THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Université Toulouse III Paul Sabatier (UT3 Paul Sabatier)

Discipline ou spécialité :
Mathématiques appliquées

Présentée et soutenue par
Fabien MONFREDA
le : 12 juillet 2013

Titre :
Étude et résolution d'équations différentielles algébriques avec applications en génie des procédés

---

École doctorale :
Mathématiques, Informatique et Télécommunications de Toulouse (MITT)

Unité de recherche :
UMR 5219

Directeurs de thèse :

Jean-Claude YAKOUBSOHN	Professeur à l'université Paul Sabatier de Toulouse
François LEMAIRE	Maître de conférences à l'université de Lille

Rapporteurs :

Moulay BARKATOU	Professeur à l'université de Limoges
Pablo SOLERNÓ	Professeur à l'université de Buenos Aires

Membres du jury :

François BOULIER	Professeur à l'université de Lille
Marc GIUSTI	Directeur de recherche au CNRS - LIX école Polytechnique
Jean-Marc LE LANN	Professeur à l'ENSIACET
Bruno SALVY	Directeur de recherche à l'INRIA - ENS Lyon

À Liam

Remerciements

Je remercie très chaleureusement mes directeurs de thèse Jean-Claude Yakoubsohn et François Lemaire pour m'avoir permis de vivre cette expérience scientifique et humaine. Ces trois dernières années furent passionnantes et je n'aurais pas pu mener à bien ce projet sans leur accompagnement. Se sentir autonome dans son travail tout en pouvant compter sur leur expérience fut un privilège plus que confortable. Merci à Jean-Claude et à François pour leurs conseils, leur disponibilité et leur écoute face à mes sempiternels doutes.

Je remercie sincèrement mes rapporteurs Moulay Barkatou et Pablo Solernó pour avoir pris le temps de lire avec attention ce mémoire. La qualité de leurs rapports a apporté un appui important à mon travail ; je leur en suis très reconnaissant.

J'exprime toute ma gratitude à François Boulier, Marc Giusti, Jean-Marc Le Lann et Bruno Salvy qui ont accepté de faire partie de mon jury.

Je tiens à saluer mon complice Karim Alloula, avec lequel j'ai pu expérimenter la méthode de déflation sur des exemples concrets provenant du génie chimique. Que nous avons déflaté ensemble ! Un grand merci pour toutes ces heures passées à travailler, à échanger et à rire. Je me souviendrai longtemps de nos pérégrinations londoniennes. J'adresse un clin d'œil amical à Jean-Pierre Belaud mais aussi à Guillaume Chèze pour des mots doux dont il a seul le secret.

Comme un doctorant est également un enseignant, je salue particulièrement Patrick Hild et Guillaume avec lesquels j'ai eu le plaisir d'accompagner pendant trois ans les élèves de l'IUT Génie Chimique Génie des Procédés dans le difficile apprentissage des mathématiques.

Un immense merci à mes camarades de l'Institut de Mathématiques de Toulouse. Je pense tout d'abord aux courageux doctorants qui ont partagé mon bureau : Stan, Usain, Pépito, Marion, Minh et Rémi. J'adresse une pensée particulière à Mathieu pour nos innombrables discussions et pour notre amour commun des oiseaux. Je pense également à nos voisines plus ou moins éloignées : Élissar, Lien et Anne-Charline. Un deuxième immense merci à mes complices Delphine et Marie-Laure.

Mes pensées affectueuses se tournent vers mes amis. Jamais ils n'ont cessé de m'encourager tout au long de ces années. Ce soutien sans faille a été une chance ; j'espère sincèrement en avoir été digne.

J'embrasse tendrement mes parents.

Table des matières

Introduction

Les équations différentielles algébriques (EDAs en abrégé) sont des équations faisant intervenir un vecteur inconnu X ainsi que sa dérivée $\dot{X}$, où la notation $\dot{X}$ désigne la dérivée de X par rapport à une variable indépendante t (qui désigne le plus souvent le temps). Ces équations sont du type

$$F\left(X, \dot{X}\right) = 0, \tag{1}$$

où la fonction F est une donnée du problème. Le terme *algébrique* indique que des équations du type $0 = G(X)$ peuvent intervenir dans (1). Ces équations algébriques peuvent apparaître explicitement dans le système étudié, ou bien de manière implicite. Les équations différentielles ordinaires (EDOs en abrégé) sont des EDAs particulières, pour lesquelles il est possible de passer de la forme générale (1) à la forme

$$\dot{X} = f(X), \tag{2}$$

où f est une fonction continue. Le passage entre (1) et (2) s'effectue lorsque la différentielle de F par rapport à $\dot{X}$ est une matrice inversible. Dans cette thèse, nous nous concentrons sur le cas où la différentielle de F par rapport à $\dot{X}$ est singulière.

Dans la plupart des ouvrages consacrés aux EDAs (par exemple [10], [28] ou encore [51]), l'aptitude de ces dernières à modéliser de nombreux phénomènes physiques ou chimiques est largement soulignée. Dans cette thèse, les EDAs modélisent plus particulièrement des systèmes étudiés en génie des procédés : la *distillation de Rayleigh* et la *distillation de Rayleigh réactive*. Par ailleurs, la modélisation de phénomènes physiques ne conduit pas nécessairement à l'obtention d'EDOs. En effet, un problème physique peut être modélisé dans différents systèmes de coordonnées (par exemple, un système de coordonnées sphériques ou cartésiennes) ; ces diverses configurations peuvent engendrer des EDOs ou des EDAs, comme dans le cadre du pendule simple. En dimension d'espace $n = 2$, les équations modélisant le mouvement d'un pendule simple, exprimées dans les coordonnées cartésiennes (x, y), forment l'EDA du second ordre

$$\begin{cases} \ddot{x} = -\lambda x \\ \ddot{y} = \mathfrak{g} - \lambda y \\ 0 = x^2 + y^2 - 1, \end{cases} \tag{3}$$

où $\mathfrak{g}$ désigne l'accélération de la pesanteur $\left(\mathfrak{g} \simeq 9,81\, m.s^{-2}\right)$ et λ (qui est une fonction de t) désigne le multiplicateur de Lagrange associé à la contrainte $0 = x^2 + y^2 - 1$ (la masse du pendule ainsi que sa longueur sont ici égales à 1). En revanche, si les coordonnées polaires sont considérées, il est bien connu que le mouvement du pendule est décrit par l'EDO du second ordre

$$\ddot{\theta} = -\mathfrak{g} \sin\theta, \tag{4}$$

où θ représente l'angle entre la verticale et le fil du pendule. Dans le cadre du pendule simple, il est ainsi possible de choisir le système de coordonnées et de privilégier les coordonnées polaires

(ou sphériques d'un point de vue plus général) fournissant (4), car on dispose de méthodes numériques robustes pour traiter les EDOs.

Qu'en est-il lorsque seules les coordonnées cartésiennes sont envisageables ? La réponse ne laisse guère le choix ; on doit être en mesure d'étudier l'EDA obtenue.

Nous allons essentiellement étudier des EDAs quasi-linéaires du type

$$E(X)\dot{X} = f(X), \tag{5}$$

où la matrice $E(X)$ est singulière. Ce type d'équations couvre notamment l'ensemble des problèmes du monde de la physique modélisés par le principe de Lagrange ou de Hamilton, et en particulier les problèmes de la mécanique du solide (comme le pendule simple). Les équations modélisant les phénomènes de distillation rentrent également dans le cadre des EDAs quasi-linéaires ; leur étude constitue un des buts premiers de cette thèse.

Les phénomènes de distillation font partie intégrante du génie des procédés. Cette discipline étudie et met en œuvre au sein d'un ensemble d'appareils (réacteurs, colonnes à distiller, mélangeurs, etc.) des phénomènes physico-chimiques, dans le but de produire de manière industrielle un ensemble de produits de caractéristiques voulues. Les EDAs décrivant ces phénomènes nous ont été fournies par une équipe du laboratoire LGC de Toulouse [1], et sont extraites des travaux de K. Alloula [1] et de R. Théry Hétreux [57]. Par exemple et de façon synthétique, considérons deux espèces chimiques composant un mélange liquide placé dans une cuve sans couvercle. On note x_1 et x_2 les fractions molaires des espèces dans le liquide. On porte celui-ci à ébullition. L'évolution de la quantité de liquide U_l, du débit vapeur V, des fractions molaires des deux espèces dans la phase vapeur y_1 et y_2 ainsi que des enthalpies molaires liquide et vapeur du système h et H est décrite par l'EDA quasi-linéaire

$$\begin{cases} \dot{U}_l = -V \\ \dot{x}_1 U_l + x_1 \dot{U}_l = -V y_1 \\ \dot{x}_2 U_l + x_2 \dot{U}_l = -V y_2 \\ \dot{h} U_l + h \dot{U}_l = Q - V H \\ 0 = y_1 - K_1 x_1 \\ 0 = y_2 - K_2 x_2 \\ 0 = x_1 + x_2 - y_1 - y_2. \end{cases} \tag{6}$$

Le système (6) modélise une distillation où les espèces chimiques ne réagissent pas entre elles. Cela étant, il est possible de modéliser des phénomènes où des réactions chimiques entrent en jeu. Ces réactions peuvent entre outre s'équilibrer instantanément ou de manière plus progressive (on parle de réactions chimiques contrôlées par la cinétique). Les EDAs induites par de tels phénomènes ne sont pas gouvernées par un principe de Lagrange. Elles sont composées de bilans de matière et d'énergie, d'équations d'équilibres thermodynamiques, éventuellement d'équilibres chimiques ainsi que de contraintes liées aux processus physiques. Par ailleurs, ces équations sont habituellement résolues de manière numérique. Grâce à de nombreux échanges avec l'équipe du LGC, nous sommes parvenu à la conclusion qu'une analyse mathématique de ces équations était nécessaire pour plusieurs raisons. Il était tout d'abord important de vérifier la validité mathématique des modèles mis en jeu. Il était ensuite intéressant d'exhiber des propriétés structurelles communes à ces systèmes, afin d'en établir une classification. Il était enfin

1. Laboratoire de Génie Chimique - Institut National Polytechnique de Toulouse - École Nationale Supérieure des Ingénieurs en Arts Chimiques Et Technologiques (ENSIACET).

pertinent d'approfondir la connaissance formelle de ces systèmes afin d'améliorer leur intégration numérique selon un ou plusieurs des points de vue suivants : le calcul des conditions initiales cohérentes, la stabilité et la performance.

La résolution des EDAs se présente généralement en deux étapes : une étape symbolique suivie d'une étape numérique. L'étape symbolique a pour but de transformer l'EDA en une EDO contrainte par des équations algébriques. L'étape numérique consiste alors à intégrer une EDO sous contraintes. La littérature mathématique actuelle traitant des EDAs propose diverses méthodes de résolution, mais leur mise en œuvre dans le contexte du génie des procédés nous semblait délicate. Nous proposons pour cette raison d'étudier une méthode de résolution, dite de *déflation*, qui à partir de l'EDA initiale fournit une suite d'EDAs de tailles décroissantes et une suite d'équations algébriques, pour finir par obtenir soit une EDO sous contraintes, soit un système uniquement composé d'équations algébriques. La méthode de déflation s'inscrit ainsi dans la lignée des méthodes de réduction (puisqu'elle réduit la taille du système à étudier), et son application aux équations modélisant la distillation de Rayleigh ainsi qu'au problème du pendule généralisé a donné des résultats satisfaisants et encourageants.

Nous présentons dans cette thèse quatre résultats :

1. *Un algorithme formel de déflation appliqué aux EDAs linéaires et quasi-linéaires.*
   La méthode de déflation est un processus symbolique itératif qui, à chaque étape, découple une EDA en une EDA de taille réduite (et de même forme que la précédente) et un ensemble d'équations algébriques. Pour cela, on sépare l'EDA initiale en deux parties : une partie différentielle et une partie algébrique. La partie algébrique est dérivée, puis utilisée pour exprimer un jeu de variables en fonction d'un autre. On substitue ensuite un de ces jeux de variables dans la partie différentielle pour parvenir à l'EDA réduite. Au bout d'un nombre fini d'étapes, on obtient soit une EDO accompagnée d'un ensemble d'équations algébriques, soit uniquement un ensemble d'équations algébriques. Nous appliquons la méthode de déflation aux EDAs linéaires à coefficients constants $E\dot{X} = AX + f$ et à coefficients variables $E(t)\dot{X} = A(t)X + f$, où les matrices E et $E(t)$ sont singulières, ainsi qu'aux EDAs quasi-linéaires (5). L'étude du cas non linéaire général (1) n'est pas abordée.
2. *Une preuve concernant la baisse de l'indice d'une EDA par la méthode de déflation dans le contexte linéaire à coefficients constants.*
   La multiplicité des techniques d'étude et de résolution des EDAs est probablement due à la notion centrale d'*indice*. Le premier indice rencontré dans l'étude des EDAs est l'*indice de Kronecker* ; il est défini dans le cadre linéaire à coefficients constants et il correspond à l'indice de nilpotence de la matrice nilpotente extraite de $(\lambda E + A)^{-1} E$, pour un certain $\lambda \in \mathbb{R}$, via la décomposition de Jordan. Si la notion d'indice de Kronecker dans le contexte des EDAs linéaires à coefficients constants est un concept commun aux différentes méthodes de résolution, il n'en est rien dans le contexte des EDAs linéaires à coefficients variables. Plusieurs notions d'indice apparaissent selon le point de vue adopté par les auteurs ; les travaux de S. L. Campbell, C. W. Gear et L. R. Petzold ([10], [21]) utilisent l'indice de différentiation, la vision géométrique de P. J. Rabier et W. C. Rheinboldt ([46], [50], [51] p. 93 – 129) observant les EDAs en tant qu'EDOs sur des variétés différentielles définit l'indice géométrique, les méthodes de projections considérées par E. Griepentrog et R. März ([17], [18]) introduisent l'indice de traçabilité ou encore la généralisation de l'indice de différentiation à des problèmes sous-déterminés ou sur-déterminés par P. Kunkel et V. Mehrmann [28] passe par la définition de l'indice d'étrangeté. Ces indices, coïncidant dans le contexte des EDAs linéaires à coefficients constants, ne sont plus égaux pour des coefficients variables. On peut également citer l'indice de perturbation [21] et l'indice structurel ([40], [44]).

La méthode de déflation produit une suite d'EDAs dont les tailles sont réduites à chaque étape, de même que les rangs des matrices coefficients. On montre que chacune de ces étapes produit une EDA dont l'indice de Kronecker est diminué de un par rapport à celui de l'EDA précédente.

3. *Une preuve du caractère géométrique de la méthode de déflation par l'étude du pendule simple en dimension n.*

   Les premières études du mouvement d'un pendule simple pesant furent réalisées par Galilée [15]. Nous nous devions de traiter cet exemple fondamental de la physique, incontournable dans le monde des EDAs. En appliquant la méthode de déflation aux équations (écrites en coordonnées cartésiennes) modélisant l'évolution du pendule et en effectuant un changement de variables sphériques à postériori, nous retrouvons bien les EDOs classiques du pendule. Ceci traduit le caractère géométrique de la méthode de déflation. Le cas du pendule en dimension n est traité, ainsi qu'une généralisation de ce dernier concernant les systèmes contraints à corps multiples [47].

4. *La résolution, via la méthode de déflation, de trois modèles de distillation de Rayleigh.*

   Nous appliquons la méthode de déflation à trois EDAs quasi-linéaires modélisant la distillation de Rayleigh ([1], [57]) : une distillation sans réactions chimiques, une distillation où les réactions chimiques s'équilibrent progressivement et une distillation où ces réactions sont instantanément équilibrées. D'un point de vue général, la distillation consiste à faire chauffer un mélange, jusqu'à évaporation, afin de séparer et récupérer les différents constituants du mélange. Au cours de la distillation, les quantités mises en avant dans l'étude telles que le volume de liquide, la température ou encore les concentrations des diverses espèces chimiques varient. On est en présence d'un système dynamique. La méthode de déflation fournit en une ou deux étapes une EDO et des contraintes algébriques, permettant ainsi de déterminer aisément des conditions initiales cohérentes. Nous montrons de plus que le modèle de distillation de Rayleigh non réactive est un cas particulier de la distillation aux réactions chimiques contrôlées par la cinétique.

Cette thèse a été menée dans le cadre du projet ANR LEDA [2] qui tente d'établir le lien entre différentes approches permettant de résoudre certaines classes d'EDAs :

- l'approche de J. Ritt qui utilise l'algèbre différentielle ;
- les approches décrites par S. L. Campbell, P. Kunkel et V. Mehrmann, initiées par R. März et E. Griepentrog qui procèdent à une réduction d'indice du problème initial ;
- les approches décrites par P. J. Rabier et W. C. Rheinboldt, orientées vers la géométrie différentielle ;
- l'approche de C. G. Jacobi, reprise par J. D. Pryce, basée sur une analyse structurelle des EDAs.

Le premier chapitre de cette thèse propose une description des principales méthodes de résolution des EDAs, notamment de celles citées ci-dessus. Nous présentons la méthode de déflation dans le second chapitre, en décrivant son déroulement pour les EDAs linéaires et quasi-linéaires. La méthode est mise en pratique sur les équations modélisant le pendule simple pour les dimensions 2, 3 et n ainsi que sur des problèmes de mécanique plus généraux. Enfin, nous étudions dans le troisième chapitre des modèles de distillation de Rayleigh.

2. Logistique des Équations Différentielles Algébriques.

Chapitre 1

Résolution d'EDAs

Dans ce chapitre, nous rappelons les méthodes de résolution les plus significatives de certaines classes d'EDAs comme les EDAs linéaires, ou encore les EDAs quasi-linéaires qui sont des problèmes non linéaires structurés. L'inventaire présenté dans ce chapitre ne se veut pas exhaustif. Nous mettons en parallèle différentes approches permettant de résoudre les systèmes différentiels implicites.

Il est démontré qu'il n'existe pas de méthode générale pour résoudre l'ensemble des EDAs (on peut se rapporter à l'ouvrage de Y. V. Matiyasevich [30], p. 176). L'investigation de méthodes de résolution des EDAs a donné naissance à de multiples techniques, empruntant de nombreux chemins mathématiques. Les premiers travaux pouvant s'appliquer à la résolution des EDAs linéaires à coefficients constants sont dus à K. Weierstrass et L. Kronecker. Il s'agit de résultats d'algèbre linéaire qui étudient les propriétés de la paire matricielle (E, A) apparaissant dans l'EDA $E\dot{X} = AX + f$. Quant aux EDAs non-linéaires, on trouve dans les travaux de C. G. Jacobi les premières idées pouvant permettre d'appréhender ces systèmes différentiels implicites. Durant la seconde moitié du vingtième siècle, la recherche sur la thématique des EDAs s'est considérablement développée. Nous proposons dans ce chapitre d'explorer différentes techniques de résolution d'EDAs, en parcourant cette période jusqu'au début de la décennie actuelle. Les ouvrages de P. Kunkel et V. Mehrmann [28] et R. Riaza [51] sont parus récemment, mais ne couvrent pas l'ensemble des techniques de résolution des EDAs, notamment les aspects d'algèbre différentielle (développés par J. Ritt) et les travaux de J. Pryce [44] (initiés par C. G. Jacobi).

Dans toute la suite de l'étude, n et m désignent deux entiers naturels non nuls.

1.1 EDAs linéaires à coefficients constants

On se propose d'étudier le problème linéaire à coefficients constants

$$E\dot{X} = AX + f, \tag{1.1}$$

où $E \in \mathbb{R}^{m\times n}$, $A \in \mathbb{R}^{m\times n}$ et $f : \mathcal{I} \to \mathbb{R}^m$, avec $\mathcal{I}$ un intervalle ouvert de $\mathbb{R}$. On suppose que $X \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^n)$ et que f est une fonction suffisamment régulière (on peut supposer que $f \in \mathcal{C}^\infty(\mathcal{I}, \mathbb{R}^m)$, même si cette hypothèse sera affaiblie par la suite). On dit que X est une solution de (1.1) si $E\dot{X}(t) = AX(t) + f(t)$ pour tout $t \in \mathcal{I}$. La configuration (1.1) rassemble les problèmes sous-déterminés ($n > m$), les problèmes sur-déterminés ($m > n$) ainsi que les problèmes carrés ($n = m$).

1.1.1 Structure des solutions des EDAs linéaires

Nous débutons l'analyse du problème (1.1) en remarquant que la différence entre deux solutions de (1.1) est une solution du problème homogène associé

$$E\dot{X} = AX. \tag{1.2}$$

Proposition 1

Soient $X_1 \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^n)$ et $X_2 \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^n)$ deux solutions du problème (1.1). La fonction $Y \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^n)$ définie par $Y(t) = X_1(t) - X_2(t)$ pour tout $t \in \mathcal{I}$ est solution du problème homogène (1.2).

Preuve - Il suffit d'écrire l'EDA satisfaite par Y. Pour tout $t \in \mathcal{I}$, on a

$$\begin{aligned} E\dot{Y}(t) &= E\left(\dot{X}_1(t) - \dot{X}_2(t)\right) \\ &= E\dot{X}_1(t) - E\dot{X}_2(t) \\ &= AX_1(t) + f(t) - AX_2(t) - f(t) \\ &= A\left(X_1(t) - X_2(t)\right) \\ &= AY(t). \end{aligned}$$

Ainsi, Y satisfait l'équation homogène (1.2). □

Le résultat précédent est également valable pour des matrices dépendant du temps $E(t)$ et $A(t)$. Du point de vue de la structure des solutions, on note ici le parallèle entre les EDAs linéaires et les EDOs linéaires ; la solution générale est obtenue en sommant les solutions du problème homogène (1.2) et une solution particulière du problème non-homogène (1.1).

On ne cherche pas par la suite à expliciter les expressions des solutions de (1.1). Dans le contexte carré, si la matrice E est inversible, l'EDA (1.1) devient l'EDO $\dot{X} = E^{-1}AX + E^{-1}f$. On détermine les expressions des solutions de cette EDO grâce aux formules habituelles (disponibles par exemple dans E. Hairer, S. P. Hørsett et G. Wanner [20]). En revanche, si la matrice E est singulière, divers travaux portant sur les inverses généralisés de E ont été effectués, en utilisant l'inverse de *Drazin* (S.L. Campbell, C.D. Meyer et N.J. Rose, [13]) ou encore l'inverse de *Moore-Penrose* (S.L. Campbell [11], P. Kunkel et V.L. Mehrmann [27]). L'ouvrage de A. Ben-Israel et T.N.E. Greville [5] aborde en détail la thématique des inverses généralisés.

1.1.2 Solvabilité des EDAs linéaires à coefficients constants

On considère une solution x du problème (1.1). On remarque aisément que la fonction $y = e^{\lambda t}x$ est solution de l'équation

$$E\dot{X} = (\lambda E + A)X + e^{\lambda t}f. \tag{1.3}$$

En effet,

$$\begin{aligned} E\dot{y} &= E\left(\lambda e^{\lambda t}x + e^{\lambda t}\dot{x}\right) \\ &= \lambda e^{\lambda t}Ex + e^{\lambda t}E\dot{x} \\ &= \lambda e^{\lambda t}Ex + e^{\lambda t}\left(Ax + f\right) \\ &= (\lambda E + A)e^{\lambda t}x + e^{\lambda t}f \\ &= (\lambda E + A)y + e^{\lambda t}f, \end{aligned}$$

ce qui montre bien que y satisfait (1.3). Travailler avec l'EDA (1.1) revient ainsi à travailler avec l'EDA (1.3). Dans le cas particulier où $\lambda = 0$, les équations (1.1) et (1.3) sont identiques. Discutons à présent des différentes hypothèses envisageables sur la matrice $E_A := \lambda E + A \in \mathbb{R}^{m\times n}$ afin d'étudier la solvabilité de (1.3).

1.1.2.1 Premier cas

Théorème 1

Si rang $E_A < n$ pour tout $\lambda \in \mathbb{R}$ et s'il existe une solution particulière x_p de (1.1), alors pour tout $v \in \ker E_A$, la fonction $\bar{x} = e^{-\lambda t}v + x_p$ est solution de l'EDA (1.1). En particulier, le problème homogène (1.2) a une solution non triviale.

Preuve - Soit v un vecteur non nul de $\mathbb{R}^n$ appartenant à $\ker E_A$. Posons $\bar{y} = e^{-\lambda t}v$. Puisque $v \in \ker E_A$, on a $\lambda Ev = -Av$. Ainsi, on a clairement

$$\begin{aligned} E\dot{\bar{y}} &= E\left(-\lambda e^{-\lambda t}v\right) \\ &= e^{-\lambda t}(-\lambda Ev) \\ &= e^{-\lambda t}(Av) \\ &= A\bar{y}. \end{aligned}$$

On en déduit que le problème homogène (1.2) admet une solution non triviale. Soit x_p une solution particulière de (1.1). On obtient alors le résultat escompté.

$$\begin{aligned} E\left(\dot{\bar{y}} + \dot{x}_p\right) &= A\bar{y} + Ax_p + f \\ &= A(\bar{y} + x_p) + f. \end{aligned}$$

□

1.1.2.2 Deuxième cas : le cas régulier

Théorème 2

S'il existe un scalaire $\lambda \in \mathbb{R}$ tel que rang $E_A = n = m$, alors l'EDA (1.1) possède une solution générale si la non-homogénéité f est suffisamment régulière.

Preuve - La matrice E_A est ici inversible. Dans ce contexte, un résultat fondamental de F. R. Gantmacher ([16], p. 28) assure l'existence de deux matrices inversibles $P \in \mathbb{R}^{n\times n}$ et $Q \in \mathbb{R}^{n\times n}$ telles que (1.1) est équivalent au problème

$$\begin{bmatrix} I_{a_1} & 0 \\ 0 & \mathcal{N} \end{bmatrix} Q^{-1}\dot{X} = \begin{bmatrix} \mathcal{J} & 0 \\ 0 & I_{n-a_1} \end{bmatrix} Q^{-1}X + P\bar{f}, \tag{1.4}$$

où $I_{a_1} \in \mathbb{R}^{a_1\times a_1}$ avec $a_1 < n$, $\mathcal{N} \in \mathbb{R}^{(n-a_1)\times(n-a_1)}$, $\mathcal{J} \in \mathbb{R}^{a_1\times a_1}$, $I_{n-a_1} \in \mathbb{R}^{(n-a_1)\times(n-a_1)}$ et $\bar{f} = E_A^{-1}f$. La matrice $\mathcal{N}$ est une matrice nilpotente d'indice $\nu_K \leqslant n - a_1$. Les matrices $\mathcal{J}$ et $\mathcal{N}$ sont de plus sous forme réduite de Jordan. En effet, en partant de la paire matricielle (E, A), on obtient

$$(E, A) = (E, \lambda E + A - \lambda E) = (E, E_A - \lambda E)\,.$$

On note $\ll \sim \gg$ [1] lorsque l'on passe d'une paire de matrices à une autre en multipliant toutes les matrices par des matrices inversibles, i.e. $(E, A) \sim (E_1, A_1)$ s'il existe deux matrices de passage

1. On montre aisément qu'il s'agit d'une relation d'équivalence entre paires matricielles.

Q_1 et Q_2 telles que $E = Q_1 E_1 Q_2$ et $A = Q_1 A_1 Q_2$. Ainsi

$$(E, A) \sim \left(E_A^{-1} E, I_n - \lambda E_A^{-1} E\right).$$

En décomposant la matrice $E_A^{-1}E$ grâce à la forme réduite de Jordan et en simplifiant les expressions, on parvient à montrer que

$$(E, A) \sim \left(\begin{bmatrix} I_{a_1} & 0 \\ 0 & \mathcal{N} \end{bmatrix}, \begin{bmatrix} \mathcal{J} & 0 \\ 0 & I_{n-a_1} \end{bmatrix}\right),$$

ce qui permet de passer de l'EDA (1.1) à l'EDA (1.4).

Le problème (1.4) se scinde en deux équations en posant $Q^{-1}X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ et $P\bar{f} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$:

$$\begin{cases} \dot{X}_1 = \mathcal{J} X_1 + f_1 & \text{(1.5a)} \\ \mathcal{N}\dot{X}_2 = X_2 + f_2. & \text{(1.5b)} \end{cases}$$

L'équation (1.5a) est une EDO. La formule de Duhamel fournit l'expression de la solution de cette équation différentielle, couplée à la condition initiale $X_1(t_0) = \bar{X}_1$:

$$X_1(t) = e^{\mathcal{J}(t-t_0)}\bar{X}_1 + \int_{t_0}^{t} e^{\mathcal{J}(t-s)} f_1(s) ds. \tag{1.6}$$

L'équation (1.5b) est quant à elle une EDA, mais sa forme spécifique permet de la résoudre. Cette équation peut se réécrire comme $(\mathcal{N}D - I_{n-a_1})\, X_2 = f_2$, où D correspond à l'opérateur de dérivation par rapport à t. Il ne reste plus qu'à inverser $\mathcal{N}D - I_{n-a_1}$:

$$(\mathcal{N}D - I_{n-a_1})^{-1} = -\sum_{i=0}^{\infty} (\mathcal{N}D)^i = -\sum_{i=0}^{\nu_K - 1} (\mathcal{N}D)^i = -\sum_{i=0}^{\nu_K - 1} \mathcal{N}^i \frac{d^i}{dt^i}.$$

Ainsi, on obtient :

$$X_2(t) = -\sum_{i=0}^{\nu_K - 1} \mathcal{N}^i f_2^{(i)}(t). \tag{1.7}$$

On remarque ici que f_2 doit être dérivable ν_K fois pour pouvoir satisfaire l'équation (1.5b). □

Définition 1 (*Paire matricielle régulière*)

Une paire de matrices carrées (E, A) est dite régulière *s'il existe un scalaire $\lambda \in \mathbb{R}$ tel que la matrice E_A est inversible. Par extension, le problème (1.1) est alors qualifié de régulier.*

Le théorème 2 s'inscrit dans le contexte de la régularité définie ci-dessus. Dans ce cadre précis, il suffit de supposer que f est une fonction de classe $\mathcal{C}^{\nu_K}(\mathcal{I}, \mathbb{R}^n)$ (et non plus de classe $\mathcal{C}^{\infty}(\mathcal{I}, \mathbb{R}^n)$). L'indice ν_K, appelé *indice de Kronecker*, a donc une influence déterminante au niveau de la résolution du problème régulier (1.1).

Remarque 1

Cet indice, indépendant du scalaire λ, est caractéristique de la paire matricielle (E, A) ; si $(E, A) \sim (E', A')$, alors les deux paires matricielles ont le même indice de Kronecker. La forme du système (1.4) est par ailleurs qualifiée de forme canonique de Kronecker [28].

Dans la théorie des EDOs, l'existence et l'unicité des solutions sont assurées par la résolution du problème de Cauchy couplant l'équation différentielle et la condition initiale [20]. Pour chaque condition initiale, il y a unicité d'une solution maximale. À la différence des EDOs, les conditions initiales d'une EDA ne sont pas libres car elles doivent satisfaire les contraintes algébriques du problème.

Exemple - Considérons l'EDA suivante :

$$\begin{cases} \dot{x}_1 = x_1 + 2x_2 \\ 0 = x_1 - x_2. \end{cases} \tag{1.8}$$

La forme matricielle de cette EDA est $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. La paire de matrices $\left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix} \right)$ est ici régulière. Les solutions de (1.8) sont :

$$\begin{cases} x_1(t) = x_{1,0} e^{3(t-t_0)} \\ x_2(t) = x_1(t). \end{cases}$$

La condition initiale $x_1(t_0) = x_{1,0}$ est libre tandis que la condition initiale $x_2(t_0) = x_{2,0}$ ne l'est pas. En d'autres termes, seules les conditions initiales de la forme $\begin{pmatrix} x_{1,0} \\ x_{1,0} \end{pmatrix}$ font sens. Par exemple, le problème (1.8) couplé de la condition initiale $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ n'admet aucune solution.

L'exemple précédent illustre le caractère non suffisant de la condition de régularité de la paire de matrices quant à l'existence et l'unicité des solutions du problème (1.1). En revanche, en considérant la fonction f de classe $\mathcal{C}^{\nu_K}(\mathcal{I}, \mathbb{R}^n)$, la régularité de la paire matricielle ainsi que la consistance des conditions initiales (ces dernières doivent satisfaire (1.7)) fournissent des conditions nécessaires et suffisantes pour obtenir l'existence et l'unicité des solutions de (1.1).

1.1.2.3 Troisième cas

Supposons enfin qu'il existe un scalaire $\lambda \in \mathbb{R}$ tel que rang $E_A = n < m$. Il existe alors une matrice de permutation P telle que l'EDA (1.3) s'écrive

$$\begin{bmatrix} E_1 \\ E_2 \end{bmatrix} \dot{X} = \begin{bmatrix} E_{A1} \\ E_{A2} \end{bmatrix} X + \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}, \tag{1.9}$$

où $\begin{bmatrix} E_1 \\ E_2 \end{bmatrix} = PE$, $\begin{bmatrix} E_{A1} \\ E_{A2} \end{bmatrix} = PE_A$ avec $E_{A1} \in \mathbb{R}^{n \times n}$ inversible et $\begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = Pe^{\lambda t} f$. Le problème (1.9) s'écrit également

$$\begin{cases} E_1 \dot{X} = E_{A1} X + f_1 & \text{(1.10a)} \\ E_2 \dot{X} = E_{A2} X + f_2, & \text{(1.10b)} \end{cases}$$

où (1.10a) est une EDA régulière puisque la matrice E_{A1} est inversible (il suffit de prendre $\lambda = 0$ pour que $\lambda E_1 + E_{A1}$ soit inversible). Le contexte de régularité indique que cette EDA possède une solution générale (si f_1 est suffisamment régulière). Il faut cependant tenir compte de la deuxième équation (1.10b). Le problème (1.1) possède une solution générale si la solution de l'EDA (1.10a) vérifie l'EDA (1.10b). Dans le cas contraire, l'EDA (1.1) ne possède aucune solution. On résume ce dernier cas dans le résultat suivant :

Théorème 3

S'il existe un scalaire $\lambda \in \mathbb{R}$ tel que rang $E_A = n < m$, alors l'EDA (1.1) possède une solution générale si la solution de l'EDA (1.10a) satisfait l'EDA (1.10b). Dans le cas contraire, l'EDA (1.1) n'admet aucune solution.

1.1.3 Un exemple de problème régulier

Observons un exemple d'EDAs provenant de la théorie des circuits électriques ([28], p. 8).

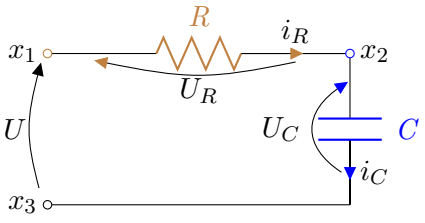


FIGURE 1.1 – Circuit RC

Modélisation - On souhaite modéliser la charge d'un condensateur de capacité constante $C \neq 0$ au travers d'un conducteur ohmique de résistance constante $R \neq 0$ (figure (1.1)). Durant cette phase de charge, la tension source U appliquée au circuit est également constante. Cette tension correspond à la différence entre les potentiels x_1 et x_3 ; on obtient donc la première équation :

$$U = x_1 - x_3.$$

En appliquant la loi des nœuds (première loi de Kirchhoff), on obtient $i_R = i_C$. La loi d'Ohm fournit la valeur de i_R : $U_R = Ri_R$ i.e. $i_R = R^{-1}(x_1 - x_2)$. Quant à i_C, on utilise la relation caractéristique d'un condensateur, à savoir $q_C = CU_C = C(x_2 - x_3)$. Or $i_C = \dot{q}_C$; on obtient ainsi $i_C = C(\dot{x}_2 - \dot{x}_3)$. La deuxième équation est :

$$C(\dot{x}_2 - \dot{x}_3) = R^{-1}(x_1 - x_2).$$

Enfin, comme U est une différence de potentiels et donc mesure un écart, on peut imposer

$$0 = x_3.$$

On parvient ainsi au système

$$\begin{cases} C(\dot{x}_2 - \dot{x}_3) = R^{-1}(x_1 - x_2) \\ 0 = x_1 - x_3 - U \\ 0 = x_3. \end{cases} \tag{1.11}$$

Résolution - On commence par écrire le système sous la forme matricielle. On considère un système de la forme (1.1) en posant

$$E = \begin{bmatrix} 0 & C & -C \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} R^{-1} & -R^{-1} & 0 \\ 1 & 0 & -1 \\ 0 & 0 & 1 \end{bmatrix}, \quad X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad \text{et} \quad f = \begin{pmatrix} 0 \\ -U \\ 0 \end{pmatrix}.$$

Puisque la matrice A est inversible ($\det A = R^{-1}$), pré-multiplions l'EDA (1.1) par A^{-1} ; on obtient

$$A^{-1}E\dot{X} = X + A^{-1}f.$$

La décomposition de Jordan de la matrice $A^{-1}E$ assure l'existence d'une matrice inversible Q ainsi que des matrices J et N telles que l'équation précédente devienne

$$\begin{bmatrix} J & 0 \\ 0 & N \end{bmatrix} Q^{-1}\dot{X} = Q^{-1}X + Q^{-1}A^{-1}f,$$

où la matrice J est une matrice inversible sous forme réduite de Jordan et N est une matrice nilpotente, également écrite sous forme réduite de Jordan. De simples calculs permettent de déterminer les expressions des matrices Q, J et N :

$$Q = \begin{bmatrix} 0 & 1 & -1 \\ -1 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad J = -(RC)^{-1} \quad \text{et} \quad N = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

En posant $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = Q^{-1}X$ et $\begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = Q^{-1}A^{-1}f$, le problème s'écrit

$$\begin{cases} -RC\dot{X}_1 = X_1 + f_1 \\ 0 = X_2 + f_2, \end{cases}$$

avec

$$X_1 = x_3 - x_2, \quad X_2 = \begin{pmatrix} x_3 \\ x_3 - x_1 \end{pmatrix}, \quad f_1 = U \quad \text{et} \quad f_2 = \begin{pmatrix} 0 \\ U \end{pmatrix}.$$

Autrement dit,

$$\begin{cases} \dot{X}_1 = -(RC)^{-1}X_1 - (RC)^{-1}U \\ X_2 = -\begin{pmatrix} 0 \\ U \end{pmatrix}. \end{cases}$$

On trouve alors

$$\begin{cases} X_1(t) = \tilde{K}e^{-t(RC)^{-1}} - U, \quad \tilde{K} \in \mathbb{R} \\ X_2(t) = -\begin{pmatrix} 0 \\ U \end{pmatrix}. \end{cases}$$

Les solutions du système (1.11) sont de la forme

$$\begin{cases} x_1(t) = U \\ x_2(t) = Ke^{-t(RC)^{-1}} + U, \quad K \in \mathbb{R} \\ x_3(t) = 0. \end{cases}$$

□

1.2 Méthode des systèmes augmentés

La complexité des EDAs réside entre autres dans l'existence des contraintes algébriques, dont certaines sont par ailleurs implicites (on parle de contraintes cachées). On peut les obtenir par des manipulations algébriques comme dans le paragraphe précédent, mais également par *différentiation*. Les travaux portant sur les systèmes augmentés sont axés sur cette dernière idée, qui consiste à dériver soit les contraintes algébriques seules, soit l'EDA dans son ensemble, le but étant d'exprimer $\dot{X}$ comme une fonction continue de X. Ce nouveau système est alors étudié et permet de résoudre l'EDA initiale.

1.2.1 EDAs linéaires à coefficients constants

On considère le problème régulier (1.1). En écrivant $E = G\begin{bmatrix} E_1 & E_2 \\ 0 & 0 \end{bmatrix}$ (E est une matrice de rang r), $A = G\begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}$ et $f = G\begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$, où $G \in \mathbb{R}^{n\times n}$ est une matrice inversible, E_1 et $A_1 \in \mathbb{R}^{r\times r}$, E_2 et $A_2 \in \mathbb{R}^{r\times(n-r)}$, $A_3 \in \mathbb{R}^{(n-r)\times r}$, $A_4 \in \mathbb{R}^{(n-r)\times(n-r)}$, $f_1 \in \mathbb{R}^r$ et $f_2 \in \mathbb{R}^{n-r}$, le système (1.1) devient

$$\begin{bmatrix} E_1 & E_2 \\ 0 & 0 \end{bmatrix} \dot{X} = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} X + \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}. \tag{1.12}$$

La matrice $\begin{bmatrix} E_1 & E_2 \end{bmatrix}$ est de rang plein par construction. La contrainte algébrique exhibée est ainsi $0 = A_3X_1 + A_4X_2 + f_2$. En dérivant cette contrainte puis en injectant l'expression obtenue dans (1.12), on parvient au nouveau système

$$\begin{bmatrix} E_1 & E_2 \\ A_3 & A_4 \end{bmatrix} \dot{X} = \begin{bmatrix} A_1 & A_2 \\ 0 & 0 \end{bmatrix} X + \begin{pmatrix} f_1 \\ -\dot{f}_2 \end{pmatrix}. \tag{1.13}$$

On réitère ce processus jusqu'à ce que la matrice coefficient de $\dot{X}$ dans (1.13) soit inversible (K. E. Brenan, S. L. Campbell et L. R. Petzold [10], p. 20).

Théorème 4

On considère le problème régulier (1.1) écrit sous la forme (1.4). Soit ν_K l'indice de nilpotence de la matrice $\mathcal{N}$ exhibée dans (1.4). La méthode précédemment décrite s'achève alors en ν_K étapes. Plus précisément, chaque étape du processus de résolution fait diminuer l'indice de 1.

Preuve - Sans perte de généralité, on peut supposer que le système régulier (1.1) est écrit sous la forme (1.4). En considérant sa forme développée (1.5), on se concentre sur l'EDA (1.5b) dont le coefficient $\mathcal{N}$ est une matrice nilpotente :

$$\mathcal{N} = \begin{bmatrix} \mathcal{N}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathcal{N}_p \end{bmatrix},$$

où chaque bloc $\mathcal{N}_k$, $k \in [\![1, p]\!]$ est écrit sous forme réduite. On obtient de ce fait p EDAs ayant la même forme :

$$\mathcal{N}_k \dot{X}_{2,k} = X_{2,k} + f_{2,k}.$$

Il suffit ainsi d'appliquer le processus de résolution à un problème de la forme $\mathcal{M}\dot{Y} = Y + h$, où $\mathcal{M} \in \mathbb{R}^{m\times m}$ est une matrice nilpotente sous forme réduite. On désigne par h_i la $i^{\text{ième}}$ coordonnée du vecteur h, où $i \in [\![1, m]\!]$. On dérive les contraintes algébriques et on réécrit ces équations dérivées dans le système ; on a

$$\begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 1 \end{bmatrix} \dot{Y} = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{bmatrix} Y + \begin{pmatrix} h_1 \\ \vdots \\ h_{m-1} \\ -\dot{h}_m \end{pmatrix}.$$

Une première étape est achevée. En appliquant (par exemple) la décomposition LU sur le coefficient de $\dot{Y}$, on obtient

$$\mathcal{M}\dot{Y} = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ & & -1 \quad 0 \end{bmatrix} Y + \begin{pmatrix} h_1 \\ \vdots \\ h_{m-1} \\ -h_{m-1} - \dot{h}_m \end{pmatrix}.$$

En reportant les contraintes dérivées dans $\mathcal{M}$, on trouve

$$\begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & -1 & 0 \end{bmatrix} \dot{Y} = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{bmatrix} Y + \begin{pmatrix} h_1 \\ \vdots \\ h_{m-1} \\ \dot{h}_{m-1} + \ddot{h}_m \end{pmatrix}.$$

Une deuxième étape est achevée. On observe ainsi qu'au fil des itérations de la méthode, le 1 présent dans la dernière ligne de la matrice coefficient se déplace vers la gauche en alternant de signe. Au bout de m étapes, on a

$$\begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ (-1)^{m+1} & & & 0 \end{bmatrix} \dot{Y} = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{bmatrix} Y + \begin{pmatrix} h_1 \\ \vdots \\ h_{m-1} \\ (-1)^m \sum_{i=1}^{m} h_i^{(i)} \end{pmatrix}.$$

On arrive donc à une matrice inversible au bout de m étapes. Or, puisque la matrice $\mathcal{M}$ est sous forme réduite, son indice est égal à sa taille, c'est-à-dire m. En appliquant ceci aux p EDAs précédemment obtenues, on retrouve que l'indice de (1.1) est égal à la plus grande dimension ν_K des matrices $\mathcal{N}_k$, pour $k \in [\![1, p]\!]$, c'est-à-dire à l'indice de nilpotence de la matrice $\mathcal{N}$. Quant à la diminution de l'indice d'une unité à chaque étape, elle découle clairement de la structure réduite de la matrice nilpotente. □

À la fin de ce type de processus, on obtient une EDO qui possède davantage de solutions que (1.1) ([10], p. 20). Il faut donc faire un tri grâce aux contraintes algébriques pour obtenir les solutions du problème de départ. D'une manière générale, on ajoute des solutions lorsqu'on dérive une équation. Par exemple et sans considérer de condition initiale, l'équation $X = 0$ n'admet que la solution nulle tandis que les solutions de l'équation $\dot{X} = 0$ sont les constantes.

Exemple - Reprenons l'exemple (1.11) de la section précédente. Le système est directement sous la forme (1.12) avec $E_1 = 0$, $E_2 = \begin{pmatrix} C & -C \end{pmatrix}$, $A_1 = R^{-1}$, $A_2 = \begin{pmatrix} -R^{-1} & 0 \end{pmatrix}$, $A_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $A_4 = \begin{bmatrix} 0 & -1 \\ 0 & 1 \end{bmatrix}$, $f_1 = 0$ et $f_2 = \begin{pmatrix} -U \\ 0 \end{pmatrix}$. Puisque $\begin{vmatrix} E_1 & E_2 \\ A_3 & A_4 \end{vmatrix} = -C \neq 0$, la méthode s'arrête au bout de la première étape et montre ainsi que le problème (1.11) possède un indice égal à 1. L'EDO donnant X est obtenue :

$$\dot{X} = \begin{bmatrix} 0 & C & -C \\ 1 & 0 & -1 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} R^{-1} & -R^{-1} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} X. \tag{1.14}$$

Les solutions du problème (1.14) sont les fonctions de la forme $X(t) = \begin{pmatrix} \alpha \\ Ke^{-t(RC)^{-1}} + \alpha \\ \beta \end{pmatrix}$, avec $(K, \alpha, \beta) \in \mathbb{R}^3$. On retrouve les solutions du problème original en prenant $\alpha = U$ et $\beta = 0$. □

Indice de différentiation. Le nombre d'itérations de la méthode relative au théorème 4 est appelé *indice de différentiation*, noté ν_D, car il correspond au nombre de différentiations effectuées sur les contraintes. Plus précisément, il indique le nombre de dérivations nécessaires pour pouvoir exprimer de manière continue l'inconnue $\dot{X}$ en fonction de X et t. Il coïncide évidemment ici avec l'indice de Kronecker. À la différence de ce dernier, il fait également sens dans le contexte linéaire à coefficients variables. Nous donnons dans la section suivante une définition plus formelle de cet indice.

1.2.2 EDAs linéaires à coefficients variables

La méthode des systèmes augmentés prend une forme plus générale dans le cadre linéaire à coefficients variables. On travaille maintenant sur des équations de la forme

$$E(t)\dot{X} = A(t)X + f, \tag{1.15}$$

où $E : \mathcal{I} \to \mathbb{R}^{n\times n}$ est une matrice singulière de *rang constant sur* $\mathcal{I} \subset \mathbb{R}$, $A : \mathcal{I} \to \mathbb{R}^{n\times n}$ et $f : \mathcal{I} \to \mathbb{R}^n$. On suppose que les fonctions E, A et f sont suffisamment régulières. Cette fois-ci, c'est l'ensemble du système (1.15) qui est dérivé autant de fois que nécessaire [12]. Grâce à ces EDAs dérivées en chaîne, on construit le système augmenté suivant :

$$E_l(t)\dot{X}_l = A_l(t)X_l + f_l, \tag{1.16}$$

où

$$E_l(t) = \begin{bmatrix} E(t) & 0 & \cdots & 0 \\ \dot{E}(t) - A(t) & E(t) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ E^{(l)}(t) - lA^{(l-1)}(t) & lE^{(l-1)}(t) - \frac{l(l-1)}{2}A^{(l-2)}(t) & \cdots & E(t) \end{bmatrix},$$

$$A_l(t) = \begin{bmatrix} A(t) & 0 & \cdots & 0 \\ \dot{A}(t) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A^{(l)}(t) & 0 & \cdots & 0 \end{bmatrix}, \quad X_l = \begin{pmatrix} X \\ \dot{X} \\ \vdots \\ X^{(l)} \end{pmatrix} \quad \text{et} \quad f_l = \begin{pmatrix} f \\ \dot{f} \\ \vdots \\ f^{(l)} \end{pmatrix}.$$

On s'intéresse plus particulièrement à la matrice $E_l(t)$. On cherche à factoriser cette matrice dans le but d'extraire une EDO portant sur X. Le concept de matrice *smoothly 1-full* est introduit par S. L. Campbell dans cette optique.

Définition 2 (*Matrice smoothly 1-full*)

Soit $M_l(t) : \mathcal{I} \to \mathbb{R}^{ln\times ln}$ une matrice par blocs. $M_l(t)$ est dite smoothly 1-full *s'il existe une matrice $R(t) : \mathcal{I} \to \mathbb{R}^{ln\times ln}$ inversible sur $\mathcal{I}$ ainsi qu'une matrice $H(t) : \mathcal{I} \to \mathbb{R}^{(l-1)n\times(l-1)n}$ telles que*

$$R(t)M_l(t) = \begin{bmatrix} I_n & 0 \\ 0 & H(t) \end{bmatrix}, \quad \forall t \in \mathcal{I}.$$

Définition 3 (*Indice de différentiation*)

Le plus petit entier ν_D pour lequel la matrice $E_{\nu_D}(t)$ issue de (1.16) est smoothly 1-full et de rang constant est appelé indice de différentiation *du problème (1.15).*

La matrice augmentée $E_{\nu_D}(t)$ est toujours singulière par construction. En conséquence, il n'est pas possible de déterminer $\dot{X}_{\nu_D}$ en fonction de X_{ν_D} et t. En revanche, si $E_{\nu_D}(t)$ est smoothly

1-full et de rang constant, elle contient suffisamment d'informations pour déterminer de manière unique $\dot{X}$ en fonction de X et t ([28], p. 97). L'équation (1.16) devient

$$R(t)E_{\nu_D}(t)\dot{X}_{\nu_D} = R(t)A_{\nu_D}(t)X_{\nu_D} + R(t)f_{\nu_D} \Leftrightarrow \begin{bmatrix} I_n & 0 \\ 0 & H(t) \end{bmatrix} \dot{X}_{\nu_D} = R(t)A_{\nu_D}(t)X_{\nu_D} + R(t)f_{\nu_D}.$$

Seules les n premières lignes sont importantes. Par un simple développement de la précédente expression, on obtient une EDO de la forme

$$\dot{X} = \bar{A}(t)X + \bar{f}. \tag{1.17}$$

Remarque 2

> *L'équation différentielle* (1.17) *est dite sous-jacente, relativement au problème* (1.15). *Comme dans le cas linéaire à coefficients constants, l'ensemble des solutions de* (1.15) *est inclus dans l'ensemble des solutions de* (1.17) *puisque une autre partie différentielle est ajoutée au problème. On retrouve les solutions du problème initial en tenant compte des équations algébriques.*

Solvabilité du problème (1.15). L'équivalence entre la régularité de la paire matricielle (E, A) et la solvabilité du problème (1.15) est perdue par rapport au cadre linéaire à coefficients constants. Les notions deviennent même indépendantes. Les raisons de cette difficulté seront détaillées dans la section suivante. En tout état de cause, il est nécessaire de définir une caractérisation convenable de la solvabilité du problème (1.15). Le théorème 2.4.7 ([10], p. 30) apporte cette caractérisation, basée entre autres sur le critère 1-full, ainsi que sur la constance du rang de $E(t)$.

1.2.3 EDAs non linéaires

Dans le contexte non linéaire, la méthode précédemment décrite est également appliquée ([10], p. 32). L'étude est ici portée sur des problèmes de la forme (1). D'un point de vue général, l'absence de forme spécifique sur (1) rend la résolution symbolique et numérique difficile. C'est pourquoi on se concentre sur des problèmes dits *structurés* comme les EDAs *semi-explicites* ([10], p. 36) ou encore les EADs sous forme *d'Hessenberg* ([10], p. 34).

Exemple (EDAs semi-explicites) - Considérons le système suivant :

$$\begin{cases} \dot{x} = a(x, y) \\ 0 = b(x, y). \end{cases}$$

Au lieu de construire un système augmenté, tirons avantage de la structure particulière du problème précédent. Par dérivation de la contrainte algébrique, on obtient

$$0 = \frac{\partial b(x, y)}{\partial x}\dot{x} + \frac{\partial b(x, y)}{\partial y}\dot{y}.$$

En substituant $\dot{x}$ par $a(x, y)$ dans l'expression précédente et en supposant l'inversibilité de $\dfrac{\partial b(x, y)}{\partial y}$, on parvient à l'EDO

$$\dot{y} = -\frac{\partial b(x, y)}{\partial y}^{-1} \frac{\partial b(x, y)}{\partial x} a(x, y).$$

Par résolution du système différentiel

$$\begin{cases} \dot{x} = a(x,y) \\ \dot{y} = -\dfrac{\partial b(x,y)}{\partial y}^{-1} \dfrac{\partial b(x,y)}{\partial x} a(x,y), \end{cases}$$

on aboutit aux solutions x et y, en prenant garde à ce que ces dernières satisfassent l'équation algébrique $0 = b(x,y)$. Obtenu après une dérivation des contraintes algébriques, ce problème possède un indice de différentiation égal à 1.

Exemple (EDAs sous forme d'Hessenberg) - Observons le cas autonome d'un système sous forme d'Hessenberg d'indice 2, qui s'écrit de la manière suivante

$$\begin{cases} \dot{x} = a(x,y) \\ 0 = b(x). \end{cases}$$

Contrairement à l'exemple précédent, l'équation algébrique ne dépend plus de l'inconnue y. On remarque rapidement qu'une seule dérivation de cette contrainte ne permettra pas de faire apparaître la dérivée de y. En effet, on obtient par dérivation

$$0 = \dot{b}(x)\dot{x}.$$

En revanche, en remplaçant $\dot{x}$ par $a(x,y)$ et en dérivant l'expression obtenue, on a

$$0 = \left(\ddot{b}(x)a(x,y) + \dot{b}(x)\frac{\partial a(x,y)}{\partial x}\right)\dot{x} + \dot{b}(x)\frac{\partial a(x,y)}{\partial y}\dot{y}.$$

Sous réserve de l'existence de $\left(\dot{b}(x)\dfrac{\partial a(x,y)}{\partial y}\right)^{-1}$, on parvient à

$$\begin{cases} \dot{x} = a(x,y) \\ \dot{y} = -\left(\dot{b}(x)\dfrac{\partial a(x,y)}{\partial y}\right)^{-1}\left(\ddot{b}(x)a(x,y) + \dot{b}(x)\dfrac{\partial a(x,y)}{\partial x}\right)\dot{x}. \end{cases}$$

Les expressions de x et y peuvent être établies après deux dérivations ; l'indice de différentiation est par conséquent égal à 2. Notons que la contrainte algébrique $0 = b(x)$ doit être satisfaite à la fin de cette résolution.

1.3 Méthode de *Kunkel-Mehrmann*

1.3.1 Motivation

On propose dans cette section une généralisation de la notion de formes canoniques aux EDAs linéaires à coefficients variables, éventuellement sous ou sur-déterminées. La forme canonique de Kronecker est un outil puissant mais limité aux EDAs linéaires à coefficients constants. Le but des formes canoniques est de transformer l'EDA afin de faciliter sa résolution tout en conservant ses quantités caractéristiques. Dans le cas des coefficients constants, le rang de la matrice E (qui est un point fondamental de par la nature même d'une EDA) ainsi que l'indice de Kronecker sont des *invariants*, c'est-à-dire qu'ils sont préservés par la forme canonique de Kronecker ([28], p. 18). Pour pouvoir appliquer cette forme canonique, la régularité de la paire matricielle (E, A) est nécessaire. Or, dans le contexte des coefficients variables, les notions de

régularité de la paire matricielle $(E(t), A(t))$ et de solvabilité du problème ne sont pas liées ([28], p. 56).

Exemple (Paire matricielle régulière & infinité de solutions) - Soit le problème de la forme (1.15), avec

$$E(t) = \begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix}, \quad A(t) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \quad \text{et} \quad f = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{1.18}$$

où $t \in \mathbb{R}$. La paire matricielle est régulière puisque $\det(\lambda E(t) + A(t)) = 1$, mais le problème possède une infinité de solutions du type $X(t) = \alpha(t) \begin{pmatrix} t \\ 1 \end{pmatrix}$, pourvu que la fonction α soit continûment dérivable sur $\mathbb{R}$ et satisfasse $\alpha(t_0) = 0$. □

Exemple (Paire matricielle singulière & unique solution) - Soit le problème de la forme (1.15), avec

$$E(t) = \begin{bmatrix} 0 & 0 \\ 1 & -t \end{bmatrix}, \quad A(t) = \begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix} \quad \text{et} \quad f = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix},$$

où $t \in \mathbb{R}$. La paire matricielle est singulière pour tout $t \in \mathbb{R}$, pourtant de simples calculs conduisent à l'unique solution

$$X(t) = \begin{pmatrix} tf_2 - t\dot{f}_1 + f_1 \\ f_2 - \dot{f}_1 \end{pmatrix}.$$

□

La régularité de la paire matricielle ne fournit plus d'informations utiles à la résolution de (1.15) ; le cœur de cette difficulté réside dans le choix de la relation d'équivalence permettant de transformer la paire matricielle.

