaisky et la technique d'analyse rétrograde de l'erreur connue sous le nom d'*Équation Modifiée* [WH74, Cor94], on montre ici comment cette dernière permet de sortir de l'interpolation numérique et d'envisager un algorithme numérico-formel d'estimation de l'erreur globale.

Dans le paragraphe §3.3, on rappelle la relation de récurrence à laquelle conduit l'intégration de l'équation variationnelle.

Pour le calcul d'une correction globale, on améliore légèrement les résultats de [Ske86, Pet86]. On montre qu'il n'est pas nécessaire d'intégrer avec une méthode d'ordre $p$ le problème qu'utilise cet estimateur pour obtenir une estimation valide. Il est possible de réaliser un ordre relatif de convergence de $r$ avec une méthode d'ordre $r$.

Le Chapitre 4 est consacré à l'extension du développement asymptotique

de l'erreur globale à pas variable. Effectivement, si une méthode d'intégration numérique conserve son ordre de convergence en fonction du pas maximum, la question se pose de savoir si cela reste vrai pour les estimateurs, i.e. si un estimateur d'ordre relatif $r$ par rapport à $h$ restera d'ordre $r$ par rapport à $H$.

Dans le cas de l'estimateur de Richardson et de l'intégration de l'équation variationnelle, les preuves de leur ordre de convergence n'utilisent que l'existence du premier terme du développement de l'erreur globale. Les résultats classiques sur le comportement asymptotique de l'erreur globale des méthodes à un pas variable sont alors suffisants pour conclure [CM84]. Ce n'est pas le cas des estimateurs de Zadunaisky et du calcul d'une correction globale, dont les preuves font un usage répété des termes suivants de ce développement.

Il s'avère que dans le cas d'une intégration à pas variable, il n'existe que des résultats partiels donnant l'existence des termes suivant du développement asymptotique de l'erreur globale en fonction de $H$. Cependant, au lieu de prendre $H$ comme paramètre de mesure de l'ordre d'un estimateur, il est aussi possible de prendre $\tau$, la tolérance utilisateur. Mais, même dans ce cas, il n'existe pas de théorème assurant l'existence de ces termes pour une heuristique classique de sélection du pas telle que l'on peut la trouver dans un code d'intégration numérique moderne. Toutefois, un résultat prouvant l'existence du premier terme de ce développement pour un contrôle classique de l'erreur locale peut être trouvé dans [Hig91].

On montre dans ce Chapitre comment il est possible d'étendre le développement de l'erreur globale en fonction de la tolérance aux termes suivants du développement dans le cas d'un contrôle théorique.

Le Chapitre 5 est un retour sur l'estimateur classique de Zadunaisky dans le cas d'une intégration à pas variable. On souhaite comprendre l'influence des variations du pas sur l'ordre relatif de cet estimateur. Dans la mesure où il affiche un ordre relatif de $p$, on souhaite savoir s'il est possible qu'il conserve cet ordre de convergence sur des grilles non-uniformes. La question se pose aussi pour le calcul d'une correction globale. Toutefois, comme l'analyse de ce dernier estimateur peut se ramener au précédent, il suffit de considérer l'estimateur classique de Zadunaisky.

On considère ici des grilles obtenues par contrôle de l'erreur locale dépendant de $\tau$ et des grilles convergentes dépendantes du seul paramètre $H$.

Dans le premier cas, il existe déjà un résultat donnant l'ordre de convergence de l'estimateur de Zadunaisky par rapport à $\tau$ dans [CHMR96]. Ce résultat est obtenue pour la famille particulière des méthodes de Runge-Kutta développées par Dormand *et al.* dans [DLMP89]. En utilisant le résultat du Chapitre 4, on énonce dans le paragraphe §5, l'ordre de convergence de cet estimateur dans le cas d'un contrôle théorique de l'erreur locale sans hypothèse sur la méthode de Runge-Kutta utilisée.

Nous nous penchons par la suite, dans le paragraphe §5.2, sur le cas particulier des méthodes Runge-Kutta développées par Dormand *et al.* dans [DDP84, DP85]. Ces méthodes ont été les premières développées par les auteurs pour

obtenir par la technique de Zadunaisky une meilleure estimation qu'avec une méthode quelconque.

On montre comment ces méthodes présentent sur des grilles quelconques un meilleur comportement asymptotique que les autres méthodes. Sans hypothèse sur les coefficients de la méthode, pour offrir une estimation valide, même d'ordre 1, l'estimateur de Zadunaisky a besoin de plus d'un terme du développement de l'erreur globale. En revanche, quand il est utilisé avec les méthodes spécifiques de Dormand et Prince, on peut montrer que pour fournir une estimation valide, il n'est plus nécessaire de supposer l'existence des termes suivants du développement de l'erreur globale par rapport au pas maximum.

Cette partie est une version augmentée de [Aïd97].

# Chapitre 2

# Rappels

## 2.1 Définitions et notations

On se donne un problème de condition initiale de la forme

$$\begin{array}{rcl} \dot{y}(t) & = & f(y(t)) \\ y(0) & = & y_0 \end{array} \tag{2.1}$$

où $t \in [t_0, T]$, $t_0 = 0$, $T > 0$, $f$ est une fonction vectorielle réelle, $\dot{y}$ est la dérivée de $y$ par rapport à $t$. Le flot de (2.1) sera noté $\psi(t_0, y_0; t)$. On supposera de plus que la fonction $f$ est aussi régulière que l'on veut.

Une méthode numérique fournit une valeur approchée $y_n$ à l'instant $t_n = t_{n-1} + h_{n-1}$, $t_M = T$, de $y(t_n)$, $n = 0, \ldots, M$. L'erreur globale commise à l'instant $t_n$ est la différence $E_n = y_n - y(t_n)$.

Pour plus de simplicité, on donne les définitions suivantes pour des méthodes explicites.

**Définition 2.1.1** *Une méthode à un pas est définie par le schéma:*

$$y_{n+1} \quad = \quad y_n + h_n \Phi(y_n, h_n, f), \tag{2.2}$$

*où $y_0$ est la condition initiale de (2.1). La fonction $\Phi$ est appelée la fonction d'incrément de la méthode.*

Quand il n'y aura pas de risque de confusion, la dépendance de $\Phi$ par rapport à $f$ sera omise.

**Définition 2.1.2** *: Soit $H = \max_n h_n$. La méthode à un pas (2.2) est* convergente *si :*

$$\lim_{H \to 0} \max_{0 \leq n \leq M} \|E_n\| \quad = \quad 0.$$

*Elle est* convergente d'ordre $p$ *si :*

$$E_n \quad = \quad \mathbf{O}(H^p).$$

**Définition 2.1.3** *Un estimateur $\hat{E}_n$ de $E_n$ sera dit* valide *d'ordre relatif $r > 0$ quand*

$$E_n = \hat{E}_n \, (1 + \mathbf{O}(H^r)). \tag{2.3}$$

**Remarques** : La valeur $\hat{y}_n = y_n - \hat{E}_n$ est alors d'ordre $p + r$ par rapport à $H$. On définit de la même manière l'ordre relatif d'un estimateur par rapport à la tolérance utilisateur $\tau$.

On rappelle la définition de *l'erreur de troncature* et de *l'erreur locale*:

**Définition 2.1.4** L'erreur de troncature *de la méthode (2.2) est*

$$\varepsilon_n \quad = \quad y(t_{n+1}) - y(t_n) - h_n \Phi(y(t_n), h_n) \tag{2.4}$$

**Définition 2.1.5** L'erreur locale *de la méthode (2.2) est*

$$\mathbf{le}_n \quad = \quad \psi(t_n, y_n; t_{n+1}) - y_{n+1}. \tag{2.5}$$

**Définition 2.1.6** *: La méthode (2.2) est dite d'ordre $p$ quand $p$ est le plus grand entier tel que $\varepsilon_n = \mathbf{O}(h_n^{p+1})$ pour toute fonction $f$ suffisamment régulière.*

On a :

**Proposition 2.1.1** *[HNW87] Dès que le problème (2.1) est suffisamment régulier, si la méthode (2.2) est d'ordre $p$, il existe des fonctions $d_i$ telles que:*

$$\varepsilon_n = d_{p+1}(y(t_n)) \, h_n^{p+1} + \cdots + d_{p+q}(y(t_n)) \, h_n^{p+q} + \mathbf{O}(h_n^{p+q+1}). \tag{2.6}$$

**Remarque** : L'entier $q$ ne dépend que de la régularité de $f$.

**Définition 2.1.7** *: La méthode (2.2) est dite* stable *s'il existe une constante $C$ indépendante de $H$ telle que, pour $H < H_0$, $y_n$ et $z_n$ donnés par :*

$$\begin{aligned} y_{n+1} &= y_n + h_n \Phi(y_n, h_n), \\ z_{n+1} &= z_n + h_n \Phi(z_n, h_n) + \epsilon_n, \end{aligned}$$

*vérifient :*

$$\max_{0 \le n \le M} \|z_n - y_n\| \quad \le \quad C \, (\|z_0 - y_0\| + \sum_{n < M} \|\epsilon_n\|).$$

La construction d'estimateurs asymptotiques repose sur le théorème suivant :

**Théorème 2.1.1** *[Gra64] On suppose que le problème (2.1) est intégré avec la méthode (2.2) stable, d'ordre $p \ge 1$, à pas constant $h$. Alors, il existe des fonctions $e_k$ telles que l'on ait*

$$y_n - y(t_n) = h^p \, e_p(t_n) + \cdots + h^{p+q} \, e_{p+q}(t_n) + \mathbf{O}(h^{p+q+1}), \tag{2.7}$$

*uniformément sur* $[0,T]$. *Les fonctions* $e_k$, $k = p, \ldots, p+q$, *sont solutions d'équations différentielles de la forme:*

$$\begin{array}{rcl} \dot{e}_k(t) & = & f'(y(t))\, e_k(t) + \Psi_k(t) \\ e_k(0) & = & 0 \end{array} \tag{2.8}$$

*où* $f'$ *est le jacobien de* $f$ *et* $\Psi_k$ *un terme inhomogène dépendant de* $\Phi$ *et de* $f$. *L'équation donnant* $e_k$ *sera appelée* la k-ième équation variationnelle.

**Preuve**: On reproduit la preuve concise de [HNW87]. On considère que les valeurs :

$$\hat{y}_n \quad = \quad y_n - h^p\, e_p(t_n),$$

sont les valeurs numériques données par le schéma:

$$\hat{y}_{n+1} \quad = \quad \hat{y}_n + h\, \hat{\Phi}(t_n, \hat{y}_n, h).$$

appliqué à (2.1). Par comparaison de ces valeurs avec $y_n$, on voit que $\hat{\Phi}$ est défini par:

$$\hat{\Phi}(t_n, \hat{y}_n, h) \quad = \quad \Phi(\hat{y}_n + h^p\, e_p(t_n), h) - h^{p-1}\, (e_p(t_n + h) - e_p(t_n)).$$

Le but est alors de montrer qu'il existe une fonction $e_p$ telle que $\hat{\Phi}$ soit d'ordre $p+1$. En utilisant le fait que $\Phi_y(y,0) = f'(y)$, le développement asymptotique de l'erreur de troncature de $\hat{\Phi}$ est :

$$\hat{\varepsilon}_n \quad = \quad (d_{p+1}(y(t_n)) - f'(y(t_n))\, e_p(t_n) + \dot{e}_p(t_n))h^{p+1} + \mathbf{O}(h^{p+2}).$$

D'où, $e_p$ défini par:

$$\begin{array}{rcl} \dot{e}_p & = & f'(y)\, e_p - d_{p+1}(y), \\ e_p(0) & = & 0, \end{array}$$

vérifie le but cherché. Les termes suivants du développement sont obtenus de manière récursive par ce procédé. □

**Remarque**:

1. Pour $k = p$, on a $\Psi_p(t) = -d_{p+1}(y(t))$ et pour $k = p+1$, avec $p \geq 2$,

$$\Psi_{p+1}(t) = -d_{p+2}(y(t)) + \frac{1}{2}\,(f'(y(t))\, d_{p+1}(y(t)) + d'_{p+1}(y(t)) f(y(t)).$$

   Les termes suivants sont d'une compléxité croissante. On les retrouvera au paragraphe §3.2.2.

2. Quand on parlera de *l'équation variationnelle* sans préciser l'indice, cela correspondra à l'équation (2.8) pour $k = p$.

On fera aussi usage du théorème suivant sur les perturbations régulières non-autonomes :

**Théorème 2.1.2** *[Sha94] Si $y$ est la solution de:*

$$\begin{aligned} y' &= f(y), \\ y(t_0) &= y_0, \end{aligned}$$

*et $u(t,\varepsilon)$ celle de:*

$$\begin{aligned} u' &= f(u) + \varepsilon\, g(t), \\ u(t_0) &= y_0 + \varepsilon\, \delta_0, \end{aligned}$$

*alors:*

$$u(t,\varepsilon) = y(t) + \varepsilon\, \delta(t) + \mathbf{O}(\varepsilon^2),$$

*où $\delta$ est donné par l'EDO linéaire inhomogène:*

$$\begin{aligned} \delta' &= f'(y)\, \delta + g(t), \\ \delta(t_0) &= \delta_0. \end{aligned}$$

On rappelle le Lemme de Gronwall :

**Lemme 2.1.1 ([Cho59])** *Soit $f(y)$ et $g(y)$ des fonctions de $\mathbb{R}^N$ dans $\mathbb{R}^N$. Soit $G$ un ouvert de $\mathbb{R}^N$. On suppose que $f$ satisfait une condition de Lipschitz par rapport à $y$ sur $G$ de constante $L$, et que :*

$$\forall y \in G, \quad \|f(y) - g(y)\| \leq K.$$

*Soient $v$ et $w$ les solutions de $v' = f(v)$, $v(0) = v_0$ et de $w' = g(w)$, $w(0) = w_0$. Alors :*

$$\forall t \in [0,T], \quad \|v(t) - w(t)\| \leq \|v_0 - w_0\| e^{Lt} + \frac{K}{L}(e^{Lt} - 1).$$

## 2.2 Méthodes de Runge-Kutta

On rappelle la définition des méthodes de Runge-Kutta (RK) :

**Définition 2.2.1** *Soit $s$ un entier, $b_i$, $c_i$, $a_{ij}$, $i,j = 1,\ldots,s$ des réels. Une* méthode de Runge-Kutta *$s$-étapes est une méthode de la forme :*

$$\begin{aligned} y_{n+1} &= y_n + h \sum_{i=1}^{s} b_i\, k_i, \\ k_i &= f(y_n + h \sum_{j=1}^{s} a_{ij}\, k_j). \end{aligned}$$

Les coefficients $c_i$ n'apparaissent pas explicitement parce que nous considérons des EDO autonomes.

Il est courant de noter la méthode précédente sous la forme du tableau suivant :

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & \vdots & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array}$$

Si la matrice $A = (a_{ij})$ est telle que $a_{ij} = 0$ pour $i \leq j$, la méthode est dite *explicite* (ERK).

Une notion importante dans l'étude des méthodes de Runge-Kutta est celle de *différentielle élémentaire*. Si $y$ est la solution de (2.1), alors, on a :

$$\ddot{y} = f'(y)f(y),$$

$$y^{(3)} = f''(y)f(y)f(y) + f'(y)f'(y)f(y),$$

etc...

Les termes $f'(y)f(y)$, $f''(y)f(y)f(y)$, $f'(y)f'(y)f(y)$ sont des différentielles élémentaires. Le problème vient de leur complexité croisante. Pour pouvoir les manipuler, il existe plusieurs notations. On peut les indicer à la manière de [Ste73], pp 114, ou de [Dor96], par des entiers.

Dans ce cas, on notera par exemple

$$\mathbf{F}_1^{(1)} = f(y),$$

$$\mathbf{F}_1^{(2)} = f'(y)f(y),$$

$$\mathbf{F}_1^{(3)} = f''(y)f(y)f(y),$$

$$\mathbf{F}_2^{(3)} = f'(y)f'(y)f(y),$$

et de manière plus générale, $\mathbf{F}_j^{(i)}$, la $j$-ème différentielle élémentaire d'ordre $i$ avec $1 \leq j \leq r_i$ où $r_i$ est le nombre de différentielles élémentaires d'ordre $i$. Il faut alors disposer d'un tableau pour établir cette correspondance entre le couple $(i, j)$ et sa différentielle élémentaire associée.

On peut aussi indicer les différentielles élémentaires par des *arbres* à la manière de Butcher [But87]. Cette méthode présente l'avantage de supprimer l'arbitraire du tableau précédent.

Soit $\bullet$ l'arbre à un nœud. On note $u = [u_1, \ldots, u_m]$ l'arbre formé par une racine connectée à $m$ branches où sont connectés les arbres $u_i$, $i = 1, \ldots, m$. On note $\mathcal{T}$, l'union de l'ensemble de tous les arbres munis d'une racine ainsi que de

l'arbre $\emptyset$ d'ordre 0. On définit alors les différentielles élémentaires de la façon suivante [But87]:

**Définition 2.2.2** *Soit $f : \mathbb{R}^N \to \mathbb{R}^N$ une fonction suffisamment différentiable. La* différentielle élémentaire $F(u) : \mathbb{R}^N \to \mathbb{R}^N$ *de $f$ associée à l'arbre $u = [u_1, \ldots, u_m] \in \mathcal{T}$ est définie par:*

$$F(u)(y) = f^{(m)}(y)(F(u_1)(y), \ldots, F(u_m)(y)),$$

*et si $u = \bullet$,*

$$F(\bullet)(y) = f(y),$$

On notera $\rho(u)$ le nombre de nœuds de l'arbre $u$.

Dans le cas où le problème (2.1) ne serait pas autonome, en ajoutant l'équation $\dot{t} = 1$, on se ramènerait au cas autonome.

Les fonctions $d_i$ du développement (2.6) s'écrivent comme une combinaison linéaire de différentielles élémentaires d'ordre $i$ appliquées en $y(t)$ :

$$d_i(y(t)) \quad = \quad \sum_{\rho(u)=i} a(u)\, F(u)(y(t)),$$

où $a : \mathcal{T} \to \mathbb{R}$ est une application ne dépendant que des coefficients de la méthode de Runge-Kutta.

L'erreur de troncature et l'erreur locale admettent alors les développements suivants :

$$\varepsilon_n \quad = \quad \sum_{u \in \mathcal{T}} a(u)\, F(u)(y(t_n))\, \frac{h_n^{\rho(u)}}{\rho(u)!},$$

$$\mathbf{le}_n \quad = \quad \sum_{u \in \mathcal{T}} a(u)\, F(u)(y_n)\, \frac{h_n^{\rho(u)}}{\rho(u)!}.$$

Ce type de développement s'appelle une $B$-série :

**Définition 2.2.3** *[HNW87] Soit $a : \mathcal{T} \to \mathbb{R}$ une application. La série formelle :*

$$B_a(x, y) \quad = \quad \sum_{u \in \mathcal{T}} a(u)\, F(u)(y)\, \frac{x^{\rho(u)}}{\rho(u)!},$$

*est appelée une* $B$-série.

## 2.3 Contrôle de l'erreur locale

Les codes d'intégration numérique effectuent leurs calculs en fonction de vecteurs de tolérances relative et absolue fournies par l'utilisateur. A chaque pas

de temps, il maintient la norme d'une estimation de l'erreur locale en dessous d'une valeur calculée d'après les tolérances fournies. Pour simplifier l'étude de ces heuristiques d'intégration, on se limitera au cas où seul un paramètre de tolérance contrôle l'intégration, la tolérance absolue $\tau$.

On considérera un contrôle de l'erreur locale par pas effectué avec une estimation valide (EPS) ou avec extrapolation locale (XEPS).

Un contrôle de l'erreur locale par pas consiste à rechercher le plus grand pas telle que l'inégalité

$$\|\mathbf{le}_n\| \leq \tau \tag{2.9}$$

reste vraie.

La norme utilisée dans un code d'intégration est d'habitude une norme euclidienne pondérée. Pour plus de simplicité, on ne considérera qu'une norme euclidienne.

Comme l'erreur locale reste une inconnue, dans l'inégalité précédente, $\mathbf{le}_n$ est remplacée par une estimation $\hat{\mathbf{le}}_n$.

Si l'estimation est valide (contrôle EPS), pour une méthode d'ordre $p$, on aura :

$$\hat{\mathbf{le}}_n \quad = \quad h_n^{p+1}\, d_{p+1}(y_n) + \sum_{k=p+2}^{r+1} h_n^k\, \hat{d}_k(y_n) + \mathbf{O}(h_n^{r+2}), \tag{2.10}$$

où $d_{p+1}$ est le terme principal de l'erreur locale de la méthode d'intégration.

Si on utilise de l'extrapolation locale (contrôle XEPS), on a :

$$\hat{\mathbf{le}}_n \quad = \quad \sum_{k=p}^{r+1} h_n^k\, \hat{d}_k(y_n) + \mathbf{O}(h_n^{r+2}). \tag{2.11}$$

On cherche donc le plus grand pas tel que l'estimation de l'erreur locale passe le test (2.9).

En pratique, un contrôle (X)EPS de l'erreur locale conduit à une relation de récurrence de la forme:

$$h_{n+1} \quad = \quad (\frac{\tau}{\|\hat{\mathbf{le}}_n\|})^{1/q} h_n,$$

où $q = p + 1$ pour un contrôle EPS et $q = p$ pour un XEPS [Sha94].

Pour éviter trop de rejets de pas, on affine la relation précédente à l'aide de coefficients de sécurité $c_M$, $c_m$ et $c$ ([HNW87], p. 167):

$$h_{n+1} \quad = \quad \min(c_M, \max(c_m, c\,(\frac{\tau}{\|\hat{\mathbf{le}}_n\|})^{1/q}))\; h_n. \tag{2.12}$$

A la relation (2.12) devrait être ajouté un moyen permettant de calculer le premier pas. Des heuristiques pour calculer un pas optimal existent (voir [HNW87], p. 182 et pour plus de détails, [GSB87]).

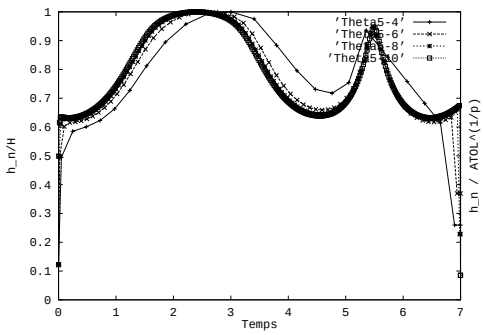



FIG. 2.1 – *Fonction de sélection - V*

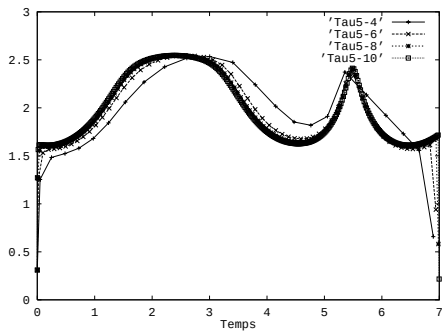



FIG. 2.2 – *Rapport $h_n/\tau^{1/5}$ - V*

La relation (2.12) définit pour chaque $\tau$ une suite unique de pas. Le contrôle de l'erreur locale qu'elle décrit sera appelé *contrôle pratique de l'erreur locale.*

On voit qu'un contrôle pratique fait intervenir de nombreux paramètres et que même en en négligeant quelques uns, il reste complexe à étudier. Une hypothèse couramment faite depuis l'ouvrage d'Henrici [Hen62] est de considérer que le pas d'intégration $h_n$ allant de $t_n$ à $t_{n+1}$ est donné par une fonction $\theta$ par $h_n = \theta(t_n)H$ où $\theta$ est une fonction fixée quand $H$ tend vers zéro.

Cependant, cette hypothèse est trop forte. D'une part, la fonction $\theta$ n'existe qu'à la limite, quand les tolérances deviennent petites, d'autre part, elle peut présenter des discontinuités. Ces deux observations conduisent à restreindre la relation donnant le pas à $h_n = \theta(t_n)\, H\, (1+\mathbf{O}(H))$ et $\theta$ admettant un nombre fini de discontinuités du premier ordre, i.e. des sauts de hauteur finie. Cela conduit à la définition d'une *fonction de sélection.*

**Définition 2.3.1** *[Sha94] Une* fonction de sélection *est une fonction $\theta : [0,T] \to ]0,1]$, continue et dérivable par morceaux, admettant un nombre fini de discontinuités, une limite à gauche et à droite de chaque discontinuité, et telle qu'il existe $\mu$ et $\zeta$ tels que : $0 < \mu \leq \theta(t)$ et $\theta(\zeta) = 1$.*

On peut voir sur les figures FIG. 2.1 et FIG. 2.3, un exemple de ce que l'on peut observer pour le rapport $h_n/H$ quand on fait diminuer la tolérance absolue d'intégration de $10^{-4}$ à $10^{-10}$ dans DOPRI5 sur un problème non-linéaire de dimension 4 et sur les problèmes A3 et A4 du package DETEST (cf Partie II). Sur la figure FIG. 2.1, on voit clairement ce rapport venir s'accumuler sur une fonction lisse sauf en $t_0$ et en $T$. En revanche, c'est beaucoup moins net pour les problèmes A3 et A4 sur les figure FIG. 2.3 et FIG. 2.5. De la même façon, en anticipant un peu sur le Chapitre 4, on a représenté sur les figures FIG. 2.2, FIG. 2.4 et FIG. 2.6, le rapport $h_n/\tau^{1/5}$. On peut remarquer la similitude des courbes entre les figures FIG. 2.1 et FIG. 2.2. C'est la même fonction sur les deux figures. La première fois, elle est divisée par $H$, la seconde, par $\tau^{1/5}$.

Avant d'essayer de traiter le cas d'un contrôle pratique de l'erreur locale, on peut étudier si le but que se fixe le contrôle de l'erreur locale préserve les

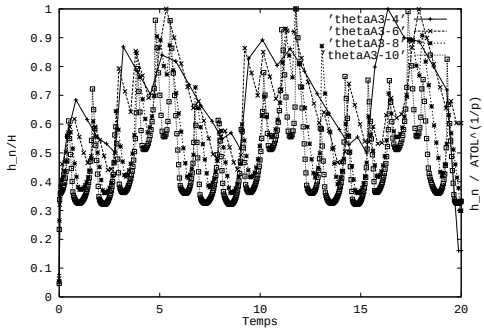



FIG. 2.3 – *Fonction de sélection - A3*

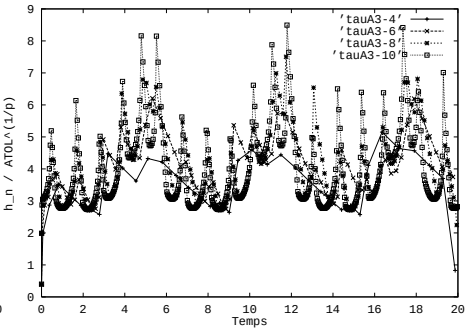



FIG. 2.4 – *Rapport $h_n/\tau^{1/5}$ - A3*

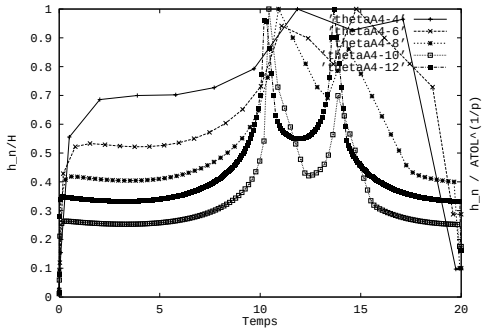



FIG. 2.5 – *Fonction de sélection - A4*

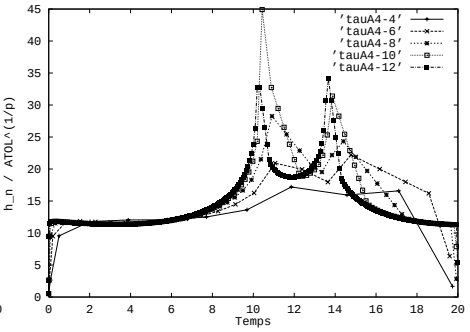



FIG. 2.6 – *Rapport $h_n/\tau^{1/5}$ - A4*

propriétés asymptotiques de l'erreur globale. Le but du contrôle de l'erreur locale est de trouver le plus grand pas $h_n$ tel que :

$$\|\hat{\mathbf{le}}_n\| \leq \tau. \tag{2.13}$$

Le problème ici est que rien ne garantit l'existence de ce plus grand pas. La situation décrite sur la figure FIG. 2.7 ne peut pas être exclue *a priori*.

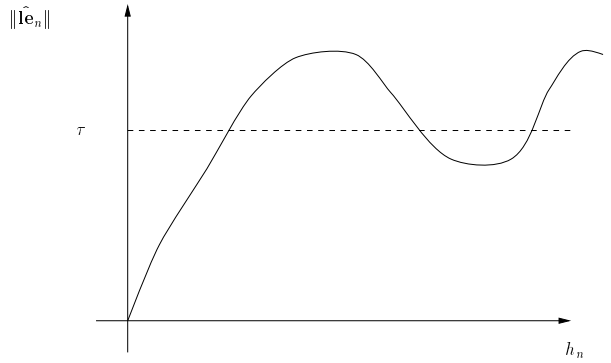



FIG. 2.7 – *Cas d'inexistence du plus grand $h_n$*

En revanche, si l'on considère un contrôle de l'erreur locale tel que ce soit le plus petit $h_n$ qui vérifie:

$$\|\hat{\mathbf{le}}_n\| \quad = \quad \tau \tag{2.14}$$

qui soit sélectionné, alors la situation précédente n'empêche plus l'existence et l'unicité du pas $h_n$. Un tel contrôle de l'erreur locale sera appelé *contrôle théorique de l'erreur locale*.

# Chapitre 3

# Estimateurs Asymptotiques

Dans ce chapitre, sont présentés les quatre estimateurs asymptotiques que nous avons retenus pour notre étude. Ils sont présentés ici dans le cas où l'intégration du problème de condition initiale (2.1) est effectuée à pas constant.

L'estimateur de Richardson et les estimateurs de Zadunaisky peuvent donner lieu à des procédés itératifs d'amélioration de la solution numérique initiale. L'estimateur de Richardson entre dans le cadre plus général des méthodes d'extrapolation, et ceux de Zadunaisky dans celui du principe de Correction Itéré du Défaut (*Iterated Defect Correction principle* [Ste80a]).

## 3.1 Estimateur de Richardson

Il se fonde sur l'existence du premier terme du développement asymptotique donné par le Théorème 2.1.1. En parallèle avec l'intégration numérique de pas $h$ est conduite une seconde intégration de pas $h/2$. On note $y^*_{2i}$ les secondes valeurs obtenues. On a:

$$\begin{aligned} y_n &= y(t_n) + h^p\, e_p(t_n) + \mathbf{O}(h^{p+1}) \\ y^*_{2n} &= y(t_n) + \frac{h^p}{2^p}\, e_p(t_n) + \mathbf{O}(h^{p+1}). \end{aligned}$$

D'où :

$$E_n = \frac{y_n - y^*_{2n}}{1 - 2^{-p}} + \mathbf{O}(h^{p+1}). \tag{3.1}$$

La valeur

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

fournit un estimateur d'ordre relatif 1. Il est d'un emploi très courant et a déjà fait l'objet d'une implantation dans un code du domaine public, GERK [SW76].

Le coût de l'estimateur de Richardson est élevé. Il multiplie par trois celui de l'intégration.

Il est possible de réduire ce coût à deux en renvoyant au lieu des valeurs $y_n$, les valeurs $y^*_{2n}$. C'est par exemple ce qui est fait dans GERK. Cependant, on n'obtient plus une estimation par pas mais seulement tous les deux pas. De plus, ce n'est pas toujours une possibilité que l'on peut adopter quand il s'agit d'implanter un estimateur dans un code déjà existant. Cela imposerait un changement indésirable dans les valeurs obtenues jusqu'alors.

## 3.2 Estimateurs de Zadunaisky

Les estimateurs de Zadunaisky se construisent sur l'idée suivante. Si l'on dispose d'un problème voisin du problème (2.1) dont on connaisse la solution exacte, alors on peut utiliser la différence entre la solution de ce problème voisin obtenue par la même méthode que celle utilisée pour le problème (2.1) et la solution exacte du problème voisin comme d'une estimation de l'erreur globale commise sur le problème (2.1).

À notre connaissance, la première référence à cette technique d'estimation de l'erreur globale remonte à [Zad66]. Ce n'est que plus tard que cette technique sera justifiée de manière heuristique dans [Zad76].

La première preuve complète de l'ordre de convergence relatif de cette technique se trouve dans [FU75]. Elle est faite dans le cas des méthodes de Runge-Kutta et de l'interpolation des valeurs $y_n$ de la solution numérique. Il est montré un résultat plus fort que la seule obtention d'une estimation valide de l'erreur globale. Il y est donné l'ordre des valeurs numériques obtenues après plusieurs itérations de cette technique d'estimation. Par la suite, dans [Fra76, Fra77, FU78], les auteurs se sont attachés à généraliser cette preuve à d'autres types de problèmes différentiels et à des classes de méthodes plus larges.

Une seconde preuve de l'ordre de convergence de l'algorithme IDeC est donnée dans [Hai78]. Elle est d'une plus grande généralité que celle que l'on trouve dans [FU75]. L'approche est différente. Les résultats sont démontrés pour toutes les méthodes numériques qui peuvent s'écrirent sous forme de $B$-séries. Ils ne se limitent pas non plus à la seule interpolation des valeurs de la solution numérique. La technique est présentée de manière explicite en terme de perturbations régulières non-autonomes et les hypothèses assurant la convergence de l'algorithme sont faites sur la perturbation. Ces hypothèses sont exprimées elles aussi à l'aide de $B$-séries. Cette preuve est d'un abord moins aisé pour un néophyte que celle de [FU75].

Enfin, on trouve dans [DDP84] une troisième preuve. On pourrait la dire intermédiaire entre celle de [FU75] et celle de [Hai78] dans le sens où elle se limite à des méthodes explicites de Runge-Kutta, mais où elle est construite pour être applicable à une classe plus large de perturbations. Nous reviendrons sur cette preuve dans le Chapitre 5, §5.2.

Le paragraphe §3.2.1 présente les estimateurs de Zadunaisky en adoptant le point de vue des perturbations régulières [Hai78]. Toutefois, les hypothèses sur la perturbation sont directement exprimées sous forme de bornes à la manière de [FU75]. En pratique, ce type de bornes est facile à montrer. Le point de vue

des perturbations est fécond et permet d'établir un lien direct entre la technique de Zadunaisky et l'Équation Modifiée (voir [Cor94] pour une introduction générale). Le paragraphe §3.2.2 reprend cette présentation dans le cas autonome. L'Équation Modifiée devient une alternative aux méthodes de perturbation fondée sur de l'interpolation numérique.

### 3.2.1 Perturbation non-autonome

On considère des perturbations du problème (2.1) de la forme:

$$\begin{aligned} \dot{\hat{y}}(t) &= f(\hat{y}(t)) + g_h(t), \\ \hat{y}(0) &= y_0. \end{aligned} \tag{3.2}$$

On suppose que l'on intègre ce problème avec la même méthode que celle utilisée sur le problème (2.1) et que l'on obtient les valeurs $\hat{y}_n$ comme valeurs approchées de $\hat{y}(t_n)$. On note $f_h(t,x) = f(x) + g_h(t)$.

S'il existe des perturbations $g_h$ telles que $\hat{y}$ soit connue, alors il est possible d'utiliser $\hat{y}_n - \hat{y}(t_n)$ comme estimation de $y_n - y(t_n)$, à condition que $g_h$ reste petite. C'est ce qui est fait dans [Zad76, FU78, DDP84, DP85].

Cette section présente les résultats connus sur l'estimateur de Zadunaisky et l'algorithme itératif qui s'en déduit, la Correction Itérée du Défaut (IDeC) [Ste78a].