1.3.2 Formes canoniques généralisées

Puisque les coefficients $E(t)$ et $A(t)$ sont variables, la transformation $\ll \sim \gg$ définie dans la section 1.1.2.2 doit pouvoir tenir compte de leurs dérivées. Ce problème ne se posait évidemment pas dans le cas des coefficients constants car si $(E_1, A_1) \sim (E_2, A_2)$, alors

$$E_1\dot{X} = A_1X + f_1 \;\Leftrightarrow\; Q_1E_2Q_2\dot{X} = Q_1A_2Q_2X + f_1 \;\Leftrightarrow\; E_2\dot{Y} = A_2Y + f_2,$$

où $Y = Q_2X$ et $f_2 = Q_1^{-1}f_1$. Dans le contexte variable, $Y = Q_2(t)X$ et par conséquent

$$\dot{Y} = \dot{Q}_2(t)X + Q_2(t)\dot{X}.$$

On note ainsi une nouvelle relation d'équivalence $\ll \sim_g \gg$ [2] ([28], p. 57) telle que

$$\begin{aligned}(E_1(t), A_1(t)) &\sim_g (E_2(t), A_2(t)) \\ &\Leftrightarrow\; E_1(t) = Q_1(t)E_2(t)Q_2(t) \quad \text{et} \quad A_1(t) = Q_1(t)A_2(t)Q_2(t) - Q_1(t)E_2(t)\dot{Q}_2(t).\end{aligned}$$

On obtient ainsi

$$E_1(t)\dot{X} = A_1(t)X + f_1 \quad \Leftrightarrow \quad E_2(t)\dot{Y} = A_2(t)Y + f_2,$$

où $f_2 = Q_1^{-1}(t)f_1$. Pour des raisons techniques de recherche d'invariants, on définit une autre relation d'équivalence notée $\ll \sim_l \gg$ [3] ([28], p. 58) telle que

$$(E_1, A_1) \sim_l (E_2, A_2) \;\Leftrightarrow\; E_1 = Q_1E_2Q_2 \quad \text{et} \quad A_1 = Q_1A_2Q_2 - Q_1E_2W,$$

2. g pour *global*.
3. l pour *local*.

où $W \in \mathbb{R}^{n\times n}$. Cette relation d'équivalence regarde le comportement local de la paire de matrices $(E_1(t), A_1(t))$ (autour d'un point t) ; on suppose que localement, les fonctions $E_1(t)$ et $A_1(t)$ sont constantes.

– Dans le cadre de la relation « $\sim_l$ », on établit que

$$(E(t), A(t)) \sim_l \left(\begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & I_{r-s} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right). \tag{1.19}$$

Cette forme de la paire de matrices permet de déterminer les invariants locaux r (le rang de E), a (le nombre d'équations algébriques) et s (l'étrangeté).

– Dans le cadre de la relation « $\sim_g$ », on établit que

$$(E(t), A(t)) \sim_g \left(\begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & I_{r-s} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & B_1(t) & 0 & B_2(t) \\ 0 & 0 & 0 & B_3(t) \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right). \tag{1.20}$$

Les expressions (1.19) et (1.20) sont des *formes canoniques généralisées*. Elles sont utilisées afin de résoudre (1.15). Notons qu'il est possible d'exhiber d'autres formes canoniques (on peut consulter par exemple les travaux de M. P. Quéré et G. Villard [45]).

Remarque 3

> *L'étrangeté est une notion qui permet de généraliser l'indice de différentiation. L'intérêt de cette notion sera abordé dans la section suivante.*

1.3.3 Résolution du problème linéaire à coefficients variables

On commence par transformer la paire matricielle $(E(t), A(t))$ (éventuellement non carrée) via (1.19), puis on transforme cette nouvelle paire obtenue via (1.20). Observons la forme développée de (1.20)

$$\begin{cases} \dot{X}_1 = B_1(t)X_2 + B_2(t)X_4 + f_1 & \text{(1.21a)} \\ \dot{X}_2 = B_3(t)X_4 + f_2 & \text{(1.21b)} \\ 0 = X_3 + f_3 & \text{(1.21c)} \\ 0 = X_1 + f_4 & \text{(1.21d)} \\ 0 = f_5. & \text{(1.21e)} \end{cases}$$

En dérivant (1.21d) et en injectant ceci dans (1.21a), on a

$$\begin{cases} 0 = B_1(t)X_2 + B_2(t)X_4 + f_1 + \dot{f}_4 \\ \dot{X}_2 = B_3(t)X_4 + f_2 \\ 0 = X_3 + f_3 \\ 0 = X_1 + f_4 \\ 0 = f_5. \end{cases} \tag{1.22}$$

On vient de définir une étape d'une méthode itérative. À chaque pas de ce processus, on exhibe un triplet (r_k, a_k, s_k). On peut montrer que la suite $(r_p, a_p, s_p)_{p\in\mathbb{N}^\star}$ mise en avant par la méthode devient stationnaire à partir d'un certain rang, noté ν_E ([28], p. 73).

Théorème 5

On considère le problème (1.15) (éventuellement sa version sous ou sur-déterminée). La méthode précédemment décrite s'achève en un nombre fini d'étapes. Plus précisément, la méthode, qui se termine au bout de ν_E itérations, où ν_E désigne le rang à partir duquel la suite $(r_p, a_p, s_p)_{p\in\mathbb{N}^\star}$ devient stationnaire, fournit une EDA de la forme

$$\begin{cases} \dot{X}_1 = A_1(t)X_3 + f_1 \\ 0 = X_2 + f_2 \\ 0 = f_3, \end{cases} \tag{1.23}$$

où les f_i sont déterminés par les dérivées de f.

Remarque 4

Pour des problèmes carrés, la partie libre X_3 ne figure plus dans (1.23). Dans ce cas, la condition de compatibilité $0 = f_3$ disparait également et sous réserve de la consistance des conditions initiales, le problème possède une unique solution ([28], p. 74 - 75).

Indice d'étrangeté. La quantité ν_E est nommée *indice d'étrangeté* [26]. Cet indice généralise la notion d'indice de différentiation pour des systèmes rectangulaires (quand il n'y a donc plus unicité de la solution), c'est-à-dire qu'il correspond au nombre de dérivations nécessaires pour écrire $\dot{X}$ comme fonction continue de X. Il possède une autre propriété : il donne un indice qui sera égal tant pour les EDOs que pour les équations algébriques. L'indice de différentiation vaut toujours 0 pour une EDO et 1 pour un système algébrique. Cette différence est gommée avec l'indice d'étrangeté ; il est nul pour les EDOs et pour les équations algébriques.

1.3.4 Illustration de la méthode de *Kunkel-Mehrmann*

Observons l'exemple (1.18) :

$$\begin{cases} -t\dot{x}_1 + t^2\dot{x}_2 = -x_1 & \text{(1.24a)} \\ -\dot{x}_1 + t\dot{x}_2 = -x_2, & \text{(1.24b)} \end{cases}$$

où la paire matricielle $(E(t), A(t)) = \left(\begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}\right)$. En multipliant l'équation (1.24b) par t et en comparant par rapport à (1.24a), on obtient l'équation algébrique $0 = -x_1 + tx_2$. Le problème (1.24) est équivalent à

$$\begin{cases} -\dot{x}_1 + t\dot{x}_2 = -x_2 \\ 0 = -x_1 + tx_2, \end{cases}$$

c'est-à-dire $(E(t), A(t)) \sim_g \left(\begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ -1 & t \end{bmatrix}\right)$. On pose à présent $y_1 = x_1 - tx_2$ et $y_2 = x_2$. Ainsi, (1.24) devient

$$\begin{cases} \dot{y}_1 = 0 \\ 0 = y_1, \end{cases}$$

où $(E(t), A(t)) \sim_g \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\right)$. Le passage de (1.21) à (1.22) fournit l'équation $0 = y_1$. L'indice d'étrangeté vaut donc 1.

1.4 Méthode des projections

La méthode des projections consiste, comme son nom tend à l'indiquer, à décomposer le vecteur X solution de (1.1) (respectivement (1.15)) en une somme de *projetés* par rapport aux noyaux (et aux images) de matrices particulières, la première étant la matrice coefficient de $\dot{X}$ dans (1.1) (respectivement (1.15)).

Par souci de clarté, nous détaillerons la décomposition effectuée pour un problème du type (1.1) d'indice de Kronecker $\nu_K = 1$, puis nous indiquerons la généralisation aux problèmes (1.15) d'indice de Kronecker $\nu_K \geqslant 1$. Enfin, nous donnerons les résultats obtenus dans le cadre de problèmes linéaires à coefficients variables où la notion d'indice de traçabilité sera introduite.

1.4.1 EDAs linéaires à coefficients constants

1.4.1.1 Indice de Kronecker égal à 1

On regarde le système linéaire à coefficients constants régulier (1.1), en commençant par un problème d'indice de Kronecker $\nu_K = 1$. Le principe de la méthode des projections peut en effet être extrait de ce contexte fondamental.

On se munit d'un *projecteur* Q (i.e. une matrice idempotente) sur le noyau de E.

Remarque 5

La propriété d'idempotence permet de montrer que pour tout projecteur Q, la relation de transversalité $\mathbb{R}^n = \ker Q \oplus im\, Q$ est satisfaite. On peut notamment se référer à l'ouvrage de C. D. Meyer *([31], p. 394).*

On établit un lien entre la régularité de la paire matricielle (E, A) et la régularité de la matrice $\bar{E} = E + AQ$. Commençons par démontrer le résultat suivant :

Proposition 2 (*[18]*)

L'indice de la paire matricielle (E, A) relative au problème (1.1) est égal à 1 si et seulement si

$$x \in \ker E \quad et \quad Ax \in im\, E \Rightarrow x = 0.$$

Preuve - Supposons que la paire matricielle (E, A) est d'indice 1, autrement dit l'indice de la matrice $\mathcal{E} := E_A^{-1}E$ vaut 1. Cela se traduit par la relation de transversalité (C. D. Meyer [31], p. 394) $\mathbb{R}^n = \ker \mathcal{E} \oplus \operatorname{im} \mathcal{E}$. On considère d'une part $x \in \mathbb{R}^n$ tel que $x \in \ker E$. Par définition de $\mathcal{E}$, on a également

$$x \in \ker \mathcal{E}. \tag{1.25}$$

D'autre part, soit $y \in \mathbb{R}^n$ tel que $Ax = Ey$. Puisque $E_A x = \lambda Ex + Ax = Ax$, on a $E_A x = Ey$ et par inversibilité de E_A, on obtient $x = \mathcal{E}y$, autrement dit

$$x \in \operatorname{im} \mathcal{E}. \tag{1.26}$$

La relation de transversalité ainsi que les résultats (1.25) et (1.26) fournissent $x = 0$.

Réciproquement, on considère que l'implication ($x \in \ker E$ et $Ax \in \operatorname{im} E \Rightarrow x = 0$) est satisfaite. Supposons que l'indice de la paire matricielle (E, A) est strictement supérieur à 1. En utilisant la forme de Jordan, on établit

$$\mathcal{E} = R \begin{bmatrix} \mathcal{J} & 0 \\ 0 & \mathcal{N} \end{bmatrix} R^{-1},$$

où $\mathcal{J}$ est une matrice inversible, $\mathcal{N}$ est une matrice nilpotente et R une matrice de passage. On sait donc par hypothèse que la matrice nilpotente $\mathcal{N}$ n'est pas la matrice nulle. De ce fait, il existe $y \neq 0$ tel que :

$$\begin{bmatrix} \mathcal{J} & 0 \\ 0 & \mathcal{N} \end{bmatrix} y \neq 0 \quad \text{et} \quad \begin{bmatrix} \mathcal{J}^2 & 0 \\ 0 & \mathcal{N}^2 \end{bmatrix} y = 0.$$

On pose $x = \mathcal{E}Ry$. Par construction, $x = R \begin{bmatrix} \mathcal{J} & 0 \\ 0 & \mathcal{N} \end{bmatrix} y \neq 0$. Or

$$\begin{aligned} Ex &= E\mathcal{E}Ry \\ &= E_A E_A^{-1} E E_A^{-1} ERy \\ &= E_A R \begin{bmatrix} \mathcal{J} & 0 \\ 0 & \mathcal{N} \end{bmatrix} R^{-1} R \begin{bmatrix} \mathcal{J} & 0 \\ 0 & \mathcal{N} \end{bmatrix} R^{-1} Ry \\ &= E_A R \begin{bmatrix} \mathcal{J}^2 & 0 \\ 0 & \mathcal{N}^2 \end{bmatrix} y \\ &= 0. \end{aligned}$$

Ainsi,

$$x \in \ker E. \tag{1.27}$$

De plus, $Ax = E_A x = E_A \mathcal{E}Ry = E_A E_A^{-1} ERy = ERy$. Autrement dit,

$$Ax \in \operatorname{im} E. \tag{1.28}$$

Grâce aux résultats (1.27) et (1.28), on en déduit que $x = 0$, ce qui est en contradiction avec la définition de x. Par conséquent, l'indice de la paire matricielle est inférieur à 1. Mais il ne peut être nul car la matrice $\mathcal{E}$ n'est pas inversible ; il vaut de fait 1. □

Nous pouvons à présent montrer le résultat suivant :

Proposition 3

Soit (E, A) une paire matricielle régulière et Q un projecteur sur $\ker E$. On a alors la caractérisation suivante :

$$(E, A) \text{ est d'indice 1} \Leftrightarrow \bar{E} \text{ est inversible.}$$

Preuve - Grâce à la proposition 2, il suffit de montrer que $x \in \ker E$ et $Ax \in \operatorname{im} E \Rightarrow x = 0$ équivaut à $\det \bar{E} \neq 0$.

On commence par supposer que la matrice $\bar{E}$ est singulière. Soit x un vecteur non nul de son noyau. On pose $x_Q = Qx$. Clairement, $x_Q \in \ker E$ (par définition de Q). De plus, $Ax_Q = AQx = -Ex$; ainsi $Ax_Q \in \operatorname{im} E$. Il reste à remarquer que x_Q ne peut être nul. Si tel était le cas, alors $x \in \ker Q$. Comme $Ex = -AQx = -Ax_Q = 0$, x appartient également à $\ker E = \operatorname{im} Q$. On en déduit que $x = 0$ ce qui contredit l'hypothèse initiale. Ainsi, $x_Q \neq 0$. Par contraposition, une première implication est prouvée.

Démontrons la seconde implication également par contraposition. Soit y un vecteur non nul tel que $Ey = 0$ et $Ay = Ex$. Puisque y est déjà dans le noyau de E, on a $Qy = y$. Il reste à fournir un vecteur y_Q non nul tel que $\bar{E}y_Q = 0$. On pose pour cela $y_Q = y - (I - Q)x$. Un tel y_Q ne peut être nul. En effet, si tel était le cas, alors $y \in \operatorname{im}(I - Q)$. Or y appartient déjà à $\operatorname{im} Q$; en conséquence $y = 0$, ce qui contredit l'hypothèse initiale. Ainsi, $\bar{E}y_Q = Ey - E(I - Q)x + AQy - AQ(I - Q)x = -Ex + Ay = 0$. La matrice $\bar{E}$ est donc singulière. □

Cette caractérisation permet de découpler le problème (1.1).

Théorème 6 (*R. März [35]*)

On considère le problème régulier (1.1). Soit Q un projecteur sur $\ker E$, $P = I - Q$, $X_Q = QX$, $X_P = PX$ et $\bar{E} = E + AQ$. Si la paire matricielle (E, A) est d'indice de Kronecker 1, alors (1.1) est équivalent à

$$\begin{cases} \dot{X}_P = P\bar{E}^{-1}AX_P + P\bar{E}^{-1}f \\ X_Q = -Q\bar{E}^{-1}AX_P - Q\bar{E}^{-1}f. \end{cases} \tag{1.29}$$

Preuve - On commence par écrire $X = (I - Q + Q)X = (P + Q)X = X_P + X_Q$. En pré-multipliant (1.1) par $\bar{E}^{-1}$ et en y injectant la nouvelle expression de X, on parvient à

$$\bar{E}^{-1}E\dot{X}_P + \bar{E}^{-1}E\dot{X}_Q = \bar{E}^{-1}AX_P + \bar{E}^{-1}AX_Q + \bar{E}^{-1}f. \tag{1.30}$$

Par définition de Q, on a $EQ = 0$ et $QP = Q(I - Q) = Q - Q^2 = 0$. Ainsi, $\bar{E}P = (E + AQ)P = EP + AQP = E(I - Q) = E$ et $\bar{E}Q = (E + AQ)Q = EQ + AQ^2 = AQ$. On en déduit $\bar{E}^{-1}E = P$ et $\bar{E}^{-1}AQ = Q$. (1.30) devient alors

$$\dot{X}_P = \bar{E}^{-1}AX_P + X_Q + \bar{E}^{-1}f. \tag{1.31}$$

En pré-multipliant (1.31) par P (respectivement par Q), on obtient la première (respectivement la deuxième) équation de (1.29). □

L'équation (1.31) est appelée EDO *sous-jacente* du problème (1.1) (bien qu'elle diffère de l'équation (1.17) pareillement nommée).

Si l'inconnue X est suffisamment régulière, le théorème précédent devient en réalité une caractérisation (R. Riaza [51], p. 30 – 33).

1.4.1.2 Exemple

La théorie des circuits électriques fournit de nombreux exemples d'utilisation des EDAs (S. Schulz [54]).

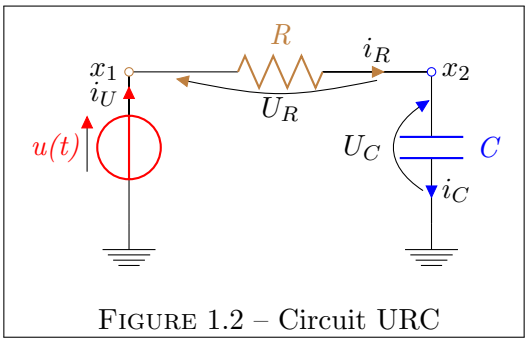


Figure 1.2 – Circuit URC

Modélisation - On ajoute au circuit RC (figure 1.1) une source de tension $u(t)$. La capacité C du condensateur est supposée strictement positive. La loi des nœuds appliquée au circuit URC (Figure 1.2) se traduit par $i_U = i_R = i_C$. Quant à la loi des mailles, elle fournit $u(t) = u_R + u_C$. Par analyse du circuit, on écrit

$$\begin{cases} u(t) = x_1 \\ u_R = x_1 - x_2 \\ u_C = x_2. \end{cases}$$

En appliquant les mêmes règles que pour le circuit RC (figure 1.1), on parvient au système

$$\begin{cases} C\dot{x}_2 = R^{-1}(x_1 - x_2) \\ 0 = i_R - R^{-1}(x_1 - x_2) \\ 0 = x_1 - u(t). \end{cases} \tag{1.32}$$

Appliquons la méthode des projections à l'EDA précédente : le problème (1.32) est modélisé par une EDA du type (1.1), où

$$E = \begin{bmatrix} 0 & C & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; A = \begin{bmatrix} R^{-1} & -R^{-1} & 0 \\ -R^{-1} & R^{-1} & 1 \\ 1 & 0 & 0 \end{bmatrix}, \; X = \begin{pmatrix} x_1 \\ x_2 \\ i_R \end{pmatrix} \text{ et } f = \begin{pmatrix} 0 \\ 0 \\ -u(t) \end{pmatrix}.$$

On note

$$Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \; X_Q = \begin{pmatrix} x_1 \\ 0 \\ i_R \end{pmatrix}, \; P = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ et } X_P = \begin{pmatrix} 0 \\ x_2 \\ 0 \end{pmatrix}.$$

Un simple calcul montre que la paire matricielle (E, A) est régulière d'indice 1. Par conséquent, la matrice $\bar{E}$ est inversible ($\det \bar{E} = C$) :

$$\bar{E}^{-1} = \begin{bmatrix} 0 & 0 & 1 \\ C^{-1} & 0 & -(RC)^{-1} \\ 0 & 1 & R^{-1} \end{bmatrix}.$$

On utilise ensuite les résultats du théorème précédent ; la première équation de (1.29) fournit

$$\dot{x}_2 = -\frac{x_2}{RC} + \frac{u(t)}{RC}.$$

Quant à l'équation algébrique de (1.29), elle devient

$$\begin{cases} x_1 = u(t) \\ i_R = -\dfrac{x_2}{R} + \dfrac{u(t)}{R}. \end{cases}$$

Il suffit alors de résoudre l'équation différentielle en x_2, puis d'injecter l'expression obtenue dans la seconde équation algébrique pour exprimer i_R.

1.4.1.3 Indices supérieurs

On peut généraliser ce procédé au problème (1.1) d'indice de Kronecker ν_K quelconque (R. März [36]). On commence par définir une suite de paires matricielles $(E_n, A_n)_{n\in\mathbb{N}}$ telle que $(E_0, A_0) = (E, A)$ et $(E_{i+1}, A_{i+1}) = (E_i + A_i Q_i, A_i Q_i)$ pour tout i, où Q_i est un projecteur sur $\ker E_i$. On obtient une caractérisation similaire entre l'indice de la paire matricielle (E, A) et les projecteurs :

Proposition 4

Soit (E, A) une paire matricielle régulière. On considère la suite $(E_n, A_n)_{n\in\mathbb{N}}$ précédemment définie. On a alors la caractérisation suivante :

$$(E, A) \text{ est d'indice } \nu_K \Leftrightarrow E_i \text{ est singulière pour tout } i < \nu_K \text{ et } E_{\nu_K} \text{ est inversible.}$$

Cette proposition permet d'établir le résultat suivant :

Théorème 7 (*R. Riaza [51]*)

On considère le problème régulier (1.1). Soit $(E_n, A_n)_{n\in\mathbb{N}}$ la suite précédemment décrite. On pose $X = U + V_1 + \cdots + V_{\nu_K-1}$. Si la paire matricielle (E, A) est d'indice de Kronecker ν_K, alors (1.1) est équivalent à

$$\begin{cases} \dot{U} = P_0 \dots P_{\nu_K-1} G_{\nu_K}^{-1} A U + P_0 \dots P_{\nu_K-1} G_{\nu_K}^{-1} f \\ V_k = -\mathcal{K}_k U + \displaystyle\sum_{j=k+1}^{\nu_K-1} \mathcal{N}_{kj} \dot{V}_j + \mathcal{L}_k f, \end{cases} \tag{1.33}$$

pour tout $k = \nu_K - 1, \dots, 1$, avec

$$\begin{aligned} \mathcal{K}_k &= P_0 \dots P_{k-1} Q_k P_{k+1} \dots P_{\nu_K-1} G_{\nu_K}^{-1} A, \\ \mathcal{N}_{kj} &= P_0 \dots P_{k-1} Q_k P_{k+1} \dots P_{j-1} Q_j, \\ \mathcal{L}_k &= P_0 \dots P_{k-1} Q_k P_{k+1} \dots P_{\nu_K-1} G_{\nu_K}^{-1}, \end{aligned}$$

$G_{\nu_K}^{-1}$ étant une correction de $E_{\nu_K}^{-1}$.

Le théorème précédent permet ainsi de résoudre le problème régulier (1.1) ; on résout en premier lieu l'EDO portant sur U dans (1.33), puis on détermine V_{ν_K-1}, V_{ν_K-2}, …, V_1.

1.4.2 Cadre linéaire à coefficients variables

La méthode des projections peut s'étendre, non sans effort, aux EDAs linéaires à coefficients variables (1.15). En suivant le même type de raisonnement, on parvient à un résultat de forme similaire à celui obtenu pour l'équation (1.1). L'indice de Kronecker n'étant plus défini dans le contexte à coefficients variables, on l'étend par la notion d'*indice de traçabilité* (on peut se référer à R. Riaza [51], S. Schulz [54] p. 14 – 19).

1.5 Méthode de *Jacobi-Pryce*

1.5.1 Motivation

On souhaite étudier un problème général de la forme $F(X, Y) = 0$. On sait que si la différentielle de F par rapport à Y est inversible, alors on peut écrire Y comme une fonction continue de X. L'ensemble $\{(X, Y) \mid F(X, Y) = 0\}$ définit dans ce cas une variété différentielle. Reprenons l'exemple (1.11)

$$\begin{cases} C\,(\dot{x}_2 - \dot{x}_3) = R^{-1}\,(x_1 - x_2) \\ 0 = x_1 - x_3 - U \\ 0 = x_3, \end{cases}$$

qui peut s'écrire comme $F(X, Y) = 0$, où

$$X = \begin{pmatrix} x_2 \\ x_3 \end{pmatrix},\ Y = \begin{pmatrix} x_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} \text{ et } F(X, Y) = \begin{pmatrix} R^{-1}\,(x_1 - x_2) - C\,(\dot{x}_2 - \dot{x}_3) \\ x_1 - x_3 - U \\ x_3 \end{pmatrix}.$$

Il n'est pas possible ici d'exprimer Y comme une fonction continue de X car la différentielle $D_Y F(X, Y) = \begin{bmatrix} R^{-1} & -C & C \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ n'est pas inversible. L'ensemble $\{(X, Y) \mid F(X, Y) = 0\}$ n'est

pas une variété différentielle.

La méthode de *Pryce* [44] propose une analyse structurelle des EDAs. Cette analyse consiste à déterminer les variables et les équations qui doivent être dérivées afin de réduire l'ensemble $\{(X,Y) \mid F(X,Y)=0\}$ à une variété différentielle. On pourra alors modifier X, Y et $F(X,Y)$ à cet effet. La méthode de *Pryce* est une généralisation de la méthode de *Pantelides* [41] et retrouve la méthode de *Jacobi* (on peut consulter [40]). Cette méthode est utilisée pour déterminer les solutions des EDAs via les séries de Taylor ; on peut se référer aux travaux de J.D. Pryce [43], de J.D. Pryce et N.S. Nedialkov ([37], [38] et [39]) et de S. Iles [25]. Nous nous concentrons uniquement dans cette section sur l'aspect de réduction des EDAs par la méthode de *Pryce*.

1.5.2 Problème initial

On pose $f_1(X,Y) = R^{-1}(x_1 - x_2) - C(\dot{x}_2 - \dot{x}_3)$, $f_2(X,Y) = x_1 - x_3 - U$ et $f_3(X,Y) = x_3$. On cherche deux vecteurs $\mathbf{c} \in \mathbb{R}^3$ et $\mathbf{d} \in \mathbb{R}^3$ tels que l'on puisse réduire l'ensemble

$$\{(X,Y) \mid f_1 = 0, f_2 = 0, f_3 = 0\}$$

à une variété différentielle du type

$$\mathfrak{V} = \{(x_1, \dots, x_1^{(d_1)}, x_2, \dots, x_2^{(d_2)}, x_3, \dots, x_3^{(d_3)}) \mid \\ f_1 = 0, \dots, f_1^{(c_1)} = 0, f_2 = 0, \dots, f_2^{(c_2)} = 0, f_3 = 0, \dots, f_3^{(c_3)} = 0\}.$$

On cherche par ailleurs à ce que la dimension de $\mathfrak{V}$ soit minimale, à condition que $d_j - c_i$ soit supérieur ou égal à l'ordre de la dérivée de la variable j dans l'équation $f_i(X,Y) = 0$ (sans quoi il y aurait une contradiction). On note cet ordre σ_{ij}. On doit considérer en définitive le problème de minimisation sous contrainte suivant

$$\min_{d_j - c_i \geqslant \sigma_{ij}} \left(\sum_{j=1}^{3} d_j - \sum_{i=1}^{3} c_i \right).$$

1.5.3 Problème dual

Pour résoudre le problème initial qui est un problème d'optimisation discrète, on passe par le problème dual suivant :

$$\max_{\mathcal{C}} \sum \sigma_{ij}\xi_{ij}, \quad \mathcal{C} = \left\{ \xi_{ij} \geqslant 0 \mid \sum_i \xi_{ij} = \sum_j \xi_{ij} = 1 \right\}.$$

On commence par établir la *matrice d'ordre des dérivées* $\Sigma = (\sigma_{ij})_{1 \leqslant i,j \leqslant 3}$. Cette matrice détermine le poids de chaque variable dans chaque équation en fonction de son degré de dérivation (quand la variable n'apparaît pas, $\sigma_{ij} = -\infty$). On obtient

$$\Sigma_1 = \begin{bmatrix} 0 & 1 & 1 \\ 0 & -\infty & 0 \\ -\infty & -\infty & 0 \end{bmatrix}.$$

On résout ensuite le problème dual en considérant les différentes façons de sommer les coefficients de la matrice d'ordre des dérivées (pour respecter la contrainte $\mathcal{C}$). Ceci revient à considérer les

six matrices (de permutation) suivantes, en leur associant les coefficients correspondants de la matrice Σ_1 et en sommant ces derniers :

$$\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} \longrightarrow \begin{bmatrix}0&1&1\\0&-\infty&0\\-\infty&-\infty&0\end{bmatrix} \longrightarrow 0-\infty+0=-\infty,$$
$$\begin{bmatrix}1&0&0\\0&0&1\\0&1&0\end{bmatrix} \longrightarrow \begin{bmatrix}0&1&1\\0&-\infty&0\\-\infty&-\infty&0\end{bmatrix} \longrightarrow 0+0-\infty=-\infty,$$
$$\begin{bmatrix}0&1&0\\1&0&0\\0&0&1\end{bmatrix} \longrightarrow \begin{bmatrix}0&1&1\\0&-\infty&0\\-\infty&-\infty&0\end{bmatrix} \longrightarrow 1+0+0=\boxed{1},$$
$$\begin{bmatrix}0&0&1\\1&0&0\\0&1&0\end{bmatrix} \longrightarrow \begin{bmatrix}0&1&1\\0&-\infty&0\\-\infty&-\infty&0\end{bmatrix} \longrightarrow 1+0-\infty=-\infty,$$
$$\begin{bmatrix}0&1&0\\0&0&1\\1&0&0\end{bmatrix} \longrightarrow \begin{bmatrix}0&1&1\\0&-\infty&0\\-\infty&-\infty&0\end{bmatrix} \longrightarrow 1+0-\infty=-\infty,$$
$$\begin{bmatrix}0&0&1\\0&1&0\\1&0&0\end{bmatrix} \longrightarrow \begin{bmatrix}0&1&1\\0&-\infty&0\\-\infty&-\infty&0\end{bmatrix} \longrightarrow 1-\infty-\infty=-\infty.$$

Le maximum (égal à 1) est atteint en faisant $\sigma_{12}+\sigma_{21}+\sigma_{33}$, soit pour le jeu de coordonnées $S_1=(\{1,2\},\{2,1\},\{3,3\})$. Le problème dual est ainsi résolu.

1.5.4 Résolution du problème initial

En considérant le jeu de coordonnées S_1, on trouve les vecteurs $\mathbf{c}$ et $\mathbf{d}$ (qui ne sont pas uniques en général) en appliquant la procédure suivante : on commence par $\mathbf{c}^0=0$, puis on définit les vecteurs $\mathbf{d}^1$ et $\mathbf{c}^1$ tels que $d_j^1=\max_i\left(\sigma_{ij}+c_i^1\right)$ et $c_i^1=d_j^1-\sigma_{ij}$. En itérant ce procédé, on décrit deux suites $\mathbf{c}^k$ et $\mathbf{d}^k$. Ces suites deviennent toujours stationnaires. En notant k_0 le rang à partir duquel ces suites sont stationnaires, on récupère $\mathbf{c}=\mathbf{c}^{k_0}$ et $\mathbf{d}=\mathbf{d}^{k_0}$.

Ainsi,

$$\begin{aligned}
c^0&=(0,0,0),\\
d^1&=\left(\max_i\left(\sigma_{i1}\right),\max_i\left(\sigma_{i2}\right),\max_i\left(\sigma_{i3}\right)\right)=(0,1,1),\\
c^1&=(d_2-\sigma_{12},d_1-\sigma_{21},d_3-\sigma_{33})=(0,0,1),\\
d^2&=\left(\max_i\left(\sigma_{i1}+c_i\right),\max_i\left(\sigma_{i2}+c_i\right),\max_i\left(\sigma_{i3}+c_i\right)\right)=(0,1,1),\\
c^2&=(d_2-\sigma_{12},d_1-\sigma_{21},d_3-\sigma_{33})=(0,0,1)=c^1.
\end{aligned}$$

Le problème initial est résolu, les valeurs recherchées sont :

$$\mathbf{d}=(d_1,d_2,d_3)=(0,1,1)\quad\text{et}\quad\mathbf{c}=(c_1,c_2,c_3)=(0,0,1).$$

La méthode de *Pryce* indique ensuite la construction du système tenant compte des équations dérivées. De par les indices précédemment trouvés, on en déduit qu'il faut dériver la dernière équation une fois, et que toutes les variables apparaissent ainsi que leurs premières dérivées sauf

$\dot{x}_1$. On obtient

$$\begin{cases} C(\dot{x}_3 - \dot{x}_2) - R^{-1}(x_2 - x_1) = 0 \\ x_1 - x_3 - U = 0 \\ x_3 = 0 \\ \dot{x}_3 = 0. \end{cases}$$

On réserve l'équation $x_3 = 0$ et le reste du problème s'écrit

$$\begin{cases} C(\dot{x}_3 - \dot{x}_2) - R^{-1}(x_2 - x_1) = 0 \\ x_1 - U = 0 \\ \dot{x}_3 = 0. \end{cases}$$

On remarque qu'il est maintenant possible d'exprimer $Y = \begin{pmatrix} x_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{pmatrix}$ en fonction de $X = \begin{pmatrix} x_2 \\ x_3 \end{pmatrix}$: en posant

$$\bar{F}(X, Y) = \begin{pmatrix} C(\dot{x}_3 - \dot{x}_2) - R^{-1}(x_2 - x_1) \\ x_1 - U \\ \dot{x}_3 \end{pmatrix},$$

on a

$$D_Y \bar{F}(X, Y) = \begin{bmatrix} R^{-1} & -C & C \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

qui est inversible. On est parvenu au but escompté.

Remarque 6

Les calculs sont simplifiés puisque l'exemple est linéaire. En toute généralité, il faut utiliser le théorème des fonctions implicites pour pouvoir exprimer certaines variables en fonction d'autres.

1.5.5 Illustration avec le pendule simple

Reprenons les équations modélisant les oscillations d'un pendule simple en dimension $n = 2$: on a $F(X, Y) = 0$, où $X = \begin{pmatrix} x \\ y \end{pmatrix}$, $Y = \begin{pmatrix} \ddot{x} \\ \ddot{y} \\ \lambda \end{pmatrix}$ et $F(X, Y) = \begin{pmatrix} \ddot{x} + \lambda x \\ \ddot{y} + \lambda y - \mathfrak{g} \\ x^2 + y^2 - 1 \end{pmatrix}$. On remarque que la matrice $D_Y \bar{F}(X, Y) = \begin{bmatrix} 1 & 0 & x \\ 0 & 1 & y \\ 0 & 0 & 0 \end{bmatrix}$ est singulière. Appliquons la méthode de *Pryce* : commençons par écrire la matrice d'ordre des dérivées :

$$\Sigma_2 = \begin{bmatrix} 2 & -\infty & 0 \\ -\infty & 2 & 0 \\ 0 & 0 & -\infty \end{bmatrix}.$$

Résolvons le problème dual :

$$\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} \longrightarrow \begin{bmatrix}2&-\infty&0\\-\infty&2&0\\0&0&-\infty\end{bmatrix} \longrightarrow 2+2-\infty=-\infty,$$
$$\begin{bmatrix}1&0&0\\0&0&1\\0&1&0\end{bmatrix} \longrightarrow \begin{bmatrix}2&-\infty&0\\-\infty&2&0\\0&0&-\infty\end{bmatrix} \longrightarrow 2+0+0=\boxed{2},$$
$$\begin{bmatrix}0&1&0\\1&0&0\\0&0&1\end{bmatrix} \longrightarrow \begin{bmatrix}2&-\infty&0\\-\infty&2&0\\0&0&-\infty\end{bmatrix} \longrightarrow -\infty-\infty-\infty=-\infty,$$
$$\begin{bmatrix}0&0&1\\1&0&0\\0&1&0\end{bmatrix} \longrightarrow \begin{bmatrix}2&-\infty&0\\-\infty&2&0\\0&0&-\infty\end{bmatrix} \longrightarrow 0-\infty+0=-\infty,$$
$$\begin{bmatrix}0&1&0\\0&0&1\\1&0&0\end{bmatrix} \longrightarrow \begin{bmatrix}2&-\infty&0\\-\infty&2&0\\0&0&-\infty\end{bmatrix} \longrightarrow -\infty+0+0=-\infty,$$
$$\begin{bmatrix}0&0&1\\0&1&0\\1&0&0\end{bmatrix} \longrightarrow \begin{bmatrix}2&-\infty&0\\-\infty&2&0\\0&0&-\infty\end{bmatrix} \longrightarrow 0+2+0=\boxed{2}.$$

Le maximum (égal à 2) est atteint pour les jeux de coordonnées $S_1 = (\{1,1\},\{2,3\},\{3,2\})$ et $S_2 = (\{1,3\},\{2,2\},\{3,1\})$. Utilisons par exemple S_1 pour résoudre le problème initial :

$$\begin{aligned}
c^0 &= (0,0,0),\\
d^1 &= \left(\max_i(\sigma_{i1}), \max_i(\sigma_{i2}), \max_i(\sigma_{i3})\right) = (2,2,0),\\
c^1 &= (d_1-\sigma_{11}, d_3-\sigma_{23}, d_2-\sigma_{32}) = (0,0,2),\\
d^2 &= \left(\max_i(\sigma_{i1}+c_i), \max_i(\sigma_{i2}+c_i), \max_i(\sigma_{i3}+c_i)\right) = (2,2,0),\\
c^2 &= (d_1-\sigma_{11}, d_3-\sigma_{23}, d_2-\sigma_{32}) = (0,0,2) = c^1.
\end{aligned}$$

Le problème d'optimisation est résolu, les valeurs recherchées sont :

$$\mathbf{d} = (d_1,d_2,d_3) = (2,2,0) \quad \text{et} \quad \mathbf{c} = (c_1,c_2,c_3) = (0,0,2).$$

On construit le système tenant compte des équations dérivées :

$$\begin{cases}
\ddot{x} + \lambda x = 0\\
\ddot{y} + \lambda y - \mathfrak{g} = 0\\
x^2 + y^2 - 1 = 0\\
2x\dot{x} + 2y\dot{y} = 0\\
2x\ddot{x} + 2y\ddot{y} + 2\dot{x}^2 + 2\dot{y}^2 = 0.
\end{cases}$$

Le système se résout en commençant par l'équation $x^2 + y^2 - 1 = 0$ (utilisation du théorème des fonctions implicites pour exprimer l'une des variables en fonction de l'autre). On exprime par exemple y en fonction de x. On utilise en suite l'équation $2x\dot{x} + 2y\dot{y} = 0$, où on exprime $\dot{y}$

en fonction de x et $\dot{x}$. On termine avec les trois dernières équations

$$\begin{cases} \ddot{x} + \lambda x = 0 \\ \ddot{y} + \lambda y - \mathfrak{g} = 0 \\ 2x\ddot{x} + 2y\ddot{y} + 2\dot{x}^2 + 2\dot{y}^2 = 0. \end{cases}$$

On pose

$$\bar{F}(X, Y) = \begin{pmatrix} \ddot{x} + \lambda x \\ \ddot{y} + \lambda y - \mathfrak{g} \\ 2x\ddot{x} + 2y\ddot{y} + 2\dot{x}^2 + 2\dot{y}^2 \end{pmatrix}.$$

La matrice $D_Y \bar{F}(X, Y) = \begin{bmatrix} 1 & 0 & x \\ 0 & 1 & y \\ 2x & 2y & 0 \end{bmatrix}$ est inversible ($\det D_Y \bar{F}(X, Y) = -2$). On peut donc exprimer $(\ddot{x}, \ddot{y}, \lambda)$ en fonction de $(x, \dot{x}, y, \dot{y})$.

1.6 Méthode de *Rabier-Rheinboldt*

Nous adoptons dans cette section un point de vue géométrique ; nous considérons les EDAs comme des équations différentielles définies sur des variétés différentielles [50]. Les EDAs sont des équations différentielles sous contraintes ; il est donc possible et naturel d'envisager leur étude vis-à-vis des variétés différentielles. Nous décrivons essentiellement les travaux de P. J. Rabier et W. C. Rheinboldt [48] repris par R. Riaza [51] concernant une méthode globale de réduction des problèmes quasi-linéaires

$$E(X)\dot{X} = f(X), \tag{1.34}$$

où $E : \mathcal{I} \to \mathbb{R}^{n \times n}$ est une matrice singulière de *rang constant* r sur $\mathcal{I} \subset \mathbb{R}$ et $f : \mathcal{I} \to \mathbb{R}^n$. On suppose que les fonctions E et f sont suffisamment régulières. L'accent est porté sur des problèmes autonomes bien qu'une généralisation aux problèmes non-autonomes soit possible ([51], p. 123 – 129).

Remarque 7

Nous omettrons les détails trop techniques de la méthode par souci de concision.

1.6.1 Méthode globale de réduction - EDAs quasi-linéaires

Commençons par remarquer que si X, solution de (1.34), est une fonction de classe $\mathcal{C}^1(\mathcal{I}, \mathcal{O}_0)$, où $\mathcal{O}_0$ est un ouvert de $\mathbb{R}^n$, alors

$$X \in \mathcal{O}_1 = \{x \in \mathcal{O}_0 \mid f(x) \in \operatorname{im} E(x)\}.$$

Pour pouvoir manipuler l'espace $\mathcal{O}_1$, il est nécessaire d'en connaître la structure. Il se trouve que sous certaines hypothèses générales, $\mathcal{O}_1$ est une sous-variété de dimension r de $\mathcal{O}_0$.

Définissons ensuite l'application $\Lambda : T\mathcal{O}_0 \simeq \mathcal{O}_0 \times \mathbb{R}^n \to \mathbb{R}^n$ telle que $\Lambda(x, p) = E(x)p - f(x)$. On désigne par $T\mathcal{O}_0$ le fibré tangent de $\mathcal{O}_0$ qui correspond à l'union disjointe des espaces tangents en tous les points de $\mathcal{O}_0$. On pose $M_0 = \Lambda^{-1}(0)$ et $\pi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ la projection sur la première composante, de sorte que $\pi(M_0) = \mathcal{O}_1$. Il est clair que si X est une solution de (1.34), alors

$$(X, \dot{X}) \in M_0. \tag{1.35}$$

On a mis ainsi en avant deux variétés différentielles $\mathcal{O}_0$ et M_0. Le mécanisme de la méthode globale de réduction va construire, par un jeu de restrictions, deux suites de variétés $(\mathcal{O}_k)_{k\in\mathbb{N}}$ et $(M_k)_{k\in\mathbb{N}}$, dont les premiers termes sont précisément $\mathcal{O}_0$ et M_0. Il est nécessaire de requérir aux deux hypothèses suivantes :

1. La matrice $E(x)$ est de rang constant $r \leqslant n$, pour tout $x \in \mathcal{O}_1$.
2. L'application $\Lambda(x, p)$ est une submersion sur M_0.

La seconde hypothèse impose à la dérivée $\dot{\Lambda}$ d'être de rang maximal n (de rang plein) pour chaque point de M_0. Ceci permet l'utilisation du théorème des fonctions implicites, notamment pour exhiber des paramétrisations locales des variétés.

L'idée est de procéder par restrictions successives. Le vecteur $\dot{X}$ doit être tangent à $\mathcal{O}_1$ (de même qu'à $\mathcal{O}_0$). Par conséquent, puisque $X \in \mathcal{O}_1$, on doit avoir

$$(X, \dot{X}) \in T\mathcal{O}_1. \tag{1.36}$$

Les conditions (1.35) et (1.36) imposent

$$(X, \dot{X}) \in M_1 = T\mathcal{O}_1 \cap M_0 = \Lambda|^{-1}_{T\mathcal{O}_1}(0).$$

On vient de déterminer les termes $\mathcal{O}_1$ et M_1 des suites $(\mathcal{O}_k)_{k\in\mathbb{N}}$ et $(M_k)_{k\in\mathbb{N}}$. Par ailleurs,

$$X \in \mathcal{O}_2 = \pi\left(M_1\right) = \{x \in O_1 \mid f(x) \in \operatorname{im} E(x)|_{T_x\mathcal{O}_1}\}.$$

Si le même type d'hypothèses s'appliquent sur les restrictions $E(x)|_{T_x\mathcal{O}_1}$ et $\Lambda|_{T\mathcal{O}_1}$, alors $\mathcal{O}_2$ définit une sous-variété de $\mathcal{O}_1$ de dimension r_1, où r_1 est égal au rang de la matrice $E(x)|_{T_x\mathcal{O}_1}$. En itérant ce procédé, on parvient à construire les suites de variétés différentielles $(\mathcal{O}_k)_{k\in\mathbb{N}}$ et $(M_k)_{k\in\mathbb{N}}$:

$$\mathcal{O}_{k+1} = \{x \in \mathcal{O}_k \mid f(x) \in \operatorname{im} E(x)|_{T_x\mathcal{O}_k}\}$$

et

$$M_{k+1} = \Lambda|^{-1}_{T\mathcal{O}_{k+1}}(0).$$

1.6.2 Indice géométrique

Dans de telles conditions (validité des hypothèses précédemment décrites), les deux suites exhibées deviennent stationnaires pour un certain entier ν_G, vérifiant

$$M_0 \supset M_1 \supset \cdots \supset M_{\nu_G} = M_{\nu_G+1}, \quad \mathcal{O}_0 \supset \mathcal{O}_1 \supset \cdots \supset \mathcal{O}_{\nu_G} \supseteq \mathcal{O}_{\nu_G+1} = \mathcal{O}_{\nu_G+2}.$$

Définition 4 (*Indice géométrique*)

Le rang ν_G à partir duquel la suite $(M_k)_{k\in\mathbb{N}}$ devient stationnaire (dans le sens d'égalité entre variétés) est appelé indice géométrique *du problème (1.34).*

Pour retrouver les solutions de (1.34), on peut faire appel aux paramétrisations locales. Il est ainsi possible de décrire les solutions de (1.34) en terme d'équations réduites.

1.6.3 Illustration

Observons le problème sous forme d'Hessenberg suivant

$$\begin{cases} \dot{x} = a(x, y, z) \\ \dot{y} = b(x, y) \\ 0 = c(y). \end{cases} \tag{1.37}$$

On se place directement sur la variété définie par l'équation $0 = c(y)$. On peut bien entendu écrire cette variété comme $\mathcal{O}_0 = \{(x, y, z) \in \mathbb{R}^3 \mid 0 = c(y)\}$. Par simple dérivation de la contrainte $0 = c(y)$, on obtient $0 = \dot{c}(y)\dot{y} = \dot{c}(y)b(x, y)$. Le fait de dériver puis de substituer $\dot{y}$ par $b(x, y)$ correspond à la phase de restriction précédemment évoquée. On obtient ainsi $\mathcal{O}_1 = \{(x, y, z) \in \mathbb{R}^3 \mid 0 = c(y), 0 = \dot{c}(y)b(x, y)\}$. De la même manière, on a $\mathcal{O}_2 = \{(x, y, z) \in \mathbb{R}^3 \mid 0 = c(y), 0 = \dot{c}(y)b(x, y), 0 = \ddot{c}(y)b^2(x, y) + \dot{c}(y)\left(b_x(x, y)a(x, y, z) + b_y(x, y)b(x, y)\right\}$. En supposant que $\dot{c}(y)b_x(x, y)a_z(x, y, z)$ est inversible, on est en mesure d'exprimer $\dot{z}$ en fonction de x et y ; en effet, on a

$$\begin{aligned}\dot{z} &= -\frac{b^3c^{(3)} + 3b\ddot{c}\left(ab_x + bb_y\right) + \dot{c}\left(a^2b_{xx} + 2abb_{xy} + b^2b_{yy} + ab_x\left(a_x + b_y\right) + b\left(a_yb_x + b^2y\right)\right)}{a_zb_x\dot{c}} \\ &= d(x, y, z).\end{aligned}$$

La solution du problème (1.37) doit donc satisfaire l'EDO

$$\begin{cases} \dot{x} = a(x, y, z) \\ \dot{y} = b(x, y) \\ \dot{z} = d(x, y, z), \end{cases}$$

tout en vivant sur la variété $\mathcal{O}_2$. Pour pouvoir exprimer cette solution, on passe par les cartes locales de la variété.

1.7 Élimination différentielle

L'algèbre différentielle, initialement développée par J. Ritt [52], fournit des outils permettant la résolution de systèmes polynomiaux différentiels. Un processus symbolique, appelé *algorithme de Rosenfeld-Gröbner*, permet entre autres de traiter certaines EDAs [56] par des techniques relatives à l'élimination différentielle. Cette dernière peut être vue comme une généralisation de l'élimination gaussienne aux systèmes d'équations différentielles polynomiales.

1.7.1 Contexte

Dressons un rapide portrait des différentes techniques d'élimination (on peut se référer à F. Boulier [6]) :

- *Élimination gaussienne* : On souhaite résoudre un système linéaire possédant n équations. À l'aide du pivot de Gauss, on triangularise le système afin de faciliter sa résolution : on élimine par exemple x_1 dans les $n-1$ dernières équations, x_2 dans les $n-2$ dernières, etc. On arrive ainsi à un système triangulaire de la forme

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ \qquad\quad a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \qquad\qquad\qquad\qquad \vdots \\ \qquad\qquad\qquad\quad a_{nn}x_n = b_n. \end{cases}$$

  Cette technique revient à déterminer une base triangulaire avec laquelle il est possible d'exprimer les solutions du système.
- *Élimination algébrique* : Pour un système polynomial, on peut généraliser le principe précédent d'au moins deux façons. La première consiste à utiliser les bases de Gröbner [14]. Dans ce cadre, on classe les monômes entre eux suivant un ordre dit admissible.

Une base de Gröbner est ainsi un système de réécriture confluent qui réécrit un monôme en une combinaison linéaire de monômes plus petits pour l'ordre admissible fixé. En un certain sens, on obtient un système triangulaire, si l'on voit les monômes comme des variables indépendantes. La deuxième consiste à fixer un ordre sur les variables et calculer un ou plusieurs ensembles triangulaires dont la réunion est équivalente au système initial. On peut se référer à M. Moreno Maza [34]. Ces ensembles sont triangulaires dans le sens où deux polynômes ont toujours des variables principales différentes. Ces ensembles triangulaires sont plus précisément des chaînes régulières ([3] et [9] ainsi que leurs références).

– *Élimination différentielle* : Dans le contexte des systèmes polynomiaux différentiels, les bases de Gröbner ne sont pas adéquates car elles ne constituent pas nécessairement des bases engendrées par un nombre fini d'éléments. Elles sont généralisées par la notion de *chaînes différentielles régulières* [6], dont le but est, comme pour l'élimination gaussienne, d'obtenir des systèmes « triangulaires » afin de résoudre le problème initial.

C'est dans ce dernier contexte que prend source l'algorithme de Rosenfeld-Gröbner, mis au point par F. Boulier, D. Lazard, F. Ollivier et M. Petitot ([7] et [8]), et notamment amélioré par F. Lemaire [29] et E. Hubert [24]. Sans rentrer dans son analyse, nous utilisons directement cet algorithme sur deux exemples afin d'en extraire les idées principales.

1.7.2 Algorithme de Rosenfeld-Gröbner illustré

Exemple 1 - Reprenons le cas (1.11)

$$\begin{cases} C\left(\dot{x}_3 - \dot{x}_2\right) = R^{-1}\left(x_2 - x_1\right) \\ 0 = x_1 - x_3 - U \\ 0 = x_3. \end{cases}$$

On commence par fixer un ordre sur les variables. Considérons par exemple l'ordre

$$x_3 < x_2 < x_1 < \dot{x}_3 < \dot{x}_2 < \dot{x}_1 < \cdots$$

Puisque nous effectuons les calculs à la main, nous tâchons de les rendre aussi simples que possible. On choisit la troisième équation du système

$$0 = x_3.$$

Elle est simple et ne peut se simplifier. Poursuivons avec la seconde équation $0 = x_1 - x_3 - U$. En la combinant avec l'équation $0 = x_3$, elle se simplifie en donnant

$$0 = x_1 - U.$$

Quant à la première équation $C\left(\dot{x}_3 - \dot{x}_2\right) = R^{-1}\left(x_2 - x_1\right)$, on peut la réduire en utilisant $0 = x_1 - U$ et $0 = \dot{x}_3$. On obtient ainsi

$$-C\dot{x}_2 = R^{-1}(x_2 - U).$$

Les solutions du système initial (1.11) sont exactement les solutions du système réduit

$$\begin{cases} -C\dot{x}_2 = R^{-1}(x_2 - U) \\ 0 = x_1 - U \\ 0 = x_3. \end{cases}$$

Remarque 8

Un tel système est une chaîne différentielle régulière. L'aspect triangulaire vis-à-vis des variables apparaît ici clairement.

Les constantes U, R et C sont considérées non nulles dans l'exemple. En supposant qu'elles sont déclarées de la sorte pour l'application de l'algorithme, ce dernier ne fait pas de distinction de cas. En revanche et ce d'une manière générale, *Rosenfeld-Gröbner* distingue toutes les configurations possibles, à savoir si les variables et les constantes sont nulles ou non. On obtient ainsi un *scindage* qu'il est parfois facile de représenter par un arbre [6].

Exemple 2 - Traitons l'exemple du pendule simple en dimension $n = 2$

$$\begin{cases} \ddot{x} = -\lambda x \\ \ddot{y} = \mathfrak{g} - \lambda y \\ 0 = x^2 + y^2 - 1. \end{cases}$$

En raison de la structure des équations à étudier, on souhaite éliminer le multiplicateur de Lagrange λ (on obtiendra ainsi un système différentiel ordinaire sur les variables x et y). On choisit ainsi l'ordre [4] d'élimination

$$y < x < \dot{y} < \dot{x} < \ddot{y} < \ddot{x} < \cdots < \lambda < \dot{\lambda} < \ddot{\lambda} < \cdots$$

Choisissons la première équation $\ddot{x} = -\lambda x$. Si l'on souhaite exprimer la plus grande variable par rapport aux autres, à savoir λ (suivant l'ordre fixé au départ), il faut être en mesure de diviser par x. Une première distinction de cas intervient à ce niveau : $x = 0$ (cas dégénéré) et $x \neq 0$ (cas non dégénéré). Le cas dégénéré conduit au système réduit

$$\begin{cases} 0 = x \\ 0 = y^2 - 1 \\ 0 = \lambda y - \mathfrak{g} \\ 0 \neq y. \end{cases}$$

En omettant les détails techniques, précisons que l'inéquation sur y ne peut être omise (entre autres parce que si $y = 0$, alors $\mathfrak{g} = 0$, ce qui est absurde). Quant au cas non dégénéré

$$\begin{cases} \lambda = -\dfrac{\ddot{x}}{x} \\ \ddot{y} = \mathfrak{g} - \lambda y \\ 0 = x^2 + y^2 - 1 \\ 0 \neq x, \end{cases}$$

il fournit par substitution de $\lambda = -\dfrac{\ddot{x}}{x}$ dans $\ddot{y} = \mathfrak{g} - \lambda y$

$$\begin{cases} \lambda = -\dfrac{\ddot{x}}{x} \\ \ddot{y}x = \mathfrak{g}x + \ddot{x}y \\ 0 = x^2 + y^2 - 1 \\ 0 \neq x. \end{cases}$$

De la même manière que pour le multiplicateur de Lagrange λ, on doit supposer que $y \neq 0$ pour pouvoir exprimer $\ddot{x}$ en fonction des autres variables ($\ddot{x}$ étant la plus grande variable de

4. Nous devrions employer le terme « classement » pour être tout à fait exact.

l'équation $\ddot{y}x = \mathfrak{g}x + \ddot{x}y$). Il y a donc deux nouvelles configurations : $y = 0$ et $y \neq 0$ menant respectivement à

$$\begin{cases} \lambda = -\dfrac{\ddot{x}}{x} \\ 0 = \mathfrak{g}x \\ 0 = x^2 - 1 \\ 0 \neq x \\ 0 = y, \end{cases} \quad \text{et} \quad \begin{cases} \lambda = -\dfrac{\ddot{x}}{x} \\ \ddot{x} = \dfrac{\ddot{y}x - \mathfrak{g}x}{y} \\ 0 = x^2 + y^2 - 1 \\ 0 \neq x \\ 0 \neq y. \end{cases}$$

Remarque 9

Le cas $y = 0$ contient clairement une contradiction.

On poursuit de fait avec le cas $y \neq 0$. En dérivant deux fois l'équation $0 = x^2 + y^2 - 1$ et en y injectant l'expression $\ddot{x} = \dfrac{\ddot{y}x - \mathfrak{g}x}{y}$, on obtient après simplification

$$\ddot{y} + y\dot{x}^2 + y\dot{y}^2 - \mathfrak{g}x^2 = 0.$$

Puisqu'une première dérivation de $0 = x^2 + y^2 - 1$ donne $\dot{x}x + \dot{y}y = 0$, on obtient $\dot{x}^2x^2 = \dot{y}^2y^2$. Ainsi, en multipliant l'expression $\ddot{y} + y\dot{x}^2 + y\dot{y}^2 - \mathfrak{g}x^2 = 0$ par x^2, on est en mesure de faire disparaitre la variable $\dot{x}$:

$$\begin{aligned} x^2\left(\ddot{y} + y\dot{x}^2 + y\dot{y}^2 - \mathfrak{g}x^2\right) = 0 &\Leftrightarrow x^2\ddot{y} + x^2y\dot{x}^2 + x^2y\dot{y}^2 - \mathfrak{g}x^4 = 0 \\ &\Leftrightarrow x^2\ddot{y} + \dot{y}^2y^3 + x^2y\dot{y}^2 - \mathfrak{g}x^4 = 0 \\ &\Leftrightarrow x^2\ddot{y} + \dot{y}^2y\left(x^2 + y^2\right) - \mathfrak{g}x^4 = 0 \\ &\Leftrightarrow x^2\ddot{y} + \dot{y}^2y - \mathfrak{g}x^4 = 0 \\ &\Leftrightarrow \left(1 - y^2\right)\ddot{y} + \dot{y}^2y - \mathfrak{g}\left(1 - y^2\right)^2 = 0 \\ &\Leftrightarrow \ddot{y} = \frac{\mathfrak{g}\left(1 - y^2\right)^2 - \dot{y}^2y}{1 - y^2}. \end{aligned}$$

En combinant les équations précédentes, on finit par obtenir le système réduit

$$\begin{cases} \lambda = \dfrac{\mathfrak{g}y(1 - y^2) + \dot{y}^2}{1 - y^2} \\ \ddot{y} = \dfrac{\mathfrak{g}\left(1 - y^2\right)^2 - \dot{y}^2y}{1 - y^2} \\ x^2 = 1 - y^2. \end{cases}$$

□

En menant les calculs différemment, on trouve $\lambda = \mathfrak{g}y + \dot{x}^2 + \dot{y}^2$ (il suffit pour cela de dériver à deux reprises la contrainte $x^2 + y^2 - 1 = 0$ et d'y injecter les expressions de $\ddot{x}$ et $\ddot{y}$), ce qui fournit le second système

$$\begin{cases} \lambda = \mathfrak{g}y + \dot{x}^2 + \dot{y}^2 \\ \ddot{y} = \mathfrak{g} - \left(\mathfrak{g}y + \dot{x}^2 + \dot{y}^2\right)y \\ x^2 = 1 - y^2, \end{cases}$$

qui est également une chaîne différentielle régulière. À l'heure actuelle, on ne sait pas si une modification de *Rosenfeld-Gröbner* permettrait de trouver ce second système.

1.8 Autres méthodes

Étant particulièrement riche, la littérature au sujet des EDAs propose d'autres approches permettant de traiter ces systèmes différentiels implicites, notamment lorsqu'ils sont linéaires.

Un algorithme de réduction des EDAs linéaires où les coefficients sont holomorphes est proposé par W. A. Harris, Y. Sibuya et L. Weinberg [23], généralisant le point de vue abordé par R. J. Hanson [22]. Le système linéaire (1.15) est découplé en une EDA et des équations algébriques de la forme

$$\begin{cases} E_1(t)\dot{X}_1 = A_1(t)X_1 + A_2(t)X_2 + f_1 \\ 0 = A_3(t)X_1 + A_4(t)X_2 + f_2, \end{cases}$$

où la matrice $E_1(t)$ est inversible. Une étude des coefficients $A_i(t)$ est alors entreprise ; dans un cas favorable (si $A_4(t)$ est inversible), une injection de la partie algébrique dans la partie différentielle est mise en avant, permettant ainsi de résoudre le système. Si $A_4(t) = 0$ ou si $A_4(t) \neq 0$ mais non inversible, on peut se ramener au système de la forme

$$\begin{cases} \bar{E}_1(t)\dot{X}_1 = \bar{A}_1(t)X_1 + \bar{A}_2(t)X_2 + \bar{f}_1 \\ 0 = \bar{A}_3(t)X_1 + \bar{f}_2. \end{cases}$$

Si $\bar{A}_3(t)$ est inversible, on peut exprimer X_1 et il reste à résoudre une équation de la forme

$$\tilde{A}_1 X_2 = \tilde{f}_1.$$

Dans le cas contraire, on répète toute la procédure précédemment décrite tant que $\bar{A}_3(t) \neq 0$. Supposons donc que $\bar{A}_3(t) = 0$; on regarde le problème

$$\bar{E}_1(t)\dot{X}_1 = \bar{A}_1(t)X_1 + \bar{A}_2(t)X_2 + \bar{f}_1. \tag{1.38}$$

Si $\bar{A}_2(t)$ est inversible, on peut exprimer X_2 en fonction du reste (X_1 est pris arbitrairement). Si $\bar{A}_2(t) = 0$, on obtient une EDO. Enfin, si $\bar{A}_2(t)$ n'est pas inversible, on peut se ramener à un problème de la forme (1.38), mais de taille réduite. Cet algorithme considère ainsi toutes les configurations possibles vis-à-vis des matrices coefficients du problème (1.15).

Plus récemment, un nouvel algorithme de réduction motivé par les travaux de W. A. Harris est décrit par M. A. Barkatou, C. El Bacha et E. Pflügel [4], permettant entre autres la classification des singularités de l'EDA.

La diversité des techniques permettant d'appréhender les EDAs illustre notamment la richesse mathématique de ces objets en même temps que leurs subtilités. Dans le chapitre suivant, nous présentons une nouvelle méthode de réduction tirant avantage d'une structure générale simple et applicable à la fois aux EDAs linéaires et quasi-linéaires.

Chapitre 2

Méthode de réduction de l'indice par déflation

Ce chapitre met en lumière une nouvelle méthode de réduction des EDAs linéaires et quasi-linéaires [32] et [33]. Nous mettons au point un algorithme formel capable de réduire, dans certains cas, ces systèmes différentiels algébriques. Cet algorithme, nommé *méthode de déflation*, est un processus symbolique itératif dont le principe consiste à déterminer une suite d'EDAs de tailles strictement décroissantes (le terme « déflation » qualifie cette décroissance des tailles). Il existe deux alternatives envisageables à la fin de ce procédé : soit on obtient un système différentiel ordinaire auquel s'ajoutent des équations algébriques, soit on parvient à un système uniquement composé d'équations algébriques. Le cœur de l'algorithme de déflation est un jeu de dérivation des contraintes algébriques et de substitution de certaines variables adéquates. Cette méthode de déflation a été initialement introduite dans le but de résoudre des EDAs quasi-linéaires intervenant dans le domaine du génie des procédés et modélisant entre autres les phénomènes de distillation. Nous présentons l'étude de ces cas dans le troisième chapitre de cette thèse.

Nous décrivons dans la suite la méthode de déflation, ainsi que tous les résultats s'y rapportant tels que la baisse de l'indice de Kronecker à chaque étape de l'algorithme pour les EDAs linéaires à coefficients constants, ou le caractère géométrique de la méthode. Nous appliquons l'algorithme à de nombreux exemples dont celui du pendule simple en dimension n.

2.1 EDAs linéaires à coefficients constants

2.1.1 Mécanisme de déflation

Nous commençons par étudier le problème linéaire à coefficients constants, i.e. de la forme

$$E\dot{X} = AX + f. \tag{2.1}$$

Notre objectif étant de séparer la partie différentielle de la partie algébrique, il convient de trouver une forme canonique générale de la paire matricielle (E, A). En écrivant $E = G\begin{bmatrix} F \\ 0 \end{bmatrix}$ et $A = G\begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$, où $G \in \mathbb{R}^{n\times n}$ est une matrice inversible, $F \in \mathbb{R}^{r\times n}$ est telle que rang F = rang $E = r$, $A_1 \in \mathbb{R}^{r\times n}$ et $A_2 \in \mathbb{R}^{(n-r)\times n}$, on a

$$\lambda E + A = \lambda G\begin{bmatrix} F \\ 0 \end{bmatrix} + G\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = G\begin{bmatrix} \lambda F + A_1 \\ A_2 \end{bmatrix}. \tag{2.2}$$

Si la paire matricielle est régulière pour un λ donné (i.e. si $\lambda E + A$ est inversible pour ce λ), on voit immédiatement que la matrice A_2 est nécessairement de rang plein.

Proposition 5

Soient E et A deux matrices carrées de taille n telles que rang $E = r$. Si le problème (2.1) est régulier, alors

$$\lambda E + A = G \begin{bmatrix} \lambda S + H & \lambda T + K \\ M & N \end{bmatrix} P^{-1}, \tag{2.3}$$

où $G \in \mathbb{R}^{n\times n}$ est une matrice inversible, $P \in \mathbb{R}^{n\times n}$ est une matrice de permutation, S et $H \in \mathbb{R}^{r\times r}$, T et $K \in \mathbb{R}^{r\times(n-r)}$, $M \in \mathbb{R}^{(n-r)\times r}$ et $N \in \mathbb{R}^{(n-r)\times(n-r)}$, où N est une matrice inversible.

Remarque 10

En d'autres termes, on extrait un bloc matriciel inversible N de la matrice A_2 par permutations des colonnes de cette dernière. Notons également que cette décomposition n'est pas unique.

Preuve - Il suffit de procéder aux transformations sur les matrices en utilisant les notations issues de (2.2) :

$$\begin{aligned} \lambda E + A &= G \left(\lambda \begin{bmatrix} F \\ 0 \end{bmatrix} + \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \right) \\ &= G \left(\lambda \begin{bmatrix} F \\ 0 \end{bmatrix} + \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \right) PP^{-1} \\ &= G \left(\lambda \begin{bmatrix} S & T \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} H & K \\ M & N \end{bmatrix} \right) P^{-1} \\ &= G \begin{bmatrix} \lambda S + H & \lambda T + K \\ M & N \end{bmatrix} P^{-1}. \end{aligned}$$

□

Définition 5 (*Transformation régulière*)

La forme de la matrice $\lambda E + A$ exhibée dans (2.3) est appelée transformation régulière *de (E, A) et par extension de (2.1).*

La transformation régulière de (2.1) donne

$$\begin{bmatrix} S & T \\ 0 & 0 \end{bmatrix} P^{-1}\dot{X} = \begin{bmatrix} H & K \\ M & N \end{bmatrix} P^{-1}X + \begin{pmatrix} g \\ h \end{pmatrix}, \tag{2.4}$$

où $\begin{pmatrix} g \\ h \end{pmatrix} = G^{-1}f$.

Remarque 11

Il est possible de faire un parallèle avec (1.12). Nous utilisons davantage le caractère de régularité de (E, A) en extrayant un bloc inversible N, tandis que la matrice A_4 présente dans (1.12) n'est pas nécessairement inversible.

En posant $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} = P^{-1}X$, on développe (2.4) :

$$\begin{bmatrix} S & T \\ 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{X}_1 \\ \dot{Y}_1 \end{pmatrix} = \begin{bmatrix} H & K \\ M & N \end{bmatrix} \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} + \begin{pmatrix} g \\ h \end{pmatrix} \Leftrightarrow \begin{cases} S\dot{X}_1 + T\dot{Y}_1 = HX_1 + KY_1 + g \\ 0 = MX_1 + NY_1 + h. \end{cases}$$

Puisque la matrice N est inversible, on a $Y_1 = -N^{-1}(MX_1 + h)$. Ainsi,

$$\begin{aligned} & S\dot{X}_1 + T\dot{Y}_1 = HX_1 + KY_1 + g \\ \Leftrightarrow\ & S\dot{X}_1 - TN^{-1}\left(M\dot{X}_1 + \dot{h}\right) = HX_1 - KN^{-1}(MX_1 + h) + g \\ \Leftrightarrow\ & \left(S - TN^{-1}M\right)\dot{X}_1 = \left(H - KN^{-1}M\right)X_1 + TN^{-1}\dot{h} - KN^{-1}h + g. \end{aligned}$$

On note

$$E_1 = S - TN^{-1}M, \quad A_1 = H - KN^{-1}M \quad \text{et} \quad f_1 = TN^{-1}\dot{h} - KN^{-1}h + g.$$

Via le changement de variable $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} = P^{-1}X$, le problème régulier (2.1) est équivalent à

$$\begin{cases} E_1\dot{X}_1 = A_1X_1 + f_1 & \text{(2.5a)} \\ \quad Y_1 = -N^{-1}(MX_1 + h). & \text{(2.5b)} \end{cases}$$

La transformation régulière de (2.1) permet ainsi d'obtenir le système (2.5) composé de l'EDA (2.5a) et des contraintes algébriques (2.5b). L'équation (2.5a) est une EDA linéaire à coefficients constants, de taille r inférieure à celle de (2.1).

Définition 6 (*EDA linéaire à coefficients constants déflatée*)

L'EDA (2.5a) *obtenue en opérant la transformation régulière est appelée* EDA déflatée, *relativement à l'EDA régulière* (2.1).

Remarque 12

L'EDA déflatée (2.5a) *n'est pas unique car elle dépend de la transformation régulière.*

Le passage de l'EDA (2.1) au système (2.5) constitue le mécanisme principal de la méthode de déflation.

2.1.2 Algorithme de déflation

L'algorithme de déflation prend en entrée la paire matricielle (E, A), la non-homogénéité f, le vecteur inconnu X ainsi que la taille du système n. Il effectue la décomposition générale (2.2) de manière à tester la régularité de la paire matricielle (E_j, A_j) courante. Dans l'affirmative, il transforme (2.1) en (2.5), passant ainsi d'une matrice coefficient E_j à une matrice coefficient E_{j+1}. Il se dessine alors trois possibilités :

1. La matrice E_{j+1} est singulière ; (2.5a) est une EDA et l'algorithme se poursuit.
2. La matrice E_{j+1} est inversible ; (2.5a) est une EDO et l'algorithme s'achève.
3. La matrice E_{j+1} est nulle ; (2.5a) est un système algébrique et l'algorithme s'achève.

Une fois achevé, l'algorithme de déflation fournit soit une EDO et un ensemble de contraintes, soit uniquement des équations algébriques. Il est présenté sous le nom `LTI_Deflation`[1].

Remarque 13

La troisième configuration est beaucoup plus rare que les autres. En effet, un tel système décrit uniquement par des équations algébriques est stationnaire. Par conséquent, sa formulation initiale contenant des dérivées est très certainement inappropriée. En revanche, il existe des exemples mathématiques illustrant cette troisième configuration.

1. Pour « Linear Time-Invariant Deflation ».

`LTI_Deflation`
begin
Input : $E_0 = E$, $A_0 = A$, $f_0 = f$, $X_0 = X$, $r_{-1} = n$.
Step $j+1$, $j \geqslant 0$:
if *E_j est singulière,*
then
Calculer le rang r_j de E_j.
Déterminer les expressions des matrices G_j, F_j, A_{1j} et A_{2j} exhibées dans la factorisation (2.2), relativement aux matrices E_j et A_j.
if *A_{2j} n'est pas de rang plein,*
then
Stop
else
Déterminer les expressions des matrices P_j, S_j, T_j, H_j, K_j, M_j et N_j exhibées dans la transformation régulière (2.3) de (E_j, A_j).
Déterminer le triplet $(E_{j+1}, A_{j+1}, f_{j+1})$ à partir des formules précédentes.
Effectuer le changement de variable $\begin{pmatrix} X_{j+1} \\ Y_{j+1} \end{pmatrix} = P_j^{-1} X_j$, où la dimension de X_{j+1} est égale à r_j.
Déterminer les contraintes algébriques $0 = M_j X_{j+1} + N_j Y_{j+1} + h_j$.
end
else
Stop
end
Output : Soit k le nombre d'étapes de l'algorithme.
if *(E_j, A_j) est régulière pour tout $j \in [\![0, k-1]\!]$,*
then
if *E_k est inversible*
then
Fournir

$$\begin{cases} \dot{X}_k = E_k^{-1}\left(A_k X_k + f_k\right) \\ 0 = M_{k-1} X_k + N_{k-1} Y_k + h_{k-1} \\ \quad \vdots \\ 0 = M_0 X_1 + N_0 Y_1 + h_0. \end{cases}$$

else
Fournir

$$\begin{cases} 0 = A_k X_k + f_k \\ 0 = M_{k-1} X_k + N_{k-1} Y_k + h_{k-1} \\ \quad \vdots \\ 0 = M_0 X_1 + N_0 Y_1 + h_0. \end{cases}$$

end
else
Le système possède une infinité de solutions ou bien il n'en possède aucune.
end
end

2.1.3 Propriétés de l'algorithme

Analysons à présent différentes propriétés qui découlent de l'algorithme dans le contexte linéaire à coefficients constants. La notation « $\sim$ » introduite dans le chapitre précédent pour les paires matricielles est employée pour des matrices ; deux matrices M_1 et M_2 sont équivalentes s'il existe deux matrices de passage R_1 et R_2 telles que $M_1 = R_1 M_2 R_2$. On note $M_1 \sim M_2$.

2.1.3.1 Transmission de la régularité

Théorème 8

Si le problème (2.1) est régulier, alors L'EDA déflatée (2.5a) est également régulière.

Preuve - Puisque le problème (2.1) est régulier, il existe par définition un réel λ tel que la matrice $\lambda E + A$ soit inversible. En utilisant la proposition 5, on a

$$\lambda E + A \sim \begin{bmatrix} \lambda S + H & \lambda T + K \\ M & N \end{bmatrix}.$$

Or, la matrice N est inversible. Par conséquent, on peut utiliser la décomposition relative au complément de Schur par rapport à la matrice N :

$$\begin{aligned}
&\begin{bmatrix} \lambda S + H & \lambda T + K \\ M & N \end{bmatrix} \\
&= \begin{bmatrix} I_r & (\lambda T + K) N^{-1} \\ 0 & I_{n-r} \end{bmatrix} \begin{bmatrix} \lambda S + H - (\lambda T + K) N^{-1} M & 0 \\ 0 & N \end{bmatrix} \begin{bmatrix} I_r & 0 \\ N^{-1} M & I_{n-r} \end{bmatrix} \\
&= \begin{bmatrix} I_r & (\lambda T + K) N^{-1} \\ 0 & I_{n-r} \end{bmatrix} \begin{bmatrix} \lambda S - \lambda T N^{-1} M + H - K N^{-1} M & 0 \\ 0 & N \end{bmatrix} \begin{bmatrix} I_r & 0 \\ N^{-1} M & I_{n-r} \end{bmatrix} \\
&= \begin{bmatrix} I_r & (\lambda T + K) N^{-1} \\ 0 & I_{n-r} \end{bmatrix} \begin{bmatrix} \lambda E_1 + A_1 & 0 \\ 0 & N \end{bmatrix} \begin{bmatrix} I_r & 0 \\ N^{-1} M & I_{n-r} \end{bmatrix}.
\end{aligned}$$

On obtient de ce fait $\lambda E + A \sim \begin{bmatrix} \lambda E_1 + A_1 & 0 \\ 0 & N \end{bmatrix}$. Ainsi $\det(\lambda E + A) = \det(\lambda E_1 + A_1) \det N$. On en déduit que la matrice $\lambda E_1 + A_1$ est inversible ; l'EDA (2.5a) est régulière. □

La propriété de régularité se transmet ainsi d'une EDA linéaire à coefficients constants à une EDA déflatée. Par conséquent, si la première paire de matrices est régulière, alors toutes les paires matricielles obtenues en chaîne par la méthode de déflation le seront également.

2.1.3.2 Invariance du rang

La transformation régulière (2.3) utilisée par l'algorithme préserve les rangs matriciels.

Théorème 9

On considère deux EDAs déflatées du problème régulier (2.1), représentées respectivement par les paires matricielles (E_1, A_1) et $\left(\tilde{E}_1, \tilde{A}_1\right)$. Alors les relations

$$\operatorname{rang} E_1 = \operatorname{rang} \tilde{E}_1 \quad \text{et} \quad \operatorname{rang} A_1 = \operatorname{rang} \tilde{A}_1$$

sont satisfaites.

Preuve - On suppose qu'il existe deux transformations régulières différentes donnant d'une part pour la matrice E :

$$E = GG^{-1}EPP^{-1} = G\begin{bmatrix} S & T \\ 0 & 0 \end{bmatrix} P^{-1} \quad \text{et} \quad E = \tilde{G}\tilde{G}^{-1}E\tilde{P}\tilde{P}^{-1} = \tilde{G}\begin{bmatrix} \tilde{S} & \tilde{T} \\ 0 & 0 \end{bmatrix} \tilde{P}^{-1}.$$

Autrement dit,

$$\begin{bmatrix} S & T \\ 0 & 0 \end{bmatrix} \sim \begin{bmatrix} \tilde{S} & \tilde{T} \\ 0 & 0 \end{bmatrix}. \tag{2.6}$$

D'autre part,

$$A = GG^{-1}APP^{-1} = G\begin{bmatrix} K & H \\ M & N \end{bmatrix} P^{-1} \quad \text{et} \quad A = \tilde{G}\tilde{G}^{-1}A\tilde{P}\tilde{P}^{-1} = \tilde{G}\begin{bmatrix} \tilde{K} & \tilde{H} \\ \tilde{M} & \tilde{N} \end{bmatrix} \tilde{P}^{-1},$$

ce qui donne

$$\begin{bmatrix} K & H \\ M & N \end{bmatrix} \sim \begin{bmatrix} \tilde{K} & \tilde{H} \\ \tilde{M} & \tilde{N} \end{bmatrix}. \tag{2.7}$$

En utilisant (2.6) et (2.7), on montre que

$$\begin{bmatrix} S & T \\ M & N \end{bmatrix} \sim \begin{bmatrix} \tilde{S} & \tilde{T} \\ \tilde{M} & \tilde{N} \end{bmatrix}.$$

Puisque les matrices N et $\tilde{N}$ sont inversibles, on obtient en utilisant le complément de Schur

$$\begin{bmatrix} E_1 & 0 \\ 0 & N \end{bmatrix} \sim \begin{bmatrix} \tilde{E}_1 & 0 \\ 0 & \tilde{N} \end{bmatrix}$$

et

$$\begin{bmatrix} A_1 & 0 \\ 0 & N \end{bmatrix} \sim \begin{bmatrix} \tilde{A}_1 & 0 \\ 0 & \tilde{N} \end{bmatrix},$$

ce qui se traduit par $\operatorname{rang} E_1 + \operatorname{rang} N = \operatorname{rang} \tilde{E}_1 + \operatorname{rang} \tilde{N}$ et $\operatorname{rang} A_1 + \operatorname{rang} N = \operatorname{rang} \tilde{A}_1 + \operatorname{rang} \tilde{N}$. Or, les matrices N et $\tilde{N}$ sont de plus de même taille ; elles sont par conséquent de même rang. On parvient ainsi à $\operatorname{rang} E_1 = \operatorname{rang} \tilde{E}_1$ et $\operatorname{rang} A_1 = \operatorname{rang} \tilde{A}_1$. □

2.1.3.3 Réduction de l'indice

Dans le contexte linéaire à coefficients constants, nous employons le terme *indice* pour désigner l'indice de Kronecker de la paire matricielle (E, A).

Théorème 10

L'indice de la paire matricielle (E, A) relative au problème (2.1) est égal à 1 si et seulement si la matrice E_1 obtenue dans l'EDA déflatée (2.5a) est inversible.

Preuve - On suppose pour débuter que la paire matricielle (E, A) est d'indice 1. Soit $x \in \mathbb{R}^n$ tel que $x \in \ker \begin{bmatrix} S & T \\ M & N \end{bmatrix}$, où les matrices S, T, M et N sont définies par la transformation

régulière (2.3). On a

$$\begin{aligned}\begin{bmatrix} S & T \\ M & N \end{bmatrix} x = 0 &\Leftrightarrow \left(\begin{bmatrix} S & T \end{bmatrix} x \right) = 0 \\ &\Leftrightarrow \begin{cases} \begin{bmatrix} S & T \end{bmatrix} x = 0 \\ \begin{bmatrix} M & N \end{bmatrix} x = 0 \end{cases} \\ &\Leftrightarrow \begin{cases} x \in \ker \begin{bmatrix} S & T \end{bmatrix} \\ x \in \ker \begin{bmatrix} M & N \end{bmatrix} \end{cases} \\ &\Leftrightarrow \begin{cases} x \in \ker \left(G^{-1} E P \right) \\ x \in \ker \begin{bmatrix} M & N \end{bmatrix} . \end{cases}\end{aligned}$$

Ainsi, $G^{-1}APx = \begin{bmatrix} H & K \\ M & N \end{bmatrix} x = \left(\begin{bmatrix} H & K \end{bmatrix} x \right)$. On note $\Theta_E = G^{-1}EP$. Puisque rang $\Theta_E =$ rang $E = r$, il existe $y \in \mathbb{R}^n$ tel que $\left(\begin{bmatrix} H & K \end{bmatrix} x \right) = \Theta_E y$. On obtient alors

$$G^{-1}APx = \Theta_E y \in \operatorname{im} \Theta_E.$$

En posant $z = Px$, on a d'une part

$$\begin{aligned} x \in \ker \Theta_E &\Leftrightarrow \Theta_E x = 0 \\ &\Leftrightarrow EPx = 0 \\ &\Leftrightarrow Ez = 0 \\ &\Leftrightarrow z \in \ker E. \end{aligned}$$

D'autre part,

$$\begin{aligned} G^{-1}APx \in \operatorname{im} \Theta_E &\Leftrightarrow G^{-1}APx = \Theta_E y \\ &\Leftrightarrow APx = EPy \\ &\Leftrightarrow Az = EPy \\ &\Leftrightarrow Az \in \operatorname{im} E. \end{aligned}$$

D'après la proposition 2, on en déduit que $z = 0$ et par conséquent $x = 0$. Ceci démontre que la matrice $\begin{bmatrix} S & T \\ M & N \end{bmatrix}$ est inversible, de même que son complément de Schur E_1.

Réciproquement, on suppose que la matrice E_1 est inversible, ce qui revient à supposer que $\begin{bmatrix} S & T \\ M & N \end{bmatrix}$ est non-singulière. Clairement, la paire matricielle (E, A) est équivalente à la paire matricielle $\left(G^{-1}EP, G^{-1}AP\right)$, autrement dit à la paire matricielle $\left(\Theta_E, G^{-1}AP\right)$. Notre but est de montrer l'implication suivante :

$$\begin{cases} x \in \ker \Theta_E \\ G^{-1}APx \in \operatorname{im} \Theta_E \end{cases} \Rightarrow x = 0.$$

Or, $\Theta_E x = 0$ implique que $\begin{bmatrix} S & T \end{bmatrix} x = 0$ et $G^{-1}APx \in \operatorname{im} \Theta_E$ implique que $\begin{bmatrix} M & N \end{bmatrix} x = 0$, ce qui se traduit par $\begin{bmatrix} S & T \\ M & N \end{bmatrix} x = 0$. La matrice $\begin{bmatrix} S & T \\ M & N \end{bmatrix}$ étant inversible, on a $x = 0$ ce qui signifie que l'indice de la paire matricielle (E, A) vaut 1. □

Le théorème 10 se généralise pour n'importe quel indice. Le passage entre l'indice 1 et l'indice 0 (une matrice inversible est d'indice 0) n'est que le reflet d'un mécanisme plus global. Avant de démontrer cette baisse générale de l'indice, montrons les deux résultats techniques suivants.

Proposition 6

On considère la matrice par blocs $\begin{bmatrix} A & B \\ CA & CB \end{bmatrix}$. *Pour tout* $k \geqslant 1$, *on a*

$$\begin{bmatrix} A & B \\ CA & CB \end{bmatrix}^k = \begin{bmatrix} (A+BC)^{k-1}A & (A+BC)^{k-1}B \\ C(A+BC)^{k-1}A & C(A+BC)^{k-1}B \end{bmatrix}.$$

Preuve - Le cas $k = 1$ est évident. Pour le cas général, supposons qu'il existe un entier p tel que

$$\begin{bmatrix} A & B \\ CA & CB \end{bmatrix}^p = \begin{bmatrix} (A+BC)^{p-1}A & (A+BC)^{p-1}B \\ C(A+BC)^{p-1}A & C(A+BC)^{p-1}B \end{bmatrix}.$$

En post-multipliant par la matrice $\begin{bmatrix} A & B \\ CA & CB \end{bmatrix}$, on obtient

$$\begin{aligned} \begin{bmatrix} A & B \\ CA & CB \end{bmatrix}^{p+1} &= \begin{bmatrix} (A+BC)^{p-1}A & (A+BC)^{p-1}B \\ C(A+BC)^{p-1}A & C(A+BC)^{p-1}B \end{bmatrix} \begin{bmatrix} A & B \\ CA & CB \end{bmatrix} \\ &= \begin{bmatrix} (A+BC)^{p-1}\left(A^2+BCA\right) & (A+BC)^{p-1}\left(AB+BCB\right) \\ C(A+BC)^{p-1}\left(A^2+BCA\right) & C(A+BC)^{p-1}\left(AB+BCB\right) \end{bmatrix} \\ &= \begin{bmatrix} (A+BC)^{p}A & (A+BC)^{p}B \\ C(A+BC)^{p}A & C(A+BC)^{p}B \end{bmatrix}. \end{aligned}$$

La propriété est héréditaire ; le raisonnement par récurrence nous fournit le résultat escompté. □

Proposition 7

On considère les notations issues de la transformation régulière (2.3) *du problème* (2.1) *et on pose* $C_j = \lambda E_j + A_j$ *pour* $j \in \mathbb{N}^\star$. *Pour tout* $k \geqslant 1$, *on a*

$$\begin{aligned} &\left(P_j^{-1}C_j^{-1}E_jP_j\right)^k \\ &= \begin{bmatrix} \left(C_{j+1}^{-1}E_{j+1}\right)^{k-1}C_{j+1}^{-1}S_j & \left(C_{j+1}^{-1}E_{j+1}\right)^{k-1}C_{j+1}^{-1}T_j \\ -N_j^{-1}M_j\left(C_{j+1}^{-1}E_{j+1}\right)^{k-1}C_{j+1}^{-1}S_j & -N_j^{-1}M_j\left(C_{j+1}^{-1}E_{j+1}\right)^{k-1}C_{j+1}^{-1}T_j \end{bmatrix}. \end{aligned}$$

Preuve - Commençons par le cas $k = 1$.

$$\begin{aligned} &P_j^{-1}C_j^{-1}E_jP_j \\ &= P_j^{-1}\left(\lambda G_jG_j^{-1}E_jP_jP_j^{-1} + G_jG_j^{-1}A_jP_jP_j^{-1}\right)^{-1}E_jP_j \\ &= P_j^{-1}P_j\left(\lambda\begin{bmatrix} S_j & T_j \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} H_j & K_j \\ M_j & N_j \end{bmatrix}\right)^{-1}G_j^{-1}E_jP_j \end{aligned}$$

$$
\begin{aligned}
&= \begin{bmatrix} \lambda S_j + H_j & \lambda T_j + K_j \\ M_j & N_j \end{bmatrix}^{-1} \begin{bmatrix} S_j & T_j \\ 0 & 0 \end{bmatrix} \\
&= \left(\begin{bmatrix} I & (\lambda T_j + K_j) N_j^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \lambda E_{j+1} + A_{j+1} & 0 \\ 0 & N_j \end{bmatrix} \begin{bmatrix} I & 0 \\ N_j^{-1} M_j & I \end{bmatrix} \right)^{-1} \begin{bmatrix} S_j & T_j \\ 0 & 0 \end{bmatrix} \\
&= \left(\begin{bmatrix} I & (\lambda T_j + K_j) N_j^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} C_{j+1} & 0 \\ 0 & N_j \end{bmatrix} \begin{bmatrix} I & 0 \\ N_j^{-1} M_j & I \end{bmatrix} \right)^{-1} \begin{bmatrix} S_j & T_j \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} I & 0 \\ N_j^{-1} M_j & I \end{bmatrix}^{-1} \begin{bmatrix} C_{j+1} & 0 \\ 0 & N_j \end{bmatrix}^{-1} \begin{bmatrix} I & (\lambda T_j + K_j) N_j^{-1} \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} S_j & T_j \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} I & 0 \\ -N_j^{-1} M_j & I \end{bmatrix} \begin{bmatrix} C_{j+1}^{-1} & 0 \\ 0 & N_j^{-1} \end{bmatrix} \begin{bmatrix} I & -(\lambda T_j + K_j) N_j^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} S_j & T_j \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} I & 0 \\ -N_j^{-1} M_j & I \end{bmatrix} \begin{bmatrix} C_{j+1}^{-1} & 0 \\ 0 & N_j^{-1} \end{bmatrix} \begin{bmatrix} S_j & T_j \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} I & 0 \\ -N_j^{-1} M_j & I \end{bmatrix} \begin{bmatrix} C_{j+1}^{-1} S_j & C_{j+1}^{-1} T_j \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} C_{j+1}^{-1} S_j & C_{j+1}^{-1} T_j \\ -N_j^{-1} M_j C_{j+1}^{-1} S_j & -N_j^{-1} M_j C_{j+1}^{-1} T_j \end{bmatrix}.
\end{aligned}
$$

Quant au cas $k \geqslant 2$, il suffit d'appliquer la proposition 6. □

On est à présent capable de démontrer la généralisation du théorème 10.

Théorème 11

On considère une EDA déflatée, représentée par la paire matricielle (E_{j+1}, A_{j+1}), d'une EDA régulière représentée par la paire matricielle (E_j, A_j), pour $j \in \mathbb{N}$. Si E_{j+1} est différente de la matrice nulle, alors l'indice de Kronecker de (E_{j+1}, A_{j+1}) est égal à celui de (E_j, A_j) diminué de 1.

Preuve - Par commodité, l'indice de Kronecker de la paire matricielle (E_j, A_j) est noté ν_{K_j}. Le cas $\nu_{K_j} = 1$ est traité par le théorème 10. Supposons donc maintenant que $\nu_{K_j} \geqslant 2$. On pose

$$
\begin{aligned}
\Theta_{E_j} &= G_j^{-1} E_j P_j \\
\Theta_{A_j} &= G_j^{-1} A_j P_j \\
\mathcal{E}_j &= (\lambda E_j + A_j)^{-1} E_j.
\end{aligned}
$$

Puisque $\mathcal{E}_j = P_j \left(\lambda \Theta_{E_j} + \Theta_{A_j}\right)^{-1} \Theta_{E_j} P_j^{-1}$, on observe que les paires de matrices (E_j, A_j) et $\left(\Theta_{E_j}, \Theta_{A_j}\right)$ sont équivalentes et qu'ainsi elles ont même indice. La matrice $\left(\lambda \Theta_{E_j} + \Theta_{A_j}\right)^{-1} \Theta_{E_j}$ est alors d'indice ν_{K_j} et vérifie

$$
\ker \left(\left(\lambda \Theta_{E_j} + \Theta_{A_j}\right)^{-1} \Theta_{E_j} \right)^{\nu_{K_j}} \oplus \operatorname{im} \left(\left(\lambda \Theta_{E_j} + \Theta_{A_j}\right)^{-1} \Theta_{E_j} \right)^{\nu_{K_j}} = \mathbb{R}^n. \tag{2.8}
$$

Prouver que $\nu_{K_{j+1}} = \nu_{K_j} - 1$ revient à montrer l'implication suivante

$$
x \in \ker \mathcal{E}_{j+1}^{\nu_{K_j}-1} \cap \operatorname{im} \mathcal{E}_{j+1}^{\nu_{K_j}-1} \Rightarrow x = 0. \tag{2.9}
$$

D'une part, si $x \in \ker \mathcal{E}_{j+1}^{\nu_{K_j}-1}$, alors

$$\begin{aligned}
\mathcal{E}_{j+1}^{\nu_{K_j}-1} x = 0 &\Leftrightarrow \mathcal{E}_{j+1}^{\nu_{K_j}-2} \mathcal{E}_{j+1} x = 0 \\
&\Leftrightarrow \mathcal{E}_{j+1}^{\nu_{K_j}-2} \left(\lambda E_{j+1} + A_{j+1}\right)^{-1} E_{j+1} x = 0 \\
&\Leftrightarrow \mathcal{E}_{j+1}^{\nu_{K_j}-2} \left(\lambda E_{j+1} + A_{j+1}\right)^{-1} \left(S_j - T_j N_j^{-1} M_j\right) x = 0 \\
&\Leftrightarrow \mathcal{E}_{j+1}^{\nu_{K_j}-2} \left(\lambda E_{j+1} + A_{j+1}\right)^{-1} \left(S_j x - T_j N_j^{-1} M_j x\right) = 0.
\end{aligned}$$

Par simple produit, on obtient

$$\mathcal{E}_{j+1}^{\nu_{K_j}-1} \left(\lambda E_{j+1} + A_{j+1}\right)^{-1} \left(S_j x - T_j N_j^{-1} M_j x\right) = 0.$$

En notant $C_{j+1} = \lambda E_{j+1} + A_{j+1}$, la proposition 7 donne

$$\begin{aligned}
&\left(\left(\lambda \Theta_{E_j} + \Theta_{A_j}\right)^{-1} \Theta_{E_j}\right)^{\nu_{K_j}} \begin{pmatrix} x \\ -N_j^{-1} M_j x \end{pmatrix} \\
&= \begin{bmatrix} \left(C_{j+1}^{-1} E_{j+1}\right)^{k-1} C_{j+1}^{-1} S_j & \left(C_{j+1}^{-1} E_{j+1}\right)^{k-1} C_{j+1}^{-1} T_j \\ -N_j^{-1} M_j \left(C_{j+1}^{-1} E_{j+1}\right)^{k-1} C_{j+1}^{-1} S_j & -N_j^{-1} M_j \left(C_{j+1}^{-1} E_{j+1}\right)^{k-1} C_{j+1}^{-1} T_j \end{bmatrix} \begin{pmatrix} x \\ -N_j^{-1} M_j x \end{pmatrix} \\
&= \begin{bmatrix} \mathcal{E}_{j+1}^{\nu_{K_j}-1} C_{j+1}^{-1} S_j & \mathcal{E}_{j+1}^{\nu_{K_j}-1} C_{j+1}^{-1} T_j \\ -N_j^{-1} M_j \mathcal{E}_{j+1}^{\nu_{K_j}-1} C_{j+1}^{-1} S_j & -N_j^{-1} M_j \mathcal{E}_{j+1}^{\nu_{K_j}-1} C_{j+1}^{-1} T_j \end{bmatrix} \begin{pmatrix} x \\ -N_j^{-1} M_j x \end{pmatrix} \\
&= \begin{pmatrix} \mathcal{E}_{j+1}^{\nu_{K_j}-1} \left(\lambda E_{j+1} + A_{j+1}\right)^{-1} \left(S_j x - T_j N_j^{-1} M_j x\right) \\ -N_j^{-1} M_j \mathcal{E}_{j+1}^{\nu_{K_j}-1} \left(\lambda E_{j+1} + A_{j+1}\right)^{-1} \left(S_j x - T_j N_j^{-1} M_j x\right) \end{pmatrix} \\
&= 0.
\end{aligned}$$

Ainsi,

$$\begin{pmatrix} x \\ -N_j^{-1} M_j x \end{pmatrix} \in \ker \left(\left(\lambda \Theta_{E_j} + \Theta_{A_j}\right)^{-1} \Theta_{E_j}\right)^{\nu_{K_j}}. \tag{2.10}$$

D'autre part, si $x \in \operatorname{im} \mathcal{E}_{j+1}^{\nu_{K_j}-1}$, alors il existe $y \in \mathbb{R}^{r_j}$, où $r_j \in \mathbb{N}^\star$ tel que $x = \mathcal{E}_{j+1}^{\nu_{K_j}-1} y$. Puisque le rang de Θ_{E_j} vaut r_j, il existe deux vecteurs $u \in \mathbb{R}^{r_j}$ et $v \in \mathbb{R}^{r_{j-1}-r_j}$, où r_{j-1} correspond à la dimension de Θ_{E_j}, tels que $y = \left(\lambda E_{j+1} + A_{j+1}\right)^{-1} \left(S_j u + T_j u\right)$. Alors, $x = \mathcal{E}_{j+1}^{\nu_{K_j}-1} \left(\lambda E_{j+1} + A_{j+1}\right)^{-1} \left(S_j u + T_j u\right)$. La proposition 7 fournit également

$$\begin{aligned}
&\left(\left(\lambda \Theta_{E_j} + \Theta_{A_j}\right)^{-1} \Theta_{E_j}\right)^{\nu_{K_j}} \begin{pmatrix} u \\ v \end{pmatrix} \\
&= \begin{bmatrix} \left(C_{j+1}^{-1} E_{j+1}\right)^{k-1} C_{j+1}^{-1} S_j & \left(C_{j+1}^{-1} E_{j+1}\right)^{k-1} C_{j+1}^{-1} T_j \\ -N_j^{-1} M_j \left(C_{j+1}^{-1} E_{j+1}\right)^{k-1} C_{j+1}^{-1} S_j & -N_j^{-1} M_j \left(C_{j+1}^{-1} E_{j+1}\right)^{k-1} C_{j+1}^{-1} T_j \end{bmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \\
&= \begin{pmatrix} x \\ -N_j^{-1} M_j x \end{pmatrix},
\end{aligned}$$

c'est-à-dire

$$\begin{pmatrix} x \\ -N_j^{-1} M_j x \end{pmatrix} \in \operatorname{im} \left(\left(\lambda \Theta_{E_j} + \Theta_{A_j}\right)^{-1} \Theta_{E_j}\right)^{\nu_{K_j}}. \tag{2.11}$$

Les expressions (2.10), (2.11) et (2.8) assure que $\begin{pmatrix} x \\ -N_j^{-1} M_j x \end{pmatrix} = 0$, et par conséquent $x = 0$. La propriété (2.9) est alors satisfaite. □

Un parallèle important doit être souligné à ce stade de l'étude. En effet, l'EDA déflatée (2.5a) correspond à la réduction via le complément de Schur de l'EDA (1.13). Si la matrice A_4 est inversible dans (1.13), alors en développant le système par rapport aux matrices par blocs, on parvient à l'EDA déflatée (2.5a).

2.1.3.4 Borne du nombre d'étapes de la méthode

Une des particularités de la méthode de déflation réside dans le lien entre le rang des matrices et la taille des systèmes réduits. Concrètement, si r_j désigne le rang de la matrice E_j, alors r_j est par définition la dimension de E_{j+1}, autrement dit du système déflaté. On s'attend ainsi à ce que le(s) rang(s) matriciel(s) intervienne(nt) dans la borne du nombre d'étapes de la méthode.

Proposition 8

On considère une EDA déflatée, représentée par la paire matricielle (E_{j+1}, A_{j+1}), d'une EDA régulière représentée par la paire matricielle (E_j, A_j), pour $j \in \mathbb{N}$. On note ν_{K_j} l'indice de (E_j, A_j). On a les trois propriétés suivantes

1. *Si $\nu_{K_j} > 1$ alors rang $E_j >$ rang E_{j+1}.*
2. *Si $\nu_{K_j} = 1$ alors rang $E_j =$ rang E_{j+1}.*
3. *Si $\nu_{K_j} > 1$ et rang $E_j = 1$ alors $E_{j+1} = 0$.*

Preuve - Démontrons les trois propriétés.

1. On suppose que $\nu_{K_j} > 1$. Par construction, $r_j = \text{rang } E_j \geqslant \text{rang } E_{j+1}$, où $E_{j+1} \in \mathbb{R}^{r_j \times r_j}$. Si rang $E_{j+1} = r_j$, alors la matrice E_{j+1} est inversible, autrement dit $\nu_{K_j} = 1$ ce qui est absurde. En conséquence, rang $E_j >$ rang E_{j+1}.
2. On suppose que $\nu_{K_j} = 1$. La matrice E_{j+1} est alors inversible, de taille r_j donc de rang r_j. On a bien rang $E_j =$ rang E_{j+1}.
3. Supposons $\nu_{K_j} > 1$ et rang $E_j = 1$. Par construction, $E_{j+1} \in \mathbb{R}$. Si $E_{j+1} \neq 0$, alors E_{j+1} est inversible. Ainsi, $\nu_{K_j} = 1$, ce qui est en contradiction avec l'hypothèse. En conclusion, $E_{j+1} = 0$. □

L'ensemble des précédents résultats nous indiquent qu'à chaque étape, l'indice est diminué, de même que le rang des matrices coefficients. Ceci mène naturellement au résultat suivant.

Théorème 12

Supposons que l'EDA (2.1) est régulière et notons $r =$ rang E. Le nombre d'étapes k de la méthode de déflation satisfait

$$k \leqslant \min(r, \nu_K),$$

où ν_K est l'indice de Kronecker de la paire matricielle (E, A).

Exemple - Considérons l'EDA

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} g \\ h \end{pmatrix},$$

qui est un problème nilpotent d'indice $\nu_K = 2$. Ainsi,

$$E = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad A = I_2, \quad X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{et} \quad f = \begin{pmatrix} g \\ h \end{pmatrix}.$$

La méthode de déflation fournit en une seule étape

$$E_1 = 0, \quad A_1 = 1, \quad X_1 = x_1 \quad \text{et} \quad f_1 = g + \dot{h}.$$

Le rang de la matrice E vaut 1 ce qui indique que le nombre d'étapes est aussi égal à 1, même si $\nu_K = 2$.

□

2.1.4 Solution du problème linéaire à coefficients constants

On résume maintenant les résultats précédents en décrivant la solution de (2.1) obtenue par l'algorithme de déflation.

Théorème 13

On suppose que l'EDA (2.1) est régulière et on note k le nombre d'étapes de la méthode de déflation. Les coordonnées de la solution de (2.1) sont données par le vecteur

$$\begin{pmatrix} X_k \\ Y_k \\ \vdots \\ Y_1 \end{pmatrix}.$$

1. *Si la matrice E_k est inversible, alors X_k satisfait l'EDO*

$$\dot{X}_k = E_k^{-1} \left(A_k X_k + f_k \right).$$

2. *Si la matrice E_k est identiquement nulle, alors*

$$X_k = -A_k^{-1} f_k.$$

De plus,

$$Y_{j+1} = -N_j^{-1} \left(M_j X_{j+1} + h_j \right),$$

pour tout $j = k-1, \ldots, 0$.

2.1.5 Exemples

Nous illustrons à présent la méthode de déflation à travers divers exemples, notamment fournis par la théorie des circuits électriques. Nous utilisons pour cela l'algorithme de déflation `LTI_Deflation` (codé grâce au logiciel Maple), qui donne à l'image du théorème précédent soit une EDO, soit une équation algébrique (suivant l'inversibilité ou la nullité de la matrice E_k), puis une liste d'équations algébriques correspondant aux contraintes exhibées à chaque étape. On note $\mathcal{E} = (\lambda E + A)^{-1} E$.

EDA d'indice 1 (G. Reissig, W. S. Martinson et P. I. Barton [49]) - Soit le problème régulier

$$\begin{cases} \dot{x}_2 + \dot{x}_3 = -x_1 + f_1 \\ \dot{x}_2 + \dot{x}_3 = -x_2 + f_2 \\ \dot{x}_4 + \dot{x}_5 = -x_3 + f_3 \\ \dot{x}_4 + \dot{x}_5 = -x_4 + f_4 \\ 0 = -x_5 + f_5. \end{cases}$$

On note

$$E = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{et} \quad A = -I_5.$$

La matrice E est de rang 2. Pour déterminer l'indice de Kronecker, on prend $\lambda = 0$ (la matrice A est inversible) et on obtient

$$\mathcal{E} = -E \sim \begin{bmatrix} -1 & 1 & & & \\ 0 & -1 & & & \\ & & 0 & 0 & 0 \\ & & 0 & 0 & 0 \\ & & 0 & 0 & 0 \end{bmatrix}.$$

L'indice de Kronecker vaut 1 puisque le bloc nilpotent est nul.

```
> E[0] := <<0|1|1|0|0>, <0|1|1|0|0>, <0|0|0|1|1>, <0|0|0|1|1>, <0|0|0|0|0>> :
> A[0] := - IdentityMatrix(5) :
> f[0] := <f[1](t), f[2](t), f[3](t), f[4](t), f[5](t)> :
> X[0] := <x[1], x[2], x[3], x[4], x[5]> :
> LTI_Deflation(E[0], A[0], f[0], X[0]) ;

[ diff(x[2](t), t) + diff(x[4](t), t) = - x[2](t) + f[2](t) - diff(f[3](t), t)
  + diff(f[4](t), t),
  diff(x[4](t), t) = - x[4](t) + f[4](t) - diff(f[5](t), t) ]

[ [ x[1] = x[2] + f[1](t) - f[2](t),
    x[3] = x[4] + f[3](t) - f[4](t),
    x[5] = f[5](t) ] ]
```

L'algorithme de déflation s'achève en une étape (un seul jeu de contraintes extraites).

EDA d'indice 1 (P. Kunkel et V. Mehrmann [28]) - Soit le problème régulier

$$\begin{cases} C\,(\dot{x}_3 - \dot{x}_2) = R^{-1}\,(x_2 - x_1) \\ 0 = x_1 - x_3 - U \\ 0 = x_3, \end{cases}$$

où $C \neq 0$ et $R \neq 0$. On note

$$E = \begin{bmatrix} 0 & -C & C \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{et} \quad A = \begin{bmatrix} -R^{-1} & R^{-1} & 0 \\ 1 & 0 & -1 \\ 0 & 0 & 1 \end{bmatrix}.$$

La matrice E est de rang 1. Pour déterminer l'indice de Kronecker, on prend $\lambda = 0$ (la matrice A est inversible) et on obtient

$$\mathcal{E} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -RC & RC \\ 0 & 0 & 0 \end{bmatrix} \sim \begin{bmatrix} -RC & & \\ & 0 & 0 \\ & 0 & 0 \end{bmatrix}.$$

Ainsi, l'indice de Kronecker vaut 1.

```
> E[0] := <<0|- C|C>, <0|0|0>, <0|0|0>> :
> A[0] := <<- 1/R|1/R|0>, <1|0|- 1>, <0|0|1>> :
> f[0] := <0, - U, 0> :
> X[0] := <x[1], x[2], x[3]> :
> LTI_Deflation(E[0], A[0], f[0], X[0]) ;

[ diff(x[2](t), t) = - (1/(R*C))*(x[2](t) - U) ]

[ [ x[1](t) = x[3](t) - U,
    x[3](t) = 0 ] ]
```

L'algorithme de déflation s'achève en une étape (un jeu de contraintes extraites).

EDA d'indice 3 (K. E. Brenan, S. L. Campbell et L. R. Petzold [10]) - Soit le problème semi-explicite

$$\begin{cases} \dot{x}_1 = -x_3 + f_1 \\ \dot{x}_2 = -x_1 + f_2 \\ 0 = -x_2 + f_3. \end{cases}$$

On note

$$E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{et} \quad A = \begin{bmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}.$$

La matrice E est de rang 2. Pour déterminer l'indice de Kronecker, on prend $\lambda = 0$ (la matrice A est inversible) et on obtient

$$\mathcal{E} = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ -1 & -\lambda & 0 \end{bmatrix} \sim \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Ainsi, l'indice de Kronecker vaut 3.

```
> E[0] := <<1|0|0>, <0|1|0>, <0|0|0>> :
> A[0] := <<0|0|- 1>, <- 1|0|0>, <0|- 1|0>> :
> f[0] := <f[1](t), f[2](t), f[3](t)> :
> X[0] := <x[1], x[2], x[3]> :
> LTI_Deflation(E[0], A[0], f[0], X[0]) ;

[ x[3](t) = f[1](t) - diff(f[2](t), t) + diff(f[3](t), t, t) ]

[ [ x[1](t) = f[2](t) - diff(f[3](t), t) ],
  [ x[2](t) = f[3](t) ] ]
```

Le nombre d'étapes peut être vu comme le nombre de jeux de contraintes algébriques (et non directement le nombre d'équations algébriques). Par conséquent, l'algorithme de déflation s'achève ici en deux étapes.

EDA d'indice 3 ([19], [55]) - Soit le problème régulier

$$\begin{cases} \dot{x}_3 + \dot{x}_7 = 0 \\ \dot{x}_4 - \dot{x}_8 = 0 \\ L\dot{x}_7 = x_6 \\ C\dot{x}_8 = x_2 \\ 0 = -x_1 - x_2 \\ 0 = -ax_1 - x_3 \\ 0 = -x_4 + V(t) \\ 0 = -x_5 + x_6, \end{cases}$$

où $L \neq 0$, $C \neq 0$ et $a \neq 0$. On note

$$E = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & L & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & C \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{et} \quad A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -a & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \end{bmatrix}.$$

La matrice E est de rang 4. Pour déterminer l'indice de Kronecker, on prend $\lambda = 1$ pour simplifier et on obtient

$$\mathcal{E} = \begin{bmatrix} 0 & 0 & 0 & -C & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & C & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & aC & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -L & aCL & 0 & 0 & 0 & 0 \\ 0 & 0 & -L & aCL & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -aC & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Ainsi, l'indice de Kronecker vaut 3.

```
> E[0] := <<0|0|1|0|0|0|1|0>, <0|0|0|1|0|0|0|- 1>, <0|0|0|0|0|0|L|0>,
>         <0|0|0|0|0|0|0|C>, <0|0|0|0|0|0|0|0>, <0|0|0|0|0|0|0|0>,
>         <0|0|0|0|0|0|0|0>, <0|0|0|0|0|0|0|0>> :
> A[0] := <<0|0|0|0|0|0|0|0>, <0|0|0|0|0|0|0|0>, <0|0|0|0|0|1|0|0>,
>         <0|1|0|0|0|0|0|0>, <- 1|- 1|0|0|0|0|0|0>, <- a|0|- 1|0|0|0|0|0>,
>         <0|0|0|- 1|0|0|0|0>, <0|0|0|0|- 1|1|0|0>> :
> f[0] := <0, 0, 0, 0, 0, 0, V(t), 0> :
> X[0] := <x[1], x[2], x[3], x[4], x[5], x[6], x[7], x[7]> :
> LTI_Deflation(E[0], A[0], f[0], X[0]) ;

[ diff(x[7](t), t) = - a*C*diff(V(t), t, t),
  diff(x[8](t), t) = diff(V(t), t) ]

[ [ x[1](t) = - (1/a)*x[3](t),
```

```
    x[2](t) = (1/a)*x[3](t),
    x[4](t) = V(t),
    x[5](t) = x[6](t) ],
  [ x[3](t) = a*C*diff(V(t), t) ],
  [ x[6](t) = - a*C*L*diff(V(t), t, t) ] ]
```

L'algorithme de déflation s'achève en trois étapes.

2.2 EDAs linéaires à coefficients variables

D'un point de vue structurel, la méthode est similaire au cas linéaire à coefficients constants. On adapte simplement le mécanisme de substitution (injection de la partie algébrique dans la partie différentielle) car de nouveaux termes provenant de la dérivation apparaissent. En revanche, le passage aux coefficients variables fait perdre la propriété de transmission de la régularité. Cette différence est somme toute attendue puisque nous savons que les notions de régularité et d'unicité des solutions deviennent indépendantes dans le contexte variable.

2.2.1 Mécanisme de déflation

On regarde le problème de la forme

$$E(t)\dot{X} = A(t)X + f. \tag{2.12}$$

La régularité n'étant plus automatiquement transmise, on est dans l'obligation de la considérer comme telle pour pouvoir appliquer la méthode de déflation. Nous adaptons ainsi la notion de régularité au problème à coefficients variables. Cette nouvelle notion a été introduite sous une autre forme dans la définition 3.1 de [46].

Définition 7 (*Régularité géométrique*)

Le problème (2.12) est dit géométriquement régulier *sur un intervalle ouvert $\mathcal{I} \in \mathbb{R}^n$ si*
- *le rang de la matrice $E(t)$ est constant sur $\mathcal{I}$;*
- *la transformation (2.3) s'applique.*

En d'autres termes, puisqu'il n'est plus possible de s'appuyer sur la régularité de la paire matricielle, on est contraint de supposer l'existence de la transformation. On exige ainsi que la matrice A_2 (dans sa version dépendante du temps) soit de rang plein quitte à permuter les colonnes de cette dernière pour extraire un bloc inversible.

La régularité géométrique de (2.12) donne

$$\begin{bmatrix} S(t) & T(t) \\ 0 & 0 \end{bmatrix} P^{-1}\dot{X} = \begin{bmatrix} H(t) & K(t) \\ M(t) & N(t) \end{bmatrix} P^{-1}X + \begin{pmatrix} g \\ h \end{pmatrix}, \tag{2.13}$$

où $\begin{pmatrix} g \\ h \end{pmatrix} = G(t)^{-1}f$. En posant $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} = P^{-1}X$, on développe (2.4) :

$$\begin{bmatrix} S(t) & T(t) \\ 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{X}_1 \\ \dot{Y}_1 \end{pmatrix} = \begin{bmatrix} H(t) & K(t) \\ M(t) & N(t) \end{bmatrix} \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} + \begin{pmatrix} g \\ h \end{pmatrix}$$
$$\Leftrightarrow \begin{cases} S(t)\dot{X}_1 + T(t)\dot{Y}_1 = H(t)X_1 + K(t)Y_1 + g \\ 0 = M(t)X_1 + N(t)Y_1 + h. \end{cases}$$

La matrice $N(t)$ est inversible par l'hypothèse de régularité géométrique. On peut ainsi extraire une expression de Y_1 de l'équation algébrique $0 = M(t)X_1 + N(t)Y_1 + h$. De plus, en dérivant cette dernière par rapport à la variable t, on obtient

$$0 = \dot{M}(t)X_1 + M(t)\dot{X}_1 + \dot{N}(t)Y_1 + N(t)\dot{Y}_1 + \dot{h}.$$

Par conséquent,

$$\begin{aligned}\dot{Y}_1 &= -N(t)^{-1}\left(\dot{M}(t)X_1 + M(t)\dot{X}_1 + \dot{N}(t)Y_1 + \dot{h}\right)\\ &= -N(t)^{-1}\left(\dot{M}(t)X_1 + M(t)\dot{X}_1 - \dot{N}(t)N(t)^{-1}\left(M(t)X_1 + h\right) + \dot{h}\right).\end{aligned}$$

Ainsi,

$$\begin{aligned}&S(t)\dot{X}_1 + T(t)\dot{Y}_1\\ &= S(t)\dot{X}_1 - T(t)N^{-1}\left(\dot{M}(t)X_1 + M(t)\dot{X}_1 - \dot{N}(t)N(t)^{-1}\left(M(t)X_1 + h\right) + \dot{h}\right)\\ &= \left(S(t) - T(t)N(t)^{-1}M(t)\right)\dot{X}_1 - T(t)N(t)^{-1}\left(\dot{M}(t) - \dot{N}(t)N(t)^{-1}M(t)\right)X_1\\ &\quad - T(t)N(t)^{-1}\left(\dot{h} - \dot{N}(t)N(t)^{-1}h\right)\end{aligned}$$

et

$$\begin{aligned}&H(t)X_1 + K(t)Y_1 + g\\ &= H(t)X_1 - K(t)N(t)^{-1}\left(M(t)X_1 + h\right) + g\\ &= \left(H(t) - K(t)N(t)^{-1}M(t)\right)X_1 + T(t)N(t)^{-1}\dot{h} - K(t)N(t)^{-1}h + g.\end{aligned}$$

On note

$$\begin{aligned}E_1(t) &= S(t) - T(t)N(t)^{-1}M(t),\\ A_1(t) &= H(t) - K(t)N(t)^{-1}M(t) + T(t)N(t)^{-1}\left(\dot{M}(t) - \dot{N}(t)N(t)^{-1}M(t)\right),\\ f_1 &= T(t)N(t)^{-1}\left(\dot{h} - \dot{N}(t)N(t)^{-1}h\right) - K(t)N(t)^{-1}h + g.\end{aligned}$$

Via le changement de variable $\begin{pmatrix}X_1\\ Y_1\end{pmatrix} = P^{-1}X$, le problème géométriquement régulier (2.12) équivaut à

$$\begin{cases}E_1(t)\dot{X}_1 = A_1(t)X_1 + f_1 & \text{(2.14a)}\\ \qquad Y_1 = -N(t)^{-1}\left(M(t)X_1 + h\right). & \text{(2.14b)}\end{cases}$$

Définition 8 (*EDA linéaire à coefficients variables déflatée*)

L'EDA (2.14a) *obtenue en opérant la régularité géométrique est appelée* EDA déflatée, *relativement à l'EDA* (2.12).

2.2.2 Algorithme de déflation

On adapte l'algorithme `LTI_Deflation` aux coefficients variables. Cette version de l'algorithme est nommée `LTV_Deflation`[2].

2. Pour « Linear Time-Varying Deflation ».

`LTV_Deflation`
begin
Input : $E_0(t) = E(t)$, $A_0(t) = A(t)$, $f_0 = f$, $X_0 = X$, $r_{-1} = n$.
Step $j+1$, $j \geqslant 0$:
if *$E_j(t)$ est singulière,*
then
if *$(E_j(t), A_j(t))$ est géométriquement régulière,*
then
Calculer le rang r_j de $E_j(t)$.
Déterminer les expressions des matrices P_j, $S_j(t)$, $T_j(t)$, $H_j(t)$, $K_j(t)$, $M_j(t)$ et $N_j(t)$ exhibées dans la transformation (2.13) de $(E_j(t), A_j(t))$.
Déterminer le triplet $(E_{j+1}(t), A_{j+1}(t), f_{j+1}(t))$ à partir des formules précédentes.
Effectuer le changement de variable $\begin{pmatrix} X_{j+1} \\ Y_{j+1} \end{pmatrix} = P_j^{-1} X_j$, où la dimension de X_{j+1} est égale à r_j.
Déterminer les contraintes algébriques $0 = M_j(t)X_{j+1} + N_j(t)Y_{j+1} + h_j$.
else
Stop
end
else
Stop
end
Output : Soit k le nombre d'étapes de l'algorithme.
if *$(E_j(t), A_j(t))$ est géométriquement régulière pour tout $j \in [\![0, k-1]\!]$,*
then
if *$E_k(t)$ est inversible*
then
Fournir

$$\begin{cases} \dot{X}_k = E_k(t)^{-1} \left(A_k(t) X_k + f_k \right) \\ 0 = M_{k-1}(t) X_k + N_{k-1}(t) Y_k + h_{k-1} \\ \quad \vdots \\ 0 = M_0(t) X_1 + N_0(t) Y_1 + h_0. \end{cases}$$

else
Fournir

$$\begin{cases} 0 = A_k(t) X_k + f_k \\ 0 = M_{k-1}(t) X_k + N_{k-1}(t) Y_k + h_{k-1} \\ \quad \vdots \\ 0 = M_0(t) X_1 + N_0(t) Y_1 + h_0. \end{cases}$$

end
end
end

2.2.3 Solution du problème linéaire à coefficients variables

A l'image du contexte linéaire à coefficients constants, rassemblons les résultats donnant la solution de (2.1) obtenue par l'algorithme de déflation `LTV_Deflation`.