#### Perturbations générales

Le principe de la preuve de l'ordre de convergence des estimateurs de Zadunaisky utilise le développement asymptotique de l'erreur globale commise sur les problèmes (2.1) et (3.2) :

$$\begin{aligned} y_n - y(t_n) &= h^p\, e_p(t_n) + \cdots + h^{p+q}\, e_{p+q}(t_n) + \mathbf{O}(h^{p+q+1}), \\ \hat{y}_n - \hat{y}(t_n) &= h^p\, \hat{e}_{h,p}(t_n) + \cdots + h^{p+q}\, \hat{e}_{h,p+q}(t_n) + \mathbf{O}(h^{p+q+1}), \end{aligned}$$

où les fonctions $\hat{e}_{h,k}$ sont elles aussi données par des équations variationnelles semblables aux relations (2.8) :

$$\begin{aligned} \dot{\hat{e}}_{h,k}(t) &= f'(\hat{y}(t))\, \hat{e}_{h,k}(t) + \hat{\Psi}_{h,k}(t), \\ \hat{e}_k(0) &= 0. \end{aligned}$$

Par différence, on a :

$$\begin{aligned} y_n - y(t_n) - (\hat{y}_n - \hat{y}(t_n)) &= h^p\, (e_p(t_n) - \hat{e}_{h,p}(t_n)) + \cdots \\ &+ h^{p+q}\, (e_{p+q}(t_n) - \hat{e}_{h,p+q}(t_n)) + \mathbf{O}(h^{p+q+1}). \end{aligned}$$

Il faut alors montrer que les différences $e_k - \hat{e}_{h,k}$ sont petites. Pour cela, il faut montrer que les fonctions $f'(y)z + \Psi_k$ et $f'(\hat{y})z + \hat{\Psi}_{h,k}$ sont voisines.

Une première étape consiste à établir la relation entre les différentielles élémentaires de $f_h$ et celles de $f$ (Lemme 3.2.1). Une fois cette borne établie pour

les différentielles élémentaires, elle s'étend aux termes inhomogènes, $\Psi_k$ et $\hat{\Psi}_{k,h}$ (Lemme 3.2.2).

Une difficulté vient de ce que même si le problème (2.1) est autonome, l'équation (3.2) ne le sera pas. Mais, comme les variables $y$ et $t$ sont séparées, les dérivées croisées en $y$ et $t$ sont toutes nulles, ce qui conduit à d'importantes simplifications. On commence donc par une proposition reliant l'ordre de la perturbation avec la différence des différentielles élémentaires perturbées et non-perturbées.

Dans les applications, les dérivées successives de la perturbation $g_h$ ne resteront pas du même ordre. Il est nécessaire de les borner, elle et toutes ses dérivées successives. On considérera que $g_h$ vérifie pour tout $k \geq 0$ :

$$g_h^{(k)} = \mathbf{O}(h^{\max(0,\min(r,m-k))}). \tag{3.3}$$

Dans les applications que nous utiliserons ici, l'ordre $r$ de la perturbation sera égal à l'ordre $p$ de la méthode d'intégration numérique. On les distingue ici dans les énoncés parce qu'il est possible de ne pas utiliser seulement des perturbations de cet ordre (voir [FU78], p 212 pour une application de ce type). L'entier $m$ correspondra à un degré de régularité dépendant de la méthode d'interpolation.

**Lemme 3.2.1** *[FU75] Soient $r \leq m$ deux entiers. Soit $\hat{y}$ la solution de (3.2). On suppose que $g_h$ vérifie la borne (3.3). On a alors:*

$$F_h(u)(\hat{y}) = F(u)(y) + \mathbf{O}(h^{\max(0,\min(r,m+1-\rho(u)))}). \tag{3.4}$$

**Lemme 3.2.2** *[FU75] Soient $r \leq m$ deux entiers. Soit $\hat{y}$ la solution de (3.2). On suppose que $g_h$ vérifie la borne (3.3). Alors, pour tout $k \geq p$,*

$$\hat{\Psi}_{k,h} = \Psi_k + \mathbf{O}(h^{\max(0,\min(r,m-k))}). \tag{3.5}$$

Une fois ces deux lemmes montrés, il est alors possible d'établir le théorème suivant. C'est une variation du théorème donné dans [Hai78], p. 410.

**Théorème 3.2.1** *[Hai78] On suppose que (2.1) est intégré avec une méthode d'ordre $p \geq 1$. On note $\hat{y}_n$ la solution numérique de (3.2) obtenue avec la même méthode. Soient $r \leq m$ deux entiers. On suppose que $g_h$ vérifie la borne (3.3). Alors,*

$$\hat{y}_n - \hat{y}(t_n) = y_n - y(t_n) + \mathbf{O}(h^{\min(p+r,m)}). \tag{3.6}$$

En règle générale, on cherche des perturbations $g_h$ telles que la solution de (3.2) soit connue. Toutefois, une hypothèse supplémentaire sur le comportement de la solution du problème perturbé permet de s'en passer.

**Corollaire 3.2.1** *[Hai78] Soient $r \leq m$ deux entiers. On suppose que (2.1) est intégré avec une méthode ERK d'ordre $p \geq 1$. On note $\hat{y}_n$, la solution numérique de (3.2) obtenue en utilisant la même méthode. On suppose que $g_h$ vérifie la borne (3.3), et que, de plus, :*

$$\hat{y}(t_n) \quad = \quad y_n + \mathbf{O}(h^m). \tag{3.7}$$

*Alors,*

$$\hat{y}_n - y_n = y_n - y(t_n) + \mathbf{O}(h^{\min(p+r,m)}). \tag{3.8}$$

**Preuve**: Il suffit de remplacer $\hat{y}(t_n)$ dans (3.6) en utilisant l'hypothèse (3.7).

□

L'intérêt de ce corollaire est d'indiquer qu'il n'est pas nécessaire de connaître la solution du problème perturbé. Il suffit de la savoir suffisamment proche de la solution numérique du problème (2.1). Il ne reste plus qu'à savoir si de telles perturbations existent.

**Exemples de perturbations**

N'importe quelle perturbation de la forme $g_h(t) = \dot{u}_h(t) - f(u_h(t))$ conduit à une solution exacte connue, $\hat{y} = u_h$. La principale difficulté consiste à trouver une fonction $u_h$ telle $g_h$ soit suffisamment petite, c'est-à-dire à trouver une solution approchée continue de (2.1). La solution continue approchée la plus simple et la plus évidente que l'on puisse construire consiste à interpoler la solution numérique elle-même [Zad76].

Soit $P_h$, la fonction définie par $P_h(t) = P_j(t)$ pour $t \in [t_{(j-1)m}, t_{jm}]$ où $P_j$ et le polynôme d'interpolation de degré $m$ des valeurs $y_i$, $i = (j-1)m, \ldots, jm$. On note $d_h(t) = \dot{P}_h(t) - f(P_h(t))$, le défaut de $P_h$ dans (2.1).

**Proposition 3.2.1** *Soit $m \geq 1$. Le défaut $d_h$ vérifie:*

$$d_h(t) \quad = \quad \sum_{k=p}^{m-1} h^k \, \Psi_k(t) + \mathbf{O}(h^{\max(0,\min(p,p+q,m))}),$$

*pour tout $k$,*

$$d_h^{(k)} \quad = \quad \mathbf{O}(h^{\max(0,\min(p,p+q-k,m-k))}).$$

**Preuve**: On reproduit l'analyse de [FU75]. D'après le Théorème (2.1.1), on a:

$$y_n - y(t_n) \quad = \quad h^p \, e_p(t_n) + \cdots + h^{p+q} \, e_{p+q}(t_n) + \mathbf{O}(h^{p+q+1}).$$

Soit $Q_h$ la fonction définie par $Q_h(t) = Q_j(t)$ pour $t \in [t_{(j-1)m}, t_{jm}]$ où $Q_j$ est le polynôme d'interpolation de degré $m$ des valeurs $y(t_i)$, $i = (j-1)m, \ldots, jm$. On a :

$$P_j \quad = \quad Q_j + h^p \, E_p^{[j]} + \cdots + h^{p+q} \, E_{p+q}^{[j]} + \mathbf{O}(h^{p+q+1}),$$

où $E_k^{[j]}$, $k = p, \ldots, p+q$ est le polynôme d'interpolation des valeurs $e_k(t_i)$, pour $i = (j-1)m, \ldots, jm$. Par conséquent, sur $[0, T]$, on a :

$$P_h \quad = \quad Q_h + h^p \, E_{h,p} + \cdots + h^{p+q} \, E_{h,p+q} + \mathbf{O}(h^{p+q+1}),$$

où $E_{h,k}(t) = E_k^{[j]}(t)$ si $t \in [t_{(j-1)\,m}, t_{j\,m}]$. D'après un résultat classique d'interpolation polynomiale [SB93], on a :

$$Q_h^{(k)} \quad = \quad y^{(k)} + \mathbf{O}(h^{\max(0,m+1-k)}). \tag{3.9}$$

D'où, en utilisant ce résultat aussi sur les $E_{h,k}$, on a :

$$P_h \quad = \quad y + h^p\, e_p + \cdots + h^m\, e_m + \mathbf{O}(h^{m+1}). \tag{3.10}$$

De plus, nous avons vu que les fonctions qui apparaissent dans le développement asymptotique de $y_i$ sont toutes différentiables. Par dérivation, on a alors:

$$\dot{P}_h \quad = \quad \dot{y} + h^p\, \dot{e}_p + \cdots + h^{m-1}\, \dot{e}_{m-1} + \mathbf{O}(h^m).$$

En utilisant le développement asymptotique de $P_h$, on a donc :

$$f(P_h) \quad = \quad f(y) + f'(y) \sum_{k=p}^{m} h^k\, e_k + \mathbf{O}(h^{\min(m+1,2p)}),$$

et avec celui de $\dot{P}_h$, on a :

$$d_h \quad = \quad \sum_{k=0}^{m-1} h^k\, (\dot{e}_k - f'(y)\, e_k) + \mathbf{O}\,(h^{\min(p,p+q-k,m-k)}).$$

Comme

$$\dot{e}_k - f'(y)\, e_k \quad = \quad \Psi_k\,.$$

on arrive au résultat souhaité.

Les bornes des dérivées de $d_h$ s'obtiennent par les relations :

$$\begin{aligned} d_h &= \dot{P}_h - \dot{y} + f(y) - f(P_h)\,, \\ \dot{d}_h &= P_h^{(2)} - y^{(2)} + f'(y)(\dot{y} - \dot{P}_h) + (f'(y) - f'(P_h))\dot{P}_h\,, \\ &\vdots \end{aligned}$$

L'ordre de $d_h^{(k)}$ est imposé par la différence $P_h^{(k)} - y^{(k)}$ donnée par les relations (3.9) et (3.10).

□

Le Théorème 3.2.1 rapproché de la Proposition 3.2.1 conduit à l'énoncé :

**Théorème 3.2.2** *On suppose que (2.1) est intégré avec une méthode ERK d'ordre $p \geq 1$. Soit $\alpha$ un réel non-nul et $m$ un entier tel que $p \leq m$. On note $\hat{y}_n$ la solution de :*

$$\begin{aligned} \dot{\hat{y}} \quad &= \quad f(\hat{y}) + \alpha\, d_h\,(t)\,, \\ \hat{y}(0) \quad &= \quad y_0\,. \end{aligned} \tag{3.11}$$

*en utilisant la même méthode que celle utilisée pour (2.1). Alors*

$$y_n + \alpha^{-1}\,(y_n - \hat{y}_n) \quad = \quad y(t_n) + \mathbf{O}(h^{\min(2p,m)}). \tag{3.12}$$

**Preuve**: En appliquant la Proposition 3.2.1 au problème (3.11), on a :

$$\begin{aligned} \dot{\hat{y}} &= f(\hat{y}) + \alpha\,(h^p \Psi_p(t) + \cdots + h^{m-1}\Psi_{m-1}(t)) + \mathbf{O}(h^m), \\ \hat{y}(0) &= y_0\,. \end{aligned}$$

En appliquant alors le Théorème 2.1.2 au problème précédent, on a :

$$\hat{y} = y + h^p \delta_p + \cdots + h^{m-1}\delta_{m-1} + \mathbf{O}(h^m),$$

avec:

$$\begin{aligned} \dot{\delta}_k(t) &= f'(y(t))\,\delta_k(t) + \alpha\,\Psi_k(t), \\ \delta_k(0) &= 0. \end{aligned}$$

Par linéarité de l'équation précédente, on a $\delta_k = \alpha\, e_k$. D'où,

$$\begin{aligned} \hat{y} - y &= \alpha \sum_{k=p}^{m-1} h^k\, e_k(t_n) + \mathbf{O}(h^m), \\ &= \alpha\,(y_n - y(t_n)) + \mathbf{O}(h^m), \end{aligned}$$

De plus, comme $\alpha\, d_h$ satisfait l'hypothèse (3.3), le Théorème 3.2.1 s'applique:

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

Cette relation peut se réécrire :

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

D'où :

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

et le résultat annoncé s'obtient alors par division par $\alpha$.

□

Une alternative à l'interpolation des valeurs $y_n$ est déjà suggérée dans [Zad76]. Elle consiste à interpoler les valeurs $f(y_i)$. Dans [DP85], il est montré que cette alternative conduit à une amélioration de l'estimateur.

**Proposition 3.2.2** *[DP85] Soit $R_h$ la fonction définie par $R_h(t) = R_j(t)$ pour $t \in [t_{(j-1)m}, t_{jm}]$ avec $R_j$ le polynôme d'interpolation de degré $m$ des valeurs $f(y_i)$, $i = (j-1)m, \ldots, jm$. Soit $S_h$ la fonction définie par $S_h(t) = S_j(t)$ pour $t \in [t_{(j-1)m}, t_{jm}]$ avec $S_j$ la fonction définie par :*

$$S_j(t) = w_j + \int_{t_{(j-1)m}}^{t} R_j,$$

*et $w_j$ définie de manière récursive par:*

$$\begin{aligned} w_{j+1} &= w_j + \int_{t_{(j-1)m}}^{t_{jm}} R_j, \\ w_1 &= y_0. \end{aligned}$$

*Soit $r_h = R_h - f(S_h)$.*

*Alors, $r_h$ vérifie:*

$$r_h^{(k)} = \mathbf{O}(h^{\max(0,\min(p,m+1-k))}).$$

Pour retrouver le résultat de [DP85], p. 487, il suffit alors d'appliquer le Théorème 3.2.1.

**Proposition 3.2.3** *[DP85] On suppose que (2.1) est intégré avec une méthode ERK d'ordre $p \geq 1$. Soit $m$ un entier tel que $p \leq m$. On note $\hat{y}_n$ la solution numérique de:*

$$\begin{aligned} \dot{\hat{y}} &= f(\hat{y}) + r_h(t), \\ \hat{y}(0) &= y_0, \end{aligned} \tag{3.13}$$

*obtenue en utilisant la même méthode. Alors,*

$$\hat{y}_n - S_h(t_n) = y_n - y(t_n) + \mathbf{O}(h^{\min(2p,m+1)}). \tag{3.14}$$

L'intérêt de cette alternative est de fournir un ordre de convergence égal à celui de l'interpolation des $y_i$ mais avec un polynôme de degré inférieur.

Un des problèmes liés aux perturbations de la forme $g_h(t) = \dot{u}_h(t) - f(u_h(t))$ est de conduire à une estimation de l'erreur globale dont le coût en terme d'évaluation de la fonction $f$, est égal à celui de l'estimateur de Richardson. Toutefois, l'hypothèse (3.3) et le Corollaire 3.2.1 peuvent être utilisés pour réduire ce coût. On a:

**Proposition 3.2.4** *[Hai78] Soit $\hat{d}_h$ la fonction polynômiale par morceaux de degré $m$ définie par $\hat{d}_h(t) = \hat{d}_j(t)$ pour $t \in [t_{(j-1)m}, t_{jm}]$ avec $\hat{d}_j$ le polynôme d'interpolation de degré $m$ des valeurs $\dot{P}_h(t_i) - f(y_i)$, $i = (j-1)m, \ldots, jm$. Alors, $\hat{d}_h$ vérifie les hypothèses (3.3) et (3.7).*

Par conséquent, l'usage d'une perturbation polynomiale par morceaux $\hat{d}_h$ conduira à une estimation de même ordre que celle obtenue avec $d_h$ mais pour un coût moindre.

### Algorithme classique de la Correction Itérée du Défaut

Le Théorème 3.2.2 permet d'améliorer la solution numérique de façon itérative, par Correction Itérative du Défaut (Iterated Defect Correction, IDeC), [FU78, Ste78a]. En fait, toutes les perturbations $g_h$ vérifiant les hypothèses du

Théorème 3.2.1 ou de son Corollaire 3.2.1 peuvent conduire elles aussi à un tel procédé itératif. On se limite ici au cas où $g_h = d_h$.

Dans [Ste80a], le principe IDeC est présenté sous deux versions différentes nommées A et B. Dans le cas de l'interpolation des valeurs $y_i$, ces deux versions correspondent respectivement à une perturbation $d_h$ et $-d_h$. Le Thérorème 3.2.1 montre qu'il n'est pas nécessaire d'établir dans ce cas, une telle distinction entre les deux versions. La première correspond à une valeur de $\alpha = 1$ et la seconde à $\alpha = -1$. Toutes les autres valeurs non nulles de $\alpha$ sont possibles.

On donne l'algorithme IDeC de [FU75] dans le cas d'une perturbation $\alpha\, d_h$. Dans [FU75], cet algorithme est envisagé de manière globale, par opposition à [Hai78] où la correction est apportée tous les $m$ pas à la solution numérique et le calcul est poursuivi à partir de cette valeur corrigée. Il devient ici :

1. intégrer (2.1) avec une méthode à un pas. Noter $y_n$ la solution obtenue. Poser $v_n^{(1)} = y_n$ et $j = 1$.

2. calculer $d_{h,j}$ à l'aide des valeurs $v_n^{(j)}$.

3. en utilisant la même méthode, intégrer :

$$\begin{aligned} \dot{\hat{y}}_{(j)} &= f(\hat{y}_{(j)}) + \alpha\, d_{h,j}(t), \\ \hat{y}_{(j)}(0) &= y_0, \end{aligned}$$

   et noter $y_n^{(j)}$ le résultat.

4. Poser $v_n^{(j+1)} = v_n^{(j)} + \alpha^{-1}\, y_n - \alpha^{-1}\, y_n^{(j)}$.

5. Aller en 2 avec $j := j + 1$.

On a alors de manière analogue au théorème 3.1, p. 7 de [FU75] :

**Théorème 3.2.3** *On a :*

$$v_n^{(j)} = y(t_n) + \mathbf{O}(h^{\min(jp,m)}). \tag{3.15}$$

Nous allons voir dans le paragrpahe précédent qu'il est possible de calquer exactement cet algorithme dans le cas autonome.

### 3.2.2 Perturbation autonome

On considère maintenant des perturbations du problème (2.1) de la forme :

$$\begin{aligned} \dot{\hat{y}} &= f(\hat{y}) + g_h(\hat{y}), \\ \hat{y}(0) &= 0. \end{aligned} \tag{3.16}$$

On note $\hat{f}(y) = f(y) + g_h(y)$ et on suppose que l'on intègre ce problème avec la même méthode que celle utilisée pour le problème (2.1). En considérant des perturbations explicites de la forme $h^r\, g_r(y) + \cdots + h^m\, g_m(y)$, on évite le

problème des bornes sur les dérivées successives de la perturbation. Ici, elles seront nécessairement vérifiées par régularité des $g_k$.

La principale différence avec le cas des perturbations non-autonomes est qu'il y a peu de chances maintenant de trouver une fonction $g_h$ telle que la solution exacte de (3.16) soit connue. Toutefois, nous avons vu que le corollaire 3.2.1 nous permet de nous affranchir de la nécessité de connaître la solution exacte du problème perturbé.

L'étape la plus importante a déjà été réalisée pour la technique d'analyse rétrograde de l'erreur que l'on appelle l'*Équation Modifiée*. Cette équation consiste précisément à chercher une EDO de la forme (3.16) telle que sa solution exacte coïncide avec la solution numérique de (2.1). Cette approche est surtout utilisée à des fins d'analyse du comportement qualitatif des méthodes numériques [GSS86, CMSS94, Hai94, HS97].

En général, une telle équation n'existe pas dès que l'on sort du cas linéaire. Par contre, il a été montré dans [Hai94, HW96] qu'il existe toujours une Équation Modifiée tronquée dont la solution passe aussi près que l'on veut de la solution numérique de (2.1). Il est possible alors de récupérer la situation du Corollaire 3.2.1 et d'en déduire un analogue du Théorème 3.2.2, puis une version autonome de l'algorithme IDeC.

### Équation Modifiée

On rappelle le théorème établissant l'existence d'une Équation Modifiée d'ordre $K$.

**Théorème 3.2.4** *[HW96] On considère une méthode stable à un pas, d'ordre $p \geq 1$. Alors, il existe des fonctions $f_i$, $i = p, \ldots, K-1$ telles que sur tout intervalle fini*

$$y_n - \hat{y}(t_n) \quad = \quad \mathbf{O}(h^K),$$

*où $\hat{y}$ est la solution de :*

$$\begin{array}{lcl} \dot{\hat{y}} & = & f(\hat{y}) + h^p\, f_p(\hat{y}) + \cdots + h^{K-1}\, f_{K-1}(\hat{y}), \\ \hat{y}(0) & = & y_0. \end{array} \tag{3.17}$$

Une formule explicite donnant les $f_i$ pour des méthodes de Runge-Kutta se trouve dans [Hai94]. Il y est montré qu'il existe des applications $c_k : \mathcal{T} \to \mathbb{R}$ telles que :

$$f_k(y) = \sum_{\rho(s)=k+1} c_k(s)\, F(s)(y). \tag{3.18}$$

D'autre part, on peut remarquer que les $f_k$ pour $k = p, \ldots, 2p-1$ sont égaux aux $\Psi_k$ apparaissant dans les équations variationnelles (2.8) (Proposition 3.2.5). Un point important à relever ici est que les $f_k$ sont des combinaisons linéaires des différentielles élémentaires de $f$ d'ordre $k+1$. Par conséquent, il suffit de

borner la différence de ces dernières avec celles de $\hat{f}$ pour obtenir la relation entre les erreurs globales commises sur (2.1) et (3.2).

**Proposition 3.2.5** *On suppose que (2.1) est intégré avec une méthode ERK d'ordre $p \geq 1$. Soient $\Psi_k$ les termes inhomogènes apparaissant dans les équations variationnelles donnant les $e_k$ de (2.8). Soient $f_k$ les fonctions apparaissant dans (3.17). Alors, pour $k = p, \ldots, 2\,p-1$, on a :*

$$\Psi_k(t) = f_k(y(t)). \tag{3.19}$$

**Preuve** : On a :

$$y_n = y(t_n) + h^p\, e_p(t_n) + \cdots + h^{N-1}\, e_{N-1}(t_n) + \mathbf{O}(h^N).$$

De manière formelle, on peut écrire:

$$y_n = y(t_n) + \sum_{k \geq p} h^k\, e_k(t_n)$$

Donc, la solution de l'Équation Modifiée $\hat{y}$ est telle que :

$$\hat{y} = y + \sum_{k \geq p} h^k\, e_k.$$

Par dérivation par rapport au temps, on a :

$$\dot{\hat{y}} = \dot{y} + \sum_{k \geq p} h^k\, \dot{e}_k.$$

Par conséquent, comme $\hat{y}$ est solution de (3.17), on a :

$$\dot{y} + \sum_{k \geq p} h^k\, \dot{e}_k \quad = \quad f(y + \sum_{k \geq p} h^k e_k) + \sum_{j \geq p} h^j f_j\,(y + \sum_{k \geq p} h^k e_k).$$

Par développement en série de Taylor de $f$ et des $f_i$ au voisinage de $y$, on a :

$$\begin{aligned}
f(\hat{y}) &= f(y) + h^p\, f'(y)\, e_p + \cdots + h^{2p-1}\, f'(y)\, e_{2p-1} \\
&+ h^{2p}(f'(y)e_{2p} + 1/2 f''(y) e_p^2) + \cdots \\
h^p\, f_p(\hat{y}) &= h^p\, f_p(y) + h^{2p}\, f_p'(y) e_p + \cdots \\
&\vdots \\
h^{2p-1}\, f_{2p-1}(\hat{y}) &= h^{2p-1}\, f_{2p-1}(y) + \cdots \\
h^{2p}\, f_{2p}(\hat{y}) &= h^{2p}\, f_{2p}(y) + \cdots
\end{aligned}$$

En rassemblant les termes de même ordre, on a $\dot{e}_k = f'(y)\, e_k + f_k(y)$ pour $k = p, \ldots, 2p-1$. Par définition de l'équation variationnelle donnant $e_k$, on a :

$$\Psi_k(t) = f_k(y(t))$$

pour $k = p, \ldots, 2p-1$, ce qui termine la preuve.

□

**Remarque**:

1. Cette relation n'est plus vraie à partir de $h^{2p}$. On a

$$\dot{e}_{2p} = f'(y)e_{2p} + f_{2p}(y) + \frac{1}{2}f''(y)e_p^2 + f_p'(y)e_p\,.$$

2. La proposition précédente avec la relation explicite donnant les $f_p$ dans [HS97] donne aussi un moyen relativement plus simple de construire les expressions algébriques donnant les termes inhomogènes que l'expression que l'on peut trouver dans [Ste73], pp 154-155.

□

On établit maintenant une proposition montrant la relation entre l'intégration du problème perturbé (3.16) et celle du problème non-perturbé. Le fait que la perturbation soit autonome complique beaucoup les relations entre les différentielles élémentaires de $f$ et celles de $\hat{f}$. Il n'y a plus les simplifications propres au cas non-autonome. Cependant, le fait de prendre la perturbation de la forme $h^r\, g(y)$ permet de dire tout de suite que toutes les différences entre les différentielles élémentaires de $f$ et celles de $\hat{f}$ seront en $\mathbf{O}(h^r)$.

**Proposition 3.2.6** *Soit $r$ un entier. On considère une perturbation de (2.1) de la forme :*

$$\begin{array}{lcl} \dot{\hat{y}} & = & f(\hat{y}) + h^r\, g(\hat{y}), \\ \hat{y}(0) & = & y_0\,, \end{array} \tag{3.20}$$

*où $g$ est infiniment différentiable. On suppose que l'on intègre ce problème avec la même méthode RK d'ordre $p \geq 1$ que pour (2.1). Alors, la solution numérique $\hat{y}_n$ de (3.20) vérifie :*

$$\hat{y}_n - \hat{y}(t_n) = y_n - y(t_n) + \mathbf{O}(h^{p+r}). \tag{3.21}$$

**Preuve**: En utilisant le Théorème 2.1.1 sur (3.20), on a :

$$\hat{y}_n - \hat{y}(t_n) \quad = \quad \sum_{k=p}^{p+q} h^k\, \hat{e}_{h,k}(t_n) + \mathbf{O}(h^{p+q+1}).$$

Les fonctions $\hat{e}_{h,k}$ sont définies cette fois par :

$$\begin{aligned} \dot{\hat{e}}_{h,k} &= \hat{f}'(\hat{y})\,\hat{e}_{h,k} + \hat{f}_{h,k}(\hat{y}), \\ \hat{e}(0) &= 0, \end{aligned}$$

où $\hat{f}(\hat{y}) = f(\hat{y}) + h^r\, g(\hat{y})$ et où la relation (3.19) a été utilisée. On a:

$$\hat{f}' = f' + h^r\, g',$$

et par régularité de $g$, $\hat{f}' = f' + \mathbf{O}(h^r)$. D'où,

$$\begin{aligned} \dot{\hat{e}}_{h,k} &= f'(\hat{y})\,\hat{e}_{h,k} + \hat{f}_{h,k}(\hat{y}) + \mathbf{O}(h^r), \\ \hat{e}(0) &= 0. \end{aligned}$$

De plus, pour tout $s \in \mathcal{T}$, par linéarité et composition, on a :

$$\hat{F}(s)(\hat{y}) = F(s)(\hat{y}) + h^r\, \hat{G}_h(s)(\hat{y}),$$

où $\hat{G}_h(s)(y)$ représente un terme contenant une combinaison de différentielles élémentaires de $f$ et de $g$ au moins borné par rapport à $h$. Comme $g$ est régulière, on a :

$$\hat{F}(s)(\hat{y}) = F(s)(\hat{y}) + \mathbf{O}(h^r).$$

Et, en utilisant le fait que les fonctions $\hat{f}_k$ sont des combinaisons linéaires de différentielles élémentaires de $\hat{f}$ d'ordre $k+1$ (relation (3.18)), on a

$$\hat{f}_{h,k}(\hat{y}) = f_k(\hat{y}) + \mathbf{O}(h^r).$$

D'où :

$$\begin{aligned} \dot{\hat{e}}_{h,k} &= f'(\hat{y})\,\hat{e}_{h,k} + f_k(\hat{y}) + \mathbf{O}(h^r), \\ \hat{e}(0) &= 0. \end{aligned}$$

Comme

$$\|F(s)(\hat{y}) - F(s)(y)\| \leq C_s \|y - \hat{y}\|,$$

et que $\|y - \hat{y}\| = \mathbf{O}(h^r)$. Par le Lemme de Gronwall 2.1.1, on a $\hat{e}_{h,k} - e_k = \mathbf{O}(h^r)$, $k = p, \ldots, 2\,p-1$. D'où:

$$\begin{aligned} \hat{y}_n - \hat{y}(t_n) &= \sum_{k=p}^{p+q} h^k\, (e_k(t_n) + \mathbf{O}(h^r)) + \mathbf{O}(h^{p+q+1}), \\ &= y_n - y(t_n) + \mathbf{O}(h^{p+r}), \end{aligned}$$

ce qui conclut la preuve.

□

Le théorème suivant établit comment l'Équation Modifiée peut être utilisée pour améliorer la solution numérique de (2.1).

**Théorème 3.2.5** *On suppose que (2.1) est intégré avec une méthode RK d'ordre $p \geq 1$. On considère :*

$$\begin{aligned} \dot{\hat{y}} &= f(\hat{y}) + \alpha \sum_{i=p}^{K-1} h^i f_i(\hat{y}), \\ \hat{y}(0) &= y_0, \end{aligned}$$

*où les $f_k$ sont les fonctions apparaissant dans l'Équation Modifiée associée à $f$ et à la méthode RK. Soit $\hat{y}_n$ sa solution numérique obtenue en utilisant la même méthode. Alors,*

$$y_n + \alpha^{-1}(y_n - \hat{y}_n) = y(t_n) + \mathbf{O}(h^{\min(2p,K)}). \tag{3.22}$$

**Preuve** : En utilisant la Proposition 3.2.6, on a :

$$\hat{y}_n - \hat{y}(t_n) = y_n - y(t_n) + \mathbf{O}(h^{2p}).$$

De plus :

$$\hat{y} = y + h^p y_p + \cdots + h^{K-1} y_{K-1} + \mathbf{O}(h^K),$$

avec

$$\dot{y}_k = f'(y) y_k + \alpha f_k(y),$$

pour $k = p, \ldots, K-1$ et $K \leq 2p$. Comme les équations variationnelles sont affines, on a $y_k = \alpha e_k$. Donc,

$$\hat{y} = y + \alpha h^p e_p + \cdots + \alpha h^{K-1} e_{K-1} + \mathbf{O}(h^K).$$

Par conséquent,

$$\hat{y}(t_n) = y(t_n) + \alpha (y_n - y(t_n)) + \mathbf{O}(h^{\min(2p,K)}).$$

En remplaçant cette relation dans la précédente, on a :

$$\frac{(\alpha - 1) y_n + \hat{y}_n}{\alpha} = y(t_n) + \mathbf{O}(h^{\min(2p,K)}).$$

□

**Algorithme autonome de la Correction Itérée du Défaut**

De même que le Théorème 3.2.2 ouvrait la porte à un algorithme itératif d'amélioration de la solution numérique, le Théorème 3.2.5 le permet aussi. L'ordre auquel on tronque l'Équation Modifiée correspond au degré des polynômes d'interpolation des $y_i$.

En prenant comme méthode de base une méthode de Runge-Kutta, on peut décrire la version autonome de l'algorithme IDeC de la façon suivante :

1. intégrer (2.1) avec la méthode à un pas. Noter $y_n$ la solution obtenue. Poser $v_n^{(1)} = y_n$ et $j = 1$.

2. calculer $F_j$, la série de l'Équation Modifiée associée aux valeurs $v_n^{(j)}$ et au problème (2.1) à l'ordre $K$.

3. en utilisant la même méthode, intégrer :

$$\begin{aligned} \dot{y} &= f(y) + \alpha\, F_j(y), \\ y(0) &= y_0, \end{aligned}$$

   et noter $y_n^{(j)}$, les valeurs obtenues.

4. Poser $v_n^{(j+1)} = v_n^{(j)} + \alpha^{-1}\, y_n - \alpha^{-1}\, y_n^{(j)}$.

5. Aller en 2 avec $j := j + 1$.

**Théorème 3.2.6** *Pour l'algorithme précédent, on a :*

$$v_n^{(j)} = y(t_n) + \mathbf{O}(h^{\min(jp,K)}). \tag{3.23}$$

**Remarque**: Pour obtenir les fonctions $F_j(y)$ à partir des valeurs $v_n^{(j)}$, on peut procéder de la manière suivante.

La fonction $F_1$ est donnée par la série de l'équation modifiée associée à $f$ et à $v_n^{(1)} = y_n$. On connait l'expression des coefficients $c_1$ de la $B$-série de $y_n$.

1. On suppose que l'on connait $F_j$ et l'expression des coefficient $c_j$ de la $B$-série en $f$ donnant $v_n^{(j)}$.

2. On a :

$$y_1^{(j)} = \sum_{u \in \mathcal{T}} \frac{h^{\rho(u)}}{\rho(u)!} \overline{\alpha}(u) a(u) \mathcal{F}_j(u)(y_0),$$

   où $\mathcal{F}_j(u)$ est la différentielle élémentaire associée a $u$ pour la fonction $f + \alpha\, F_j$. Elle dépend de $h$. Le nombre $\overline{\alpha}(u)$ est le nombre de manière d'indicer de manière croissante l'arbre $u$.

3. On la développe en fonction de $h$, on remplace dans l'expression précédente et on ré-arrange les termes pour obtenir l'expression de la $B$-série de $y_1^{(j)}$ en fonction des différentielles élémentaires de $f$. On note $c_j$ ses coefficients.

4. On a alors l'expression de la $B$-série de $v_1^{(j+1)}$. Ses coefficients sont donnés par $c_{j+1} = c_j + \dfrac{a - a_j}{\alpha}$.

5. A partir de $c_{j+1}$, on peut calculer $F_{j+1}$ à l'aide de la relation (2.12) de [Hai94].

L'étape 3 donne lieu à des calculs lourds.

**Exemple**: Soit $\dot{y} = \lambda\, y$, $y(0) = 1$, que l'on intègre avec la méthode d'Euler explicite. Dans ce cas simple, l'Équation Modifiée peut se calculer sous forme close:

$$\dot{z} \quad = \quad \frac{\ln(1+\lambda\, h)}{h}\, z.$$

Sa solution passe exactement par les valeurs $y_n = (1+\lambda\, h)^n$. Son intégration donne :

$$\hat{y}_n = (1+\hat{\lambda}\, h)^n,$$

avec $\hat{\lambda} = (1/h)\, \ln(1+\lambda\, h)$. La valeur améliorée en utilisant l'algorithme précédent pour $\alpha = 1$ est:

$$2\,(1+\lambda\, h)^n - (1+\hat{\lambda}\, h)^n.$$

L'Équation Modifiée associée à $v_n^{(2)}$ n'est pas d'un calcul direct, alors que celle recherchée pour $\alpha = -1$, c'est-à-dire pour les valeurs

$$v_n^{(2)} = \hat{y}_n$$

elles-mêmes, est :

$$\dot{z} \quad = \quad (\lambda + \lambda - \frac{\ln(1+\lambda\, h)}{h})\, z.$$

On peut alors calculer les valeurs améliorées de manière récursive par l'algorithme :

$$\begin{array}{lcl} \lambda_k & = & \lambda + \lambda_{k-1} - \frac{1}{h}\, \ln(1+\lambda_{k-1}\, h), \\ \lambda_0 & = & \lambda. \end{array} \tag{3.24}$$

On peut vérifier alors que :

$$(1+\lambda_k\, h)^{(t/h)} - e^{\lambda\, t} = \mathbf{O}(h^{k+1}).$$

□

## 3.3 Intégration de l'équation variationnelle

Les méthodes d'estimation précédentes (RS et ZD) sont coûteuses. Il est suggéré dans [Pro80] d'intégrer l'équation variationnelle donnant le premier terme du développement de l'erreur globale. L'estimation obtenue est alors identique à celle proposée dans [Ste74].

A première vue, l'équation variationnelle de l'erreur a l'air de peu d'intérêt pour estimer l'erreur globale. D'une part, elle nécessite le calcul du jacobien, d'autre part, elle nécessite la solution exacte. Toutefois, ces deux problèmes se règlent assez vite en faisant des approximations simples. On remplace le jacobien par une différence divisée et la solution exacte par la solution numérique obtenue. De plus, l'intérêt de cette équation est de pouvoir être intégrée par une méthode d'ordre peu élevé.