Théorème 14

On suppose que les EDAs $E_j(t)\dot{X}_j = A_j(t)X_j + f_j$ sont géométriquement régulières pour tout $j \in [\![0, k-1]\!]$, où k désigne le nombre d'étapes de la méthode de déflation. Les coordonnées de la solution de (2.12) sont données par le vecteur

$$\begin{pmatrix} X_k \\ Y_k \\ \vdots \\ Y_1 \end{pmatrix}.$$

1. *Si la matrice $E_k(t)$ est inversible, alors X_k satisfait l'EDO*

$$\dot{X}_k = E_k(t)^{-1}\left(A_k(t)X_k + f_k\right).$$

2. *Si la matrice $E_k(t)$ est identiquement nulle, alors*

$$X_k = -A_k(t)^{-1} f_k.$$

De plus,

$$Y_{j+1} = -N_j(t)^{-1}\left(M_j(t)X_{j+1} + h_j\right),$$

pour tout $j = k-1, \ldots, 0$.

2.2.4 Exemple

On résout les systèmes de la même manière que dans le contexte des coefficients non variables. Le terme *indice* fait ici référence à l'indice de différentiation ; on observe que dans cet exemple, il coïncide avec k. On omet d'écrire la dépendance en t pour ne pas alourdir les notations, mais on garde en mémoire que toutes les quantités manipulées dépendent du temps. L'exemple suivant, décliné en trois cas, est extrait de R. Riaza [51], qui le traite par la méthode des projections.

EDA d'indice 1 - Soit le problème homogène géométriquement régulier suivant

$$\begin{cases} C_1\dot{x}_1 = -\dot{C}_1 x_1 + x_4 - x_5 \\ C_2\dot{x}_2 = -\dot{C}_2 x_2 - x_3 - x_4 \\ L\dot{x}_3 = x_2 - \dot{L}x_3 \\ 0 = x_1 - x_2 + R_1 x_4 \\ 0 = x_1 - R_2 x_5, \end{cases}$$

où $C_1 \neq 0$, $C_2 \neq 0$, $L \neq 0$, $R_1 \neq 0$ et $R_2 \neq 0$. On note

$$E = \begin{bmatrix} C_1 & 0 & 0 & 0 & 0 \\ 0 & C_2 & 0 & 0 & 0 \\ 0 & 0 & L & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{et} \quad A = \begin{bmatrix} -\dot{C}_1 & 0 & 0 & 1 & -1 \\ 0 & -\dot{C}_2 & -1 & -1 & 0 \\ 0 & 1 & -\dot{L} & 0 & 0 \\ 1 & -1 & 0 & R_1 & 0 \\ 1 & 0 & 0 & 0 & -R_2 \end{bmatrix}.$$

On applique à présent l'algorithme au système précédent :

```
> E[0] := <<C[1]|0|0|0|0>, <0|C[2]|0|0|0>, <0|0|L|0|0>, <0|0|0|0|0>,
>         <0|0|0|0|0>> :
> A[0] := <<- diff(C[1], t)|0|0|1|- 1>, <0|- diff(C[2], t)|- 1|- 1|0>,
>         <0|1|- diff(L, t)|0|0>, <1|- 1|0|R[1]|0>, <1|0|0|0|- R[2]>> :
> f[0] := <0, 0, 0, 0, 0> :
> X[0] := <x[1], x[2], x[3], x[4], x[5]> :
> LTV_Deflation(E[0], A[0], f[0], X[0]) ;

[ L*diff(x[3](t), t) = - diff(L, t)*x[3](t) + R[1]*x[4](t) + R[2]*x[5](t),
  C[2]*R[1]*diff(x[4](t), t) + C[2]*R[2]*diff(x[5](t), t) = - x[3](t)
  - (1 + diff(C[2], t)*R[1] + C[2]*diff(R[1], t))*x[4](t)
  - (diff(C[2], t)*R[2] + C[2]*diff(R[2], t))*x[5](t),
  C[1]*R[2]*diff(x[5](t), t) = x[4](t) - (1 + diff(C[1], t)*R[2]
  + C[1]*diff(R[2], t))*x[5](t) ]

[ [ x[1](t) = R[2]*x[5](t),
    x[2](t) = x[1](t) + R[1]*x[4](t) ] ]
```

L'algorithme de déflation s'achève en une étape (un jeu de contraintes extraites). La solution du problème satisfait

$$\begin{cases} L\dot{x}_3 = -\dot{L}x_3 + R_1x_4 + R_2x_5 \\ C_2R_1\dot{x}_4 + C_2R_2\dot{x}_5 = -x_3 - \left(1 + \dot{C}_2R_1 + C_2\dot{R}_1\right)x_4 - \left(\dot{C}_2R_2 + C_2\dot{R}_2\right)x_5 \\ C_1R_2\dot{x}_5 = x_4 - \left(1 + \dot{C}_1R_2 + C_1\dot{R}_2\right)x_5 \\ x_1 = R_2x_5 \\ x_2 = x_1 + R_1x_4. \end{cases}$$

EDA d'indice 2 - Soit le problème homogène géométriquement régulier suivant

$$\begin{cases} C_1\dot{x}_1 = -\dot{C}_1x_1 + x_4 - x_5 \\ C_2\dot{x}_2 = -\dot{C}_2x_2 - x_3 - x_4 \\ L\dot{x}_3 = x_2 - \dot{L}x_3 \\ 0 = x_1 - x_2 \\ 0 = x_1 - Rx_5, \end{cases}$$

où $C_1 \neq 0$, $C_2 \neq 0$, $L \neq 0$ et $R \neq 0$. On suppose que $C_1 + C_2 \neq 0$ et on note

$$E = \begin{bmatrix} C_1 & 0 & 0 & 0 & 0 \\ 0 & C_2 & 0 & 0 & 0 \\ 0 & 0 & L & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{et} \quad A = \begin{bmatrix} -\dot{C}_1 & 0 & 0 & 1 & -1 \\ 0 & -\dot{C}_2 & -1 & -1 & 0 \\ 0 & 1 & -\dot{L} & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -R \end{bmatrix}.$$

L'algorithme appliqué au système précédent donne :

```
> E[0] := <<C[1]|0|0|0|0>, <0|C[2]|0|0|0>, <0|0|L|0|0>, <0|0|0|0|0>,
>         <0|0|0|0|0>> :
```

```
> A[0] := <<- diff(C[1], t)|0|0|1|- 1>, <0|- diff(C[2], t)|- 1|- 1|0>,
>         <0|1|- diff(L, t)|0|0>, <1|- 1|0|0|0>, <1|0|0|0|- R> > :
> f[0] := <0, 0, 0, 0, 0> :
> X[0] := <x[1], x[2], x[3], x[4], x[5]> :
> alias(a = - C[2]*diff(R, t) - diff(C[2], t)*R) :
> alias(c = C[1]/C[2]) :
> alias(b = - 1 - diff(C[1], t)*R + diff(C[2], t)*R*c) :
> LTV_Deflation(E[0], A[0], f[0], X[0]) ;

[ L*(1 + c)*c*diff(x[4](t), t) + b*c*L*diff(x[5](t), t) =
  - (diff(L, t)*(c + c^2) - L*diff(c, t))*x[4](t) - (R*c^2 + diff(L, t)*b*c
  + L*c*diff(b, t) - L*b*diff(c, t))*x[5](t),
  C[1]*R*diff(x[5](t), t) = x[4](t) + (a*c + b)*x[5](t) ]

[ [ x[1](t) = R[2]*x[5](t),
    x[2](t) = x[1](t) ],
  [ 0 = (1 + c)*x[4](t) + b*x[5](t) + c*x[3](t) ] ]
```

L'algorithme de déflation s'achève en deux étapes. En effet, la matrice E_2 correspondante s'écrit $\begin{bmatrix} L(1+c)c & Lbc \\ 0 & C_1R \end{bmatrix}$ et son déterminant vaut $LRC_1^2C_2^{-2}\,(C_1+C_2)$. La solution du problème satisfait

$$\begin{cases} L(1+c)c\dot{x}_4 + Lbc\dot{x}_5 = -(\dot{L}(c+c^2) - L\dot{c})x_4 - (R_2c^2 + \dot{L}cb + Lc\dot{b} - L\dot{c}b)x_5 \\ C_1R_2\dot{x}_5 = x_4 + (ca+b)x_5 \\ x_1 = R_2x_5 \\ x_2 = x_1 \\ x_3 = -c^{-1}\left((1+c)x_4 + bx_5\right). \end{cases}$$

EDA d'indice 3 - On reprend le problème précédent en supposant $C_1+C_2=0$. Une troisième étape est nécessaire pour achever l'algorithme qui fournit ainsi

```
> E[0] := <<C[1]|0|0|0|0>,<0|C[2]|0|0|0>,<0|0|L|0|0>,<0|0|0|0|0>,<0|0|0|0|0>> :
> A[0] := <<- diff(C[1], t)|0|0|1|- 1>, <0|- diff(C[2], t)|- 1|- 1|0>,
>         <0|1|- diff(L, t)|0|0>, <1|- 1|0|0|0>, <1|0|0|0|- R> > :
> f[0] := <0, 0, 0, 0, 0> :
> X[0] := <x[1], x[2], x[3], x[4], x[5]> :
> alias(a = - C[2]*diff(R, t) - diff(C[2], t)*R) :
> alias(c = C[1]/C[2]) :
> alias(b = - 1 - diff(C[1], t)*R + diff(C[2], t)*R*c :
> alias(d = 1 + diff(C[1], t)*R + C[1]*diff(R, t)) :
> alias(e = C[1]*R^2 + C[1]*R*diff(L, t) - L - diff(C[1], t)*R*L
> - C[1]*diff(R, t)*L) :
> LTV_Deflation(E[0], A[0], f[0], X[0]) ;

[ L*C[1]*R*diff(x[5](t), t) = - (L*d + e)*x[5](t) ]

[ [ x[1](t) = R[2]*x[5](t),
    x[2](t) = x[1](t) ],
  [ x[3](t) = - x[5](t) ],
  [ L*x[4](t) = - e*x[5](t) ] ]
```

La solution du problème satisfait par conséquent

$$\begin{cases} \dot{x}_5 = -\left(LC_1R\right)^{-1}\left(Ld+e\right)x_5 \\ x_1 = Rx_5 \\ x_2 = x_1 \\ x_3 = -x_5 \\ x_4 = -L^{-1}ex_5. \end{cases}$$

2.3 EDAs quasi-linéaires

2.3.1 Mécanisme de déflation

On généralise à présent la méthode de déflation au contexte quasi-linéaire

$$E\left(X\right)\dot{X} = f\left(X\right), \tag{2.15}$$

déjà rencontré dans le premier chapitre. En raison de la non-linéarité des expressions et en particulier du vecteur $f\left(X\right)$, la méthode fait appel au théorème des fonctions implicites pour exprimer à partir des contraintes algébriques un jeu de variables par rapport à un autre. Dans le cas linéaire, la méthode exige l'inversibilité de la matrice coefficient $N(t)$ dans la contrainte algébrique

$$0 = M(t)X_1 + N(t)Y_1 + h(t).$$

Dans ce nouveau contexte, le théorème des fonctions implicites exige l'inversibilité de la matrice Jacobienne $J_{Y_1}\left(h\right)$, où

$$0 = h(X_1, Y_1),$$

ce qui revient à considérer la linéarisation de la contrainte, *i.e.*

$$0 = J_{X_1}\left(h\right)\dot{X}_1 + J_{Y_1}\left(h\right)\dot{Y}_1.$$

On commence par supposer que le rang r de la matrice $E\left(X\right)$ est constant sur un intervalle $\mathcal{I} \in \mathbb{R}$. Cette hypothèse permet de décomposer la matrice via les transformations usuelles de l'algèbre linéaire. En posant $E\left(X\right) = G\left(X\right)\begin{bmatrix} F\left(X\right) \\ 0 \end{bmatrix}$ et $f\left(X\right) = G\left(X\right)\begin{pmatrix} g\left(X\right) \\ h\left(X\right) \end{pmatrix}$, on a

$$\begin{aligned} E\left(X\right)\dot{X} = f\left(X\right) &\Leftrightarrow G\left(X\right)\begin{bmatrix} F\left(X\right) \\ 0 \end{bmatrix}\dot{X} = G\left(X\right)\begin{pmatrix} g\left(X\right) \\ h\left(X\right) \end{pmatrix} \\ &\Leftrightarrow \begin{bmatrix} F\left(X\right) \\ 0 \end{bmatrix}\dot{X} = \begin{pmatrix} g\left(X\right) \\ h\left(X\right) \end{pmatrix} \end{aligned}$$

L'attention se porte maintenant sur la contrainte algébrique $0 = h\left(X\right)$. On suppose qu'il existe une matrice de permutation $\mathcal{P}$, éventuellement égale à la matrice identité, induisant

$$\mathcal{P}X = \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix},$$

et telle que la matrice Jacobienne $J_{Y_1}\left(h\right)$ est inversible. Ainsi

$$\begin{aligned} \begin{bmatrix} F\left(X\right) \\ 0 \end{bmatrix}\dot{X} = \begin{pmatrix} g\left(X\right) \\ h\left(X\right) \end{pmatrix} &\Leftrightarrow \begin{bmatrix} F\left(\mathcal{P}^{-1}\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}\right) \\ 0 \end{bmatrix}\mathcal{P}^{-1}\begin{pmatrix} \dot{X}_1 \\ \dot{Y}_1 \end{pmatrix} = \begin{pmatrix} g\left(\mathcal{P}^{-1}\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}\right) \\ h\left(\mathcal{P}^{-1}\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}\right) \end{pmatrix} \\ &\Leftrightarrow \begin{bmatrix} S\left(X_1, Y_1\right) & Y\left(X_1, Y_1\right) \\ 0 & 0 \end{bmatrix}\begin{pmatrix} \dot{X}_1 \\ \dot{Y}_1 \end{pmatrix} = \begin{pmatrix} \bar{g}\left(X_1, Y_1\right) \\ \bar{h}\left(X_1, Y_1\right) \end{pmatrix}, \end{aligned}$$

où

$$\begin{bmatrix} S\left(X_1, Y_1\right) & Y\left(X_1, Y_1\right) \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} F\left(\mathcal{P}^{-1}\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}\right) \\ 0 \end{bmatrix} \mathcal{P}^{-1},$$

$$\bar{g}\left(X_1, Y_1\right) = g\left(\mathcal{P}^{-1}\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}\right) \quad \text{et} \quad \bar{h}\left(X_1, Y_1\right) = h\left(\mathcal{P}^{-1}\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}\right).$$

Définition 9 (1-*Régularité*)

Le problème (2.15) *est dit* 1-régulier *sur un intervalle ouvert* $\mathcal{I} \in \mathbb{R}$ *si*

- *le rang de la matrice* $E\left(X\right)$ *est constant sur* $\mathcal{I}$ *;*
- *il existe une matrice de permutation* $\mathcal{P}$ *telle que* $\mathcal{P}X = \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}$ *et la matrice Jacobienne* $J_{Y_1}\left(h\right)$ *est inversible.*

Dans le cadre de la 1-régularité, il existe une fonction $\varphi_1 : \mathcal{I}_r \to \mathcal{I}_{n-r}$ telle que

$$Y_1 = \varphi_1\left(X_1\right).$$

En injectant ceci dans l'expression transformée de (2.15) et en développant cette dernière, on parvient au système

$$\begin{cases} \left(S\left(X_1, \varphi_1\left(X_1\right)\right) - T\left(X_1, \varphi_1\left(X_1\right)\right) J_{Y_1}\left(h\right)^{-1} J_{X_1}\left(h\right)\right) \dot{X}_1 = \bar{g}\left(X_1, \varphi_1\left(X_1\right)\right) \\ 0 = \bar{h}\left(X_1, Y_1\right). \end{cases} \tag{2.16}$$

On note

$$\begin{aligned} E_1\left(X_1\right) &= S\left(X_1, \varphi_1\left(X_1\right)\right) - T\left(X_1, \varphi_1\left(X_1\right)\right) J_{Y_1}\left(h\right)^{-1} J_{X_1}\left(h\right), \\ f_1\left(X_1\right) &= \bar{g}\left(X_1, \varphi_1\left(X_1\right)\right). \end{aligned}$$

Le problème (2.15) est alors équivalent à

$$\begin{cases} E_1\left(X_1\right) \dot{X}_1 = f_1\left(X_1\right) & \text{(2.17a)} \\ Y_1 = \varphi_1\left(X_1\right). & \text{(2.17b)} \end{cases}$$

Définition 10 (*EDA quasi-linéaire déflatée*)

L'EDA (2.17a) *obtenue en opérant la* 1*-régularité est appelée* EDA déflatée, *relativement à l'EDA* (2.15).

Remarque 14

L'utilisation du théorème des fonctions implicites est une difficulté pratique de la méthode de déflation quasi-linéaire, notamment parce qu'il faut manipuler des fonctions implicites emboîtées. Dans certains cas concrets comme pour l'exemple du pendule simple, il est cependant possible d'exprimer à moindre coup les fonctions implicites (qui en conséquence ne le sont plus). De plus, l'utilisation du théorème des fonctions implicites introduit des difficultés algorithmiques car les logiciels de calcul formel peuvent difficilement gérer ce théorème.

2.3.2 Algorithme de déflation

On applique en chaîne le mécanisme précédemment décrit en supposant qu'à chaque étape, le problème déflaté obtenu est 1-régulier. L'algorithme `QL_Deflation`[3] s'en suit.

3. Pour « QuasiLinear Deflation ».

`QL_Deflation`
begin
Input : $X_0 = X$, $E_0(X_0) = E(X)$, $f_0(X_0) = f(X)$ et $r_{-1} = n$.
Step $j+1$, $j \geqslant 0$:
if *$E_j(X_j)$ est singulière,*
then
Calculer le rang r_j de $E_j(X_j)$.
Déterminer les expressions des matrices $G_j(X_j)$ et $F_j(X_j)$, ainsi que des vecteurs $g_j(X_j)$ et $h_j(X_j)$.
if *Le théorème des fonctions implicites ne s'applique pas sur la contrainte algébrique* $0 = h_j(X_j)$,
then
Stop
else
Déterminer les expressions des matrices $\mathcal{P}_j$, $S_j(X_j)$, $T_j(X_j)$, $H_j(X_j)$.
Effectuer le changement de variable $\begin{pmatrix} X_{j+1} \\ Y_{j+1} \end{pmatrix} = \mathcal{P}_j X_j$, où la dimension de X_{j+1} est égale à r_j.
Déterminer la matrice $E_{j+1}(X_{j+1})$ et le vecteur $f_{j+1}(X_{j+1})$ à partir des formules précédentes.
Appliquer le théorème des fonctions implicites sur la contrainte algébrique $0 = \bar{h}_j(X_{j+1}, Y_{j+1})$ et extraire formellement une fonction φ_{j+1} telle que $Y_{j+1} = \varphi_{j+1}(X_{j+1})$.
end
else
Stop
end
Output : Soit k le nombre d'étapes de l'algorithme.
if *$E_j(X_j)\dot{X}_j = f_j(X_j)$ est 1-régulière pour tout* $j \in [\![0, k-1]\!]$,
then
if *$E_k(X_k)$ est inversible*
then
Fournir

$$\begin{cases} \dot{X}_k = E_k(X_k)^{-1} f_k(X_k) \\ 0 = \bar{h}_{k-1}(X_k, Y_k) \\ \quad \vdots \\ 0 = \bar{h}_0(X_1, Y_1). \end{cases}$$

else
Fournir

$$\begin{cases} 0 = f_k(X_k) \\ 0 = \bar{h}_{k-1}(X_k, Y_k) \\ \quad \vdots \\ 0 = \bar{h}_0(X_1, Y_1). \end{cases}$$

end
end
end

2.3.3 Application aux problèmes mécaniques

On utilise l'algorithme `QL_Deflation` pour étudier des problèmes issus de la mécanique. On commence par réduire via l'algorithme les équations modélisant le mouvement d'un pendule simple pesant en dimension 2 et 3. Ces équations réduites (EDOs sous contraintes) sont de plus traitées numériquement. On étend enfin l'étude à des systèmes dynamiques plus généraux. Ceci permet en outre de retrouver et de résoudre le problème du pendule en dimension quelconque n.

2.3.3.1 Pendule en dimension 2

2.3.3.1.1 Modélisation et résolution formelle On s'intéresse à l'étude standard du pendule simple dans le plan. Considérons un mobile P de masse $\mathfrak{m} = 1$, relié à un point fixe par un fil de masse négligeable et de longueur constante $\mathfrak{l} = 1$. Les frottements sont négligés. On utilise un système de coordonnées cartésiennes (x, y), où l'axe des abscisses est orienté vers la droite et l'axe des ordonnées vers le bas. Ainsi, on établit aisément les expressions des énergies de ce système, à savoir l'énergie cinétique E_c et l'énergie potentielle E_p :

$$E_c = \frac{v^2}{2} = \frac{\dot{x}^2 + \dot{y}^2}{2} \quad \text{et} \quad E_p = \mathfrak{g}(1 - y).$$

On désigne par $\mathfrak{g}$ l'accélération de la pesanteur. On choisit l'énergie potentielle de telle sorte qu'elle soit nulle lorsque le pendule est à sa position d'équilibre stable, c'est-à-dire quand $y = 1$. On est dans le cadre d'un problème sous contrainte ; posons $h_0(x, y) = x^2 + y^2 - 1$. Puisque la longueur du fil est supposée constante, les coordonnées (x, y) de P vérifient $h_0(x, y) = 0$. On peut alors écrire le Lagrangien de ce problème :

$$L(x, \dot{x}, y, \dot{y}, \lambda) = E_c - E_p - \lambda h_0(x, y) = \frac{1}{2}\left(\dot{x}^2 + \dot{y}^2\right) - \mathfrak{g}(1 - y) - \lambda\left(x^2 + y^2 - 1\right),$$

où $\lambda \in \mathbb{R}$ est le multiplicateur de Lagrange associé à la contrainte du problème. Il ne reste plus qu'à obtenir les équations du mouvement :

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}}\right) = \frac{\partial L}{\partial x}, \quad \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{y}}\right) = \frac{\partial L}{\partial y} \quad \text{et} \quad 0 = \frac{\partial L}{\partial \lambda}.$$

De simples calculs amènent à :

$$\ddot{x} = -2\lambda x, \quad \ddot{y} = \mathfrak{g} - 2\lambda y \quad \text{et} \quad 0 = x^2 + y^2 - 1.$$

On modélise donc le mouvement d'un pendule simple par le système dynamique suivant :

$$\begin{cases} \ddot{x} = -2\lambda x \\ \ddot{y} = g - 2\lambda y \\ 0 = x^2 + y^2 - 1. \end{cases} \tag{2.18}$$

On introduit les variables x_1, x_2, x_3, x_4 et x_5 pour mettre (2.18) sous la forme (2.15) :

$$x_1 = x, \quad x_2 = y, \quad x_3 = \dot{x}, \quad x_4 = \dot{y} \quad \text{et} \quad x_5 = 2\lambda.$$

Le système (2.18) devient donc :

$$\begin{cases} \dot{x}_1 = x_3 \\ \dot{x}_2 = x_4 \\ \dot{x}_3 = -x_5 x_1 \\ \dot{x}_4 = \mathfrak{g} - x_5 x_2 \\ 0 = x_1^2 + x_2^2 - 1. \end{cases} \tag{2.19}$$

On note

$$X_0 = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}, \quad E_0(X_0) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{et} \quad f_0(X_0) = \begin{pmatrix} x_3 \\ x_4 \\ -x_5 x_1 \\ \mathfrak{g} - x_5 x_2 \\ x_1^2 + x_2^2 - 1 \end{pmatrix}.$$

Théorème 15

La réduction de (2.19) est obtenue après trois étapes de la méthode de déflation.

1. *Si $|x_2| > |x_1|$, alors la solution générale de (2.19) satisfait l'EDO sous contraintes suivante :*

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_3 \end{pmatrix} = \begin{pmatrix} x_3 \\ -\left[\mathfrak{g}\varphi_1(x_1) + \left(\varphi_1(x_1)^{-1} x_3\right)^2\right] x_1 \end{pmatrix}, \tag{2.20}$$

   avec
   - $x_2 = \varphi_1(x_1)$ *tel que* $0 = x_1^2 + x_2^2 - 1$;
   - $x_4 = -\varphi_1(x_1)^{-1} x_1 x_3$;
   - $x_5 = \mathfrak{g}\varphi_1(x_1) + \left(\varphi_1(x_1)^{-1} x_3\right)^2$.

2. *Si $|x_1| > |x_2|$, alors la solution générale de (2.19) satisfait l'EDO sous contraintes suivante :*

$$\begin{pmatrix} \dot{x}_2 \\ \dot{x}_4 \end{pmatrix} = \begin{pmatrix} x_4 \\ \mathfrak{g} - \left[\mathfrak{g} x_2 + \left(\phi_1(x_2)^{-1} x_4\right)^2\right] x_2 \end{pmatrix}, \tag{2.21}$$

   avec
   - $x_1 = \phi_1(x_2)$ *tel que* $0 = x_1^2 + x_2^2 - 1$;
   - $x_3 = -\phi_1(x_2)^{-1} x_2 x_4$;
   - $x_5 = \mathfrak{g} x_2 + \left(\phi_1(x_2)^{-1} x_4\right)^2$.

Preuve - On applique directement l'algorithme de déflation.

1. ***Étape 1*** - La matrice $E_0(X_0)$ est déjà écrite sous forme décomposée. Puisque $|x_2| > |x_1|$, on sait que $x_2 \neq 0$. Par le théorème des fonctions implicites, il existe une fonction φ_1 telle que $x_2 = \varphi_1(x_1)$, où $0 = x_1^2 + x_2^2$. On applique le mécanisme de déflation et on obtient le système déflaté :

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\varphi_1(x_1)^{-1} x_1 & 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{x}_1 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{pmatrix} = \begin{pmatrix} x_3 \\ -x_5 x_1 \\ \mathfrak{g} - x_5 \varphi_1(x_1) \\ x_4 \end{pmatrix}. \tag{2.22}$$

   On note

$$X_1 = \begin{pmatrix} x_1 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}, \quad E_1(X_1) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\varphi_1(x_1)^{-1} x_1 & 0 & 0 & 0 \end{bmatrix} \quad \text{et} \quad f_1(X_1) = \begin{pmatrix} x_3 \\ -x_5 x_1 \\ \mathfrak{g} - x_5 \varphi_1(x_1) \\ x_4 \end{pmatrix}.$$

   Puisque la matrice $E_1(X_1)$ n'est pas inversible, on itère le processus.

   Étape 2 - Le complément de Schur fournit :

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\varphi_1(x_1)^{-1} x_1 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\varphi_1(x_1)^{-1} x_1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Après de simples calculs, (2.22) devient :

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{x}_1 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{pmatrix} = \begin{pmatrix} x_3 \\ -x_5 x_1 \\ \mathfrak{g} - x_5 \varphi_1(x_1) \\ h_1\left(x_1, x_3, x_4\right) \end{pmatrix},$$

où $h_1\left(x_1, x_3, x_4\right) = \varphi_1(x_1)^{-1} x_1 x_3 + x_4$. Par le théorème des fonctions implicites, il existe une fonction φ_2 telle que $x_4 = \varphi_2(x_1, x_3)$, où $0 = h_1\left(x_1, x_3, x_4\right)$. Dans le cas présent, cette fonction est explicite. En dérivant la contrainte et en injectant cette nouvelle expression dans le précédent système, on arrive à

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\left(\frac{x_3}{\varphi_1(x_1)} - \frac{x_1 x_3 \dot{\varphi}_1(x_1)}{\varphi_1^2(x_1)}\right) & -\frac{x_1}{\varphi_1(x_1)} & 0 \end{bmatrix} \begin{pmatrix} \dot{x}_1 \\ \dot{x}_3 \\ \dot{x}_5 \end{pmatrix} = \begin{pmatrix} x_3 \\ -x_5 x_1 \\ \mathfrak{g} - x_5 \varphi_1(x_1) \end{pmatrix}. \tag{2.23}$$

On note

$$X_2 = \begin{pmatrix} x_1 \\ x_3 \\ x_5 \end{pmatrix}, \quad f_2\left(X_2\right) = \begin{pmatrix} x_3 \\ -x_5 x_1 \\ \mathfrak{g} - x_5 \varphi_1(x_1) \end{pmatrix}$$

et

$$E_2\left(X_2\right) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\left(\frac{x_3}{\varphi_1(x_1)} - \frac{x_1 x_3 \dot{\varphi}_1(x_1)}{\varphi_1^2(x_1)}\right) & -\frac{x_1}{\varphi_1(x_1)} & 0 \end{bmatrix}.$$

Encore une fois, le processus continue.

Étape 3 - Le complément de Schur fournit :

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\left(\frac{x_3}{\varphi_1(x_1)} - \frac{x_1 x_3 \dot{\varphi}_1(x_1)}{\varphi_1^2(x_1)}\right) & -\frac{x_1}{\varphi_1(x_1)} & 0 \end{bmatrix}$$
$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\left(\frac{x_3}{\varphi_1(x_1)} - \frac{x_1 x_3 \dot{\varphi}_1(x_1)}{\varphi_1^2(x_1)}\right) & -\frac{x_1}{\varphi_1(x_1)} & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Après de simples calculs, (2.23) devient :

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{x}_1 \\ \dot{x}_3 \\ \dot{x}_5 \end{pmatrix} = \begin{pmatrix} x_3 \\ -x_5 x_1 \\ h_2(x_1, x_3, x_5) \end{pmatrix},$$

où $h_2(x_1, x_3, x_5) = \left(\frac{x_3}{\varphi_1(x_1)} - \frac{x_1 x_3 \dot{\varphi}_1(x_1)}{\varphi_1^2(x_1)}\right) x_3 - \frac{x_1}{\varphi_1(x_1)} x_5 x_1 + \mathfrak{g} - x_5 \varphi_1(x_1)$. Le théorème des fonctions implicites permet d'exprimer x_5 en fonction de x_1 et x_3 via une fonction $\varphi_3 : x_5 =$

$\varphi_3(x_1, x_3)$. Comme à l'étape précédente, cette fonction est explicite. Plus précisément :

$$\begin{aligned}
&\left(\frac{x_3}{\varphi_1(x_1)} - \frac{x_1 x_3 \dot{\varphi}_1(x_1)}{\varphi_1^2(x_1)}\right) x_3 - \frac{x_1}{\varphi_1(x_1)} x_5 x_1 + \mathfrak{g} - x_5 \varphi_1(x_1) = 0 \\
\Leftrightarrow\ &\frac{x_3^2}{\varphi_1^2(x_1)} \left(\varphi_1(x_1) - x_1 \dot{\varphi}_1(x_1)\right) - x_5 \left(\frac{x_1^2}{\varphi_1(x_1)} + \varphi_1(x_1)\right) + \mathfrak{g} = 0 \\
\Leftrightarrow\ &\left(\frac{x_3}{\varphi_1(x_1)}\right)^2 \left(\varphi_1(x_1) + \frac{x_1^2}{\varphi_1(x_1)}\right) - x_5 \left(\frac{x_1^2 + \varphi_1^2(x_1)}{\varphi_1(x_1)}\right) + \mathfrak{g} = 0 \\
\Leftrightarrow\ &\left(\frac{x_3}{\varphi_1(x_1)}\right)^2 \left(\frac{1}{\varphi_1(x_1)}\right) - \frac{x_5}{\varphi_1(x_1)} + \mathfrak{g} = 0 \\
\Leftrightarrow\ &\left(\frac{x_3}{\varphi_1(x_1)}\right)^2 - x_5 + \mathfrak{g}\varphi_1(x_1) = 0 \\
\Leftrightarrow\ &x_5 = \mathfrak{g}\varphi_1(x_1) + \left(\varphi_1(x_1)^{-1} x_3\right)^2.
\end{aligned}$$

On termine la troisième étape en obtenant l'équation différentielle :

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_3 \end{pmatrix} = \begin{pmatrix} x_3 \\ -\left[\mathfrak{g}\varphi_1(x_1) + \left(\varphi_1(x_1)^{-1} x_3\right)^2\right] x_1 \end{pmatrix}.$$

2. Le déroulement de la méthode étant similaire, on ne donne pas les résultats intermédiaires.

Étape 1 - Puisque $|x_1| > |x_2|$, $x_1 \neq 0$. Par conséquent, il existe une fonction ϕ_1 telle que $x_1 = \phi_1(x_2)$, où $0 = x_1^2 + x_2^2$. La première étape de la méthode fournit :

$$X_1 = \begin{pmatrix} x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}, \quad E_1(X_1) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\phi_1(x_2)^{-1} x_2 & 0 & 0 & 0 \end{bmatrix} \quad \text{et} \quad f_1(X_1) = \begin{pmatrix} x_4 \\ -x_5 \phi_1(x_2) \\ \mathfrak{g} - x_5 x_2 \\ x_3 \end{pmatrix}.$$

Étape 2 - La nouvelle contrainte algébrique obtenue est $x_3 = -\phi_1(x_2)^{-1} x_2 x_4$. La seconde étape fournit :

$$X_2 = \begin{pmatrix} x_2 \\ x_4 \\ x_5 \end{pmatrix}, \quad f_2(X_2) = \begin{pmatrix} x_4 \\ \mathfrak{g} - x_5 x_2 \\ -x_5 \phi_1(x_2) \end{pmatrix}$$

et

$$E_2(X_2) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\left(\frac{x_4}{\phi_1(x_2)} - \frac{x_2 x_4 \dot{\phi}_1(x_2)}{\phi_1^2(x_2)}\right) & -\frac{x_2}{\phi_1(x_2)} & 0 \end{bmatrix}.$$

Étape 3 - La dernière contrainte algébrique est $0 = \left(\frac{x_4}{\phi_1(x_2)} - \frac{x_2 x_4 \dot{\phi}_1(x_2)}{\phi_1^2(x_2)}\right) x_4 + \frac{x_2}{\phi_1(x_2)}(\mathfrak{g} - x_5 x_2) - x_5 \phi_1(x_2)$. Elle permet d'exprimer la variable x_5 en fonction des variables x_2 et x_4 :

$$x_5 = \mathfrak{g} x_2 + \left(\phi_1(x_2)^{-1} x_4\right)^2.$$

On récupère l'EDO

$$\begin{pmatrix} \dot{x}_2 \\ \dot{x}_4 \end{pmatrix} = \begin{pmatrix} x_4 \\ \mathfrak{g} - \left[\mathfrak{g} x_2 + \left(\phi_1(x_2)^{-1} x_4\right)^2\right] x_2 \end{pmatrix}.$$

□

Remarque 15

On montre aisément que les deux expressions de x_5 (fournies par les points 1 et 2 du théorème 15) sont égales.

2.3.3.1.2 Coordonnées polaires Il est possible de retrouver l'EDO du pendule simple en dimension $n = 2$. Montrons-le par exemple pour le premier point du théorème précédent. Soit $\theta \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$. On pose $x_1 = \sin\theta$. On a alors :

- $x_2 = \varphi_1(x_1) = \sqrt{1 - x_1^2} = \cos\theta \neq 0$;
- $x_3 = \dot{x}_1 = \dot{\theta}\cos\theta$;
- $\dot{x}_3 = \ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta$.

On a donc d'une part :

$$\ddot{x}_1 = \ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta. \tag{2.24}$$

D'autre part, le système (2.20) fournit :

$$\ddot{x}_1 = -\left[\mathfrak{g}\varphi_1(x_1) + \left(\varphi_1(x_1)^{-1}x_3\right)^2\right]x_1.$$

On remplace dans l'expression précédente les valeurs de x_1, $\varphi_1(x_1)$ et x_3. On obtient après simplifications :

$$\ddot{x}_1 = -\mathfrak{g}\sin\theta\cos\theta - \dot{\theta}^2\sin\theta. \tag{2.25}$$

En comparant (2.24) et (2.25), on parvient à :

$$\ddot{\theta} = -\mathfrak{g}\sin\theta.$$

On retrouve bien l'équation classique du pendule simple.

2.3.3.1.3 Résolution numérique Nous avons écrit une procédure, nommée `simulation` (disponible en annexe, programmée grâce au logiciel MAPLE), qui permet d'exploiter de manière numérique les équations fournies par le théorème 15. Ce théorème donne en réalité les équations du mouvement du pendule sur les deux cartes locales nécessaires pour décrire l'ensemble du cercle d'équation $0 = x_1^2 + x_2^2 - 1$. Le point d'intérêt de `simulation` consiste à se positionner sur la bonne carte locale au moment adéquat. La procédure `simulation` commence par déterminer la carte locale ($|x_1| > |x_2|$ ou $|x_2| > |x_1|$) correspondant aux conditions initiales fournies. L'intégrateur démarre et s'exécute sur un petit pas de temps fixé au départ. Il s'agit d'intégrer l'EDO mise en avant par la méthode de déflation, et non une EDA (via un schéma de Runge-Kutta d'ordre 4 et 5). La position du point courant obtenu est analysée : s'il appartient à la même carte locale, l'intégrateur s'exécute sur un pas de temps supplémentaire, sinon il bascule sur la deuxième carte avant de s'exécuter. Tout au long de l'intégration, `simulation` bascule d'une carte locale à l'autre et décrit ainsi l'ensemble du système.

La procédure `simulation` prend en entrée les deux systèmes à intégrer (correspondants aux deux cartes locales), l'intervalle d'intégration, le pas d'intégration et les conditions initiales. Elle fournit en sortie les temps discrétisés ainsi que les itérés des $x_1, \ldots, x_5$.

On commence par lâcher le pendule de la position $(1, 0)$ avec une vitesse initiale nulle $(0, 0)$. On prend $x_5 = \mathfrak{g}x_2 + \left(x_1^{-1}x_4\right)^2 = 0$. Les systèmes `sys1` et `sys2` sont donnés par les équations du théorème 15. On trace les solutions grâce à la commande `traceSimulation`.

```
> L := [sys1, sys2] :
> CI := [1, 0, 0, 0, 0] :
> P := simulation(L, 5, 0.01, CI) :
> traceSimulation(P) :
```

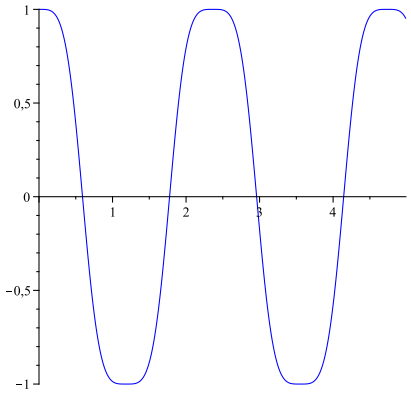


FIGURE 2.1 – $x_1(t)$ en fonction de t

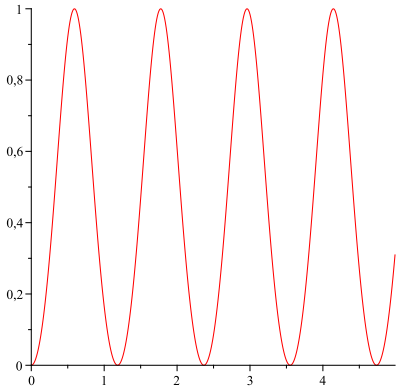


FIGURE 2.2 – $x_2(t)$ en fonction de t

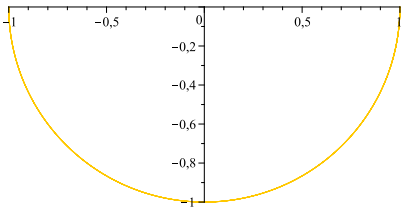


FIGURE 2.3 – Courbe paramétrée $(x_1(t), x_2(t))$

On prend un autre jeu de conditions initiales. Cette fois-ci, on donne une vitesse non nulle $(0, 5)$ au pendule. On prend $x_5 = \mathfrak{g}x_2 + \left(x_1^{-1}x_4\right)^2 = 25$.

```
> L := [sys1, sys2] :
> CI := [1, 0, 0, 5, 25] :
> P := simulation(L, 5, 0.01, CI) :
> traceSimulation(P) :
```

On obtient alors les courbes suivantes :

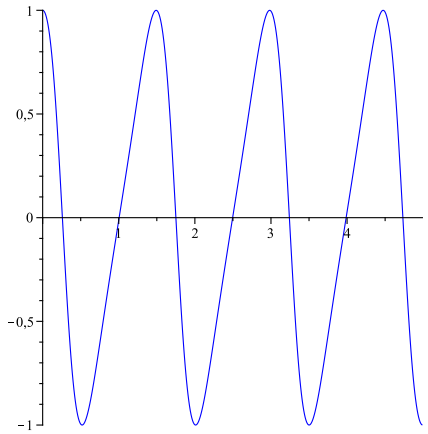


FIGURE 2.4 – $x_1(t)$ en fonction de t

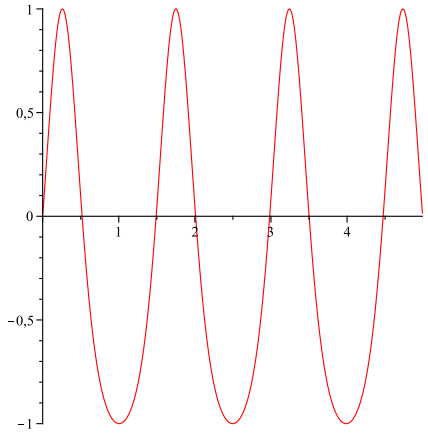


FIGURE 2.5 – $x_2(t)$ en fonction de t

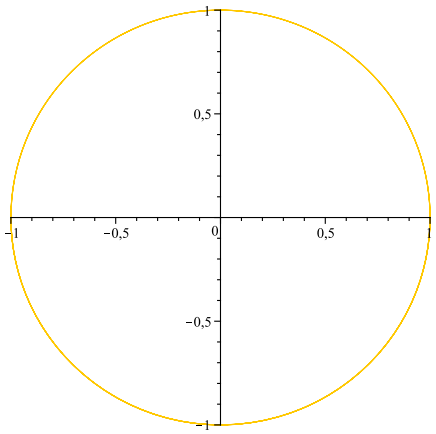


FIGURE 2.6 – Courbe paramétrée $(x_1(t), x_2(t))$

La vitesse initiale est ici suffisante pour que le pendule effectue une révolution.

2.3.3.2 Pendule en dimension 3

On ne donne ici que les résultats obtenus par la méthode de déflation, les calculs étant similaires à ceux de la dimension $n = 2$. On effectue par ailleurs le changement de coordonnées sphériques afin de retrouver les équations classiques du pendule 3D.

2.3.3.2.1 Résolution formelle Le mouvement d'un pendule simple pesant en dimension $n = 3$ est décrit par l'EDA (2.15), où

$$X_0 = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix}, \quad E_0(X_0) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{et} \quad f_0(X_0) = \begin{pmatrix} x_4 \\ x_5 \\ x_6 \\ -x_7 x_1 \\ -x_7 x_2 \\ \mathfrak{g} - x_7 x_3 \\ x_1^2 + x_2^2 + x_3^2 - 1 \end{pmatrix}.$$

Théorème 16

La réduction du problème précédent est obtenue après trois étapes de la méthode de déflation.

1. *Si $|x_3| > |x_1|$ et $|x_3| > |x_2|$, alors on obtient l'EDO sous contraintes suivante :*

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_4 \\ \dot{x}_5 \end{pmatrix} = \begin{pmatrix} x_4 \\ x_5 \\ -\left[\mathfrak{g}\varphi_1(x_1, x_2) + x_4^2 + x_5^2 + \left(\varphi_1(x_1, x_2)^{-1}(x_1 x_4 + x_2 x_5)\right)^2\right] x_1 \\ -\left[\mathfrak{g}\varphi_1(x_1, x_2) + x_4^2 + x_5^2 + \left(\varphi_1(x_1, x_2)^{-1}(x_1 x_4 + x_2 x_5)\right)^2\right] x_2 \end{pmatrix}, \quad (2.26)$$

   avec $x_3 = \varphi_1(x_1, x_2)$ tel que $0 = x_1^2 + x_2^2 + x_3^2 - 1$, $x_6 = -\varphi_1(x_1, x_2)^{-1}(x_1 x_4 + x_2 x_5)$ et $x_7 = \mathfrak{g}\varphi_1(x_1, x_2) + x_4^2 + x_5^2 + \left(\varphi_1(x_1, x_2)^{-1}(x_1 x_4 + x_2 x_5)\right)^2$.

2. *Si $|x_2| > |x_1|$ et $|x_2| > |x_3|$, alors on obtient l'EDO sous contraintes suivante :*

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_6 \end{pmatrix} = \begin{pmatrix} x_4 \\ x_6 \\ -\left[\mathfrak{g}x_3 + x_4^2 + \left(\phi_1(x_1, x_3)^{-1}(x_1 x_4 + x_3 x_6)\right)^2 + x_6^2\right] x_1 \\ \mathfrak{g} - \left[\mathfrak{g}x_3 + x_4^2 + \left(\phi_1(x_1, x_3)^{-1}(x_1 x_4 + x_3 x_6)\right)^2 + x_6^2\right] x_3 \end{pmatrix}, \quad (2.27)$$

   avec $x_2 = \phi_1(x_1, x_3)$ tel que $0 = x_1^2 + x_2^2 + x_3^2 - 1$, $x_5 = -\phi_1(x_1, x_3)^{-1}(x_1 x_4 + x_3 x_6)$ et $x_7 = \mathfrak{g}x_3 + x_4^2 + \left(\phi_1(x_1, x_3)^{-1}(x_1 x_4 + x_3 x_6)\right)^2 + x_6^2$.

3. *Si $|x_1| > |x_2|$ et $|x_1| > |x_3|$, alors on obtient l'EDO sous contraintes suivante :*

$$\begin{pmatrix} \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_5 \\ \dot{x}_6 \end{pmatrix} = \begin{pmatrix} x_5 \\ x_6 \\ -\left[\mathfrak{g}x_3 + \left(\psi_1(x_2, x_3)^{-1}(x_2 x_5 + x_3 x_6)\right)^2 + x_5^2 + x_6^2\right] x_2 \\ \mathfrak{g} - \left[\mathfrak{g}x_3 + \left(\psi_1(x_2, x_3)^{-1}(x_2 x_5 + x_3 x_6)\right)^2 + x_5^2 + x_6^2\right] x_3 \end{pmatrix}, \quad (2.28)$$

   avec $x_1 = \psi_1(x_2, x_3)$ tel que $0 = x_1^2 + x_2^2 + x_3^2 - 1$, $x_4 = -\psi_1(x_2, x_3)^{-1}(x_2 x_5 + x_3 x_6)$ et $x_7 = \mathfrak{g}x_3 + \left(\psi_1(x_2, x_3)^{-1}(x_2 x_5 + x_3 x_6)\right)^2 + x_5^2 + x_6^2$.

2.3.3.2.2 Coordonnées sphériques Observons par exemple le changement de coordonnées sur le premier point du théorème précédent. Soit $\theta \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$. On pose :

- $x_1 = \sin\theta \sin\psi$;
- $x_2 = \sin\theta \cos\psi$.

On a alors :

- $x_3 = \varphi_1(x_1, x_2) = \sqrt{1-(\sin\theta\sin\psi)^2-(\sin\theta\cos\psi)^2} = \cos\theta$;
- $x_4 = \dot{x}_1 = \dot{\theta}\cos\theta\sin\psi + \dot{\psi}\sin\theta\cos\psi$;
- $x_5 = \dot{x}_2 = \dot{\theta}\cos\theta\cos\psi - \dot{\psi}\sin\theta\sin\psi$.

On a donc d'une part :

$$\begin{cases} \ddot{x}_1 = \ddot{\theta}\cos\theta\sin\psi + \ddot{\psi}\sin\theta\cos\psi - \left(\dot{\theta}^2+\dot{\psi}^2\right)\sin\theta\sin\psi + 2\dot{\theta}\dot{\psi}\cos\theta\cos\psi & (2.29a) \\ \ddot{x}_2 = \ddot{\theta}\cos\theta\cos\psi - \ddot{\psi}\sin\theta\sin\psi - \left(\dot{\theta}^2+\dot{\psi}^2\right)\sin\theta\cos\psi - 2\dot{\theta}\dot{\psi}\cos\theta\sin\psi. & (2.29b) \end{cases}$$

D'autre part, le système (2.26) fournit après simplifications :

$$\begin{cases} \ddot{x}_1 = -\dot{\theta}^2\sin\theta\sin\psi - \dot{\psi}^2\sin^3\theta\sin\psi - \mathfrak{g}\cos\theta\sin\theta\sin\psi & (2.30a) \\ \ddot{x}_2 = -\dot{\theta}^2\sin\theta\cos\psi - \dot{\psi}^2\sin^3\theta\cos\psi - \mathfrak{g}\cos\theta\sin\theta\cos\psi. & (2.30b) \end{cases}$$

En comparant (2.29a) à (2.30a) et (2.29b) à (2.30b), on parvient après simplifications à :

$$\ddot{\theta}\cos\theta\sin\psi + \ddot{\psi}\sin\theta\cos\psi - \dot{\psi}^2\cos^2\theta\sin\theta\sin\psi + 2\dot{\theta}\dot{\psi}\cos\theta\cos\psi + \mathfrak{g}\cos\theta\sin\theta\sin\psi = 0 \quad (2.31)$$

$$\ddot{\theta}\cos\theta\cos\psi - \ddot{\psi}\sin\theta\sin\psi - \dot{\psi}^2\cos^2\theta\sin\theta\cos\psi - 2\dot{\theta}\dot{\psi}\cos\theta\sin\psi + \mathfrak{g}\cos\theta\sin\theta\cos\psi = 0 \quad (2.32)$$

On multiplie (2.31) par $\sin\psi$, (2.32) par $\cos\psi$ puis on somme les deux. On obtient :

$$\ddot{\theta}\cos\theta - \dot{\psi}^2\cos^2\theta\sin\theta + \mathfrak{g}\cos\theta\sin\theta = 0.$$

Enfin, on multiplie (2.31) par $\cos\psi$, (2.32) par $-\sin\psi$ puis on somme les deux. On obtient :

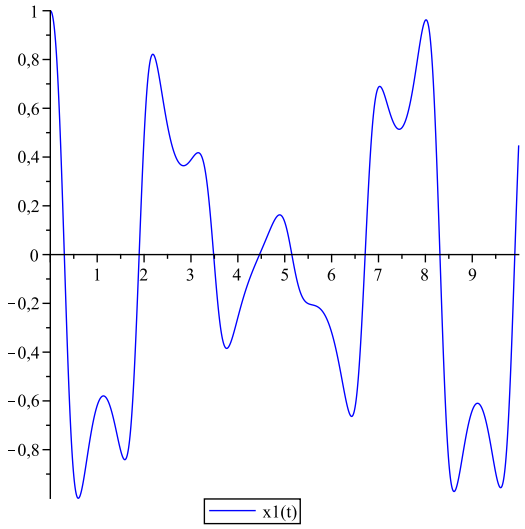
$$\ddot{\psi}\sin\theta + 2\dot{\theta}\dot{\psi}\cos\theta = 0.$$

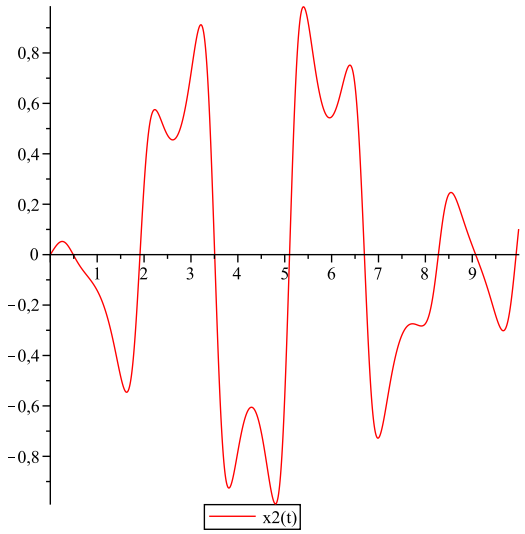
On retrouve bien les équations classiques du pendule sphérique, à savoir :

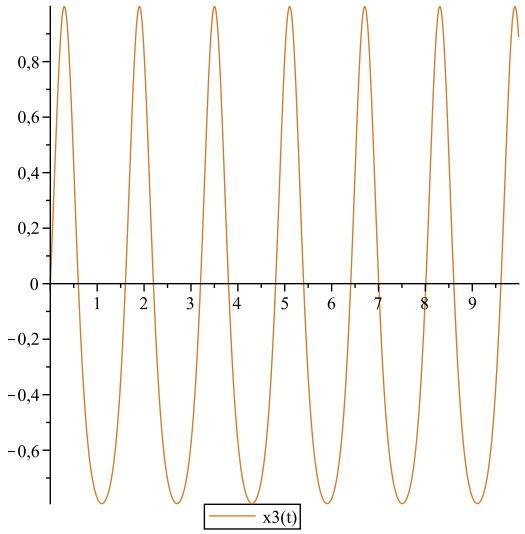
$$\begin{cases} \ddot{\theta} = \dot{\psi}^2\cos\theta\sin\theta - \mathfrak{g}\sin\theta \\ \ddot{\psi} = -2\dot{\theta}\dot{\psi}\tan^{-1}\theta. \end{cases}$$

2.3.3.2.3 Résolution numérique On adapte la procédure `simulation` pour le pendule en trois dimensions (on garde les mêmes notations par souci de simplicité). Il est à propos de considérer cette fois trois systèmes. Ces derniers sont décrits dans le théorème 16. Il faut donc jongler entre trois cartes locales. On part pour commencer de la position initiale $(1, 0, 0)$ et de la vitesse initiale $(0, 0.3, 4)$.

```
> L := [sys1, sys2, sys3] :
> CI := [1, 0, 0, 0, 0.3, 4, 16.09] :
> P := simulation(L, 10, 0.01, CI) :
> traceSimulation(P) :
```







Les trois courbes ci-dessus représentent les coordonnées $x_1(t)$, $x_2(t)$ et $x_3(t)$. On termine par la représentation paramétrique donnant le mouvement du pendule sur la sphère :

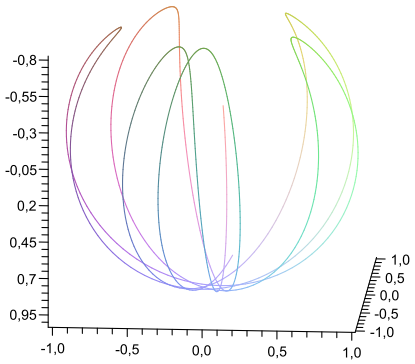


FIGURE 2.7 – Courbe paramétrée $(x_1(t), x_2(t), x_3(t))$

2.3.3.3 Problèmes multi-corps et pendule en dimension n

Les équations du mouvement des systèmes contraints multi-corps peuvent être obtenues par le formalisme d'Euler-Lagrange [2]. Si les coordonnées cartésiennes sont utilisées, les équations engendrent alors une EDA quasi-linéaire de la forme

$$\begin{cases} \dot{p} = v \\ M(p)\dot{v} = \alpha(p,v) - D_p g(p)^\top \lambda \\ 0 = g(p), \end{cases} \tag{2.33}$$

où $p \in \mathbb{R}^n$ est le vecteur position, $v \in \mathbb{R}^n$ est le vecteur vitesse et $\lambda \in \mathbb{R}^m$ représente le multiplicateur de Lagrange associé à la contrainte $0 = g(p) \in \mathbb{R}^m$. Soit $m \leqslant n$. Supposons que la matrice de masse $M(p)$ est symétrique définie positive et que la matrice Jacobienne $D_p g(p)$ est de rang plein. La forme matricielle du système correspond au problème (2.15), où

$$X_0 = \begin{pmatrix} p \\ v \\ \lambda \end{pmatrix}, \quad E_0(X_0) = \begin{bmatrix} I_n & 0 & 0 \\ 0 & I_n & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{et} \quad f_0(X_0) = \begin{pmatrix} v \\ M(p)^{-1}\left(\alpha(p,v) - D_p g(p)^\top \lambda\right) \\ g(p) \end{pmatrix}.$$

On applique le méthode de déflation à cette EDA. On introduit dans ce but les notations suivantes : $\mathbf{x}_1 := x_{1:n-m}$ et $\mathbf{x}_2 := x_{n-m+1:n}$ où $x \in \mathbb{R}^n$.

Théorème 17

Supposons que la matrice $D_2 g(\mathbf{p}_1, \mathbf{p}_2)$ est inversible et introduisons les fonctions $\mathbf{p}_2 = \varphi_1(\mathbf{p}_1)$ telle que $g(\mathbf{p}_1, \mathbf{p}_2) = 0$ et $\varphi_2(\mathbf{p}_1, \mathbf{v}_1) = D\varphi_1(\mathbf{p}_1)\mathbf{v}_1$. Introduisons également les matrices $\left[M(p)^{-1}\right]_1 \in \mathbb{R}^{(n-m)\times n}$ et $\left[M(p)^{-1}\right]_2 \in \mathbb{R}^{m\times n}$ telles que $M(p)^{-1} = \begin{pmatrix} \left[M(p)^{-1}\right]_1 \\ \left[M(p)^{-1}\right]_2 \end{pmatrix}$, et pour terminer la matrice $\Delta := \left[M(p)^{-1}\right]_2 - D_2\varphi_2(\mathbf{p}_1, \mathbf{v}_1)\left[M(p)^{-1}\right]_1 \in \mathbb{R}^{m\times n}$. Alors

- *la matrice $\Delta D_p g(p)^\top$ est inversible ;*
- *le nombre d'étapes de l'algorithme de déflation appliqué à l'EDA (2.33) est égal à 3. De plus, la solution générale de cette EDA satisfait l'EDO*

$$\begin{pmatrix} \dot{\mathbf{p}}_1 \\ \dot{\mathbf{v}}_1 \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1 \\ \left[M(p)^{-1}\right]_1 \left(\alpha(p,v) - D_p g(p)^\top \lambda\right) \end{pmatrix}, \quad \text{avec} \quad p = (\mathbf{p}_1, \varphi_1(\mathbf{p}_1)), \tag{2.34}$$

avec $v = (\mathbf{v}_1, \varphi_2(\mathbf{p}_1, \mathbf{v}_1))$ et $\lambda = \left(\Delta D_p g(p)^\top\right)^{-1}\left(\Delta\alpha(p,v) - D_1\varphi_2(\mathbf{p}_1, \mathbf{v}_1)\mathbf{v}_1\right)$.