Ainsi, en appliquant la méthode d'Euler à (2.8) pour $k = p$, on obtient :

$$\begin{aligned} e_{p,i+1} &= e_{p,n} + h\, f'(y(t_n))e_{p,n} + h\, d_{p+1}(t_n), \\ e_{p,0} &= 0. \end{aligned}$$

On considère maintenant $v_n = h^p\, e_{p,n}$. Il est clair que $v_i$ est une estimation valide de $E_n$ puisque $v_n = h^p\, (e_p(t_n) + \mathbf{O}(h))$ et que donc, $v_n = E_n + \mathbf{O}(h^{p+1})$.

En utilisant alors le premier terme du développement de l'erreur de troncature $\varepsilon_n$, on a :

$$\begin{aligned} v_{n+1} &= v_n + h\,(f(y_n) - f(y_n - v_n)) + \varepsilon_n, \\ v_0 &= 0. \end{aligned} \tag{3.25}$$

Cette relation est décrite dans [Ste78b] comme un moyen peu coûteux permettant d'estimer l'erreur globale. Elle ne demande qu'une seule évaluation supplémentaire de la fonction $f$ par pas de calcul, $f(y_n - v_n)$.

Toutefois, elle nécessite encore le remplacement de l'erreur de troncature par une estimation. Elle restera valide si on la remplace par une estimation valide de l'erreur de troncature $\hat{\varepsilon}_n = \varepsilon_n\,(1 + \mathbf{O}(h))$.

Le fait que la plupart des codes d'intégration numérique effectue de l'extrapolation locale pour estimer l'erreur locale conduit à une estimation non-valide de l'erreur locale. Par conséquent, il est à prévoir que l'utilisation de ces valeurs dans la relation (3.25) ne conduise pas à des résultats brillants.

Il est possible d'utiliser le défaut du polynôme d'interpolation $P_h$ vu pour ZD. Effectivement, on vu que

$$\begin{aligned} h\, d_h(t_n) &= h^{p+1}\, \Psi_p(t_n) + \mathbf{O}(h^{p+2}), \\ &= h^{p+1}\, d_{p+1}(t_n) + \mathbf{O}(h^{p+2}), \\ &= \varepsilon_n + \mathbf{O}(h^{p+2}). \end{aligned}$$

## 3.4 Calcul d'une Correction Globale

La fonction d'interpolation $P_h$ décrite au paragraphe §3.2 conduit directement à un autre algorithme présenté dans [Ske86]. On considère :

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

La fonction $\mathcal{E}_h$ donne la valeur de l'erreur globale à chaque instant $t_n, n = 0, \ldots$. Elle satisfait l'équation différentielle :

$$\begin{aligned} \dot{\mathcal{E}}_h(t) &= \dot{P}_h(t) - f(P_h(t) - \mathcal{E}_h(t)), \\ \mathcal{E}_h(0) &= 0. \end{aligned} \tag{3.26}$$

L'équation (3.26) peut alors être intégrée avec une autre méthode que celle utilisée pour (2.1). On note $\varepsilon_n$ sa solution numérique. Elle fournit directement une estimation de l'erreur globale. Pour que cette estimation soit valide, il est courant de demander que la méthode utilisée pour intégrer (3.26) soit d'un ordre au moins égal à celui de la méthode employée pour (2.1) [Ske86, Pet86, DLMP89, DP89, DGP94]. Ce n'est pas nécessaire.

De plus, une remarque dans [Pet86] indique que d'autres types d'interpolation peuvent être envisagés. Dans [Pet86], l'interpolation des $f(y_i)$ est utilisée avec succès dans des tests numériques.

Par conséquent, au lieu de se limiter à $P_h$ et $S_h$, nous reprenons des perturbations de la forme $g_h(t) = \dot{u}_h(t) - f(u_h(t))$ vues au paragraphe §3.2.1. Nous supposons qu'elle vérifie une borne semblable à (3.3), pour tout $k \geq 0$ :

$$g_h^{(k)} \quad = \quad \mathbf{O}(h^{\max(0,\min(p,m-k))}). \tag{3.27}$$

Cette fois, on ne considère pas de perturbations qui seraient d'un ordre inférieur à celui de la méthode.

Et, nous posons :

$$\mathcal{E}_h = u_h - y,$$

ce qui conduit à :

$$\begin{array}{lcl} \dot{\mathcal{E}}_h(t) & = & \dot{u}_h(t) - f(u_h(t) - \mathcal{E}_h(t)), \\ \mathcal{E}_h(0) & = & u_h(0) - y(0). \end{array} \tag{3.28}$$

La condition initiale peut maintenant être non nulle.

On continue à noter $\varepsilon_n$ sa solution numérique. Maintenant, l'estimateur SC est :

$$\hat{\varepsilon}_n = \varepsilon_n - u_h(t_n) + y_n. \tag{3.29}$$

Pour montrer son ordre de convergence, on commence par un lemme :

**Lemme 3.4.1** *Soit $\hat{f}_h(t,x) = \dot{u}_h(t) - f(u_h(t) - x)$. On suppose que $g_h$ vérifie la borne (3.27). Alors pour tout $s \in \mathcal{T}$ :*

$$\hat{F}_h(s)(\mathcal{E}_h) = \mathbf{O}(h^{\max(0,\min(p,m+1-\rho(s)))}). \tag{3.30}$$

**Preuve**: En reprenant l'analyse de [Dor96], p. 236, on définit :

$$f_h(t,x) \quad = \quad f(x) + g_h(t).$$

On a alors :

$$\hat{f}_h(t,\mathcal{E}_h) \quad = \quad f_h(t,u_h) - f(y).$$

L'auteur montre alors que :

$$\hat{F}_h(s)(\mathcal{E}_h) \quad = \quad F_h(s)(u_h) - F(s)(y) + \mathbf{O}(h^p).$$

En utilisant alors le Lemme 3.2.1, on a :

$$F_h(s)(u_h) \quad = \quad F(s)(y) + \mathbf{O}(h^{\max(0,\min(p,m-\rho(s)+1))}).$$

D'où :

$$\hat{F}_h(s)(\mathcal{E}_h) \quad = \quad \mathbf{O}(h^{\max(0,\min(p,m+1-\rho(s)))}).$$

□

On peut maintenant établir l'ordre de convergence du Calcul d'une Correction Globale.

**Théorème 3.4.1** *On suppose que (2.1) est intégré avec une RK d'ordre $p \geq 1$. On suppose que $g_h$ vérifie la borne (3.27). On suppose que le problème (3.28) est intégré avec une RK d'ordre $q \geq 1$. Alors,*

$$\hat{\varepsilon}_n = y_n - y(t_n) + \mathbf{O}(h^{\min(m,p+q)}). \tag{3.31}$$

**Preuve** : L'usage d'une RK d'ordre $q$ sur (3.28) conduit à une valeur $\varepsilon_n$ vérifiant :

$$\varepsilon_n \quad = \quad \mathcal{E}_h(t_n) + h^q\, l_{h,q}(t_n) + \cdots + h^{q+p}\, l_{h,q+p}(t_n) + \mathbf{O}(h^{q+p+1}),$$

où les fonctions $l_{h,i}$, $i = q, \ldots, q+p$ sont données par :

$$\begin{aligned} \dot{l}_{h,i} &= \hat{f}'_h(\mathcal{E}_h)\, l_{h,i} + \hat{\Psi}_{h,i} \\ l_{h,i}(0) &= 0. \end{aligned}$$

et $\hat{f}_h(t,x) = \dot{u}_h(t) - f(u_h(t) - x)$. Par conséquent, $\hat{f}'_h(x) = f'(u_h - x)$ et $\hat{f}'_h(\mathcal{E}_h) = f'(y)$, qui ne dépend ni de $h$, ni de $\mathcal{E}_h$. Les équations variationnelles se réduisent alors à :

$$\begin{aligned} \dot{l}_{h,i} &= f'(y)\, l_{h,i} + \hat{\Psi}_{h,i} \\ l_{h,i}(0) &= 0. \end{aligned}$$

Ces équations sont des perturbations régulières non-autonomes de :

$$\begin{aligned} \dot{X} &= f'(y)\, X \\ X(0) &= 0, \end{aligned}$$

dont la solution est zéro. Le Lemme 3.4.1 permet d'obtenir alors la relation équivalente de la Proposition 3.2.2, pour $i \geq q$ :

$$\hat{\Psi}_i(t) \quad = \quad \mathbf{O}(h^{\max(0,\min(p,m-i))}).$$

D'où pour $i \geq q$ :

$$l_{h,i}(t) \quad = \quad \mathbf{O}(h^{\max(0,\min(p,m-i))}).$$

Et,

$$h^i\, l_{h,i}(t) \quad = \quad \mathbf{O}(h^{\max(i,\min(p+i,m))}).$$

De plus, comme :

$$\mathcal{E}_h(t_n) \quad = \quad u_h(t_n) - y_n + y_n - y(t_n),$$

on a :

$$\varepsilon_n - u_h(t_n) + y_n \quad = \quad y_n - y(t_n) + \mathbf{O}(h^{\min(p+q,m)}).$$

□

Cette proposition montre qu'il n'est pas nécessaire d'utiliser des polynômes de degré élevé pour peu qu'on l'on intègre (3.28) avec une méthode d'ordre $r < p$. Elle établit aussi que par exemple, dans le cas de l'interpolation des $y_i$, il suffit de prendre des polynômes de degré $p + r$ et une méthode de réintégration d'ordre $r$ pour garantir un ordre $p+r$. Et, dans le cas de l'interpolation des $f(y_i)$, cela signifie qu'il suffit d'utiliser une méthode d'ordre $r+1$ et des polynômes de degrés $p + r$ pour garantir un ordre $p + r + 1$.

**Remarque** : L'intégration de l'équation variationnelle n'est qu'un cas particulier de cet estimateur. On a vu au paragraphe §3.3 que quand la méthode d'Euler est appliquée à l'équation variationnelle, on obtient la relation de récurrence :

$$\begin{aligned} e_{n+1} &= e_n + h\,(f(y_n) - f(y_n - e_n)) + \hat{\varepsilon}_n, \\ e_0 &= 0, \end{aligned}$$

où $\hat{\varepsilon}_n$ est une estimation valide de l'erreur de troncature. On peut retrouver cette relation par application de la méthode d'Euler sur (3.26).

On a :

$$\mathcal{E}_{n+1} \quad = \quad \mathcal{E}_n + h\,(\dot{P}_h(t_n) - f(y_n - \mathcal{E}_n)).$$

En développant $f$ au voisinage de $P_h(t_n)$, on a :

$$\mathcal{E}_{n+1} \quad = \quad \mathcal{E}_n + h\,(\dot{P}_h(t_n) - f(y_n) + f'(y_n)\,\mathcal{E}_n)) + \mathbf{O}(h^{2p}).$$

En utilisant l'approximation

$$f'(y_n)\,\mathcal{E}_n \quad = \quad f(y_n) - f(y_n - \mathcal{E}_n) + \mathbf{O}(\mathcal{E}_n^2),$$

on obtient la relation :

$$\mathcal{E}_{n+1} \quad = \quad \mathcal{E}_n + h\,(f(y_n) - f(y_n - \mathcal{E}_n) + h\,d_h(t_n, h).$$

Comme $m \geq p + 1$, le terme $h\,d_h(t_n, h)$ est une estimation valide de l'erreur locale.

□

**Illustration Numérique**: Pour vérifier l'ordre de l'estimateur énoncé plus haut, nous avons effectué des calculs symboliques sur l'équation $\dot{y} = y^2$, $y(0) = 1$. Nous avons utilisé Maple pour calculer le développement en série de Taylor de la solution numérique et de sa valeur corrigée à l'aide de l'estimateur SC sur les $m$ premiers pas. Les résultats suivants montrent l'ordre local de cet estimateur.

Nous avons utilisé des méthodes ERK classiques d'ordre 2 à 4, ainsi que des polynômes d'interpolation des $y_i$ et des $f(y_i)$. On peut voir qu'avec $p = 3$, $m = 4$ et $q = 1$ un terme supplémentaire est obtenu dans les valeurs corrigées et qu'avec les mêmes valeurs de $p$ et de $m$, si $q = 2$ on ne gagne pas de termes supplémentaires par rapport au calcul précédent. Ici, on est limité par le degré du polynôme d'interpolation.

```
>  read SolvCor ;
```

```
 Solving For The Correction with parameters
```

$$p := 3$$

$$m := 4$$

$$q := 1$$

```
Error committed on the y[4]
```

$$ER_4 := -\frac{4}{3}\,h^4 - \frac{33}{2}\,h^5 + \mathrm{O}(h^6)$$

```
Error on the corrected value y[4] - eps[4]
```

$$sc_4 := \frac{34}{3}\,h^5 + \mathrm{O}(h^6)$$

```
>  read SolvCor ;
```

```
 Solving For The Correction with parameters
```

$$p := 3$$

$$m := 4$$

$$q := 2$$

```
Error committed on the y[4]
```

$$ER_4 := -\frac{4}{3}\,h^4 - \frac{33}{2}\,h^5 + \mathrm{O}(h^6)$$

```
Error on the corrected value y[4] - eps[4]
```

$$sc_4 := -\frac{31}{4}\,h^5 + \mathrm{O}(h^6)$$

Cela s'observe aussi avec $m = 5$ et $q$ variant de 1 à 3.

```
>  read SolvCor;
```

```
 Solving For The Correction with parameters
```

$$p := 3$$

$$m := 5$$

$$q := 1$$

```
Error committed on the y[5]
```

$$ER_5 := -\frac{5}{3}\,h^4 - \frac{635}{24}\,h^5 - \frac{6745}{24}\,h^6 + \mathrm{O}(h^7)$$

```
Error on the corrected value y[5] - eps[5]
```

$$sc_5 := -\frac{35}{6}\,h^5 - \frac{1355}{6}\,h^6 + \mathrm{O}(h^7)$$

```
>  read SolvCor ;
```

```
 Solving For The Correction with parameters
```

$$p := 3$$

$$m := 5$$

$$q := 2$$

```
Error committed on the y[5]
```

$$ER_5 := -\frac{5}{3}\,h^4 - \frac{635}{24}\,h^5 - \frac{6745}{24}\,h^6 + \mathrm{O}(h^7)$$

```
Error on the corrected value y[5] - eps[5]
```

$$sc_5 := -\frac{25}{4}\,h^6 + \mathrm{O}(h^7)$$

```
>  read SolvCor ;
```

```
 Solving For The Correction with parameters
```

$$p := 3$$

$$m := 5$$

$$q := 3$$

```
Error committed on the y[5]
```

$$ER_5 := -\frac{5}{3}\,h^4 - \frac{635}{24}\,h^5 - \frac{6745}{24}\,h^6 + \mathrm{O}(h^7)$$

```
Error on the corrected value y[5] - eps[5]
```

$$sc_5 := -\frac{3475}{384}\,h^6 + \mathrm{O}(h^7)$$

Il est aussi possible de voir que l'interpolation des valeurs $f(y_i)$ conduit à des valeurs corrigées meilleures. L'ordre de l'estimateur peut être ici $p+q-1$ avec des polynômes de degré $p+q$.

Sur le même exemple, cette propriété est observée avec une méthode de base d'ordre 2 et $m = 2$. Il suffit maintenant d'intégrer (3.28) avec la méthode d'Euler pour obtenir un terme supplémentaire de la solution exacte. Le même calcul est conduit avec $m = 3$ and $q = 2$. L'ordre local est maintenant de 5.

```
>  read SolvCor ;

 Solving For The Correction with parameters
```

$$p := 2$$

$$m := 2$$

$$q := 1$$

```
 And Interpolation of the f(y[i]) values
Error committed on the y[2]
```

$$ER_2 := -\frac{3}{2}\,h^3 - \frac{13}{2}\,h^4 + \mathrm{O}(h^5)$$

```
Errors on the corrected value y[2] - eps[2]
```

$$sc_2 := -\frac{3}{2}\,h^4 + \mathrm{O}(h^5)$$

```
>  read SolvCor ;

 Solving For The Correction with parameters
```

$$p := 2$$

$$m := 3$$

$$q := 2$$

```
 And Interpolation of the f(y[i]) values
Error committed on the y[3]
```

$$ER_3 := -\frac{9}{4}\,h^3 - \frac{33}{2}\,h^4 - \frac{339}{4}\,h^5 + \mathrm{O}(h^6)$$

```
Errors on the corrected value y[3] - eps[3]
```

$$sc_3 := -\frac{75}{16}\,h^5 + \mathrm{O}(h^6)$$

On reprend ces calculs avec une méthode de base d'ordre 3. Si l'on compare les valeurs corrigées pour $m = 3$, $q = 1$ et pour $m = 3$, $q = 2$, on remarque que dans les deux cas, seul un terme de plus est obtenu. Enfin, un calcul avec $m = 4$ et $q = 2$ montre que maintenant deux termes sont gagnés.

```
>  read SolvCor ;
```

```
 Solving For The Correction with parameters
```

$$p := 3$$

$$m := 3$$

$$q := 1$$

```
 And Interpolation of the f(y[i]) values
Error committed on the y[3]
```

$$ER_3 := -h^4 - \frac{71}{8}\,h^5 + \mathrm{O}(h^6)$$

```
Errors on the corrected value y[3] - eps[3]
```

$$sc_3 := \frac{7}{2}\,h^5 + \mathrm{O}(h^6)$$

```
>  read SolvCor ;
```

```
 Solving For The Correction with parameters
```

$$p := 3$$

$$m := 3$$

$$q := 2$$

```
 And Interpolation of the f(y[i]) values
Error committed on the y[3]
```

$$ER_3 := -h^4 - \frac{71}{8}\,h^5 - \frac{429}{8}\,h^6 + \mathrm{O}(h^7)$$

```
Errors on the corrected value y[3] - eps[3]
```

$$sc_3 := -\frac{33}{16}\,h^5 - \frac{311}{16}\,h^6 + \mathrm{O}(h^7)$$

```
>  read SolvCor ;
```

```
 Solving For The Correction with parameters
```

$$p := 3$$

$$m := 4$$

$$q := 2$$

```
 And Interpolation of the f(y[i]) values
Error committed on the y[4]
```

$$ER_4 := -\frac{4}{3}\,h^4 - \frac{33}{2}\,h^5 - \frac{275}{2}\,h^6 + \mathrm{O}(h^7)$$

```
Errors on the corrected value y[4] - eps[4]
```

$$sc_4 := -\frac{5}{3}\,h^6 + \mathrm{O}(h^7)$$

# Chapitre 4

# Analyse asymptotique d'une intégration à pas variable

Les preuves de l'ordre relatif de convergence de ZD et SC effectuées dans le Chapitre 3 l'ont été en supposant que le pas d'intégration était constant. Cette hypothèse n'est pas vérifiée dans EUROSTAG et elle ne l'est pas non plus dans les codes d'intégration numérique disponibles sur la NetLib [1].

Si l'on souhaite calquer les preuves faites dans le cas d'une intégration à pas constant, il faut pouvoir étendre le Théorème 2.1.1 à une grille convergente, c'est-à-dire dont le pas maximum tend vers zéro. Il faut tout d'abord faire une croix sur un énoncé qui nous dirait que l'on peut écrire :

$$y_n \quad = \quad y(t_n) + H^p\, e_p(t_n) + \cdots + H^{p+q}\, e_{p+q}(t_n) + \mathbf{O}(H^{p+q+1}),$$

dès que la grille serait convergente. Il est extrêmement facile d'exhiber une suite de grilles convergentes pour laquelle même le premier terme de ce développement n'existe pas. Pour cela, il suffit d'utiliser des suites extraites. On considère deux suites de grilles $(h_{N,i})_{1\leq i\leq N}$ et $(\hat{h}_{N,i})_{1\leq i\leq N}$ telles que les limites

$$\lim_{N\to\infty} \frac{y_N - y(t_N)}{H_N^p}$$

et

$$\lim_{N\to\infty} \frac{\hat{y}_N - y(t_N)}{\hat{H}_N^p}$$

existent, vaillent respectivement $E$ et $\hat{E}$ et soient différentes. Alors, la suite de grilles $(h^*_{N,i})_{1\leq i\leq N}$ définie par $h^*_{N,i} = h_{N,i}$ si $N$ est pair et $h^*_{N,i} = \hat{h}_{N,i}$ si $N$ est

1. http://www.netlib.org/ode/index.html

impair, est telle que

$$\lim_{N\to\infty} \frac{y_N^* - y(t_N)}{H_N^{*p}}$$

n'existe pas.

Une fois ce point admis, on peut ou bien chercher à caractériser les grilles pour lesquelles le développement asymptotique de l'erreur globale est préservé, ou bien chercher à montrer dans quelle mesure les algorithmes qu'utilisent les codes d'intégration numérique préservent ce développement. Il n'existe pas de théorème caractérisant les grilles pour lesquelles le développement asymptotique de l'erreur globale est préservé. Et, c'est plutôt la seconde approche, plus pragmatique, qui est adoptée dans la littérature. Cependant, même cette voie est d'un abord difficile. Les algorithmes de sélection du pas sont nombreux.

Fondés sur des approximations des solutions locales, ils utilisent des paramètres de lissage et des heuristiques optimisant leurs variations. Ils sont développés pour optimiser le coût de l'intégration et validés par l'expérience numérique, et non pas construits pour maintenir l'existence du développement de l'erreur globale.

Les extensions du Théorème 2.1.1 au cas d'une intégration à pas variable peuvent s'écrire en considérant comme paramètre indépendant de développement ou bien le pas maximum $H$ [Hen62, HNW87, Sha94], ou bien la tolérance utilisateur $\tau$ [Ste80b, Hig91, CHMR97, Stu97]. Dans les deux cas, il faut ajouter à ce paramètre un procédé pour déterminer le raffinement de la grille. Dans le premier cas, il s'agit d'une fonction $\theta$ de sélection du pas, dans le second, d'une heuristique du contrôle de l'erreur locale de type (2.12). Dans le cas où $\theta$ est supposé suffisamment régulière, on retrouve tous les termes du développement asymptotique de l'erreur globale [HNW87].

Mais, comme nous l'avons vu au paragraphe §2.3, cette hypothèse est trop forte, dans la mesure où une fonction de sélection du pas n'est observable qu'à la limite, quand la tolérance tend vers zéro.

Dans le cas où le pas courant est donné par une relation de la forme $h_n = \theta(t_n)\, H\, (1 + \mathbf{O}(H))$ avec $\theta$ une fonction de sélection du pas (cf. Définition 2.3.1, on peut montrer que le premier terme du développement est conservé. On a alors [CM84, Sha94] :

$$y_n \quad = \quad y(t_n) + H^p\, e_p(t_n) + \mathbf{O}(H^{p+1}), \tag{4.1}$$

où $e_p$ vérifie maintenant :

$$\begin{array}{lcl} \dot{e}_p & = & f'(y)\, e_p + \theta^p\, \mathbf{\Psi}_p \\ e_p(0) & = & 0 \end{array} \tag{4.2}$$

Ces résultats sont importants. Ils permettent de justifier la validité de l'estimateur de Richardson et de l'intégration de l'équation variationnelle pour une intégration à pas variable. Mais, comme rien ne peut plus être affirmé quant aux termes suivants, il n'est pas possible de calquer la preuve de [FU75] pour justifier l'ordre de ZD et celui de SC par rapport au pas maximum.

La restriction la plus forte n'est pas imposée par la régularité de la fonction $\theta$. L'hypothèse qui restreint la preuve au seul premier terme est la prise en compte du fait que la fonction de sélection n'est observable qu'à la limite, quand la tolérance utilisateur est suffisamment petite par la relation précédente $h_n = \theta(t_n)\, H + \mathbf{O}(H^2)$. La présence du $\mathbf{O}(H^2)$ empêche de montrer l'existence des termes suivants.

Un paramètre tenant mieux compte de ce qui se passe effectivement dans un code d'intégration numérique est la tolérance utilisateur. A ce paramètre, il faut ajouter un algorithme comme (2.12) ou (2.14), permettant d'assurer l'existence d'une application $\tau \mapsto (h_n)$.

Des résultats existent déjà qui expriment une relation entre l'erreur globale et la tolérance utilisateur. Ils ont été développés dans le but de trouver des conditions qui garantissent que l'erreur globale est proportionnelle à la tolérance [Sha77, Ste80b, Hig91, CHMR97, Stu97]. On dit que l'erreur globale est proportionnelle à la tolérance quand il existe une constante $C$ telle que $\|y_n - y(t_n)\|_\infty \sim C\tau$.

Cette recherche a été reprise plus récemment dans [Hig91] dans le cas d'un contrôle XEPS de l'erreur locale. Il y est montré que l'on a :

$$y_n \quad = \quad y(t_n) + \tau\, e_p(t_n) + \mathbf{O}(\tau^{1+1/p}),$$

où $e_p$ vérifie maintenant :

$$\begin{aligned} \dot{e}_p &= f'(y)\, e_p + \hat{\theta}^p\, \Psi_p, \\ e_p(0) &= 0, \end{aligned}$$

avec $\hat{\theta}(t) = 1/\|\hat{d}_p(y(t))\|$. La preuve de ce résultat fait appel à deux hypothèses :

**H1** $\lim_{\tau \to 0} H = 0$,

**H2** $\forall t \in [0,T], \hat{d}_p(y(t)) \neq 0$.

Ce dernier résultat est très similaire au précédent. Dans [Sha94], Chapitre 7, §4, le même résultat est montré de manière plus directe, en établissant une relation entre le pas maximun et la tolérance, puis en composant ce résultat avec la relation (4.1). Ce résultat explique entre autre que l'on retrouve les mêmes courbes pour les rapports $h_n/H$ et $h_n/\tau^{1/p}$ (cf. §2.3), puisque par comparaison des deux développements, on voit que $H$ est proportionnel à $\tau^{1/p}$.

Si l'intérêt de ce résultat est de tenir compte du contrôle de l'erreur locale (2.12), il n'en reste pas moins vrai que l'on ne dispose encore que d'un seul terme du développement de l'erreur globale.

Le but de ce chapitre est de montrer que l'erreur globale d'une méthode à un pas admet un développement asymptotique dans le cas d'un contrôle théorique de l'erreur locale. L'idée qui préside au travail de ce paragraphe est de montrer que dans certaines conditions le pas courant admet un développement asymptotique en fonction de la tolérance et que par conséquent, par composition, il est possible d'en déduire un développement de l'erreur globale.

Cette idée a déjà été utilisée dans [HS97] dans le cadre d'une classe particulière de méthodes à un pas, les méthodes symétriques, et d'un contrôle particulier conservant les bonnes propriétés de ces méthodes à pas variable. Il y est alors montré que l'on a :

$$y_n - y(t_n) \quad = \quad \varepsilon^p e_p(t_n) + \varepsilon^{p+1} e_{p+1}(t_n) + \cdots ,$$

avec $\varepsilon = \tau^{1/q}$, et $q$ dépend de l'utilisation ou non d'extrapolation locale. Ce résultat est montré en passant par une analyse rétrograde de l'erreur et l'utilisation de l'équation modifiée.

D'une manière plus générale, on montre dans le paragraphe §4.1, que si $h_n$ admet un développement asymptotique en fonction d'un paramètre $\delta$ et si les fonctions qui apparaissent dans ce développement sont d'une régularité suffisante, alors l'erreur globale admettra aussi un développement par rapport à $\delta$ de la forme (2.7). Ce résultat est montré de manière directe et calculatoire, sans faire usage d'analyse rétrograde de l'erreur.

Dans un deuxième temps, on montre dans le paragraphe §4.2, qu'au moins le contrôle théorique de l'erreur locale donné par (2.14) conduit à un pas qui vérifie les hypothèses nécessaires au résultat précédent. Nous faisons aussi usage des hypothèses H1 et H2.

Enfin, dans le paragraphe §4.3, on discute les restrictions possibles des hypothèses utilisées à montrer ce résultat.

## 4.1 Proposition de base

La proposition suivante montre qu'une condition suffisante pour que l'erreur globale admette un développement asymptotique est que le pas courant d'intégration en admette un aussi.

**Proposition 4.1.1** *On suppose que $h_n$ admet le développement suivant:*

$$h_n = \rho_1(t_n)\,\delta + \cdots + \rho_{2p}(t_n)\,\delta^{2p} + \mathbf{O}(\delta^{2p+1}),$$

*où $\delta \to 0$ et où les fonctions $\rho_i$ sont régulières. Alors, on a :*

$$y_n - y(t_n) = \delta^p\,\epsilon_p(t_n) + \cdots + \delta^{2p-1}\,\epsilon_{2p-1}(t_n) + \mathbf{O}(\delta^{2p}) \tag{4.3}$$

**Preuve**: Comme dans la preuve du Théorème 2.1.1, on considère que les valeurs :

$$\hat{y}_n \quad = \quad y_n - \delta^p\,\epsilon_p(t_n),$$

correspondent à la solution numérique de :

$$\hat{y}_{n+1} \quad = \quad \hat{y}_n + h_n\,\hat{\Phi}(t_n, \hat{y}_n, h_n).$$

appliquée à (2.1). On cherche alors une fonction $\epsilon_p$ telle que ces valeurs soient d'ordre $p+1$ par rapport à $\delta$.

La fonction d'incrément $\hat{\Phi}$ est définie par :

$$\hat{\Phi}(t_n, \hat{y}_n, h_n) \quad = \quad \Phi(\hat{y}_n + \delta^p\, \epsilon_p(t_n), h_n) - \frac{\delta^p}{h_n}\,(\epsilon_p(t_n + h_n) - \epsilon_p(t_n)).$$

L'erreur de troncature de $\Phi$ admet le développement :

$$\varepsilon_n \quad = \quad d_{p+1}(y(t_n))\, h_n^{p+1} + \cdots + d_{2p}(y(t_n))\, h_n^{2p} + \mathbf{O}(h_n^{2p+1}).$$

Par composition avec le développement de $h_n$, on a :

$$\begin{aligned} \varepsilon_n \quad &= \quad d_{p+1}(y(t_n))\, \rho_1^{p+1}(t_n)\, \delta^{p+1} \\ &+ \quad ((p+1)\, d_{p+1}(y(t_n))\, \rho_1^p(t_n)\, \rho_2(t_n) + d_{p+2}(y(t_n))\, \rho_1^{p+2}(t_n))\, \delta^{p+2} \\ &+ \quad \cdots + \mathbf{O}(\delta^{2p+1}). \end{aligned}$$

Il est possible de réécrire cette relation sous la forme :

$$\varepsilon_n = \sum_{k=p+1}^{2p} \gamma_k(t_n)\, \delta^k + \mathbf{O}(\delta^{2p+1}). \tag{4.4}$$

L'erreur de troncature de $\hat{\Phi}$ est :

$$\hat{\varepsilon}_n \quad = \quad y(t_n + h_n) - y(t_n) - h_n\, \Phi(y(t_n) + \delta^p\, \epsilon_p(t_n), h_n) + \delta^p\, (\epsilon_p(t_n + h_n) - \epsilon_p(t_n)).$$

Par composition de développement pour $\epsilon_p$, on a :

$$\delta^p\, (\epsilon_p(t_n + h_n) - \epsilon_p(t_n)) = \rho_1\, \dot{\epsilon}_p(t_n)\, \delta^{p+1} + \mathbf{O}(\delta^{p+2}).$$

Et, de plus, on a :

$$\begin{aligned} h_n\, \Phi(y(t_n) + \delta^p\, \epsilon_p(t_n), h_n) \quad &= \quad h_n\, \Phi(y(t_n), h_n) \\ &\quad + h_n\, \delta^p\, \frac{\partial \Phi}{\partial y}(y(t_n), h_n) + \mathbf{O}(h_n \delta^{p+1}), \end{aligned}$$

et

$$\frac{\partial \Phi}{\partial y}(y(t_n), h_n) \quad = \quad \frac{\partial \Phi}{\partial y}(y(t_n), 0) + \mathbf{O}(h_n).$$

En utilisant le fait que $h_n = \mathbf{O}(\delta)$, et par injection de ces relations dans l'erreur de troncature de $\hat{\Phi}$, on a :

$$\hat{\varepsilon}_n = (\gamma_{k+1}(t_n) - \rho_1(t_n)\, f'(y(t_n))\, \epsilon_p(t_n) + \rho_1(t_n)\, \dot{\epsilon}_p(t_n))\, \delta^{p+1} + \mathbf{O}(\delta^{p+2}).$$

Comme, $\gamma_{k+1} = \rho_1^{p+1}\, d_{p+1}(y)$, il suffit de prendre $\epsilon_p$ solution de :

$$\begin{aligned} \dot{\epsilon}_p \quad &= \quad f'(y)\, \epsilon_p - \rho_1^p\, d_{p+1}(y), \\ \epsilon_p(0) \quad &= \quad 0, \end{aligned} \tag{4.5}$$

pour garantir que $\hat{\varepsilon}_n = \mathbf{O}(\delta^{p+2})$.

Et la preuve du Théorème 2.1.1 se généralise de la même façon. □

## 4.2 Développement asymptotique du pas d'intégration

Dans ce paragraphe, on supposera vérifiées les hypothèses H1 et H2.

Avant de montrer qu'un contrôle théorique de l'erreur locale sélectionne un pas admettant un développement asymptotique par rapport à $\tau$, nous avons besoin d'un lemme. Ce lemme établit une relation entre le paramètre local $h_n$ du développement de l'erreur locale, et le paramètre global $H$.

**Lemme 4.2.1** *Si $h_n$ est sélectionné suivant un contrôle théorique de l'erreur locale, alors $H = \max_n h_n$ vérifie :*

$$H \quad = \quad \mathbf{O}(h_n), \tag{4.6}$$

*pour tout $n$.*

**Preuve** : Il suffit de montrer que $H$ et le pas minimum sont du même ordre de grandeur par rapport à $\tau$.

Pour $\tau$ suffisamment petit, par l'hypothèse H1, $h_n$ tend vers zéro. Il est par conséquent donné par :

$$h_n^{p+1} \, \|d_{p+1}(y_n)\| + \mathbf{O}(h_n^{p+2}) \quad = \quad \tau.$$

En développant $d_{p+1}$ au voisinage de $y(t_n)$, on a :

$$h_n^{p+1} \, \|d_{p+1}(y(t_n))\| + \mathbf{O}(h_n^{p+1}(y_n - y(t_n))) + \mathbf{O}(h_n^{p+2}) \quad = \quad \tau.$$

D'où, en faisant usage de l'hypothèse H2, on a :

$$1 + \mathbf{O}((y_n - y(t_n))) + \mathbf{O}(h_n) \quad = \quad \tau / (h_n^{p+1} \, \|d_{p+1}(y(t_n))\|),$$

où tous les $\mathbf{O}$ tendent vers zéro quand $\tau$ tend vers zéro. Donc, $h_n$ est équivalent :

$$h_n \quad \sim \quad (\frac{\tau}{\|d_{p+1}(y(t_n))\|})^{1/(p+1)}.$$

D'où, $H$ vérifie :

$$H \quad \sim \quad (\frac{\tau}{\mu})^{1/(p+1)},$$

avec $\mu = \min_{t \in [0,T]} \|d_{p+1}(y(t))\|$.

Et, le pas minimum $\hat{h}$ vérifie :

$$\hat{h} \quad \sim \quad (\frac{\tau}{\hat{\mu}})^{1/(p+1)},$$

avec $\hat{\mu} = \max_{t \in [0,T]} \|d_{p+1}(y(t))\|$.

D'où :

$$\frac{H}{\hat{h}} \quad \sim \quad (\frac{\hat{\mu}}{\mu})^{1/(p+1)},$$

qui est le résultat souhaité.

□

On discute maintenant du développement asymptotique du pas courant dans le cas d'un contrôle théorique fondé sur une stratégie EPS.

**Proposition 4.2.1** *On suppose que $h_n$ est sélectionné suivant un contrôle théorique et une stratégie EPS, alors, on a :*

$$\begin{aligned} h_n &= \frac{\delta}{\|d_{p+1}(y(t_n))\|^{1/(p+1)}} \\ &- \frac{1}{p+1}\frac{\hat{D}_{p+2}(y(t_n))}{\|d_{p+1}(y(t_n))\|^{1+2/(p+1)}}\,\delta^2 + \cdots + \mathbf{O}(\delta^{2\,p+1}). \end{aligned} \tag{4.7}$$

*où* $\delta = \tau^{1/(p+1)}$.

**Preuve** : Le principe de la preuve est simple. Il consiste à inverser le développement asymptotique de $\|\hat{\mathbf{le}}_n\|$ à la manière de ce que l'on peut lire dans [Car61]. Tout d'abord, en faisant usage du Lemme 4.2.1, on se débarrasse de la dépendance de l'estimation locale,$\hat{\mathbf{le}}_n$, en $y_n$ et on la remplace par une dépendance en $y(t_n)$, c'est-à-dire par une fonction qui ne dépend plus de la tolérance.