Preuve - Commençons par démontrer que $\Delta D_p g(p)^\top$ est une matrice inversible. Puisque $D_2\varphi_2(\mathbf{p}_1, \mathbf{v}_1) = D\varphi_1(\mathbf{p}_1)$, on a

$$\begin{aligned}
\Delta D_p g(p)^\top &= \left(\left[M(p)^{-1}\right]_2 - D_2\varphi_2(\mathbf{p}_1, \mathbf{v}_1)\left[M(p)^{-1}\right]_1\right) D_p g(p)^\top \\
&= \left(\left[M(p)^{-1}\right]_2 - D\varphi_1(\mathbf{p}_1)\left[M(p)^{-1}\right]_1\right) D_p g(p)^\top \\
&= \left(\left[D_2 g(\mathbf{p}_1, \mathbf{p}_2)\right]^{-1} D_1 g(\mathbf{p}_1, \mathbf{p}_2)\left[M(p)^{-1}\right]_1 + \left[M(p)^{-1}\right]_2\right) D_p g(p)^\top \\
&= \left[D_2 g(\mathbf{p}_1, \mathbf{p}_2)\right]^{-1}\left(D_1 g(\mathbf{p}_1, \mathbf{p}_2)\left[M(p)^{-1}\right]_1 + D_2 g(\mathbf{p}_1, \mathbf{p}_2)\left[M(p)^{-1}\right]_2\right) D_p g(p)^\top \\
&= D_2 g(p)^{-1} D_p g(p) M(p)^{-1} D_p g(p)^\top.
\end{aligned}$$

Par un résultat classique d'algèbre linéaire ([31], 210), on observe

$$\begin{aligned}
rang\left(\Delta D_p g(p)^\top\right) &= rang\left(D_p g(p) M(p)^{-1} D_p g(p)^\top\right) \\
&= rang\left(D_p g(p)^\top\right) - dim\left(\ker\left(D_p g(p) M(p)^{-1}\right) \cap im\left(D_p g(p)^\top\right)\right).
\end{aligned}$$

Montrons à présent que $\ker\left(D_p g(p) M(p)^{-1}\right) \cap \operatorname{im}\left(D_p g(p)^\top\right) = \{0\}$. Soit un vecteur x tel que $D_p g(p) M(p)^{-1} x = 0$ et $x = D_p g(p)^\top y$. Alors $x^\top M(p)^{-1} x = y^\top D_p g(p) M(p)^{-1} x = 0$. Soit $x = M(p)z$; ceci est équivalent à $z^\top M(p) z = 0$. Puisque $M(p)$ est une matrice symétrique définie positive, on déduit que $z = 0$ et ainsi $x = 0$. On obtient alors $rang\left(\Delta D_p g(p)^\top\right) = rang\left(D_p g(p)^\top\right) = m$ puisque $D_p g(p)^\top$ est de rang plein. La matrice $\Delta D_p g(p)^\top$ est ainsi non singulière.

On applique l'algorithme `QL_Deflation` au problème (2.33) :
Étape 1 - Par le théorème des fonctions implicites, on peut écrire

$$\mathbf{p}_2 = \varphi_1(\mathbf{p}_1) \quad \text{et} \quad g(\mathbf{p}_1, \varphi_1(\mathbf{p}_1)) = 0.$$

On écrit dès lors $\mathbf{p}_2$ pour désigner $\varphi_1(\mathbf{p}_1)$. Il s'en suit $\dot{\mathbf{p}}_2 = D\varphi_1(\mathbf{p}_1)\dot{\mathbf{p}}_1$, *i.e.*

$$D\varphi_1(\mathbf{p}_1)\dot{\mathbf{p}}_1 = \mathbf{v}_2,$$

avec $D\varphi_1(\mathbf{p}_1) = -D_2 g(\mathbf{p}_1, \mathbf{p}_2)^{-1} D_1 g(\mathbf{p}_1, \mathbf{p}_2)$. Alors, l'EDA déflatée est

$$\begin{bmatrix} I_{n-m} & 0 & 0 \\ 0 & I_n & 0 \\ D\varphi_1(\mathbf{p}_1) & 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{\mathbf{p}}_1 \\ \dot{v} \\ \dot{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1 \\ M(p)^{-1}\left(\alpha(p, v) - D_p g(p)^\top \lambda\right) \\ \mathbf{v}_2 \end{pmatrix} \tag{2.35}$$

et la première contrainte algébrique est $\mathbf{p}_2 = \varphi_1(\mathbf{p}_1)$.

Étape 2 - On a

$$\begin{bmatrix} I_{n-m} & 0 & 0 \\ 0 & I_n & 0 \\ D\varphi_1(\mathbf{p}_1) & 0 & 0 \end{bmatrix} = \begin{bmatrix} I_{n-m} & 0 & 0 \\ 0 & I_n & 0 \\ D\varphi_1(\mathbf{p}_1) & 0 & I_m \end{bmatrix} \begin{bmatrix} I_{n-m} & 0 & 0 \\ 0 & I_n & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

L'équation (2.35) est alors équivalente à

$$\begin{bmatrix} I_{n-m} & 0 & 0 \\ 0 & I_n & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{\mathbf{p}}_1 \\ \dot{v} \\ \dot{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1 \\ M(p)^{-1}\left(\alpha(p, v) - D_p g(p)^\top \lambda\right) \\ -D\varphi_1(\mathbf{p}_1)\mathbf{v}_1 + \mathbf{v}_2 \end{pmatrix}.$$

La contrainte $\mathbf{v}_2 = D\varphi_1(\mathbf{p}_1)\mathbf{v}_1 := \varphi_2(\mathbf{p}_1, \mathbf{v}_1)$ implique

$$\dot{\mathbf{v}}_2 = D_1\varphi_2(\mathbf{p}_1, \mathbf{v}_1)\dot{\mathbf{p}}_1 + D_2\varphi_2(\mathbf{p}_1, \mathbf{v}_1)\dot{\mathbf{v}}_1.$$

Des calculs directs amènent à

$$D_1\varphi_2(\mathbf{p}_1, \mathbf{v}_1) = D_2g^{-1}\left((D_{21}gD_2g^{-1}D_1g - D_{11}g)\mathbf{v}_1 + (D_{12}g - D_{22}gD_2g^{-1}D_1g)\mathbf{v}_1D_2g^{-1}D_1g\right),$$
$$D_2\varphi_2(\mathbf{p}_1, \mathbf{v}_1) = -D_2g^{-1}D_1g,$$

où g signifie $g(\mathbf{p}_1, \mathbf{p}_2)$. La seconde équation déflatée est alors

$$\begin{bmatrix} I_{n-m} & 0 & 0 \\ 0 & I_{n-m} & 0 \\ D_1\varphi_2(\mathbf{p}_1, \mathbf{v}_1) & D_2\varphi_2(\mathbf{p}_1, \mathbf{v}_1) & 0 \end{bmatrix} \begin{pmatrix} \dot{\mathbf{p}}_1 \\ \dot{\mathbf{v}}_1 \\ \dot{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{h}(p, v, \lambda)_1 \\ \mathbf{h}(p, v, \lambda)_2 \end{pmatrix}, \tag{2.36}$$

où $h(p, v, \lambda) = M(p)^{-1}\left(\alpha(p, v) - D_pg(p)^\top\lambda\right)$.

Étape 3 - On a

$$\begin{bmatrix} I_{n-m} & 0 & 0 \\ 0 & I_{n-m} & 0 \\ D_1\varphi_2 & D_2\varphi_2 & 0 \end{bmatrix} = \begin{bmatrix} I_{n-m} & 0 & 0 \\ 0 & I_{n-m} & 0 \\ D_1\varphi_2 & D_2\varphi_2 & I_m \end{bmatrix} \begin{bmatrix} I_{n-m} & 0 & 0 \\ 0 & I_{n-m} & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

où $D_i\varphi_2$ signifie $D_i\varphi_2(\mathbf{p}_1, \mathbf{v}_1)$, pour $i = 1, 2$. L'équation (2.36) est ainsi équivalente à

$$\begin{bmatrix} I_{n-m} & 0 & 0 \\ 0 & I_{n-m} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{\mathbf{p}}_1 \\ \dot{\mathbf{v}}_1 \\ \dot{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{h}(p, v, \lambda)_1 \\ A(\mathbf{p}_1, \mathbf{v}_1, \lambda) \end{pmatrix},$$

où

$$\begin{aligned} A(\mathbf{p}_1, \mathbf{v}_1, \lambda) &= -D_1\varphi_2(\mathbf{p}_1, \mathbf{v}_1)\mathbf{v}_1 - D_2\varphi_2(\mathbf{p}_1, \mathbf{v}_1)\mathbf{h}(p, v, \lambda)_1 + \mathbf{h}(p, v, \lambda)_2 \\ &= -D_1\varphi_2(\mathbf{p}_1, \mathbf{v}_1)\mathbf{v}_1 + \Delta\left(\alpha(p, v) - D_pg(p)^\top\lambda\right). \end{aligned}$$

Puisque la matrice $\Delta D_pg(p)^\top$ est inversible, on déduit l'expression de λ de $A(\mathbf{p}_1, \mathbf{v}_1, \lambda) = 0$:

$$\lambda = \left(\Delta D_pg(p)^\top\right)^{-1}\left(\Delta\alpha(p, v) - D_1\varphi_2(\mathbf{p}_1, \mathbf{v}_1)\mathbf{v}_1\right).$$

On peut enfin conclure par la dernière équation déflatée

$$\begin{pmatrix} \dot{\mathbf{p}}_1 \\ \dot{\mathbf{v}}_1 \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1 \\ \left[M(p)^{-1}\right]_1\left(\alpha(p, v) - D_pg(p)^\top\lambda\right) \end{pmatrix}.$$

□

Le besoin de régularité dans la méthode de déflation exige l'inversibilité de la matrice $D_2g(\mathbf{p}_1, \mathbf{p}_2)$ où $\mathbf{p}_2 = p_{n-m+1:n}$. Dans un cas plus général, considérons une matrice de permutation γ telle que $\gamma p = \bar{p} = \left(\bar{\mathbf{p}}_1^\top, \bar{\mathbf{p}}_2^\top\right)^\top$, où la matrice $D_2g(\bar{\mathbf{p}}_1, \bar{\mathbf{p}}_2)$ est inversible. En utilisant $p = \gamma^{-1}\bar{p}$ et $v = \gamma^{-1}\bar{v}$, le système (2.33) devient

$$\begin{bmatrix} \gamma^{-1} & 0 & 0 \\ 0 & \gamma^{-1} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} \dot{\bar{p}} \\ \dot{\bar{v}} \\ \dot{\lambda} \end{pmatrix} = \begin{pmatrix} \gamma^{-1}\,\bar{v} \\ M(p)^{-1}\left(\alpha(p, v) - D_pg(p)^\top\lambda\right) \\ g(p) \end{pmatrix}. \tag{2.37}$$

En pré-multipliant (2.37) par $\begin{bmatrix}\gamma & 0 & 0\\ 0 & \gamma & 0\\ 0 & 0 & I_m\end{bmatrix}$, on obtient

$$\begin{bmatrix}I_n & 0 & 0\\ 0 & I_n & 0\\ 0 & 0 & 0\end{bmatrix}\begin{pmatrix}\dot{\bar{p}}\\ \dot{\bar{v}}\\ \dot{\lambda}\end{pmatrix} = \begin{pmatrix}\bar{v}\\ \gamma M(p)^{-1}\left(\alpha(p,v) - D_p g(p)^\top \lambda\right)\\ g(p)\end{pmatrix}.$$

Par conséquent, le théorème 17 s'applique avec

$$\begin{cases} \bar{p} = \gamma p = \begin{pmatrix}\bar{\mathbf{p}}_1\\ \bar{\mathbf{p}}_2\end{pmatrix}\\ \bar{v} = \gamma v = \begin{pmatrix}\bar{\mathbf{v}}_1\\ \bar{\mathbf{v}}_2\end{pmatrix}\\ M_\gamma(\bar{p}) = \gamma M(\gamma^{-1}\bar{p})\gamma^{-1} = \gamma M(p)\gamma^{-1}\\ \alpha_\gamma(\bar{p},\bar{v}) = \gamma\alpha(\gamma^{-1}\bar{p},\gamma^{-1}\bar{v}) = \gamma\alpha(p,v)\\ D_{\bar{p}}g(\bar{p}) = D_p g(\gamma^{-1}\bar{p})\gamma^{-1} = D_p g(p)\gamma^{-1}. \end{cases}$$

On retrouve par application du théorème 17 le cas du pendule simple en dimension n, modélisé par le système

$$\begin{cases} \dot{x}_{1:n} = x_{n+1:2n}\\ \dot{x}_{n+1:2n-1} = -x_{2n+1}x_{1:n-1}\\ \dot{x}_{2n} = \mathfrak{g} - x_{2n+1}x_n\\ 0 = \sum_{i=1}^{n} x_i^2 - 1. \end{cases} \tag{2.38}$$

Théorème 18

La réduction de (2.38) est obtenue après 3 étapes de l'algorithme de déflation.

1. *Si $|x_n| \geqslant |x_i|$ pour tout $i \neq n$, la solution générale de (2.38) satisfait le système différentiel ordinaire*

$$\begin{cases} \dot{x}_{1:n-1} = x_{n+1:2n-1}\\ \dot{x}_{n+1:2n-1} = -x_{2n+1}x_{1:n-1}, \end{cases} \tag{2.39}$$

avec $x_{2n+1} = \mathfrak{g}x_n + \sum_{i=1}^{n} x_{n+i}^2$, $\quad x_{2n} = -\frac{1}{x_n}\sum_{i=1}^{n-1} x_i x_{n+i}$ $\quad$ *et* $\quad x_n = \pm\sqrt{1 - \sum_{i=1}^{n-1} x_i^2}$.

2. *S'il existe $k < n$ tel que $|x_k| \geqslant |x_i|$ pour tout $i \neq k$, la solution générale de (2.38) satisfait le système différentiel ordinaire*

$$\begin{cases} \dot{x}_{1:k-1} = x_{n+1:n+k-1}\\ \dot{x}_{k+1:n} = x_{n+k+1:2n}\\ \dot{x}_{n+1:n+k-1} = -x_{2n+1}x_{1:k-1}\\ \dot{x}_{n+k+1:2n-1} = -x_{2n+1}x_{k+1:n-1}\\ \dot{x}_{2n} = \mathfrak{g} - x_{2n+1}x_n, \end{cases} \tag{2.40}$$

avec $x_{2n+1} = \mathfrak{g}x_n + \sum_{i=1}^{n} x_{n+i}^2$, $x_{n+k} = -\frac{1}{x_k}\sum_{i=1,i\neq k}^{n} x_i x_{n+i}$ *et* $x_k = \pm\sqrt{1 - \sum_{i=1,i\neq k}^{n} x_i^2}$.

Preuve - Le premier point est une application directe du théorème 17. Précisément, $m = 1$ et

$$\begin{aligned} p &= x_{1:n} \quad \text{avec} \quad \mathbf{p}_1 = x_{1:n-1}, \\ v &= x_{n+1:2n} \quad \text{avec} \quad \mathbf{v}_1 = x_{n+1:2n-1}, \\ \lambda &= \frac{x_{2n+1}}{2}. \end{aligned}$$

De plus, on a

$$\begin{aligned} M(p) &= I_n, \\ \alpha(p,v) &= \begin{pmatrix} 0 & \cdots & 0 & \mathfrak{g} \end{pmatrix}^\top, \\ g(p) &= \sum_{i=1}^{n} x_i^2 - 1. \end{aligned}$$

Clairement,

$$x_n = \mathbf{p}_2 = \varphi_1(\mathbf{p}_1) = \pm\sqrt{1 - \sum_{i=1}^{n-1} x_i^2} \neq 0 \quad \text{et} \quad x_{2n} = \mathbf{v}_2 = \varphi_2(\mathbf{p}_1, \mathbf{v}_1) = -\frac{1}{x_n}\sum_{i=1}^{n-1} x_i x_{n+i}.$$

Le théorème 17 fournit une expression du multiplicateur de Lagrange. Tout d'abord, on a

$$\begin{aligned} \left(\Delta D_p g(p)^\top\right)^{-1} &= \left[D_p g(p) D_p g(p)^\top\right]^{-1} D_2 g(p) \\ &= \left[\begin{pmatrix} 2x_1 & \cdots & 2x_n \end{pmatrix}\begin{pmatrix} 2x_1 & \cdots & 2x_n \end{pmatrix}^\top\right]^{-1} 2x_n \\ &= \left[4\sum_{i=1}^{n} x_i^2\right]^{-1} 2x_n \\ &= \frac{x_n}{2}. \end{aligned}$$

Ensuite, on obtient

$$D_1\varphi_2(\mathbf{p}_1, \mathbf{v}_1) = \begin{pmatrix} -\frac{x_{n+1}}{x_n} - \frac{x_1^2 x_{n+1}}{x_n^3} & \cdots & -\frac{x_{2n-1}}{x_n} - \frac{x_{n-1}^2 x_{2n-1}}{x_n^3} \end{pmatrix}.$$

Ainsi,

$$\begin{aligned} D_1\varphi_2(\mathbf{p}_1, \mathbf{v}_1)\mathbf{v}_1 &= \sum_{i=1}^{n-1}\left(-\frac{x_{n+i}}{x_n} - \frac{x_i^2 x_{n+i}}{x_n^3}\right) x_{n+i} \\ &= -\frac{1}{x_n}\sum_{i=1}^{n-1} x_{n+i}^2 - \frac{1}{x_n^3}\sum_{i=1}^{n-1} x_i^2 x_{n+i}^2 \\ &= -\frac{1}{x_n}\sum_{i=1}^{n-1} x_{n+i}^2 - \frac{x_{2n}^2}{x_n} \\ &= -\frac{1}{x_n}\sum_{i=1}^{n} x_{n+i}^2. \end{aligned}$$

Ceci conduit à

$$\begin{aligned} \Delta\alpha(p,v) - D_1\varphi_2(\mathbf{p}_1, \mathbf{v}_1)\mathbf{v}_1 &= D_2 g(p)^{-1} D_p g(p)\alpha(p,v) - D_1\varphi_2(\mathbf{p}_1, \mathbf{v}_1)\mathbf{v}_1 \\ &= \frac{1}{2x_n}\begin{pmatrix} 2x_1 & \cdots & 2x_{n-1} & 2x_n \end{pmatrix}\begin{pmatrix} 0 & \cdots & 0 & \mathfrak{g} \end{pmatrix}^\top + \frac{1}{x_n}\sum_{i=1}^{n} x_{n+i}^2 \\ &= \mathfrak{g} + \frac{1}{x_n}\sum_{i=1}^{n} x_{n+i}^2. \end{aligned}$$

Enfin,

$$x_{2n+1} = 2\lambda = 2\frac{x_n}{2}\left(\mathfrak{g} + \frac{1}{x_n}\sum_{i=1}^{n} x_{n+i}^2\right) = \mathfrak{g}x_n + \sum_{i=1}^{n} x_{n+i}^2.$$

On en déduit le système différentiel ordinaire

$$\begin{aligned}
\begin{pmatrix} \dot{x}_{1:n-1} \\ \dot{x}_{n+1:2n-1} \end{pmatrix} &= \begin{pmatrix} \dot{\mathbf{p}}_1 \\ \dot{\mathbf{v}}_1 \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{v}_1 \\ \left[M(p)^{-1}\right]_1 \left(\alpha(p,v) - D_p g(p)^\top \lambda\right) \end{pmatrix} \\
&= \begin{pmatrix} x_{n+1:2n-1} \\ \begin{pmatrix} I_{n-1} & 0_{1:n-1} \end{pmatrix} \begin{pmatrix} -x_{2n+1}x_1 & \cdots & -x_{2n+1}x_{n-1} & \mathfrak{g} - x_{2n+1}x_n \end{pmatrix}^\top \end{pmatrix} \\
&= \begin{pmatrix} x_{n+1:2n-1} \\ -x_{2n+1}x_{1:n-1} \end{pmatrix}.
\end{aligned}$$

Pour le second point, on utilise une simple permutation des variables. Considérons la matrice de permutation $\gamma = \begin{bmatrix} I_{k-1} & 0 & 0 \\ 0 & 0 & I_{n-k} \\ 0 & 1 & 0 \end{bmatrix}$. On a toujours $m = 1$ et

$$\begin{aligned}
\bar{p} = \gamma p = \begin{pmatrix} x_{1:k-1} \\ x_{k+1:n} \\ x_k \end{pmatrix} \quad \text{avec} \quad \bar{\mathbf{p}}_1 = \begin{pmatrix} x_{1:k-1} \\ x_{k+1:n} \end{pmatrix}, \\
\bar{v} = \gamma v = \begin{pmatrix} x_{n+1:n+k-1} \\ x_{n+k+1:2n} \\ x_{n+k} \end{pmatrix} \quad \text{avec} \quad \bar{\mathbf{v}}_1 = \begin{pmatrix} x_{n+1:n+k-1} \\ x_{n+k+1:2n} \end{pmatrix}, \\
\lambda = \frac{x_{2n+1}}{2}.
\end{aligned}$$

De plus, on obtient après permutation

$$\begin{aligned}
M_\gamma(\bar{p}) &= \gamma\, I_n \gamma^{-1} = I_n, \\
\alpha_\gamma(\bar{p},\bar{v}) &= \gamma \begin{pmatrix} 0 & \cdots & 0 & \mathfrak{g} \end{pmatrix}^\top = \begin{pmatrix} 0_{1:k-1}^\top & 0_{1:n-k-1}^\top & \mathfrak{g} & 0 \end{pmatrix}^\top, \\
g(\bar{p}) &= g(p) = \sum_{i=1}^{n} x_i^2 - 1, \\
D_{\bar{p}} g(\bar{p}) &= \begin{pmatrix} 2x_1 & \cdots & 2x_n \end{pmatrix} \gamma^{-1} = \begin{pmatrix} 2x_{1:k-1}^\top & 2x_{k+1:n}^\top & 2x_k \end{pmatrix}.
\end{aligned}$$

Clairement

$$x_k = \bar{\mathbf{p}}_2 = \varphi_1(\bar{\mathbf{p}}_1) = \pm\sqrt{1 - \sum_{i=1, i\neq k}^{n} x_i^2} \neq 0 \quad \text{et} \quad x_{n+k} = \bar{\mathbf{v}}_2 = \varphi_2(\bar{\mathbf{p}}_1, \bar{\mathbf{v}}_1) = -\frac{1}{x_k}\sum_{i=1, i\neq k}^{n} x_i x_{n+i}.$$

Le même procédé s'applique pour exprimer le multiplicateur de Lagrange. Après calculs, on parvient à

$$x_{2n+1} = \mathfrak{g}x_n + \sum_{i=1}^{n} x_{n+i}^2.$$

On termine par le système différentiel ordinaire :

$$
\begin{aligned}
&\begin{pmatrix} \dot{x}_{1:k-1} \\ \dot{x}_{k+1:n} \\ \dot{x}_{n+1:n+k-1} \\ \dot{x}_{n+k+1:2n-1} \\ \dot{x}_{2n} \end{pmatrix} \\
&= \begin{pmatrix} \dot{\bar{\mathbf{p}}}_1 \\ \dot{\bar{\mathbf{v}}}_1 \end{pmatrix} \\
&= \begin{pmatrix} \bar{\mathbf{v}}_1 \\ \left[M_\gamma(\bar{p})^{-1}\right]_1 \left(\alpha_\gamma(\bar{p},\bar{v}) - D_{\bar{p}} g(\bar{p})^\top \lambda\right) \end{pmatrix} \\
&= \begin{pmatrix} x_{n+1:n+k-1} \\ x_{n+k+1:2n} \\ \begin{bmatrix} I_{k-1} & 0 & 0 \\ 0 & I_{n-k} & 0 \end{bmatrix} \begin{pmatrix} -x_{2n+1}x_{1:k-1}^\top & -x_{2n+1}x_{k+1:n-1}^\top & \mathfrak{g} - x_{2n+1}x_n & -x_{2n+1}x_k \end{pmatrix}^\top \end{pmatrix} \\
&= \begin{pmatrix} x_{n+1:n+k-1} \\ x_{n+k+1:2n} \\ -x_{2n+1}x_{1:k-1} \\ -x_{2n+1}x_{k+1:n-1} \\ \mathfrak{g} - x_{2n+1}x_n \end{pmatrix}.
\end{aligned}
$$

□

2.3.4 Caractère géométrique

Pour clore ce deuxième chapitre, nous montrons le caractère géométrique de la méthode de déflation. En effet, nous allons établir que la méthode conserve la géométrie du système étudié en comparant les équations du pendule simple en dimension n fournies par cette dernière et celles provenant du formalisme d'Euler-Lagrange en coordonnées sphériques.

2.3.4.1 Coordonnées sphériques

Soient $X \in \mathbb{S}^{n-1}$ et $\theta \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[\times]0, 2\pi[^{n-2}$. On pose

$$
\begin{cases} c_k = \cos\theta_k, & k \in [\![1, n-1]\!], \\ c_n = 1, & \\ s_k = \sin\theta_k, & k \in [\![1, n-1]\!]. \end{cases}
$$

Le changement de variable $x = F(\theta)$ (que nous noterons simplement F) suivant établit une équivalence entre les coordonnées sphériques et les coordonnées cartésiennes :

$$
x_k = F_k = \left(\prod_{p=1}^{n-k} s_p\right) c_{n-k+1}, \quad k \in [\![1, n]\!].
$$

2.3.4.2 Lagrangien en coordonnées sphériques

On commence par écrire les équations d'Euler-Lagrange en coordonnées sphériques, via l'étude du lagrangien $\mathcal{L}$. En notant $\langle , \rangle$ le produit scalaire usuel, le lagrangien s'écrit

$$
\mathcal{L}\left(\theta, \dot{\theta}\right) = \frac{1}{2}\left\langle D_\theta F \dot{\theta}, D_\theta F \dot{\theta} \right\rangle - \mathfrak{g}(1 - c_1).
$$

On pose $v(\theta) = D_\theta F^\top D_\theta F$. Ainsi,

$$\mathcal{L}\left(\theta, \dot{\theta}\right) = \frac{1}{2}\dot{\theta}^\top v(\theta)\dot{\theta} - \mathfrak{g}(1 - c_1). \tag{2.41}$$

Théorème 19

L'équation d'Euler-Lagrange appliquée à l'expression (2.41) fournit les équations suivantes :

$$\begin{cases} \ddot{\theta}_1 = \dfrac{c_1}{s_1}\displaystyle\sum_{k=2}^{n-1} v_{kk}(\theta)\dot{\theta}_k^2 - \mathfrak{g}s_1 \\ \ddot{\theta}_i = \dfrac{1}{v_{ii}(\theta)}\left(\dfrac{c_i}{s_i}\displaystyle\sum_{k=1+i}^{n-1} v_{kk}(\theta)\dot{\theta}_k^2 - 2v_{ii}(\theta)\dot{\theta}_i\displaystyle\sum_{k=1}^{i-1}\frac{c_k}{s_k}\dot{\theta}_k\right), \end{cases} \tag{2.42}$$

pour tout $i \in [\![2, n-1]\!]$.

Preuve - En injectant le lagrangien défini par (2.41) dans l'équation d'Euler-Lagrange

$$\frac{d}{dt}D_{\dot{\theta}}\mathcal{L}\left(\theta, \dot{\theta}\right) - D_\theta\mathcal{L}\left(\theta, \dot{\theta}\right) = 0,$$

on obtient

$$\ddot{\theta} = v(\theta)^{-1}\left(\frac{1}{2}\dot{\theta}^\top D_\theta v(\theta)\dot{\theta} - \left(\sum_{k=1}^{n-1}\frac{\partial v(\theta)}{\partial\theta_k}\dot{\theta}_k\right)\dot{\theta} - \mathfrak{g}s_1 e\right), \tag{2.43}$$

où $e = (1, 0, \ldots, 0)^\top$. Il faut à présent procéder au calcul de (2.43). Nous opérons celui-ci en plusieurs étapes :

1. Détermination de $v(\theta)$.
2. Calcul de $\left(\displaystyle\sum_{k=1}^{n-1}\frac{\partial v(\theta)}{\partial\theta_k}\dot{\theta}_k\right)\dot{\theta}$.
3. Calcul de $\dot{\theta}^\top D_\theta v(\theta)\dot{\theta}$.

Détermination de $v(\theta)$ - Par définition, $v(\theta)$ est une matrice symétrique et on a

$$v_{ij}(\theta) = \sum_{k=1}^{n}\frac{\partial F_k}{\partial\theta_i}\frac{\partial F_k}{\partial\theta_j},$$

pour tout $(i, j) \in [\![1, n]\!]^2$. Les fonctions F_k satisfont

$$\frac{\partial F_k}{\partial\theta_i} = \begin{cases} F_k\dfrac{c_i}{s_i}, & i \in [\![1, n-k]\!] \\ -F_k\dfrac{s_i}{c_i}, & i = n-k+1 \\ 0 \quad \text{sinon.} \end{cases}$$

De plus,

$$\sum_{k=1}^{n-i}F_k^2 = \sum_{k=1}^{n-i}\left(\prod_{p=1}^{n-k}s_p^2\right)c_{n-k+1}^2 = \prod_{p=1}^{n-1}s_p^2 + \sum_{k=2}^{n-i}\left(\prod_{p=1}^{n-k}s_p^2\right)c_{n-k+1}^2 = \prod_{p=1}^{i+1}s_p^2 + \left(\prod_{p=1}^{i}s_p^2\right)c_{i+1}^2.$$

Ainsi

$$\sum_{k=1}^{n-i}F_k^2 = (c_{i+1}^2 + s_{i+1}^2)\prod_{p=1}^{i}s_p^2 = \prod_{p=1}^{i}s_p^2.$$

On commence par déterminer les termes diagonaux de $v(\theta)$. Pour tout $i \in [\![1, n-1]\!]$, on a

$$v_{ii}(\theta) = \sum_{k=1}^{n} \left(\frac{\partial F_k}{\partial \theta_i}\right)^2 = \sum_{k=1}^{n-i+1} \left(\frac{\partial F_k}{\partial \theta_i}\right)^2 = \sum_{k=1}^{n-i} \left(\frac{\partial F_k}{\partial \theta_i}\right)^2 + \left(\frac{\partial F_{n-i+1}}{\partial \theta_i}\right)^2.$$

Par conséquent

$$v_{ii}(\theta) = \frac{c_i^2}{s_i^2} \sum_{k=1}^{n-i} F_k^2 + \frac{s_i^2}{c_i^2} F_{n-i+1} = \left(\prod_{p=1}^{i-1} s_p^2\right) c_i^2 + \left(\prod_{p=1}^{i-1} s_p^2\right) s_i^2 = \left(c_i^2 + s_i^2\right) \prod_{p=1}^{i-1} s_p^2 = \prod_{p=1}^{i-1} s_p^2.$$

On détermine ensuite les termes sur-diagonaux (on récupère les autres par symétrie). Pour tout $i \neq j$, $i < j$ on a

$$v_{ij}(\theta) = \sum_{k=1}^{n} \frac{\partial F_k}{\partial \theta_i} \frac{\partial F_k}{\partial \theta_j} = \sum_{k=1}^{n-j+1} \frac{\partial F_k}{\partial \theta_i} \frac{\partial F_k}{\partial \theta_j} = \sum_{k=1}^{n-j} \frac{\partial F_k}{\partial \theta_i} \frac{\partial F_k}{\partial \theta_j} + \frac{\partial F_{n-j+1}}{\partial \theta_i} \frac{\partial F_{n-j+1}}{\partial \theta_j}.$$

Alors

$$v_{ij}(\theta) = \frac{c_i c_j}{s_i s_j} \sum_{k=1}^{n-j} F_k^2 - \frac{c_i s_j}{s_i c_j} F_{n-j+1}^2 = s_1^2 \dots s_i \dots s_j c_i c_j - s_1^2 \dots s_i \dots s_{j-1}^2 s_j c_i c_j = 0.$$

En résumé, on obtient

$$v(\theta) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & l^2 s_1^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l^2 s_1^2 \dots s_{N-2}^2 \end{bmatrix}.$$

Calcul de $\left(\sum_{k=1}^{n-1} \frac{\partial v(\theta)}{\partial \theta_k} \dot{\theta}_k\right) \dot{\theta}$ - On note cette expression α. Comme $\frac{\partial v_{ii}(\theta)}{\partial \theta_k} = 0$ pour tout $k \geqslant i$, on a :

$$\alpha_i = \left[\left(\sum_{k=1}^{n-1} \frac{\partial v(\theta)}{\partial \theta_k} \dot{\theta}_k\right) \dot{\theta}\right]_i = \sum_{k=1}^{n-1} \frac{\partial v_{ii}(\theta)}{\partial \theta_k} \dot{\theta}_k \dot{\theta}_i = \sum_{k=1}^{i-1} \frac{\partial v_{ii}(\theta)}{\partial \theta_k} \dot{\theta}_k \dot{\theta}_i = 2 v_{ii}(\theta) \dot{\theta}_i \sum_{k=1}^{i-1} \frac{c_k}{s_k} \dot{\theta}_k.$$

Calcul de $\dot{\theta}^\top D_\theta v(\theta) \dot{\theta}$ - On note cette expression β. Pour tout $i \in [\![1, n]\!]$, on obtient

$$\beta_i = \dot{\theta}^\top D_\theta v(\theta) \dot{\theta} = \sum_{k=1}^{n-1} \frac{\partial v_{kk}(\theta)}{\partial \theta_i} \dot{\theta}_k^2 = \sum_{k=1+i}^{n-1} \frac{\partial v_{kk}(\theta)}{\partial \theta_i} \dot{\theta}_k^2 = 2 \frac{c_i}{s_i} \sum_{k=1+i}^{n-1} v_{kk} \dot{\theta}_k^2.$$

En rassemblant les trois informations récoltées et en les remplaçant dans (2.43), on parvient au résultat attendu

$$\begin{cases} \ddot{\theta}_1 = \dfrac{c_1}{s_1} \displaystyle\sum_{k=2}^{n-1} v_{kk}(\theta) \dot{\theta}_k^2 - \mathfrak{g} s_1 \\ \ddot{\theta}_i = \dfrac{1}{v_{ii}(\theta)} \left(\dfrac{c_i}{s_i} \displaystyle\sum_{k=1+i}^{n-1} v_{kk}(\theta) \dot{\theta}_k^2 - 2 v_{ii}(\theta) \dot{\theta}_i \sum_{k=1}^{i-1} \dfrac{c_k}{s_k} \dot{\theta}_k \right), \end{cases}$$

pour tout $i \in [\![2, n-1]\!]$. □

2.3.4.3 Équations fournies par la méthode de déflation

À présent, on utilise les équations différentielles du pendule simple en dimension n (2.39) exhibées par l'algorithme. En effectuant le changement de variable sphérique, on obtient d'une part

$$\begin{aligned}
\ddot{x} &= \begin{pmatrix} \ddot{x}_1 \\ \vdots \\ \ddot{x}_{n-1} \\ \ddot{x}_n \end{pmatrix} \\
&= \begin{pmatrix} -x_{2n+1}x_1 \\ \vdots \\ -x_{2n+1}x_{n-1} \\ \dot{\theta}^\top D^2_\theta F_n \dot{\theta} + D_\theta F_n \ddot{\theta} \end{pmatrix} \\
&= \begin{pmatrix} -x_{2n+1}F_1 \\ \vdots \\ -x_{2n+1}F_{n-1} \\ -c_1\dot{\theta}_1^2 \end{pmatrix} - \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \\ s_1 & 0 & \cdots & 0 \end{bmatrix} \ddot{\theta} \\
&= G(\theta) - \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \\ s_1 & 0 & \cdots & 0 \end{bmatrix} \ddot{\theta}.
\end{aligned}$$

D'autre part :

$$\ddot{x} = \dot{\theta}^\top D^2_\theta F \dot{\theta} + D_\theta F \ddot{\theta}.$$

On compare les deux expressions de $\ddot{x}$ et on a :

$$\left(v(\theta) + D^t F \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \\ s_1 & 0 & \cdots & 0 \end{bmatrix} \right) \ddot{\theta} = D_\theta^\top F \left(G(\theta) - \dot{\theta}^\top D^2_\theta F \dot{\theta} \right).$$

Ainsi,

$$\ddot{\theta} = \begin{bmatrix} c_1^2 & 0 & \cdots & 0 \\ 0 & v_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_{n-1n-1} \end{bmatrix}^{-1} D_\theta^\top F \left(G(\theta) - \dot{\theta}^\top D^2_\theta F \dot{\theta} \right). \tag{2.44}$$

Théorème 20

L'expression (2.44) est égale à

$$\begin{cases} \ddot{\theta}_1 = \dfrac{c_1}{s_1} \displaystyle\sum_{k=2}^{n-1} v_{kk}(\theta)\dot{\theta}_k^2 - \mathfrak{g} s_1 \\ \ddot{\theta}_i = \dfrac{1}{v_{ii}(\theta)} \left(\dfrac{c_i}{s_i} \displaystyle\sum_{k=1+i}^{n-1} v_{kk}(\theta)\dot{\theta}_k^2 - 2v_{ii}(\theta)\dot{\theta}_i \displaystyle\sum_{k=1}^{i-1} \dfrac{c_k}{s_k}\dot{\theta}_k \right), \end{cases} \tag{2.45}$$

pour tout $i \in [\![2, n-1]\!]$.

Preuve - On part de l'expression (2.44)

$$\ddot{\theta} = \begin{bmatrix} c_1^2 & 0 & \cdots & 0 \\ 0 & v_{22}(\theta) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_{n-1n-1}(\theta) \end{bmatrix}^{-1} \left(D_\theta^\top FG(\theta) - D_\theta^\top F\dot{\theta}^\top D_\theta^2 F\dot{\theta} \right).$$

Il suffit de développer l'expression précédente en procédant comme suit :

1. Calcul de $D_\theta^\top FG(\theta)$.
2. Calcul de $D_\theta^\top F\dot{\theta}^\top D_\theta^2 F\dot{\theta}$.

Calcul de $D_\theta^\top FG(\theta)$ - On note cette expression γ. On introduit le symbole de Kronecker δ_1^i tel que $\delta_1^i = 1$ si $i = 1$ et $\delta_1^i = 0$ si $i \neq 1$. Pour tout $i \in [\![1, n-1]\!]$, on obtient

$$\begin{aligned}
\gamma_i &= \left[D_\theta^\top FG(\theta) \right]_i \\
&= \sum_{k=1}^{n-1} \frac{\partial F_k}{\partial \theta_i} \left(-x_{2n+1} F_k \right) - c_1 \frac{\partial F_n}{\partial \theta_i} \dot{\theta}_1^2 \\
&= \left(x_{2n+1} \frac{\partial F_n}{\partial \theta_1} F_n - c_1 \frac{\partial F_n}{\partial \theta_1} \dot{\theta}_1^2 \right) \delta_1^i \\
&= \left[\frac{\partial F_n}{\partial \theta_1} F_n \left(x_{2n+1} - \dot{\theta}_1^2 \right) \right] \delta_1^i \\
&= -c_1 s_1 \left[\left(\|\dot{x}\|^2 + \mathfrak{g} x_n \right) - \dot{\theta}_1^2 \right] \delta_1^i \\
&= -c_1 s_1 \left(\sum_{k=1}^{n-1} v_{kk}(\theta) \dot{\theta}_k^2 + \mathfrak{g} c_1 - \dot{\theta}_1^2 \right) \delta_1^i \\
&= -c_1 s_1 \left(\sum_{k=2}^{n-1} v_{kk}(\theta) \dot{\theta}_k^2 + \mathfrak{g} c_1 \right) \delta_1^i.
\end{aligned}$$

Calcul de $D_\theta^\top F\dot{\theta}^\top D_\theta^2 F\dot{\theta}$ - On note cette expression η. Pour tout $i \in [\![1, n-1]\!]$, on a

$$\begin{aligned}
\eta_i &= \left[D_\theta^\top F\dot{\theta}^t D_\theta^2 F\dot{\theta} \right]_i \\
&= \sum_{k=1}^{n-i+1} \frac{\partial F_k}{\partial \theta_i} \dot{\theta}^\top D_\theta^2 F_k \dot{\theta} \\
&= \sum_{k=1}^{n-i+1} \frac{\partial F_k}{\partial \theta_i} \left(\sum_{m,p} \frac{\partial^2 F_k}{\partial \theta_m \partial \theta_p} \dot{\theta}_m \dot{\theta}_p \right) \\
&= \sum_{k=1}^{n-i+1} \frac{\partial F_k}{\partial \theta_i} \left(\sum_{m=1}^{n-1} \frac{\partial^2 F_k}{\partial \theta_m^2} \dot{\theta}_m^2 + 2 \sum_{p=1}^{n-2} \sum_{m=1}^{n-p-1} \frac{\partial^2 F_k}{\partial \theta_m \partial \theta_{n-p}} \dot{\theta}_m \dot{\theta}_{n-p} \right) \\
&= \sum_{m=1}^{n-1} \sum_{k=1}^{n-i+1} \frac{\partial F_k}{\partial \theta_i} \frac{\partial^2 F_k}{\partial \theta_m^2} \dot{\theta}_m^2 + 2 \sum_{p=1}^{n-2} \sum_{m=1}^{n-p-1} \sum_{k=1}^{n-i+1} \frac{\partial F_k}{\partial \theta_i} \frac{\partial^2 F_k}{\partial \theta_m \partial \theta_{n-p}} \dot{\theta}_m \dot{\theta}_{n-p} \\
&= \sum_{m=1}^{n-1} \sum_{k=1}^{\min(n-i+1, n-m+1)} \frac{\partial F_k}{\partial \theta_i} \frac{\partial^2 F_k}{\partial \theta_m^2} \dot{\theta}_m^2 \\
&\quad + 2 \sum_{p=1}^{n-2} \sum_{m=1}^{n-p-1} \sum_{k=1}^{\min(p+1, n-i+1)} \frac{\partial F_k}{\partial \theta_i} \frac{\partial^2 F_k}{\partial \theta_m \partial \theta_{n-p}} \dot{\theta}_m \dot{\theta}_{n-p}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{m=1}^{i} \sum_{k=1}^{n-i+1} \frac{\partial F_k}{\partial \theta_i} \frac{\partial^2 F_k}{\partial \theta_m^2} \dot{\theta}_m^2 + \sum_{m=i+1}^{n-1} \sum_{k=1}^{n-m+1} \frac{\partial F_k}{\partial \theta_i} \frac{\partial^2 F_k}{\partial \theta_m^2} \dot{\theta}_m^2 \\
&\quad + 2 \sum_{m=1}^{i-1} \left(\sum_{k=1}^{n-i+1} \frac{\partial F_k}{\partial \theta_i} \frac{\partial^2 F_k}{\partial \theta_m \partial \theta_i} \right) \dot{\theta}_m \dot{\theta}_i + 2 \sum_{p=1}^{n-i-1} \sum_{m=1}^{n-p-1} \sum_{k=1}^{p+1} \frac{\partial F_k}{\partial \theta_i} \frac{\partial^2 F_k}{\partial \theta_m \partial \theta_{n-p}} \dot{\theta}_m \dot{\theta}_{n-p} \\
&\quad + 2 \sum_{p=n-i+1}^{n-2} \sum_{m=1}^{n-p-1} \sum_{k=1}^{n-i+1} \frac{\partial F_k}{\partial \theta_i} \frac{\partial^2 F_k}{\partial \theta_m \partial \theta_{n-p}} \dot{\theta}_m \dot{\theta}_{n-p} \\
&= \zeta_1 + \zeta_2 + \zeta_3 + \zeta_4 + \zeta_5.
\end{aligned}
$$

On montre aisément que :

$$\sum_{k=1}^{n-i+1} \frac{\partial F_k}{\partial \theta_i} F_k = 0.$$

Détaillons à présent les calculs des ζ_l, $l \in [\![1, 5]\!]$ séparément :

1. $$\zeta_1 = - \sum_{m=1}^{i} \sum_{k=1}^{n-i+1} \frac{\partial F_k}{\partial \theta_i} F_k \dot{\theta}_m^2 = - \sum_{m=1}^{i} \dot{\theta}_m^2 \left(\sum_{k=1}^{n-i+1} \frac{\partial F_k}{\partial \theta_i} F_k \right) = 0,$$

   car $m \leqslant n - k + 1$.

2. $$\zeta_2 = \sum_{m=i+1}^{n-1} \sum_{k=1}^{n-m+1} \frac{c_i}{s_i} F_k \frac{\partial^2 F_k}{\partial \theta_m^2} \dot{\theta}_m^2 = \frac{c_i}{s_i} \sum_{m=i+1}^{n-1} \left(\sum_{k=1}^{n-m+1} F_k \frac{\partial^2 F_k}{\partial \theta_m^2} \right) \dot{\theta}_m^2 = - \frac{c_i}{s_i} \sum_{m=i+1}^{n-1} v_{mm}(\theta) \dot{\theta}_m^2,$$

   car $i \leqslant n - k + 1$.

3. $$\zeta_3 = 2 \dot{\theta}_i \sum_{m=1}^{i-1} \frac{1}{2} \frac{\partial v_{ii}(\theta)}{\partial \theta_m} \dot{\theta}_m = 2 v_{ii}(\theta) \dot{\theta}_i \sum_{m=1}^{i-1} \frac{c_m}{s_m} \dot{\theta}_m,$$

   car $m \leqslant i - 1$.

4. $$\begin{aligned}
   \zeta_4 &= 2 \sum_{p=1}^{n-i-1} \sum_{m=1}^{n-p-1} \frac{c_m}{s_m} \left(\sum_{k=1}^{p+1} \frac{\partial F_k}{\partial \theta_i} \frac{\partial F_k}{\partial \theta_{n-p}} \right) \dot{\theta}_m \dot{\theta}_{n-p} \\
   &= 2 \sum_{p=1}^{n-i-1} \sum_{m=1}^{n-p-1} \frac{c_m}{s_m} \left(\sum_{k=1}^{n-(n-p)+1} \frac{\partial F_k}{\partial \theta_i} \frac{\partial F_k}{\partial \theta_{n-p}} \right) \dot{\theta}_m \dot{\theta}_{n-p} \\
   &= 0,
   \end{aligned}$$

   car $n - p > i$.

5. $$\zeta_5 = 2 \sum_{p=n-i+1}^{n-2} \sum_{m=1}^{n-p-1} \frac{c_m}{s_m} \frac{c_{n-p}}{s_{n-p}} \left(\sum_{k=1}^{n-i+1} \frac{\partial F_k}{\partial \theta_i} F_k \right) \dot{\theta}_m \dot{\theta}_{n-p} = 0,$$

   car $m < n - k + 1$ et $n - p < n - k + 1$.

En résumé, on a pour tout $i \in [\![1, n-1]\!]$

$$\eta_i = -\frac{c_i}{s_i} \sum_{m=i+1}^{n-1} v_{mm}(\theta)\dot{\theta}_m^2 + 2v_{ii}(\theta)\dot{\theta}_i \sum_{m=1}^{i-1} \frac{c_m}{s_m}\dot{\theta}_m.$$

En conclusion, on arrive à

$$\begin{aligned}
\ddot{\theta}_1 &= \frac{1}{c_1^2}(\gamma_1 - \eta_1) \\
&= \frac{1}{c_1^2}\left[-c_1 s_1 \left(\sum_{k=2}^{n-1} v_{kk}(\theta)\dot{\theta}_k^2 + \mathfrak{g}c_1\right) + \frac{c_1}{s_1}\sum_{m=2}^{n-1} v_{mm}(\theta)\dot{\theta}_m^2\right] \\
&= \frac{1}{c_1}\left[-s_1 \left(\sum_{k=2}^{n-1} v_{kk}(\theta)\dot{\theta}_k^2 + \mathfrak{g}c_1\right) + \frac{1}{s_1}\sum_{k=2}^{n-1} v_{kk}(\theta)\dot{\theta}_k^2\right] \\
&= \left(\frac{1}{c_1 s_1} - \frac{s_1}{c_1}\right)\sum_{k=2}^{n-1} v_{kk}(\theta)\dot{\theta}_k^2 - \mathfrak{g}s_1 \\
&= \frac{c_1^2}{c_1 s_1}\sum_{k=2}^{n-1} v_{kk}(\theta)\dot{\theta}_k^2 - \mathfrak{g}s_1 \\
&= \frac{c_1}{s_1}\sum_{k=2}^{n-1} v_{kk}(\theta)\dot{\theta}_k^2 - \mathfrak{g}s_1
\end{aligned}$$

et

$$\ddot{\theta}_i = \frac{1}{v_{ii}(\theta)}\left(\frac{c_i}{s_i}\sum_{m=i+1}^{n-1} v_{mm}(\theta)\dot{\theta}_m^2 - 2v_{ii}(\theta)\dot{\theta}_i \sum_{m=1}^{i-1}\frac{c_m}{s_m}\dot{\theta}_m\right),$$

pour tout $i \in [\![2, n-1]\!]$. □

Grâce aux théorèmes 19 et 20 fournissant les mêmes expressions, on vient de prouver que la méthode de déflation conserve la géométrie des systèmes étudiés.

Dans ce chapitre, nous avons introduit une nouvelle méthode de résolution des EDAs linéaires et quasi-linéaires. De par sa structure, la méthode de déflation peut ainsi s'appliquer à une grande variété d'EDAs. L'architecture de la méthode est décrite par des algorithmes dont diverses propriétés ont été étudiées.

Dans sa version linéaire, la méthode de déflation est proche de l'algorithme introduit par W. A. Harris, Y. Sibuya et L. Weinberg [23]. À la différence de cet algorithme, la méthode de déflation se concentre sur les problèmes réguliers. La méthode de déflation partage certaines propriétés avec les méthodes des systèmes augmentés et des projections comme la baisse caractéristique de l'indice (diminution de 1 à chaque étape quand les coefficients sont constants) ou encore l'invariance du rang et de l'indice de Kronecker. Elle conserve une structure identique pour les EDAs linéaires et quasi-linéaires comme la méthode des systèmes augmentés, mais en revanche diminue la taille des systèmes au lieu de les augmenter.

Dans sa version quasi-linéaire, la méthode de déflation peut faire écho à une traduction algorithmique de la méthode globale de réduction géométrique de P. J. Rabier et W. C. Rheinboldt [48], bien que cette dernière ne contienne pas une phase de substitution des variables. Enfin, la proximité entre la méthode de déflation et la démarche d'élimination différentielle

mise en œuvre dans l'algorithme Rosenfeld-Gröbner doit être soulignée. L'algorithme Rosenfeld-Gröbner est capable de traiter des EDAs générales (à condition qu'elles soient polynomiales) mais également des équations aux dérivées partielles. Son approche n'est pas matricielle : il analyse une liste d'équations. La méthode de déflation conserve et manipule la structure matricielle des problèmes étudiés. De plus, elle ne nécessite pas de fixer un ordre sur les variables, comme l'exige l'algorithme Rosenfeld-Gröbner.

On applique dans le troisième chapitre la méthode de déflation à des EDAs quasi-linéaires modélisant des phénomènes de distillation.

Chapitre 3

Application à la résolution de modèles de distillation de Rayleigh

Dans ce chapitre, nous étudions grâce à la méthode de déflation trois systèmes quasi-linéaires extraits des travaux de K. Alloula [1] et de R. Théry Hétreux [57] modélisant la *distillation de Rayleigh*. La *distillation de Rayleigh réactive*, qui fait intervenir des réactions chimiques, est aussi étudiée. Après avoir présenté les principes physiques, nous introduisons les systèmes différentiels algébriques et nous analysons ces derniers.

Les EDAs rencontrées dans le domaine du génie des procédés sont généralement traitées par des méthodes numériques. Si ces EDAs conduisent à la plupart des difficultés rencontrées lors de la résolution numérique dans d'autres domaines (tels que la mécanique), elles ont néanmoins quelques particularités :

- la modélisation d'installations industrielles composées de plusieurs opérations unitaires reliées en un réseau complexe (au sein desquelles ont lieu divers phénomènes physico-chimiques entre plusieurs constituants) conduit toujours à des EDAs de grande taille ;
- des événements de temps et d'état peuvent survenir en cours de l'intégration numérique et notamment conduire à un nouveau calcul de conditions initiales cohérentes ;
- la diversité de nature des équations composant le système rend difficile une formulation évitant à priori l'écriture d'EDAs d'indice élevé.

Ces particularités rendent les difficultés d'intégration numérique des EDAs encore plus aigües dans le domaine du génie des procédés. La recherche dans cette discipline s'est entre autres concentrée sur l'analyse des structures des EDAs afin de mieux appréhender ces difficultés. Comme le souligne P. Rouchon [53], les méthodes de résolution numériques font généralement défaut pour les EDAs d'indice élevé. Dans ces travaux, il propose une méthode de résolution formelle qui réduit les EDAs d'indice élevé en EDAs d'indice un. Ces EDAs réduites sont ensuite traitées d'un point de vue numérique. Pour cette étape d'intégration numérique, au-delà des codes déjà utilisés pour l'intégration des EDOs, des intégrateurs numériques basés sur les méthodes à pas multiples sont utilisés (se reporter notamment aux travaux de L. Petzold, [42]) afin de disposer d'outils efficaces (au niveau de la stabilité et de la convergence). D'autres processus similaires de résolution des EDAs ont été mis en œuvre ; on peut consulter les travaux de C. C. Pantelides [41] et J. Unger [58].

La méthode de déflation propose une réduction d'indice complète puisque les EDAs réduites sont d'indice zéro, c'est-à-dire des EDOs, pour lesquelles il est possible d'utiliser des méthodes numériques classiques très performantes, et de contourner les difficultés d'intégration des EDAs. Nous présentons dans ce chapitre le pré-traitement des équations, soit la phase formelle de la résolution. Rappelons que notre analyse conduit à la production de modèles équivalents aux

modèles de départ, pour lesquels l'intégration numérique sera améliorée, notamment du point de vue du calcul des conditions initiales cohérentes.

3.1 Présentation

La distillation est un procédé chimique qui consiste à séparer les différents constituants d'un mélange liquide par un processus de vaporisation puis de condensation. Cette technique de séparation exploite le fait que les espèces chimiques qui constituent le mélange ont des températures d'ébullition différentes. Le mélange est chauffé dans un bouilleur au dessus duquel est placée une colonne à distiller. Les constituants se vaporisent au fur et à mesure, le bouilleur s'appauvrissant d'abord en constituants les plus volatils. Ces différents gaz vont ensuite être condensés à des hauteurs différentes dans la colonne. On parvient ainsi à séparer les constituants avec une certaine pureté. Les produits de cette distillation sont nommés distillats.

La *distillation de Rayleigh* est une distillation sans reflux (les gaz condensés ne retombent pas dans le bouilleur) et discontinue (le bouilleur n'est pas alimenté pendant le processus). Elle est qualifiée de *réactive* quand les constituants du mélange réagissent entre eux dans la phase liquide (et éventuellement dans la phase vapeur). La distillation de Rayleigh est notamment utilisée dans le domaine de la chimie de spécialité, par exemple dans la chimie pharmaceutique, afin de synthétiser des produits à haute valeur ajoutée, pour un faible tonnage.

Nous considérons deux types de réactions chimiques :
- les réactions chimiques dites *instantanément équilibrées* : à tout instant t, les réactions chimiques sont à l'équilibre. Cette hypothèse est retenue quand l'échelle de temps des réactions chimiques est très petite devant celle des autres phénomènes ;
- les réactions chimiques dites *contrôlées par la cinétique* assurent une transition progressive vers un état d'équilibre chimique. Les vitesses des réactions chimiques contrôlées par la cinétique dépendent notamment de la température. Le cas précédent peut être envisagé comme un cas limite pour lequel les vitesses de réactions sont très rapides.

On peut également considérer une distillation de Rayleigh réactive où certaines réactions chimiques sont équilibrées instantanément et d'autres sont contrôlées par la cinétique. Nous ne prenons pas en compte cette configuration dans nos travaux.

Dans notre étude, nous distinguons deux étapes du procédé de distillation de Rayleigh :
- l'étape *monophasique* qui correspond à la période de chauffe avant l'ébullition. Bien que nous considérions une mince pellicule de phase vapeur en équilibre avec la phase liquide, c'est principalement cette dernière qui est mise en avant pendant le régime monophasique. Durant cette étape, le débit vapeur reste nul ;
- l'étape *diphasique* qui débute au moment de l'ébullition. Il existe alors un flux de vapeur alimenté par la phase liquide. Les phases liquide et vapeur co-existent et sont supposées être à l'équilibre thermodynamique.

Remarque 16

> *Une troisième étape pourrait être considérée ; au moment où la phase liquide disparaît, le système devient uniquement une phase vapeur. Cette étape monophasique vapeur n'est pas abordée dans cette étude.*

Enfin, nous considérons pour la phase liquide deux classes de comportements thermodynamiques :
- le comportement *idéal*, dans lequel les interactions entre les différentes espèces du mélange sont négligées. Ce dernier se comporte alors comme un corps pur ;

– le comportement *non idéal*, dans lequel des interactions existent entre les différentes espèces selon leur nature.

Dans cette étude, la phase vapeur est toujours considérée comme un gaz parfait, c'est-à-dire une phase idéale.

3.2 Notations

Les systèmes modélisant le phénomène de distillation de Rayleigh étudiés dans ce chapitre sont des EDAs quasi-linéaires, c'est-à-dire de la forme

$$E_0(X_0)\dot{X}_0 = f_0(X_0). \tag{3.1}$$

Dans la partie 3.2.1, nous donnons la liste des notations intervenant lors de la première étape de l'algorithme de déflation. Ces notations se généralisent pour les étapes de déflation suivantes. Ainsi, l'étape k produit elle le système $E_k(X_k)\dot{X}_k = f_k(X_k)$, à partir du système $E_{k-1}(X_{k-1})\dot{X}_{k-1} = f_{k-1}(X_{k-1})$. Dans la partie 3.2.2, nous fournissons une nomenclature métier des symboles utilisés dans les différents modèles. La partie 3.2.3 liste des sous-expressions communes présentes dans les modèles obtenus par déflation, et leur associe des notations.