Dans le cas d'un contrôle fondé sur une stratégie EPS, l'estimation de l'erreur locale vérifie :

$$\hat{\mathbf{le}}_n = d_{p+1}(y_n)\,h_n^{p+1} + \cdots + \hat{d}_{2p}(y_n)\,h_n^{2p} + \mathbf{O}(h_n^{2p+1}).$$

On a de plus : $y_n = y(t_n) + \mathbf{O}(H^p)$. D'où :

$$\hat{\mathbf{le}}_n = (d_{p+1}(y(t_n)) + \mathbf{O}(H^p))\,h_n^{p+1} + \cdots + (\hat{d}_{2p}(y(t_n)) + \mathbf{O}(H^p))\,h_n^{2p} + \mathbf{O}(h_n^{2p+1}).$$

En faisant usage du Lemme 4.2.1, on a donc :

$$\hat{\mathbf{le}}_n = d_{p+1}(y(t_n))\,h_n^{p+1} + \cdots + \hat{d}_{2p}(y(t_n))\,h_n^{2p} + \mathbf{O}(h_n^{2p+1}).$$

Comme $d_{p+1}(y(t))$ ne s'annule jamais, on a :

$$\|\hat{\mathbf{le}}_n\| = \|d_{p+1}(y(t_n))\|\,h_n^{p+1} + \sum_{k=p+2}^{2p} h_n^k\,\hat{D}_k(t_n) + \mathbf{O}(h_n^{2p+1}),$$

où les fonctions $\hat{D}_k$ sont régulières et s'expriment en fonction de $d_{p+1}$ à $\hat{d}_i$, $i \leq k$.

A l'aide de ce développement, on recherche $h_n$ tel que :

$$\|d_{p+1}(y(t_n))\|\,h_n^{p+1} + \hat{D}_{p+2}(t_n)\,h_n^{p+2} + \cdots + \mathbf{O}(h_n^{2p+1}) = \tau.$$

En utilisant le fait que $d_{p+1}$ ne s'annule pas, cela conduit à :

$$h_n\,(1 + \frac{\hat{D}_{p+2}(t_n)}{\|d_{p+1}(y(t_n))\|}\,h_n + \cdots + \mathbf{O}(h_n^{q+2}))^{\frac{1}{p+1}} = \omega,$$

où $\omega = (\tau/\|d_{p+1}(y(t_n))\|)^{\frac{1}{p+1}}$.

Maintenant, nous devons développer une expression de la forme $(1+u)^\alpha$ avec $\alpha = 1/(p+1)$. On a :

$$\begin{aligned}(1+\sum_{k=1}^{2p}\frac{\hat{D}_{p+1+k}(t_n)}{\|d_{p+1}(y(t_n))\|}\,h_n^k+\mathbf{O}(h_n^{2p+1}))^\alpha &= 1+\alpha\,\frac{\hat{D}_{p+2}(t_n)}{\|d_{p+1}(y(t_n))\|}\,h_n \\ &\quad +(\,\alpha\,\frac{\hat{D}_{p+3}(t_n)}{\|d_{p+1}(y(t_n))\|} \\ &\quad +\frac{\alpha(\alpha-1)}{2}\,(\frac{\hat{D}_{p+2}}{\|d_{p+1}(y(t_n))\|})^2\,)\,h_n^2 \\ &\quad +\cdots+\mathbf{O}(h_n^{2p+1}),\end{aligned}$$

ce qui peut se réécrire sous la forme :

$$(1+\sum_{k=1}^{2p}\frac{\hat{D}_{p+1+k}(t_n)}{\|d_{p+1}(y(t_n))\|}\,h_n^k+\mathbf{O}(h_n^{2p+1}))^\alpha = 1+\sum_{k=1}^{2p}\gamma_{p+1+k}(t_n)\,h_n^k+\mathbf{O}(h_n^{2p+1}).$$

On a :

$$h_n\,(1+\sum_{k=1}^{2p}\gamma_{p+1+k}(t_n)h_n^k+\mathbf{O}(h_n^{2p+1})) = \omega.$$

Nous sommes maintenant dans une position où nous pouvons appliquer le théorème d'inversion d'un développement asymptotique([Car61], pp. 26). Pour $\delta$ suffisamment petit, l'équation précédente admet une solution unique donnée par :

$$h_n = \omega-\alpha\,\frac{\hat{D}_{p+2}(t_n)}{\|d_{p+1}(y(t_n))\|}\,\omega^2+\cdots+\mathbf{O}(\omega^{2p+1}).$$

Le $\mathbf{O}(\omega^{2p+1})$ se réduit à un $\mathbf{O}(\tau^{\frac{2p+1}{p+1}})$, ce qui conclut la preuve. □

**Remarques** : En utilisant les mêmes hypothèses, il est possible de montrer que dans le cas d'une stratégie XEPS, le pas vérifie :

$$h_n = \frac{\tau^{1/p}}{\|\hat{d}_p(y(t_n))\|^{1/p}}-\frac{1}{p}\,\frac{\hat{D}_{p+1}(t_n)}{\|\hat{d}_p\,(y(t_n))\|^{1+2/p}}\,\tau^{2/p}+\cdots+\mathbf{O}(\tau^{\frac{2p+1}{p}}).$$

De plus, il faut noter que les fonctions qui interviennent dans ces deux développement sont continues, grâce à l'hypothèse H2. La conjonction des Propositions 4.1.1 et 4.2.1 conduit au théorème :

**Théorème 4.2.1** *Si l'on suppose que le pas d'intégration est sélectionné suivant un contrôle théorique de l'erreur locale, alors l'erreur globale vérifie :*

$$y_n-y(t_n) = \delta^p\,\epsilon_p(t_n)+\cdots+\delta^{2p}\,\epsilon_{2p}(t_n)+\mathbf{O}(\delta^{2p+1}), \tag{4.8}$$

*où $\delta = \tau^{1/(p+1)}$ pour une stratégie EPS et $\delta = \tau^{1/p}$ pour une stratégie XEPS, et où les fonctions $\epsilon_k$ sont solutions d'équations variationnelles similaires à celle de (2.8). Et, la fonction $\epsilon_p$ vérifie :*

$$\begin{aligned} \dot{\epsilon}_p &= f'(y)\,\epsilon_p - \hat{\Psi}_p, \\ \epsilon_p(0) &= 0. \end{aligned}$$

*où $\hat{\Psi}_p = \frac{d_{p+1}}{\|d_{p+1}\|^{p/(p+1)}}$ pour une stratégie EPS et $\hat{\Psi}_p = \frac{d_{p+1}}{\|\hat{d}_p\|}$ pour une stratégie XEPS.*

## 4.3 Discussion

L'approche développée dans ce chapitre pour obtenir le développement asymptotique de l'erreur globale dans le cas d'une intégration à pas variable peut être simplifiée [Hai97]. Au lieu d'utiliser l'inversion de séries dans la Proposition 4.2.1, il est possible d'utiliser directement le théorème des fonctions implicites. Dans ce cas, au lieu d'obtenir un développement asymptotique du pas en termes de fonction du temps, on obtient des coefficients fonctions de la solution numériques. Il faut alors étendre la Proposition 4.1.1 à ce cas, ce qui peut se faire sans grandes modifications de la preuve de cette proposition.

Le résultat démontré dans ce chapitre l'a été au prix de plusieurs hypothèses restrictives des comportements possibles d'un contrôle du pas d'intégration. Nous voulons dire ici dans quelle mesure il nous semble possible d'étendre ce résultat.

La première question est de savoir si en conservant les hypothèses H1 et H2, le pas sélectionné par un contrôle pratique admet un développement asymptotique à plusieurs termes. Un résultat qui va dans ce sens se trouve dans [SN91]. Dans l'hypothèse où il n'y aurait pas depas rejeté et en supposant H1 vrai, il est montré qu'asympotiquement le pas d'intégration ne dépend que de la solution numérique et plus du pas précédent. On peut alors se ramener à la Proposition 4.1.1.

La deuxième question est de savoir si l'on peut réduire les hypothèses H1 et H2. L'hypothèse H1 n'est pas très exigeante. Le problème ici est qu'au voisinage d'un point où le terme dominant de l'estimation de l'erreur locale s'annule, le pas d'intégration risque de prendre de trop grande valeur. Typiquement, sur un problème où la méthode d'intégration numérique est exacte, le contrôle du pas donné par (2.12) conduit à une intégration en quelques pas, et le nombre de pas n'augmente pas quand on fait décroître la tolérance. Ce cas n'est malgré tout pas génant pour l'existence du développement asymptotique de l'erreur globale en fonction de la tolérance. Dans ce cas, tous les termes du développement sont nuls et ce cas ne donc pas à être traité.

En revanche, pour un contrôle de type (2.12), l'hypothèse H2 est nécessaire. Dans [CHMR97], des exemples sont donnés pour lesquels l'erreur n'est plus proportionnelle à la tolérance quand $\hat{d}_p(y(t))$ s'annule. On peut essayer de se rassurer sur la rareté de ces situations en faisant une analyse probabiliste de la

fréquence de ces situations comme dans [Stu97] ou bien proposer des contrôles du pas pour lequel même si le terme dominant de l'erreur locale s'annule, la solution de l'équation variationnelle continue à exister [CHMR97].

# Chapitre 5

# Estimateur de Zadunaisky à pas variable

Nous avons vu au paragraphe §3.2 et dans le Théorème 3.2.2 que l'ordre relatif maximum de convergence de la technique de Zadunaisky à pas constant et pour une perturbation $d_h$ est égal à l'ordre de la méthode d'intégration.

Nous nous posons maintenant la question de l'extension de ce résultat à des intégrations à pas variable.

La question est alors de savoir s'il est possible de démontrer que les variations du pas ne détruisent pas l'ordre de cet estimateur. Nous souhaitons savoir dans quelle mesure on conserve la relation $\hat{y}_n - y_n = y_n - y(t_n) + \mathbf{O}(H^{p+r})$ avec $m = p + r$, ou bien une relation semblable avec $\tau^{1/q}$ à la place de $H$ pour une intégration utilisant un contrôle de l'erreur locale. Toutefois, deux points posent problème.

Le premier est que la preuve du Théorème 3.2.2 telle que l'on peut la trouver dans [FU75] fait usage des termes de $p$ à $2p-1$ du développement de l'erreur globale. L'existence de ces termes n'étant pas démontrée dans le cas d'un contrôle pratique de l'erreur locale de la forme (2.12), il n'est pas possible de calquer la preuve de [FU75] à ce cas.

Le second est qu'il existe un contre-exemple montrant que l'ordre relatif de convergence de cet estimateur ne s'étend pas à toutes les grilles [Haï78]. Ce contre-exemple peut se reformuler et se résumer de la manière suivante.

On applique la méthode d'Euler à $y' = y$, $y(0) = 1$, avec un premier pas de longueur $h$ et un deuxième de longueur $\alpha\, h$. On a :

$$\begin{aligned} y_1 &= (1+h), \\ y_2 &= (1+h)(1+\alpha\, h). \end{aligned}$$

On considère alors $P_h$, le polynôme d'interpolation de degré 2 des valeurs $y_0$, $y_1$ et $y_2$. L'intégration de $\dot{\hat{y}} = \hat{y} + d_h(t)$, $\hat{y}(0) = 1$ fournit les valeurs $\hat{y}_1$ et $\hat{y}_2$. Les valeurs $z_i = 2\, y_i - \hat{y}_i$, $i = 1, 2$ devraient être des valeurs approchées de $e^h$ et de

$e^{(1+\alpha)h}$ à l'ordre 3. Or, on a :

$$\begin{aligned} z_1 &= 1 + h + \frac{h^2}{1+\alpha}, \\ z_2 &= 1 + (1+\alpha)\,h + (1+\alpha+2\alpha^2)\,\frac{h^2}{1+\alpha} + \mathbf{O}(h^3), \end{aligned}$$

et ces valeurs ne sont d'ordre trois que si $\alpha = 1$. Il faut relever ici que dans cet exemple, les hypothèses H1 et H2 sont vérifiées. Elles ne sont donc pas suffisantes.

Ce contre-exemple montre qu'il est déjà certain que l'ordre de l'estimateur de Zadunaisky n'est pas conservé par rapport au pas maximum. Toutefois, ce contre-exemple n'est plus valable dès que $\alpha$ n'est plus une constante, ce qui est le cas des codes d'intégration numérique. Par conséquent, il reste possible qu'en ajoutant des hypothèses de régularité dans les variations du pas, on retrouve le bon ordre.

Il est tout d'abord clair que si l'on se contente de variations de pas de la forme $h_n = \theta(t_n)\,H$ avec $\theta$ une fonction dérivable, la question ne se pose pas. En utilisant la preuve de [HNW87], p. 213-214, on se trouve dans la même situation qu'à pas constant. Le problème se présente si l'on veut tenir compte des variations observées dans un code classique d'intégration numérique.

Dans ce domaine, on trouve dans [CHMR96] un résultat particulier pour les méthodes ERK spéciales développées dans [DLMP89]. Les auteurs considèrent un contrôle de l'erreur locale de type XEPS (2.12) vérifiant les hypothèses H1 et H2 du Chapitre 4. Leurs énoncés sont établis pour des méthodes ERK particulières appelées *formules d'estimation à r-termes* de [DLMP89]. Ces méthodes prolongent celles déjà construites dans [DDP84, DP85] (cf. §5.2 de ce mémoire). Elles sont conçues pour pouvoir fournir une estimation de l'erreur globale à chaque pas de temps, et non pas tout les $m$ pas de temps dans la version classique décrite au paragraphe 3.2.1. Les méthodes développées dans [DLMP89] utilisent pour cela comme polynôme d'interpolation une extension continue fournie avec la méthode de Runge-Kutta.

Il est alors montré que, pour ces formules et pour ce contrôle du pas, l'estimateur de Zadunaisky est valide et vérifie :

$$\hat{y}_n - y_n = y_n - y(t_n) + \mathbf{O}(\tau^{\frac{p+r}{p}}).$$

Ce résultat particulier utilise le fait que l'ordre de l'estimation est donné par celui de l'expression donnée dans la relation (12) de [DLMP89], p. 839. Cette relation établit que la différence entre les erreurs locales commises par une méthode de Runge-Kutta sur les problèmes (2.1) et (3.11) s'écrit :

$$\mathbf{le}_n - \hat{\mathbf{le}}_n = \sum_{i=p+1}^{\infty} h^{i-1} \sum_{j=1}^{r_i} \tau_j^{(i)} (\mathbf{F}_{hj}^{(i)}(y(t_n)) - \mathbf{F}_j^{(i)}(y(t_n)) + \mathbf{O}(h_n^{2p}), \quad (5.1)$$

où $\mathbf{le}_n$ et $\hat{\mathbf{le}}_n$ désigne respectivement les erreurs locales commises sur (2.1) et (3.11), $\mathbf{F}_j^{(i)}$ et $\mathbf{F}_{hj}^{(i)}$ sont les différentielles élémentaires associées à (2.1) et (3.11),

$r_i$ est le nombre de différentielles élémentaires d'ordre $i$, et les $\tau_j^{(i)}$ sont des coefficients qui ne dépendent que des coefficients de la méthode de Runge-Kutta.

La preuve utilise alors le fait que pour les formules d'estimation à $r$-termes utilisant une extension continue, cette différence se réduit à un $\mathbf{O}(h_n^{p+r+1})$. Ce fait est très dépendant du polynôme d'interpolation choisi et de la méthode ERK choisie. L'estimation obtenue par ce procédé n'utilise que la dernière valeur calculée pour poursuivre son estimation.

On peut se demander s'il est possible d'étendre ce résultat à d'autres méthodes et à d'autres types d'interpolation, et, d'une manière générale, ce qui est démontrable pour cet estimateur, dans le cas d'une intégration à pas variable. Les tests numériques que l'on peut trouver dans [Zad76, AL97] et dans la Partie II de ce mémoire sont encourageants et laissent penser que même pour un code d'intégration à pas variable, et même pour des ERK quelconques, l'estimateur de Zadunaisky estime asymptotiquement de manière correcte l'erreur globale.

On montre ici deux résultats.

Dans le paragraphe §5.1, on considère que l'intégration est effectuée avec une ERK quelconque et un contrôle théorique du pas. Il est alors possible de montrer que l'estimateur classique de Zadunaisky reste valide et de donner son ordre relatif par rapport à $\tau$. Dans ce cas, il n'est plus possible d'affirmer que la relation (5.1) se réduit à un $\mathbf{O}(h_n^{\min(m,p)+1})$ dans le cas d'une intégration à pas variable fondée sur un contrôle de l'erreur locale et une tolérance $\tau$ (cf. §4). Effectivement, d'après la Proposition 3.2.1, si pour une intégration à pas constant, la relation (5.1) est de cet ordre, cette assertion repose néanmoins elle-même sur la Proposition 3.2.1 qui fait usage des termes $e_p$ à $e_{2p-1}$ du développement de l'erreur globale pour borner les dérivées successives du défaut $d_h$.

Dans le paragraphe §5.2, on se penche sur le cas des méthodes spéciales de Dormand et Prince [DDP84, DP85]. On peut montrer que leur efficacité dans l'usage de l'estimateur de Zadunaisky peut être reliée à leur propriété sur des grilles quelconques. Il n'est pas nécessaire de considérer les méthodes utilisant des extensions continues construites dans [DLMP89, DP89]. On montre que déjà, celles construites dans [DDP84, DP85], permettent d'obtenir une estimation valide de l'erreur globale par rapport au pas maximum, sans contrôle de l'erreur locale, si le premier terme du développement asymptotique de l'erreur globale par rapport au pas maximum est préservé. Cette condition ne suffit pas à assurer la validité de l'estimation de Zadunaisky quand celle-ci est employée pour une méthode de Runge-Kutta quelconque.

## 5.1 Contrôle de l'erreur locale

On ne considère ici que le cas de l'interpolation des $y_i$ telle qu'elle est décrite dans §3.2.1. De plus, on suppose que le pas courant $h_n$ est sélectionné suivant un contrôle théorique de l'erreur locale et par rapport à une tolérance $\tau$. On note $P_\tau$ l'analogue de $P_h$ de §3.2.1 et $d_\tau$ l'analogue de $d_h$.

Pour calquer la preuve de [FU75] à cette situation, il suffit de montrer que

l'on a encore :

$$F(u)(y_n) - F_\tau(u)(\hat{y}_n) \quad = \quad \mathbf{O}(\tau^{\max(0,\min(p,m+1-\rho(u)))/p}), \qquad (5.2)$$

pour un contrôle de type XEPS, par exemple. Pour prouver la relation (5.2), il suffit de montrer que le défaut $d_\tau$ et ses dérivées admettent une borne par rapport à $\tau$ analogue à celle de $d_h$ par rapport à $h$. Une fois qu'une telle borne est montrée, le Théorème 3.2.1 peut lui être appliquée.

**Proposition 5.1.1** *On suppose que (2.1) est intégré à l'aide d'une méthode de Runge-Kutta d'ordre $p \geq 1$, et que le pas d'intégration est obtenu par un contrôle théorique de l'erreur locale. On suppose que les hypothèses H1 et H2 sont vérifiées. On a :*

$$d_\tau^{(k)} \quad = \quad \mathbf{O}(\delta^{\max(0,\min(p,2p-k,m-k))}),$$

*avec $\delta = \tau^{1/(p+1)}$ pour une stratégie EPS et $\delta = \tau^{1/p}$ pour une stratégie XEPS.*

**Preuve**: Par application du Théorème 4.2.1, on a :

$$y_n - y(t_n) \quad = \quad \delta^p\, \epsilon_p(t_n) + \cdots + \delta^{2p}\, \epsilon_{2p}(t_n) + \mathbf{O}(\delta^{2p+1}),$$

avec $\delta = \tau^{1/(p+1)}$ pour une stratégie EPS et $\delta = \tau^{1/p}$ pour une stratégie XEPS.

Soit $Q_\tau$ la fonction définie par $Q_\tau(t) = Q_j(t)$ pour $t \in [t_{(j-1)m}, t_{jm}]$ avec $Q_j$ le polynôme d'interpolation de degré $m$ des valeurs $y(t_i)$, $i = (j-1)m, \ldots, jm$. On a :

$$P_j \quad = \quad Q_j + \delta^p\, E_p^{[j]} + \cdots + \delta^{2p}\, E_{2p}^{[j]} + \mathbf{O}(\delta^{2p+1}),$$

avec $E_k^{[j]}$, $k = p, \ldots, 2p$ est le polynôme d'interpolation de degré $m$ des valeurs $\epsilon_k(t_i)$, pour $i = (j-1)m, \ldots, jm$. Donc, sur $[0, T]$, on a :

$$P_\tau \quad = \quad Q_\tau + \delta^p\, E_{\tau,j} + \cdots + \delta^{2p}\, E_{\tau,2p} + \mathbf{O}(\delta^{2p+1}).$$

Le résultat sur l'interpolation lagrangienne est maintenant :

$$Q_\tau^{(k)} \quad = \quad y^{(k)} + \mathbf{O}(H^{\max(0,m+1-k)}),$$

avec $H = \max_n h_n$. En utilisant, la Proposition 4.7, on sait que $H = \mathbf{O}(\delta)$.

Par composition, on a donc :

$$P_\tau^{(k)} \quad = \quad y^{(k)} + \mathbf{O}(\delta^{\max(0,\min(p,2p-k,m+1-k))}).$$

Et la preuve se conclut de la même façon que celle de la Proposition 3.2.1.

□

Il est maintenant possible de calquer la preuve du Théorème 3.2.2 pour exprimer l'ordre de convergence de l'estimateur de Zadunaisky par rapport à $\tau$.

**Théorème 5.1.1** *On suppose que (2.1) est intégré à l'aide d'une méthode de Runge-Kutta d'ordre $p \geq 1$, et que le pas d'intégration est obtenu par un contrôle théorique de l'erreur locale. On suppose que les hypothèses H1 et H2 sont vérifiées. On note $\hat{y}_n$ la solution numérique de :*

$$\begin{aligned} \dot{\hat{y}} &= f(\hat{y}) + d_\tau(t), \\ \hat{y}(0) &= y_0. \end{aligned}$$

*obtenue en utilisant la même méthode sur la même grille. On a :*

$$\hat{y}_n - y_n = y_n - y(t_n) + \mathbf{O}(\tau^{\min(2p,m)/q}). \tag{5.3}$$

*avec $q = p + 1$ pour une stratégie EPS et $q = p$ pour une stratégie XEPS.*

## 5.2 Méthodes ERK de Dormand et Prince

Des méthodes spécifiques de Runge-Kutta on été développées dans [DDP84, DP85] pour améliorer les performances de l'estimateur de Zadunaisky. Effectivement, on le verra dans la Partie II, le fait d'utiliser de l'interpolation numérique induit des erreurs importantes dans le cas de tolérances d'intégration larges, c'est-à-dire quand l'intégration est effectuée avec de grands pas.

L'idée de la construction de ces méthodes spécifiques repose sur l'utilisation des degrés de liberté des méthodes de Runge-Kutta pour réduire le degré du polynôme d'interpolation. Le but que se fixent les auteurs est :

- ou bien, d'obtenir un ordre meilleur que celui prévu par le Théorème 3.2.2 en conservant le même degré pour les polynômes d'interpolation,
- ou bien d'obtenir le même ordre d'estimation que celui prévu par le Théorème 3.2.2 en utilisant des polynômes d'interpolation de degré inférieur.

Dans [DLMP89, DP89], cette recherche est étendue au cas où le polynôme d'interpolation est fourni par l'extension continue de la méthode de Runge-Kutta. En ce qui nous concerne on peut se limiter aux méthodes de [DDP84, DP85] et à l'interpolation lagrangienne des $y_i$ par blocs de degré $m$.

Dans ce paragraphe, nous voulons montrer que cette optimisation des méthodes de Runge-Kutta vis-à-vis de l'estimateur de Zadunaisky leur procure de meilleurs propriétés sur des grilles non-équidistantes. Pour y parvenir, nous commençons par reproduire les principales étapes de l'analyse conduite dans [DDP84] dans le cas d'une intégration à pas constant. Puis, nous montrons comment cette analyse s'adapte au cas des grilles non-équidistantes et ce qu'elle implique.

Nous adoptons ici la notation des différentielles élémentaires de [DDP84, DP85] et, de même, nous noterons $\mathbf{E}_n = y_n - y(t_n) - \hat{y}_n + y_n$, la différence entre

l'erreur globale et l'estimation de Zadunaisky. Ce changement de notations par rapport au paragraphe 3.2 s'explique par le fait que je souhaite rester au plus près des calculs et des expressions [DDP84]. La raison en est que l'on admet les détails de leur analyse et en particulier, la formule donnant $\mathbf{p}_n$ (cf. plus bas) en fonction des différentielles élémentaires.

Ainsi, après quelques calculs, il leur est possible d'exprimer sous forme d'une relation de récurrence la propagation de l'erreur commise sur l'estimation de l'erreur globale :

$$\mathbf{E}_{n+1} \quad = \quad \mathbf{E}_n + h\,\mathbf{G}_n(\mathbf{E}_n) + h\,\mathbf{p}_n. \tag{5.4}$$

Et, en utilisant le Lemme de Gragg [Gra64], on obtient la majoration :

$$\|\mathbf{E}_n\| \quad \leq \quad \frac{\exp(C\,(t_n - t_0)) - 1}{C}\, max_n \|\mathbf{p}_n\|, \tag{5.5}$$

où $C$ est une constante telle que $\|\mathbf{G}_n(\mathbf{E}_n)\| \leq C\,\|\mathbf{E}_n\|$. Le terme $\mathbf{p}_n$ est donné par l'expression :

$$\begin{aligned} \mathbf{p}_n \quad &= \quad \sum_{i \geq p+1} h^{i-1} \sum_{j=1}^{r_i} \tau_j^{(i)}\,(\mathbf{F}_{hj}^{(i)} - \mathbf{F}_j^{(i)}) \\ &+ \quad \sum_{i \geq 1} h^{i-1} \sum_{j=1}^{r_i} \zeta_j^{(i)} (\mathbf{F}_{hjk1}^{(i)} - \mathbf{F}_{jk1}^{(i)})^{k1} (y_n - y(t_n)) + \mathbf{O}(h^{2p}), \end{aligned} \tag{5.6}$$

où $\mathbf{F}_{hj}^{(i)}$ sont les différentielles élémentaires correspondant au problème perturbé (3.2) avec comme perturbation $d_h$, $r_i$ est le nombre de différentielles élémentaires d'ordre $i$, $\tau_j^{(i)}$ et $\zeta_j^{(i)}$ sont des coefficients dépendants des coefficients de la méthode de Runge-Kutta, l'indice $k1$ marque la dérivée partielle par rapport à la première composante de $y_n - y(t_n)$ et l'exposant $k1$ devant $y_n - y(t_n)$ dénote la première composante du vecteur. Toutes les différentielles élémentaires sont prises en $y(t_n)$.

Dès que $p > 1$, il est possible de simplifier le terme $\mathbf{p}_n$. Comme $y_n - y(t_n) = \mathbf{O}(h^p)$, le terme

$$\sum_{i \geq 1} h^{i-1} \sum_{j=1}^{r_i} \zeta_j^{(i)} (\mathbf{F}_{hjt}^{(i)} - \mathbf{F}_{jt}^{(i)})^t (y_n - y(t_n)) + \mathbf{O}(h^{2p})$$

n'a pas d'influence sur $\mathbf{E}_n$.

Par conséquent, par abus de notation et en se fendant de la même lettre pour désigner deux termes différents, le terme dont l'ordre détermine celui de $\mathbf{E}_n$ se ramène à :

$$\mathbf{p}_n \quad = \quad \sum_{i \geq p+1} h^{i-1} \sum_{j=1}^{r_i} \tau_j^{(i)}\,(\mathbf{F}_{hj}^{(i)} - \mathbf{F}_j^{(i)}) \tag{5.7}$$

où toutes les différentielles élémentaires sont prises en $y(t_n)$.

Maintenant, pour borner $\mathbf{p}_n$, on se sert de la majoration de la Proposition 3.2.1 :

$$d_h^{(j)} = \begin{cases} \mathbf{O}(h^{\min(p,p+q-j,m-j)}) & j = 0, \ldots, m-1 \\ \mathbf{O}(1) & m \leq j, \end{cases} \tag{5.8}$$

où $p+q$ est le nombre de termes existant dans le développement asymptotique de l'erreur globale. La majoration (5.8) peut être utilisée pour trier les termes en $h^p$ et en $h^{p+1}$ de (5.7). Pour comprendre comment s'effectue ce tri, nous le décrivons brièvement sur un exemple. On prend $p = 2$. L'expression (5.7) devient :

$$\mathbf{p}_n = h^2 \sum_{j=1}^{r_3} \tau_j^{(3)} (\mathbf{F}_{hj}^{(3)} - \mathbf{F}_j^{(3)}) + h^3 \sum_{j=1}^{r_4} \tau_j^{(4)} (\mathbf{F}_{hj}^{(4)} - \mathbf{F}_j^{(4)}) + \cdots .$$

On utilise alors le fait que :

$$\begin{aligned}
\mathbf{F}_{h1}^{(3)} - \mathbf{F}_1^{(3)} &= d_h'' + 2\, f'' f d_h + f'' d_h^2, \\
\mathbf{F}_{h2}^{(3)} - \mathbf{F}_2^{(3)} &= f' f' d_h + f' d_h', \\
\mathbf{F}_{h1}^{(4)} - \mathbf{F}_1^{(4)} &= 3\, f''' f^2 d_h + 3\, f''' f d_h^2 + f''' d_h^3 + d_h''', \\
\mathbf{F}_{h2}^{(4)} - \mathbf{F}_2^{(4)} &= f'' f' f d_h + f'' f' d_h f + f'' f' d_h^2 + f'' d_h' f + f'' d_h d_h', \\
\mathbf{F}_{h3}^{(4)} - \mathbf{F}_3^{(4)} &= 2\, f' f'' f d_h + f' f'' d_h^2 + f' d_h'', \\
\mathbf{F}_{h4}^{(4)} - \mathbf{F}_4^{(4)} &= f' f' f' d_h + f' f' d_h'.
\end{aligned}$$

D'où :

$$\begin{aligned}
\mathbf{p}_n &= h^2 \{\tau_1^{(3)} (d_h'' + 2\, f'' f d_h + f'' d_h^2) + \tau_2^{(3)} (f' f' d_h + f' d_h')\} \\
&+ h^3 \{\tau_1^{(4)} (3\, f''' f^2 d_h + 3\, f''' f d_h^2 + f''' d_h^3 + d_h''') \\
&+ \tau_2^{(4)} (f'' f' f d_h + f'' f' d_h f + f'' f' d_h^2 + f'' d_h' f + f'' d_h d_h') \\
&+ \tau_3^{(4)} (2\, f' f'' f d_h + f' f'' d_h^2 + f' d_h'') + \tau_4^{(4)} (f' f' f' d_h + f' f' d_h')\} + \cdots .
\end{aligned}$$

Maintenant, si $m = 2$, la borne (5.8) indique que $d_h = \mathbf{O}(h^2)$, $d_h' = \mathbf{O}(h)$, $d_h'' = \mathbf{O}(1)$ et $d_h''' = \mathbf{O}(1)$. Donc, $h^2\, d_h''$ produira un terme au moins d'ordre deux, et par conséquent, le coefficient $\tau_1^{(3)}$ apparaîtra dans tous les termes de $\mathbf{p}_n$. En gardant cela à l'esprit, on a :

$$\mathbf{p}_n = h^2 \{\mathbf{A}\, \tau_1^{(3)}\} + h^3 \{\mathbf{B}_1\, \tau_1^{(3)} + \mathbf{B}_2\, \tau_2^{(3)} + \mathbf{B}_3\, \tau_1^{(4)} + \mathbf{B}_4\, \tau_3^{(4)}\} + \cdots ,$$

où $\mathbf{A}$ et les $\mathbf{B}_i$ sont indépendants de $h$, et où tous les autres termes sont au moins d'ordre quatre.

| $p$ | $m$ | $h^p$ | $h^{p+1}$ |
|---|---|---|---|
| 2 | 2 | $\tau_1^{(3)}$ | $\tau_i^{(3)}$,$\tau_j^{(4)}$ $i=1,2$, $j=1,3$ |
| | 3 | 0 | $\tau_1^{(3)}$, $\tau_1^{(4)}$ |
| 3 | 2 | $\tau_1^{(4)}$, $\tau_3^{(4)}$ | $\tau_i^{(4)}$,$\tau_j^{(5)}$, $i=1,2,3,4$, $j=1,4,5,8$ |
| | 3 | $\tau_1^{(4)}$ | $\tau_1^{(4)}$,$\tau_3^{(4)}$,$\tau_1^{(5)}$,$\tau_5^{(5)}$ |
| | 4 | 0 | $\tau_1^{(4)}$,$\tau_1^{(5)}$ |
| 4 | 2 | $\tau_i^{(5)}$, i=1,4,5,8 | $\tau_i^{(5)}$,$\tau_j^{(6)}$,$i=1,...,9$,$j=1,4,5,6,7,13,14,15,19$ |
| | 3 | $\tau_1^{(5)}$,$\tau_5^{(5)}$ | $\tau_i^{(5)}$,$\tau_j^{(6)}$, $i=1,4,5,8$, $j=1,6,7,15$ |
| | 4 | $\tau_1^{(5)}$ | $\tau_1^{(5)}$,$\tau_5^{(5)}$,$\tau_1^{(6)}$,$\tau_7^{(6)}$ |
| | 5 | 0 | $\tau_1^{(5)}$,$\tau_1^{(6)}$ |

TAB. 5.1 – *Coefficients RK pour l'interpolation des $y_n$*

En utilisant cette règle, il est possible d'établir la liste des $\tau_j^{(i)}$ qui apparaissent dans les termes $h^p$ et $h^{p+1}$. Cette table est la Table 3 dans [DDP84] et la Table 1 dans [DP85]. On la reproduit ici sur la table TAB. 5.1. On y lit que pour $p=2$ et $m=2$, seul $\tau_1^{(3)}$ apparaît dans le terme $h^2$.

Cette procédure permet de construire des ERK d'ordre $p$ donnant des estimations à $r$-termes à l'aide de l'estimation de Zadunaisky, c'est-à-dire telles que $\hat{y}_n - y_n = y_n - y(t_n) + \mathbf{O}(h^{p+r})$. Le nombre $r$ dépend à la fois de $p$ et de $m$. Dans les cas que nous verrons, il n'excède jamais deux. Les méthodes spécifiques développées de cette manière seront appelées des *formules d'estimation à $r$ termes*.

Typiquement, dans [DDP84], cinq de ces méthodes sont construites. Elles sont désignées par les sigles suivants: RK2(1)2G, RK3(2)3G1, RK3(2)3G2, RK4(3)5FG, RK4(3)5G, où RKp(q)S signifie que l'on a affaire à une méthode RK d'ordre p à S étapes avec une formule emboîtée d'ordre q. On peut lire sur la table TAB. 5.2 l'ordre relatif de l'estimation de Zadunaisky pour l'interpolation des $y_i$ et différents types de méthodes RK.

Nous en venons maintenant à leur propriété sur des grilles non-équidistantes qui préservent le premier terme du développement asymptotique de l'erreur globale.

**Proposition 5.2.1** *Même si l'intégration ne préserve que le premier terme du*

| $p$ | $m$ | RK classique | RKDP | |
|---|---|---|---|---|
| 2 | 2 | 0 | 1 | |
| | 3 | 1 | 1 | |
| | 4 | 2 | 2 | |
| | | | RK3(2)3G1 | RK3(2)3G2 |
| 3 | 2 | 0 | 1 | 0 |
| | 3 | 0 | 1 | 0 |
| | 4 | 1 | 1 | 2 |
| | | | RK4(3)5FG | RK4(3)5G |
| 4 | 2 | 0 | 0 | 1 |
| | 3 | 0 | 1 | 2 |
| | 4 | 0 | 2 | 2 |
| | 5 | 1 | 2 | 2 |

TAB. 5.2 – *Ordre relatif de l'estimation pour des ERK classiques et des ERK de Dormand et Prince pour l'interpolation des $y_n$ à pas constant*

*développement asymptotique par rapport au pas maximum $H = \max_n h_n$, une formule d'estimation à r-termes continuera à fournir une estimation valide de l'erreur globale, i.e. on aura au moins $\hat{y}_n - y_n = y_n - y(t_n) + \mathbf{O}(H^{p+1})$*

**Remarque**: Cette proposition présente un intérêt, par exemple, dans le cas $p = 2$ et $m = 3$. Dans ce cas, la simple existence du premier terme du développement asymptotique ne suffit pas à garantir une estimation valide pour une ERK quelconque. Par contre, cela suffit à la RK3(2)3G1.