3.2.1 Schéma de déflation

On commence naturellement par écrire les vecteurs X_0 et $f_0(X_0)$ ainsi que la matrice $E_0(X_0)$:

$$\boxed{X_0, E_0(X_0) \text{ et } f_0(X_0)}$$

On exhibe la sous-matrice de rang plein $\tilde{E}_0(X_0)$ de $E_0(X_0)$ telle que

$$E_0(X_0) = \begin{bmatrix} \tilde{E}_0(X_0) & 0 \\ 0 & 0 \end{bmatrix},$$

ainsi que le sous-vecteur $\tilde{f}_0(X_0)$ de $f_0(X_0)$ correspondant aux contraintes algébriques explicites *i.e.* $\tilde{f}_0(X_0) = 0$:

$$\boxed{\tilde{E}_0(X_0) \text{ et } \tilde{f}_0(X_0)}$$

Remarque 17

> *La matrice $\tilde{E}_0(X_0)$ n'est pas nécessairement carrée. De plus, exhiber cette matrice peut nécessiter au préalable la transformation de la matrice $E_0(X_0)$ via des décompositions usuelles (décomposition LU par exemple).*

On fixe également le vecteur $g_0(X_0)$ tel que $f_0(X_0) = \begin{pmatrix} g_0(X_0) \\ \tilde{f}_0(X_0) \end{pmatrix}$.

$$\boxed{g_0(X_0)}$$

On détermine ensuite, si nécessaire, la matrice de permutation $\mathcal{P}_1$ des variables ; cette matrice définit deux nouveaux vecteurs X_1 et Y_1 tels que $\mathcal{P}_1 X_0 = \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}$.

$$\boxed{\mathcal{P}_1, X_1 \text{ et } Y_1}$$

On note $\bar{g}_0(X_1, Y_1) = g_0(\mathcal{P}_1^{-1}\mathcal{P}_1 X_0)$ et $\bar{f}_0(X_1, Y_1) = \tilde{f}_0(\mathcal{P}_1^{-1}\mathcal{P}_1 X_0)$. Le théorème des fonctions implicites appliqué aux contraintes algébriques $\bar{f}_0(X_1, Y_1) = 0$ nous permet d'écrire

$Y_1 = \varphi_1(X_1)$. Par dérivation, $\dot{Y}_1 = \dot{\varphi}_1(X_1)\,\dot{X}_1 = -J_{Y_1}\left(\bar{f}_0\right)^{-1} J_{X_1}\left(\bar{f}_0\right)\dot{X}_1$. Le choix de la matrice de permutation est fait pour assurer l'inversibilité de la matrice Jacobienne $J_{Y_1}\left(\bar{f}_0\right)$. Les deux matrices Jacobiennes ont besoin d'être déterminées :

$$\boxed{J_{Y_1}\left(\bar{f}_0\right) \text{ et } J_{X_1}\left(\bar{f}_0\right)}$$

En appliquant la matrice de permutation $\mathcal{P}_1$ au système de départ, on définit deux nouvelles matrices $\begin{bmatrix} A_0(X_1,Y_1) & B_0(X_1,Y_1) \end{bmatrix} = \begin{bmatrix} \tilde{E}_0(\mathcal{P}_1^{-1}\mathcal{P}_1\,X_0) & 0 \end{bmatrix}\mathcal{P}_1^{-1}$.

$$\boxed{A_0(X_1,Y_1) \text{ et } B_0(X_1,Y_1)}$$

Remarque 18

La matrice $A_0(X_1,Y_1)$ est nécessairement carrée.

Remarque 19

On exhibera souvent des structures par blocs des quatre matrices précédentes pour simplifier les calculs.

L'étape de déflation s'achève en calculant la nouvelle matrice coefficient $E_1(X_1)$:

$$E_1(X_1) = A_0(X_1,\varphi_1(X_1)) - B_0(X_1,\varphi_1(X_1))\,J_{Y_1}\left(\bar{f}_0\right)^{-1} J_{X_0}\left(\bar{f}_0\right),$$

ainsi que le nouveau membre de droite $f_1(X_1) = \bar{g}_0(X_1,\varphi(X_1))$:

$$\boxed{X_1, E_1(X_1) \text{ et } f_1(X_1)}$$

3.2.2 Nomenclature métier

Étant donné un mélange composé de N constituants, on définit les notations suivantes :
- U_l : Rétention liquide en *Mole* (quantité totale de moles présentes dans le liquide).
- V : Débit vapeur en *Mole par seconde* (variation de la rétention liquide au cours du temps).
- x_i (pour tout $i \in [\![1,N]\!]$) : Fraction molaire du constituant i dans la phase liquide. On note $x = (x_1,\ldots,x_N)^\top$.
- y_i (pour tout $i \in [\![1,N]\!]$) : Fraction molaire du constituant i dans la phase vapeur. On note $y = (y_1,\ldots,y_N)^\top$.
- T : Température supposée homogène du système en *Kelvin*.
- P : Pression imposée au système en *Bar*. En pratique, nous considérons dans nos travaux cette pression égale à la pression atmosphérique.
- $h(T,P,x)$: Enthalpie molaire de la phase liquide en *Joule par mole*.
- $H(T,P,y)$: Enthalpie molaire de la phase vapeur en *Joule par mole*.
- Q : Puissance de chauffe en *Watt*. La variation de Q en fonction du temps définit la politique de chauffe.
- $K_i(T,P)$ ou $K_i(T,P,x)$ (pour tout $i \in [\![1,N]\!]$) : Constante d'équilibre thermodynamique. On note $K = (K_1,\ldots,K_N)^\top$.
- y_{inerte} : Fraction molaire de l'air.

Soit R le nombre de réactions chimiques et $\tilde{R}$ le nombre de réactifs :
- ξ_j (pour tout $j \in [\![1,R]\!]$) : Avancement de la réaction chimique j en *Mole par seconde*.
- $\nu_{i,j}$ (pour tout $i \in [\![1,N]\!]$ et $j \in [\![1,R]\!]$) : Coefficient stœchiométrique du constituant i dans la réaction j.
- $Q_{r_j}(T)$ (pour tout $j \in [\![1,R]\!]$) : Chaleur de la réaction j en *Joule par mole*.

- $a_i(T, P, x)$ (pour tout $i \in [\![1, N]\!]$) : Activité du constituant i en phase liquide.
- $K_{c_j}(T)$ (pour tout $j \in [\![1, R]\!]$) : Constante cinétique de la réaction j.
- $k_j(T)$ (pour tout $j \in [\![1, R]\!]$) : Constante de vitesse de la réaction j.

Remarque 20

- $\nu_{i,j} = 0$ *lorsque le constituant* i *n'apparaît pas dans la réaction* j.
- $\nu_{i,j} > 0$ *si le constituant* i *est un produit de la réaction* j.
- $\nu_{i,j} < 0$ *si le constituant* i *est un réactif de la réaction* j.

À l'exception des constantes d'équilibres thermodynamiques ($K_i(T, P)$, $K_i(T, P, x)$, $K(T, P)$ ou $K(T, P, x)$), nous omettrons dans la suite les variables dont dépendent les expressions précédemment définies.

3.2.3 Nomenclature des sous-expressions communes

Lors de l'application de la méthode de déflation, des sous-expressions génériques par rapport aux constituants et/ou aux réactions apparaissent. Ces expressions ont des formes similaires, quelles que soient les hypothèses retenues (régime monophasique ou diphasique, phase liquide idéale ou non, réactions chimiques instantanément équilibrées ou contrôlées par la cinétique). Il est donc intéressant d'en définir une nomenclature, à la fois pour alléger les expressions manipulées et pour mieux analyser les similitudes et les différences entre les modèles obtenus par application de la méthode de déflation. Il faut noter que les notations du type $\bar{\xi}$ correspondent à une hypothèse de non idéalité de la phase liquide, par opposition aux notations du type ξ adoptées lorsque l'on suppose la phase liquide idéale.

Distillation de Rayleigh non réactive

$$\mathcal{D}_i = 1 - K_i(T, P), \quad \forall i \in [\![1, N]\!] \tag{3.2}$$

$$\bar{\mathcal{D}}_i = 1 - K_i(T, P, x) - \nabla_{x_i} K(T, P, x) \cdot x, \quad \forall i \in [\![1, N]\!], \quad \text{où} \quad \nabla_a f = \begin{pmatrix} \dfrac{\partial f_1}{\partial a} & \cdots & \dfrac{\partial f_N}{\partial a} \end{pmatrix} \tag{3.3}$$

$$\mathcal{K}_i = -\mathcal{D}_N^{-1} \mathcal{D}_i, \quad \forall i \in [\![1, N]\!] \tag{3.4}$$

$$\bar{\mathcal{K}}_i = -\bar{\mathcal{D}}_N^{-1} \bar{\mathcal{D}}_i, \quad \forall i \in [\![1, N]\!] \tag{3.5}$$

$$\mathcal{H}_i = \frac{\partial h}{\partial x_i} + \frac{\partial h}{\partial x_N} \mathcal{K}_i, \quad \forall i \in [\![1, N]\!] \tag{3.6}$$

$$\bar{\mathcal{H}}_i = \frac{\partial h}{\partial x_i} + \frac{\partial h}{\partial x_N} \bar{\mathcal{K}}_i, \quad \forall i \in [\![1, N]\!] \tag{3.7}$$

$$\mathcal{L} = \frac{\partial h}{\partial T} + \mathcal{D}_N^{-1} \frac{\partial h}{\partial x_N} \left(\nabla_T K(T, P) \cdot x \right) \tag{3.8}$$

$$\bar{\mathcal{L}} = \frac{\partial h}{\partial T} + \bar{\mathcal{D}}_N^{-1} \frac{\partial h}{\partial x_N} \left(\nabla_T K(T, P, x) \cdot x \right) \tag{3.9}$$

$$\mathcal{G} = H - h + \sum_{i=1}^{N} \left(\mathcal{H}_i + \frac{\mathcal{L}\mathcal{D}_i}{\nabla_T K(T,P)\cdot x} \right) (1 - K_i(T,P)) x_i \tag{3.10}$$

$$\bar{\mathcal{G}} = H - h + \sum_{i=1}^{N} \left(\bar{\mathcal{H}}_i + \frac{\bar{\mathcal{L}}\bar{\mathcal{D}}_i}{\nabla_T K(T,P,x)\cdot x} \right) (1 - K_i(T,P,x)) x_i \tag{3.11}$$

Réactions chimiques contrôlées par la cinétique

$$\Delta_j = k_j \left(\prod_{i=1}^{\tilde{R}} a_i^{|\nu_{i,j}|} - \frac{1}{K_{c_j}} \prod_{i=\tilde{R}+1}^{N} a_i^{|\nu_{i,j}|} \right) U_l, \quad \forall j \in [\![1, R]\!] \tag{3.12}$$

$$\mathcal{M}_i = \sum_{j=1}^{R} \nu_{i,j} \Delta_j - \left(\sum_{k=1}^{N} \sum_{j=1}^{R} \nu_{k,j} \Delta_j \right) x_i, \quad \forall i \in [\![1, N]\!] \tag{3.13}$$

$$\mathcal{Q} = Q - \sum_{j=1}^{R} Q_{r_j} \Delta_j - \left(\sum_{i=1}^{N} \sum_{j=1}^{R} \nu_{i,j} \Delta_j \right) h - \sum_{i=1}^{N} \left(\frac{\mathcal{L}\mathcal{D}_i}{\nabla_T K(T,P)\cdot x} + \mathcal{H}_i \right) \mathcal{M}_i \tag{3.14}$$

$$\bar{\mathcal{Q}} = Q - \sum_{j=1}^{R} Q_{r_j} \Delta_j - \left(\sum_{i=1}^{N} \sum_{j=1}^{R} \nu_{i,j} \Delta_j \right) h - \sum_{i=1}^{N} \left(\frac{\bar{\mathcal{L}}\bar{\mathcal{D}}_i}{\nabla_T K(T,P,x)\cdot x} + \bar{\mathcal{H}}_i \right) \mathcal{M}_i \tag{3.15}$$

Réactions chimiques instantanément équilibrées

$$\mathcal{A}_j = \prod_{i=1}^{N} a_i^{\nu_{i,j}} - K_{c_j}, \quad \forall j \in [\![1, R]\!] \tag{3.16}$$

$$\Gamma = \begin{bmatrix} \mathcal{D}_{N-R} & \cdots & \mathcal{D}_N \\ \dfrac{\partial \mathcal{A}_1}{\partial x_{N-R}} & \cdots & \dfrac{\partial \mathcal{A}_1}{\partial x_N} \\ \vdots & & \vdots \\ \dfrac{\partial \mathcal{A}_R}{\partial x_{N-R}} & \cdots & \dfrac{\partial \mathcal{A}_R}{\partial x_N} \end{bmatrix} \in \mathbb{R}^{(R+1)\times(R+1)} \tag{3.17}$$

$$\bar{\Gamma} = \begin{bmatrix} \bar{\mathcal{D}}_{N-R} & \cdots & \bar{\mathcal{D}}_N \\ \dfrac{\partial \mathcal{A}_1}{\partial x_{N-R}} & \cdots & \dfrac{\partial \mathcal{A}_1}{\partial x_N} \\ \vdots & & \vdots \\ \dfrac{\partial \mathcal{A}_R}{\partial x_{N-R}} & \cdots & \dfrac{\partial \mathcal{A}_R}{\partial x_N} \end{bmatrix} \in \mathbb{R}^{(R+1)\times(R+1)} \tag{3.18}$$

$$\mathbb{A}_{ij} = U_l \left[\Gamma_{i,1}^{-1} \mathcal{D}_j + \sum_{k=1}^{R} \Gamma_{i,k+1}^{-1} \frac{\partial \mathcal{A}_k}{\partial x_j} \right], \quad \forall (i,j) \in [\![1, R+1]\!] \times [\![1, N_R]\!] \tag{3.19}$$

$$\bar{\mathbb{A}}_{ij} = U_l \left[\bar{\Gamma}_{i,1}^{-1} \bar{\mathcal{D}}_j + \sum_{k=1}^{R} \bar{\Gamma}_{i,k+1}^{-1} \frac{\partial \mathcal{A}_k}{\partial x_j} \right], \quad \forall (i,j) \in [\![1, R+1]\!] \times [\![1, N_R]\!] \tag{3.20}$$

$$\mathbb{T}_i = U_l \left[\sum_{k=1}^{R} \Gamma_{i,k+1}^{-1} \frac{\partial \mathcal{A}_k}{\partial T} - \Gamma_{i,1}^{-1} \left(\nabla_T K(T,P) \cdot x \right) \right], \quad \forall i \in [\![1, R+1]\!] \tag{3.21}$$

$$\bar{\mathbb{T}}_i = U_l \left[\sum_{k=1}^{R} \bar{\Gamma}_{i,k+1}^{-1} \frac{\partial \mathcal{A}_k}{\partial T} - \bar{\Gamma}_{i,1}^{-1} \left(\nabla_T K(T,P,x) \cdot x \right) \right], \quad \forall i \in [\![1, R+1]\!] \tag{3.22}$$

$$\mathbb{H}_j = U_l \left[\left(\sum_{k=1}^{R+1} \frac{\partial h}{\partial x_{N_R+k}} \Gamma_{k,1}^{-1} \right) \mathcal{D}_j + \sum_{m=1}^{R} \left(\sum_{k=1}^{R+1} \frac{\partial h}{\partial x_{N_R+k}} \Gamma_{k,m+1}^{-1} \right) \frac{\partial \mathcal{A}_m}{\partial x_j} \right], \quad \forall j \in [\![1, N_R]\!] \tag{3.23}$$

$$\bar{\mathbb{H}}_j = U_l \left[\left(\sum_{k=1}^{R+1} \frac{\partial h}{\partial x_{N_R+k}} \bar{\Gamma}_{k,1}^{-1} \right) \bar{\mathcal{D}}_j + \sum_{m=1}^{R} \left(\sum_{k=1}^{R+1} \frac{\partial h}{\partial x_{N_R+k}} \bar{\Gamma}_{k,m+1}^{-1} \right) \frac{\partial \mathcal{A}_m}{\partial x_j} \right], \quad \forall j \in [\![1, N_R]\!] \tag{3.24}$$

$$\mathbb{S} = \sum_{m=1}^{R} \left(\sum_{k=1}^{R+1} \frac{\partial h}{\partial x_{N_R+k}} \Gamma_{k,m+1}^{-1} \right) \frac{\partial \mathcal{A}_m}{\partial T} - \left(\nabla_T K(T,P) \cdot x \right) \sum_{k=1}^{R+1} \frac{\partial h}{\partial x_{N_R+k}} \Gamma_{k,1}^{-1} \tag{3.25}$$

$$\bar{\mathbb{S}} = \sum_{m=1}^{R} \left(\sum_{k=1}^{R+1} \frac{\partial h}{\partial x_{N_R+k}} \bar{\Gamma}_{k,m+1}^{-1} \right) \frac{\partial \mathcal{A}_m}{\partial T} - \left(\nabla_T K(T,P,x) \cdot x \right) \sum_{k=1}^{R+1} \frac{\partial h}{\partial x_{N_R+k}} \bar{\Gamma}_{k,1}^{-1} \tag{3.26}$$

$$\mathcal{N}_a = \begin{bmatrix} -\sum_{i=1}^{N} \nu_{i,1} & \cdots & -\sum_{i=1}^{N} \nu_{i,R} & 1 & 0 & \cdots & 0 & 0 \\ -\nu_{1,1} & \cdots & -\nu_{1,R} & x_1 & U_l & \cdots & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\nu_{N_R,1} & \cdots & -\nu_{N_R,R} & x_{N_R} & 0 & \cdots & U_l & 0 \\ -\nu_{N-R,1} & \cdots & -\nu_{N-R,R} & x_{N_R} & -\mathbb{A}_{1,1} & \cdots & -\mathbb{A}_{1,N_R} & -\mathbb{T}_1 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ -\nu_{N,1} & \cdots & -\nu_{N,R} & x_N & -\mathbb{A}_{R+1,1} & \cdots & -\mathbb{A}_{R+1,N_R} & -\mathbb{T}_{R+1} \end{bmatrix} \tag{3.27}$$

$$\bar{\mathcal{N}}_a = \begin{bmatrix} -\sum_{i=1}^{N} \nu_{i,1} & \cdots & -\sum_{i=1}^{N} \nu_{i,R} & 1 & 0 & \cdots & 0 & 0 \\ -\nu_{1,1} & \cdots & -\nu_{1,R} & x_1 & U_l & \cdots & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\nu_{N_R,1} & \cdots & -\nu_{N_R,R} & x_{N_R} & 0 & \cdots & U_l & 0 \\ -\nu_{N-R,1} & \cdots & -\nu_{N-R,R} & x_{N_R} & -\bar{\mathbb{A}}_{1,1} & \cdots & -\bar{\mathbb{A}}_{1,N_R} & -\bar{\mathbb{T}}_1 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ -\nu_{N,1} & \cdots & -\nu_{N,R} & x_N & -\bar{\mathbb{A}}_{R+1,1} & \cdots & -\bar{\mathbb{A}}_{R+1,N_R} & -\bar{\mathbb{T}}_{R+1} \end{bmatrix} \tag{3.28}$$

$$\mathcal{N}_b = \begin{pmatrix} Q_{r_1} & \cdots & Q_{r_R} & h & \frac{\partial h}{\partial x_1} U_l - \mathbb{H}_1 & \cdots & \frac{\partial h}{\partial x_{N_R}} U_l - \mathbb{H}_{N_R} & \frac{\partial h}{\partial T} U_l - \mathbb{S} \end{pmatrix} \tag{3.29}$$

$$\bar{\mathcal{N}}_b = \begin{pmatrix} Q_{r_1} & \cdots & Q_{r_R} & h & \frac{\partial h}{\partial x_1} U_l - \bar{\mathbb{H}}_1 & \cdots & \frac{\partial h}{\partial x_{N_R}} U_l - \bar{\mathbb{H}}_{N_R} & \frac{\partial h}{\partial T} U_l - \bar{\mathbb{S}} \end{pmatrix} \tag{3.30}$$

3.3 Distillation de Rayleigh non réactive

Nous nous intéressons successivement à deux hypothèses thermodynamiques, donnant lieu à deux modèles mathématiques. Dans un premier temps, les phases liquide et vapeur sont considérées comme étant idéales, puis on suppose que la phase liquide est non idéale. La méthode de déflation peut s'appliquer localement à chacun des modèles obtenus sous réserve de certaines hypothèses, dont la principale est la nullité ou la non nullité du débit vapeur. La discussion distingue donc tout naturellement le régime monophasique du régime diphasique. Nous montrons également que la méthode ne peut pas être déroulée au point de bulle, point de transition entre les deux régimes.

3.3.1 Phases liquide et vapeur idéales

Le problème de la distillation de Rayleigh non réactive pour lequel les deux phases sont considérées comme idéales peut être décrit par le système suivant :

$$\begin{cases} \dot{U}_l = -V & \text{(3.31a)} \\ \dot{\overline{x_i U_l}} = -V y_i, \quad \forall\, i \in [\![1, N]\!] & \text{(3.31b)} \\ \dot{\overline{hU_l}} = Q - VH & \text{(3.31c)} \\ 0 = y_i - K_i(T,P)x_i, \quad \forall\, i \in [\![1, N]\!] & \text{(3.31d)} \\ 0 = \sum_{i=1}^{N} (x_i - y_i) - y_{\text{inerte}} & \text{(3.31e)} \\ 0 = V y_{\text{inerte}}. & \text{(3.31f)} \end{cases}$$

Ce modèle est présenté dans [1]. L'équation (3.31a) constitue le *bilan matière global*. Elle indique que toute variation du nombre de moles en phase liquide se traduit par une variation, égale en valeur absolue, mais de signe opposé du nombre de moles en phase vapeur. Au lieu de considérer la variation du nombre total de moles dans la phase liquide, il est possible de considérer la variation du nombre de moles de chaque constituant dans la phase liquide. On obtient ainsi les N bilans matière partiels (3.31b). Le *bilan énergétique* (ou *enthalpique*) est quant à lui décrit par (3.31c). Les *équilibres thermodynamiques* entre la phase liquide et la phase vapeur pour chaque espèce chimique sont fournis pas (3.31d). L'équation de sommation (3.31e) caractérise les fractions molaires des constituants étudiés. (3.31f) caractérise l'état de la vapeur juste au dessus de la phase liquide : lorsque le débit vapeur est nul, en régime monophasique, ce film de vapeur contient une certaine concentration de vapeur de gaz inerte (air) ; avec l'ébullition, cette concentration devient nulle et un débit vapeur non nul peut s'établir.

3.3.1.1 Régime monophasique

Le régime monophasique correspond à la montée en température du système, jusqu'à l'apparition de la première bulle caractérisant le début du changement de phase. Pendant le régime monophasique, on suppose qu'il n'y a pas de débit vapeur ; l'énergie fournie par la chauffe fait monter le liquide en température. Tant que la température d'ébullition n'est pas atteinte, l'énergie s'accumule mais n'est pas suffisante pour l'apparition d'un débit vapeur. Puisque $V = 0$, on a $y_{\text{inerte}} \neq 0$. On considère la notation définie par (3.2).

Théorème 21

Supposons que $y_{\text{inerte}} \neq 0$, $U_l \neq 0$ *et* $\dfrac{\partial h}{\partial T} \neq 0$. *Alors l'EDA* (3.31) *est équivalente au*

système différentiel

$$\begin{pmatrix} \dot{U}_l \\ \dot{x}_1 \\ \vdots \\ \dot{x}_N \\ \dot{T} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \left(\dfrac{\partial h}{\partial T} U_l\right)^{-1} Q \end{pmatrix}, \quad \text{avec} \quad \begin{pmatrix} y_1 \\ \vdots \\ y_N \\ V \\ y_{\text{inerte}} \end{pmatrix} = \begin{pmatrix} K_1(T,P)x_1 \\ \vdots \\ K_N(T,P)x_N \\ 0 \\ \displaystyle\sum_{i=1}^{N} \mathcal{D}_i x_i \end{pmatrix}.$$

Dans ces conditions, le problème (3.31) est une EDA d'indice 1.

Preuve - On écrit le système (3.31) sous la forme matricielle (3.1), avec

$$X_0 = \begin{pmatrix} U_l \\ x \\ T \\ y \\ V \\ y_{\text{inerte}} \end{pmatrix} \in \mathbb{R}^{2N+4}, \quad f_0(X_0) = \begin{pmatrix} -V \\ -Vy \\ Q - VH \\ y_1 - K_1(T,P)x_1 \\ \vdots \\ y_N - K_N(T,P)x_N \\ \displaystyle\sum_{i=1}^{N} (x_i - y_i) - y_{\text{inerte}} \\ V y_{\text{inerte}} \end{pmatrix} \in \mathbb{R}^{2N+4}$$

et

$$E_0(X_0) = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ x_1 & U_l & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots \\ x_N & 0 & \cdots & U_l & 0 & 0 & \cdots & 0 \\ h & \dfrac{\partial h}{\partial x_1} U_l & \cdots & \dfrac{\partial h}{\partial x_N} U_l & \dfrac{\partial h}{\partial T} U_l & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{(2N+4)\times(2N+4)}.$$

Puisque $U_l \neq 0$ et $\dfrac{\partial h}{\partial T} \neq 0$, le rang de la matrice $E_0(X_0)$ vaut $N+2$. Ainsi,

$$\tilde{E}_0(X_0) = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ x_1 & U_l & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_N & 0 & \cdots & U_l & 0 \\ h & \dfrac{\partial h}{\partial x_1} U_l & \cdots & \dfrac{\partial h}{\partial x_N} U_l & \dfrac{\partial h}{\partial T} U_l \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}$$

et

$$\tilde{f}_0(X_0) = \begin{pmatrix} y_1 - K_1(T,P)x_1 \\ \vdots \\ y_N - K_N(T,P)x_N \\ \displaystyle\sum_{i=1}^{N} (x_i - y_i) - y_{\text{inerte}} \\ V y_{\text{inerte}} \end{pmatrix} \in \mathbb{R}^{N+2}. \tag{3.32}$$

De plus, on a

$$g_0(X_0) = \begin{pmatrix} -V \\ -Vy \\ Q - VH \end{pmatrix} \in \mathbb{R}^{N+2}.$$

Il n'est pas nécessaire de permuter les composantes du vecteur X_0 dans ce contexte. X_1 et Y_1 s'écrivent :

$$X_0 = \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \quad \text{où } X_1 = \begin{pmatrix} U_l \\ x \\ T \end{pmatrix} \in \mathbb{R}^{N+2} \text{ et } Y_1 = \begin{pmatrix} y \\ V \\ y_{\text{inerte}} \end{pmatrix} \in \mathbb{R}^{N+2}.$$

Ainsi, par dérivation de (3.32) par rapport à Y_1, on a

$$J_{Y_1}\left(\bar{f}_0\right) = \begin{bmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ -1 & \cdots & -1 & 0 & -1 \\ 0 & \cdots & 0 & y_{\text{inerte}} & V \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}.$$

On peut aisément établir que

$$\det J_{Y_1}\left(\bar{f}_0\right) = y_{\text{inerte}}.$$

De ce fait, on est en mesure d'appliquer le théorème des fonctions implicites au système algébrique $\tilde{f}_0\left(X_0\right) = 0$, où $\tilde{f}_0\left(X_0\right)$ est définie par (3.32). Plus précisément, il existe une fonction φ_1 telle que :

$$Y_1 = \varphi_1\left(X_1\right) \Leftrightarrow \begin{pmatrix} y_1 \\ \vdots \\ y_N \\ V \\ y_{\text{inerte}} \end{pmatrix} = \begin{pmatrix} K_1(T,P)x_1 \\ \vdots \\ K_N(T,P)x_N \\ 0 \\ \displaystyle\sum_{i=1}^{N} \mathcal{D}_i x_i \end{pmatrix} \tag{3.33}$$

Par ailleurs, on obtient

$$A_0(X_1, Y_1) = \tilde{E}_0(X_1, Y_1) \in \mathbb{R}^{(N+2)\times(N+2)},$$

ainsi que

$$B_0(X_1, Y_1) = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}.$$

La nouvelle matrice coefficient $E_1(X_1)$ est alors la matrice $\tilde{E}_0(X_1, \varphi_1(X_1))$. La première étape de la méthode de déflation se termine en fournissant

$$E_1(X_1) = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ x_1 & U_l & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_N & 0 & \cdots & U_l & 0 \\ h & \dfrac{\partial h}{\partial x_1} U_l & \cdots & \dfrac{\partial h}{\partial x_N} U_l & \dfrac{\partial h}{\partial T} U_l \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)},$$

$$X_1 = \begin{pmatrix} U_l \\ x \\ T \end{pmatrix} \in \mathbb{R}^{N+2} \text{ et } f_1(X_1) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ Q \end{pmatrix} \in \mathbb{R}^{N+2}.$$

Comme la matrice $E_1(X_1)$ est inversible, le processus de déflation s'achève. Le problème (3.31) est alors équivalent au système

$$\begin{cases} \dot{X}_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ x_1 & U_l & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_N & 0 & \cdots & U_l & 0 \\ h & \dfrac{\partial h}{\partial x_1} U_l & \cdots & \dfrac{\partial h}{\partial x_N} U_l & \dfrac{\partial h}{\partial T} U_l \end{bmatrix}^{-1} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ Q \end{pmatrix} \\ 0 = \tilde{f}_0\left(X_1, Y_1\right). \end{cases}$$

Les calculs étant simples dans ce cas, il est possible d'écrire explicitement le système :

$$\begin{pmatrix} \dot{U}_l \\ \dot{x}_1 \\ \vdots \\ \dot{x}_N \\ \dot{T} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \left(\dfrac{\partial h}{\partial T} U_l\right)^{-1} Q \end{pmatrix},$$

avec

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \\ V \\ y_{\text{inerte}} \end{pmatrix} = \begin{pmatrix} K_1(T,P) x_1 \\ \vdots \\ K_N(T,P) x_N \\ 0 \\ \displaystyle\sum_{i=1}^{N} \mathcal{D}_i x_i \end{pmatrix}.$$

Le processus de déflation ne comporte qu'une seule étape. Nous sommes en présence d'un problème d'indice 1. □

Remarque 21

- *Quand il n'y a pas de permutation des composantes du vecteur X_0, les calculs sont grandement simplifiés grâce à la nullité de la matrice $B_0(X_1, Y_1)$. Il n'est pas nécessaire de calculer la matrice Jacobienne $J_{X_1}\left(\bar{f}_0\right)$. De plus, l'expression explicite de la matrice Jacobienne $J_{Y_1}\left(\bar{f}_0\right)$ ne sert qu'à donner son déterminant.*
- *Les composantes du vecteur Y_1, à savoir y, V et y_{inerte} sont exprimées en fonction de celles du vecteur X_1, à savoir U_l, x et T grâce à (3.33).*
- *D'un point de vue structurel, le système (3.31) satisfaisant les hypothèses du théorème précédent est un problème* semi-explicite d'indice 1, *c'est-à-dire de la forme*

$$\begin{cases} \dot{x} = f(x, y) \\ 0 = g(x, y), \end{cases}$$

où la matrice Jacobienne $g_y(x, y)$ est inversible.

3.3.1.2 Régime diphasique

Observons à présent le comportement du système à l'ébullition, lorsque les deux phases coexistent et que le débit vapeur est non nul. Puisque $V \neq 0$, on a $y_{\text{inerte}} = 0$. On considère les notations définies par (3.2), (3.4), (3.6), (3.8) et (3.10).

Théorème 22

Supposons que $V \neq 0$, $U_l \neq 0$, $\dfrac{\partial h}{\partial T} \neq 0$, $\mathcal{D}_N \neq 0$, $\nabla_T K(T,P)\cdot x \neq 0$ et $\mathcal{G} \neq 0$. Alors l'EDA (3.31) est équivalente au système différentiel

$$\begin{pmatrix} \dot{U}_l \\ \dot{x}_1 \\ \vdots \\ \dot{x}_{N-1} \\ \dot{T} \end{pmatrix} = \begin{pmatrix} -\mathcal{G}^{-1}Q \\ U_l^{-1}\mathcal{G}^{-1}Qx_1\mathcal{D}_1 \\ \vdots \\ U_l^{-1}\mathcal{G}^{-1}Qx_{N-1}\mathcal{D}_{N-1} \\ U_l^{-1}\left(\nabla_T K(T,P)\cdot x\right)^{-1}\mathcal{G}^{-1}Q\displaystyle\sum_{i=1}^{N}\mathcal{D}_i^2 x_i \end{pmatrix},$$

avec

$$\begin{pmatrix} x_N \\ y_1 \\ \vdots \\ y_{N-1} \\ y_N \\ y_{\text{inerte}} \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{i=1}^{N-1}\mathcal{K}_i x_i \\ K_1(T,P)x_1 \\ \vdots \\ K_{N-1}(T,P)x_{N-1} \\ K_N(T,P)\displaystyle\sum_{i=1}^{N-1}\mathcal{K}_i x_i \\ 0 \end{pmatrix} \text{ et } V = \mathcal{G}^{-1}Q.$$

Dans ces conditions, le problème (3.31) est une EDA d'indice 2.

Preuve - On est amené à déterminer l'expression de la variable V ; or la seule équation purement algébrique dans laquelle cette dernière figure est l'équation (3.31f). Cette équation ne permet pas d'exprimer V en fonction des autres variables. Ainsi, il est nécessaire de procéder à une permutation des variables afin de faire figurer V dans les variables différentielles. Précisons que les expressions X_0, $E_0(X_0)$, $f_0(X_0)$, $\tilde{E}_0(X_0)$, $\tilde{f}_0(X_0)$ et $g_0(X_0)$ sont bien entendu les mêmes que dans le régime monophasique.

Soit $\mathcal{P}_1 \in \mathbb{R}^{(2N+4)\times(2N+4)}$ la matrice de permutation telle que

$$\mathcal{P}_1 X_0 = \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} \text{ où } X_1 = \begin{pmatrix} U_l \\ x_1 \\ \vdots \\ x_{N-1} \\ T \\ V \end{pmatrix} \in \mathbb{R}^{N+2} \text{ et } Y_1 = \begin{pmatrix} x_N \\ y \\ y_{\text{inerte}} \end{pmatrix} \in \mathbb{R}^{N+2}.$$

D'une part, par dérivation de (3.32) par rapport à Y_1, on a

$$J_{Y_1}\left(\bar{f}_0\right) = \begin{bmatrix} \gamma_1 & \gamma_2 \\ 0 & V \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}, \tag{3.34}$$

avec

$$\gamma_1 = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -K_N(T,P) & 0 & \cdots & 1 \\ 1 & -1 & \cdots & -1 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)} \text{ et } \gamma_2 = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -1 \end{pmatrix} \in \mathbb{R}^{N+1}.$$

On peut aisément montrer que $\det J_{Y_1}\left(\bar{f}_0\right) = (-1)^N V \mathcal{D}_N$. Le théorème des fonctions implicites s'applique donc au système algébrique $\tilde{f}_0(X_0) = 0$, où $\tilde{f}_0(X_0)$ est définie dans (3.32). Plus précisément, il existe une fonction $\varphi_1 : \mathcal{I}_0 \subset \mathbb{R}^{N+2} \to \mathbb{R}^{N+2}$ telle que :

$$Y_1 = \varphi_1(X_1) \Leftrightarrow \begin{pmatrix} x_N \\ y_1 \\ \vdots \\ y_{N-1} \\ y_N \\ y_{\text{inerte}} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N-1} \mathcal{K}_i\, x_i \\ K_1(T,P)\, x_1 \\ \vdots \\ K_{N-1}(T,P)\, x_{N-1} \\ K_N(T,P) \sum_{i=1}^{N-1} \mathcal{K}_i\, x_i \\ 0 \end{pmatrix}. \tag{3.35}$$

D'autre part, par dérivation de (3.32) par rapport à X_1, on obtient

$$J_{X_1}\left(\bar{f}_0\right) = \begin{bmatrix} \delta_1 & 0 \\ 0 & y_{\text{inerte}} \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}, \tag{3.36}$$

où

$$\delta_1 = \begin{bmatrix} 0 & -K_1(T,P) & \cdots & 0 & -\dfrac{\partial K_1}{\partial T}\, x_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -K_{N-1}(T,P) & -\dfrac{\partial K_{N-1}}{\partial T}\, x_{N-1} \\ 0 & 0 & \cdots & 0 & -\dfrac{\partial K_N}{\partial T}\, x_N \\ 0 & 1 & \cdots & 1 & 0 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}.$$

Par ailleurs, on obtient

$$A_0(X_1, Y_1) = \begin{bmatrix} \alpha_1 & 0 \\ \alpha_3 & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}, \tag{3.37}$$

avec

$$\alpha_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ x_1 & U_l & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{N-1} & 0 & \cdots & U_l & 0 \end{bmatrix} \in \mathbb{R}^{N\times(N+1)}$$

et

$$\alpha_3 = \begin{bmatrix} x_N & 0 & \cdots & 0 & 0 \\ h & \dfrac{\partial h}{\partial x_1} U_l & \cdots & \dfrac{\partial h}{\partial x_{N-1}} U_l & \dfrac{\partial h}{\partial T} U_l \end{bmatrix} \in \mathbb{R}^{2\times(N+1)}$$

ainsi que

$$B_0(X_1, Y_1) = \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}, \tag{3.38}$$

où

$$\beta_3 = \begin{bmatrix} U_l & 0 & \cdots & 0 \\ \dfrac{\partial h}{\partial x_N} U_l & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{2\times(N+1)}.$$

Déterminons à présent la matrice $E_1(X_1)$ grâce aux matrices (3.37), (3.38), (3.34) et (3.36) :

$$\begin{aligned} E_1(X_1) &= \begin{bmatrix} \alpha_1 & 0 \\ \alpha_3 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} \gamma_1 & \gamma_2 \\ 0 & V \end{bmatrix}^{-1} \begin{bmatrix} \delta_1 & 0 \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & 0 \\ \alpha_3 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} \gamma_1^{-1} & -\gamma_1^{-1}\gamma_2 V^{-1} \\ 0 & V^{-1} \end{bmatrix} \begin{bmatrix} \delta_1 & 0 \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & 0 \\ \alpha_3 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} \gamma_1^{-1}\delta_1 & 0 \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & 0 \\ \alpha_3 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3\gamma_1^{-1}\delta_1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & 0 \\ \bar{\alpha}_3 & 0 \end{bmatrix}, \end{aligned}$$

où $\bar{\alpha}_3 = \alpha_3 - \beta_3\gamma_1^{-1}\delta_1$. Par de simples calculs, on montre que

$$\gamma_1^{-1} = \begin{bmatrix} \mathcal{D}_N^{-1} & \cdots & \mathcal{D}_N^{-1} & \mathcal{D}_N^{-1} & \mathcal{D}_N^{-1} \\ 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ \mathcal{D}_N^{-1} K_N(T,P) & \cdots & \mathcal{D}_N^{-1} K_N(T,P) & \mathcal{D}_N^{-1} & \mathcal{D}_N^{-1} K_N(T,P) \end{bmatrix}.$$

On obtient ainsi

$$\bar{\alpha}_3 = \begin{bmatrix} x_N & \mathcal{K}_1 U_l & \cdots & \mathcal{K}_{N-1} U_l & \mathcal{D}_N^{-1} \left(\nabla_T K(T,P) \cdot x\right) U_l \\ h & \mathcal{H}_1 U_l & \cdots & \mathcal{H}_{N-1} U_l & \mathcal{L} U_l \end{bmatrix}.$$

La première étape de la méthode se termine avec

$$X_1 = \begin{pmatrix} U_l \\ x_1 \\ \vdots \\ x_{N-1} \\ T \\ V \end{pmatrix} \in \mathbb{R}^{N+2} \text{ et } Y_1 = \begin{pmatrix} x_N \\ y \\ y_{\text{inerte}} \end{pmatrix} \in \mathbb{R}^{N+2},$$

$$E_1(X_1) = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ x_1 & U_l & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{N-1} & 0 & \cdots & U_l & 0 & 0 \\ x_N & \mathcal{K}_1 U_l & \cdots & \mathcal{K}_{N-1} U_l & \mathcal{D}_N^{-1} \left(\nabla_T K(T,P) \cdot x\right) U_l & 0 \\ h & \mathcal{H}_1 U_l & \cdots & \mathcal{H}_{N-1} U_l & \mathcal{L} U_l & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}$$

et

$$f_1(X_1) = \begin{pmatrix} -V \\ -V\, K_1(T,P)\, x_1 \\ \vdots \\ -V\, K_{N-1}(T,P)\, x_{N-1} \\ -V\, K_N(T,P)\, x_N \\ Q - V\, H \end{pmatrix} \in \mathbb{R}^{N+2}.$$

La matrice E_1 étant singulière, on itère le processus de déflation.

On considère maintenant le problème quasi-linéaire $E_1(X_1)\,\dot{X}_1 = f_1(X_1)$. Procédons en premier lieu à la transformation de la matrice coefficient $E_1(X_1)$ en appliquant la décomposition LU à cette dernière. Le problème précédent est alors équivalent au système $U_1(X_1)\,\dot{X}_1 = L_1(X_1)^{-1}\,f_1(X_1)$. En gardant les mêmes notations ($E_1(X_1)$ et $f_1(X_1)$), on a

$$E_1(X_1) = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & U_l & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & U_l & 0 & 0 \\ 0 & 0 & \cdots & 0 & \mathcal{D}_N^{-1}\,(\nabla_T K(T,P)\cdot x)\,U_l & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}$$

et

$$f_1(X_1) = \begin{pmatrix} -V \\ V\,x_1\,\mathcal{D}_1 \\ \vdots \\ V\,x_{N-1}\,\mathcal{D}_{N-1} \\ -V\sum_{i=1}^{N}\mathcal{K}_i\,\mathcal{D}_i\,x_i \\ Q - \mathcal{G}\,V \end{pmatrix} \in \mathbb{R}^{N+2}.$$

Puisque $\mathcal{G} \neq 0$ par hypothèse, il est possible d'extraire une expression de V ; il existe une fonction $\varphi_2 : \mathcal{I}_1 \subset \mathbb{R}^{N+1} \to \mathbb{R}$ telle que

$$V = \varphi_2 \begin{pmatrix} U_l \\ x_1 \\ \vdots \\ x_{N-1} \\ T \end{pmatrix} = \mathcal{G}^{-1}\,Q.$$

Ainsi, il n'est pas nécessaire de permuter les inconnues de X_1 :

$$X_2 = \begin{pmatrix} U_l \\ x_1 \\ \vdots \\ x_{N-1} \\ T \end{pmatrix} \text{ et } Y_2 = V = \varphi_2(X_2).$$

La matrice $B_1(X_2, Y_2)$ étant nulle, on a

$$\begin{aligned} E_2(X_2) &= A_1(X_2, \varphi_2(X_2)) \\ &= \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & U_l & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & U_l & 0 \\ 0 & 0 & \cdots & 0 & \mathcal{D}_N^{-1}\,(\nabla_T K(T,P)\cdot x)\,U_l \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}. \end{aligned}$$

Cette matrice est inversible par hypothèse. La méthode de déflation se termine en fournissant l'EDO sous contraintes suivante

$$\begin{pmatrix} \dot{U}_l \\ \dot{x}_1 \\ \vdots \\ \dot{x}_{N-1} \\ \dot{T} \end{pmatrix} = \begin{pmatrix} -\mathcal{G}^{-1}Q \\ U_l^{-1}\mathcal{G}^{-1}Qx_1\mathcal{D}_1 \\ \vdots \\ U_l^{-1}\mathcal{G}^{-1}Qx_{N-1}\mathcal{D}_{N-1} \\ U_l^{-1}\left(\nabla_T K(T,P)\cdot x\right)^{-1}\mathcal{G}^{-1}Q\sum_{i=1}^{N}\mathcal{D}_i^2 x_i \end{pmatrix}, \begin{pmatrix} x_N \\ y_1 \\ \vdots \\ y_{N-1} \\ y_N \\ y_{\text{inerte}} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N-1}\mathcal{K}_i x_i \\ K_1(T,P)x_1 \\ \vdots \\ K_{N-1}(T,P)x_{N-1} \\ K_N(T,P)\sum_{i=1}^{N-1}\mathcal{K}_i x_i \\ 0 \end{pmatrix}$$

et $V = \mathcal{G}^{-1}Q$. Deux étapes du processus de déflation sont nécessaires pour résoudre le problème ; ainsi nous sommes en présence d'une EDA d'indice 2. □

Remarque 22

- *Dans la première étape de la méthode, on peut toujours supposer que la variable x_N est amenée dans la partie algébrique, quitte à changer la numérotation des indices. On se place ainsi dans le cadre le plus général possible.*
- *Seule la première ligne de la matrice γ_1^{-1} est en réalité utilisée. Il est donc possible d'éviter le calcul complet de l'inverse de γ_1.*

À partir de l'ébullition et contrairement au cas monophasique, les fractions molaires de chaque constituant dans chaque phase peuvent varier au cours du temps. C'est le cas en général, à l'exception des mélanges azéotropiques pour lesquels les proportions des fractions molaires dans chaque phase sont constantes.

L'hypothèse d'idéalité de la phase vapeur est fréquemment utilisée. Elle correspond à la modélisation des gaz parfaits. En revanche, la phase liquide est en général considérée comme non idéale. En nous appuyant sur les résultats précédents, nous pouvons généraliser l'étude au cas d'une phase liquide considérée comme non idéale.

3.3.2 Phase liquide non idéale et phase vapeur idéale

Le problème de la distillation de Rayleigh non réactive pour lequel la phase liquide est considérée comme non idéale est décrit par le système suivant :

$$\begin{cases} \dot{U}_l = -V \\ \dot{\overline{x_i U_l}} = -V y_i, \quad \forall\, i \in [\![1, N]\!] \\ \dot{\overline{h U_l}} = Q - VH \\ 0 = y_i - K_i(T,P,x)x_i, \quad \forall\, i \in [\![1, N]\!] \\ 0 = \sum_{i=1}^{N}(x_i - y_i) - y_{\text{inerte}} \\ 0 = V y_{\text{inerte}}. \end{cases} \tag{3.39}$$

Dans ce contexte de phase liquide non idéale, les équilibres thermodynamiques K_i deviennent également dépendants de x.

3.3.2.1 Régime monophasique

Théorème 23

Supposons que $y_{\text{inerte}} \neq 0$, $U_l \neq 0$ *et* $\dfrac{\partial h}{\partial T} \neq 0$. *Alors l'EDA* (3.39) *est équivalente au système différentiel*

$$\begin{pmatrix} \dot{U}_l \\ \dot{x}_1 \\ \vdots \\ \dot{x}_N \\ \dot{T} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \left[\dfrac{\partial h}{\partial T}\, U_l\right]^{-1} Q \end{pmatrix},$$

avec

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \\ V \\ y_{\text{inerte}} \end{pmatrix} = \begin{pmatrix} K_1(T,P,x)\, x_1 \\ \vdots \\ K_N(T,P,x)\, x_N \\ 0 \\ \displaystyle\sum_{i=1}^{N} (1 - K_i(T,P,x))\, x_i \end{pmatrix}.$$

Dans ces conditions, le problème (3.39) *est une EDA d'indice 1.*

Preuve - La preuve est en tout point similaire à celle dans le cas idéal. □

3.3.2.2 Régime diphasique

On considère les notations définies par (3.3), (3.5), (3.7), (3.9) et (3.11).

Théorème 24

Supposons que $V \neq 0$, $U_l \neq 0$, $\dfrac{\partial h}{\partial T} \neq 0$, $\bar{\mathcal{D}}_N \neq 0$, $\nabla_T K(T,P,x) \cdot x \neq 0$ *et* $\bar{\mathcal{G}} \neq 0$. *Alors il existe une fonction* $\phi_1 : \mathcal{I} \subset \mathbb{R}^N \to \mathbb{R}$ *où* $x_N = \phi_1(T, x_1, \ldots, x_{N-1})$ *telle que l'EDA* (3.39) *soit équivalente au système différentiel*

$$\begin{pmatrix} \dot{U}_l \\ \dot{x}_1 \\ \vdots \\ \dot{x}_{N-1} \\ \dot{T} \end{pmatrix} = \begin{pmatrix} -\bar{\mathcal{G}}^{-1} Q \\ U_l^{-1}\, \bar{\mathcal{G}}^{-1} Q\, x_1\, (1 - K_1(T,P,x)) \\ \vdots \\ U_l^{-1}\, \bar{\mathcal{G}}^{-1} Q\, x_{N-1}\, (1 - K_{N-1}(T,P,x)) \\ U_l^{-1}\, \left(\nabla_T K(T,P,x) \cdot x\right)^{-1}\, \bar{\mathcal{G}}^{-1} Q \displaystyle\sum_{i=1}^{N} \bar{\mathcal{D}}_i\, (1 - K_i(T,P,x))\, x_i \end{pmatrix},$$

avec

$$\begin{pmatrix} x_N \\ y_1 \\ \vdots \\ y_{N-1} \\ y_N \\ y_{\text{inerte}} \end{pmatrix} = \begin{pmatrix} \phi_1(T, x_1, \ldots, x_{N-1}) \\ K_1(T,P,x)\, x_1 \\ \vdots \\ K_{N-1}(T,P,x)\, x_{N-1} \\ K_N(T,P,x)\, \phi_1(T, x_1, \ldots, x_{N-1}) \\ 0 \end{pmatrix} \text{ et } V = \bar{\mathcal{G}}^{-1} Q.$$

Dans ces conditions, le problème (3.39) *est une EDA d'indice 2.*

Preuve - Le processus de résolution dans le cas non idéal est en tout point similaire au cas idéal jusqu'à la détermination de la matrice Jacobienne $J_{Y_1}\left(\bar{f}_0\right)$. En effet, on a

$$\tilde{f}_0(X_0) = \begin{pmatrix} y_1 - K_1(T,P,x)\,x_1 \\ \vdots \\ y_N - K_N(T,P,x)\,x_N \\ \sum_{i=1}^{N}(x_i - y_i) - y_{\text{inerte}} \\ V\,y_{\text{inerte}} \end{pmatrix} \in \mathbb{R}^{N+2}. \tag{3.40}$$

Par dérivation de (3.40) par rapport à Y_1, on a

$$J_{Y_1}\left(\bar{f}_0\right) = \begin{bmatrix} \gamma_1 & \gamma_2 \\ 0 & V \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}, \tag{3.41}$$

avec

$$\gamma_1 = \begin{bmatrix} -\dfrac{\partial K_1(T,P,x)}{\partial x_N}\,x_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\dfrac{\partial K_N(T,P,x)}{\partial x_N}\,x_N - K_N(T,P,x) & 0 & \cdots & 1 \\ 1 & -1 & \cdots & -1 \end{bmatrix} \text{ et } \gamma_2 = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -1 \end{pmatrix} \in \mathbb{R}^{N+1}.$$
$$\in \mathbb{R}^{(N+1)\times(N+1)}$$

On peut facilement montrer que $\det J_{Y_1}\left(\bar{f}_0\right) = (-1)^N\,V\,\bar{\mathcal{D}}_N$. Le théorème des fonctions implicites s'applique donc au système algébrique $\tilde{f}_0\left(X_0\right) = 0$, où $\tilde{f}_0\left(X_0\right)$ est définie dans (3.40). Plus précisément, il existe deux fonctions $\varphi_1 : \mathcal{I}_0 \subset \mathbb{R}^{N+2} \to \mathbb{R}^{N+2}$ et $\phi_1 : \mathcal{I} \subset \mathbb{R}^N \to \mathbb{R}$ telles que :

$$Y_1 = \varphi_1(X_1) \Leftrightarrow \begin{pmatrix} x_N \\ y_1 \\ \vdots \\ y_{N-1} \\ y_N \\ y_{\text{inerte}} \end{pmatrix} = \begin{pmatrix} \phi_1(T,x_1,\ldots,x_{N-1}) \\ K_1(T,P,x)\,x_1 \\ \vdots \\ K_{N-1}(T,P,x)\,x_{N-1} \\ K_N(T,P,x)\,\phi_1(T,x_1,\ldots,x_{N-1}) \\ 0 \end{pmatrix}. \tag{3.42}$$

Également, par dérivation de (3.40) par rapport à X_1, on obtient

$$J_{X_1}\left(\bar{f}_0\right) = \begin{bmatrix} \delta_1 & 0 \\ 0 & y_{\text{inerte}} \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}, \tag{3.43}$$

où

$$\delta_1 = \begin{bmatrix} 0 & -\dfrac{\partial K_1}{\partial x_1}\,x_1 - K_1 & \cdots & -\dfrac{\partial K_1}{\partial x_{N-1}}\,x_1 & -\dfrac{\partial K_1}{\partial T}\,x_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & -\dfrac{\partial K_{N-1}}{\partial x_1}\,x_{N-1} & \cdots & -\dfrac{\partial K_{N-1}}{\partial x_{N-1}}\,x_{N-1} - K_{N-1} & -\dfrac{\partial K_{N-1}}{\partial T}\,x_{N-1} \\ 0 & -\dfrac{\partial K_N}{\partial x_1}\,x_N & \cdots & -\dfrac{\partial K_N}{\partial x_{N-1}}\,x_N & -\dfrac{\partial K_N}{\partial T}\,x_N \\ 0 & 1 & \cdots & 1 & 0 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)},$$

où les K_i dépendent de T, P et x. Déterminons à présent la matrice $E_1(X_1)$ grâce aux matrices (3.37), (3.38), (3.41) et (3.43) :

$$\begin{aligned}
E_1(X_1) &= \begin{bmatrix} \alpha_1 & 0 \\ \alpha_3 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} \gamma_1 & \gamma_2 \\ 0 & V \end{bmatrix}^{-1} \begin{bmatrix} \delta_1 & 0 \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} \alpha_1 & 0 \\ \alpha_3 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} \gamma_1^{-1} & -\gamma_1^{-1}\gamma_2 V^{-1} \\ 0 & V^{-1} \end{bmatrix} \begin{bmatrix} \delta_1 & 0 \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} \alpha_1 & 0 \\ \alpha_3 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} \gamma_1^{-1}\delta_1 & 0 \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} \alpha_1 & 0 \\ \alpha_3 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3\gamma_1^{-1}\delta_1 & 0 \end{bmatrix} \\
&= \begin{bmatrix} \alpha_1 & 0 \\ \bar{\alpha}_3 & 0 \end{bmatrix},
\end{aligned}$$

où $\bar{\alpha}_3 = \alpha_3 - \beta_3\,\gamma_1^{-1}\,\delta_1$. Par de simples calculs, on montre que la première ligne de la matrice γ_1^{-1} est :

$$\bar{\mathcal{D}}_N^{-1} \begin{bmatrix} 1 & \cdots & 1 & 1 & 1 \end{bmatrix}.$$

On obtient ainsi

$$\bar{\alpha}_3 = \begin{bmatrix} x_N & \bar{\mathcal{K}}_1\,U_l & \cdots & \bar{\mathcal{K}}_{N-1}\,U_l & \bar{\mathcal{D}}_N^{-1}\,(\nabla_T K(T,P,x)\cdot x)\,U_l \\ h & \bar{\mathcal{H}}_1\,U_l & \cdots & \bar{\mathcal{H}}_{N-1}\,U_l & \bar{\mathcal{L}}\,U_l \end{bmatrix}.$$

La première étape de la méthode se termine avec

$$X_1 = \begin{pmatrix} U_l \\ x_1 \\ \vdots \\ x_{N-1} \\ T \\ V \end{pmatrix} \in \mathbb{R}^{N+2} \text{ et } Y_1 = \begin{pmatrix} x_N \\ y \\ y_{\text{inerte}} \end{pmatrix} \in \mathbb{R}^{N+2},$$

$$E_1(X_1) = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ x_1 & U_l & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{N-1} & 0 & \cdots & U_l & 0 & 0 \\ x_N & \bar{\mathcal{K}}_1\,U_l & \cdots & \bar{\mathcal{K}}_{N-1}\,U_l & \bar{\mathcal{D}}_N^{-1}\,(\nabla_T K(T,P,x)\cdot x)\,U_l & 0 \\ h & \bar{\mathcal{H}}_1\,U_l & \cdots & \bar{\mathcal{H}}_{N-1}\,U_l & \bar{\mathcal{L}}\,U_l & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}$$

et

$$f_1(X_1) = \begin{pmatrix} -V \\ -V\,K_1(T,P,x)\,x_1 \\ \vdots \\ -V\,K_{N-1}(T,P,x)\,x_{N-1} \\ -V\,K_N(T,P,x)\,x_N \\ Q - V\,H \end{pmatrix} \in \mathbb{R}^{N+2}.$$

La matrice E_1 n'est pas inversible, le processus de déflation se poursuit.