**Preuve**: L'analyse conduite plus haut dans le cas d'une intégration à pas constant reste vraie avec $h$ changé en $h_n$ dans la relation de récurrence (5.4). Elle devient maintenant:

$$\mathbf{E}_{n+1} \quad = \quad \mathbf{E}_n + h_n\, \mathbf{G}_n(\mathbf{E}_n) + h_n\, \mathbf{p}_n.$$

Le Lemme de Gragg s'applique aussi à une grille non-équidistante ([CM84], p. 76) et l'on obtient la même borne que (5.5). Le terme dont l'ordre détermine celui de $\mathbf{E}_n$ est toujours $\mathbf{p}_n$. Il est donné par l'expression:

$$\begin{aligned} \mathbf{p}_n \quad &= \quad \sum_{i \geq p+1} h_n^{i-1} \sum_{j=1}^{r_i} \tau_j^{(i)} \, (\mathbf{F}_{\mathcal{H}j}^{(i)} - \mathbf{F}_j^{(i)}) \\ &+ \quad \sum_{i \geq 1} h_n^{i-1} \sum_{j=1}^{r_i} \zeta_j^{(i)} (\mathbf{F}_{\mathcal{H}jt}^{(i)} - \mathbf{F}_{jt}^{(i)})^t (y_n - y(t_n)) + \mathbf{O}(h_n^{2p}), \end{aligned}$$

où les différentielles élémentaires associées aux problèmes (2.1) et (3.11) dépendent maintenant de tous les pas d'intégration, ce que l'on résume par l'indice $\mathcal{H}$.

Maintenant, si la grille ne préserve que le premier terme du développement asymptotique de l'erreur globale, on peut appliquer la Proposition 3.2.1 avec $q = 0$, et la borne devient :

$$d_{\mathcal{H}}^{(j)} = \begin{cases} \mathbf{O}(H^{\min(p-j,m-j)}) & j = 0, \ldots, m-1 \\ \mathbf{O}(1) & m \leq j, \end{cases} \tag{5.9}$$

La simplification faite dans la relation (5.6) reste toujours valable dès que $p > 1$, car quel que soit la grille, on a $y_n - y(t_n) = \mathbf{O}(H^p)$. De plus, la différence $\mathbf{F}_{\mathcal{H}jt}^{(i)} - \mathbf{F}_{jt}^{(i)}$ reste un $\mathbf{O}(H^p)$. Par conséquent, le terme qui contrôle l'ordre de l'estimation reste :

$$\mathbf{p}_n = \sum_{i \geq p+1} h_n^{i-1} \sum_{j=1}^{r_i} \tau_j^{(i)} \, (\mathbf{F}_{\mathcal{H}j}^{(i)} - \mathbf{F}_j^{(i)}). \tag{5.10}$$

Un changement important entre la borne donnée dans la Proposition (3.2.1) pour le pas constant et la borne (5.9) est que maintenant il est possible d'obtenir une dérivé $k$-ième non-bornée en prenant $m$ trop grand par rapport à $p$.

Néanmoins, la borne (5.9) demeure identique à la borne (5.8) dès que $m \leq p$. Par conséquent, les mêmes $\tau_j^{(i)}$ feront leur apparition dans les deux cas du développement de $\mathbf{p}_n$. Par conséquent, on aura :

$$\mathbf{p}_n = \mathbf{O}(H^{p+r})$$

pour une formule d'estimation à $r$-termes avec $m \leq p$ et $p = 2, \ldots, 4$.

Cela montre que les formules RK2(1)2G avec $m = 2$, RK3(2)3G1 avec $m = 2, 3$, et RK4(3)5FG avec $m = 3$ et RK4(3)5G avec $m = 2$ fournissent une estimation à un terme.

De plus, la formule RK4(3)5FG avec $m = 4$ continuera à fournir une estimation valide à deux termes, de même que la formule RK4(3)5G en fournira deux avec $m = 3, 4$.

Les cas des formules RK3(2)3G2 avec $m = 4$, RK4(3)5FG avec $m = 5$ et RK4(3)5G avec $m = 5$ doivent être considérés à part. Il suffit de remarquer que la borne (5.9) est la même pour $m = p + 1$ que pour $m = p$. Par conséquent, la formule RK3(2)3G2 continuera à fournir une estimation à un terme avec $m = 4$. Par contre, elle n'en fournira plus deux comme dans le cas d'une intégration à pas constant. Le même phénomène se produit pour les formules RK4(3)5FG et RK4(3)5G avec $m = 5$. Elles continuent à fournir une estimation à deux termes.

□

On résume sur la table Tab. 5.3 ces différents résultats. Il est intéressant de noter ce qui se passe quand on utilise des valeurs de $m$ qui excèdent celles données dans la figure Fig. 5.1. Par exemple, pour la méthode RK2(1)2G avec

| $p$ | $m$ | RK classique | RKDP | |
|---|---|---|---|---|
| 2 | 2 | 0 | 1 | |
| | 3 | 0 | 1 | |
| | 4 | -1 | 0 | |
| | | | RK3(2)3G1 | RK3(2)3G2 |
| 3 | 2 | 0 | 1 | 0 |
| | 3 | 0 | 1 | 0 |
| | 4 | 0 | 1 | 2 |
| | | | RK4(3)5FG | RK4(3)5G |
| 4 | 2 | 0 | 0 | 1 |
| | 3 | 0 | 1 | 2 |
| | 4 | 0 | 2 | 2 |
| | 5 | 0 | 2 | 2 |

TAB. 5.3 – *Ordre relatif de l'estimation pour des ERK classiques et des ERK de Dormand et Prince pour l'interpolation des $y_n$ à pas variable conservant un terme du développement de l'erreur globale*

$m = 4$, la borne (5.9) donne $d'''_H = \mathbf{O}(1/H)$. Par conséquent, le coefficient $\tau_1^{(4)}$ va passer du terme en $h^3$ au terme en $h^2$. Par suite, la formule RK2(1)2G ne fournira plus une estimation valide. La même remarque peut être faite pour les formules d'estimation à $r$-termes. On ne peut plus garantir qu'elles conserveront la propriété énoncée dans le Théorème 5.2.1 quand elles sont utilisées avec des polynômes d'un degré qui excède $p + 2$.

Une conséquence de l'analyse ci-dessus est que l'estimateur de Zadunaisky utilisé avec une méthode de Runge-Kutta quelconque ne fournira pas nécessairement une estimation valide si seul existe le premier terme du développement asymptotique de l'erreur globale par rapport à $H$. Par exemple, pour $p = 2$ et $m = 3$, la borne (5.9) établit que $d''_H = \mathbf{O}(1)$ alors qu'à pas constant, on avait $d''_h = \mathbf{O}(h)$. Par conséquent, le terme en $h^2$ de $\mathbf{p}_n$ ne sera plus supprimé.

Pour conclure cette remarque, nous noterons que même les formules d'estimation à $r$ termes perdront leur validité sur des grilles qui ne préservent même pas le premier terme du développement de l'erreur globale.

**Illustration Numérique**: Pour illustrer la Proposition 5.2.1, on considère l'équation $\dot{y} = y^2$, $y(0) = 1$, $t \in [0, 0.5]$ et la suite convergente de grilles définies par $h_{2p} = h_0$ et $h_{2p+1} = 2\,h_0$ et $h_0 = 0.5/(3P)$ si le nombre de pas $N = 2P$ , et $h_0 = 0.5/(3P + 1)$, si $N = 2P + 1$.

Le pas maximum est $2\,h_0$ et il tend vers zéro quand $N$ tend vers l'infini. Une telle suite de grille ne peut pas être décrite par une fonction de sélection et ne peut pas non plus être décrite par une relation de la forme $h_n = \theta(t_n)\,H + \mathbf{O}(H^2)$. Le pas n'admet pas non plus un développement asymptotique suivant

les hypothèses nécessaires à l'application du Théorème 4.2.1.

Cela peut se voir en remarquant que si une relation de la forme $h_n = \theta(t_n)\, H$ existait, alors $\theta$ devrait satisfaire $\theta(t_1) = 1$ si $t_1$ est atteint avec $t_1 = h_0$ et $\theta(t_1) = 1/2$ si $t_1$ est atteint avec $t_1 = h_0/3 + 2\,h_0/3$. Ce raisonnement peut s'étendre aux deux autres cas.

Néanmoins, cette suite de grilles ne détruit pas le premier terme du développement asymptotique de l'erreur globale, comme cela peut se voir sur la figure Fig. 5.1. Ce fait peut aussi se comprendre en remarquant que l'intégration sur cette grille est équivalente à une intégration à pas constant de longueur $3\,h_0$ [Hai97].

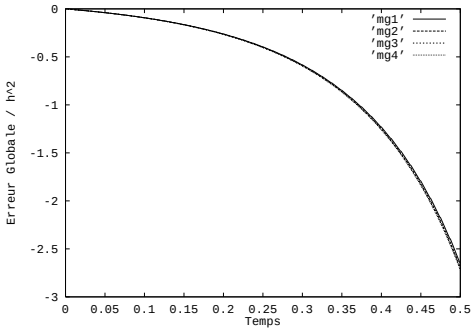



Fig. 5.1 – *Erreur magnifiée pour ERK2*

Sur les figures Fig. 5.2 et Fig. 5.3, on montre le comportement des valeurs

$$\max_n \frac{|2\,y_n - z_n - y(t_n)|}{(2\,h_0)^r}$$

où $z_n$ est la valeur obtenue par intégration du problème perturbé, pour un nombre croissant de pas $N$, pour respectivement la méthode classique de Runge-Kutta d'ordre 2 RK2 ($c_2 = 1/2$) et pour la formule RK2(1)2G avec $m = 3$. On a pris $r = 3$ pour RK2 et $r = 4$ pour RK2(1)2G.

Le fait que dans les deux cas on obtienne une belle ligne droite montre que pour la formule RK2, les valeurs $2\,y_n - \hat{y_n}$ ne sont que d'ordre deux alors que pour RK2(1)2G, elles sont d'ordre trois.

La même expérience à été faite avec $m = 4$ sur les figures Fig. 5.4 et Fig. 5.5 pour respectivement RK2 and RK2(1)2G. Cette fois, on a pris $r = 3$ pour les deux formules. Et, on voit que quelle que soit la formule, les valeurs obtenues sont d'ordre deux.

Et, le fait que la formule RK2(1)2G se comporte mieux que la classique RK2 ne peut être imputé à un meilleur défaut. Sur les figures Fig. 5.6 et Fig. 5.7, on a tracé pour les deux méthodes, respectivement les rapports

$$\max_{t\in[0,0.5]} \|d_h(t)\|/h^3$$

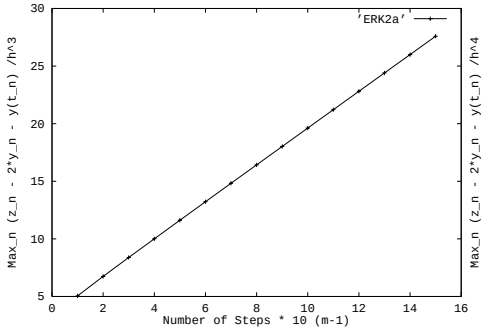



FIG. 5.2 – *RK2 & m=3*

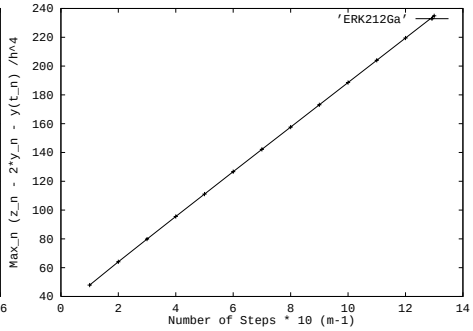



FIG. 5.3 – *RK212G & m=3*

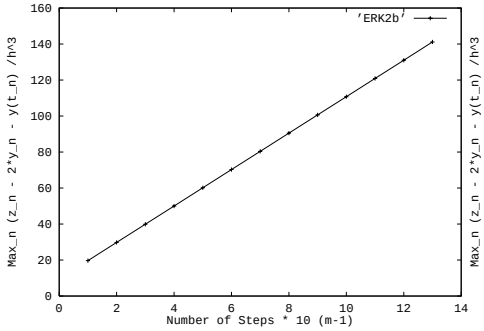



FIG. 5.4 – *RK2 & m=4*

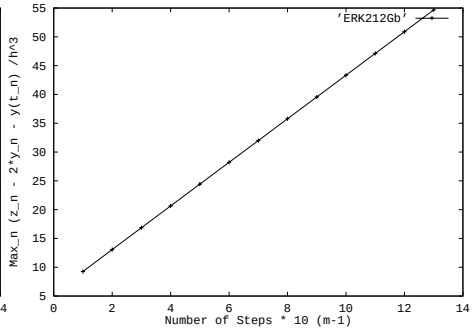



FIG. 5.5 – *RK212G & m=4*

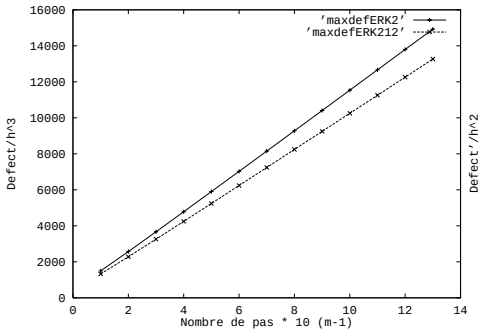



FIG. 5.6 – *Défaut - RK2 et RK212G*

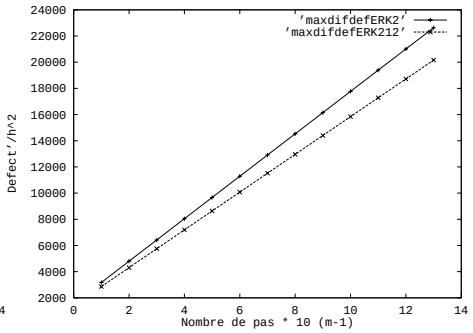



FIG. 5.7 – *Dérivé du défaut - RK2 et 212G*

et

$$\max_{t\in[0,0.5]} \|d'_h(t)\|/h^2$$

en fonction d'un nombre croissant de pas.

Il apparaît clairement d'une part, que dans les deux cas les défauts des deux méthodes sont du même ordre et leurs dérivées aussi, d'autre part, que leur dérivée n'est que d'ordre un, comme la borne (5.9) l'indique.

## 5.3 Discussion

Nous voulons ici juste faire une remarque et une conjecture.

La remarque concerne la raison pour laquelle sur une grille non-uniforme l'estimateur de Zadunaisky peut perdre sa validité. Si on reprend le contre-exemple qui entame ce chapitre, on remarque que la valeur obtenue par intégration du problème perturbé $z_1 = 1+h+h^2/(1+\alpha)$ dépend de $\alpha$ alors que l'erreur $y_1 - y(t_1)$ n'en dépend pas. Les méthodes développées dans [DP85] utilisent comme polynôme d'interpolation une extension continue de la méthode de Runge-Kutta. Par conséquent, elles n'utilisent que des informations locales et ne souffrent pas de ce problème.

Cette remarque étant faite, on peut conjecturer que l'estimation obtenue par l'équation modifiée reste convergente d'ordre relatif $p$ à pas variable.

Dans ce cas, on note $z_{n+1}$ la solution numérique en $t_{n+1}$ de :

$$\begin{array}{lcl} \dot{z} & = & f(Z_n) + \alpha\left(h_n^p f_p(Z_n) + \cdots + h_n^{2p-1} f_{2p-1}(Z_n)\right), \\ z(t_n) & = & z_n, \end{array} \tag{5.11}$$

obtenue avec la même méthode numérique et $t_{n+1} = t_n + h_n$, $Z_0(t_0) = z_0 = y_0$..

Et, on devrait avoir :

$$y_n + \alpha^{-1}(y_n - z_n) = y(t_n) + \mathbf{O}(H^{\min(2p,K)}), \tag{5.12}$$

avec $H = \max_n h_n$.

Le cas $\alpha = -1$ est équivalent à une méthode à un pas. Effectivement, dans ce cas, la valeur $y_n + \alpha^{-1}(y_n - z_n)$ vaut $z_n$ et est indépendante de $y_n$. Pour $\alpha \neq -1$, le résultat n'est pas direct.

Cependant, l'estimation ainsi obtenue ne l'est qu'à partir d'informations locales. Si on reprend la preuve de Dormand et Prince ci-dessus, on remarque qu'elle reste entièrement valable à ceci près que l'on s'épargne cette fois le terme $\mathbf{F}^{(i)}_{\mathcal{H}jt} - \mathbf{F}^{(i)}_{jt}$. Le système perturbé restant autonome, il n'y a plus de raison de distinguer les dérivées partielles par rapport à $t$. Le terme déterminant l'ordre de l'estimation demeure encore

$$\mathbf{p}_n = \sum_{i \geq p+1} h_n^{i-1} \sum_{j=1}^{r_i} \tau_j^{(i)} \left(\mathbf{F}^{(i)}_{h_n, j} - \mathbf{F}^{(i)}_j\right).$$

Mais, cette fois, les différentielles élémentaires du problème perturbé ne dépendent plus que du pas courant $h_n$.

Il n'est pas alors nécessaire de faire appel au développement asymptotique de l'erreur globale pour montrer que $\mathbf{p}_n$ est un $\mathbf{O}(h_n^{2p})$. Nous avons vu au paragraphe §3.2.2 qu'il suffisait que la perturbation du problème soit d'ordre $p$ pour que cela se répercute sur les différentielles élémentaires.

Il resterait un dernier point pour pouvoir conclure. Il reste à savoir si, sur une grille quelconque, l'équation modifié précédente passe aussi près que l'on veut de la solution numérique de (2.1), c'est-à-dire si l'on a encore

$$Z_n(t_n) = y_n + \mathbf{O}(H^{2p}).$$

# Deuxième partie

# Expérimentations numériques

# Chapitre 1

# Introduction

Le but des tests numériques présentés dans cette partie est simple. Il s'agit pour nous de savoir si l'on peut se fier à au moins un des estimateurs asymptotiques étudiés dans la partie précédente.

Les tests s'organisent de la façon suivante.

Tout d'abord, dans le Chapitre 2, nous reproduisons les tests publiés dans [AL97, Aïd96]. Ils constituent la base sur laquelle s'est fait le choix de l'implantation de l'estimateur de Richardson dans EUROSTAG. Ils utilisent différents systèmes différentiels de petites tailles dont la solution sous forme close est connue. Nous avons ajouté à ceux-là, des tests particuliers de l'estimation du Calcul d'une Correction Globale dans le cas où la seconde méthode d'intégration est d'ordre inférieur à celui de la première. Claude-Pierre Jeannerod et Josselin Visconti se sont chargés d'effectuer les mêmes tests sur des équations algébro-différentielles d'indice un et deux pour confirmer que les résultats observés n'étaient pas fondamentalement différents du cas ordinaire [Jea97, JV97]. Ces tests ne sont pas reproduits ici.

Bien que l'analyse de la Partie I ait été conduite pour des méthodes à un pas, nous avons voulu aussi tester l'estimateur de Richardson et l'intégration de l'équation variationnelle sur des codes employant des méthodes multipas. Nous nous sommes limités à ces deux estimateurs dans la mesure où l'analyse conduite à pas variable pour l'estimateur de Zadunaisky et pour le Calcul d'une Correction Globale ne laissait présager rien de bon dans ce cas. Effectivement, les codes que nous avons utilisés (DVODE [BBH89] et DSTEP [SG73]) implantent des méthodes à pas et ordre variables fondées sur respectivement la famille BDF et la famille Adams. Se posent alors deux problèmes. D'une part, à pas constant, des conditions supplémentaires sont nécessaires sur la stabilité de la méthode numérique et sur les valeurs l'initialisant pour garantir l'existence d'un développement de la forme (2.7). Dès que l'on sort de ce cas, il n'est plus possible de découpler l'indice du pas de sa valeur [Gra64, HNW87]. D'autre part, nous avons vu dans la Partie I que le passage d'une intégration à pas constant à une intégration à pas variable ne se faisait pas sans le risque d'une perte de convergence des estimateurs. Dans le cas des codes multipas où l'ordre varie, les

résultats sont encore plus délicats. Les principaux résultats sur cette question se trouvent dans [GT74, GW74, CL84, SZ90]. D'une manière générale, dans le cas de codes effectifs, les termes successifs sont perdus.

Puis, dans le Chapitre 3, nous nous intéressons au cas particulier de l'utilisation de l'Équation Modifiée comme alternative aux estimateurs de Zadunaisky employant des procédés d'interpolation numérique. Bien que l'on dispose dans [HS97] d'une relation générale donnant les termes successifs de la perturbation de l'équation modifiée, nous n'avons pas codé cet algorithme. Nous nous sommes contentés de calculer ces termes dans les cas qui nous intéressaient.

Enfin, dans le Chapitre 4, nous présentons les tests réalisés sur le simulateur EUROSTAG lui-même. Ces tests ont été effectués par Rodolphe Chopinet au cours de l'été 97 pour son stage industriel [Cho97].

# Chapitre 2

# Comparaison des estimateurs

Pour effectuer ces tests, nous avons choisi plusieurs codes du domaine public disponible ou bien sur `http://www.netlib.org/ode/index.html`, ou bien sur `http://www.unige.ch/folks/hairer/index.html` :

1. DOPRI5 (version d'avril 96) [DP80], utilise une paire ERK d'ordre 4 et 5 et un contrôle de l'erreur locale par extrapolation locale,

2. DVODE [BH75, BBH89], utilise la famille de méthodes multipas BDF. Démarre à l'ordre un. Ne fait pas d'extrapolation locale.

3. DSTEP [SG73], utilise la famille de méthodes multipas Adams. Démarre à l'ordre un. Effectue de l'extrapolation locale.

Dans DOPRI5, tous les estimateurs du Chapitre 3 ont été implantés. Par contre, dans DVODE et DSTEP, seuls celui de Richardson et l'intégration de l'équation variationnelle l'ont été.

Nous avons utilisé les six systèmes suivants :

**I.** Un système linéaire instable de dimension 2 [Hal69] :

$$y' = \begin{bmatrix} -1 + \frac{3}{2}\cos^2 t & 1 - \frac{3}{2}\sin t \cos t \\ -1 - \frac{3}{2}\sin t \cos t & -1 + \frac{3}{2}\sin^2 t \end{bmatrix} y$$

avec $y(0) = (1, 0)^t$, et $t \in [0, 10]$. Sa solution exacte est :

$$Y(t) = \begin{bmatrix} e^{t/2}\cos t & e^{-t}\sin t \\ -e^{t/2}\sin t & e^{-t}\cos t \end{bmatrix} y(0)$$

**II.** Une équation scalaire instable [SW76] :

$$\begin{aligned} y' &= 10\,(y - t^2) \\ y(0) &= 0.02 \end{aligned}$$

avec $t \in [0, 2]$. Sa solution exacte est :

$$y(t) = 0.02 + 0.2\,t + t^2.$$

**III.** Un système stable non-linéaire de dimension 4 :

$$\begin{aligned} y_1' &= -y_3 y_1 + y_2 \\ y_2' &= -y_1 - y_3 y_2 \\ y_3' &= y_4 \\ y_4' &= -y_3 \end{aligned}$$

avec $y(t_0) = (1, 1, 1, 1)^t$, $t \in [0, 7]$.

Sa solution exacte est :

$$\begin{aligned} y_1(t) &= (\cos t + \sin t)e^{-1+\cos t-\sin t} \\ y_2(t) &= (\cos t - \sin t)e^{-1+\cos t-\sin t} \\ y_3(t) &= \cos t + \sin t \\ y_4(t) &= \cos t - \sin t. \end{aligned}$$

**IV.** Un système linéaire raide de dimension 4 [Pro80] :

$$y' = \begin{bmatrix} -0.1 & -49.9 & 0 \\ 0 & -50 & 0 \\ 0 & 70 & -120 \end{bmatrix} y$$

avec $y(0) = (2, 1, 2)^t$, et $t \in [0, 1]$.

Sa solution exacte est :

$$\phi(t_0, y_0; t) = \begin{bmatrix} e^{-(t-t_0)/10} & e^{-50(t-t_0)} - e^{-(t-t_0)/10} & 0 \\ 0 & e^{-50(t-t_0)} & 0 \\ 0 & e^{-50(t-t_0)} - e^{-120(t-t_0)} & e^{-120(t-t_0)} \end{bmatrix} y_0$$

**V.** Le problème A3 du package DETEST [HEFS72], $\dot{y}(t) = \cos(t)\,y(t)$, $y(0) = 1$, $t \in [0, 20]$.

**VI.** Le problème A4 du package DETEST $\dot{y} = 0.25\,y\,(1 - 0.05\,y)$, $y(0) = 1$, $t \in [0, 20]$.

Pour ZD et SC, nous avons utilisé des polynômes d'interpolation de degré 10 et des différences divisées.

Pour EV, nous avons utilisé l'estimation de l'erreur de troncature donnée par le défaut du polynôme d'interpolation utilisé pour ZD (EVd). Pour des codes fondés sur des méthodes multipas, nous avons utilisé à la fois l'estimation qu'ils

utilisent pour la sélection du pas (EV) et l'erreur locale exacte (EVE), pour avoir une idée de ce que peut donner cet estimateur au mieux.

Pour pouvoir effectuer des comparaisons de manière systématique, nous avons utilisé la méthode suivante. Nous avons attribué à chaque estimateur, à chaque instant $t_n$, 0, s'il ne fournissait pas même l'ordre de grandeur de l'erreur et 1 sinon, plus le nombre de chiffres significatifs de l'erreur dans le premier cas. Puis, nous avons fait la moyenne de ces valeurs sur tout l'intervalle d'intégration. La valeur obtenue est une mesure de *l'efficacité* de l'estimateur. Une valeur proche de un signifie que l'estimateur donne correctement l'ordre de grandeur de l'erreur globale et un chiffre significatif. Toutefois, une valeur voisine de 0.5 suffit à garantir qu'il donne cet ordre de grandeur. Si cette méthode ne donne pas une vision dynamique de l'estimation, c'est-à-dire de son comportement au cours du temps, elle fournit une idée correcte de sa précision.

De plus, pour rester au plus proche du cadre théorique développé dans le Chapitre 4, les tests ont été tous effectués en fixant la tolérance relative des codes à zéro. Cette manière de procéder est plus contraignante pour le code, mais elle correspond mieux à ce qui a été supposé au chapitre 4.

Pour tester SC dans le cas d'un ré-intégration avec une méthode d'ordre inférieur à 5, et pour limiter le nombre de cas possibles, nous n'avons pas repris l'ensemble des problèmes. Nous nous sommes limités aux problèmes III, V et VI. Nous avons utilisé des méthodes ERK d'ordre $q = 2, ..., 5$ pour intégrer l'équation (3.26) et des polynômes d'interpolation prenant les valeurs $m = 6, 8, 10$. La méthode d'ordre 2 est obtenue avec $c_2 = 1/2$. La méthode d'ordre 3 est donnée par le tableau :

| | | | |
|---|---|---|---|
| 1/3 | 1/3 | | |
| 2/3 | 0 | 2/3 | |
| | 1/4 | 0 | 3/4 |

La méthode d'ordre 4 est la méthode classique de Runge-Kutta. La méthode d'ordre 5 utilisée ici est différente de celle de DOPRI5. Il s'agit de la ERKF5 [Feh66]. Nous avons utilisé la même mesure de l'efficacité de l'estimation que dans les autres cas.

Tous les calculs ont été conduits en arithmétique flottante double précision sur un SUN4.

## 2.1 DOPRI5

Avant de commenter l'ensemble des résultats, nous commençons par illustrer le problème que pose l'utilisation d'une méthode d'interpolation polynomiale élémentaire pour ZD et SC. Sur les figures FIG. 2.1 à FIG. 2.4 sont représentées à la fois l'erreur globale commise dans DOPRI5 sur le système V, ainsi que

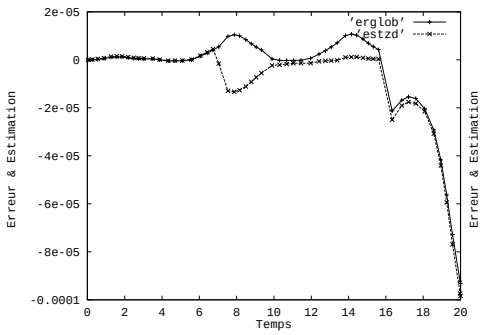



FIG. 2.1 – *ZD - système V - atol*=$10^{-5}$

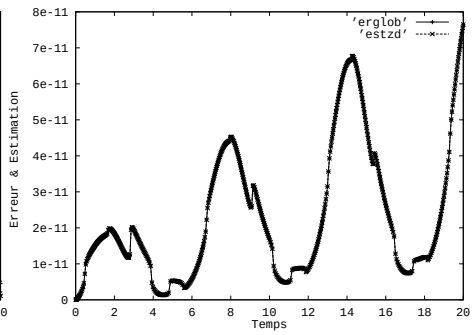



FIG. 2.2 – *ZD - système V - atol*=$10^{-10}$

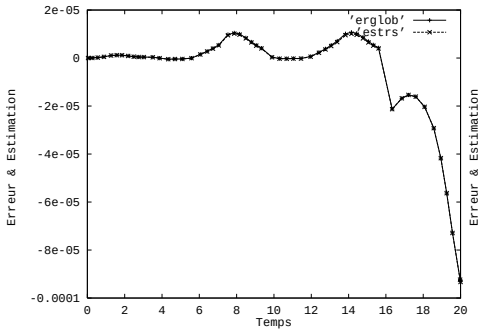



FIG. 2.3 – *RS - système V - atol*=$10^{-5}$

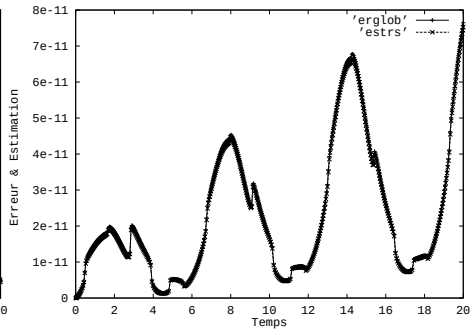



FIG. 2.4 – *RS - système V - atol*=$10^{-10}$

les estimations obtenues par ZD et RS. Ces figures permettent de comparer le comportement de ZD et RS pour des tolérances grandes ($10^{-5}$, FIG. 2.1 et FIG. 2.3) et petites ($10^{-10}$, FIG. 2.2 et FIG. 2.4).

Si dans les deux situations, l'estimation fournie par RS se comporte de manière satisfaisante, ce n'est pas le cas de celle obtenue par ZD. Elle montre clairement un comportement irrégulier sur la figure FIG. 2.1. Une fois que la tolérance a été suffisamment diminuée pour que les pas d'intégration soient petits, ce phénomène disparaît et l'estimation est alors bien meilleure que celle de RS.

Cela illustre un des avantages de RS sur ZD. Alors que RS n'a besoin que de l'existence du premier terme du développement asymptotique, ZD et SC nécessitent plus de régularité des termes $e_k$ de ce développement pour permettre une convergence rapide des termes perturbés $e_{h,k}$. On voit que sur le problème V, au moins la première de ces fonctions n'est que dérivable par morceaux à cause des variations du pas d'intégration (voir [CHMR96] pour plus de détails).

Les tables TAB. 2.1 et TAB. 2.2 donnent l'efficacité des estimateurs telle que nous l'avons définie plus haut. D'une manière générale, ZD et SC fournissent plus de chiffres significatifs que RS ou EVd. L'estimation RS donne de manière

| -log atol | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 3.2 | 1.8 | 1.6 | 2.0 | 2.0 | 2.1 | 2.2 | 2.0 | 2.3 | 1.5 |
| I | 4.3 | 5.5 | 6.8 | 6.6 | 6.4 | 6.0 | 4.7 | 3.9 | 3.0 | 1.1 |
| | 4.3 | 5.5 | 6.8 | 6.5 | 6.4 | 6.5 | 6.0 | 4.9 | 4.1 | 3.1 |
| | 1.1 | 0.4 | 0.6 | 1.2 | 1.5 | 1.7 | 1.8 | 2.2 | 1.6 | 0.7 |
| | 1.0 | 1.8 | 1.8 | 2.1 | 2.1 | 2.0 | 2.0 | 2.1 | 2.0 | 2.1 |
| II | 0.1 | 3.5 | 4.7 | 5.4 | 6.0 | 6.5 | 6.9 | 6.2 | 5.6 | 4.5 |
| | 0.1 | 3.4 | 4.7 | 5.4 | 6.0 | 6.5 | 6.8 | 6.9 | 6.7 | 5.6 |
| | 0.1 | 0.0 | 0.0 | 0.2 | 0.5 | 0.7 | 0.9 | 1.1 | 1.3 | 1.5 |
| | 2.3 | 2.3 | 2.4 | 2.3 | 2.3 | 2.4 | 2.3 | 2.2 | 2.2 | 2.1 |
| III | 2.4 | 1.3 | 2.5 | 3.3 | 4.2 | 5.2 | 6.0 | 4.1 | 3.0 | 2.3 |
| | 2.3 | 1.1 | 2.5 | 3.3 | 4.1 | 5.0 | 6.0 | 5.5 | 4.4 | 3.4 |
| | 0.2 | 0.1 | 0.4 | 0.3 | 0.7 | 1.5 | 2.0 | 2.1 | 1.8 | 0.8 |

TAB. 2.1 – *RS, ZD, SC, EVd - DOPRI5 - I/III*

constante en fonction de la tolérance l'ordre de grandeur de l'erreur globale et un chiffre significatif. Sur la plupart des exemples, il faut attendre d'avoir atteint la zone des tolérances petites pour que ZD et SC deviennent bien meilleurs que RS. Ils peuvent alors être utilisés pour améliorer de manière drastique la solution numérique. En règle générale, SC est plus précis que ZD. Si le besoin de l'utilisateur se limite à l'ordre de grandeur de l'erreur globale, EVd peut faire l'affaire.

On remarque aussi une dégradation de l'estimation pour des tolérances inférieures ou égales à $10^{-11}$ qui peut être mise sur le compte d'un effet de l'erreur arithmétique.

En ce qui concerne SC utilisé à un ordre inférieur à son maximum, une première lecture des tables TAB. 2.3 et TAB. 2.4 montre que, dans l'ensemble, on n'obtient pas une estimation satisfaisante pour des tolérances grandes. C'est nettement visible sur la table TAB. 2.3.

Il n'est pas possible d'utiliser conjointement une interpolation de faible degré et une méthode d'ordre un pour obtenir une estimation fiable. En revanche, sur les problèmes A3 et III, l'estimation obtenue avec des polynômes de degré 8 devient suffisante pour des tolérances petites.

| -log atol | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2.2 | 3.9 | 3.6 | 2.2 | 2.2 | 2.3 | 2.3 | 2.7 | 2.3 | 1.2 |
| IV | 1.7 | 2.0 | 3.2 | 3.7 | 4.5 | 6.3 | 7.0 | 4.7 | 1.2 | 3.7 |
| | 1.7 | 2.0 | 3.0 | 3.8 | 4.5 | 5.1 | 5.0 | 4.3 | 3.4 | 2.2 |
| | 0.3 | 0.4 | 0.7 | 0.8 | 1.2 | 1.7 | 2.0 | 2.1 | 0.8 | 0.4 |
| | 2.2 | 2.3 | 2.2 | 2.3 | 2.5 | 2.1 | 2.7 | 2.2 | 2.1 | 1.5 |
| V | 0.9 | 0.1 | 0.8 | 1.7 | 2.2 | 2.4 | 3.5 | 4.9 | 3.7 | 2.5 |
| | 0.9 | 0.1 | 0.8 | 1.7 | 2.2 | 2.3 | 3.2 | 4.7 | 3.5 | 3.8 |
| | 0.0 | 0.0 | 0.0 | 0.1 | 0.2 | 0.2 | 0.6 | 1.0 | 1.6 | 0.4 |
| | 2.4 | 2.4 | 2.3 | 2.9 | 2.4 | 2.2 | 2.3 | 2.4 | 2.4 | 2.3 |
| VI | 0.0 | 2.0 | 3.2 | 4.0 | 4.2 | 4.9 | 6.4 | 5.8 | 5.1 | 1.9 |
| | 0.0 | 2.2 | 3.4 | 3.3 | 3.7 | 4.5 | 5.0 | 5.0 | 4.0 | 3.1 |
| | 0.0 | 0.4 | 0.6 | 0.3 | 0.9 | 1.1 | 1.2 | 1.9 | 2.2 | 1.2 |

TAB. 2.2 – *RS, ZD, SC, EVd - DOPRI5 - III/VI*

| -log Atol | | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| m | q | | | | | | | | | | |
| 6 | 1 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | 5 | 0.0 | 0.0 | 0.2 | 0.5 | 0.3 | 0.8 | 0.7 | 0.7 | 1.2 | 1.5 |
| 8 | 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.2 | 0.5 | 0.5 | 0.3 |
| | 2 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.2 | 0.5 | 0.5 | 1.1 | 0.9 |
| | 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.2 | 0.6 | 0.7 | 1.6 | 1.2 |
| | 4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.4 | 0.9 | 0.8 | 0.7 |
| 10 | 1 | 0.0 | 0.0 | 0.0 | 0.1 | 0.2 | 0.2 | 0.6 | 0.9 | 1.6 | 0.4 |
| | 2 | 0.0 | 0.0 | 0.0 | 0.2 | 0.1 | 0.7 | 0.8 | 1.3 | 1.1 | 1.0 |
| | 3 | 0.0 | 0.0 | 0.0 | 0.1 | 0.3 | 0.7 | 1.0 | 1.6 | 1.7 | 1.2 |
| | 4 | 0.0 | 0.0 | 0.0 | 0.1 | 0.3 | 0.5 | 0.5 | 1.2 | 1.2 | 0.7 |
| | 5 | 0.0 | 0.0 | 0.2 | 0.9 | 1.1 | 1.6 | 2.5 | 3.3 | 3.3 | 2.9 |

TAB. 2.3 – *SC problème A3*

| -log Atol | | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| m | q | | | | | | | | | | |
| 6 | 1 | 0.1 | 0.4 | 0.0 | 0.0 | 0.3 | 0.1 | 0.1 | 0.1 | 0.2 | 0.3 |
| | 5 | 1.1 | 1.8 | 1.5 | 1.4 | 1.6 | 2.2 | 2.1 | 2.5 | 2.7 | 2.1 |
| 8 | 1 | 0.1 | 0.2 | 0.4 | 0.2 | 0.3 | 0.4 | 1.0 | 1.4 | 1.4 | 1.5 |
| | 2 | 0.1 | 0.6 | 0.4 | 0.8 | 0.7 | 1.1 | 1.3 | 1.7 | 2.0 | 2.2 |
| | 3 | 0.5 | 1.1 | 0.7 | 0.8 | 1.1 | 1.5 | 2.0 | 2.6 | 2.7 | 2.7 |
| | 4 | 0.5 | 0.3 | 0.3 | 0.3 | 0.6 | 0.7 | 0.7 | 0.9 | 0.9 | 1.3 |
| 10 | 1 | 0.0 | 0.4 | 0.6 | 0.3 | 0.9 | 1.1 | 1.2 | 1.9 | 2.2 | 1.2 |
| | 2 | 0.0 | 0.2 | 1.2 | 1.2 | 1.6 | 1.9 | 2.0 | 2.6 | 2.5 | 1.7 |
| | 3 | 0.0 | 0.9 | 1.5 | 1.2 | 1.7 | 2.1 | 2.3 | 2.8 | 2.5 | 2.4 |
| | 4 | 0.0 | 0.6 | 0.6 | 0.6 | 0.7 | 0.8 | 1.0 | 1.1 | 1.0 | 1.1 |
| | 5 | 0.0 | 1.4 | 2.2 | 2.6 | 3.0 | 3.3 | 3.9 | 4.5 | 4.0 | 3.1 |

TAB. 2.4 – *SC problème A4*

| -log Atol | | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| m | q | | | | | | | | | | |
| 6 | 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | 5 | 0.9 | 0.8 | 1.1 | 1.1 | 1.2 | 1.6 | 1.9 | 2.1 | 2.3 | 2.7 |
| 8 | 1 | 0.1 | 0.0 | 0.0 | 0.1 | 0.3 | 0.7 | 0.9 | 1.3 | 2.0 | 1.2 |
| | 2 | 0.1 | 0.2 | 0.1 | 0.3 | 0.6 | 0.8 | 1.3 | 1.6 | 2.2 | 2.2 |
| | 3 | 0.2 | 0.2 | 0.2 | 0.4 | 0.8 | 1.4 | 2.0 | 2.7 | 3.1 | 2.1 |
| | 4 | 0.1 | 0.2 | 0.1 | 0.0 | 0.7 | 0.7 | 0.9 | 1.0 | 1.1 | 1.1 |
| 10 | 1 | 0.2 | 0.1 | 0.4 | 0.3 | 0.7 | 1.5 | 2.0 | 2.1 | 1.8 | 0.8 |
| | 2 | 0.3 | 0.3 | 0.6 | 0.6 | 1.3 | 2.1 | 3.1 | 3.4 | 3.0 | 2.0 |
| | 3 | 0.6 | 0.3 | 0.7 | 1.1 | 1.6 | 2.6 | 3.5 | 3.7 | 2.7 | 1.6 |
| | 4 | 0.2 | 0.2 | 0.2 | 0.7 | 0.7 | 0.8 | 0.9 | 1.1 | 1.1 | 1.2 |
| | 5 | 1.5 | 0.7 | 1.4 | 2.3 | 3.0 | 4.0 | 5.2 | 5.2 | 4.2 | 3.1 |

TAB. 2.5 – *SC problème III*

Bien que DOPRI5 ne soit pas prévu pour être utilisé à des tolérances plus petites que $10^{-12}$, on peut voir sur les figures FIG. 2.5 et FIG. 2.6 ce qui se produit pour ces estimateurs pour des tolérances très petites sur respectivement les problèmes II et VI. Sur ces figures, est représenté le maximum de l'erreur et des estimateurs en échelle logarithmique. On constate que pour des tolérances situées entre $10^{-3}$ et $10^{-12}$, l'erreur et les estimateurs décroissent linéairement et que leurs courbes sont confondues. En revanche, pour des tolérances plus faibles, l'erreur arithmétique conduit à une stagnation de l'erreur locale. Les estimateurs se distinguent alors. L'estimateur de Richardson - à gauche de chaque figure - reste voisin de l'erreur globale alors que les estimateurs SC et ZD - à droite de

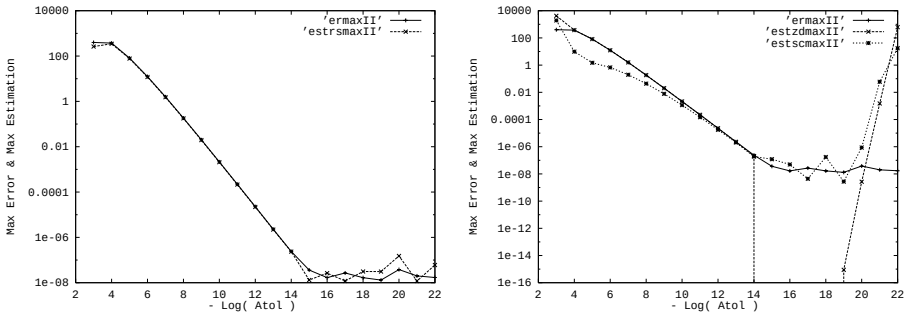



FIG. 2.5 – *Comportement du maximum des estimations et du maximum de l'erreur globale sur le système II*

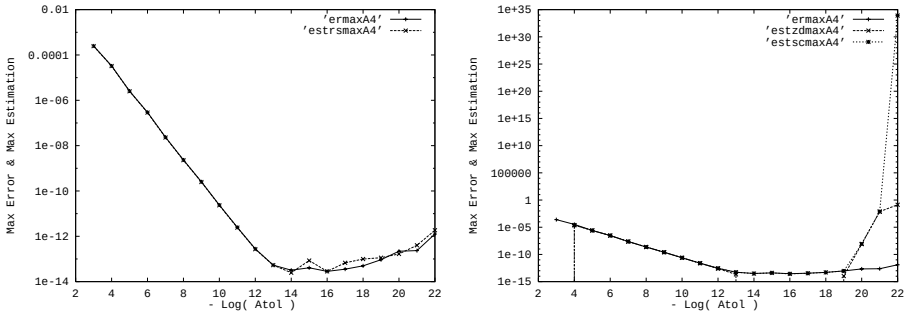



FIG. 2.6 – *Comportement du maximum des estimations et du maximum de l'erreur globale sur le système V*

chaque figure -, beaucoup plus sensibles aux erreurs d'arrondis, divergent.

## 2.2 DSTEP et DVODE

En arguant du fait que l'estimateur de Richardson n'a besoin que du premier terme du développement asymptotique pour donner une estimation valide, il était susceptible de continuer à donner une estimation correcte de l'erreur globale dans des codes aussi sophistiqués que DVODE et DSTEP quant à la sélection de l'ordre et du pas d'intégration. Un détail restait toutefois à fixer. Il fallait choisir l'ordre $p$ à utiliser dans la relation donnant l'estimation de Richardson. Nous avons choisi de le prendre égal à 2. La raison qui justifie ce choix est que le premier pas étant effectué avec une méthode d'ordre 1, nous avons considéré qu'au pire l'erreur globale était d'ordre deux. Autre point intéressant à noter, l'ordre n'intervient que comme facteur multiplicatif $(1-2^{-p})^{-1}$. Au pire, on se trompera d'un rapport constant par rapport à $H$ si le code maintient le premier terme du développement de l'erreur.

Comme on peut le voir sur les figures Fig. 2.7 et Fig. 2.8, l'estimation RS obtenue sur les systèmes III et IV intégrés avec DSTEP suit bien le comportement de l'erreur globale. C'est aussi le cas pour le système raide (Fig. 2.8). Si l'on compare ce résultat à ce que l'on obtient avec DVODE pour les mêmes systèmes et les mêmes valeurs de la tolérance (figures Fig. 2.9 et Fig. 2.10), on remarque que ce phénomène est encore plus flagrant sur DVODE. L'estimation RS suit le comportement erratique de l'erreur globale.

Dans les tables Tab. 2.6 et Tab. 2.7 on donne sur la première ligne, la mesure de l'efficacité de l'estimateur RS. Si on la compare avec ce que l'on obtenait avec DOPRI5, il est clair qu'elle est moins bonne. On peut imputer cette dégradation de l'estimation aux variations d'ordre. Néanmoins, il fournit encore une estimation satisfaisante dans les deux codes, puisqu'il donne au moins l'ordre de grandeur de l'erreur globale et un premier chiffre.

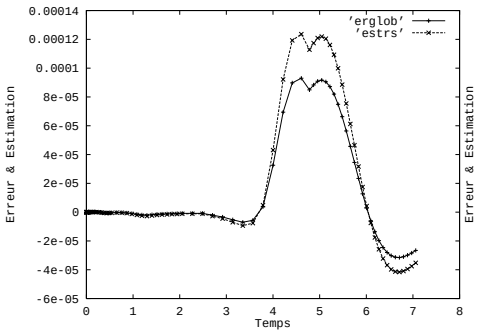



Fig. 2.7 – *RS - système III - DSTEP,* $atol=10^{-5}$

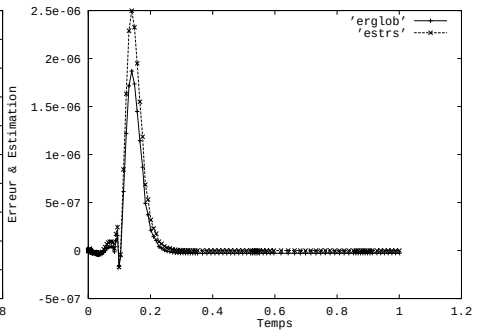



Fig. 2.8 – *RS - système IV - DSTEP,* $atol=10^{-5}$

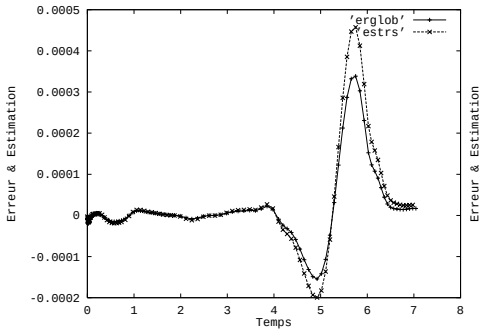



FIG. 2.9 – *RS, système III, DVODE,* $atol=10^{-5}$

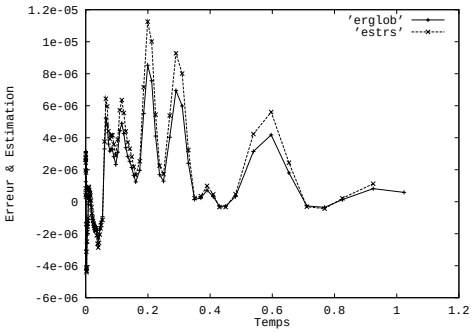



FIG. 2.10 – *RS - système IV - DVODE* - $atol=10^{-5}$

Pour ce type de codes d'intégration, EV donne de meilleurs résultats que dans le cas précédent. Les tables TAB. 2.6 et TAB. 2.7 montrent pour chaque code et chaque système, sur la seconde ligne, la qualité de l'estimation obtenue en utilisant l'estimation de l'erreur locale calculée par le code, et sur la troisième ligne, celle que l'on peut obtenir au mieux en prenant l'erreur locale exacte (EEV). Il apparaît alors clairement que cet estimateur peut fournir une très bonne estimation de l'erreur locale si l'on améliore l'estimation de l'erreur locale. C'est ce que l'on peut trouver dans [Ste79].

| -log atol | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1.2 | 1.5 | 1.1 | 1.1 | 1.1 | 1.2 | 1.1 | 0.9 | 1.2 | 1.2 |
| I | 0.8 | 1.2 | 0.7 | 0.6 | 0.7 | 0.9 | 0.8 | 0.8 | 0.6 | 0.5 |
| | 1.4 | 1.7 | 1.4 | 1.7 | 1.9 | 2.2 | 2.2 | 2.5 | 2.6 | 2.7 |
| | 2.1 | 1.8 | 2.0 | 2.4 | 2.5 | 2.7 | 2.0 | 2.7 | 1.1 | 2.0 |
| II | 0.3 | 0.3 | 0.4 | 0.4 | 0.5 | 0.5 | 0.5 | 0.7 | 0.3 | 0.4 |
| | 0.4 | 0.5 | 0.6 | 0.6 | 0.7 | 0.8 | 0.9 | 0.9 | 0.8 | 1.2 |
| | 1.2 | 1.1 | 1.3 | 1.3 | 1.2 | 1.1 | 1.1 | 1.1 | 1.2 | 1.2 |
| III | 0.7 | 0.8 | 0.7 | 0.8 | 0.9 | 0.9 | 0.8 | 0.7 | 0.5 | 0.7 |
| | 1.6 | 1.9 | 2.2 | 2.3 | 2.3 | 2.4 | 2.5 | 2.7 | 3.0 | 3.0 |
| | 1.4 | 1.3 | 1.3 | 1.2 | 1.2 | 1.2 | 1.4 | 1.3 | 1.0 | 1.1 |
| IV | 1.0 | 0.8 | 0.7 | 0.7 | 1.0 | 0.6 | 0.4 | 1.0 | 0.8 | 0.8 |
| | 1.7 | 2.2 | 2.2 | 2.3 | 2.4 | 2.7 | 2.9 | 2.7 | 3.0 | 2.8 |
| | 1.5 | 1.0 | 1.9 | 1.4 | 1.3 | 1.3 | 1.3 | 1.2 | 1.1 | 1.1 |
| V | 0.6 | 0.6 | 0.8 | 0.8 | 0.9 | 0.6 | 0.7 | 0.7 | 0.7 | 0.7 |
| | 1.5 | 1.5 | 2.1 | 1.8 | 2.1 | 2.2 | 2.3 | 2.6 | 2.5 | 2.7 |
| | 1.3 | 1.4 | 1.9 | 1.5 | 1.0 | 1.1 | 1.3 | 0.8 | 1.1 | 1.2 |
| VI | 1.3 | 1.3 | 1.3 | 0.6 | 1.5 | 1.2 | 0.7 | 0.9 | 0.8 | 0.3 |
| | 1.9 | 2.3 | 2.6 | 2.6 | 3.2 | 3.1 | 3.4 | 3.3 | 3.6 | 3.9 |

TAB. 2.6 – *RS, EV, EEV - DVODE*

| -log atol | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1.5 | 1.5 | 1.2 | 1.1 | 1.1 | 0.7 | 0.6 | 0.5 | 0.7 | 0.8 |
| I | 0.7 | 0.3 | 0.3 | 0.3 | 0.2 | 0.3 | 0.4 | 0.2 | 0.0 | 0.0 |
| | 1.8 | 1.7 | 1.8 | 2.3 | 2.5 | 2.5 | 2.5 | 2.9 | 2.8 | 3.0 |
| | 1.1 | 1.3 | 1.5 | 0.7 | 1.0 | 1.2 | 2.6 | 1.5 | 1.2 | 1.2 |
| II | 0.6 | 0.7 | 0.4 | 0.0 | 0.6 | 0.6 | 0.6 | 0.6 | 0.3 | 0.3 |
| | 0.5 | 0.3 | 0.8 | 0.7 | 1.1 | 1.3 | 0.8 | 1.5 | 0.8 | 1.4 |
| | 1.3 | 1.5 | 1.1 | 1.1 | 1.2 | 0.9 | 0.7 | 0.6 | 0.7 | 0.8 |
| III | 0.5 | 0.7 | 0.4 | 0.4 | 0.5 | 0.5 | 0.7 | 0.1 | 0.0 | 0.0 |
| | 1.3 | 1.8 | 1.8 | 2.1 | 2.3 | 2.5 | 2.4 | 2.6 | 2.9 | 2.7 |
| | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.2 | 1.1 |
| IV | 0.7 | 0.3 | 0.3 | 0.3 | 0.4 | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 |
| | 2.7 | 2.4 | 3.1 | 3.4 | 3.4 | 3.6 | 4.0 | 4.3 | 4.7 | 4.9 |
| | 1.4 | 1.2 | 1.1 | 1.2 | 1.2 | 1.0 | 0.7 | 0.5 | 0.5 | 0.6 |
| V | 0.3 | 0.2 | 0.6 | 0.6 | 0.0 | 0.7 | 0.5 | 0.2 | 0.0 | 0.0 |
| | 1.4 | 1.7 | 1.8 | 1.7 | 1.8 | 1.9 | 2.2 | 2.6 | 2.4 | 2.8 |
| | 1.4 | 1.5 | 1.3 | 1.4 | 1.3 | 1.0 | 1.0 | 0.9 | 0.9 | 0.9 |
| VI | 0.0 | 0.3 | 0.7 | 0.5 | 0.3 | 0.2 | 0.7 | 0.2 | 0.0 | 0.0 |
| | 1.7 | 2.6 | 2.7 | 3.2 | 3.2 | 3.6 | 3.6 | 3.7 | 4.0 | 3.9 |

TAB. 2.7 – *RS, EV, EEV - DSTEP*

# Chapitre 3

# Estimation par Équation Modifiée

Le but de ces tests n'étant que de comparer l'estimation obtenue en utilisant l'équation modifiée (EM) avec celle obtenue par la technique de même ordre de Zadunaisky (ZD), nous nous sommes limités à une méthode de Runge-Kutta d'ordre 2 avec $c_2 = 1/2$.

Dans [HW96], sont donnés les termes $f_p$ et $f_{p+1}$ de l'équation modifiée :

$$\begin{aligned} f_p(y) &= d_{p+1}(y), \\ f_{p+1}(y) &= d_{p+2}(y) - \frac{1}{2}(f_p'(y)f(y) + f'(y)f_p(y)). \end{aligned}$$

Pour cette méthode Runge-Kutta, on a :

$$\begin{aligned} d_{p+1}(y) &= -\frac{1}{24}f''(y)f^2(y) - \frac{1}{6}f'^2(y)f(y), \\ d_{p+2}(y) &= -\frac{1}{48}f'''(y)f^3(y) - \frac{1}{6}f''(y)f'(y)f^2(y) - \frac{1}{24}f'^3(y)f(y), \end{aligned}$$

où nous avons utilisé une simplification pour le cas scalaire.

Nous avons considéré les problèmes A3 et A4 du package DETEST vu dans le Chapitre 2. À ceux-ci, nous avons ajouté les problèmes simples suivants : $\dot{y} = y^2$, $y(0) = 1$, $t \in [0, 0.5]$ (problème VII) et $\dot{y} = y^3$, $y(0) = 1/2$, $t \in [0, 1]$ (problème VIII).

Nous avons comparé tout d'abord la qualité des estimations EM et ZD sur le problème A3. Celui-ci est le plus difficile des quatre problèmes à intégrer. Nous n'avons utilisé pour ce test que le premier terme de l'équation modifiée et pour l'estimateur de Zadunaisky, des polynômes de degré 3. L'ordre de convergence des deux estimateurs est alors le même et, ce que l'on mesure dans cette situation, est l'influence de la constante dans le grand **O**.

L'équation modifiée pour A3 est alors :

$$\begin{aligned} \dot{z} &= \cos(t)\, z - \frac{1}{24}\, h^2\, (-\cos(t)\, z - 2\, \sin(t)\, \cos(t)\, z) \\ &- \frac{1}{6}\, (-\cos(t) \sin(t)\, z + \cos^3(t)\, z) \end{aligned}$$

et pour A4 :

$$\begin{aligned} \dot{z} &= \frac{1}{4}\, z - \frac{1}{80}\, z^2 \\ &+ h^2\, (-\frac{1}{384} z + \frac{11}{15360} z^2 - \frac{3}{51200} z^3 - \frac{1}{384} z^4) \\ &+ h^3\, (\frac{1}{2048} z - \frac{1}{5120} z^2 + \frac{11}{409600} z^3 - \frac{1}{655360} z^4 + \frac{1}{32768000} z^5). \end{aligned}$$

Ensuite, nous avons comparé les ordres maximum atteignables sur la grille $h$, $2h$, ..., vue au §5.3. Nous avons déjà vu au §5.3, que l'estimateur de Zadunaisky perd sa validité asymptotique dans ce cas. Pour ces tests, nous avons utilisé les problèmes autonomes parce qu'il est alors beaucoup plus simple de calculer $f_{p+1}$.

Dans le cas où l'intégration est conduite à pas variable, l'estimation EM est obtenue par la suite des problèmes :

$$\begin{aligned} \dot{Z}_n &= f(Z_n) + h_n^p f_p(Z_n) + \cdots, \\ Z_n(t_n) &= z_n, \end{aligned}$$

avec $t_{n+1} = t_n + h_n$, $Z_0(t_0) = z_0 = y_0$ et $z_n$ la valeur numérique obtenue en $t_n$ par la méthode de Runge-Kutta.

Les figures FIG. 3.1, FIG. 3.2 et FIG. 3.3 correspondent à l'estimation de l'erreur globale commise sur le problème A3 pour une intégration à pas constant de 80 pas. On peut relever deux traits caractéristiques qui se reproduisent à chacune des expériences que nous avons réalisées. D'une part, la figure FIG. 3.1 montre que l'estimation obtenue par l'équation modifiée est moins précise que celle obtenue par l'estimateur de Zadunaisky. On peut voir sur les figures FIG. 3.2 et FIG. 3.3 la différence entre l'erreur globale et respectivement, l'estimation obtenue par ZD et celle obtenue par l'équation modifiée. On constate que l'estimateur ZD fournit un chiffre significatif de plus que EM. D'autre part, la perte de ce chiffre significatif est compensée par une plus grande régularité de l'estimateur EM en fonction du temps.

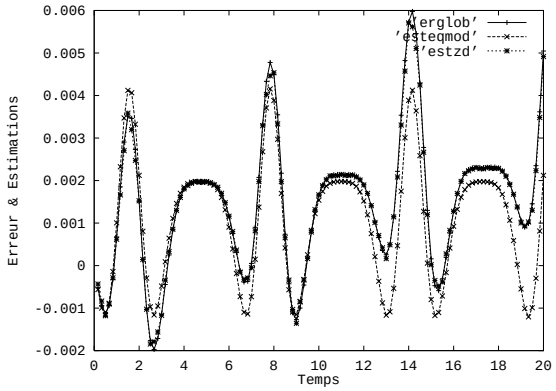



FIG. 3.1 – *Erreur et estimation sur A3 - 80 pas*

En ce qui concerne le caractère asymptotique de ces deux estimateurs, on observe directement ce qui se produit sur la grille $h$, $2h$, quand on augmente le nombre de pas. Les figures FIG. 3.4 et FIG. 3.5 montrent le comportement du rapport

$$\frac{\max_n |z_n - 2\,y_n + y(t_n)|}{H^r}$$

avec $H = 2h$ et respectivement $r = 3$ et $r = 5$ sur le problème $\dot{y} = y^2$. Pour ZD, on a pris $m = 4$ et pour EM, on a pris deux termes, ce qui fait que les deux estimateurs devraient exhiber le même ordre relatif de convergence. L'obtention de droites montre clairement que les valeurs obtenues par ZD ne sont que d'ordre 2 alors que celles obtenues par EM sont d'ordre 4. Même sur cette grille qui détruit le deuxième terme du développement asymptotique de l'erreur globale, EM conserve le même ordre qu'à pas constant.

Cette expérience est reproduite sur le problème $\dot{y} = y^3$ sur les figures FIG. 3.6 et FIG. 3.7. Elle conduit à la même remarque.

Toutefois, le fait que EM présente un meilleur caractère asymptotique que ZD n'est pas incompatible avec le fait qu'il puisse fournir une bonne estimation de l'erreur globale même sur cette grille. On peut observer qu'en fait, ZD reste plus précis que EM même sur cette grille. Les figures FIG. 3.8, FIG. 3.9 et FIG. 3.10 reprennent le test qui a été fait sur A3. On constate que ZD conserve toujours un chiffre significatif de plus que EM, même s'il est moins régulier.

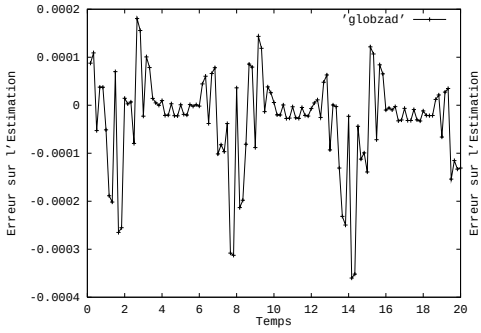



FIG. 3.2 – *Erreur ZD sur A3 $m = 3$*

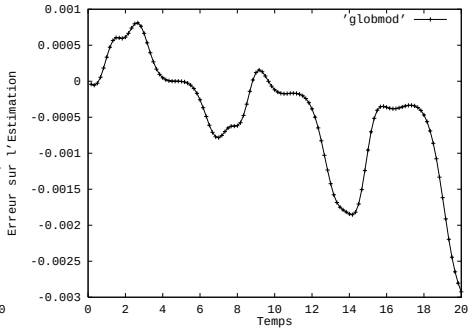



FIG. 3.3 – *Erreur EM sur A3 avec un terme*

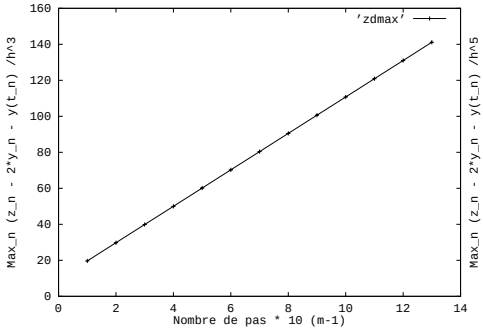



FIG. 3.4 – *Ordre de l'estimation ZD - VII*

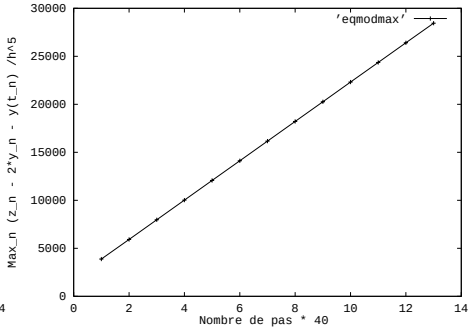



FIG. 3.5 – *Ordre de l'estimation EM - VII*

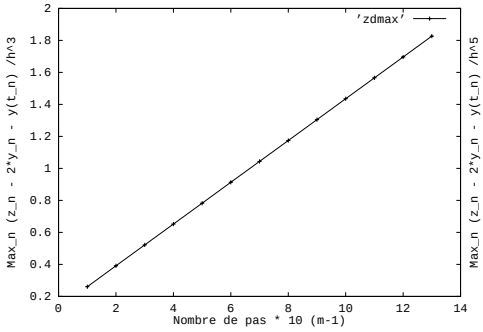



FIG. 3.6 – *Ordre de l'estimation ZD - VIII*

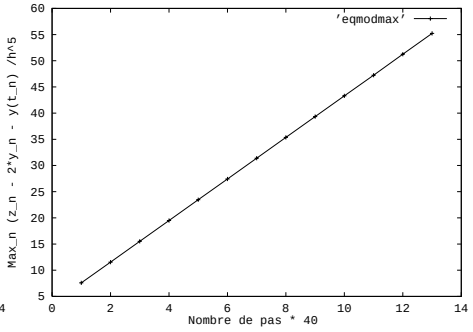



FIG. 3.7 – *Ordre de l'estimation EM - VIII*

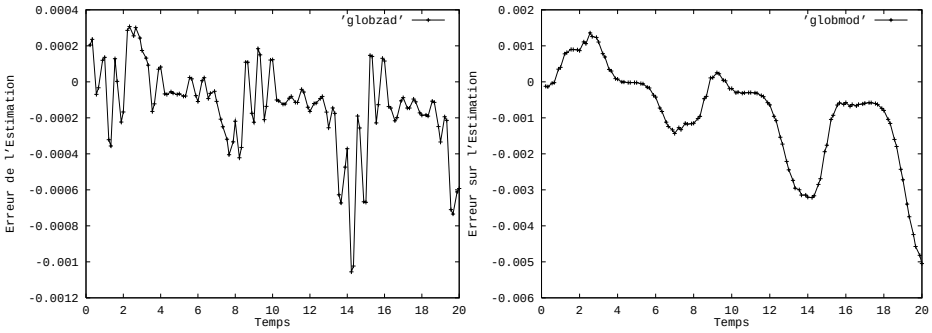



FIG. 3.9 – *Erreur de ZD - A3 - pas alterné*

FIG. 3.10 – *Erreur EM - A3 - pas alterné*

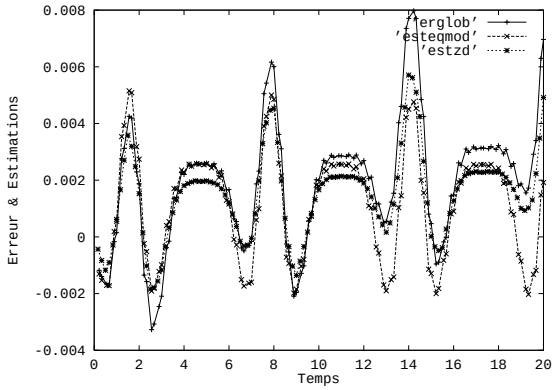



FIG. 3.8 – *Problème A3 - pas alterné*

Une dernière mesure permet de comprendre pourquoi l'estimateur ZD continue à fournir une estimation aussi précise à pas variable que EM alors qu'il a un caractère asymptotique moins bon. Il ne perd son caractère asymptotique que de peu. Parce que sur ce système le phénomène est très visible, on compare sur le problème A4 la convergence du rapport

$$\frac{\max_n \|Estimation(t_n)\|}{\max_n \|Erreur(t_n)\|}$$

vers un dans le cas d'une intégration à pas constant, puis sur la grille $h$, $2h$ pour ZD et pour EM.

La table TAB. 3.1 montre que dans le cas d'une intégration à pas constant, ZD et EM présentent un rapport voisin de un à $10^{-4}$ près. Et, quand on passe

| Nbr pas × 40 | ZD | | EM | |
|---|---|---|---|---|
| | PC | PV | PC | PV |
| 1 | $8.3\,10^{-4}$ | $2.2\,10^{-2}$ | $2.2\,10^{-4}$ | $8.9\,10^{-4}$ |
| 2 | $2.1\,10^{-4}$ | $2.2\,10^{-2}$ | $8.7\,10^{-5}$ | $3.7\,10^{-4}$ |
| 3 | $7.0\,10^{-5}$ | $2.0\,10^{-2}$ | $7.6\,10^{-5}$ | $2.0\,10^{-4}$ |
| 4 | $8.0\,10^{-5}$ | $1.9\,10^{-2}$ | $8.5\,10^{-5}$ | $1.5\,10^{-4}$ |
| 5 | $3.9\,10^{-5}$ | $1.9\,10^{-2}$ | $1.1\,10^{-4}$ | $1.4\,10^{-4}$ |
| 6 | $5.1\,10^{-5}$ | $1.9\,10^{-2}$ | $1.3\,10^{-4}$ | $1.5\,10^{-4}$ |
| 7 | $1.3\,10^{-4}$ | $1.9\,10^{-2}$ | $1.7\,10^{-4}$ | $1.6\,10^{-4}$ |
| 8 | $1.8\,10^{-4}$ | $1.9\,10^{-2}$ | $2.1\,10^{-4}$ | $1.8\,10^{-4}$ |
| 9 | $2.3\,10^{-4}$ | $1.9\,10^{-2}$ | $2.6\,10^{-4}$ | $2.1\,10^{-4}$ |
| 10 | $2.9\,10^{-4}$ | $1.8\,10^{-2}$ | $3.1\,10^{-4}$ | $2.4\,10^{-4}$ |
| 11 | $3.4\,10^{-4}$ | $1.8\,10^{-2}$ | $3.7\,10^{-4}$ | $2.9\,10^{-4}$ |
| 12 | $4.1\,10^{-4}$ | $1.8\,10^{-2}$ | $4.3\,10^{-4}$ | $3.3\,10^{-4}$ |
| 13 | $4.8\,10^{-4}$ | $1.8\,10^{-2}$ | $5.0\,10^{-4}$ | $3.8\,10^{-4}$ |

TAB. 3.1– — *Estimateur/Erreur - 1— pour un nombre croissant de pas constant (PC) et variable (PV)*

à la grille $h$, $2h$, seul EM conserve un aussi bon rapport. L'estimation fournie par ZD n'est plus voisine de un qu'à $10^{-2}$. Ce rapport ne tend donc plus vers un mais vers une valeur voisine, ce qui explique que l'estimation reste encore bonne.

# Chapitre 4

# Tests effectués sur Eurostag

Nous présentons les conclusions que l'on peut tirer de l'implantation de l'estimateur de Richardson dans EUROSTAG. Les simulations ont été effectuées en bloquant l'ordre maximum de la méthode d'intégration à deux. Cette restriction n'est pas faite uniquement pour ces tests. Pour des raisons de stabilité de l'intégration, il est préférable de se limiter à cet ordre. Par contre, le contrôle du pas d'intégration a été laissé inchangé.

Ces tests correspondent à trois scénarii. Le premier scénario (sim1) est constitué par un réseau de 40 variables. C'est un cas simple qui sert à tester le bon fonctionnement du logiciel et de l'estimation. Le deuxième (cigretri) comporte 73 variables dont une divergente. C'est cette variable que l'on surveille. Le troisième (voltcol) compte 270 variables. C'est un scénario au cours duquel est simulée une brusque chute de tension. Les automates du réseau essaient de maintenir la tension constante au nœud où se produit la chute de tension, sans succès.

De manière générale, les systèmes d'équations différentielles que doit traiter EUROSTAG sont fortement non-linéaires, mais leur jacobien est creux. Ces équations sont générées automatiquement à partir d'une interface graphique. Et, sauf sur des exemples simples de cas d'écoles, leur structure algébrique n'est pas disponible.

**Scénario Sim1**

La figure FIG. 4.1 montre le comportement d'une des variables de ce système. On peut voir sur la figure FIG. 4.2 que selon l'estimateur de Richardson l'erreur commise est de l'ordre de $2\,10^{-4}$ dans la phase transitoire. L'estimation relative, i.e. le rapport de la valeur absolue de l'estimation sur la valeur de la solution numérique, présente sur la figure FIG. 4.3 un pic au moment où la variable passe par zéro. En dehors de ce pic, l'estimation relative est de l'ordre de $10^{-3}$, ce qui fait moins de 0.1% d'erreur.