On considère maintenant le problème quasi-linéaire $E_1(X_1)\,\dot{X}_1 = f_1(X_1)$. Procédons en premier lieu à la transformation de la matrice coefficient $E_1(X_1)$ en appliquant la décomposition

LU à cette dernière. Le problème précédent est alors équivalent au système $U_1(X_1)\,\dot{X}_1 = L_1(X_1)^{-1}\,f_1(X_1)$. En gardant les mêmes notations ($E_1(X_1)$ et $f_1(X_1)$), on a

$$E_1(X_1) = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & U_l & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & U_l & 0 & 0 \\ 0 & 0 & \cdots & 0 & \bar{\mathcal{D}}_N^{-1}\,(\nabla_T K(T,P,x)\cdot x)\,U_l & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}$$

et

$$f_1(X_1) = \begin{pmatrix} -V \\ V\,x_1\,(1-K_1(T,P,x)) \\ \vdots \\ V\,x_{N-1}\,(1-K_{N-1}(T,P,x)) \\ -V\sum_{i=1}^{N}\bar{\mathcal{K}}_i\,x_i\,(1-K_i(T,P,x)) \\ Q-\bar{\mathcal{G}}\,V \end{pmatrix} \in \mathbb{R}^{N+2}.$$

Puisque $\bar{\mathcal{G}} \neq 0$ par hypothèse, il est possible d'extraire une expression de V ; il existe une fonction $\varphi_2 : \mathcal{I}_1 \subset \mathbb{R}^{N+1} \to \mathbb{R}$ telle que

$$V = \varphi_2\begin{pmatrix} U_l \\ x_1 \\ \vdots \\ x_{N-1} \\ T \end{pmatrix} = \bar{\mathcal{G}}^{-1}\,Q.$$

Ainsi, il n'est pas nécessaire de permuter les inconnues de X_1 : $X_2 = \begin{pmatrix} U_l & x_1 & \dots & x_{N-1} & T \end{pmatrix}^\top$ et $Y_2 = V = \varphi_2(X_2)$. Comme la matrice $B_1(X_2, Y_2)$ est nulle, on a

$$\begin{aligned} E_2(X_2) &= A_1(X_2, \varphi_2(X_2)) \\ &= \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & U_l & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & U_l & 0 \\ 0 & 0 & \cdots & 0 & \bar{\mathcal{D}}_N^{-1}\,(\nabla_T K(T,P,x)\cdot x)\,U_l \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}. \end{aligned}$$

La méthode de déflation se termine puisque la matrice précédente est inversible. On obtient l'EDO

$$\begin{pmatrix} \dot{U}_l \\ \dot{x}_1 \\ \vdots \\ \dot{x}_{N-1} \\ \dot{T} \end{pmatrix} = \begin{pmatrix} -\bar{\mathcal{G}}^{-1}\,Q \\ U_l^{-1}\,\bar{\mathcal{G}}^{-1}\,Q\,x_1\,(1-K_1(T,P,x)) \\ \vdots \\ U_l^{-1}\,\bar{\mathcal{G}}^{-1}\,Q\,x_{N-1}\,(1-K_1(T,P,x)) \\ U_l^{-1}\,(\nabla_T K(T,P,x)\cdot x)^{-1}\,\bar{\mathcal{G}}^{-1}\,Q\sum_{i=1}^{N}\bar{\mathcal{D}}_i\,(1-K_i(T,P,x))\,x_i \end{pmatrix},$$

avec

$$\begin{pmatrix} x_N \\ y_1 \\ \vdots \\ y_{N-1} \\ y_N \\ y_{\text{inerte}} \end{pmatrix} = \begin{pmatrix} \phi_1 \\ K_1(T,P,x)\, x_1 \\ \vdots \\ K_{N-1}(T,P,x)\, x_{N-1} \\ K_N(T,P,x)\, \phi_1 \\ 0 \end{pmatrix} \text{ et } V = \bar{\mathcal{G}}^{-1}\, Q.$$

Deux étapes du processus de déflation sont nécessaires pour résoudre le problème ; ainsi nous sommes en présence d'une EDA d'indice 2. □

Remarque 23

- *Ce modèle, pour lequel la phase liquide est considérée comme non idéale, généralise le cas idéal ; si les K_i ne dépendent plus de x, les expressions $\bar{\mathcal{D}}_i$, $\bar{\mathcal{K}}_i$, $\bar{\mathcal{H}}_i$, $\bar{\mathcal{L}}$ et $\bar{\mathcal{G}}$ deviennent $\mathcal{D}_i$, $\mathcal{K}_i$, $\mathcal{H}_i$, $\mathcal{L}$ et $\mathcal{G}$.*
- *À partir de l'expression (3.42), la dépendance des K_i passe aussi par la fonction ϕ_1 ; $K_i(T,P,x) = K_i(T,P,x_1,\ldots,x_{N-1},\phi_1(T,x_1,\ldots,x_{N-1}))$.*

3.3.3 Transition entre les régimes monophasique et diphasique

Les modèles des sections 3.3.1 et 3.3.2 ne traitent pas de la transition entre le régime monophasique (avant l'ébulition) et le régime diphasique (pendant l'ébullition) qui est nommée point de bulle. Ce point de bulle correspond au cas où V et y_{inerte} sont nuls simultanément. Contrairement aux configurations étudiées dans les sections 3.3.1 et 3.3.2, la méthode de déflation ne peut s'appliquer si les variables V et y_{inerte} sont toutes les deux nulles. En effet, la matrice Jacobienne $J_{X_0}(\bar{f}_0)$ ne peut plus être de rang plein :

$$J_{X_0}(\bar{f}_0) = \begin{bmatrix} 0 & -K_1(T,P) & \cdots & 0 & -\dfrac{\partial K_1(T,P)}{\partial T}x_1 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -K_N(T,P) & -\dfrac{\partial K_N(T,P)}{\partial T}x_N & 0 & \cdots & 1 & 0 & 0 \\ 0 & 1 & \cdots & 1 & 0 & -1 & \cdots & -1 & 0 & -1 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}$$
$$\in \mathbb{R}^{(N+2)\times(2N+4)}.$$

Il est ainsi impossible d'extraire, via des permutations de colonnes, un bloc matriciel inversible de taille $N+2$. Les conditions requises pour appliquer le théorème des fonctions implicites ne sont alors pas réunies ; la méthode de déflation reste muette. Cette dernière ne s'applique que si l'une des deux variables V ou y_{inerte} est différente de 0.

3.4 Distillation de Rayleigh réactive

3.4.1 Réactions chimiques contrôlées par la cinétique

Dans toute la suite de ce chapitre, nous nous concentrons sur le régime diphasique (la période d'ébullition), où rappelons-le, $V \neq 0$ et $y_{\text{inerte}} = 0$. Puisque nous ne regardons que cette configuration, nous ne faisons plus apparaître la variable y_{inerte} dans les systèmes que nous considérons. Étudions à présent un modèle de distillation de Rayleigh réactive en régime diphasique, issu de [57].

Dans cette modélisation, les réactions chimiques sont dites contrôlées par la cinétique.

3.4.1.1 Phases liquide et vapeur idéales

En faisant appel aux notations recensées en (3.2), (3.4), (3.6), (3.8), (3.10), (3.12), (3.13) et (3.14), on considère le modèle suivant :

$$\begin{cases} \dot{U}_l = \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{R} \nu_{i,j}\,\Delta_j - V \\ \dot{\overline{x_i\,U_l}} = \displaystyle\sum_{j=1}^{R} \nu_{i,j}\,\Delta_j - V\,y_i, \quad \forall i \in [\![1,N]\!] \\ \dot{\overline{h\,U_l}} = Q - V\,H - \displaystyle\sum_{j=1}^{R} Q_{r_j}\,\Delta_j \\ 0 = y_i - K_i(T,P)\,x_i, \quad \forall i \in [\![1,N]\!] \\ 0 = \displaystyle\sum_{i=1}^{N}(x_i - y_i). \end{cases} \tag{3.44}$$

Théorème 25

Supposons que $U_l \neq 0$, $\dfrac{\partial h}{\partial T} \neq 0$, $\mathcal{D}_N \neq 0$, $\nabla_T K(T,P)\cdot x \neq 0$ *et* $\mathcal{G} \neq 0$. *Alors l'EDA* (3.44) *est équivalente au système différentiel*

$$\begin{pmatrix} \dot{U}_l \\ \dot{x}_1 \\ \vdots \\ \dot{x}_{N-1} \\ \dot{T} \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{R} \nu_{i,j}\,\Delta_j - \mathcal{G}^{-1}\,\mathcal{Q} \\ U_l^{-1}\left(\mathcal{G}^{-1}\,\mathcal{Q}\,x_1\,\mathcal{D}_1 + \mathcal{M}_1\right) \\ \vdots \\ U_l^{-1}\left(\mathcal{G}^{-1}\,\mathcal{Q}\,x_{N-1}\,\mathcal{D}_{N-1} + \mathcal{M}_{N-1}\right) \\ U_l^{-1}\left(\nabla_T K(T,P)\cdot x\right)^{-1}\displaystyle\sum_{i=1}^{N}\left(\mathcal{G}^{-1}\,\mathcal{Q}\,\mathcal{D}_i\,x_i + \mathcal{M}_i\right)\,\mathcal{D}_i \end{pmatrix},$$

avec

$$\begin{pmatrix} x_N \\ y_1 \\ \vdots \\ y_{N-1} \\ y_N \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{i=1}^{N-1} \mathcal{K}_i\,x_i \\ K_1(T,P)\,x_1 \\ \vdots \\ K_{N-1}(T,P)\,x_{N-1} \\ K_N(T,P)\displaystyle\sum_{i=1}^{N-1} \mathcal{K}_i\,x_i \end{pmatrix}$$

et

$$V = \mathcal{G}^{-1}\mathcal{Q}.$$

Dans ces conditions, le problème (3.44) *est une EDA d'indice 2.*

Preuve - On écrit le système (3.44) sous la forme matricielle (3.1), avec

$$X_0 = \begin{pmatrix} U_l \\ x \\ T \\ y \\ V \end{pmatrix} \in \mathbb{R}^{2N+3}, \quad f_0(X_0) = \begin{pmatrix} \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{R} \nu_{i,j}\,\Delta_j - V \\ \displaystyle\sum_{j=1}^{R} \nu_{1,j}\,\Delta_j - V\,y_1 \\ \vdots \\ \displaystyle\sum_{j=1}^{R} \nu_{N,j}\,\Delta_j - V\,y_N \\ Q - V\,H - \displaystyle\sum_{j=1}^{R} Q_{r_j}\,\Delta_j \\ y_1 - K_1(T,P)\,x_1 \\ \vdots \\ y_N - K_N(T,P)\,x_N \\ \displaystyle\sum_{i=1}^{N}(x_i - y_i) \end{pmatrix} \in \mathbb{R}^{2N+3}$$

et

$$E_0(X_0) = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ x_1 & U_l & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots \\ x_N & 0 & \cdots & U_l & 0 & 0 & \cdots & 0 \\ h & \dfrac{\partial h}{\partial x_1}U_l & \cdots & \dfrac{\partial h}{\partial x_N}U_l & \dfrac{\partial h}{\partial T}U_l & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{(2N+3)\times(2N+3)}.$$

Puisque $U_l \neq 0$ et $\dfrac{\partial h}{\partial T} \neq 0$, le rang de la matrice $E_0(X_0)$ vaut $N+2$. Ainsi,

$$\tilde{E}_0(X_0) = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ x_1 & U_l & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_N & 0 & \cdots & U_l & 0 \\ h & \dfrac{\partial h}{\partial x_1}U_l & \cdots & \dfrac{\partial h}{\partial x_N}U_l & \dfrac{\partial h}{\partial T}U_l \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}$$

et

$$\tilde{f}_0(X_0) = \begin{pmatrix} y_1 - K_1(T,P)\,x_1 \\ \vdots \\ y_N - K_N(T,P)\,x_N \\ \displaystyle\sum_{i=1}^{N}(x_i - y_i) \end{pmatrix} \in \mathbb{R}^{N+2}. \tag{3.45}$$

De plus, on a

$$g_0(X_0) = \begin{pmatrix} \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{R} \nu_{i,j}\,\Delta_j - V \\ \displaystyle\sum_{j=1}^{R} \nu_{1,j}\,\Delta_j - V\,y_1 \\ \vdots \\ \displaystyle\sum_{j=1}^{R} \nu_{N,j}\,\Delta_j - V\,y_N \\ Q - V\,H - \displaystyle\sum_{j=1}^{R} Q_{r_j}\,\Delta_j \end{pmatrix} \in \mathbb{R}^{N+2}.$$

À ce stade, nous sommes dans l'obligation de permuter les composantes du vecteur X_0 car la variable V n'apparait pas dans la partie algébrique (3.45). Soit $\mathcal{P}_1 \in \mathbb{R}^{(2N+3)\times(2N+3)}$ la matrice de permutation telle que

$$\mathcal{P}_1 X_0 = \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \text{ où } X_1 = \begin{pmatrix} U_l \\ x_1 \\ \vdots \\ x_{N-1} \\ T \\ V \end{pmatrix} \in \mathbb{R}^{N+2} \text{ et } Y_1 = \begin{pmatrix} x_N \\ y \end{pmatrix} \in \mathbb{R}^{N+1}.$$

D'une part, par dérivation de (3.45) par rapport à Y_1, on a

$$J_{Y_1}\left(\bar{f}_0\right) = \begin{bmatrix} \gamma_1 & I_N \\ 1 & \gamma_4 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}, \tag{3.46}$$

avec

$$\gamma_1 = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -K_N(T,P) \end{pmatrix} \in \mathbb{R}^N \text{ et } \gamma_4 = \begin{pmatrix} -1 & \cdots & -1 \end{pmatrix} \in \mathbb{R}^{1\times N}.$$

On peut aisément établir que $\det J_{Y_1}\left(\bar{f}_0\right) = (-1)^N\,\mathcal{D}_N$. Le théorème des fonctions implicites s'applique donc au système algébrique $\tilde{f}_0(X_0) = 0$, où $\tilde{f}_0(X_0)$ est définie par (3.45). Plus précisément, il existe une fonction $\varphi_1 : \mathcal{I}_0 \subset \mathbb{R}^{N+2} \to \mathbb{R}^{N+1}$ telle que

$$Y_1 = \varphi_1(X_1) \Leftrightarrow \begin{pmatrix} x_N \\ y_1 \\ \vdots \\ y_{N-1} \\ y_N \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{i=1}^{N-1} \mathcal{K}_i\,x_i \\ K_1(T,P)\,x_1 \\ \vdots \\ K_{N-1}(T,P)\,x_{N-1} \\ K_N(T,P)\displaystyle\sum_{i=1}^{N-1} \mathcal{K}_i\,x_i \end{pmatrix}. \tag{3.47}$$

D'autre part, par dérivation de (3.45) par rapport à X_1, on obtient

$$J_{X_1}\left(\bar{f}_0\right) = \begin{bmatrix} 0 & \delta_2 \\ 0 & \delta_4 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+2)}, \tag{3.48}$$

où

$$\delta_2 = \begin{bmatrix} -K_1(T,P) & \cdots & 0 & -\dfrac{\partial K_1(T,P)}{\partial T}\, x_1 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & -K_{N-1}(T,P) & -\dfrac{\partial K_{N-1}(T,P)}{\partial T}\, x_{N-1} & 0 \\ 0 & \cdots & 0 & -\dfrac{\partial K_N(T,P)}{\partial T}\, x_N & 0 \end{bmatrix} \in \mathbb{R}^{N\times(N+1)}$$

et

$$\delta_4 = \begin{pmatrix} 1 & \cdots & 1 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{1\times(N+1)}.$$

Par ailleurs, on a

$$A_0(X_1,Y_1) = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}, \tag{3.49}$$

avec

$$\alpha_1 = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix} \in \mathbb{R}^N,\ \alpha_2 = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 \\ U_l & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & U_l & 0 & 0 \end{bmatrix} \in \mathbb{R}^{N\times(N+1)},$$

$$\alpha_3 = \begin{pmatrix} x_N \\ h \end{pmatrix} \in \mathbb{R}^2 \text{ et } \alpha_4 = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 \\ \dfrac{\partial h}{\partial x_1} U_l & \cdots & \dfrac{\partial h}{\partial x_{N-1}} U_l & \dfrac{\partial h}{\partial T} U_l & 0 \end{bmatrix} \in \mathbb{R}^{2\times(N+1)},$$

ainsi que

$$B_0(X_1,Y_1) = \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+1)}, \tag{3.50}$$

où

$$\beta_3 = \begin{pmatrix} U_l \\ \dfrac{\partial h}{\partial x_N} U_l \end{pmatrix} \in \mathbb{R}^{2\times(N+1)}.$$

Déterminons à présent la matrice $E_1(X_1)$ à partir des matrices (3.49), (3.50), (3.46) et (3.48) :

$$\begin{aligned} E_1(X_1) &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} \gamma_1 & I_N \\ 1 & \gamma_4 \end{bmatrix}^{-1} \begin{bmatrix} 0 & \delta_2 \\ 0 & \delta_4 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} \bar\gamma_1 & \bar\gamma_2 \\ \bar\gamma_3 & \bar\gamma_4 \end{bmatrix} \begin{bmatrix} 0 & \delta_2 \\ 0 & \delta_4 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} 0 & \bar\gamma_1\,\delta_2 + \bar\gamma_2\,\delta_4 \\ 0 & \bar\gamma_3\,\delta_2 + \bar\gamma_4\,\delta_4 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & \beta_3\,(\bar\gamma_1\,\delta_2 + \bar\gamma_2\,\delta_4) \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \bar\alpha_4 \end{bmatrix}, \end{aligned}$$

où $\bar\alpha_4 = \alpha_4 - \beta_3\,(\bar\gamma_1\,\delta_2 + \bar\gamma_2\,\delta_4)$. Pour calculer ce terme, on commence par déterminer les expressions $\bar\gamma_1$ et $\bar\gamma_2$ à l'aide du complément de Schur :

$$\bar\gamma_1 = -\,(1-\gamma_4\,\gamma_1)^{-1}\,\gamma_4 \in \mathbb{R}^{1\times(N+1)} \text{ et } \bar\gamma_2 = (1-\gamma_4\,\gamma_1)^{-1} \in \mathbb{R}.$$

On parvient aisément à simplifier l'expression de $\bar{\alpha}_3$:

$$\bar{\alpha}_4 = \alpha_4 - \beta_3 \, (1 - \gamma_4 \, \gamma_1)^{-1} \, (\delta_4 - \gamma_4 \, \delta_2).$$

Déterminons à présent les expressions de chaque terme :

$$(1 - \gamma_4 \, \gamma_1)^{-1} = \mathcal{D}_N^{-1} \in \mathbb{R}$$

et

$$\begin{aligned} \delta_4 - \gamma_4 \, \delta_2 &= \begin{pmatrix} 1 & \cdots & 1 & 0 & 0 \end{pmatrix} - \begin{pmatrix} K_1(T,P) & \cdots & K_{N-1}(T,P) & \nabla_T K(T,P) \cdot x & 0 \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{D}_1 & \cdots & \mathcal{D}_{N-1} & -(\nabla_T K(T,P) \cdot x) & 0 \end{pmatrix} \in \mathbb{R}^{1 \times (N+1)}. \end{aligned}$$

On établit ainsi

$$\bar{\alpha}_4 = \begin{bmatrix} \mathcal{K}_1 \, U_l & \cdots & \mathcal{K}_{N-1} \, U_l & \mathcal{D}_N^{-1} \, (\nabla_T K(T,P) \cdot x) \, U_l & 0 \\ \mathcal{H}_1 \, U_l & \cdots & \mathcal{H}_{N-1} \, U_l & \mathcal{L} \, U_l & 0 \end{bmatrix}.$$

La première étape de la méthode de déflation s'achève en fournissant

$$E_1(X_1) = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ x_1 & U_l & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{N-1} & 0 & \cdots & U_l & 0 & 0 \\ x_N & \mathcal{K}_1 \, U_l & \cdots & \mathcal{K}_{N-1} \, U_l & \mathcal{D}_N^{-1} \, (\nabla_T K(T,P) \cdot x) \, U_l & 0 \\ h & \mathcal{H}_1 \, U_l & \cdots & \mathcal{H}_{N-1} \, U_l & \mathcal{L} \, U_l & 0 \end{bmatrix} \in \mathbb{R}^{(N+2) \times (N+2)},$$

$$X_1 = \begin{pmatrix} U_l \\ x_1 \\ \vdots \\ x_{N-1} \\ T \\ V \end{pmatrix} \in \mathbb{R}^{N+2}$$

et

$$f_1(X_1) = \begin{pmatrix} \displaystyle\sum_{i=1}^{N} \sum_{j=1}^{R} \nu_{i,j} \, \Delta_j - V \\ \displaystyle\sum_{j=1}^{R} \nu_{1,j} \, \Delta_j - V \, K_1(T,P) \, x_1 \\ \vdots \\ \displaystyle\sum_{j=1}^{R} \nu_{N,j} \, \Delta_j - V \, K_N(T,P) \, x_N \\ \displaystyle Q - V \, H - \sum_{j=1}^{R} Q_{r_j} \, \Delta_j \end{pmatrix} \in \mathbb{R}^{N+2}.$$

Comme la matrice $E_1(X_1)$ n'est pas inversible, la méthode de déflation se poursuit.

On considère à présent le problème quasi-linéaire $E_1(X_1) \, \dot{X}_1 = f_1(X_1)$. Transformons la matrice coefficient $E_1(X_1)$ via une décomposition LU. Le problème précédent est alors équivalent

au système $U_1(X_1)\,\dot{X}_1 = L_1(X_1)^{-1}\,f_1(X_1)$. En gardant les mêmes notations ($E_1(X_1)$ et $f_1(X_1)$), on a

$$E_1(X_1) = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & U_l & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & U_l & 0 & 0 \\ 0 & 0 & \cdots & 0 & \mathcal{D}_N^{-1}\,(\nabla_T K(T,P)\cdot x)\,U_l & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}$$

et

$$f_1(X_1) = \begin{pmatrix} \sum_{i=1}^{N}\sum_{j=1}^{R} \nu_{i,j}\,\Delta_j - V \\ V\,x_1\,\mathcal{D}_1 + \mathcal{M}_1 \\ \vdots \\ V\,x_{N-1}\,\mathcal{D}_{N-1} + \mathcal{M}_{N-1} \\ -\sum_{i=1}^{N} (V\,x_i\,\mathcal{D}_i + \mathcal{M}_i)\,\mathcal{K}_i \\ \mathcal{Q} - \mathcal{G}\,V \end{pmatrix}.$$

Puisque $\mathcal{G} \neq 0$ par hypothèse, il est possible d'extraire une expression de V ; il existe une fonction $\varphi_2 : \mathcal{I}_1 \subset \mathbb{R}^{N+1} \to \mathbb{R}$ telle que

$$V = \varphi_2 \begin{pmatrix} U_l \\ x_1 \\ \vdots \\ x_{N-1} \\ T \end{pmatrix} = \mathcal{G}^{-1}\,\mathcal{Q}.$$

Il n'est par conséquent pas nécessaire de permuter les inconnues de X_1 (*i.e.* la matrice de permutation est la matrice identité) :

$$X_2 = \begin{pmatrix} U_l \\ x_1 \\ \vdots \\ x_{N-1} \\ T \end{pmatrix} \text{ et } Y_2 = V = \varphi_2(X_2).$$

Comme la matrice $B_1(X_2, Y_2)$ est nulle, on a directement

$$\begin{aligned} E_2(X_2) &= A_1(X_2, \varphi_2(X_2)) \\ &= \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & U_l & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & U_l & 0 \\ 0 & 0 & \cdots & 0 & \mathcal{D}_N^{-1}\,(\nabla_T K(T,P)\cdot x)\,U_l \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}. \end{aligned}$$

Cette matrice est inversible par hypothèse. La deuxième étape de la méthode de déflation clôt le processus itératif et donne l'EDO sous contraintes

$$\begin{pmatrix} \dot{U}_l \\ \dot{x}_1 \\ \vdots \\ \dot{x}_{N-1} \\ \dot{T} \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{R} \nu_{i,j}\Delta_j - \mathcal{G}^{-1}\mathcal{Q} \\ U_l^{-1}\left(\mathcal{G}^{-1}\mathcal{Q}x_1\mathcal{D}_1 + \mathcal{M}_1\right) \\ \vdots \\ U_l^{-1}\left(\mathcal{G}^{-1}\mathcal{Q}x_{N-1}\mathcal{D}_{N-1} + \mathcal{M}_{N-1}\right) \\ U_l^{-1}\left(\nabla_T K\cdot x\right)^{-1}\displaystyle\sum_{i=1}^{N}\left(\mathcal{G}^{-1}\mathcal{Q}\mathcal{D}_i x_i + \mathcal{M}_i\right)\mathcal{D}_i \end{pmatrix}, \begin{pmatrix} x_N \\ y_1 \\ \vdots \\ y_{N-1} \\ y_N \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{i=1}^{N-1}\mathcal{K}_i x_i \\ K_1 x_1 \\ \vdots \\ K_{N-1}x_{N-1} \\ \displaystyle K_N\sum_{i=1}^{N-1}\mathcal{K}_i x_i \end{pmatrix}$$

et

$$V = \mathcal{G}^{-1}\mathcal{Q}.$$

La méthode de déflation est achevée en deux étapes ; l'EDA considérée est d'indice 2. □

Remarque 24

- *La structure de ce problème réactif est assez similaire à celle du modèle non réactif étudié précédemment. Une des raisons de cette similitude réside dans le caractère non différentiel des termes régissant l'évolution des réactions chimiques. De plus, le nombre de réactions chimiques R n'influe pas sur le nombre d'équations du système.*
- *Si on supprime les termes Δ_j, c'est-à-dire si on supprime les réactions chimiques, $\mathcal{Q}$ devient simplement Q et les termes $\mathcal{M}_i$ disparaissent. Par ailleurs, sans les réactions chimiques, on retrouve les résultats du modèle non réactif en régime diphasique ($y_{\text{inerte}} = 0$).*

3.4.1.2 Phase liquide non idéale et phase vapeur idéale

Les constantes d'équilibres thermodynamiques dépendent désormais de T, de P et de x. On considère le modèle suivant en tenant compte des notations définies par (3.3), (3.5), (3.7), (3.9), (3.11), (3.12), (3.13) et (3.15).

$$\begin{cases} \dot{U}_l = \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{R}\nu_{i,j}\,\Delta_j - V \\ \dot{\overline{x_i\,U_l}} = \displaystyle\sum_{j=1}^{R}\nu_{i,j}\,\Delta_j - V\,y_i, \quad \forall i \in [\![1,N]\!] \\ \dot{\overline{h\,U_l}} = Q - V\,H - \displaystyle\sum_{j=1}^{R} Q_{r_j}\,\Delta_j \\ 0 = y_i - K_i(T,P,x)\,x_i, \quad \forall i \in [\![1,N]\!] \\ 0 = \displaystyle\sum_{i=1}^{N}(x_i - y_i). \end{cases} \tag{3.51}$$

Théorème 26

Supposons que $U_l \neq 0$, $\dfrac{\partial h}{\partial T} \neq 0$, $\bar{\mathcal{D}}_N \neq 0$, $\nabla_T K(T,P,x)\cdot x \neq 0$ et $\bar{\mathcal{G}} \neq 0$. Alors il existe une

fonction $\phi_1 : \mathcal{I} \subset \mathbb{R}^N \to \mathbb{R}$ *telle que l'EDA* (3.44) *soit équivalente au système différentiel*

$$\begin{pmatrix} \dot{U}_l \\ \dot{x}_1 \\ \vdots \\ \dot{x}_{N-1} \\ \dot{T} \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{R} \nu_{i,j}\Delta_j - \bar{\mathcal{G}}^{-1}\bar{\mathcal{Q}} \\ U_l^{-1}\left(\bar{\mathcal{G}}^{-1}\bar{\mathcal{Q}}x_1\left(1 - K_1(T,P,x)\right) + \mathcal{M}_1\right) \\ \vdots \\ U_l^{-1}\left(\bar{\mathcal{G}}^{-1}\bar{\mathcal{Q}}x_{N-1}\left(1 - K_{N-1}(T,P,x)\right) + \mathcal{M}_{N-1}\right) \\ U_l^{-1}\left(\nabla_T K(T,P,x)\cdot x\right)^{-1}\displaystyle\sum_{i=1}^{N}\left(\bar{\mathcal{G}}^{-1}\bar{\mathcal{Q}}\left(1 - K_i(T,P,x)\right)x_i + \mathcal{M}_i\right)\bar{\mathcal{D}}_i \end{pmatrix},$$

avec

$$\begin{pmatrix} x_N \\ y_1 \\ \vdots \\ y_{N-1} \\ y_N \end{pmatrix} = \begin{pmatrix} \phi_1(T,x_1,\ldots,x_{N-1}) \\ K_1(T,P,x)\,x_1 \\ \vdots \\ K_{N-1}(T,P,x)\,x_{N-1} \\ K_N(T,P,x)\,\phi_1(T,x_1,\ldots,x_{N-1}) \end{pmatrix} \text{ et } V = \bar{\mathcal{G}}^{-1}\,\bar{\mathcal{Q}}.$$

Dans ces conditions, le problème (3.51) *est une EDA d'indice 2.*

Preuve - Jusqu'à la détermination de la matrice Jacobienne $J_{Y_1}\left(\bar{f}_0\right)$, le mécanisme de déflation est identique. Dans ce contexte non idéal, on a :

$$\tilde{f}_0(X_0) = \begin{pmatrix} y_1 - K_1(T,P,x)\,x_1 \\ \vdots \\ y_N - K_N(T,P,x)\,x_N \\ \displaystyle\sum_{i=1}^{N}(x_i - y_i) \end{pmatrix} \in \mathbb{R}^{N+2}. \tag{3.52}$$

Par dérivation de (3.52) par rapport à Y_1, on a

$$J_{Y_1}\left(\bar{f}_0\right) = \begin{bmatrix} \gamma_1 & I_N \\ 1 & \gamma_4 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}, \tag{3.53}$$

avec

$$\gamma_1 = \begin{pmatrix} -\dfrac{\partial K_1(T,P,x)}{\partial x_N}\,x_1 \\ \vdots \\ -\dfrac{\partial K_{N-1}(T,P,x)}{\partial x_N}\,x_{N-1} \\ -\dfrac{\partial K_N(T,P,x)}{\partial x_N}\,x_N - K_N(T,P,x) \end{pmatrix} \in \mathbb{R}^N \text{ et } \gamma_4 = \begin{pmatrix} -1 & \cdots & -1 \end{pmatrix} \in \mathbb{R}^{1\times N}.$$

On a naturellement $\det J_{Y_1}\left(\bar{f}_0\right) = (-1)^N\,\bar{\mathcal{D}}_N$. Le théorème des fonctions implicites s'applique donc au système algébrique $\tilde{f}_0(X_0) = 0$, où $\tilde{f}_0(X_0)$ est définie par (3.45). Plus précisément, il existe deux fonctions $\varphi_1 : \mathcal{I}_0 \subset \mathbb{R}^{N+2} \to \mathbb{R}^{N+1}$ et $\phi_1 : \mathcal{I} \subset \mathbb{R}^N \to \mathbb{R}$ telles que

$$Y_1 = \varphi_1(X_1) \Leftrightarrow \begin{pmatrix} x_N \\ y_1 \\ \vdots \\ y_{N-1} \\ y_N \end{pmatrix} = \begin{pmatrix} \phi_1(T,x_1,\ldots,x_{N-1}) \\ K_1(T,P,x)\,x_1 \\ \vdots \\ K_{N-1}(T,P,x)\,x_{N-1} \\ K_N(T,P,x)\,\phi_1(T,x_1,\ldots,x_{N-1}) \end{pmatrix}. \tag{3.54}$$

De plus, par dérivation de (3.52) par rapport à X_1, on obtient

$$J_{X_1}\left(\bar{f}_0\right) = \begin{bmatrix} 0 & \delta_2 \\ 0 & \delta_4 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+2)}, \tag{3.55}$$

où

$$\delta_2 = \begin{bmatrix} -\dfrac{\partial K_1}{\partial x_1}\, x_1 - K_1 & \cdots & -\dfrac{\partial K_1}{\partial x_{N-1}}\, x_1 & -\dfrac{\partial K_1}{\partial T} x_1 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ -\dfrac{\partial K_{N-1}}{\partial x_1}\, x_{N-1} & \cdots & -\dfrac{\partial K_N}{\partial x_{N-1}}\, x_{N-1} - K_{N-1} & -\dfrac{\partial K_{N-1}}{\partial T} x_{N-1} & 0 \\ -\dfrac{\partial K_N}{\partial x_1}\, x_N & \cdots & -\dfrac{\partial K_N}{\partial x_{N-1}}\, x_N & -\dfrac{\partial K_N}{\partial T} x_N & 0 \end{bmatrix} \in \mathbb{R}^{N\times(N+1)},$$

avec $K_i = K_i(T, P, x)$ et

$$\delta_4 = \begin{pmatrix} 1 & \cdots & 1 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{1\times(N+1)}.$$

Déterminons à présent la matrice $E_1(X_1)$ à partir des matrices (3.49), (3.50), (3.53) et (3.55) :

$$\begin{aligned} E_1(X_1) &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} \gamma_1 & I_N \\ 1 & \gamma_4 \end{bmatrix}^{-1} \begin{bmatrix} 0 & \delta_2 \\ 0 & \delta_4 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} \bar{\gamma}_1 & \bar{\gamma}_2 \\ \bar{\gamma}_3 & \bar{\gamma}_4 \end{bmatrix} \begin{bmatrix} 0 & \delta_2 \\ 0 & \delta_4 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} 0 & \bar{\gamma}_1\,\delta_2 + \bar{\gamma}_2\,\delta_4 \\ 0 & \bar{\gamma}_3\,\delta_2 + \bar{\gamma}_4\,\delta_4 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & \beta_3\left(\bar{\gamma}_1\,\delta_2 + \bar{\gamma}_2\,\delta_4\right) \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \bar{\alpha}_4 \end{bmatrix}, \end{aligned}$$

où $\bar{\alpha}_4 = \alpha_4 - \beta_3\left(\bar{\gamma}_1\,\delta_2 + \bar{\gamma}_2\,\delta_4\right)$. On a :

$$\bar{\alpha}_4 = \alpha_4 - \beta_3\left(1 - \gamma_4\,\gamma_1\right)^{-1}\left(\delta_4 - \gamma_4\,\delta_2\right).$$

Déterminons à présent les expressions de chaque terme :

$$\left(1 - \gamma_4\,\gamma_1\right)^{-1} = \bar{\mathcal{D}}_N^{-1} \in \mathbb{R}$$

et

$$\delta_4 - \gamma_4\,\delta_2 = \begin{pmatrix} \bar{\mathcal{D}}_1 & \cdots & \bar{\mathcal{D}}_{N-1} & -\left(\nabla_T K(T,P,x)\cdot x\right) & 0 \end{pmatrix} \in \mathbb{R}^{1\times(N+1)}.$$

On établit ainsi

$$\bar{\alpha}_4 = \begin{bmatrix} \bar{\mathcal{K}}_1\,U_l & \cdots & \bar{\mathcal{K}}_{N-1}\,U_l & \bar{\mathcal{D}}_N^{-1}\left(\nabla_T K(T,P,x)\cdot x\right) U_l & 0 \\ \bar{\mathcal{H}}_1\,U_l & \cdots & \bar{\mathcal{H}}_{N-1}\,U_l & \bar{\mathcal{L}}\,U_l & 0 \end{bmatrix}.$$

La première étape de la méthode de déflation s'achève en fournissant

$$E_1(X_1) = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ x_1 & U_l & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{N-1} & 0 & \cdots & U_l & 0 & 0 \\ x_N & \bar{\mathcal{K}}_1\,U_l & \cdots & \bar{\mathcal{K}}_{N-1}\,U_l & \bar{\mathcal{D}}_N^{-1}\left(\nabla_T K(T,P,x)\cdot x\right) U_l & 0 \\ h & \bar{\mathcal{H}}_1\,U_l & \cdots & \bar{\mathcal{H}}_{N-1}\,U_l & \bar{\mathcal{L}}\,U_l & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)},$$

$$X_1 = \begin{pmatrix} U_l \\ x_1 \\ \vdots \\ x_{N-1} \\ T \\ V \end{pmatrix} \in \mathbb{R}^{N+2}$$

et

$$f_1(X_1) = \begin{pmatrix} \sum_{i=1}^{N}\sum_{j=1}^{R} \nu_{i,j}\,\Delta_j - V \\ \sum_{j=1}^{R} \nu_{1,j}\,\Delta_j - V\,K_1(T,P,x)\,x_1 \\ \vdots \\ \sum_{j=1}^{R} \nu_{N,j}\,\Delta_j - V\,K_N(T,P,x)\,x_N \\ Q - V\,H - \sum_{j=1}^{R} Q_{r_j}\,\Delta_j \end{pmatrix} \in \mathbb{R}^{N+2}.$$

Comme la matrice $E_1(X_1)$ n'est pas inversible, la méthode de déflation se poursuit.

On considère à présent le problème quasi-linéaire $E_1(X_1)\,\dot{X}_1 = f_1(X_1)$. Transformons la matrice coefficient $E_1(X_1)$ via une décomposition LU. Le problème précédent est alors équivalent au système $U_1(X_1)\,\dot{X}_1 = L_1(X_1)^{-1}\,f_1(X_1)$. En gardant les mêmes notations ($E_1(X_1)$ et $f_1(X_1)$), on a

$$E_1(X_1) = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & U_l & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & U_l & 0 & 0 \\ 0 & 0 & \cdots & 0 & \bar{\mathcal{D}}_N^{-1}\,(\nabla_T K(T,P,x)\,\cdot\,x)\,U_l & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}$$

et

$$f_1(X_1) = \begin{pmatrix} \sum_{i=1}^{N}\sum_{j=1}^{R} \nu_{i,j}\,\Delta_j - V \\ V\,x_1\,(1 - K_1(T,P,x)) + \mathcal{M}_1 \\ \vdots \\ V\,x_{N-1}\,(1 - K_{N-1}(T,P,x)) + \mathcal{M}_{N-1} \\ -\sum_{i=1}^{N} (V\,x_i\,(1 - K_i(T,P,x)) + \mathcal{M}_i)\,\bar{\mathcal{K}}_i \\ \bar{\mathcal{Q}} - \bar{\mathcal{G}}\,V \end{pmatrix}.$$

Puisque $\bar{\mathcal{G}} \neq 0$ par hypothèse, il est possible d'extraire une expression de V ; il existe une fonction $\varphi_2 : \mathcal{I}_1 \subset \mathbb{R}^{N+1} \to \mathbb{R}$ telle que

$$V = \varphi_2 \begin{pmatrix} U_l \\ x_1 \\ \vdots \\ x_{N-1} \\ T \end{pmatrix} = \bar{\mathcal{G}}^{-1}\,\bar{\mathcal{Q}}.$$

Il n'est par conséquent pas nécessaire de permuter les inconnues de X_1 (*i.e.* la matrice de permutation est la matrice identité) :

$$X_2 = \begin{pmatrix} U_l \\ x_1 \\ \vdots \\ x_{N-1} \\ T \end{pmatrix} \text{ et } Y_2 = V = \varphi_2(X_2).$$

Comme la matrice $B_1(X_2, Y_2)$ est nulle, on a directement

$$\begin{aligned} E_2(X_2) &= A_1(X_2, \varphi_2(X_2)) \\ &= \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & U_l & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & U_l & 0 \\ 0 & 0 & \cdots & 0 & \bar{\mathcal{D}}_N^{-1}\ (\nabla_T K(T,P,x)\ \cdot\ x)\ U_l \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}. \end{aligned}$$

Cette matrice est inversible par hypothèse. La deuxième étape de la méthode de déflation clôt le processus itératif et donne l'EDO

$$\begin{pmatrix} \dot{U}_l \\ \dot{x}_1 \\ \vdots \\ \dot{x}_{N-1} \\ \dot{T} \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{R} \nu_{i,j}\,\Delta_j - \bar{\mathcal{G}}^{-1}\,\bar{\mathcal{Q}} \\ U_l^{-1}\ \left(\bar{\mathcal{G}}^{-1}\,\bar{\mathcal{Q}}\,x_1\ (1 - K_1(T,P,x)) + \mathcal{M}_1\right) \\ \vdots \\ U_l^{-1}\ \left(\bar{\mathcal{G}}^{-1}\,\bar{\mathcal{Q}}\,x_{N-1}\ (1 - K_{N-1}(T,P,x)) + \mathcal{M}_{N-1}\right) \\ U_l^{-1}\ (\nabla_T K(T,P,x)\ \cdot\ x)^{-1} \displaystyle\sum_{i=1}^{N} \left(\bar{\mathcal{G}}^{-1}\,\bar{\mathcal{Q}}\ (1 - K_i(T,P,x))\ x_i + \mathcal{M}_i\right)\ \bar{\mathcal{D}}_i \end{pmatrix},$$

avec

$$\begin{pmatrix} x_N \\ y_1 \\ \vdots \\ y_{N-1} \\ y_N \end{pmatrix} = \begin{pmatrix} \phi_1(T, x_1, \ldots, x_{N-1}) \\ K_1(T,P,x)\,x_1 \\ \vdots \\ K_{N-1}(T,P,x)\,x_{N-1} \\ K_N(T,P,x)\,\phi_1(T, x_1, \ldots, x_{N-1}) \end{pmatrix}$$

et

$$V = \bar{\mathcal{G}}^{-1}\,\bar{\mathcal{Q}}.$$

La méthode de déflation est achevée en deux étapes ; l'EDA considérée est d'indice 2. □

3.4.2 Réactions chimiques instantanément équilibrées

Terminons l'étude par un type de modélisation axé sur les réactions chimiques instantanément équilibrées. Là encore, ce modèle de distillation de Rayleigh réactive est considéré en régime diphasique. Les réactions chimiques ont toujours lieu dans la phase liquide. On pose $N_R = N - R - 1$.

3.4.2.1 Phases liquide et vapeur idéales

On considère le modèle réactif suivant :

$$\begin{cases} \dot{U}_l - \displaystyle\sum_{j=1}^{R}\left(\sum_{i=1}^{N}\nu_{i,j}\right)\dot{\xi}_j = -V \\ \dot{\overline{x_i\,U_l}} - \displaystyle\sum_{j=1}^{R}\nu_{i,j}\,\dot{\xi}_j = -V\,y_i, \quad \forall i \in [\![1,N]\!] \\ \dot{\overline{h\,U_l}} + \displaystyle\sum_{j=1}^{R} Q_{r_j}\,\dot{\xi}_j = Q - V\,H \\ 0 = y_i - K_i(T,P)\,x_i, \quad \forall i \in [\![1,N]\!] \\ 0 = \displaystyle\sum_{i=1}^{N}(x_i - y_i) \\ 0 = \mathcal{A}_j, \quad \forall j \in [\![1,R]\!]. \end{cases} \tag{3.56}$$

On utilise les notations définies par (3.2), (3.16), (3.17), (3.19), (3.21), (3.23), (3.25), (3.27) et (3.29).

Théorème 27

Supposons que $U_l \neq 0$, $\dfrac{\partial h}{\partial T} \neq 0$, $\det\Gamma \neq 0$, $\det\mathcal{N}_a \neq 0$ *et* $H - \mathcal{N}_b\mathcal{N}_a^{-1}\begin{pmatrix}1\\ y\end{pmatrix} \neq 0$. *Alors il existe des fonctions* $\phi_j : \mathcal{I} \subset \mathbb{R}^{N-R} \to \mathbb{R}$ *pour tout* $j \in [\![1, R+1]\!]$ *telles que l'EDA* (3.56) *soit équivalente au système différentiel*

$$\begin{pmatrix} \dot{\xi} \\ \dot{U}_l \\ \dot{x}_1 \\ \vdots \\ \dot{x}_{N_R} \\ \dot{T} \end{pmatrix} = -\mathcal{N}_a^{-1}\left(H - \mathcal{N}_b\mathcal{N}_a^{-1}\begin{pmatrix}1\\ y\end{pmatrix}\right)^{-1} Q \begin{pmatrix}1\\ y\end{pmatrix},$$

avec

$$\begin{pmatrix} x_{N-R} \\ \vdots \\ x_N \\ y_1 \\ \vdots \\ y_{N_R} \\ y_{N-R} \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \phi_1(x_1,\dots,x_{N_R},T) \\ \vdots \\ \phi_{R+1}(x_1,\dots,x_{N_R},T) \\ K_1(T,P)\,x_1 \\ \vdots \\ K_{N_R}(T,P)\,x_{N_R} \\ K_{N-R}(T,P)\,\phi_1(x_1,\dots,x_{N_R},T) \\ \vdots \\ K_N(T,P)\,\phi_{R+1}(x_1,\dots,x_{N_R},T) \end{pmatrix} \quad \text{et } V = \left(H - \mathcal{N}_b\mathcal{N}_a^{-1}\begin{pmatrix}1\\ y\end{pmatrix}\right)^{-1} Q.$$

Dans ces conditions, le problème (3.56) *est une EDA d'indice 2.*

Preuve - On écrit le système (3.56) sous la forme matricielle (3.1), avec

$$X_0 = \begin{pmatrix} \xi \\ U_l \\ x \\ T \\ y \\ V \end{pmatrix} \in \mathbb{R}^{2N+R+3},$$

$$E_0(X_0) = \begin{bmatrix} \tilde{E}_0(X_0) & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{(2N+R+3)\times(2N+R+3)},$$

avec

$$\tilde{E}_0(X_0) = \begin{bmatrix} -\sum_{i=1}^{N}\nu_{i,1} & \cdots & -\sum_{i=1}^{N}\nu_{i,R} & 1 & 0 & \cdots & 0 & 0 \\ -\nu_{1,1} & \cdots & -\nu_{1,R} & x_1 & U_l & \cdots & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\nu_{N,1} & \cdots & -\nu_{N,R} & x_N & 0 & \cdots & U_l & 0 \\ Q_{r_1} & \cdots & Q_{r_R} & h & \dfrac{\partial h}{\partial x_1}U_l & \cdots & \dfrac{\partial h}{\partial x_N}U_l & \dfrac{\partial h}{\partial T}U_l \end{bmatrix}$$
$$\in \mathbb{R}^{(N+2)\times(N+R+2)}$$

et

$$f_0(X_0) = \begin{pmatrix} -V \\ -V\,y \\ Q - V\,H \\ y_1 - K_1(T,P)\,x_1 \\ \vdots \\ y_N - K_N(T,P)\,x_N \\ \sum_{i=1}^{N}(x_i - y_i) \\ \mathcal{A}_1 \\ \vdots \\ \mathcal{A}_R \end{pmatrix} \in \mathbb{R}^{2N+R+3}.$$

Puisque $U_l \neq 0$ et $\dfrac{\partial h}{\partial T} \neq 0$, le rang de la matrice $E_0(X_0)$ vaut $N+2$. Ainsi,

$$\tilde{f}_0(X_0) = \begin{pmatrix} y_1 - K_1(T,P)\,x_1 \\ \vdots \\ y_N - K_N(T,P)\,x_N \\ \sum_{i=1}^{N}(x_i - y_i) \\ \mathcal{A}_1 \\ \vdots \\ \mathcal{A}_R \end{pmatrix} \in \mathbb{R}^{N+R+1}. \tag{3.57}$$

De plus, on a

$$g_0(X_0) = \begin{pmatrix} -V \\ -V\,y \\ Q - V\,H \end{pmatrix} \in \mathbb{R}^{N+2}.$$

Soit $\mathcal{P}_1 \in \mathbb{R}^{(2N+R+3)\times(2N+R+3)}$ la matrice de permutation telle que

$$\mathcal{P}_1\, X_0 = \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \text{ où } X_1 = \begin{pmatrix} \xi \\ U_l \\ x_1 \\ \vdots \\ x_{N_R} \\ T \\ V \end{pmatrix} \in \mathbb{R}^{N+2} \text{ et } Y_1 = \begin{pmatrix} x_{N-R} \\ \vdots \\ x_N \\ y \end{pmatrix} \in \mathbb{R}^{N+R+1}.$$

D'une part, par dérivation de (3.57) par rapport à Y_1, on a

$$J_{Y_1}\left(\bar{f}_0\right) = \begin{bmatrix} \gamma_1 & I_N \\ \gamma_3 & \gamma_4 \end{bmatrix} \in \mathbb{R}^{(N+R+1)\times(N+R+1)}, \tag{3.58}$$

avec

$$\gamma_1 = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ -K_{N-R}(T,P) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -K_N(T,P) \end{bmatrix} \in \mathbb{R}^{N\times(R+1)},$$

$$\gamma_3 = \begin{bmatrix} 1 & \cdots & 1 \\ \dfrac{\partial \mathcal{A}_1}{\partial x_{N-R}} & \cdots & \dfrac{\partial \mathcal{A}_1}{\partial x_N} \\ \vdots & & \vdots \\ \dfrac{\partial \mathcal{A}_R}{\partial x_{N-R}} & \cdots & \dfrac{\partial \mathcal{A}_R}{\partial x_N} \end{bmatrix} \in \mathbb{R}^{(R+1)\times(R+1)}$$

et

$$\gamma_4 = \begin{bmatrix} -1 & \cdots & -1 \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{(R+1)\times N}.$$

On établit aisément que $\gamma_3 - \gamma_4\,\gamma_1 = \Gamma \in \mathbb{R}^{(R+1)\times(R+1)}$, où Γ est une matrice inversible par hypothèse, définie par (3.17). Puisque Γ est le complément de Schur de la matrice $J_{Y_1}\left(\bar{f}_0\right)$ relativement au bloc I_N, on en déduit que $J_{Y_1}\left(\bar{f}_0\right)$ est inversible. De ce fait, on est capable d'appliquer le théorème des fonctions implicites au système algébrique (3.57). En d'autres termes, il existe des fonctions $\varphi_1 : \mathcal{I}_0 \subset \mathbb{R}^{N+2} \to \mathbb{R}^{N+R+1}$ et $\phi_j : \mathcal{I} \subset \mathbb{R}^{N-R} \to \mathbb{R}$ pour tout $j \in [\![1, R+1]\!]$ telles que :

$$Y_1 = \varphi_1\left(X_1\right) \Leftrightarrow \begin{pmatrix} x_{N-R} \\ \vdots \\ x_N \\ y_1 \\ \vdots \\ y_{N_R} \\ y_{N-R} \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \phi_1(x_1,\ldots,x_{N_R},T) \\ \vdots \\ \phi_{R+1}(x_1,\ldots,x_{N_R},T) \\ K_1(T,P)\,x_1 \\ \vdots \\ K_{N_R}(T,P)\,x_{N_R} \\ K_{N-R}(T,P)\,\phi_1(x_1,\ldots,x_{N_R},T) \\ \vdots \\ K_N(T,P)\,\phi_{R+1}(x_1,\ldots,x_{N_R},T) \end{pmatrix}. \tag{3.59}$$

D'autre part, par dérivation de (3.57) par rapport à X_1, on obtient

$$J_{X_1}\left(\bar{f}_0\right) = \begin{bmatrix} 0 & \delta_2 \\ 0 & \delta_4 \end{bmatrix} \in \mathbb{R}^{(N+R+1)\times(N+R+1)}, \tag{3.60}$$

où

$$\delta_2 = \begin{bmatrix} -K_1(T,P) & \cdots & 0 & -\dfrac{\partial K_1(T,P)}{\partial T}\, x_1 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & -K_{N_R}(T,P) & -\dfrac{\partial K_{N_R}(T,P)}{\partial T}\, x_{N_R} & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & -\dfrac{\partial K_N(T,P)}{\partial T}\, x_N & 0 \end{bmatrix} \in \mathbb{R}^{N\times(N-R+1)}$$

et

$$\delta_4 = \begin{bmatrix} 1 & \cdots & 1 & 0 & 0 \\ \dfrac{\partial \mathcal{A}_1}{\partial x_1} & \cdots & \dfrac{\partial \mathcal{A}_1}{\partial x_{N_R}} & \dfrac{\partial \mathcal{A}_1}{\partial T} & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ \dfrac{\partial \mathcal{A}_R}{\partial x_1} & \cdots & \dfrac{\partial \mathcal{A}_R}{\partial x_{N_R}} & \dfrac{\partial \mathcal{A}_R}{\partial T} & 0 \end{bmatrix} \in \mathbb{R}^{(R+1)\times(N-R+1)}.$$

Par ailleurs, on obtient

$$A_0(X_1, Y_1) = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}, \tag{3.61}$$

avec

$$\alpha_1 = \begin{bmatrix} -\sum_{i=1}^{N} \nu_{i,1} & \cdots & -\sum_{i=1}^{N} \nu_{i,R} & 1 \\ -\nu_{1,1} & \cdots & -\nu_{1,R} & x_1 \\ \vdots & & \vdots & \vdots \\ -\nu_{N_R,1} & \cdots & -\nu_{N_R,R} & x_{N_R} \end{bmatrix} \in \mathbb{R}^{(N-R)\times(R+1)},$$

$$\alpha_2 = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 \\ U_l & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & U_l & 0 & 0 \end{bmatrix} \in \mathbb{R}^{(N-R)\times(N-R+1)},$$

$$\alpha_3 = \begin{bmatrix} -\nu_{N-R,1} & \cdots & -\nu_{N-R,R} & x_{N_R} \\ \vdots & & \vdots & \vdots \\ -\nu_{N,1} & \cdots & -\nu_{N,R} & x_N \\ Q_{r_1} & \cdots & Q_{r_R} & h \end{bmatrix} \in \mathbb{R}^{(R+2)\times(R+1)}$$

et

$$\alpha_4 = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 \\ \dfrac{\partial h}{\partial x_1} U_l & \cdots & \dfrac{\partial h}{\partial x_{N_R}} U_l & \dfrac{\partial h}{\partial T} U_l & 0 \end{bmatrix} \in \mathbb{R}^{(R+2)\times(N-R+1)}$$

ainsi que

$$B_0(X_1, Y_1) = \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+R+1)}, \tag{3.62}$$

où

$$\beta_3 = \begin{bmatrix} U_l & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_l \\ \dfrac{\partial h}{\partial x_{N-R}} U_l & \cdots & \dfrac{\partial h}{\partial x_N} U_l \end{bmatrix} \in \mathbb{R}^{(R+2)\times(R+1)}.$$

Déterminons à présent la matrice $E_1(X_1)$ à partir des matrices (3.61), (3.62), (3.58) et (3.60) :

$$\begin{aligned} E_1(X_1) &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} \gamma_1 & \gamma_2 \\ \gamma_3 & \gamma_4 \end{bmatrix}^{-1} \begin{bmatrix} 0 & \delta_2 \\ 0 & \delta_4 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} \bar\gamma_1 & \bar\gamma_2 \\ \bar\gamma_3 & \bar\gamma_4 \end{bmatrix} \begin{bmatrix} 0 & \delta_2 \\ 0 & \delta_4 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} 0 & \bar\gamma_1\,\delta_2 + \bar\gamma_2\,\delta_4 \\ 0 & \bar\gamma_3\,\delta_2 + \bar\gamma_4\,\delta_4 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & \beta_3\,(\bar\gamma_1\,\delta_2 + \bar\gamma_2\,\delta_4) \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \bar\alpha_4 \end{bmatrix}, \end{aligned}$$

où $\bar\alpha_4 = \alpha_4 - \beta_3\,(\bar\gamma_1\,\delta_2 + \bar\gamma_2\,\delta_4)$. On détermine ensuite les expressions $\bar\gamma_1$ et $\bar\gamma_2$ en utilisant le complément de Schur de la matrice $J_{Y_1}\left(\bar f_0\right)$:

$$\bar\gamma_1 = -\Gamma^{-1}\,\gamma_4 \in \mathbb{R}^{(R+1)\times N} \text{ et } \bar\gamma_2 = \Gamma^{-1} \in \mathbb{R}^{(R+1)\times(R+1)}.$$

L'expression de $\bar\alpha_4$ devient :

$$\bar\alpha_4 = \alpha_4 - \beta_3\,\Gamma^{-1}(\delta_4 - \gamma_4\,\delta_2).$$

Explicitions davantage l'expression précédente ; tout d'abord

$$\begin{aligned} \delta_4 - \gamma_4\,\delta_2 &= \begin{bmatrix} 1 & \cdots & 1 & 0 & 0 \\ \dfrac{\partial \mathcal{A}_1}{\partial x_1} & \cdots & \dfrac{\partial \mathcal{A}_1}{\partial x_{N_R}} & \dfrac{\partial \mathcal{A}_1}{\partial T} & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ \dfrac{\partial \mathcal{A}_R}{\partial x_1} & \cdots & \dfrac{\partial \mathcal{A}_R}{\partial x_{N_R}} & \dfrac{\partial \mathcal{A}_R}{\partial T} & 0 \end{bmatrix} - \begin{bmatrix} K_1 & \cdots & K_{N_R} & \nabla_T K \cdot x & 0 \\ 0 & \cdots & 0 & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{D}_1 & \cdots & \mathcal{D}_{N_R} & -\nabla_T K \cdot x & 0 \\ \dfrac{\partial \mathcal{A}_1}{\partial x_1} & \cdots & \dfrac{\partial \mathcal{A}_1}{\partial x_{N_R}} & \dfrac{\partial \mathcal{A}_1}{\partial T} & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ \dfrac{\partial \mathcal{A}_R}{\partial x_1} & \cdots & \dfrac{\partial \mathcal{A}_R}{\partial x_{N_R}} & \dfrac{\partial \mathcal{A}_R}{\partial T} & 0 \end{bmatrix} \in \mathbb{R}^{(R+1)\times(N-R+1)}, \end{aligned}$$

où $K = K(T,P)$ et $K_i = K_i(T,P)$ pour tout $i \in [\![1, N_R]\!]$. En posant $\left(\Gamma^{-1}\right)_{i,j} = \Gamma^{-1}_{i,j}$ pour tout $(i,j) \in [\![1, R+1]\!]^2$, on obtient après calculs

$$\bar{\alpha}_4 = \begin{bmatrix} -\mathbb{A}_{1,1} & \cdots & -\mathbb{A}_{1,N_R} & -\mathbb{T}_1 & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ -\mathbb{A}_{R+1,1} & \cdots & -\mathbb{A}_{R+1,N_R} & -\mathbb{T}_{R+1} & 0 \\ \dfrac{\partial h}{\partial x_1} U_l - \mathbb{H}_1 & \cdots & \dfrac{\partial h}{\partial x_{N_R}} U_l - \mathbb{H}_{N_R} & \dfrac{\partial h}{\partial T} U_l - \mathbb{S} & 0 \end{bmatrix} \in \mathbb{R}^{(R+2)\times(N-R+1)}.$$

La première étape de la méthode de déflation s'achève en donnant

$$E_1(X_1) = \begin{bmatrix} \mathcal{N}_a & 0 \\ \mathcal{N}_b & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)},$$

$$X_1 = \begin{pmatrix} \xi \\ U_l \\ x_1 \\ \vdots \\ x_{N_R} \\ T \\ V \end{pmatrix} \in \mathbb{R}^{N+2}$$

et

$$f_1(X_1) = \begin{pmatrix} -V \\ -V\, y \\ Q - V\, H \end{pmatrix} \in \mathbb{R}^{N+2}.$$

On regarde maintenant le problème quasi-linéaire $E_1(X_1)\,\dot{X}_1 = f_1(X_1)$. On transforme la matrice coefficient $E_1(X_1)$ à l'aide du complément de Schur :

$$E_1(X_1) = \begin{bmatrix} I_{N+1} & 0 \\ \mathcal{N}_b \mathcal{N}_a^{-1} & 1 \end{bmatrix} \begin{bmatrix} \mathcal{N}_a & 0 \\ 0 & 0 \end{bmatrix}.$$

Ainsi, en écrivant $f_1(X_1) = \begin{pmatrix} -V \begin{pmatrix} 1 \\ y \end{pmatrix} \\ Q - V\, H \end{pmatrix}$, on est amené à étudier le problème

$$\begin{bmatrix} \mathcal{N}_a & 0 \\ 0 & 0 \end{bmatrix} \dot{X}_1 = \begin{pmatrix} -V \begin{pmatrix} 1 \\ y \end{pmatrix} \\ Q - V \left(H - \mathcal{N}_b \mathcal{N}_a^{-1} \begin{pmatrix} 1 \\ y \end{pmatrix} \right) \end{pmatrix}.$$

Sans permutation des inconnues de X_1, on pose :

$$X_2 = \begin{pmatrix} \xi \\ U_l \\ x_1 \\ \vdots \\ x_{N_R} \\ T \end{pmatrix} \in \mathbb{R}^{N+2} \text{ et } Y_2 = V.$$

Puisque $H - \mathcal{N}_b \mathcal{N}_a^{-1} \begin{pmatrix} 1 \\ y \end{pmatrix} \neq 0$ par hypothèse, il est possible d'extraire une expression de V ; il existe une fonction $\varphi_2 : \mathcal{I}_1 \subset \mathbb{R}^{N+1} \to \mathbb{R}$ telle que

$$Y_2 = \varphi_2(X_2) \Leftrightarrow V = \left(H - \mathcal{N}_b \mathcal{N}_a^{-1} \begin{pmatrix} 1 \\ y \end{pmatrix} \right)^{-1} Q.$$

Comme la matrice $B_2(X_2, Y_2)$ est nulle, on a $E_2(X_2) = \mathcal{N}_a$. La méthode de déflation s'achève ainsi. On obtient l'EDO

$$\begin{pmatrix} \dot{\xi} \\ \dot{U}_l \\ \dot{x}_1 \\ \vdots \\ \dot{x}_{N_R} \\ \dot{T} \end{pmatrix} = -\mathcal{N}_a^{-1} \left(H - \mathcal{N}_b \mathcal{N}_a^{-1} \begin{pmatrix} 1 \\ y \end{pmatrix} \right)^{-1} Q \begin{pmatrix} 1 \\ y \end{pmatrix},$$

avec

$$\begin{pmatrix} x_{N-R} \\ \vdots \\ x_N \\ y_1 \\ \vdots \\ y_{N_R} \\ y_{N-R} \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \phi_1(x_1, \ldots, x_{N_R}, T) \\ \vdots \\ \phi_{R+1}(x_1, \ldots, x_{N_R}, T) \\ K_1(T,P)\, x_1 \\ \vdots \\ K_{N_R}(T,P)\, x_{N_R} \\ K_{N-R}(T,P)\, \phi_1(x_1, \ldots, x_{N_R}, T) \\ \vdots \\ K_N(T,P)\, \phi_{R+1}(x_1, \ldots, x_{N_R}, T) \end{pmatrix} \text{ et } V = \left(H - \mathcal{N}_b \mathcal{N}_a^{-1} \begin{pmatrix} 1 \\ y \end{pmatrix} \right)^{-1} Q.$$

La méthode de déflation est exécutée en deux étapes : le problème est une EDA d'indice 2. □

Remarque 25

- *Les activités chimiques A_j sont modélisées par des équations purement algébriques. Elles fournissent donc des équations supplémentaires reliant les fractions molaires en phase liquide x_i entre elles. Il est donc cohérent d'obtenir davantage d'expressions implicites décrivant ces fractions molaires.*
- *À partir de la fin de la première étape, les fractions molaires en phase vapeur y_i sont normalement exprimées en fonction de $T, P, x_1, \ldots, x_{N_R}$. Afin de ne pas alourdir les notations, nous laissons l'expression y, mais il faut clairement la considérer comme une fonction implicite des variables précédentes.*

3.4.2.2 Phase liquide non idéale et phase vapeur idéale

On considère le modèle réactif suivant :

$$\begin{cases} \dot{U}_l - \displaystyle\sum_{j=1}^{R} \left(\sum_{i=1}^{N} \nu_{i,j} \right) \dot{\xi}_j = -V \\ \dot{\overline{x_i\, U_l}} - \displaystyle\sum_{j=1}^{R} \nu_{i,j}\, \dot{\xi}_j = -V\, y_i, \quad \forall i \in [\![1, N]\!] \\ \dot{\overline{h\, U_l}} + \displaystyle\sum_{j=1}^{R} Q_{r_j}\, \dot{\xi}_j = Q - V\, H \\ 0 = y_i - K_i(T, P, x)\, x_i, \quad \forall i \in [\![1, N]\!] \\ 0 = \displaystyle\sum_{i=1}^{N} (x_i - y_i) \\ 0 = \mathcal{A}_j, \quad \forall j \in [\![1, R]\!]. \end{cases} \tag{3.63}$$

On utilise les notations définies par (3.3), (3.16), (3.18), (3.20), (3.22), (3.24), (3.26), (3.28) et (3.30).