**Scénario Cigretri** Sur ce scénario, on observe une variable divergente,

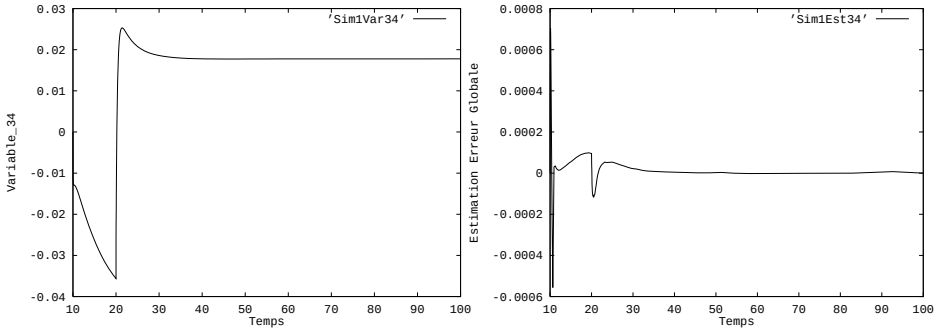



FIG. 4.1 – *Scénario Sim1 - variable 34*

FIG. 4.2 – *Scénario Sim1 - Estimation erreur variable 34*

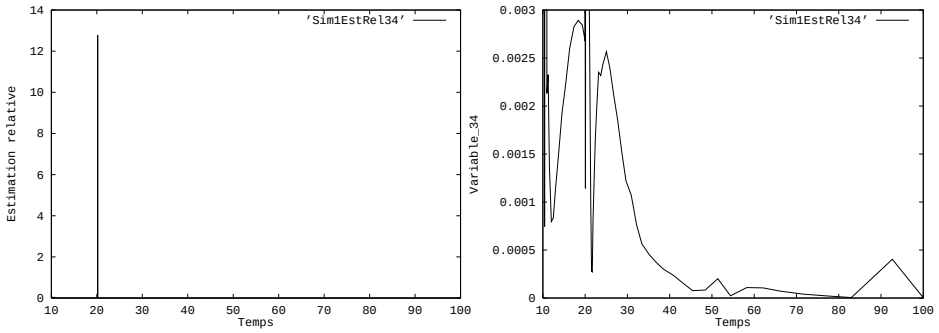



FIG. 4.3 – *Scénario Sim1 - Estimation relative*

FIG. 4.4 – *Scénario Sim1 - Estimation relative (zoom)*

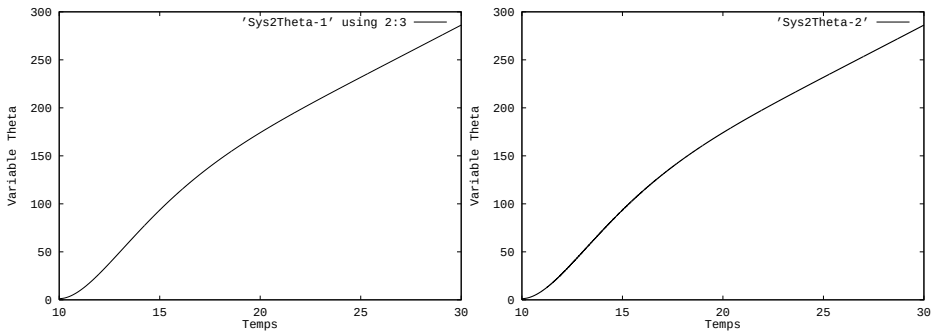



FIG. 4.5 – *Scénario Cigretri - Variable Theta - Tol.* $10^{-4}$

FIG. 4.6 – *Scénario Cigretri - Variable Theta - Tol.* $10^{-5}$

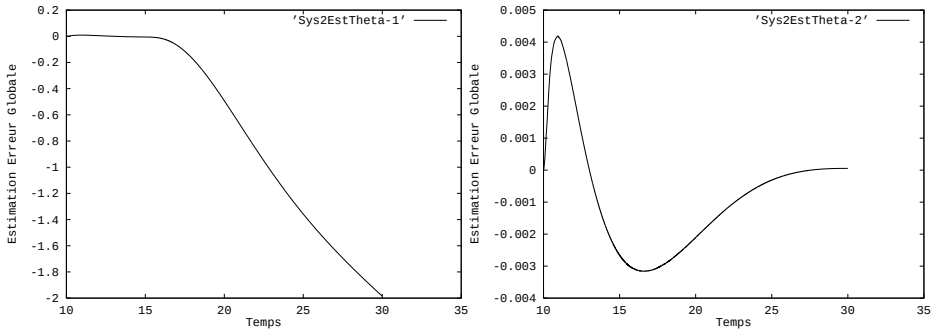



FIG. 4.7 – *Scénario Cigretri - Estimation Erreur Variable Theta - Tol.* $10^{-4}$

FIG. 4.8 – *Scénario Cigretri - Estimation Erreur Variable Theta - Tol.* $10^{-5}$

Theta, et une variable tendant vers zéro, IQ. On effectue une première simulation avec une tolérance de $10^{-4}$ puis avec une tolérance de $10^{-5}$. Le fait de faire varier les tolérances n'induit pas de comportement qualitativement différent sur la variable Theta (FIG. 4.5 et FIG. 4.6). Ce n'est pas le cas en revanche, des estimations. Elles sont visibles sur les figures FIG. 4.7 et FIG. 4.8.

La différence de comportement des estimations entre 10 et 15 secondes n'est qu'apparente et cela se voit en effectuant un zoom (FIG. 4.9 et FIG. 4.10). En revanche, entre 15 et 25 secondes, les deux estimations sont qualitativement différentes (FIG. 4.11 et FIG. 4.12), alors que rien de tel n'est visible entre les deux courbes de la variable.

Les variations de la variable IQ données sur les figures FIG. 4.13 et FIG. 4.14 permettent de mieux comprendre ce phénomène. Les estimations d'erreur obtenues sur les figures FIG. 4.15 et FIG. 4.16 exhibent des comportements très différents au-delà de 15s. Avant cet instant, un zoom sur l'intervalle $[10, 15]$ montrerait comme pour la variable Theta que leur forme est la même. La figure FIG. 4.18 montre non seulement des oscillations erratiques de l'estimation,

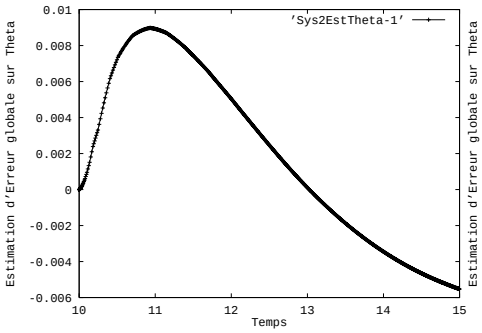



FIG. 4.9 – *Scénario Cigretri - Estimation Erreur Variable Theta - Tol.* $10^{-4}$ *- zoom*

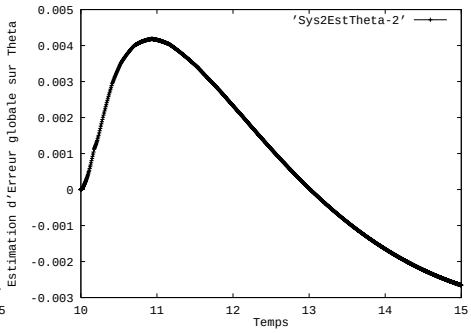



FIG. 4.10 – *Scénario Cigretri - Estimation Erreur Variable Theta - Tol.* $10^{-5}$ *(zoom)*

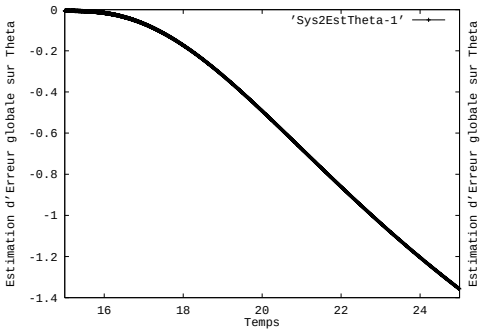



FIG. 4.11 – *Scénario Cigretri - Estimation Erreur Variable Theta - Tol.* $10^{-4}$ *(zoom 2)*

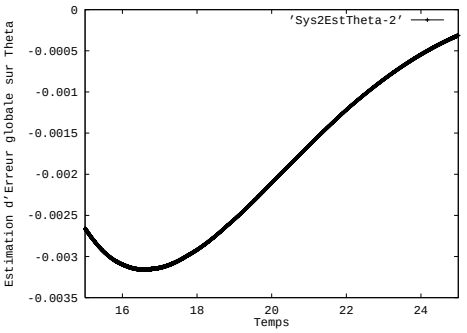



FIG. 4.12 – *Scénario Cigretri - Estimation Erreur Variable Theta - Tol.* $10^{-5}$ *(zoom 2)*

mais celle-ci estime l'erreur globale à un ordre de grandeur de $10^{-6}$ alors que pour les tolérances de $10^{-4}$ celle-ci était estimée à une valeur de l'ordre de $10^{-2}$ (Fig. 4.17). De plus, si l'on compare le comportement de l'estimation pour une tolérance de $10^{-4}$ avec la variation de IQ entre $[15 : 25]$ (Fig. 4.19), on remarque que l'estimation reproduit bien la brutale variation de la dérivée de la variable.

Cela peut s'interpréter comme un effet de la limite de la précision machine. Au moins pour ces variables, il est probable qu'au-delà de 15 secondes, l'erreur numérique soit prépondérante sur l'erreur de la méthode d'intégration.

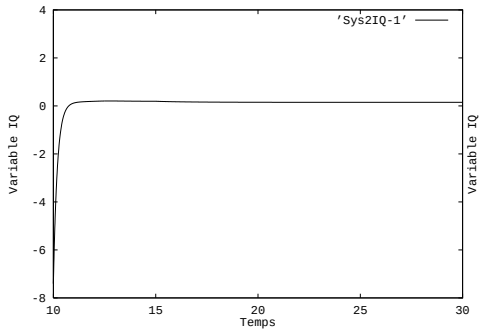



Fig. 4.13 – *Scénario Cigretri - Variable IQ - Tol.* $10^{-4}$

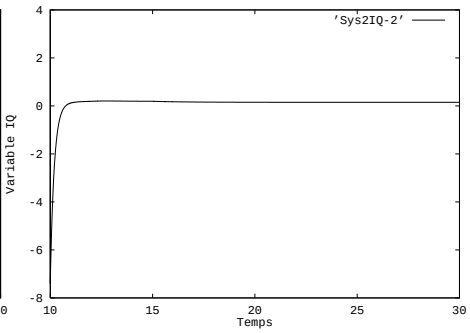



Fig. 4.14 – *Scénario Cigretri - Variable IQ - Tol.* $10^{-5}$

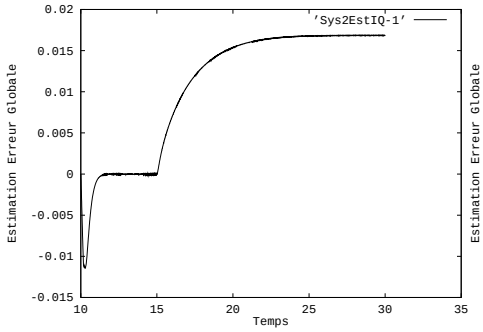



Fig. 4.15 – *Scénario Cigretri - Estimation Erreur Variable IQ Tol.* $10^{-4}$

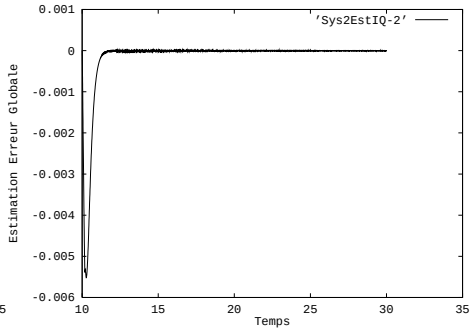



Fig. 4.16 – *Scénario Cigretri - Estimation Erreur Variable IQ - Tol.* $10^{-5}$

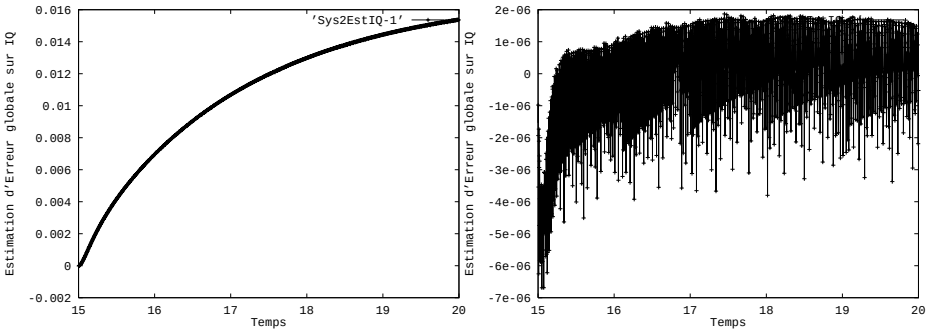



FIG. 4.17 – *Scénario Cigretri - Estimation Erreur Variable IQ - Tol.* $10^{-4}$

FIG. 4.18 – *Scénario Cigretri - Estimation Erreur Variable IQ - Tol.* $10^{-5}$

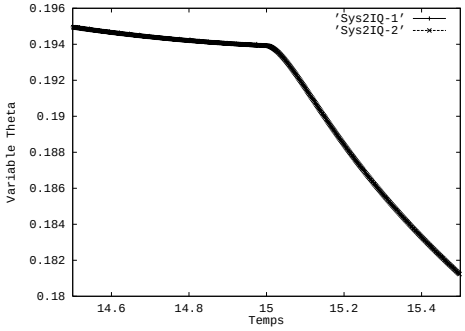



FIG. 4.19 – *Scénario Cigretri - Variable IQ - Tol.* $10^{-4}$ *et* $10^{-5}$

**Scénario Voltcol - Décalage des automates** Pour le troisième scénario, on peut voir sur la figure FIG. 4.20 comment la tension au nœud $N_{14}$ décroît petit à petit au cours de la simulation. Cette simulation dure plus de 7500 secondes. Peu après l'instant $t = 7500\,s$, la tension finit par chuter complètement. Les pics du graphe correspondent à des passages de seuils des automates. A ces instants, la variable connaît des sauts. Sur la figure FIG. 4.22, on peut voir comment l'implantation de l'estimateur de Richardson permet d'obtenir une estimation de l'erreur commise sur l'instant du déclenchement de l'automate.

Typiquement, les deux courbes montrent un $\Delta t$ de l'ordre de la seconde.

Toutefois, sur ce même scénario, du fait de la longueur de la simulation, on a aussi pu observer sur les variables de certains automates une estimation relative de l'ordre de un et ce, non pas de manière ponctuelle, mais bien constante. Cela

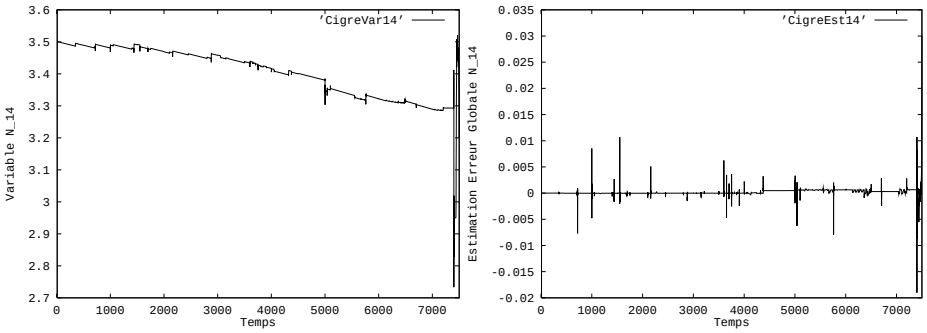



FIG. 4.20 – *Scénario Voltcol variable 14*




FIG. 4.21 – *Scénario Voltcol - Estimation Erreur variable 14*

correspond à un doute complet sur la valeur réelle de la variable de l'automate.

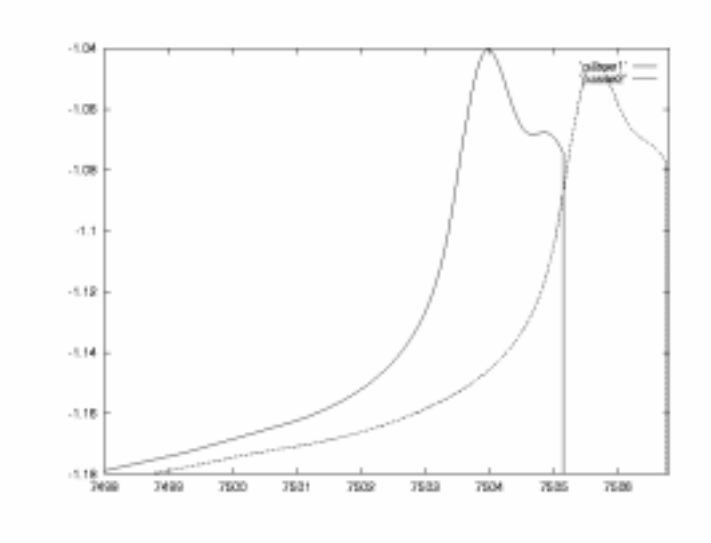



FIG. 4.22 – *Décalage d'une variable d'un automate*

# Chapitre 5

# Conclusion

L'analyse asymptotique de la Partie I et les tests de cette partie conduisent aux conclusions suivantes.

D'une part, quitte à utiliser un estimateur asymptotique, autant utiliser l'estimateur de Richardson. Malgré son coût élevé, il offre des avantages sérieux sur les autres estimateurs. Non seulement, il s'adapte facilement à de nombreuses méthodes d'intégration et demande peu d'hypothèses sur le comportement de l'erreur globale, mais aussi, il présente une estimation utile pour toute la gamme des tolérances possibles d'intégration. Ne supprimer qu'un terme du développement de l'erreur globale en faisant varier le pas est une manière plus sûre et plus directe d'obtenir une estimation fiable de l'erreur que de passer par de l'interpolation polynômiale pour essayer de gagner plusieurs termes. Enfin, cet estimateur semble plus robuste vis à vis des erreurs arithmétiques.

Les premiers tests effectués sur EUROSTAG ne doivent pas être sur-interprétés. Il s'agit avant tout d'une étude préliminaire visant à vérifier que l'estimateur implanté peut fournir une information utile à l'utilisateur. Notre rôle s'arrête ici. Une étude plus sytématique des simulations conduites par EUROSTAG ne peut être réalisée que de manière interne à la DER.

Dautre part, d'un point de vue théorique, si l'on souhaite utiliser les techniques développées ici pour doubler l'ordre de la méthode d'intégration, il nous semble que l'équation modifiée peut être une alternative intéressante au calcul d'une correction globale et aux estimateurs de Zadunaisky à perturbations non-autonomes.

En ce qui concerne l'erreur arithmétique dont nous avons supposé négligeables les effets, des techniques de vérification de l'influence des arrondis sur le calcul des solutions numériques peuvent être mises en œuvre pour déceler les pertes de précision (CADNA [AV96, BBC+97], AQUARELS [BBC+97]). Ce type de techniques a déjà fait l'objet d'implantations dans d'autres simulateurs de la DER. Elles nous semblent être un contrepoint utile aux techniques d'estimation de l'erreur globale.

# Troisième partie

# Intégration numérique dans $\mathbb{C}$

# Chapitre 1

# Introduction

On considère l'équation différentielle ordinaire :

$$\begin{array}{rcl} y'(t) & = & f(t, y(t)) \\ y(t_0) & = & y_0 \end{array} \tag{1.1}$$

où $t$ prend ses valeurs dans $\mathbb{C}$ et $f$ est une fonction holomorphe de $n+1$ variables complexes dans un voisinage de $(t_0, y_0)$. Dans ces conditions, la solution du problème (1.1) est holomorphe et définie dans un voisinage de $t_0$ [Car61]. Cette solution est égale à son développement en série entière sur un disque de rayon $R$. On s'intéresse au cas où $R$ est fini, c'est-à-dire au cas où la solution admet au moins une singularité à distance finie de $t_0$, et où il est possible de prolonger cette solution au-delà de son disque de convergence.

**Exemple 1** : *Singularité polaire*. Soit l'équation différentielle scalaire :

$$\begin{array}{rcl} y' & = & y^2, \\ y(0) & = & 1. \end{array} \tag{1.2}$$

Elle admet pour solution sous forme close $y(t) = 1/(1-t)$. Pour une méthode d'intégration numérique qui reste sur l'axe réel, il ne sera pas possible de calculer la valeur de cette solution au-delà de 1, car la solution devient infinie en 1. Pourtant, la forme close montre que les valeurs de la solution existent au-delà de $t = 1$. Si l'on se donne dans $\mathbb{C}$ un chemin $\phi : [0,1] \to \mathbb{C}$, $t \mapsto \phi(t)$ tel que $\phi(0) = 0$, et $\phi(1) = 2$, et tel que pour tout $t$, $\phi(t)$ ne prend pas la valeur 1, alors l'intégration de (1.2) le long de ce chemin fournira la valeur de la solution en 2, qui est ici -1.

Il peut ne pas y avoir unicité du prolongement analytique.

**Exemple 2** : *Singularité algébrique*. Soit l'équation différentielle scalaire suivante :

$$\begin{array}{rcl} y' & = & y^3, \\ y(0) & = & 1. \end{array} \tag{1.3}$$

Elle admet deux solutions sous forme close qui sont :

$$y_1(t) = \frac{\sqrt{1-2\,t}}{2\,t-1}, \quad y_2(t) = -\frac{\sqrt{1-2\,t}}{2\,t-1}$$

Ici, non seulement ses deux solutions deviennent infinies au point 1/2, mais les valeurs de la solution de (1.3) dépendent du chemin sur lequel l'intégration est effectuée autour de 1/2.

Le problème est alors de calculer *un* prolongement analytique de la solution de (1.1) en un point $T$ de $\mathbb{C}$.

Une méthode de résolution numérico-formelle de ce problème a été implantée par Chaffy [Cha95] dans le cas d'équations différentielles scalaires non-linéaires pour lesquelles la position des singularités est connue à l'avance. Sa méthode se fonde sur le procédé de Weierstrass. Le calcul des valeurs de la série entière est effectué grâce à un langage de calcul formel (REDUCE) qui permet d'exprimer les dérivées successives de la solution. Un prolongement analytique est alors effectué le long d'un chemin spécifié en entrée du programme.

Si l'on souhaite calculer les valeurs de la solution de (1.1) à l'aide d'une méthode d'intégration numérique classique, il semble difficile de se passer de cette spécification *a priori* du chemin d'intégration $\phi$ et donc, de la détermination des singularités de la solution. Toutefois, si dans le cas linéaire, la détermination des singularités est possible, il n'est pas toujours simple de le faire pour un système non-linéaire d'équations différentielles. Dans le cas complexe, les données que doit prendre en entrée un code d'intégration numérique semblent donc être : la fonction $f$, la condition initiale $(t_0, y_0)$, la tolérance utilisateur $\tau$, et le chemin d'intégration $\phi$.

Cependant, il existe des heuristiques purement numériques pour parvenir à détecter la présence de singularités sur la droite réelle à partir de la donnée seule de $f$ [SE92].

On montre dans le Chapitre 2 comment il est possible, en s'inspirant de ces heuristiques, de légèrement modifier les codes d'intégration numérique pour permettre de s'affranchir de la donnée *a priori* du chemin d'intégration et de la remplacer par la spécification d'un paramètre réel. Ainsi, le cas de l'implémentation d'un intégrateur dans le cas complexe se ramène au cas réel.

Les expérimentations numériques du Chapitre 3 sont là pour attester du bon fonctionnement de l'algorithme présenté dans le chapitre précédent (§3.1). De plus, on s'intéresse à l'influence d'une singularité isolée sur l'erreur globale d'une méthode d'intégration numérique (§3.2).

# Chapitre 2

# Algorithme d'intégration

Il est possible de montrer à travers un exemple simple comment le contrôle du pas d'intégration peut permettre d'effectuer une intégration dans $\mathbb{C}$ sans spécifier de chemin.

On reprend l'exemple (1.2) et on lui applique la méthode d'Euler explicite. La relation de récurrence obtenue est :

$$\begin{aligned} y_{n+1} &= y_n + h_n\, y_n^2, \\ y_0 &= 1. \end{aligned}$$

Son erreur de troncature est :

$$\varepsilon_n = y(t_n + h_n) - y(t_n) - h_n y(t_n)^2 .$$

Connaissant la solution de (1.2), cette erreur de troncature se réécrit :

$$\varepsilon_n = \frac{h_n^2}{(1-t_n)^2(1-t_n-h_n)}. \tag{2.1}$$

Le minimum du module de cette erreur est commandé par le maximum du module de $J = 1-t_n-h_n$. Soit $h_n = \rho_n \exp(i\,\theta_n)$, et soit $t_n = r_n \exp(i\,s_n)$. Soit $S$ le cercle de centre $t_n$ et de rayon $\rho_n$. Soit $M$ l'intersection de $S$ et du segment $[t_n, 1]$. Soit $M'$ le symétrique de M par rapport à $t_n$. Pour $\rho_n$ fixé, le maximum de $J$ est atteint au point $t_{n+1} = t_n + h_n$ qui est l'affixe de $M'$. Cela est résumé sur la figure Fig. 2.1.

Le résultat de ce calcul ne semble pas très encourageant. Pour le pas $h_0$, il conseille de prendre $h_0$ réel négatif. Cela ne nous rapproche pas de $T = 2$.

Toutefois, $M$ correspond au maximun de l'erreur de troncature et les points entre $M$ et $M'$ sont des points où elle décroît. On voit sur la figure 2.2 les variations de l'erreur de troncature commise par la méthode d'Euler explicite appliquée au problème (1.2) sur le premier pas. Le module du pas est fixé et son argument varie de 0 à $2\,\pi$. On constate que le minimum est bien atteint en $\pi$ et qu'entre ces deux valeurs, l'erreur de troncature diminue de manière

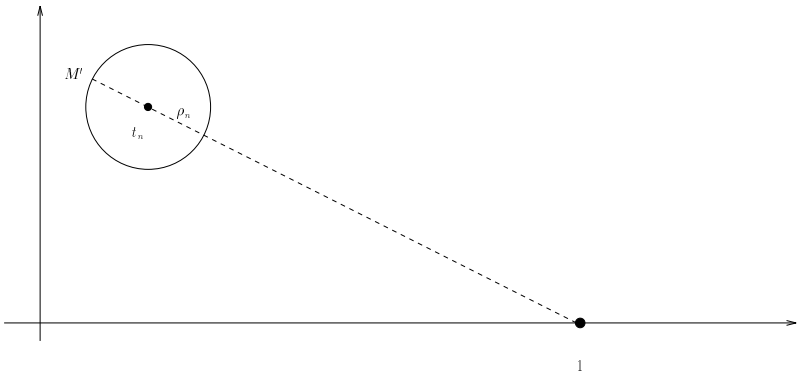



FIG. 2.1 – *Minimum de l'erreur de troncature*

régulière. Par conséquent, il sera possible de choisir une direction intermédiaire dans laquelle on diminuera suffisament l'erreur de troncature de façon à passer le test de tolérance.

D'une manière plus générale, dans le cas où il existerait plusieurs points singuliers, les variations de l'erreur de troncature ne seront pas aussi simples. Mais, en se limitant à un secteur bien choisi, il doit être possible de trouver à chaque instant une direction pour se faufiler entre ces points.

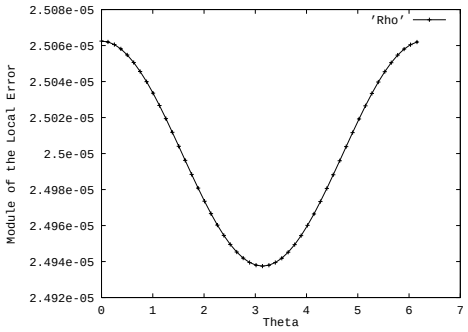



FIG. 2.2 – *Variation de l'erreur de troncature sur le premier pas du problème (1)*

On considère une méthode à un pas (2.2).

On note aussi $\psi(u,v,t)$, la fonction de $\mathbb{C}\times\mathbb{C}^N\times\mathbb{C}\to\mathbb{C}$ vérifiant $\psi(u,v,u)=v$ et $\psi_t(u,v,t)=f(t,\psi(u,v,t))$.

Le contrôle du pas permet à l'intégrateur de détecter la présence d'une singularité du type de l'exemple 1. Effectivement, au voisinage de ce point, il n'est plus possible de passer le test (2.9) avec un pas supérieur au pas $h_{min}$ imposé par l'arithmétique de la machine.

Dans le cas de l'intégration dans le champ complexe, il est possible d'utiliser cette sensibilité de l'estimation de l'erreur locale aux variations du flot de l'équation différentielle pour contourner d'éventuelles singularités.

L'idée consiste à essayer de changer de direction d'intégration avant d'effectuer une réduction du pas. On note $Pas_d$, le pas dans la direction courante et $\varepsilon_d$, le module de l'estimation de l'erreur locale dans cette direction. Le premier pas, $PasInit$, est pris dans la direction de $t_0$ à $T$ et son module est déterminé selon la règle donnée dans [HNW87], p. 182. Cela conduit à l'algorithme général de la figure FIG. 2.3.

---

Tcourant := $t_0$ ;
PasCourant := PasInit ;
Ycourant := $y_0$ ;
**Tant que** Tcourant $\neq$ T **faire**
    (Ysuivant, $\varepsilon$) := FaireUnPas(Tcourant, Ycourant, PasCourant) ;
    **Si** $\varepsilon \leq \tau$ **alors**
        Tcourant := Tcourant + Pascourant ;
        Ycourant := Ysuivant ;
        PasCourant := Allonger(PasCourant) ;
    **sinon**
        (Ysuivant, $Pas_d$, $\varepsilon_d$) := ChangerDeDirection(Tcourant, Ycourant, PasCourant) ;
        **Tant que** $\varepsilon_d > \tau$ **et** $\|Pas\| > H_{min}$ **faire**
            (Ysuivant, $\varepsilon_d$) := FaireUnPas(Tcourant, Ycourant, $Pas_d$) ;
            **Si** $\varepsilon_d \leq \tau$ **alors**
                Tcourant := Tcourant + $Pas_d$ ;
                Ycourant := Ysuivant ;
                PasCourant := $\dfrac{|Pas_d|}{|T - Tcourant|}\,(Tcourant - T)$ ;
            **sinon**
                $Pas_d$ := Réduire($Pas_d$) ;
            **fsi**
        **ftq**
    **fsi**
**ftq**

FIG. 2.3 – *Algorithme Général*

---

On remarque sur l'algorithme général qu'après avoir effectué un changement de direction, l'intégration reprend le long de la ligne droite qui joint l'instant courant et l'instant final. De plus, on remarquera qu'en cas d'échec dans la direction courante, la réduction du pas se fait dans la direction $Pas_d$, et non dans la direction courante.

Il existe plusieurs manières de concevoir un changement de direction. On peut par exemple effectuer une rotation d'angle $\theta_{max}$ (Figure FIG. 2.4), ou bien, se placer sur une demi-droite orthogonale à la direction courante (Figure FIG. 2.5). On ajoute dans ce cas, à $T - T_{courant}$, $\lambda \times i\,(T - T_{courant})$. Dans les deux cas, les points testés se trouveront sur la nouvelle direction choisie.

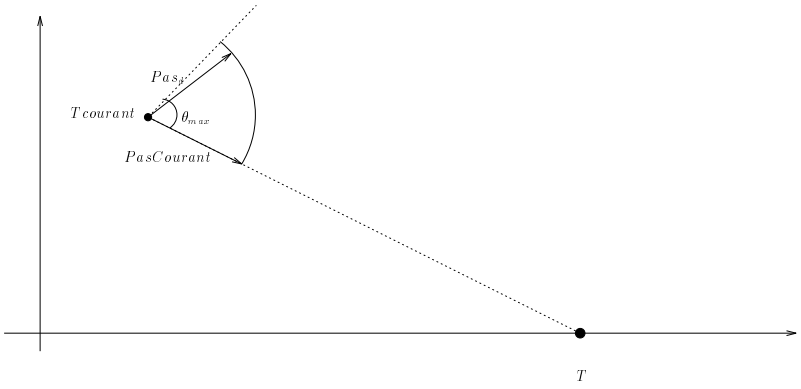



FIG. 2.4 – *Première stratégie de changement de direction*

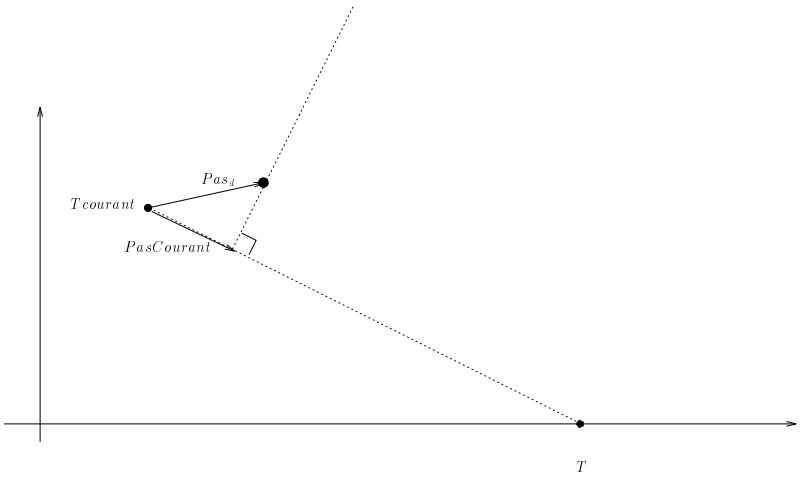



FIG. 2.5 – *Seconde stratégie de changement de direction*

Dans la mesure où l'on ne se préoccupe pas du coût de l'intégration, il est possible d'affiner les stratégies précédentes en effectuant une minimisation de la norme de l'estimation sur l'arc de cercle ou le segment considéré. Cette minimisationn peut se faire par échantillonage.

# Chapitre 3

# Expérimentations numériques

Les idées décrites dans la section précédente ont conduit à l'écriture d'un code d'intégration numérique d'équations différentielles ordinaires sur $\mathbb{C}$. Ce code est écrit en C et appelé *Ccart*. Nous avons implanté deux méthodes d'intégration. La première est la RKF5(4) [SW76]. La seconde est la RKDP8(7) [HNW87], p. 195.

Ces méthodes ont été choisies pour la simplicité de l'estimation de l'erreur locale qu'elles permettent par extrapolation locale. Les deux stratégies précédentes y sont implantées.

Les données que l'utilisateur doit fournir sont :

- la fonction $f$,
- la condition initiale $(t_0, y_0)$,
- l'instant final $T$,
- une tolérance absolue Atol, et une relative Rtol,
- un paramètre réel caractérisant la stratégie d'intégration.

Pour la première stratégie, ce paramètre est $\theta_{max}$, l'angle maximum de déviation autorisé. L'utilisateur peut le fixer comme il le souhaite, mais il est recommandé de ne pas lui donner de valeurs supérieures à $\pi/2$. Dans le cas contraire, l'intégrateur présente des oscillations superflues.

Pour la seconde stratégie, ce paramètre est une constante $\lambda$. On recommande de ne pas lui donner de valeur trop grande. Un $\lambda$ de 4 ou 5 est tout à fait suffisant.

Une capacité d'estimation de l'erreur globale par extrapolation de Richardson est aussi implantée.

Les cas d'échec du code d'intégration sont :

- impossibilité de trouver un pas de module supérieur à $h_{min}$ et permettant de passer le test sur l'erreur de troncature,

– dépassement du nombre d'évaluations de fonctions.

Le second cas garantit l'arrêt du programme. Ces cas d'échec sont les mêmes que pour un intégrateur réel.

Les expérimentations numériques ont été conduites d'abord dans le but de vérifier que l'intégrateur trouvait un chemin reliant $t_0$ à $T$ ne passant pas par une singularité (§3.1), ensuite pour mesurer l'influence d'une singularité sur l'erreur globale d'intégration (§3.2).

Pour effectuer ces expérimentations nous avons pris de nombreuses équations différentielles présentant des singularités. Nous présentons ici les résultats observés sur les exemples 1 et 2 de notre introduction qui sont typiques du comportement du code d'intégration Cart. Les deux exemples sont intégrés jusqu'à $T = 2$ pour le premier et jusqu'à $T = 1$ pour le second. La valeur de la solution pour le premier exemple est 1 et pour le second exemple, deux valeurs sont possibles suivant la nature du chemin, $i$ et $-i$.