Théorème 28

Supposons que $U_l \neq 0$, $\dfrac{\partial h}{\partial T} \neq 0$, $\det \bar{\Gamma} \neq 0$, $\det \bar{\mathcal{N}_a} \neq 0$ *et* $H - \bar{\mathcal{N}_b} \bar{\mathcal{N}_a}^{-1} \begin{pmatrix} 1 \\ y \end{pmatrix} \neq 0$. *Alors il existe des fonctions* $\phi_j : \mathcal{I} \subset \mathbb{R}^{N-R} \to \mathbb{R}$ *pour tout* $j \in [\![1, R+1]\!]$ *telles que l'EDA* (3.63) *soit équivalente au système différentiel*

$$\begin{pmatrix} \dot{\xi} \\ \dot{U}_l \\ \dot{x}_1 \\ \vdots \\ \dot{x}_{N_R} \\ \dot{T} \end{pmatrix} = -\bar{\mathcal{N}_a}^{-1} \left(H - \bar{\mathcal{N}_b} \bar{\mathcal{N}_a}^{-1} \begin{pmatrix} 1 \\ y \end{pmatrix} \right)^{-1} Q \begin{pmatrix} 1 \\ y \end{pmatrix},$$

avec

$$\begin{pmatrix} x_{N-R} \\ \vdots \\ x_N \\ y_1 \\ \vdots \\ y_{N_R} \\ y_{N-R} \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \phi_1(x_1, \ldots, x_{N_R}, T) \\ \vdots \\ \phi_{R+1}(x_1, \ldots, x_{N_R}, T) \\ K_1(T, P, x)\, x_1 \\ \vdots \\ K_{N_R}(T, P, x)\, x_{N_R} \\ K_{N-R}(T, P, x)\, \phi_1(x_1, \ldots, x_{N_R}, T) \\ \vdots \\ K_N(T, P, x)\, \phi_{R+1}(x_1, \ldots, x_{N_R}, T) \end{pmatrix} \quad \text{et } V = \left(H - \bar{\mathcal{N}_b} \bar{\mathcal{N}_a}^{-1} \begin{pmatrix} 1 \\ y \end{pmatrix} \right)^{-1} Q.$$

Dans ces conditions, le problème (3.56) *est une EDA d'indice 2.*

Preuve - Le début de la preuve est similaire à celle du cas idéal, jusqu'au calcul de la matrice Jacobienne $J_{Y_1}\left(\bar{f}_0\right)$. On écrit ici

$$\tilde{f}_0(X_0) = \begin{pmatrix} y_1 - K_1(T,P,x)\,x_1 \\ \vdots \\ y_N - K_N(T,P,x)\,x_N \\ \sum_{i=1}^{N}(x_i - y_i) \\ \mathcal{A}_1 \\ \vdots \\ \mathcal{A}_R \end{pmatrix} \in \mathbb{R}^{N+R+1}. \tag{3.64}$$

D'une part, par dérivation de (3.64) par rapport à Y_1, on a

$$J_{Y_1}\left(\bar{f}_0\right) = \begin{bmatrix} \gamma_1 & I_N \\ \gamma_3 & \gamma_4 \end{bmatrix} \in \mathbb{R}^{(N+R+1)\times(N+R+1)}, \tag{3.65}$$

avec

$$\gamma_1 = \begin{bmatrix} -\dfrac{\partial K_1(T,P,x)}{\partial x_{N-R}}\,x_1 & \cdots & -\dfrac{\partial K_1(T,P,x)}{\partial x_N}\,x_1 \\ \vdots & & \vdots \\ -\dfrac{\partial K_{N-R}(T,P,x)}{\partial x_{N-R}}\,x_{N-R} - K_{N-R}(T,P,x) & \cdots & -\dfrac{\partial K_{N-R}(T,P,x)}{\partial x_N}\,x_{N-R} \\ \vdots & \ddots & \vdots \\ -\dfrac{\partial K_N(T,P,x)}{\partial x_{N-R}}\,x_N & \cdots & -\dfrac{\partial K_N(T,P,x)}{\partial x_N}\,x_N - K_N(T,P,x) \end{bmatrix} \in \mathbb{R}^{N\times(R+1)},$$

$$\gamma_3 = \begin{bmatrix} 1 & \cdots & 1 \\ \dfrac{\partial \mathcal{A}_1}{\partial x_{N-R}} & \cdots & \dfrac{\partial \mathcal{A}_1}{\partial x_N} \\ \vdots & & \vdots \\ \dfrac{\partial \mathcal{A}_R}{\partial x_{N-R}} & \cdots & \dfrac{\partial \mathcal{A}_R}{\partial x_N} \end{bmatrix} \in \mathbb{R}^{(R+1)\times(R+1)} \text{ et } \gamma_4 = \begin{bmatrix} -1 & \cdots & -1 \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{(R+1)\times N}.$$

On établit aisément que $\gamma_3 - \gamma_4\,\gamma_1 = \bar{\Gamma} \in \mathbb{R}^{(R+1)\times(R+1)}$, où $\bar{\Gamma}$ est définie par (3.18). Cette matrice est supposée inversible. Puisque $\bar{\Gamma}$ est le complément de Schur de la matrice $J_{Y_1}\left(\bar{f}_0\right)$ relativement au bloc I_N, on en déduit que $J_{Y_1}\left(\bar{f}_0\right)$ est inversible. De ce fait, on est capable d'appliquer le théorème des fonctions implicites au système algébrique (3.57). En d'autres termes, il existe des fonctions $\varphi_1 : \mathcal{I}_0 \subset \mathbb{R}^{N+2} \to \mathbb{R}^{N+R+1}$ et $\phi_j : \mathcal{I} \subset \mathbb{R}^{N-R} \to \mathbb{R}$ pour tout $j \in [\![1, R+1]\!]$ telles que :

$$Y_1 = \varphi_1\left(X_1\right) \Leftrightarrow \begin{pmatrix} x_{N-R} \\ \vdots \\ x_N \\ y_1 \\ \vdots \\ y_{N_R} \\ y_{N-R} \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \phi_1(x_1,\ldots,x_{N_R},T) \\ \vdots \\ \phi_{R+1}(x_1,\ldots,x_{N_R},T) \\ K_1(T,P,x)\,x_1 \\ \vdots \\ K_{N_R}(T,P,x)\,x_{N_R} \\ K_{N-R}(T,P,x)\,\phi_1(x_1,\ldots,x_{N_R},T) \\ \vdots \\ K_N(T,P,x)\,\phi_{R+1}(x_1,\ldots,x_{N_R},T) \end{pmatrix}. \tag{3.66}$$

D'autre part, par dérivation de (3.64) par rapport à X_1, on obtient

$$J_{X_1}\left(\bar{f}_0\right) = \begin{bmatrix} 0 & \delta_2 \\ 0 & \delta_4 \end{bmatrix} \in \mathbb{R}^{(N+R+1)\times(N+R+1)}, \tag{3.67}$$

où

$$\delta_2 = \begin{bmatrix} -\dfrac{\partial K_1}{\partial x_1}x_1 - K_1 & \cdots & -\dfrac{\partial K_1}{\partial x_{N_R}}x_1 & -\dfrac{\partial K_1}{\partial T}x_1 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ -\dfrac{\partial K_{N_R}}{\partial x_1}x_{N_R} & \cdots & -\dfrac{\partial K_{N_R}}{\partial x_{N_R}}x_{N_R} - K_{N_R} & -\dfrac{\partial K_{N_R}}{\partial T}x_{N_R} & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ -\dfrac{\partial K_N}{\partial x_1}x_N & \cdots & -\dfrac{\partial K_N}{\partial x_{N_R}}x_N & -\dfrac{\partial K_N}{\partial T}x_N & 0 \end{bmatrix} \in \mathbb{R}^{N\times(N-R+1)},$$

avec $K_i = K_i(T, P, x)$ et

$$\delta_4 = \begin{bmatrix} 1 & \cdots & 1 & 0 & 0 \\ \dfrac{\partial \mathcal{A}_1}{\partial x_1} & \cdots & \dfrac{\partial \mathcal{A}_1}{\partial x_{N_R}} & \dfrac{\partial \mathcal{A}_1}{\partial T} & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ \dfrac{\partial \mathcal{A}_R}{\partial x_1} & \cdots & \dfrac{\partial \mathcal{A}_R}{\partial x_{N_R}} & \dfrac{\partial \mathcal{A}_R}{\partial T} & 0 \end{bmatrix} \in \mathbb{R}^{(R+1)\times(N-R+1)}.$$

Par ailleurs, on obtient

$$A_0(X_1, Y_1) = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}, \tag{3.68}$$

avec

$$\alpha_1 = \begin{bmatrix} -\sum_{i=1}^{N}\nu_{i,1} & \cdots & -\sum_{i=1}^{N}\nu_{i,R} & 1 \\ -\nu_{1,1} & \cdots & -\nu_{1,R} & x_1 \\ \vdots & & \vdots & \vdots \\ -\nu_{N_R,1} & \cdots & -\nu_{N_R,R} & x_{N_R} \end{bmatrix} \in \mathbb{R}^{(N-R)\times(R+1)},$$

$$\alpha_2 = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 \\ U_l & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & U_l & 0 & 0 \end{bmatrix} \in \mathbb{R}^{(N-R)\times(N-R+1)},$$

$$\alpha_3 = \begin{bmatrix} -\nu_{N-R,1} & \cdots & -\nu_{N-R,R} & x_{N_R} \\ \vdots & & \vdots & \vdots \\ -\nu_{N,1} & \cdots & -\nu_{N,R} & x_N \\ Q_{r_1} & \cdots & Q_{r_R} & h \end{bmatrix} \in \mathbb{R}^{(R+2)\times(R+1)}$$

et

$$\alpha_4 = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 \\ \dfrac{\partial h}{\partial x_1}U_l & \cdots & \dfrac{\partial h}{\partial x_{N_R}}U_l & \dfrac{\partial h}{\partial T}U_l & 0 \end{bmatrix} \in \mathbb{R}^{(R+2)\times(N-R+1)}$$

ainsi que

$$B_0(X_1, Y_1) = \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+R+1)}, \tag{3.69}$$

où

$$\beta_3 = \begin{bmatrix} U_l & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_l \\ \dfrac{\partial h}{\partial x_{N-R}} U_l & \cdots & \dfrac{\partial h}{\partial x_N} U_l \end{bmatrix} \in \mathbb{R}^{(R+2)\times(R+1)}.$$

Déterminons à présent la matrice $E_1(X_1)$ à partir des matrices (3.68), (3.69), (3.65) et (3.67) :

$$\begin{aligned} E_1(X_1) &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} \gamma_1 & \gamma_2 \\ \gamma_3 & \gamma_4 \end{bmatrix}^{-1} \begin{bmatrix} 0 & \delta_2 \\ 0 & \delta_4 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} \bar\gamma_1 & \bar\gamma_2 \\ \bar\gamma_3 & \bar\gamma_4 \end{bmatrix} \begin{bmatrix} 0 & \delta_2 \\ 0 & \delta_4 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} 0 & \bar\gamma_1\,\delta_2 + \bar\gamma_2\,\delta_4 \\ 0 & \bar\gamma_3\,\delta_2 + \bar\gamma_4\,\delta_4 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & \beta_3\,(\bar\gamma_1\,\delta_2 + \bar\gamma_2\,\delta_4) \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \bar\alpha_4 \end{bmatrix}, \end{aligned}$$

où $\bar\alpha_4 = \alpha_4 - \beta_3\,(\bar\gamma_1\,\delta_2 + \bar\gamma_2\,\delta_4)$. On détermine les expressions $\bar\gamma_1$ et $\bar\gamma_2$ en utilisant le complément de Schur de la matrice $J_{Y_1}\left(\bar f_0\right)$:

$$\bar\gamma_1 = -\bar\Gamma^{-1}\,\gamma_4 \in \mathbb{R}^{(R+1)\times N} \text{ et } \bar\gamma_2 = \bar\Gamma^{-1} \in \mathbb{R}^{(R+1)\times(R+1)}.$$

L'expression de $\bar\alpha_4$ devient :

$$\bar\alpha_4 = \alpha_4 - \beta_3\,\bar\Gamma^{-1}(\delta_4 - \gamma_4\,\delta_2).$$

Explicitions davantage l'expression précédente ; tout d'abord

$$\begin{aligned} \delta_4 - \gamma_4\,\delta_2 &= \begin{bmatrix} 1 & \cdots & 1 & 0 & 0 \\ \dfrac{\partial \mathcal{A}_1}{\partial x_1} & \cdots & \dfrac{\partial \mathcal{A}_1}{\partial x_{N_R}} & \dfrac{\partial \mathcal{A}_1}{\partial T} & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ \dfrac{\partial \mathcal{A}_R}{\partial x_1} & \cdots & \dfrac{\partial \mathcal{A}_R}{\partial x_{N_R}} & \dfrac{\partial \mathcal{A}_R}{\partial T} & 0 \end{bmatrix} - \begin{bmatrix} 1-\bar{\mathcal{D}}_1 & \cdots & 1-\bar{\mathcal{D}}_{N_R} & \nabla_T K\cdot x & 0 \\ 0 & \cdots & 0 & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \bar{\mathcal{D}}_1 & \cdots & \bar{\mathcal{D}}_{N_R} & -\nabla_T K\cdot x & 0 \\ \dfrac{\partial \mathcal{A}_1}{\partial x_1} & \cdots & \dfrac{\partial \mathcal{A}_1}{\partial x_{N_R}} & \dfrac{\partial \mathcal{A}_1}{\partial T} & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ \dfrac{\partial \mathcal{A}_R}{\partial x_1} & \cdots & \dfrac{\partial \mathcal{A}_R}{\partial x_{N_R}} & \dfrac{\partial \mathcal{A}_R}{\partial T} & 0 \end{bmatrix} \in \mathbb{R}^{(R+1)\times(N-R+1)}, \end{aligned}$$

où $K = K(T, P, x)$. En posant $\left(\bar{\Gamma}^{-1}\right)_{i,j} = \bar{\Gamma}^{-1}_{i,j}$ pour tout $(i,j) \in [\![1, R+1]\!]^2$, on obtient après calculs

$$\bar{\alpha}_4 = \begin{bmatrix} -\bar{\mathbb{A}}_{1,1} & \cdots & -\bar{\mathbb{A}}_{1,N_R} & -\bar{\mathbb{T}}_1 & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ -\bar{\mathbb{A}}_{R+1,1} & \cdots & -\bar{\mathbb{A}}_{R+1,N_R} & -\bar{\mathbb{T}}_{R+1} & 0 \\ \dfrac{\partial h}{\partial x_1} U_l - \bar{\mathbb{H}}_1 & \cdots & \dfrac{\partial h}{\partial x_{N_R}} U_l - \bar{\mathbb{H}}_{N_R} & \dfrac{\partial h}{\partial T} U_l - \bar{\mathbb{S}} & 0 \end{bmatrix} \in \mathbb{R}^{(R+2)\times(N-R+1)}.$$

La première étape de la méthode de déflation s'achève en donnant

$$E_1(X_1) = \begin{bmatrix} \bar{\mathcal{N}}_a & 0 \\ \bar{\mathcal{N}}_b & 0 \end{bmatrix} \in \mathbb{R}^{(N+2)\times(N+2)},$$

$$X_1 = \begin{pmatrix} \xi \\ U_l \\ x_1 \\ \vdots \\ x_{N_R} \\ T \\ V \end{pmatrix} \in \mathbb{R}^{N+2} \text{ et } f_1(X_1) = \begin{pmatrix} -V \\ -V\, y \\ Q - V\, H \end{pmatrix} \in \mathbb{R}^{N+2}.$$

On regarde maintenant le problème quasi-linéaire $E_1(X_1)\,\dot{X}_1 = f_1(X_1)$. On transforme la matrice coefficient $E_1(X_1)$ à l'aide du complément de Schur :

$$E_1(X_1) = \begin{bmatrix} I_{N+1} & 0 \\ \bar{\mathcal{N}}_b \bar{\mathcal{N}}_a^{-1} & 1 \end{bmatrix} \begin{bmatrix} \bar{\mathcal{N}}_a & 0 \\ 0 & 0 \end{bmatrix}.$$

Ainsi, en écrivant $f_1(X_1) = \begin{pmatrix} -V \begin{pmatrix} 1 \\ y \end{pmatrix} \\ Q - V\, H \end{pmatrix}$, on est amené à étudier le problème

$$\begin{bmatrix} \bar{\mathcal{N}}_a & 0 \\ 0 & 0 \end{bmatrix} \dot{X}_1 = \begin{pmatrix} -V \begin{pmatrix} 1 \\ y \end{pmatrix} \\ Q - V \left(H - \bar{\mathcal{N}}_b \bar{\mathcal{N}}_a^{-1} \begin{pmatrix} 1 \\ y \end{pmatrix} \right) \end{pmatrix}.$$

Sans permutation des inconnues de X_1, on pose :

$$X_2 = \begin{pmatrix} \xi \\ U_l \\ x_1 \\ \vdots \\ x_{N_R} \\ T \end{pmatrix} \in \mathbb{R}^{N+2} \text{ et } Y_2 = V.$$

Puisque $H - \bar{\mathcal{N}}_b \bar{\mathcal{N}}_a^{-1} \begin{pmatrix} 1 \\ y \end{pmatrix} \neq 0$ par hypothèse, il est possible d'extraire une expression de V ; il existe une fonction $\varphi_2 : \mathcal{I}_1 \subset \mathbb{R}^{N+1} \to \mathbb{R}$ telle que

$$Y_2 = \varphi_2(X_2) \Leftrightarrow V = \left(H - \bar{\mathcal{N}}_b \bar{\mathcal{N}}_a^{-1} \begin{pmatrix} 1 \\ y \end{pmatrix} \right)^{-1} Q.$$

Comme la matrice $B_2(X_2, Y_2)$ est nulle, on a

$$E_2(X_2) = \bar{\mathcal{N}}_a.$$

La méthode de déflation s'achève ainsi. On obtient l'EDO

$$\begin{pmatrix} \dot{\xi} \\ \dot{U}_l \\ \dot{x}_1 \\ \vdots \\ \dot{x}_{N_R} \\ \dot{T} \end{pmatrix} = -\bar{\mathcal{N}}_a^{-1} \left(H - \bar{\mathcal{N}}_b \bar{\mathcal{N}}_a^{-1} \begin{pmatrix} 1 \\ y \end{pmatrix} \right)^{-1} Q \begin{pmatrix} 1 \\ y \end{pmatrix},$$

avec

$$\begin{pmatrix} x_{N-R} \\ \vdots \\ x_N \\ y_1 \\ \vdots \\ y_{N_R} \\ y_{N-R} \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \phi_1(x_1, \ldots, x_{N_R}, T) \\ \vdots \\ \phi_{R+1}(x_1, \ldots, x_{N_R}, T) \\ K_1(T, P, x)\, x_1 \\ \vdots \\ K_{N_R}(T, P, x)\, x_{N_R} \\ K_{N-R}(T, P, x)\, \phi_1(x_1, \ldots, x_{N_R}, T) \\ \vdots \\ K_N(T, P, x)\, \phi_{R+1}(x_1, \ldots, x_{N_R}, T) \end{pmatrix} \text{ et } V = \left(H - \bar{\mathcal{N}}_b \bar{\mathcal{N}}_a^{-1} \begin{pmatrix} 1 \\ y \end{pmatrix} \right)^{-1} Q.$$

La méthode de déflation est exécutée en deux étapes : le problème est une EDA d'indice 2. □

Remarque 26

- *Comme pour le cas idéal, la variable y est à substituer par les expressions faisant intervenir $T, P, x_1, \ldots, x_{N_R}$ données par (3.66) à partir de la fin de la première étape. De plus, il faut comprendre que les expressions des constantes d'équilibres thermodynamiques ne dépendent plus explicitement des x_{N-R}, …, x_N, mais de $\phi_1(x_1, \ldots, x_{N_R}, T)$, …, $\phi_{R+1}(x_1, \ldots, x_{N_R}, T)$. Là encore, nous n'écrivons pas explicitement ce changement de dépendance afin de ne pas alourdir les notations.*
- *La matrice $\bar{\mathcal{N}}_b$ est supposée inversible. Cette hypothèse devient coûteuse si R est grand car on a $N_R + 1 \leqslant \operatorname{rang} \bar{\mathcal{N}}_b \leqslant N + 1$. Ceci s'applique bien évidemment à la matrice $\mathcal{N}_b$.*

3.5 Quelques remarques

3.5.1 Sommation des fractions molaires dans la phase liquide

Les trois modèles de distillation de Rayleigh (3.31), (3.44) et (3.56) (ainsi que leurs versions non idéales) ont une propriété commune : la somme des fractions molaires x_i est conservée au cours du temps. Cette propriété se retrouve aisément en manipulant les équations des modèles.

1. ***Distillation non réactive*** - On commence par sommer les N bilans matière partiels (3.31b) pour obtenir

$$U_l \sum_{i=1}^{N} \dot{x}_i + \dot{U}_l \sum_{i=1}^{N} x_i = -V \sum_{i=1}^{N} y_i.$$

En utilisant (3.31a), on a

$$U_l \sum_{i=1}^{N} \dot{x}_i - V \sum_{i=1}^{N} (x_i - y_i) = 0.$$

En considérant (3.31e), (3.31f) et $U_l \neq 0$, on parvient à

$$\sum_{i=1}^{N} \dot{x}_i = 0.$$

Ainsi, la somme des fractions molaires x_i est constante. Par définition des fractions molaires, on prend comme valeur initiale 1 et par conséquent

$$\sum_{i=1}^{N} x_i = 1.$$

Cette information peut bien entendu être utilisée afin de simplifier les expressions obtenues dans les théorèmes 21 - 24.

2. ***Réactions chimiques contrôlées par la cinétique*** - Dans ce contexte réactif, on parvient aisément à montrer que

$$U_l \sum_{i=1}^{N} \dot{x}_i + \left(\dot{U}_l + V\right) \left(\sum_{i=1}^{N} x_i - 1\right) = 0. \tag{3.70}$$

On observe à présent $\sum_{i=1}^{N} x_i$ en tant que fonction suffisamment régulière de la variable t ; on note $s(t) = \sum_{i=1}^{N} x_i$. De ce fait, si $s(0) = 1$, on obtient bien par (3.70) que $s'(0) = 0$. En dérivant une première fois (3.70), on montre de plus que $s''(0) = 0$. Ainsi, en procédant par dérivations successives, on établit en utilisant le développement de Taylor de $s(t)$ en $t = 0$ que $s(t) = s(0)$. Puisque $s(0) = 1$, on en conclut le résultat escompté.

3. ***Réactions chimiques instantanément équilibrées*** - Cette configuration réactive agit de la même manière sur la somme des fractions molaires x_i ; le schéma de la preuve est en tout point similaire.

3.5.2 Détermination initiale du débit vapeur V

Pour initialiser la valeur de V en début de régime diphasique, on prend habituellement comme approximation la valeur $(H - h)^{-1} Q$. La méthode de déflation améliore cette approximation en fournissant des expressions plus précises tenant compte du contexte d'étude :

– ***distillation non réactive idéale***

$$V = \mathcal{G}^{-1} Q;$$

– ***distillation non réactive non idéale***

$$V = \bar{\mathcal{G}}^{-1} Q;$$

– ***réactions chimiques contrôlées par la cinétique - Cas idéal***

$$V = \mathcal{G}^{-1} \mathcal{Q};$$

– ***réactions chimiques contrôlées par la cinétique - Cas non idéal***

$$V = \bar{\mathcal{G}}^{-1}\bar{\mathcal{Q}};$$

– ***réactions chimiques instantanément équilibrées - Cas idéal***

$$V = \left(H - \mathcal{N}_b\mathcal{N}_a^{-1}\begin{pmatrix}1\\ y\end{pmatrix}\right)^{-1} Q;$$

– ***réactions chimiques instantanément équilibrées - Cas non idéal***

$$V = \left(H - \bar{\mathcal{N}}_b\bar{\mathcal{N}}_a^{-1}\begin{pmatrix}1\\ y\end{pmatrix}\right)^{-1} Q.$$

Dans ce chapitre, nous avons étudié trois modèles d'EDAs quasi-linéaires modélisant des phénomènes de distillation de Rayleigh. L'application de la méthode de déflation à ces modèles a permis de mettre en avant différentes configurations physiques. Mis à part le modèle de distillation non réactif en régime monophasique qui est un problème d'indice 1, nous avons montré que toutes les autres configurations sont des EDAs d'indice 2.

En tenant compte des remarques concernant la sommation des fractions molaires x_i, il est possible d'exprimer de façon *explicite* les équations obtenues dans les théorèmes 21 - 26 puisqu'on peut écrire $x_N = 1 - \sum_{i=1}^{N-1} x_i$. De ce fait, on peut se passer d'un calcul effectif des fonctions implicites dans l'application de la méthode. Par ailleurs, la structure particulière de ces deux configurations (distillation non réactive et distillation où les réactions chimiques sont contrôlées par la cinétique) évite d'inverser dans sa globalité la matrice Jacobienne $J_{Y_1}\left(\bar{f}_0\right)$. Dans le cas réactif, il suffit d'inverser un scalaire ($\mathcal{D}_N$ ou $\bar{\mathcal{D}}_N$). Dans le cas non réactif, le calcul de l'inverse est simple (car la matrice Jacobienne est creuse) et au demeurant, il suffit de déterminer la première ligne de cette matrice. En revanche, ces simplifications sont impossibles pour la distillation où les réactions chimiques sont instantanément équilibrées. Le calcul des fonctions implicites ainsi que la détermination de l'inverse de la matrice Jacobienne sont nécessaires.

Pour terminer, soulignons le fait que la distinction entre les régimes monophasique et diphasique est nécessaire pour l'étude des modèles. En effet, il existe une discontinuité entre ces deux configurations. Dans le régime monophasique, le débit vapeur V est nul tandis que sa valeur devient strictement positive en régime diphasique. Il est donc cohérent de traiter séparément les deux régimes. La méthode de déflation prend ici son sens en décrivant deux cartes locales d'intersection nulle.

Annexe A

Pendule

A.1 Dimension 2

On présente la procédure `simulation` dans le contexte du pendule en dimension 2.

1. `sys1` décrit la deuxième configuration (deuxième point du théorème 15) :
   (a) `estEtatValide` décide si le point courant appartient ou non à cette configuration ($|x_1| > |x_2|$) ;
   (b) `calculeSolution` intègre le système (2.21) ;
   (c) `evalSolution` fournit le nouveau point à évaluer.
2. `sys2` est le miroir de `sys1` pour la première configuration ($|x_2| > |x_1|$).
3. `trouveBonSysteme` décide à quel système affecter le nouveau point calculé.
4. `simulation` procède à l'intégration sur l'intervalle choisi initialement en basculant d'une configuration à l'autre.
5. `traceSimultaion` permet d'obtenir les différents graphes.

```
###################
# Cas |x1| > |x2| #
###################

> sys1 := module()
>   export estEtatValide, calculeSolution, evalSolution :
>   estEtatValide := proc(x)
>   return evalb(abs(x[1]) >= 0.9*abs(x[2])) :
>   end :
>
>   calculeSolution := proc(t0, x)
>   local sys, IC, sol, sub :
>   sys := [diff(x2(t), t) = x4(t), diff(x4(t),t) = 9.81 - x5(t)*x2(t)] :
>   sub :=
>   x5(t) = 9.81*x2(t) + (x4(t)/x1(t))^2,
>   x3(t) = (-1/x1(t))*(x2(t)*x4(t)),
>   x1(t) = sign(x[1])*sqrt(1 - x2(t)^2) :
>   sys := subs(sub, sys) :
>   IC := [x2(t0) = x[2], x4(t0) = x[4]] :
>   sol := dsolve([op(sys), op(IC)], numeric) :
```

```
>   return [eval(sol), [sub]] :
>   end :
>
>   evalSolution := proc(solution, t1)
>   local sol, sub, solt1, etat, valautres, toutesval :
>   sol, sub := op(solution) :
>   solt1 := sol(t1) :
>   solt1 := solt1 [2..-1] :
>   valautres := subs(solt1, sub) :
>   toutesval := [op(solt1), op(valautres)] :
>   etat := subs(op(toutesval), [x1(t), x2(t), x3(t), x4(t), x5(t)]) :
>   return etat :
>   end :
> end module :

####################
# Cas |x2| > |x1| #
####################

> sys2 := module()
>
>   export estEtatValide, calculeSolution, evalSolution :
>   estEtatValide := proc(x)
>   return evalb(abs(x[2]) >= 0.9*abs(x[1])) :
>   end :
>
>   calculeSolution := proc(t0, x)
>   local sys, IC, sol, sub :
>   sys := [diff(x1(t), t) = x3(t), diff(x3(t), t) = -x5(t)*x1(t)] :
>   sub :=
>   x5(t) = 9.81*x2(t) + (x3(t)/x2(t))^2,
>   x4(t) = (-1/x2(t))*(x1(t)*x3(t)),
>   x2(t) = sign(x[2])*sqrt(1 - x1(t)^2) :
>   sys := subs(sub, sys) :
>   IC := [x1(t0) = x[1], x3(t0) = x[3]] :
>   return [eval(sol), [sub]] :
>   end :
>
>   evalSolution := proc(solution, t1)
>   local sol, sub, solt1, etat, valautres, toutesval :
>   sol, sub := op(solution) :
>   solt1 := sol(t1) :
>   solt1 := solt1 [2..-1] :
>   valautres := subs(solt1, sub) :
>   toutesval := [op(solt1), op(valautres)] :
>   etat := subs(op(toutesval), [x1(t), x2(t), x3(t), x4(t), x5(t)]) :
>   return etat :
>   end :
> end module :
```

```
> trouveBonSysteme := proc(L, etat)
> local sys :
> for sys in L
>   do  if sys :-estEtatValide(etat)
>       then return sys :
>       fi :
>   od :
> error "L'état n'est valide pour aucun système" :
> end :

#########################
# Systèmes à considérer #
#########################

> L := [sys1, sys2] :

###############
# Intégration #
###############

> simulation := proc(L, tfin, Delta, IC)
> local t0, sys, P, etat, solution :
> t0 := 0 :
> sys := trouveBonSysteme(L, IC) :
> solution := sys:-calculeSolution(t0, IC) :
> P := [] :
> etat := IC :
> while (t0 < tfin)
>   do while (t0 < tfin) and sys :-estEtatValide(etat)
>      do
>      P := [op(P), [t0, etat]] :
>      t0 := t0 + Delta :
>      etat := sys :-evalSolution(solution, t0) :
>      od :
>   sys := trouveBonSysteme(L, etat) :
>   solution := sys :-calculeSolution(t0, etat) :
>   od :
> return P :
> end :

###########
# Graphes #
###########

> traceSimulation := proc(P)
> local Coor1, Coor2, Param, Energie :
> Coor1 := map(u -> [u[1], u[2][1]], P) :
> print(plot(Coor1)) :
> Coor2 := map(u -> [u[1], u[2][2]], P) :
```

```
> print(plot(Coor2)) :
> Param := map(u -> [u[2][1], -u[2][2]], P) :
> print(plot(Param)) :
> Energie := map(u -> [u[1], 1/2*(u[2][3]^2 + u[2][4]^2) - 9.81*u[2][2]], P) :
> print(plot(Energie)) :
> end :
```

Il est possible de tracer l'énergie $E_m(t)$ du système au cours de l'intégration. Cette énergie est théoriquement constante pour le pendule (conservation de l'énergie mécanique). Pour le premier exemple (point de départ $[1, 0, 0, 0, 0]$), on observe :

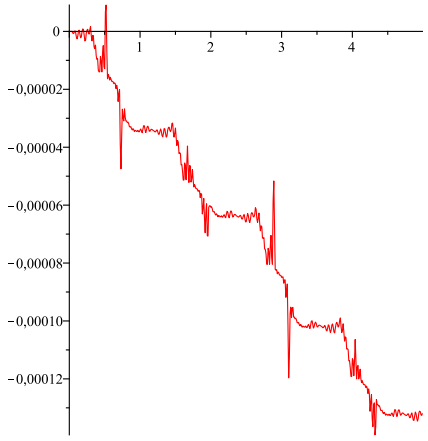


FIGURE A.1 – $E_m(t)$ en fonction de t

Il faut se référer à l'échelle pour voir que cette énergie est quasi-constante (diminution de 0.00012 sur l'intervalle $[0, 5]$). Les paliers observés correspondent à l'intégration sur une des cartes locales. Le passage d'un palier à l'autre correspond au basculement d'une carte à l'autre.

A.2 Dimension 3

On étend la procédure précédente à la dimension 3. Ici, trois systèmes sont à considérer.

```
############################
# Cas |x1| > |x2| et |x3| #
############################

> sys1 := module()
>   export estEtatValide, calculeSolution, evalSolution :
>   estEtatValide := proc(x)
>   return evalb(abs(x[1]) >= 0.9*abs(x[2]) and abs(x[1]) >= 0.9*abs(x[3])) :
>   end :
>
>   calculeSolution := proc(t0, x)
>   local sys, IC, sol, sub :
>   sys := [
>   diff(x2(t), t) = x5(t),
>   diff(x3(t), t) = x6(t),
>   diff(x5(t), t) = -x7(t)*x2(t),
```

```
>   diff(x6(t), t) = 9.81 -x7(t)*x3(t)] :
>   sub :=
>   x7(t) = 9.81*x3(t) + x4(t)^2 + x5(t)^2 + x6(t)^2,
>   x4(t) = (-1/x1(t))*(x2(t)*x5(t) + x3(t)*x6(t)),
>   x1(t) = sign(x[1])*sqrt(1 - x2(t)^2 - x3(t)^2) :
>   sys := subs(sub, sys) :
>   IC := [x2(t0) = x[2], x3(t0) = x[3], x5(t0) = x[5], x6(t0) = x[6]] :
>   sol := dsolve([op(sys), op(IC)], numeric) :
>   return [eval(sol), [sub]] :
>   end :
>
>   evalSolution := proc(solution, t1)
>   local sol, sub, solt1, etat,  valautres, toutesval :
>   sol, sub := op(solution) :
>   solt1 := sol(t1) :
>   solt1 := solt1 [2..-1] :
>   valautres := subs(solt1, sub) :
>   toutesval := [op(solt1), op(valautres)] :
>   etat := subs(op(toutesval), [x1(t), x2(t), x3(t), x4(t), x5(t), x6(t),
>   x7(t)]) :
>   return etat :
>   end :
> end module:

############################
# Cas |x2| > |x1| et |x3| #
############################

> sys2 := module()
>   export estEtatValide, calculeSolution, evalSolution :
>   estEtatValide := proc(x)
>   return evalb(abs(x[2]) >= 0.9*abs(x[1]) and abs(x[2]) >= 0.9*abs(x[3])) :
>   end :
>
>   calculeSolution := proc(t0, x)
>   local sys, IC, sol, sub :
>   sys := [
>   diff(x1(t), t) = x4(t),
>   diff(x3(t), t) = x6(t),
>   diff(x4(t), t) = -x7(t)*x1(t),
>   diff(x6(t), t) = 9.81 - x7(t)*x3(t)] :
>   sub :=
>   x7(t) = 10*x3(t) + x4(t)^2 + x5(t)^2 + x6(t)^2 ,
>   x5(t) = (-1/x2(t))*(x1(t)*x4(t) + x3(t)*x6(t)),
>   x2(t) = sign(x[2])*sqrt(1 - x1(t)^2 - x3(t)^2) :
>   sys := subs(sub, sys) :
>   IC := [x1(t0) = x[1], x3(t0) = x[3], x4(t0) = x[4], x6(t0) = x[6]] :
>   sol := dsolve([op(sys), op(IC)], numeric) :
>   return [eval(sol), [sub]] :
```

```
>   end :
>
>   evalSolution := proc(solution, t1)
>   local sol, sub, solt1, etat,  valautres, toutesval :
>   sol, sub := op(solution) :
>   solt1 := sol(t1) :
>   solt1 := solt1 [2..-1] :
>   valautres := subs(solt1, sub) :
>   toutesval := [op(solt1), op(valautres)] :
>   etat := subs(op(toutesval), [x1(t), x2(t), x3(t), x4(t), x5(t), x6(t),
>   x7(t)]) :
>   return etat :
>   end :
> end module :

############################
# Cas |x3| > |x1| et |x2| #
############################

> sys3 := module()
>   export estEtatValide, calculeSolution, evalSolution :
>   estEtatValide := proc(x)
>   return evalb(abs(x[3]) >= 0.9*abs(x[1]) and abs(x[3]) >= 0.9*abs(x[2])) :
>   end :
>
>   calculeSolution := proc(t0, x)
>   local sys, IC, sol, sub :
>   sys := [
>   diff(x1(t), t) = x4(t),
>   diff(x2(t), t) = x5(t),
>   diff(x4(t), t) = -x7(t)*x1(t),
>   diff(x5(t), t) = -x7(t)*x2(t)] :
>   sub :=
>   x7(t) = 9.81*x3(t) + x4(t)^2 + x5(t)^2 + x6(t)^2,
>   x6(t) = (-1/x3(t))*(x1(t)*x4(t) + x2(t)*x5(t)),
>   x3(t) = sign(x[3])*sqrt(1 - x1(t)^2 - x2(t)^2) :
>   sys := subs(sub, sys) :
>   IC := [x1(t0) = x[1], x2(t0) = x[2], x4(t0) = x[4], x5(t0) = x[5]] :
>   sol := dsolve([op(sys), op(IC)], numeric) :
>   return [eval(sol), [sub]] :
>   end :
>
>   evalSolution := proc(solution, t1)
>   local sol, sub, solt1, etat,  valautres, toutesval :
>   sol, sub := op(solution) :
>   solt1 := sol(t1) :
>   solt1 := solt1 [2..-1] :
>   valautres := subs(solt1, sub) :
>   toutesval := [op(solt1), op(valautres)] :
```

```
>   etat := subs(op(toutesval), [x1(t), x2(t), x3(t), x4(t), x5(t), x6(t),
>   x7(t)]) :
>   return etat :
>   end :
> end module :

> trouveBonSysteme := proc(L, etat)
> local sys :
> for sys in L
>   do  if sys :-estEtatValide(etat)
>       then return sys :
>       fi :
>   od :
> error "L'état n'est valide pour aucun système" :
> end :

#########################
# Systèmes à considérer #
#########################

> L := [sys1, sys2, sys3] :

###############
# Intégration #
###############

> simulation := proc(L, tfin, Delta, IC)
> local t0, sys, P, etat, solution :
> t0 := 0 :
> sys := trouveBonSysteme(L, IC) :
> solution := sys:-calculeSolution(t0, IC) :
> P := [] :
> etat := IC :
> while (t0 < tfin)
>   do while (t0 < tfin) and sys :-estEtatValide(etat)
>      do
>      P := [op(P), [t0, etat]] :
>      t0 := t0 + Delta :
>      etat := sys :-evalSolution(solution, t0) :
>      od :
>   sys := trouveBonSysteme(L, etat) :
>   solution := sys :-calculeSolution(t0, etat) :
>   od :
> return P :
> end :


###########
# Graphes #
```

```
############

> traceSimulation := proc(P)
> local Coor1, Coor2, Coor3, Param, Energie :
> Coor1 := map(u -> [u[1], u[2][1]], P) :
> print(plot(Coor1)) :
> Coor2 := map(u -> [u[1], u[2][2]], P) :
> print(plot(Coor2)) :
> Coor3 := map(u -> [u[1], u[2][3]], P) :
> print(plot(Coor3)) :
> Param := map(u -> u[2][1..3], P) :
> print(spacecurve(Param)) :
> Energie := map(u -> [u[1], 1/2*(u[2][4]^2 + u[2][5]^2 + u[2][6]^2) -
> 9.81*u[2][3]], P) :
> print(plot(Energie)) :
> end :
```

Comme pour la dimension 2, il est possible de tracer l'énergie $E_m(t)$ du système au cours de l'intégration. Pour le point de départ $[1, 0, 0, 0, 0.3, 4, 16.09]$, on observe :

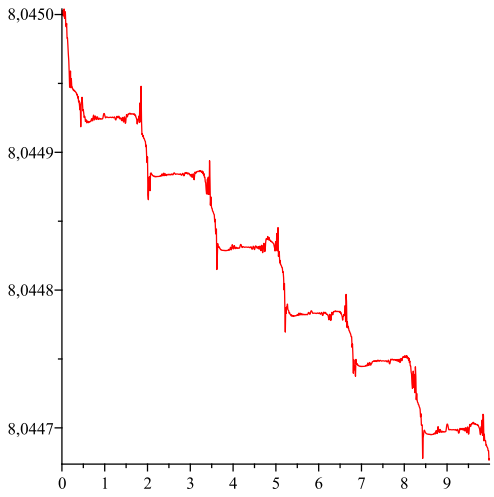


FIGURE A.2 – $E_m(t)$ en fonction de t

On peut faire le même genre de remarque que dans le cas de la dimension 2 concernant l'erreur.

Conclusion

Dans cette thèse, nous avons mis en avant une nouvelle méthode de résolution des EDAs linéaires et quasi-linéaires. Cette méthode, dite de déflation, est un processus de résolution symbolique ; elle fournit dans le cas des modèles de distillation de Rayleigh des EDOs sous contraintes. Cette réduction à des EDOs est le cœur de la méthode étudiée. En effet, à l'heure actuelle, les méthodes de résolution numérique les plus performantes concernent les EDOs. Il existe des méthodes numériques applicables aux EDAs, mais elles sont généralement mises en défaut pour des indices élevés. En réduisant les EDAs à des EDOs sous contraintes, on évite ainsi les problèmes d'intégration numérique des EDAs.

Dans le premier chapitre, nous avons parcouru les méthodes majeures de résolution des EDAs, en se focalisant sur l'aspect formel. La méthode de déflation a été introduite et étudiée dans le second chapitre. Nous avons décrit l'algorithme appliqué aux EDAs linéaires à coefficients constants et variables, puis aux EDAs quasi-linéaires. Nous avons notamment résolu les équations modélisant le mouvement du pendule en dimension 2 et 3. Afin de vérifier la validité de la méthode, la résolution formelle a été dans ce cas complétée d'une résolution numérique. Les résultats obtenus illustrent la pertinence de la méthode de déflation. Le caractère géométrique de cette dernière a été démontré par l'étude du pendule en dimension n. Ce problème est en réalité un cas particulier des systèmes mécaniques contraints à multi-corps. Dans le troisième chapitre, nous avons effectué le pré-traitement de trois modèles de distillation de Rayleigh. Cette résolution formelle a fourni à elle seule plusieurs résultats sur les systèmes étudiés. Grâce aux équations réduites, on peut notamment déterminer les conditions initiales cohérentes. Par ailleurs, on connait mieux la structure mathématique des modèles. On sait à présent que la distillation de Rayleigh non réactive n'est qu'un cas particulier de la distillation de Rayleigh aux réactions chimiques contrôlées par la cinétique. De même, d'un point de vue mathématique, le cas où la phase liquide est considérée idéale n'est qu'un cas particulier du cas où elle n'est pas idéale. On sait aussi qu'il n'est pas possible de traiter en une seule fois les phases monophasique et diphasique ; elles doivent être traitées séparément.

Plusieurs points sont encore à l'étude : nous travaillons actuellement avec K. Alloula sur la résolution numérique des équations de la distillation réduites par la méthode de déflation. Cette phase est bien plus subtile que pour le pendule car il faut pouvoir gérer convenablement les fonctions implicites au cours de l'intégration. La méthode de déflation, en tant que pré-traitement, nous a permis d'obtenir de nombreuses informations sur les modèles mis en jeu. Nous prolongeons ainsi l'étude par un traitement numérique afin d'exploiter d'un point de vue pratique les résultats fournis par la méthode de déflation. Nous souhaitons de plus rendre le traitement du processus de déflation automatique. Il est envisageable de s'appuyer sur la structure même des matrices (la matrice coefficient E, les matrices Jacobiennes intermédiaires, etc.) pour améliorer l'exécution de la méthode de déflation.

Bibliographie

[1] K. Alloula. *Modèle de coopération entre calcul formel et calcul numérique pour la simulation et l'optimisation des systèmes.* PhD thesis, Institut National Polytechnique de Toulouse, 2007.

[2] V. I. Arnold. *Mathematical methods of classical mechanics.* Springer, 1989.

[3] P. Aubry, D. Lazard, and M. Moreno Maza. On the Theories of Triangular Sets. *Journal of Symbolic Computation*, 28 :105–124, 1999.

[4] M.A. Barkatou, C. El Bacha, and E. Pflügel. Simultaneously row and column reduced higher order linear differential systems. In *International Symposium on Symbolic and Algebraic Computation*, pages 45–52, 2010.

[5] A. Ben-Israel and T. N.E. Greville. *Generalized inverses. Theory and applications.* Canadian Mathematical Society, 2003.

[6] F. Boulier. *Réécriture algébrique dans les systèmes d'équations différentielles polynomiales en vue d'application dans les sciences du vivant.* Habilitation à diriger des recherches, Université des sciences et technologies de Lille, LIFL, 2006.

[7] F. Boulier, D. Lazard, F. Ollivier, and M. Petitot. Representation for the radical of finitely generated differential ideal. In *International Symposium on Symbolic and Algebraic Computation*, pages 158–166, 1995.

[8] F. Boulier, D. Lazard, F. Ollivier, and M. Petitot. Computing representations for radicals of finitely generated differential ideals. *Applicable Algebra in Engineering, Communication and Computing*, 20(1) :73–121, 2009. (1997 Techrep. IT306 of the LIFL).

[9] F. Boulier, F Lemaire, and M. Moreno Maza. Well known theorems on triangular systems and the D^5 principle. In *Proceedings of Transgressive Computing 2006*, pages 79–91, Granada, Spain, 2006.

[10] K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical solution of initial-value problems in differential-algebraic equations.* Society for Industrial and Applied Mathematics, 1996.

[11] S.L. Campbell. Linear systems of differential equations with singular coefficients. *SIAM Journal on Mathematical Analysis*, 8 :1057–1066, 1977.

[12] S.L. Campbell. A general form for solvable linear time varying singular systems of differential equations. *SIAM Journal on Mathematical Analysis*, 18 :1101–1115, 1987.

[13] S.L. Campbell, C.D. Meyer, and N.J. Rose. Applications of the drazin inverse to linear systems of differential equations with singular constant coefficients. *SIAM Journal on Applied Mathematics*, 31 :411–425, 1976.

[14] D. Cox, John Little, and D O'Shea. *Ideals, varieties and algorithms. An introduction to computational algebraic geometry and commutative algebra.* Springer, 2007.

[15] Galileo Galilei. *Dialogue sur les deux grands systèmes du monde.* Seuil, 2000.

[16] F.R. Gantmacher. *The theory of matrices*, volume 2. American Mathematical Society Chelsea publishing, 1959.

[17] E. Griepentrog and R. März. *Differential-algebraic equations and their numerical treatment.* BSB Teubner, 1986.

[18] E. Griepentrog and R. März. Basic properties of some differential-algebraic equations. *Z. Anal. Anwendungen*, 8 :25–41, 1989.

[19] M. Günther and P. Rentrop. The differential-algebraic index concept in electric circuit simulation. *Zeitschrift für angewandte Mathematik und Mechanik*, 76 :91–94, 1996.

[20] E. Hairer, S.P. Hørsett, and G. Wanner. *Solving ordinary differential equations I. Nonstiff problems.* Springer, 1993.

[21] E. Hairer and G. Wanner. *Solving ordinary differential equations II. Stiff and differential-algebraic problems.* Springer, 2010.

[22] R.J. Hanson. Analytic linear systems of differential equations in implicit form. *Funkcialaj Ekvacioj*, 10 :123–131, 1967.

[23] W.A. Harris, Y. Sibuya, and L Weinberg. A reduction algorithm for linear differential systems. *Funkcialaj Ekvacioj*, 11 :59–67, 1968.

[24] E. Hubert. *Étude algébrique et algorithmique des singularités des équations différentielles implicites.* PhD thesis, Institut National Polytechnique de Grenoble, 1997.

[25] S. Iles. *Computational complexity of numerical solutions of initial value problems for differential algebraic equations.* PhD thesis, Faculty of Graduate Studies, The University of Western Ontario, London, Ontario, 2005.

[26] P. Kunkel and V.L. Mehrmann. *Canonical forms for linear differential-algebraic equations with variable coefficients*, volume 56. 1994.

[27] P. Kunkel and V.L. Mehrmann. Generalized inverses of differential-algebraic operators. *SIAM Journal on Matrix Analysis and Applications*, 17 :426–442, 1996.

[28] P. Kunkel and V.L. Mehrmann. *Differential-algebraic equations : analysis and numerical solution.* European Mathematical Society, 2006.

[29] F. Lemaire. *Contribution à l'algorithmique en algèbre différentielle.* PhD thesis, Université des sciences et technologies de Lille, LIFL, 2002.

[30] Y. V. Matiyasevich. *Hilbert's tenth problem.* The MIT Press, 1993.

[31] C.D. Meyer. *Matrix analysis and applied linear algebra.* Society for Industrial and Applied Mathematics, 2000.

[32] F. Monfreda and J.C. Yakoubsohn. On an index reduction method by deflation for linear differential-algebraic equations. *Submitted to Applicable Algebra in Engineering, Communication and Computing (AAECC)*, 2013.

[33] F. Monfreda and J.C. Yakoubsohn. On an index reduction method by deflation for quasi-linear differential-algebraic equations. *Preprint*, 2013.

[34] M. Moreno Maza. On Triangular Decompositions of Algebraic Varieties. Technical report, NAG Ltd, Oxford, UK, 2000. Presented at the MEGA2000 conference. Technical Report TR 4/99.

[35] R. März. On linear differential-algebraic equations and linearizations. *Applied Numerical Mathematics*, 18 :267–292, 1995.

[36] R. März. Canonical projectors for linear differential-algebraic equations. *Computers and Mathematics with applications*, 31 :121–135, 1996.

[37] N.S. Nedialkov and J.D. Pryce. Solving daes by taylor series (i) : Computing taylor coefficients. *BIT Numerical Mathematics*, pages 1–30, 2005.

[38] N.S. Nedialkov and J.D. Pryce. Solving daes by taylor series (ii) : Computing the system jacobian. *BIT Numerical Mathematics*, 2005.

[39] N.S. Nedialkov and J.D. Pryce. Solving daes by taylor series (iii) : the daets code. *Journal of Numerical Analysis, Industrial and Applied Mathematics*, 1 :1–30, 2007.

[40] F. Ollivier. Jacobi's bound and normal forms computations. a historical survey. *Arxiv preprint arXiv : 0911.2674*, 2009.

[41] C. C. Pantelides. The consistent initialization of differential-algebraic systems. *Society for Industrial and Applied Mathematics*, 9 :213–231, 1988.

[42] L. R. Petzold. A description of dassl : A differential/algebraic system solver. *R. S. Stepleman et al., editors, IMACS Trans. Scient. Comp.*, 1 :65—68, 1983.

[43] J.D. Pryce. Solving high-index daes by taylor series. *Numerical Algorithms*, 19 :195–211, 1998.

[44] J.D. Pryce. A simple structural analysis method for daes. *BIT Numerical Mathematics*, 41 :364–394, 2001.

[45] M.P. Quéré and G. Villard. An algorithm for the reduction of linear differential algebraic equation. In *International Symposium on Symbolic and Algebraic Computation*, pages 223–231, 1995.

[46] P.J. Rabier and W.C. Rheinboldt. Classical and generalized solutions of time-dependent linear differential-algebraic equations. *Linear algebra and its applications*, 245 :259–293, 1996.

[47] P.J. Rabier and W.C. Rheinboldt. *Nonholonomic motion of rigid mechanical systems from a DAE viewpoint*. Society for Industrial and Applied Mathematics, 2000.

[48] P.J. Rabier and W.C. Rheinboldt. Theoretical and numerical analysis of differential-algebraic equations. 8 :183–540, 2002.

[49] G. Reissig, W.S. Martinson, and P.I. Barton. Differential-algebraic equations of index 1 may have an arbitrarily high structural index. *SIAM Journal on Scientific Computing*, 21 :1987–1990, 2000.

[50] W.C. Rheinboldt. Differential-algebraic systems as differential equations on manifolds. *Mathematics of computation*, 43 :473–482, 1984.

[51] R. Riaza. *Differential-algebraic systems : analytical aspects and circuit applications*. World Scientific Publishing Co. Pte. Ltd., 2008.

[52] J. F. Ritt. *Differential Algebra*. American Mathematical Society, 1950.

[53] P. Rouchon. *Simulation dynamique et commande non linéaire des colonnes à distiller*. PhD thesis, École Nationale Supérieur des Mines de Paris, 1990.

[54] S. Schulz. Four lectures on differential-algebraic equations. *Department of Mathematics - Research Reports*, 497, 2003.

[55] M. Takamatsu and S. Iwata. Index reduction for differential-algebraic equations by substitution method. *Linear Algebra and its Applications*, 429 :2268–2277, 2008.

[56] G. Thomas. *Contributions théoriques et algorithmiques à l'étude des équations différentielles algébriques. Approche par le calcul formel*. PhD thesis, Institut universitaire polytechnique de Grenoble, 1997.

[57] R. Théry. *Analyse de faisabilité, synthèse et conception de procédés de distillation réactive*. PhD thesis, Institut National Polytechnique de Toulouse, 2002.

[58] J. Unger, A. Kroner, and W. Marquardt. Structural analysis of differential-algebraic equation systems. theory and applications. *Computers Chemical Engineering*, 19 :867–882, 1995.

Résumé :

Cette thèse propose d'étudier et de résoudre certaines classes d'équations différentielles algébriques (EDAs), intervenant notamment dans le domaine du génie des procédés.

Les EDAs sont des systèmes différentiels généraux qui englobent en outre les équations différentielles ordinaires. On met au point dans cette thèse une nouvelle méthode de résolution des EDAs linéaires et quasi-linéaires. Cette méthode, nommée méthode de déflation, est un processus symbolique itératif dont le but consiste à transformer une EDA, pour obtenir soit une équation différentielle sous contraintes, soit un système d'équations algébriques. La méthode de déflation est donnée par le biais d'un algorithme formel ; on analyse les propriétés de ce dernier en détail.

Le premier chapitre de cette thèse parcourt les méthodes de résolution des EDAs les plus significatives de la littérature. Ces méthodes de résolution sont présentées et illustrées. Dans le second chapitre, la méthode de déflation est décrite et analysée. On montre notamment le caractère géométrique de la méthode, à savoir qu'elle préserve la géométrie des systèmes étudiés, à travers l'étude des équations modélisant le mouvement d'un pendule simple en dimension n. La méthode de déflation est mise en pratique sur des systèmes mécaniques contraints à corps multiples. On montre également la baisse caractéristique de l' indice de Kronecker durant l'application de la méthode de déflation. Plus précisément, on prouve que l'indice de Kronecker diminue de un entre chaque étape de la méthode. Enfin, nous résolvons formellement dans le troisième chapitre des EDAs quasi-linéaires modélisant des phénomènes de distillation de Rayleigh.

Abstract :

This thesis deals with the study and the resolution of several classes of differential algebraic equations (DAEs), especially involved in the process engineering field. DAEs are general differential systems which include ordinary differential equations. We establish in this work a new resolution method for linear and quasilinear DAEs. The method, called the deflation method, is an iterative symbolic process which transforms DAEs into either constrained differential equations or algebraic equations. The deflation method is provided by a symbolic algorithm. We analyse properties of this algorithm in detail.

The first chapter of the thesis describes the most significant resolution methods of DAEs known in the actual literature. These methods are presented and illustrated. In the second chapter, the deflation method is studied. We show the geometric aspect of the deflation method (the method preserves the geometry of the studied systems) through the study of the equations of the n-pendulum. The deflation method is used on constrained multibody systems. We also show how the Kronecker index decreases during the application of the method. In the last chapter, we solve quasilinear DAEs provided by Rayleigh distillation models.