A ceux deux exemples, nous ajoutons un exemple pour lequel il y a deux singularités :

$$\begin{array}{lcl} y' & = & t\,y^3, \\ y(-3) & = & 4\,i, \end{array} \tag{3.1}$$

dont la solution est :

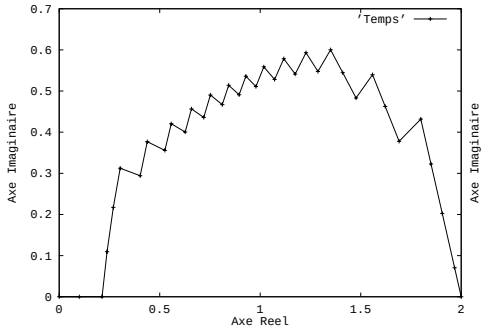
$$y(t) = \frac{\sqrt{143/16 - t^2}}{t^2 - 143/16}.$$

La première singularité se trouve sur l'axe réel et vaut environ -2.989 et 2.989. La première singularité est donc très proche de l'instant initial. De plus, nous l'intégrons jusqu'à $T = 3$.

Nous avons fixé les valeurs des paramètres d'intégration $\theta_{max}$ et $\lambda$ à 1.55 et 4, respectivement.

## 3.1 Contournement automatique d'une singularité

Les figures Fig. 3.1 et Fig. 3.3 montrent le chemin trouvé par l'intégrateur dans le cas de l'exemple 1, avec la première et la seconde stratégie respectivement. L'intégration a été effectuée avec la méthode RKF5(4) et une tolérance absolue de $10^{-5}$. On constate que dans les deux cas, l'intégrateur atteint $T = 2$ et que la valeur de la solution approchée est voisine de -1. Les chemins diffèrent peu d'une stratégie à une autre. Toutefois, la seconde stratégie a été conçue pour essayer de lisser le chemin d'intégration. On remarque sur les figures Fig. 3.2 et Fig. 3.4 qui représentent l'image du chemin d'intégration, i.e. les valeurs de la solution approchée en fonction du temps, que la seconde stratégie fournit un résultat légèrement plus lisse que la première. Ce fait s'accentue sur des exemples plus difficiles.



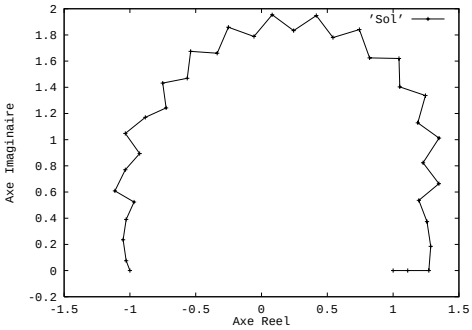
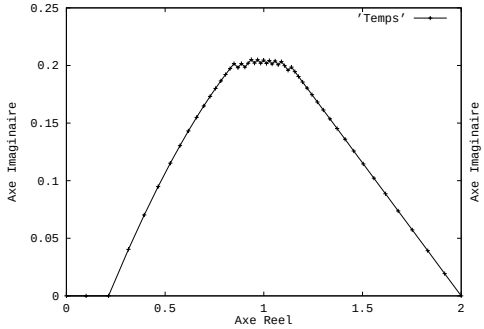




FIG. 3.1 – *Chemin d'intégration - Exemple 1 - Méthode 1 - Stratégie 1 - atol* $= 10^{-5}$

FIG. 3.2 – *Solution - Exemple 1 - Méthode 1 - Stratégie 1 - atol* $= 10^{-5}$



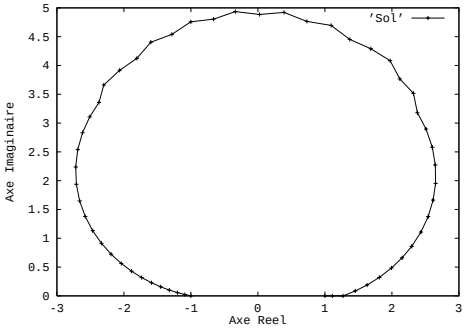
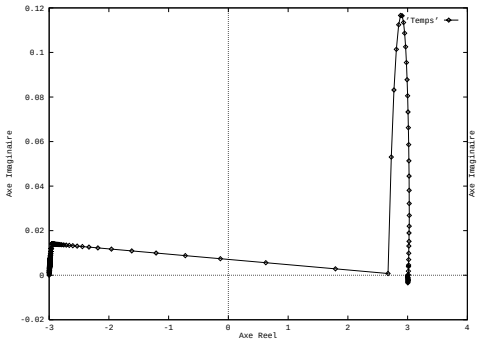




FIG. 3.3 – *Chemin d'intégration - Exemple 1 - Méthode 1 - Stratégie 2 - atol* $= 10^{-5}$

FIG. 3.4 – *Solution - Exemple 1 - Méthode 1 - Stratégie 2 - atol* $= 10^{-5}$



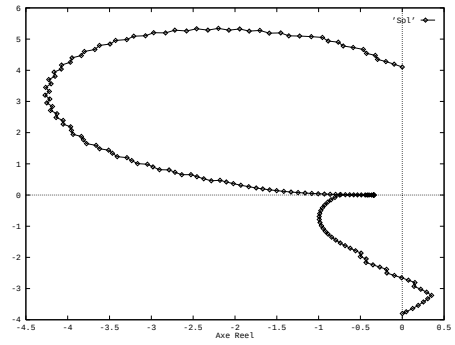
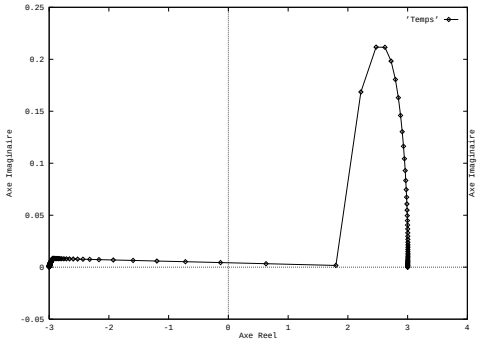




FIG. 3.5 – *Chemin d'intégration - Exemple 2 - Méthode 1 - Stratégie 1 - atol* $= 10^{-5}$

FIG. 3.6 – *Solution - Exemple 2 - Méthode 1 - Stratégie 1 - atol* $= 10^{-5}$



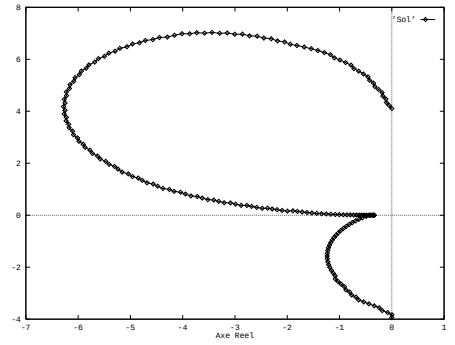




FIG. 3.7 – *Chemin d'intégration - Exemple 2 - Méthode 2 - Stratégie 2 - atol* $= 10^{-5}$

FIG. 3.8 – *Solution - Exemple 2 - Méthode 2 - Stratégie 2 - atol* $= 10^{-5}$

Les figures FIG. 3.5 et FIG. 3.7 montrent les chemins trouvés par les deux stratégies sur l'exemple 3. Les singularités qui se situent proches de l'instant initial et de l'instant final n'empêchent pas l'intégrateur de trouver un chemin, même si ceux-ci ont une forme plus complexe que dans le cas précédent.

La différence entre les deux stratégies est plus manifeste. Les figures FIG. 3.6 et FIG. 3.8 montrent que la seconde stratégie fournit une solution plus lisse que la première.

De plus, au voisinage de l'instant final $T = 3$, la figure FIG. 3.10 montre que la première stratégie conduit à un chemin qui spirale légèrement autour de $T$, alors que celui fourni par la seconde est plus direct. Ce phénomène d'enroulement s'observe souvent avec la première stratégie quand l'instant final est voisin d'une singularité.

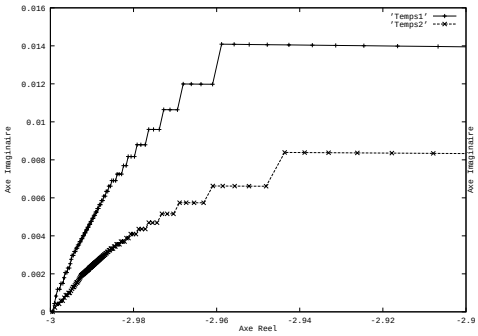



FIG. 3.10 – *Agrandissement au voisinage de $t_0$*

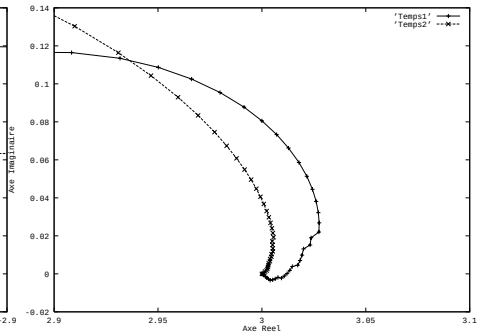



FIG. 3.11 – *Agrandissement au voisinage de $T$*

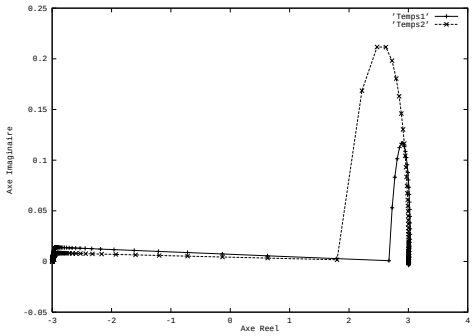
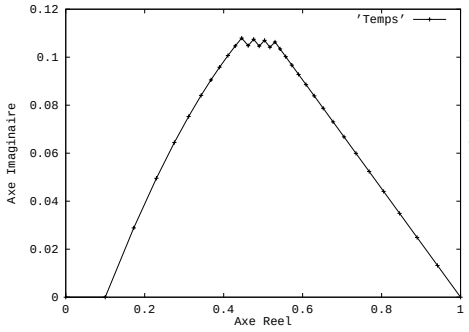



FIG. 3.9 – *Chemin selon les deux stratégies*

Les figures FIG. 3.12 et FIG. 3.14 montrent ce qui se produit quand la solution de l'équation différentielle est multiforme. Nous avons effectué un tour autour de la singularité en donnant comme condition initiale pour la seconde partie de l'intégration (FIG. 3.14 et FIG. 3.13) la valeur trouvée à la fin de la première intégration (figures FIG. 3.12 et FIG. 3.13). On constate qu'après un premier tour (FIG. 3.12 à FIG. 3.15), la solution n'a pas repris sa valeur initiale de 1 et qu'elle vaut -1.

Une question intéressante qui se pose pour cette technique d'intégration est de savoir ce qui se passe pour le chemin d'intégration choisi par l'intégrateur quand la tolérance devient de plus en plus petite. La figure FIG. 3.16 montre ce qui se produit sur l'exemple 1 quand on fait passer la tolérance utilisateur de $10^{-5}$ à $10^{-11}$. L'intégration a été effectuée avec la méthode RKDP8(7) qui supporte mieux des tolérances élevées que la RKF5(4), et avec la seconde stratégie. On constate que la suite des chemins s'accumule sur une courbe continue et hérissée de dentelures. Ce phénomène peut être visualisé sur les autres exemples et avec l'autre stratégie.



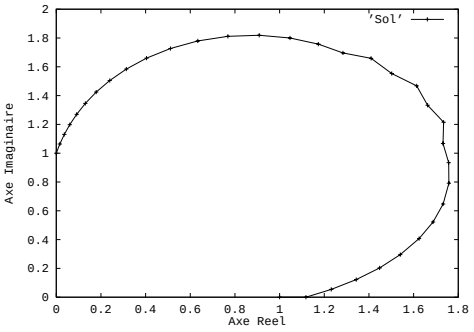
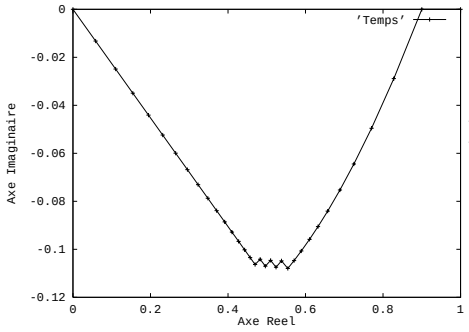




FIG. 3.12 – *Chemin d'intégration - Exemple 2 - atol* $= 10^{-5}$

FIG. 3.13 – *Solution - Exemple 2 - atol* $= 10^{-5}$



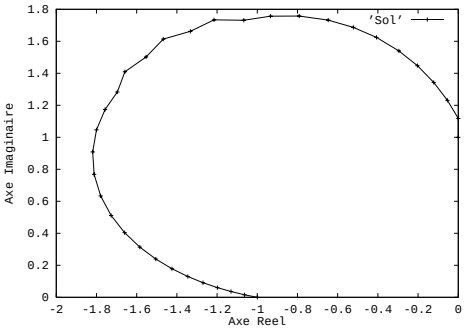




FIG. 3.14 – *Chemin d'intégration - Exemple 2 - atol* $= 10^{-5}$

FIG. 3.15 – *Solution - Exemple 2 - atol* $= 10^{-5}$

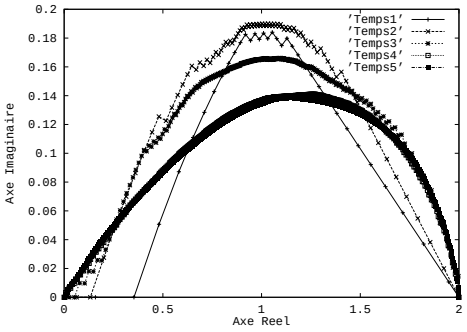



FIG. 3.16 – *convergence du chemin en fonction de la tolérance*

Les solutions d'équations différentielles dans le champs complexe peuvent présenter une *frontière naturelle* de singularités autour de la condition initiale. Dans un tel cas, il n'existe pas de prolongement analytique de la solution au-delà de cette frontière.

Un exemple de système différentiel donnant naissance à un phénomène de ce type est fourni par l'équation de Chazy [CO96] :

$$\begin{aligned} y_1' &= y_2 y_3 - y_1 y_2 - y_1 y_3, \\ y_2' &= y_1 y_3 - y_2 y_1 - y_2 y_3, \\ y_3' &= y_1 y_2 - y_3 y_1 - y_3 y_2. \end{aligned} \tag{3.2}$$

Sur cette équation, le code d'intégration sort après avoir atteint le pas $H_{min}$.

Bien que ce ne soit pas exactement une limitation, nous insistons sur le fait que l'intégrateur prendra aussi des valeurs complexes si les variations de la solution sont trop importantes sur la droite réelle. Par exemple, des tests effectués sur $y' = 2x\,y$ nous ont montré que sur une gaussienne, l'intégrateur contourne le sommet de la solution.

## 3.2 Influence d'une singularité sur l'erreur globale

On suppose ici que le segment d'intégration $[0, T]$ ne contient plus de singularités de la solution de l'équation différentielle (1.1). En revanche, on s'intéresse au cas où ce segment est voisin d'une singularité.

Notre propos ici est de savoir si cette stratégie permet d'obtenir un gain sur l'erreur globale à l'instant final d'intégration. Effectivement, il n'est plus possible de comparer ailleurs qu'à cet instant les différentes valeurs obtenues sur différents chemins.

Un premier travail dans le sens de l'étude de l'influence d'une singularité sur un intégrateur numérique a été réalisé dans [ATV97]. On y montrait comment à l'aide d'un outil de visualisation d'ensembles de solutions, un utilisateur pouvait

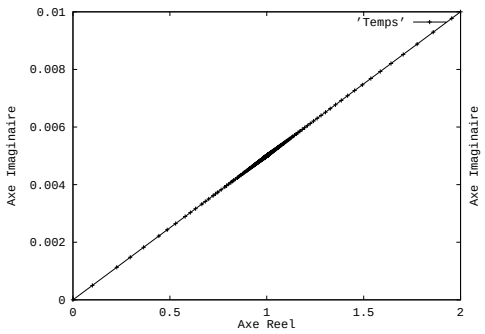



FIG. 3.17 – *Chemin d'intégration rectiligne*

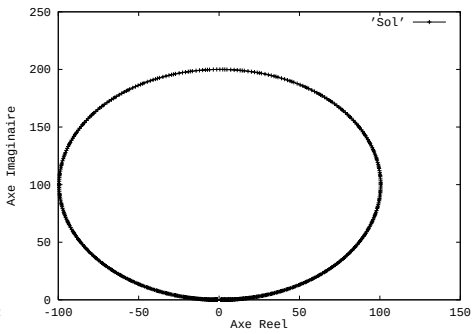



FIG. 3.18 – *Valeurs de la solution*

de manière interactive modifier les chemins d'intégrations pour faire décroître une estimation de l'erreur globale.

Cependant, l'intégrateur utilisé alors ne bénéficiait pas d'une stratégie de minimisation de l'erreur locale. Cette stratrégie conduit à des écarts par rapport à la trajectoire rectiligne que peut prendre un intégrateur classique. On peut alors se demander si le coût supplémentaire que représente cette minimisation est au moins compensé par une erreur globale plus petite à l'arrivée.

Pour effectuer alors nos comparaisons, nous utilisons une équation à singularité essentielle $\dot{y} = \dfrac{1}{(1-t)^2} y$, $y(0) = 1$, avec $T = 2.0 + 0.1\,i$, ainsi que l'équation de l'exemple 1 que nous intégrons jusqu'à $T = 2.0 + 0.01\,i$.

Il est possible d'effectuer avec notre code d'intégration ou bien une intégration sur le segment $[0, T]$ en mettant à zéro le paramètre de contrôle de changement de direction - on retrouve alors une intégration numérique réelle classique - ou bien en mettant ce paramètre à une valeur strictement positive, on autorise l'intégrateur à dévier de sa trajectoire rectiligne.

Ainsi, on peut voir sur les figures FIG. 3.17 et FIG. 3.19 ce que l'on obtient comme variations du pas d'intégration quand on fixe respectivement les valeurs du paramètre de déviation à zéro et à 4,0 sur le problème de l'exemple 1. On constate d'une part que l'intégration forcée le long du segment conduit à une accumulation de points au voisinage de la singularité (FIG. 3.17), d'autre part, que même si la singularité ne se trouve pas exactement sur le segment $[0, T]$, la stratégie de contournement conduit à s'en écarter (FIG. 3.19).

Si on compare les valeurs que prennent les solutions numériques dans les deux cas (FIG. 3.18 et FIG. 3.20), on constate que celles-ci sont beaucoup plus grandes en norme sur le chemin rectiligne que sur le chemin dévié.

Les mesures du nombre d'évaluations de fonctions, du nombre de pas et de l'erreur commise en $T$ en fonction de la tolérance absolue et pour différentes valeurs du paramètre de déviation sont données dans les figures FIG. 3.21 à FIG. 3.24.

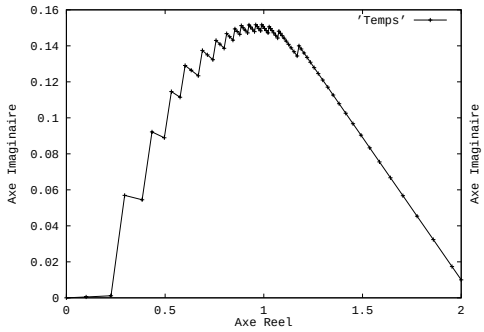



FIG. 3.19 – *Chemin d'intégration avec déviation*

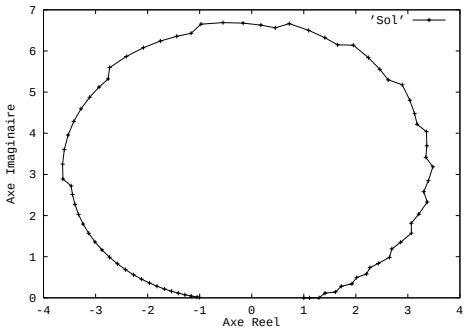



FIG. 3.20 – *Valeurs de la solution*

On constate très clairement qu'avec un paramètre de déviation non nul, on obtient une erreur plus petite avec moins de pas et d'évaluations de fonctions.

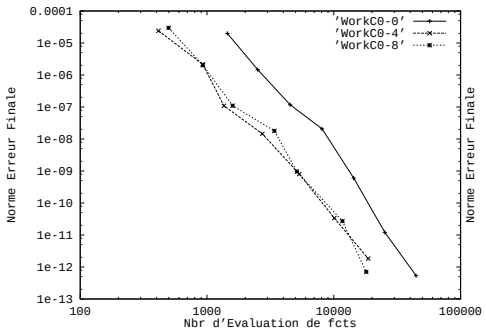



FIG. 3.21 – *Nombre d'évaluations de fonctions vs Erreur - exemple 1*

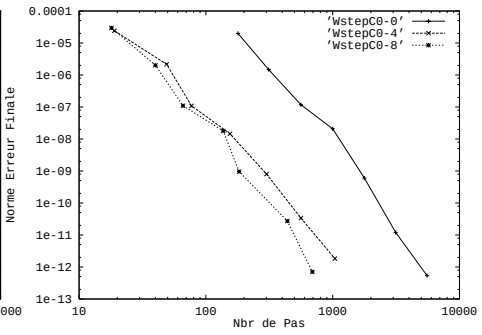



FIG. 3.22 – *Nombre de pas vs Erreur - exemple 1*

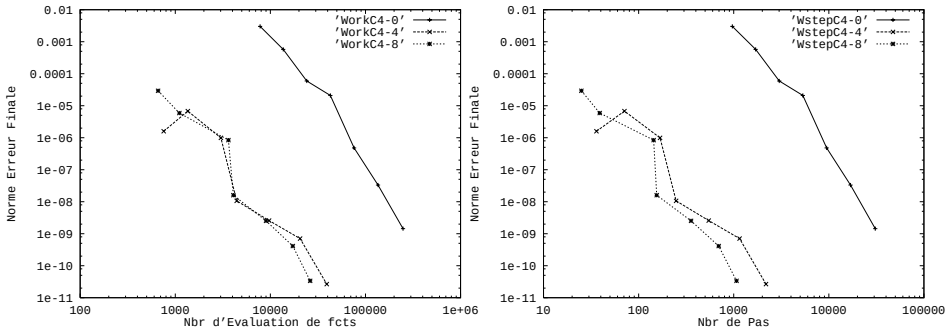



FIG. 3.23 – *Nombre d'évaluations de fonctions vs Erreur - sing. ess.*




FIG. 3.24 – *Nombre de pas vs Erreur - sing. ess.*

# Chapitre 4

# Conclusion

Nous espérons avoir montré dans cette partie les possibilités techniques que présentait le contrôle de l'erreur locale pour l'intégration numérique des équations différentielles ordinaires dans le champ complexe. Quand il existe, le prolongement analytique d'une solution est un procédé purement local qui s'effectue de proche en proche. Il n'est pas étonnant alors de pouvoir le réaliser de manière automatique en évitant les singularités de la solution.

Bien que le code que nous ayons écrit ne soit qu'un prototype, il permet de montrer qu'une minimisation locale dans $\mathbb{C}$ de l'erreur locale conduit à des valeurs en sortie plus précises et à un moindre coût qu'une intégration imposée sur un segment.

Enfin, il faut reconnaître que l'intégration numérique dans $\mathbb{C}$ n'est pas un domaine très couru de l'analyse numérique. Les problèmes de simulation numérique sont généralement posés avec un temps réel. Une tâche que nous nous donnons pour l'avenir est de rechercher des applications liées à ce type d'intégration.

# Bibliographie


[ABJ93] J. Y. Astic, A. Bihain, and M. Jerosolimski. The mixed Adams-BDF variable step size algorithm to simulate transient and long term phenomena in power systems. In *IEEE PES Summer Meeting Vancouver*, 1993.

[Aïd96] R. Aïd. Estimation de l'erreur globale pour l'intégration numérique d'équations différentielles ordinaires. Technical Report 159, LMC-IMAG, Grenoble, Mars 1996.

[Aïd97] R. Aïd. Asymptotic global error estimation for variable step-size and one-step methods. Technical Report RR 984-M, LMC-IMAG, Octobre 1997.

[AL97] R. Aïd and L. Levacher. Numerical investigations on global error estimation for ordinary differential equations. *J. of Comput. and Appl. Math.*, 82(1-2):21–39, sept. 1997.

[Alt82] R. Alt. Evaluation de l'erreur de discrétisation des méthodes à pas séparés à l'aide d'interpolation rationnelle. In *Les mathématiques de l'informatique*, AFCET Colloq. Paris, pages 515–524, 1982.

[ATV97] R. Aïd, L. Testard, and G. Villard. Global error visualization. *soumis à J. of Universal Computer Science*, 1997.

[AV96] R. Alt and J. Vignes. Validation of results of collocation methods for ODEs with the CADNA library. *Applied Numerical Mathematics*, 21:119–139, 1996.

[BBC+97] J-C. Bajard, O. Beaumont, J-M. Chesneaux, M. Daumas, J. Erhel, D. Michelucci, J-M. Muller, B. Philippe, N. Revol, J-L. Roch, and J. Vignes. *Qualité des Calculs sur Ordinateurs.* Masson, 1997.

[BBH89] P. N. Brown, G. D. Byrne, and A. C. Hindmarsh. VODE : A variable coefficient ODE solver. *SIAM J. Sci. Stat. Comput.*, 10(5):1038–1051, 1989.

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

[BH75] G. D. Byrne and A. C. Hindmarsh. A polyalgorithm for the numerical solution of ordinary differential equations. *ACM Trans. Math. Software*, 1:71–96, 1975.

[BS66] R. Burlisch and J. Stoer. Asymptotic upper and lower bounds for results of extrapolation methods. *Numer. Math*, 8:93–104, 1966.

[But87] J. C. Butcher. *The Numerical Analysis of Ordinary Differential Equations*. Wiley-Interscience Publication, 1987.

[Car61] H. Cartan. *Théorie élémentaire des fonctions analytiques d'une ou plusieurs variables complexes*. Hermann, 1961.

[Cha95] C. Chaffy. The analytic continuation process : from computer algebra to numerical analysis. In *ISSAC'95*, pages 216–222. ACM, 1995.

[CHMR96] M. Calvo, D. J. Higham, J. I. Montijano, and L. Randez. Global error estimation with adaptive explicit Runge-Kutta methods. *IMA Jour. of Numer. Anal.*, 16:47–63, 1996.

[CHMR97] M. Calvo, D. J. Higham, J. I. Montijano, and L. Randez. Step-size selection fot tolerance proportionality in explicit Runge-Kutta codes. *Advances in Computational Mathematics*, 7(3):361–382, 1997.

[Cho59] M. Choquet. *Équations différentielles*. Centre de Documentation de la faculté de la Sorbonne, 5, Place de la Sorbonne, 1959.

[Cho97] R. Chopinet. Implantation d'un estimateur de l'erreur numérique globale dans le code EUROSTAG. Technical Report HR-41/97-072, Électricité de France-Direction des Études et Recherches, Départements FCR, Clamart, France, 1997.

[CK63] F. Ceschino and J. Kuntzmann. *Problèmes différentiels de conditions initiales*. Dunod (Paris), 1963.

[CL84] M. Crouzeix and F. J. Lisbona. The convergence of variable-stepsize variable-formula multistep methods. *SIAM J. Numer. Anal.*, 21:512–534, 1984.

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

[CMSS94] M. P. Calvo, A. Murua, and J. M. Sanz-Serna. Modified equations for odes. In P. E. Kloeden and K. J. Palmer, editors, *Chaotic Numerics*, volume 172 of *Contemporary Mathematics*, pages 63–74. AMS, July 1994.

[CO96] P. A. Clarkson and P. J. Olver. Symmetry and the Chazy equation. *J. Differ. Equations*, 124(1):225–246, 1996.

[Cor94] R. E. Corless. *Error Backward*, volume 172 of *Contemporary Mathematics*, pages 31–62. AMS, 1994.

[DDP84] J. R. Dormand, R.R Duckers, and P. J. Prince. Global error estimation with Runge-Kutta methods. *IMA Journal of Numerical Analysis*, 4:169–184, 1984.

[DGP94] J. R. Dormand, J. P. Gilmore, and P. J. Prince. *Globally embedded Runge-Kutta schemes*, volume 1 of *Annals of Numerical Mathematics*, pages 97–106. Baltzer A. G. Science Publishers, 1994.

[DLMP89] J. R. Dormand, M. A. Lockyer, N. E. McGorrigan, and P. J. Prince. Global error estimation with Runge-Kutta triples. *Computers Math. Apllic.*, 18(9):835–846, 1989.

[Dor96] J. R. Dormand. *Numerical Methods for Differential Equations*. CRC Press Inc, New York, 1996.

[DP80] J. R. Dormand and P. J. Prince. A family of embedded Runge-Kutta formulae. *J. Comp. Appl. Math.*, 5:977–989, 1980.

[DP85] J. R. Dormand and P. J. Prince. Global error estimation with Runge-Kutta methods II. *IMA Journal of Numerical Analysis*, 5:481–497, 1985.

[DP89] J. R. Dormand and P. J. Prince. Practical Runge-Kutta processes. *SIAM J. Sci. Stat. Comput.*, 10(5):977–989, 1989.

[DS76] D. P. Davey and N. F. Stuart. Guaranted error bounds for initial value problem using polytope arithmetic. *BIT*, 16:257–268, 1976.

[EH79] B. Epstein and D. L. Hicks. Comparison between two error estimation procedures. In P. C. C. Wang, editor, *Information Linkage Between Applied Mathematics and Industry*, pages 293–298. Academic Press, 1979.

[Feh66] E. Fehlberg. Low-order classical runge-kunta formulas with stepsize control and their application to some heat transfer problems. Technical Report TR R-315, Nasa, 1966.

[Fra76] R. Frank. The method of iterated defect-correction and its application to two-point boundary value problems, Part I. *Num. Math.*, 25:409–419, 1976.

[Fra77] R. Frank. The method of iterated defect correction and its application to two-point boundary value problems, Part II. *Num. Math.*, 27:407–420, 1977.

[FU75] R. Frank and C. W. Ueberhuber. Iterated defect correction for Runge-Kutta methods. Technical Report 14/75, Institut für Numerische Mathematik, T. U. Wien, 1975.

[FU78] R. Frank and C. W. Ueberhuber. Iterated defect correction for differential equations, Part I: Theoretical results. *Computing*, 20:207–228, 1978.

[Gra64] W. B. Gragg. *Repeated extrapolation to the limit in the numerical solution of ordinary differential equations.* Phd thesis, University of California, 1964. see also SIAM J. Numer. Anal., ser. B, vol. 2, p. 384-403, 1965.

[GSB87] I. Gladwell, L. F. Shampine, and R.W. Brankin. Automatic selection of the initial step size for an ODE solver. *J. Comput. and Appl. Math.*, 18:175–192, 1987.

[GSS86] D. F. Griffiths and J. M. Sanz-Serna. On the scope of the method of the modified equations. *SIAM J. Sci. Stat. Comput.*, 7:994–1008, 1986.

[GT74] C. W. Gear and K. W. Tu. The effect of variable mesh size on the stability of multistep methods. *SIAM J. Numer. Anal.*, 11:1044–1058, 1974.

[GW74] C. W. Gear and D. S. Watanabe. Stability and convergence of variable order multistep methods. *SIAM J. Numer. Anal.*, 11:1024–1043, 1974.

[Hai78] E. Hairer. On the order of iterated defect correction - An algebraic proof. *Num. Math.*, 29:409–424, 1978.

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

[Hai97] E. Hairer. Communication privée, juin 1997.

[Hal69] J. K. Hale. *Ordinary Differential Equations.* John wiley and Sons, 1969.

[HEFS72] T. E. Hull, W. H. Enright, B. M. Fellen, and A. E. Sedgwick. Comparing numerical methods for ordinary differential equations. *SIAM J. Numer. Anal.*, 9(4):603–637, 1972.

[Hen62] P. Henrici. *Discrete Variable Methods in Ordinary Differential Equations.* John Wiley & Sons, Inc., 1962.

[Hig91] D. J. Higham. Global error versus tolerance for explicit Runge-Kutta methods. *IMA J. Numer. Anal.*, pages 457–480, 1991.

[HNW87] E. Hairer, S.P. Norsett, and G. Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems.* Springer-Verlag, 1987.

[HS97] E. Hairer and D. Stoffer. Reversible long-term integration with variable step sizes. *SIAM on Scientific Computing*, 18(1), 1997.

[HW96] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II - Stiff and Differential-Algebraic Problems - Second Revised Edition.* Springer, 1996.

[Jea97] C.P. Jeannerod. Estimation de l'erreur globale lors de l'intégration numérique des systèmes algbro-différentiels d'indice un et deux. Rapport de stage de DEA-Ensimag, juin 1997.

[JV97] C.-P. Jeannerod and J. Visconti. Global error estimation for index 1 and 2 DAEs. *Numerical Algorithm, à paraître*, 1997.

[Moo76] R. E. Moore. *Interval Analysis.* Prentice-Hall, Englewood Cliffs, N. J., 1976.

[Pet86] P. J. Peterson. Global error estimation using defect correction techniques for explicit Runge-Kutta methods. Tech. Rep. 192/86, Departement of Computer Science, University of Toronto, Canada, 1986.

[Pro80] A. Prothero. Estimating the accuracy of numerical solutions to ordinary differential equations. In I. Gladwell and D. K. Sayers, editors, *Computational Techniques for Ordinary Differential Equations*, pages 103–128. Academic Press London, 1980.

[Rhi97] R. Rhim. Implicit methods for enclosing solutions of odes. 1997. SCAN97.

[Rih94] R. Rihm. Interval methods for initial value problems. In J. Hertzberger, editor, *Topics in Validated Numerics*, Proceedings of IMACS-GAMM International Workshop on Validated Computation, pages 173–207, 1994.

[SB93] J. Stoer and R. Burlisch. *Introduction to Numerical Analysis - Second Edition.* Springer-Verlag, 1993.

[SE92] H. Suhartanto and W. Enright. Detecting and locating a singular point in the numerical solution of IVPs for ODEs. *Computing*, 48:161–175, 1992.

[SG73] L. F. Shampine and M. K. Gordon. Solving ordinary differential equations with ODE, STEP and INTRP. Technical Report SLA-73-1060, Sandia Laboratories, 1973.

[Sha77] L. F. Shampine. Local error estimation for ordinary differential equations. *Applied Mathematics and Computation*, 3:189–210, 1977.

[Sha84] L. F. Shampine. Asymptotic bounds on the errors of one-step methods. *Numer. Math.*, 45:201–206, 1984.

[Sha94] L. F. Shampine. *Numerical Solution of Ordinary Differential Equations.* Chapman and Hall, 1994.

[Ske86] R. D. Skeel. Thirteen ways to estimate global error. *Numer. Math.*, 48:1–20, 1986.

[SN91] D. Stoffer and K. Nipp. Invariant curves for variable step-size integrators. *BIT*, 31:169–180, 1991.

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

[Ste74] H. J. Stetter. Economic global error estimation. In Willoughby, editor, *Stiff Differential Systems*, pages 245–258. Plenum Press, 1974.

[Ste78a] H. J. Stetter. The defect correction principle and discretisation methods. *Numer. Math.*, 29:425–443, 1978.

[Ste78b] H. J. Stetter. Global error estimation in ode-solvers. In G. A. Watson, editor, *Numerical Analysis*, volume 630 of *Lecture Notes in Math.* Springer Verlag, 1978.

[Ste79] H. J. Stetter. Global error estimation in Adams PC-codes. *ACM Trans. on Math. Soft.*, 5(4):415–430, Dec. 1979.

[Ste80a] H. J. Stetter. Global error estimation in ordinary initial value problems. In *Lecture Notes in Mathematic*, volume 968. Springer Verlag, 1980.

[Ste80b] H. J. Stetter. Tolerance proportionality in ODE-codes. In R. März, editor, *Proc. Second Conf. on Numerical Treatment of Ordinary Differential Equations*. Seminarberichte 32, 1980.

[Stu97] A. M. Stuart. Probabilistic and deterministic convergence proofs for sofware for initial value problems. *Numerical Algorithms*, 14:227–260, 1997.

[SW76] L. F. Shampine and H. A. Watts. Global error estimation for ordinary differential equations. *ACM Trans. Math. Softw.*, 2:172–186, 1976.

[SZ90] L. F. Shampine and W. Zhang. Rate of convergence of multistep codes started by variation of order and stepsize. *SIAM J. Numer. Anal.*, 27(6):1506–1518, 1990.

[WH74] R. F. Warmimg and B. J. Hyett. The modified equation approach to the stability and accuracy analysis of finite-difference methods. *J. of Computational Physics*, 14:159–179, 1974.

[Zad66] P. Zadunaisky. A method for the estimation of the errors propagated in the numerical solution of a system of ordinary differential equations. In *Proc. Int. Astronomical Union*, volume 25, pages 281–287. Academic Press, New York, 1966.

[Zad76] P. E. Zadunaisky. On the estimation of error propagated in the numerical solution of a system of ordinary differential equations. *Num. Math.*, 27:21–39, 1976.